Pontifícia Universidade Católica
do Rio de Janeiro

Adrian Manresa Pérez

# Machine Learning to Predict High-cost Hospitalizations

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção.

Advisor : Prof. Fernanda Araujo Baião Amorim
Co-advisor: Prof. Silvio Hamacher

Rio de Janeiro
March 2020

Pontifícia Universidade Católica
do Rio de Janeiro

Adrian Manresa Pérez

# Machine Learning to Predict High-cost Hospitalizations

Dissertation presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção. Approved by the Examination Committee.

**Prof. Fernanda Araujo Baião Amorim**
Advisor
Departamento de Engenharia Industrial – PUC-Rio

**Prof. Silvio Hamacher**
Co-advisor
Departamento de Engenharia Industrial – PUC-Rio

**Prof. Fernando Augusto Bozza**
FIOCRUZ

**Prof. Julia Lima Fleck**
Departamento de Engenharia Industrial – PUC-Rio

Rio de Janeiro, March the 20th, 2020

**Adrian Manresa Pérez**

Bachelor in Industrial Engineering by the Technological University of Havana José Antonio Echeverría (Havana, Cuba) in 2016.

Bibliographic data

Manresa Pérez, Adrian

Machine Learning to Predict High-cost Hospitalizations / Adrian Manresa Pérez; advisor: Fernanda Araujo Baião Amorim; co-advisor: Silvio Hamacher. – 2020.

147 f.: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2020.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Custo em Saúde;. 3. Aprendizado de Máquina;. 4. Modelos Preditivos.. I. Amorim, Fernanda Araujo Baião. II. Hamacher, Silvio. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

To my parents, for their support
and encouragement.

# Acknowledgments

First of all, I wish to acknowledge the support of all my family, especially the dedication and the great love that I have always received from my parents. They give me all the confidence and comprehension I need to fight for my dreams. They are my strength, my guide, the engine of my life. Thanks to my sister for her love and happiness, and for taking care of our parents during this time that I have been away.

Then I wish to thank the most lovely, comprehensive, and intelligent woman I know: my wife. Thanks for putting a smile on my face every day, for being my friend, my inspiration, for enjoying my dreams, for making them real. Thanks for the immense help with this thesis and for always being my safe place. Thanks to my wife's family, especially to his parents and brother. They have made me part of the family since day one, always offering me their love and unconditional support.

I am so grateful to my advisors, Professor Fernanda Baião and Silvio Hamacher, for trusting me and giving me the opportunity of doing this interesting work. They have provided me excellent guidance during my master's studies, contributing to my formation, and to accomplish the goals of this thesis.

I wish to express my deepest gratitude to the Cuban family I found here. In special to Julio, Kirenia, David, Lorena, Rocio, Emilio, Daylis, Randy, Adila, and Félix, and to those that arrive later, Madaine, Raul, Gretel, and Anabel. Thanks for being my family in Rio de Janeiro and for being there for everything.

I would also like to thank to my colleagues and friends from DEI and the Tecgraf Institute, for all the moments that we share together, especially to Bianca, Leonardo, Amanda, Leila, Igor, and Iuri.

I am truly thankful to Pontifical Catholic University of Rio de Janeiro, for giving me the opportunity of pursuing my master's degree and for the excellent professors I had here.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

This study was also finaced by PUC-Rio, and by the Brazilian Agency FAPERJ under grant E-26/200.878/2019.

## Abstract

Manresa Pérez, Adrian; Amorim, Fernanda Araujo Baião (Advisor); Hamacher, Silvio (Co-Advisor). **Machine Learning to Predict High-cost Hospitalizations**. Rio de Janeiro, 2020. 147p. Dissertação de mestrado – Departamento de Engenharia Industrial, Pontífícia Universidade Católica do Rio de Janeiro.

Healthcare providers are evolving their management models, developing proactive programs to improve the quality and efficiency of their health services, considering the available historical information. Proactive strategies seek not only to prevent and detect diseases but also to enhance hospitalization outcomes. In this sense, one of the most challenging tasks is to identify which patients should be included in proactive health programs. To this end, forecasting and modeling cost-related variables are among the most widely used approaches for identifying such patients, since these variables are potential indicators of the patients' hospitalization risk, their severity, and their medical resources consumption. Most of the existing research works in this area aim to model cost variables from an overall perspective and predict cost variations for specific periods. In contrast, this work focuses on predicting the costs of a particular event. Specifically, this thesis prescribes a solution for identifying high-cost hospitalizations, to support health service managers in their proactive actions. To this end, the Design Science Research (DSR) methodology was combined with the Data Science life cycle in a real scenario of a health consulting company. The data provided describes patients' hospitalizations through their demographic characteristics and their medical resource consumption. Different statistical and Machine Learning techniques were used to predict high-cost hospitalizations, such as Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression Trees (CART), Random Forest (RF), and Extreme Gradient Boosting (XGB). The experimental results showed that RF and XGB presented the best performance, reaching an Area Under the Curve Precision-Recall (AUCPR) of 0.732 and 0.644, respectively. In the case of RF, the model was able to detect, on average, 72% of the high-cost hospitalizations with a 33% of Precision, which represents 78.7% of the total cost generated by the high-cost hospitalizations. Moreover, the obtained results showed that the use of prior cost and aggregated variables of resource consumption increased the model's ability to predict high-cost hospitalizations.

## Keywords

Healthcare Cost; Machine Learning; Predictive Model.

# Resumo

Manresa Pérez, Adrian; Amorim, Fernanda Araujo Baião; Hamacher, Silvio. **Aplicação de Técnicas de Aprendizado de Máquina para a Predição de Internações de Alto Custo**. Rio de Janeiro, 2020. 147p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Empresas do ramo da Saúde vêm evoluindo seus modelos de gestão, desenvolvendo programas proativos para melhorar a qualidade e a eficiência dos seus serviços considerando informações históricas. Estratégias proativas buscam prevenir e detectar doenças precocemente e também melhorar os resultados das internações. Nesse sentido, uma tarefa desafiadora é identificar quais pacientes devem ser incluídos em programas proativos de saúde. Para isso, a previsão e a modelagem de variáveis relacionadas aos custos estão entre as abordagens mais amplamente utilizadas, uma vez que essas variáveis sào potenciais indicadores do risco, da gravidade e do consumo de recursos médicos de uma internação. A maioria das pesquisas nesta área têm como foco modelar variáveis de custo em uma perspectiva geral e prever variações de custos para períodos específicos. Por outro lado, este trabalho se concentra na previsão dos custos de um evento específico. Em particular, esta dissertação prescreve uma solução para a predição de internações de alto custo, visando dar apoio a gestores de serviços em saúde em suas ações proativas. Para esse fim, foi seguida a metodologia de pesquisa Design Science Research (DSR), aliada ao ciclo de vida de projeto de Ciência de Dados, sobre um cenário real de uma empresa de consultoria em saúde. Os dados fornecidos descrevem internações de pacientes através de suas características demográficas e do histórico de consumo de recursos médicos. Diferentes técnicas estatísticas e de Aprendizado de Máquina foram aplicadas, como Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression Trees (CART), Random Forest (RF) e Extreme Gradient Boosting (XGB). Os resultados experimentais evidenciaram que as técnicas RF e XGB apresentaram o melhor desempenho, atingindo AUCPR de 0,732 e 0,644, respectivamente. O modelo de predição da técnica RF foi capaz de detectar até 72%, em média, das internações de alto custo com 33% de precisão, o que representa 78,7% do custo total gerado por tais internações. Além disso, os resultados monstraram que o uso de custo prévio e variáveis agregadas de consumo de recursos aumentaram a capacidade de predição do modelo.

# Palavras-chave

Custo em Saúde; Aprendizado de Máquina; Modelos Preditivos.

# Table of contents

# List of figures

## List of tables

# List of Abreviations

AUROC – Area Under the Receiver Operating Characteristics

AUCPR – Area Under the Precision Recall curve

CART – Classification and Regression Tree

CRISP-DM – Cross Industry Standard Process for Data Mining

CV – Cross-Validation

DMAIC – Define, Measure, Analyze, Improve, Control

DS – Design Science

DSR – Design Science Research

FP – False Positive

FPR – False Positive Rate

FN – False Negative

GBM – Gradient Boosting Machine

KNNs – K-Nearest Neighbors

KDD – Knowledge Discovery in Databases

KNNs – K-nearest neighbors

LASSO – Least Absolute Shrinkage and Selection Operator

LR – Logistic Regression

ML – Machine Lerning

MCC – Matthews Correlation Coefficient

MLE – Maximum Likelihood Estimation

PPV – Positive Predicted Value

PR – Precision-Recall

RF – Random Forest

ROC – Receiver Operating Characteristic

ROSE – Random Over Sampling Examples

RR – Ridge Regression

SEMMA – Sample, Explore, Modify, Model, Assess

SMOTE – Synthetic Minority Over-Sampling Technique

TN – True Negative

TNR – True Negative Rate

TP – True Positive

TPR – True Positive Rate

XGB – Extreme Gradient Boosting

*I believe in intuitions and inspirations. I sometimes feel that I am right. I do not know that I am.*

**Albert Einstein**

# 1 Introduction

Organizations have been serving as sources of large amounts of data with the capability to transform their entire business processes. This data provides insights on processes behaviors and could serve as a guide to create business value and to assist decision-making processes [3]. Particularly the Healthcare industry manages an extensive amount of data from a wide variety of sources, such as traditional interviews, clinical and laboratory processes, medical equipment, hospital bills, pharmacy deliveries, insurance companies, wearable devices for telemonitoring, etc. [4].

In the last decades, data-driven approaches and techniques (such as forecasting, optimization, simulation, machine learning, process mining, etc.) have increasingly become the focus of academic and corporate investigations in the Health field because of its capacity to provide enhanced visibility of the operations and to improve performance [5, 6].

In this direction, current healthcare systems face a variety of challenges, including financial demands for cost optimization and proper use of resources, aging populations, and an always increasing demand for health services. Proper use of the vast amount of data generated in health organizations is becoming the approach for meeting these challenges, and thus improve the quality and efficiency of their services [7]. Furthermore, healthcare systems are changing their vision of health models from a reactive to a proactive approach [8], with the objective of improving patient health, reduce the risk of hospitalizations, shorten the in-hospital stay, avoid readmissions and optimize the allocation of available resources. The proactive strategy attempts to create a commitment between caregivers and patients, where both take an active role in managing the patient's health. For that, the healthcare providers should design health models to prevent and early detect diseases, and also improve the hospitalization's outcomes. Moreover, the continuous increase of health expenses is one of the most critical problems in the world [9–11], especially when it is not reflected in better health but is caused by inefficient services as unnecessary medical tests, frequent readmissions and high-cost of patented prescription drugs [12, 13].

One of the most challenging tasks to address these problems is to correctly identify patients who may incur high expenses and thus should be part of a proactive health plan to receive appropriate care and to optimize the allocation of existing resources [14]. The selection of this group of people involves the analysis of a large number of variables and data available for each patient. This overwhelming amount of information makes this task a real challenge for doctors and health managers, compromising the effectiveness of this critical decision [8]. A first step in identifying the target patients could be to analyze the patient's expected cost since it has been widely used as an indicator of the patient severity, of a high complexity hospitalization, or of high resource consumption [9].

In healthcare systems, costs may refer to the overall cost of a patient during a determined period, or the specific cost a patient will generate when hospitalized. The latter refers, for example, to the value charged to a patient (or insurance company) when admitted into the hospital to perform a scheduled surgery. In both cases, cost values present high variability caused by a variety of factors, such as the type of procedure, insurance plan, hospital characteristics, patient medical record, etc. In this context, forecasting and modeling cost variables are valuable but also challenging tasks, which results are of huge importance for all actors involved. For the beneficiaries, knowing their future health expenses would allow them to make a better choice of insurance plans [15], and even guide them to adopt proactive strategies to reduce health risks. Health service providers will benefit by identifying highly complex patients and thus allow them to take preventive actions and prioritize the limited medical resources [16]. Also, the insurance providers will improve accountability and their business planning to attract new members and also manage the existing ones via preventive care [10].

The present dissertation is developed in conjunction with a Health Consulting organization that serves as a connection for diverse corporations with healthcare providers and health insurers. The organization's goals are to reduce waste of resources and enhance financial health, acting as an efficient and centralized healthcare manager. One of their strengths is the ability to customize their final products. For that, they have improved their capability to obtain the raw data generated by the involved companies and transform it into valuable information, thus aggregating value to their services.

Recently, the organization got involved in designing a proactive plan, for which the identification of suitable patients is a key issue. For the organization, there are two situations of interest to implement the named plan, with a long and short-term vision. The first, as a strategic plan, seeks to analyze the whole population within the organization and is focused, for example, on detecting and treating chronic diseases. In contrast, the second situation is more operational and emerges at the moment that a patient is already admitted to a hospital institution, which is the focus of the present research. Here the problem is to identify patients deserving special attention to avoid unexpected behaviors in their health condition and increased complexity of the medical procedure.

Thus, the problem addressed in this research is "how to detect an unexpected high-cost of a patient treatment during his/her hospitalization"? Hence, the main objective of this dissertation is to predict, at the moment of a hospitalization, which patients may incur an unexpected high-cost hospitalization.

For this purpose, statistical and Machine Learning (ML) models were considered to predict high-cost hospitalizations. The study was performed on a database of patients' hospitalizations provided by the organization, which has information about their demographic characteristics, their record of medical resource consumption for three years before the last hospital admission, and the motive and procedures of it. The complementary objectives of this research are:

- Perform a literature review to identify the main concepts and variables used in previous researches to model healthcare costs and medical resource consumption.
- Define the variable that describes unexpected cost in this research context.
- Implement statistical and ML techniques to model healthcare costs and to make predictions.

This research contributes primarily to the development of a predictive model that is both relevant and rigorous. Relevant because it prescribes a solution to a real problem in a notable context and rigorous because it has a foundation supported in the existing literature, which provides the necessary knowledge to conduct the research. Furthermore, it contributes to the academic field by addressing the cost modeling task from a different perspective, not considering the continuous cost variable in a temporal fashion, but instead forecasting whether and unexpected costs can occur in a specific event. Besides, the work method followed in this research combines the application of the Design Science Research (DSR) methodology with the Data Science life cycle in a real scenario of a health consulting organization, thus extending the boundaries of the Design Science (DS) theory and practice. Moreover, the empirical evaluation of statistical and ML techniques in a Brazilian database of patient's in-hospital admissions is a novel approach for predicting high-cost hospitalizations in the country.

Following Chapter 1, this thesis is structured as follows: Chapter 2 presents the main status and contributions in the literature to model healthcare costs and medical resource consumption; Chapter 3 describes the scientific methodology followed in this research to develop a Data Science approach for predicting high-cost hospitalizations; in Chapter 4 the results and procedure explaining the ML model's development is described; finally, Chapter 5 presents conclusions and future works.

# 2 Related Works on Modeling Healthcare Cost and Resource Consumption Variables

This chapter describes the research works found in the literature regarding the development of predictive models in the healthcare field, specifically to model cost and resource consumption data.

We first explain how the perspectives of previous investigations in this field differ from the methodology followed by the present study. Despite the different possibilities to address the problem, past work findings still contribute to the base knowledge of this dissertation. In this sense, the different types of explanatory variables used and their characteristics, the statistical and ML techniques, as well as the methods followed to evaluate them, are summarized.

The related studies reviewed were selected from two recent systematic literature reviews on the topic. Wammes et al. [17] reviewed studies on the characteristics and healthcare utilization patterns of high-cost users, whereas Morid et al. [16] analyzed supervised learning methods used for predicting healthcare costs. From a total of 60 papers analyzed by both, 14 were selected after filtering out the works that did not focus on cost prediction, i.e., the ones that just intended to describe the characteristics of high-cost patients.

## 2.1 Perspectives when modeling healthcare cost variables

As stated in Chapter 1, this thesis prescribes a solution for the demand of a Health Consulting organization, which aims to create a proactive plan that aggregates value to their services. The proposed solution seeks to model healthcare costs as the methodology to identify suitable patients that will integrate the named plan. In this context, the objective is to build a predictive model to detect high-cost hospitalizations when a patient is already admitted to a hospital, that is, the cost modeling procedure is intended for a specific event. In contrast, related works which also model healthcare costs do not look to a particular event in the patient's record, but intend to forecast whether they will increase their expenses for a defined period [9,18]. For example, Sushmita et al. [19] aim to predict health costs in four future scenarios (3, 6, 9, and 12 months). Lahiri et al. [15] predict whether or not the individual's healthcare expense will rise in the following year.

Moreover, previous investigations model the cost variable as a numeric value or create discrete variables that characterize the problem under study. The study of Duncan et al. [10] exemplify the use of the continuous cost variable, where the objective is to predict next year's total expenditures using last year's information. On the other hand, the discretization of the cost variable is a characteristic of researches aiming to identify likely patients to incur in high-expenses, where a classification model is more appropriated. Discretization approaches go from binary split to cost bucketing [11]. In the later, researchers aim to isolate high-cost hospitalizations by grouping observations in a manner that the total sum of the hospitalizations' cost within each bucket is nearly the same. The bucketing approach implemented by Guo et al. [14] intends to group patients by degree of severity considering their health expenses, and then asses the risk of a patient transition from a less costly bucket to a higher one.

## 2.2 Features used to explain healthcare cost variables

Different types of variables have been used as inputs of predictive models to explain the behavior of costs in the healthcare field. Variables encountered in this context relate to patients' medical conditions and different types of medical resource consumption, where the latter present challenging characteristics regarding their data distribution. Resource consumption data is typically sparse (mostly zero-valued indicating no use of resources), with few observations presenting extreme values. In other words, this data is characterized by highly non-Gaussian distributions, with a median value nearly to zero and a heavy right-hand tail. This poses great difficulties to classical statistical methods, which has led to novel statistical approaches and the introduction of more advanced ML techniques [19].

Clinical variables have been used to show the effect of specific characteristics and patients' conditions on cost prediction [16], and also Kuo et al. [20] tested pharmacy-related features. In addition, since the main objective of cost predicting scenarios is to propose a correct medical intervention, models with higher interpretability of the medical variables are preferred in the literature. On the other hand, there have been studies that confirm relevant results using features of prior costs to predict future expenses [11,19,21]. Cost information could give a global picture of the patient's health, capturing its behavior of health resource consumption and medical conditions. Table 2.1 shows examples of prior cost variables used in the literature.

Table 2.1: Examples of past cost variables used in the literature to predict future costs.

| Source | Past cost inputs |
|---|---|
| Duncan et al. [21] | Professional costs<br>Pharmacy costs<br>Outpatient costs<br>Inpatient costs |
| Sushmita et al. [19] | Total cost |
| Frees et al. [22] | Total cost |
| Kuo et al. [20] | Total medication cost |
| Bertsimas et al. [11] | Monthly cost<br>Total pharmacy cost<br>Total medical cost<br>Total cost<br>Total cost in the last 6 months<br>Total in last 3 month<br>Number of months above average<br>Cost of the highest month |

The work of Bertsimas et al. [11] was the earliest research found that intends to model past cost variables by creating a set of related features to capture hidden patterns in cost time series. Their objective was to reflect in the variables not only the aggregated cost value but to model behaviors of past hospitalizations, indicating whether there was a constant increase in the resource consumption or if the temporal series present abrupt changes that may indicate an acute event. Recently, Morid et al. [23] developed more advanced approaches to model this type of behavior in past hospitalizations using their corresponding cost. The authors argue that there may be two situations of interest, one where a patient shows a constant high-cost consumption, which may indicate the presence of a chronic condition and thus is likely to incur high-cost in the future. The second situation refers to patients with a low-cost profile who suddenly have a spike in their temporal cost series due to an exceptional situation (e.g., accident or pregnancy), while still maintaining a low risk of high future costs. In these conditions, relying solely on aggregated values of past costs may prevent the correct identification of these patterns. For that, they proposed a method to extract temporal patterns from a patient time series data using change point detection methods [24].

## 2.3 Statistical and Machine Learning techniques applied to health cost predictions

Different data-driven approaches to model health costs have been used in the literature, such as rule-based methods, statistical models, and machine learning techniques [16]. Rule-based models require a vast knowledge domain of the theme under study and are used mostly to create risk-based models [19]. On the other hand, statistical models and machine learning algorithms are powerful tools for detecting the relationships between the diverse variables involved and have been widely used to model healthcare costs from different perspectives.

Table 2.2 presents a review of statistical and ML techniques used in prior studies to predict high-cost hospitalizations, detailing the objectives pursued, the approach followed to define the outcome, the set of variables used to explain healthcare cost, and the evaluation metrics employed to assess predictive model performance.

In general, the objectives pursued refer to model cost from a temporal perspective, where predictive models are trained on the data related to the oldest events, and then tested in the newest set of samples. Moreover, the definition of high-cost patients is made by analyzing the cost distribution of the global population under study and defining patients included in the highest percentiles as such. On the other hand, a very heterogeneous set of predictors have been tested to explain healthcare cost, including demographic and administrative information, medical diagnosis and resource consumption, medical check-ups, insurance profiles, and self-reported health status.

A variety of predictive techniques have been employed, with the oldest researches relying on the statistical ones aiming to assess the relationship between the predictors and the outcome variable. In contrast, novel approaches use ML techniques, taking advantage of their predictive power to forecast healthcare expenditures. In the two more recent works reported in this study, the results obtained with the ANN and the RNN outperforms the other techniques [9, 18]. Finally, evaluation metrics used in the literature intends to measure model performance from an overall perspective (e.g., AUROC and Accuracy) and also from a cost view, using variations of the classic metrics specified in terms of costs (e.g., Cost Capture (CC)).

In summary, the related works found in the literature and presented in this Chapter, address the healthcare cost predictions from a perspective different from the research objective of this study. They focus on the temporal variations of health expenses and medical resource consumption. In contrast,

Table 2.2: Summary of related works on predicting high-cost in the healthcare field.

| **Paper** | **Objectives** | **Response Variable** | **Type of Variables** | **Statistical and ML Techniques** | **Evaluation Metrics** |
|---|---|---|---|---|---|
| Kim and Park [9] | Predict high-cost patients | Patients whose costs are in the upper 10% of the subsequent years expenditure distribution | Medical check-up, insurance eligibility, diagnosis, and health care utilization | Logistic Regression (LR), Random Forest (RF), Artificial Neural Network (ANN) | Area Under the Receiver Operating Characteristic Curve (AUROC), Cost Capture (CC) (equals to Sensitivity in terms of cost) |
| Yang et al. [18] | Predict future health expenditures in 4 time scenarios (1, 3, 6, 12 month) | The upper decile of the population cost distribution | Administrative insurance claims from the Medicaid program | Linear Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO), Gradient Boosting Machine (GBM), Recurrent Neural Network (RNN) | $R^2$, Root Mean Square Error (RMSE) |
| Tamang et al. [25] | Predict future high-cost patients within 1 year in the upper decile of the cost distribution | The upper decile of the population cost distribution | Demographic, healthcare utilization, procedures codes, pharmacy | LR, Elastic-net (EN) | CC |
| Chang et al. [26] | Predict patients being consistent high-cost users | Patients with plan-specific in the top 20% of medical costs across 2008 and 2009 | Demographic, costs, diagnosis codes, procedures codes, pharmacy | LR | AUROC, Sensitivity, Positive Predicted Value (PPV), Negative Predicted Value (NPV) |
| Duncan et al. [10] | Predict next-year total costs | Continuous cost variable | Demographic, procedures, claim cost | LR, LASSO, Multivariate Adaptive Regression Splines (MARS), RF, Regression Tree (RT) M5, GBM | $R^2$, Trunc $R^2$, Mean Absolute Error (MAE), Trunc MAE |
| Robst [27] | Predict future persistent high-cost cases for 5 scenarios in a 6-year time frame | Individual costs being in the top 1% of expenditure distribution | Demographic, diagnosis codes, insurance payments | LR | Accuracy |
| Boscardin et al. [28] | Predict new high cost individuals | Total annual costs within the top 10 % of the population | Demographic, self-reported health status, health services use | Step-wise multivariable LR | AUROC |
| Sushmita et al. [19] | Predict future cost value in 4 scenarios (3, 6, 9, 12 month) | Continuous cost variable | Demographic, procedures, previous cost | RT M5, RF regression | MAE, RMSE |
| IzadShenas et al. [29] | Identify high-cost patients | Patients in the top 5 percentile among the general population | Medical Expenditure Panel Survey (MEPS)* | DecisionTrees (DT) (C5.0 CHAID), ANN | Accuracy, G-mean, AUROC |
| Lahiri, Agarwal [15] | Predict whether patients may increase healthcare expenses for the subsequent year | Continuous cost variable | Demographic, procedures, pharmacy, previous cost | GBM, Conditional Inference Tree (CIT), ANN, Support Vector Machine (SVM), LR, Naive Bayes (NB) | False Negative Rate (FNR), Flse Positive Rate (FPR) |
| Fleishman and Cohen [30] | Predict whether a patient will incur in high medical expenditures fos a specific year | Patients in the upper expenditure decile | MEPS | LR | Bayesian information criterion (BIC), AUROC |
| Bertsimas et al. [11] | Classify cost bucket and estimate next year total costs | Continuous cost variable & cost bucketing (5 buckets equal total cost) | Demographic, procedures, previous cost | DT, Clustering | Hit Ratio, Absolute Prediction Error, $R^2$ |
| Cohen et al. [31] | Predict the likelihood of incurring high levels of medical expenditures in a subsequent year | Patients in the top decile of the expenditure distribution | MEPS | LR | Accuracy, Sensitivity, Specificity |
| Crawford et al. [32] | Predict increase of medical costs for specific diseases in a year | Patients in the top 15% of the expenditure distribution | Demographic, healthcare costs, medical events | ANN | AUROC, Sensitivity, Specificity |

*: https://www.ahrq.gov/data/meps.html

as stated in the Introduction, this research aims to identify whether a patient would incur an unexpected high-cost hospitalization; thus, the focus is on a specific event. Moreover, past studies use as response variable either the continuous cost variable or discretizes it to label high-cost patients. The discretization approaches consider the entire population for the high-cost definition; however, this study refines this definition considering high-cost hospitalizations within groups of medical procedures with common characteristics.

# 3
# Scientific Methodology and Theoretical Foundations

This chapter describes the scientific methodology followed in this research. A detailed description of the steps performed to conduct scientific research is provided. Moreover, the work method adopted to guide the study is outlined, including theoretical foundations on the techniques applied.

## 3.1
## DS Paradigm

There are different paradigms to guide scientific research which can be classified according to their purpose and research goals [33]. On the one hand, Natural and Social Sciences (known as explanatory science [34]) have as their mission the search for the truth, aiming to describe, explain, and predict to advance the knowledge in a given area [35]. Natural Sciences are those motivated to understand complex phenomena, to discover and to explain their behavior (e.g., Physics, Chemistry, and Biology); while Social Sciences seek to describe and reflect on human beings, their actions and social relations (e.g., Anthropology, Economics, and Politics).

On the other hand, the Design Science(DS) paradigm emerges to guide research studies aiming to find solutions to given problems or to design and create artifacts that improve the daily routine of professionals [33]. DS prescribes solutions to real problems in different domains (e.g., Medicine, Engineering, and Management), thus reducing the existing gap between theory and practice [34]. Furthermore, DS emphasizes the connection between knowledge and practice by showing that it is possible to produce scientific knowledge by designing useful artifacts. In this sense this paradigm is intended for situation where studies focused in the design, conception and problem solving cannot rely solely on the paradigms of natural and social sciences. This limitation occurs because the goals of traditional science are to explore, to describe and to explain, but in this scenario the objectives are to prescribe solutions and methods for solving a given problem or designing new artifacts [34]. In this context, the DS paradigm was adopted in this research work as the scientific methodology to guide the development of a Machine Learning (ML) model to automatically predict high-cost hospitalizations. In order to guarantee the reliability of results, the main points to conduct this research work are described as follows, following the structure proposed by Dresch et al. [33].

The reasons and objectives of the research are the starting point and, in this context, the motivation comes from the need to give a solution to a practical problem, and it is concerned with prescribing solutions and designing artifacts. These steps were defined in the introduction, establishing the motivation, study problem, and the research goals.

The next step defines the scientific methods that will help to reach the proposed goals. The abductive method is considered a creative process, searching for explanatory hypotheses to a given phenomenon or situation, thus allowing the introduction of a new idea [33]. Research under the DS paradigm relies on the abductive scientific method when proposing solutions; however, it is not restricted to it. For example, when the researcher uses previous knowledge to build and evaluate the artifact, the deductive method [36] could be more suitable. Then, the research conducted could be guided by more than one scientific method, depending on the current step and goals being developed.

Once the scientific approach is established, it is time to define the research method, which ensures that the investigation will provide a solution to the research problem. The Design Science Research (DSR) method arises under the DS paradigm as a way to create knowledge in the form of a prescription (when solving a particular real problem) or a design (when building a new artifact).

DSR can be used to create and evaluate design artifacts (such as frameworks, models, methods and instantiations) with scientific rigor [37]. It aims to solve problems which are relevant to practice, while the development and evaluation process should contribute to the state of the art. According to Hevner [1], DSR is inherently iterative and comprises three core research cycles (Design, Relevance and Rigor), as illustrated in Figure 3.1. The main tasks are developed around the central Design Cycle, where the artifacts are built, evaluated, and refined using the succeeding feedback. The Relevance Cycle constitutes a bridge between the environment (people, organizations, and technology) and the DS activities, where the environment provides the opportunities and requirements to design the artifact, and also establishes the acceptance criteria for the evaluation of the research result. The Rigor Cycle relates to the scientific foundations, experience, and expertise that compose the base knowledge of the research project. Also, the Rigor cycle ensures that the designs produced are research contributions, hence guaranteeing its innovation.

The last step of the DS paradigm is to compose the work method, which is the methodological guidelines with the logical steps to reach the goals of the

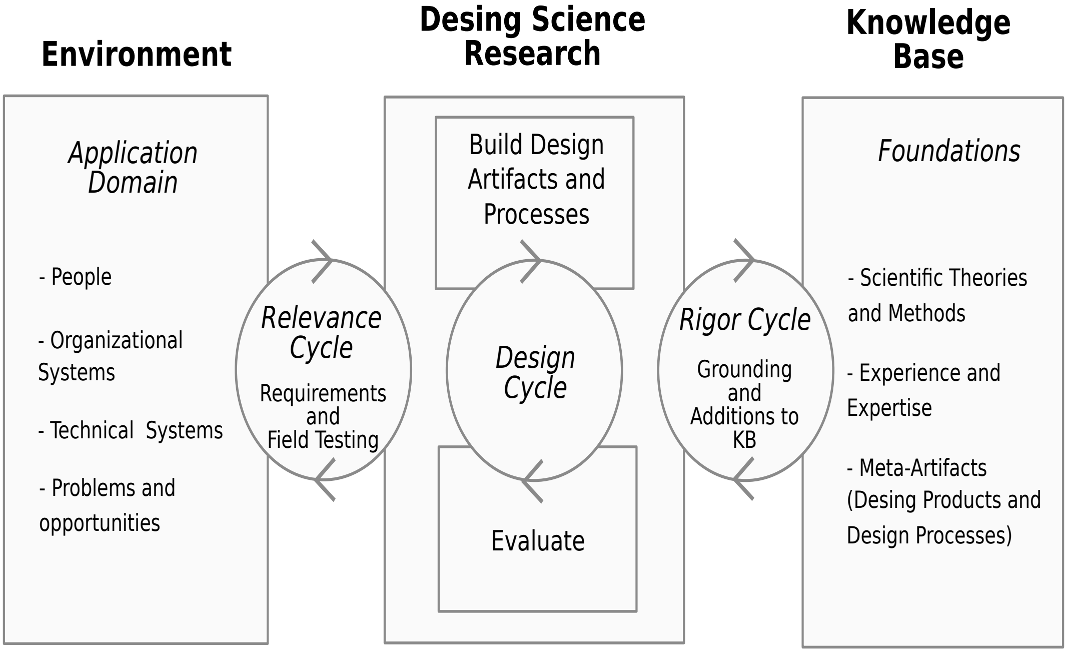

Environment
Desing Science Research
Knowledge Base
Application Domain
- People
- Organizational Systems
- Technical Systems
- Problems and opportunities
Relevance Cycle
Requirements and Field Testing
Build Design Artifacts and Processes
Design Cycle
Evaluate
Rigor Cycle
Grounding and Additions to KB
Foundations
- Scientific Theories and Methods
- Experience and Expertise
- Meta-Artifacts (Desing Products and Design Processes)

Figure 3.1: DSR cycles [1].

study. This method provides clarity and transparency to the research process and ensures its later reproducibility. In the context of the present research, the work method will represent the life cycle of a Data Science project, and will be based on a framework to conduct a predictive analytic (data mining) process, the CRoss Industry Standard Process for Data Mining (CRISP-DM) [2]; which will be detailed in the following Section.

## 3.2 Data Science Life Cycle

Different approaches have already been proposed to group data science activities within frameworks that standardizes the steps of the entire process. Examples of the most comprehensive frameworks are the Knowledge Discovery in Databases (KDD) process by Fayyad [38], the Sample, Explore, Modify, Model, Assess (SEMMA) process from the SAS Institute [39], and the Cross Industry Standard Process for Data Mining (CRISP-DM) [40]. These frameworks have the common purpose of guiding the development of methods for making sense of raw data by applying data-mining techniques for pattern discovery and extraction [38]. Hence, due to the generality of these approaches, they could be considered equivalent to the purpose of this research. In this sense, the CRISP-DM process, defined by its creators [2] as *"... a comprehensive data mining methodology and process model that provides (...) with a complete blueprint for conducting a data mining project"*, was followed.

CRISP-DM phases are shown in Figure 3.2, where it is important to notice the cyclical nature and the interactions between the core stages of the process, allowing the inclusion of the new experiences gained during the design process to refine and trigger further business questions.

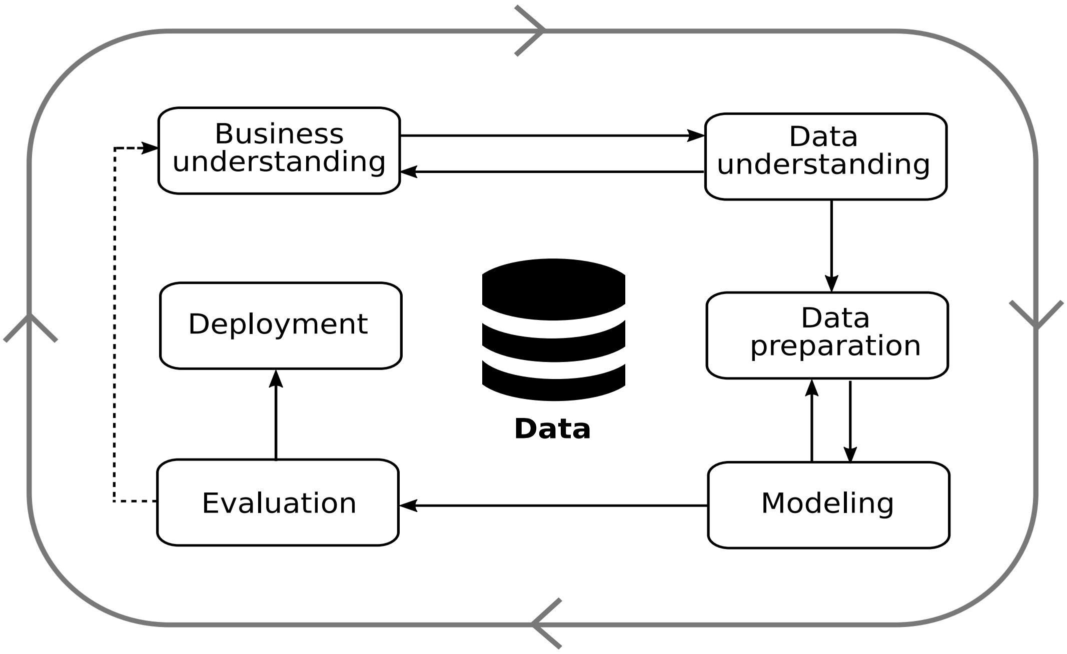


Figure 3.2: Data Science Life Cycle [2].

### 3.2.1 Business understanding

This phase focuses on finding and understanding the questions of interest from a business perspective, and transforming these demands into a data mining problem definition. This first phase is related to the DSR's Relevance Cycle, due to the information exchange with the environment (organization and stakeholders), and is decisive to make all the subsequent decisions during the project. From this cooperation, the researcher needs to compose a set of success criteria (metrics) to know what a "good" model will look like and thus ensure that the answers to the problem would assist a decision-making process. Within this phase, there are three tasks described next.

**a) Determine business objectives**

Make the main objectives of the business to be pursued explicit, as well as the related questions that the organization wishes to address and thus avoid giving answers to incorrect questions.

**b) Evaluate the situation**

This task includes planning the resources available, discovering which data will be accessible to meet the proposed goals, and listing all the assumptions coming from the characteristics of the data.

**c) Determine the data mining goals**

The business objectives are transformed into data mining goals, stating what is going to be done, how, and with which data. If this goal transformation is not successful, the business objectives should be redefined. The success measure for the data mining results should also be defined, indicating the metrics and levels of acceptance.

### 3.2.2 Data Understanding

The data understanding phase includes collecting the available data, describing its characteristics to get familiar with it, discovering initial insights and identifying data quality issues.

This phase interacts with the previous phase (Business understanding) as it is possible to reformulate or refine research questions and objectives after getting familiar with the available data.

It is also the beginning of the so-called Internal Cycle of Data Science [41] (which also includes Data Preparation and Data Modeling). In this research context, the Internal Cycle of Data Science is the core process of the Design Cycle within the DSR methodology.

The Data Understanding phase comprises the tasks explained next.

**a) Collect and describe data**

Here the first step is to load and integrate the data coming from different sources and to report any issue encountered in the process, to avoid future errors on replications of the project. Once the data is accessible, the analyst should describe the properties of the acquired data, examining the format, the number of records and variables, in order to check whether it satisfies the project specifications.

**b) Explore data**

This step comprises a deep understanding of all variables in the data set. The analyst elaborates a report outlining the findings of the exploratory analysis and potential hypotheses, which will also serve as a guide to define the actions to be undertaken in the next phase (Data Preparation). Related tasks are the analysis of descriptive statistical metrics for individual variables, assessment of relationships (correlations) between pairs of variables, and the use of visual techniques (graphs, tables, histograms) to analyze more complex relationships between variables.

The descriptive analysis of each variable is carried out by calculating the following statistics regarding their data distribution: for quantitative variables, the median and interquartile range (i.e., the difference between 1st quartile and the 3rd quartile) or mean and standard deviation; for the qualitative variables, the frequency and proportions are calculated.

A correlation analysis evaluates the relationship between pairs of variables. In the present research, for continuous variables, the correlation was measured using the Spearman's rank correlation coefficient ($r_s$) [42], which is the Pearson's product-moment correlation coefficient but between rank variables ($r_x, r_y$), thus being robust when extreme values are present and do not make any assumption of the data distribution [43]. $r_s$ is defined as:

$$r_s = \frac{cov(r_x, r_y)}{\sigma(r_x)\sigma(r_y)} \tag{3-1}$$

where $cov(r_x, r_y)$ is the covariance [44] of the rank variables and $\sigma(r_x)$, $\sigma(r_y)$ their standard deviation. $r_s$ is a measure of the strengths and direction of the monotonic[1] relationship between two variables and can take values from -1 to 1, where the closer $r_s$ is to zero, the weaker is the association between the ranks.

The association between categorical variables, in the present research, was measured using the Cramér's V coefficient ($V_c$) [45], which is an extension of the phi coefficient $(\varphi)^2$ for categorical variables with more than two classes. $V_c$ measures the strength of the association between two variables after conducting the chi-squared ($\chi^2$) test of independence. $V_c$ ranges from 0 to 1, with higher values indicating higher strengths of association, and is defined as [46]:

$$V_c = \sqrt{\frac{\varphi^2}{L-1}} \tag{3-2}$$

where $\varphi^2 = \chi^2/N$ ($N$ is the sample size), and $L$ is the smaller value of either the number of columns or the number of rows of the contingency table.

After the correlation analysis, highly-correlated variables are further investigated using visualization techniques such as scatter plots (to display

[1] It is less restrictive than linear

[2] Pearson's product-moment correlation coefficient for binary variables

relationships between numerical variables) or bar plots (to depict categorical variables proportions).

**c) Assess data quality**

At this point, the analyst examines the quality of the data. The previous task could point to some problems such as the presence of outliers, attributes with redundant meanings or missing values. Outliers, in this case, refer to an observation or set of observations which appears to be inconsistent with the remainder of the data distribution typically caused by imputation or human errors, and consequently, need to be removed. Moreover, it is also necessary to check the plausibility of values and any suspicious data, such as miscodes or spelling mistakes. All issues should be reported and discussed with the business stakeholders to agree on the treatment that they will receive.

When predicting cost

### 3.2.3 Data Preparation

The data preparation phase comprises all activities to construct the data set to be mined, using the knowledge obtained in the previous phases. The objective is that the final data set meets all the input requirements of the modeling techniques to be used in the next phase (Data Modeling) of the Data Science life cycle. It is also possible to cycle back and forth between data understanding and data preparation activities as required by learning more of the data set and performing additional operations on it. Tasks within this phase include selecting, cleansing, constructing, integrating and formatting data.

**a) Select data**

In this task, the goal is to select both records and attributes that will be used in the posterior analysis. The selection criteria are based on the data mining goals, as well as on quality and technical constraints. All explanations that justifies data inclusion or exclusion should be reported and validated by business stakeholders.

In addition, this task applies sampling techniques following two purposes: to split the data set into subsets for model learning and for model evaluation, and to overcome class imbalance issues due to the occurrence of rare events.

The objective of splitting the data set is to obtain fair evaluations of the model's performance, by ensuring that observations used for evaluation were not considered during the training phase. An effective approach is to partition the whole data set into three parts: training, validation, and test sets. The training set is used for model learning, during which a validation procedure is conducted (using the validation set) to assess the performance of ML models (during the hyper-parameter tuning and model selection) and to avoid overfitting the model to the training data. Then, the test set is used for evaluating the learned model, to estimate the effectiveness of the model on previously unseen data.

This approach has some drawbacks, especially when the sample is not large. Several authors [47–49] have pointed out that the use of a test set limits the number of examples for the training process, and its size may not have sufficient power or precision to make reasonable judgments.

In this context, resampling techniques emerge as a solution for obtaining honest estimates of the model's performance during the training phase without requiring and extra test set. Generally, resampling techniques operate in a repeated process, where a subset of samples is used to fit a model, and the remaining samples are used to measure its efficacy. Once the iterations are finished, the results are aggregated and summarized. Examples of resampling techniques are the k-fold Cross-Validation (CV) [48], Monte Carlo CV [50] and the Bootstrap [51], differing on how subsamples are chosen. Deciding the resampling technique could be a tough task, often relying on the bias-variance trade-off. The bias is the difference between the average estimation and the true values, while the variance relates to uncertainty (noise); it is the effect of obtaining different results when repeating the resampling procedure.

The k-fold CV (and its variant, repeated k-fold CV) presents good bias and variance properties, and reasonable computational cost [52, 53]. In the k-fold CV procedure, the samples are randomly partitioned into $k$ sets (folds) of approximately equal size. Then, $1-1/k$ of the samples are used to fit the model, and the remaining $1/k$ samples (not used for training) is predicted and used to estimate performance measures. This procedure is repeated $k$ times, using each fold for prediction exactly in one of the $k$ iterations. Then the error estimations are summarized, usually providing the mean and standard deviation. The repeated k-fold CV follows the same procedure, but it is repeated a defined number of times, aiming to reduce the variance.

Another aspect to consider when using resampling techniques is the way the samples are selected to be part of each set. Random sampling is the usual technique, but in some cases it is crucial to account for the outcome variable distribution when splitting the data, meaning that it is desirable to make the training and validation sets as homogeneous as possible. A solution is to use stratified random sampling, which applies random sampling within subgroups; e.g., the outcome classes in classification problems, and in regression problems, the numeric response could be broken into similar groups, and then the randomization is executed within these groups.

As pointed out before, sampling techniques are also used to deal with class imbalance problems. A significant difference in the relative frequencies of the classes can have a notable impact on the effectiveness of the model because it tends to focus on the prevalent class and to ignore the rare events. There are different approaches to overcome class imbalance, including altering the probability cutoff when making class predictions and modifying case weights [54] and prior probabilities [55] during the training process. Moreover, when the class imbalance problem is known in advance, a useful method to reduce its impact is to sample the training set to have roughly equal event rates (i.e., to balance the class frequencies). Solutions following this approach include a variety of techniques to sample the data, such as random oversampling (up-sampling) the rare class, random undersampling (down-sampling) the prevalent class, and generating new artificial examples with some similarities with the observations belonging to the minority class. Both up-sampling and down-sampling approaches present drawbacks: undersampling discards valuable data by reducing the sampling size, while oversampling may increase the likelihood of overfitting since observations are duplicated [56]. Figure 3.3 *a*) illustrates the behavior of these techniques.

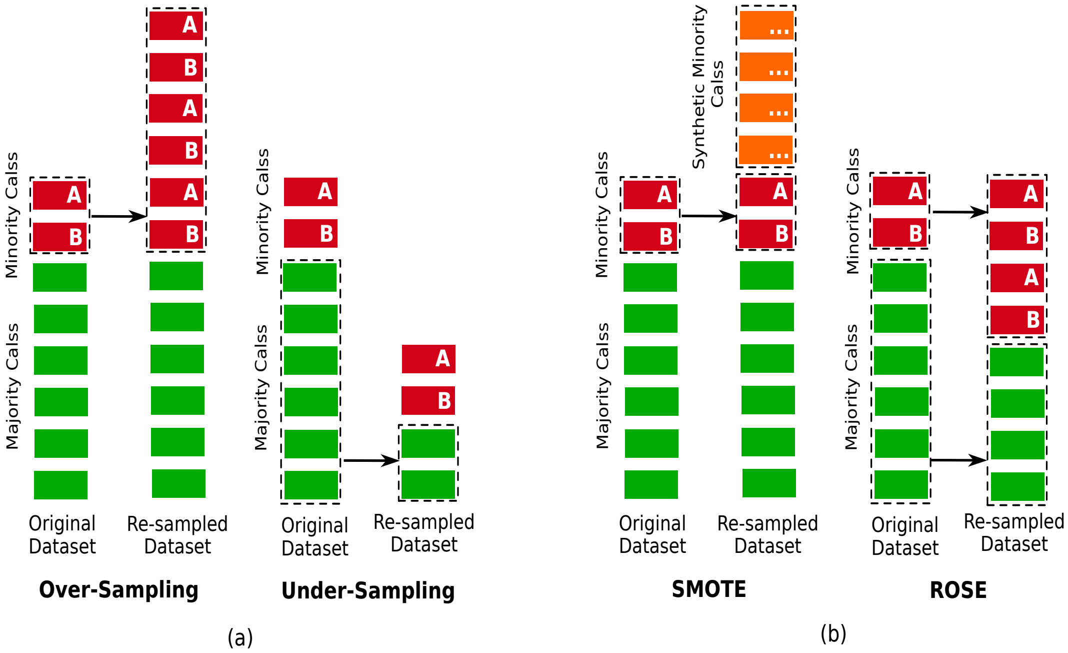


Figure 3.3: Resampling technique for imbalanced data. *a*) up-sampling and down-sampling. *b*) SMOTE and ROSE

In order to address these disadvantages, other approaches create synthetic observations by creating a set of artificial events of the minority class instead of simply increasing their multiplicity, hence having the desired effect of causing the classifier to identify larger decision regions associated to the minority class [57], such as the Synthetic Minority Over-Sampling Technique (SMOTE) [57] and Random Over Sampling Examples (ROSE) [58] (see Figure 3.3 *b*)). SMOTE uses both up-sampling and down-sampling, depending on the class; however, the up-sampling is conducted by creating synthetic observations. First, it selects a random data point of the minority class and its K-Nearest Neighbors (KNNs) to create the new observation as a random combination of the predictors of the selected data point and its neighbors. Besides, it is possible to down-sample the prevalent class in order to help balance the training set. ROSE combines oversampling and undersampling by generating an augmented sample of data (especially the rare class). The procedure is to draw with equal probability an observation belonging to one of the two classes and generate a new example in its neighborhood [59]. This technique ensures that the same attention is addressed to both classes during the resampling procedure, hence helping the classifier in estimating a more accurate classification rule. If the training set is sampled to be balanced, the validation set should not, because it must reflect the imbalance so that honest estimates of future performance can be computed. Researches on the effectiveness of using sampling procedures to overcome skewed class distributions agree that these techniques mitigate the imbalance issue in many cases [60, 61], but there is no clear winner among the various approaches, as modeling techniques can react differently to sampling in different data sets.

**b) Clean data**

Data cleansing techniques aim to solve each quality problem outlined in the data understanding phase. To do that, the analyst could either select a clean subset of data (i.e., discard the mistaken observations), drop zero or near-zero variance predictors across samples (i.e., predictors values constant or almost constant), apply imputations techniques to estimate missing data or standardize predictors' spelling. This task is essential to ensure reliable data mining results and subsequent analysis.

**c) Construct and format data**

The next step after getting the data cleaned is to undertake operations related to developing entirely new records or composing derived attributes (i.e., feature engineering). New observations may be necessary for scenarios where the objectives of the data mining problem and the ML technique require examples that are not available in the original data. In contrast, derived attributes are constructed from the existing ones in order to reduce the number of input variables and to ease the model process. Another possibility is to perform single-attribute transformations, which include binning, standardization, normalization, and creation of dummy variables.

### 3.2.4 Data Modeling

In the final step of the Internal Cycle of Data Science, various modeling techniques are selected and implemented. As stated before, due to algorithms' specific requirements in the data format, it may be necessary to step back to the data preparation stage. The related tasks are the selection of the modeling technique, the creation of models, and the assessment of models.

**a) Select and build the modeling technique**

The first step is to choose and report one or more ML models to address the problem under study, together with all its requirements and assumptions. ML techniques can be grouped following different criteria (e.g., their learning style or their functional similarity, that is, similar methodologies to solve a problem).

The first grouping criterion is useful to get an idea of the roles of the input data and the model preparation process, clearly distinguishing what types of techniques can be used according to the characteristics of the problem. The most commonly used are supervised and unsupervised learning, but there are others, such as semi-supervised and reinforcement learning. In supervised learning, the value of the response variable is known in advance (labeled data), so the objective is to learn patterns in the data from examples, to be able to generalize and predict the outcome variable in future observations. In unsupervised learning, the result is not known beforehand (unlabeled data), thus there is no predefined target; therefore, its objective is to model the underlying structure or distribution in the data [62]. On the other hand, ML techniques can also be grouped by problem categories (functional similarities), such as classification, regression, clustering, and association rules. According to this dissertation's objective, the research can be classified as a supervised learning problem, precisely a binary classification problem, since the target variable is categorical with two possibles labels (i.e., high and low-cost).

The specific algorithms chosen[3] to solve the problem under study are

[3] The algorithms were selected guided by those used in the related works, see table 2.2.

Logistic Regression (LR), Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression Tree (CART), Random Forest (RF) and Extreme Gradient Boosting (XGB). In the following subsections, these algorithms are described in detail.

– *Logistic Regression (LR)*

LR estimates the relationship between a dependent categorical variable $Y \in \{0; 1\}$ and an independent one $X$. Like ordinary linear regression, it falls into a larger class of techniques called Generalized Linear Models (GLMs) that comprise many different probability distributions (named link functions). For LR, the logistic function is used to meet the characteristics of this problem, and is defined as [63]:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{3-3}$$

where the term $p(X)$ represents the conditional probability $Pr(Y = 1|X)$, and $\beta_0$, $\beta_1$ are the unknown regression coefficients. In the case of LR, $\beta_0$ and $\beta_1$ are estimated using the Maximum Likelihood Estimation (MLE) method [64]. After some manipulations on equation 3-3 we obtain the link function called *logit* (log-odds):

$$\ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3-4}$$

where $p(X)/[1 - p(X)]$ is called the *odds*, and its values range between 0 and $\infty$ (both exclusive). Values close to the extremes indicate very low and very high probabilities of falling into a specific class of the response variable $\big(Pr(Y = 1|X)\big)$, respectively. Then, from equation 3-4 is clear that the LR model 3-3 has a *logit* that is linear in X, but it is important to emphasize that it is the logarithm of the *odds* that is a linear function of the predictors, so the interpretation of the coefficient values is also a function of this relationship. Hence, increasing $X$ by one unit changes the *log-odds* by $\beta_1$ or equivalently it multiplies the odds by $e^{\beta_1}$.

– *Ridge Regression and Lasso*

MLE estimator for logistic regression often does poorly in both prediction and interpretation [65]. These weak predictions are mainly caused by the ratio between the number of observations $n$ and preditors $p$ (e.g., $n$ not much larger than $p$ or $p > n$), resulting in great variability in the MLE fit. Also, the presence of non-informative and highly correlated features leads to unnecessary complexity and unstable parameter estimates of the fitted model [53, 66]. In this context, regularizations (shrinkage) methods emerge as methodologies to improve the fitting procedure of simple linear regression.

RR adds a shrinkage penalty to the objective function of the MLE method, the log-likelihood function $\ell(\beta_0, \beta_j)$ [66], as follows:

$$\ell^{\lambda}(\beta_0, \beta_j) = \ell(\beta_0, \beta_j) - \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3-5}$$

where $\lambda$ is a tuning parameter to be determined from the data (within the validation procedure). The second term of the equation $\lambda \sum_{j=1}^{p} \beta_j^2$ is the shrinkage penalty, which increases as the values of $\beta_j$ grow, having the effect of shrinking penalty the estimates of $\beta_j$ towards zero. The ridge parameter $\lambda$ controls the amount of shrinkage, when $\lambda = 0$ the resulting estimates will be the ordinary MLE, whereas if $\lambda \to \infty$ all $\beta_j$ tend to 0.

The RR could shrink all of the coefficients towards zero, but it will not set any of them exactly to zero. This may be a challenge for model interpretation in settings in which the number of variables $p$ is quite large, and a coefficient with a value near 0 may be misleading [65]. In this context, the LASSO technique appears as an alternative to the RR to overcome this limitation. LASSO formulation is similar to the RR in 3-5, being the penalty term the only difference [67]:

$$\ell^{\lambda}(\beta_0, \beta_j) = \ell(\beta_0, \beta_j) - \lambda \sum_{j=1}^{p} |\beta_j| \tag{3-6}$$

In the case of the LASSO, the penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ has the effect of forcing some of the coefficient estimates to be exactly equal to zero when $\lambda$ is sufficiently large, thus performing variable selection. An important observation of the RR and LASSO techniques is that unlike the ordinary MLE, the estimates are not scale equivariant (i.e., the scale of the predictors affect the coefficient estimates) [53]. Therefore it is best to apply RR and LASSO after standardizing the predictors, so that they would have zero mean and unitary standard deviation [67].

– *Classification and Regression Tree*

Tree-based models seek to partition the data into smaller groups that are more homogenous with respect to the response [68]. To construct a decision tree, the first step is to divide the $p$ predictor' space into $r$ distinct and non-overlapping regions, and then, to make the same prediction for each observation that falls within a region. The first step is performed following the recursive binary splitting approach [69], which consists in searching the predictor and its split value that partitions the data into two groups optimizing a specific measure. In the case of classification trees, the splitting criterion could be the Gini index (G):

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3-7}$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the $m^{th}$ region that are from the $k^{th}$ class. Equation 3-7 results in small values when all $\hat{p}_{mk}$'s are close to 0 or 1, which means that the node contains most of the observations coming from a class. For this reason the Ginix index is referred to as a measure of the node purity. Other measure used as splitting criterion is the Cross-entropy (C), given by:

$$c = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}) \tag{3-8}$$

The Cross-entropy has a behavior similar to the Gini index, the closer are all $\hat{p}_{mk}$ values to 0 or 1, the smaller is the Cross-entropy.

Following one of this criteria, the splitting process continues within each newly created partition until some stop condition is met, such as the minimum number of samples in a node or the maximum tree depth. This approach can lead to large trees that are likely to over-fit the training data, leading to too complex trees that have poor test set performance. This approach can lead to large trees that are likely to over-fit the training data, leading to too complex trees that have poor test set performance. To overcome this issue, after growing a deep tree, it is pruned back using a cost-complexity pruning approach proposed by [68], where the purity criterion (Gini Index or Cross-entropy) is penalized by a factor of the total number of terminal nodes in the tree. At last, in the classification setting, predictions are made classifying each observation to the most commonly occurring class of training observations in the region to which it belongs.

– *Random Forest*

Traditional decision-tree based algorithms suffer from high variance[4] [69]. This unwanted behavior is caused because the splitting variable and the exact position of the cutpoint fully depend on the distribution of the observations in the learning set. Then, as the recursive partitioning approach is used to

[4] Instability to small changes in the learning data.

build the tree, little variation in data could change the first split variable or the cutpoint, leading to an entirely new tree structure. RF overcomes the variability generated by a single tree using the bagging[5] method [70], which is based on the fact that averaging over a set of observations, reduces the variance. The bagging procedure randomly draws different bootstrap samples from the learning set, and individual trees are grown on each sample. Then, the resulting predictions are averaged, and for the classification case, is calculated following the majority vote approach, which is to take the most commonly occurring class among all predictions. Also, these tree-based ensemble procedures do not have a pruning step, and so the trees grow deeply, resulting in individual trees with high variance and low bias. The advantage is that the resulting prediction combines the output of a diverse set of trees that are instable but produce unbiased predictions [69].

The RF algorithm [71] generalizes the bagging procedure, allowing to restrict the number of candidate predictors. In RF, the set of predictors to select from is randomly limited at each split, thus producing decorrelated (i.e., diverse) trees. Less correlated[6] trees help reduce the variance of the predicted values since trees with more diverse structures are created [52]. Bagging is considered a special case of RF, where the number of randomly selected splitting variables equals the entire set of variables.

The RF algorithm has a set of parameters to be configured as their optimal values depend on the data. As stated by Breiman [71], the randomness (diversity) added to a model should reflect a balance between correlation and the strength of the trees. This trade-off is reflected in the optimization of three hyper-parameters[7], the number of candidate variables at each split ($n_{try}$), the sample size ($n_{sample}$), and the node size ($n_{node}$). In the case of $n_{try}$, low values lead to more diverse trees and hence produce better stability when aggregating the result. However, it comes at the cost of low average performance (e.g., accuracy) as the selected variables are likely to be suboptimal. To set $n_{sample}$ is to determine the number of observations sampled for training each tree. As with $n_{try}$, decreasing the value of $n_{sample}$ decreases the model variance, but since fewer observations are used for training single trees their performance worsens. Moreover, it is necessary to determine if the procedure of selecting the $n_{sample}$ is made with or without replacement. Researches [72] on this topic may suggest that there is no significant difference sampling scheme when the optimal value of $n_{sample}$ is used. Nevertheless, Janitza et al. [73] argues

[5] Bootstrap Aggregation.

[6] Lower correlation is achieved by adding a source of randomness when growing the decision trees.

[7] Tuned during the cross-validation procedure.

that, in fact, sampling with replacement introduces a variables selection bias when there are categorical variables with more than two categories. The other hyper-parameter, the $n_{node}$, specifies the minimum number of observations in a terminal node; therefore, it determines the depth of the tree, serving as stopping criterion during the tree construction process.

Moreover, in the RF algorithm, the number of trees ($n_{tree}$) and the splitting rule needs to be set beforehand. Setting the $n_{tree}$ is challenging since it highly depends on the dataset properties and the computational time available. As a recommended path to set the $n_{tree}$, related researches [74] suggest to inspect the performance[8] estimation during the validation procedure, showing the behavior for a growing number of trees. On the other hand, there are options to set the splitting rule, like the conditional inference forests, which assess the p-values of statistical tests for both, the variable selection and split value [75], or the randomized splitting rule introduced by Geurts et al. [76], where only a randomly selected subset of possible splitting values is considered. These different choices intend to overcome a selection preference resulting from the Gini index criterion originally proposed [71]. Using the Gini index favors the selection of variables with many possible splits, such as the continuous variables over the binary one [77].

– *Extreme Gradient Boosting*

Analogously to bagging and RF, boosting is an ensemble algorithm seeking to improve the performance of individual models (e.g., CART). However, boosting algorithms seek models that complement one another, turning a weak learner[9] into a strong one, whereas bagging and RF build independent models that are not aware of the performance of the others [78].

The first implemented algorithm of boosting, known as Adaptative Boosting (AdaBoost) [79], trains classifiers on weighted versions of the training sample, giving higher weight to cases that are currently misclassified. Then the final model is defined to be a linear combination of the classifiers from each stage and the final prediction, a weighted sum of the model's predictions. Further improvements over the AdaBoost algorithm led to a generalization named Gradient Boosting Machine (GBM) [80], which transforms the boosting definition to an optimization problem. The major difference between AdaBoost and GBM is the approach to identify the faults of weak learners. While the AdaBoost model uses high weights to misclassified data points, GBM uses gradients to minimize a loss function[10]. This generalization allows arbitrary

[8] Performance is measured using a metric of interes

[9] Models that performs barely better than chance, e.g., decision stump

[10] Determines the error between the prediction and the observed value

differentiable loss functions to be used, expanding the technique beyond binary classification problems to support regression and multi-class classification.

XGB is an improved implementation GBM since it enhances the costly process of estimating the potential loss for all possible splits when growing additive decision trees. This improvement is carried out using the distribution of all data points in a leaf, and with this information, reduce the search space of all possible splits [81]. The XGB implementation has a set of hyper-parameters to be tunned, like the number of boosting iteration ($n_{rounds}$), the learning rate ($eta$), the maximum tree depth each tree can grow ($max_{depth}$), the minimum loss reduction required to make a further partition ($gamma$), the number of predictors to be subsampled each time a tree is trained ($n_{predictor}$) and the subsample ratio of the training instances ($n_{rows}$).

### b) Assess the model

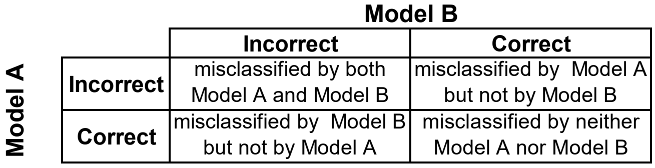
Once the ML techniques are applied, the different experimental scenarios are evaluated and compared, defining the best parameter configuration for the data set and the technique with the best overall performance. There are two primary purposes when evaluating the learned models, one side is focused on defining which one perform better based on statistical significance tests, and the other side of the evaluation process relies on specific metrics to measure model performance.

The first point of view is interested in selecting models based on the estimated performance for which it is necessary to know whether there is a statistically significant difference among them. In this research, the McNemar's test [82] is used to compare the predictions of two models to each other. The McNemar's test is a non-parametric statistical test for paired comparisons that can be applied to compare the performance of two ML classifiers [83]. The null hypothesis states that the two algorithms should have the same error rate, and it is applied to paired nominal data based on a version of 2x2 contingency table, where a $\chi^2$ test for goodness of fit is applied. The contingency table contains the number of hitz and misclassifications of both models, as depicted in Figure 3.4.

On the other hand, there are an extensive number of metrics to measure models performance especially for classifiers (i.e., the response variable is categorical), which can be grouped into three families, qualitative, rank-based, and probabilities-based metrics [84].

– *Qualitative metrics*

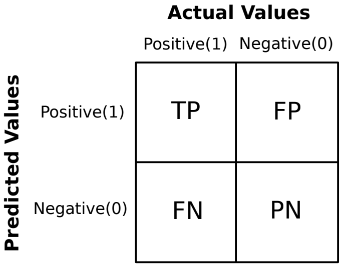
Metrics with a qualitative understanding of the errors aims to find a model that minimizes the number of faults. These types of metrics can be



| Model A | Model B: Incorrect | Model B: Correct |
|---|---|---|
| Incorrect | misclassified by both Model A and Model B | misclassified by Model A but not by Model B |
| Correct | misclassified by Model B but not by Model A | misclassified by neither Model A nor Model B |

Figure 3.4: Contingency Table for McNemar's test.

derived from a confusion matrix, as illustrated in Figure 3.5 for a binary classification problem.



| Predicted Values | Actual Values: Positive(1) | Actual Values: Negative(0) |
|---|---|---|
| Positive(1) | TP | FP |
| Negative(0) | FN | PN |

Figure 3.5: Confusion Matrix

In the matrix, rows display the predicted class and columns the real class value, and in both, positive and negative refers to the classes of the response variable. Besides, each cell represents either a success or a mistake in the prediction as follows:

– *True Positive (TP)*: actual positives observations that are correctly predicted as positives

– *False Positive (FP)*: actual negatives observations that are wrongly predicted as positives

– *True Negative (TN)*: actual negatives observations that are correctly predicted as negatives

– *False Negative (FN)*: actual positives observations that are wrongly predicted as negatives

Among the most common metrics are Accuracy, Sensitivity, Positive Predicted Value (PPV), F-measure and Matthews Correlation Coefficient (MCC). Accuracy, possibly the most used metric in the ML field [84], is a simple measure of the proportions of all samples classified correctly:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-9}$$

In the context of databases with unbalanced classes, Accuracy does not provide a fair estimate of the model performance, for example, in a dataset with 1% of the observations belonging to the positive class, a useless classifier[11] will obtain an accuracy of 99%. In this case another metric that can be useful because it consider information from both classes is the geometric mean (G-mean) [85] of the True Positive Rate $[TPR = TP/(TP + FN)]$ and the True Negative Rate[12] $[TNR = TN/(TN + FP)]$, defined as:

$$G\text{-}mean = \sqrt{TPR \times TNR} \tag{3-10}$$

Using G-mean ensures, unlike Accuracy[13], that poor performance in prediction of the positive examples will lead to a low G-mean value, even if the negative examples are correctly classified.

Moreover, there are metrics focused on single classes (positive class) separately, such as Sensitivity and PPV. Sensitivity[14] is the fraction of total positive instances correctly classified as positive, and PPV[15] is the proportion of correctly classified instances among the ones classified as positive:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3-11}$$

$$PPV = \frac{TP}{TP + FP} \tag{3-12}$$

The equations 3-11 and 3-12 show two perspectives, first if looking for model's Sensitivity, the objective may be reducing the FN, and second when looking for model's PPV the aim is to reduce the FP. These two values, (FN and FP), are known as the classifier errors and has the opposite effect on each other. This effect is shown in Figure 3.6, where it also appears another metric that balances and combines Sensitivity and PPV in one value, the F-measure $(F_\beta)$ [86]. It is defined as the weighted harmonic mean of Sensitivity and PPV:

$$\begin{gathered} F_\beta = \frac{(1 + \beta^2) \times TP}{(1 + \beta^2) \times TP + \beta^2 \times FN + FP} \\ \\ F_\beta = \frac{(1 + \beta^2) \times Sensitivity \times PPV}{\beta^2 \times PPV + Sensitivity} \end{gathered} \tag{3-13}$$

[11] Classify all observations as negative.

[12] Also known as Specificity.

[13] Arithmetic mean between TPR and TNR.

[14] Also known as *TPR* or *Recall*.

[15] Also known as *Precision*.

Where $\beta$ is a parameter that controls the influence of Sensitivity and PPV in the score. As illustrated in Figure 3.6 common values of $\beta$ are 1 to seek weights equilibrium, 0.5 to favor PPV and 2 to augment Sensitivity's influence.

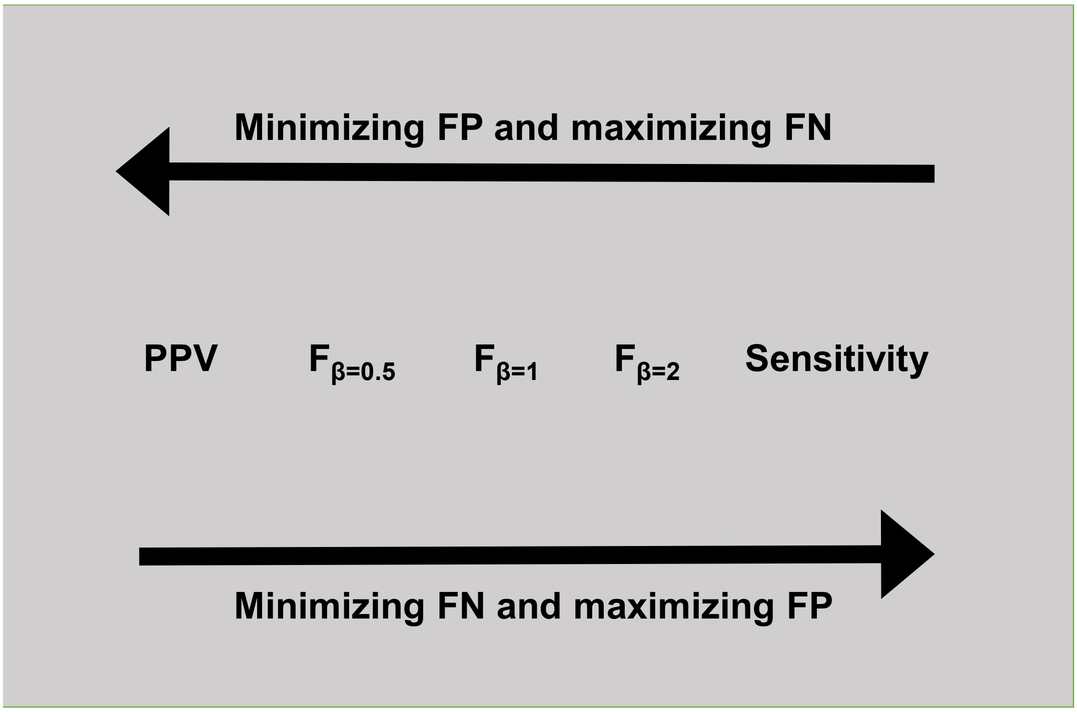

Minimizing FP and maximizing FN
PPV
$F_{\beta=0.5}$
$F_{\beta=1}$
$F_{\beta=2}$
Sensitivity
Minimizing FN and maximizing FP

Figure 3.6: Effect of optimizing one of the two errors *FP* or *FN*.

One issue with the F-measure (also true for Sensitivity and PPV) is that it depends on which class is defined as the positive, changing its value depending on this decision. The Matthews Correlation Coefficient (MCC) [87] overcomes this issue. Besides, it is unaffected by the unbalanced datasets and takes into account all ratios of the four confusion matrix categories (TP, FP, TN, FN). MCC is a contingency matrix method of calculating $\varphi$ (see Section 3.2.2) between the actual and predicted values (Powers2011), and can be derived from the Confusion Matrix as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{3-14}$$

MCC values ranges int he interval $[-1, 1]$, where a value of 0 is the performance of a useless classifier, and extreme values are related to perfect classification (1) and to total misclassification (-1).

All the measures described above depend on an arbitrarily selected classification threshold[16]. In other words, these metrics do not account for the magnitude of the error, for example, if the decision threshold is set to 0.5, the difference between the predicted probabilities of two observation (e.g., 0.1 and 0.4) are not reflected in this kind of metrics, both will be classified with the same label.

– *Rank-based metrics*

Another important family of performance measures is based on ranks (separability) of predictions. These metrics are not committed to a specific threshold as the previously discussed qualitative metrics; then, it is possible to have a good classifier in terms of rankings and yield bad accuracy as a particular threshold is selected to separate the classes. Rank-based metrics use the ordering given by the predicted score[17] to rank the instances. The ranking is made according to the decreasing order of the predicted score for an observation being positive, and then, this list is compared to the real class label in the test set. It is common to use graphical techniques to summarize the classification performance, such as Receiver Operating Characteristic (ROC) curve [88] and the Precision-Recall (PR) curve [89].

ROC graphs are two-dimensional plots, where the x-axis represents the Sensitivity (see 3-11), and the y-axis denotes the False Positive Rate[18] (FPR) [FPR = FP/(TN + FP)]. Each point in the plot results from the confusion matrix created with the ranking list of the classifier's score for each instance in the test set. Then, the complete plot is created changing the decision threshold to decide which class an observation belongs to, thus having a different confusion matrix for each pair Sensitivity and FPR.

Figure 3.7 shows an example of the ROC Graph. In the ROC space, a good classifier should reach as close to the top left corner as possible, corresponding to a perfect classification. In contrast, the diagonal that connects the lower-left corner (0, 0) and the top-right one (1, 1) indicates a random performance. The point (0, 0) reflects a scenario where all classifications belongs to the negative class (i.e., TP = FP = 0, caused by a high threshold). On the other hand, the point (1, 1) indicates the opposite, all classifications correspond to the positive class (i.e., FN = TN = 0, as a consequence of a very low threshold).

Comparing classifiers visually through a ROC plot can be difficult, given that curves can overlap and even cross each other, not making clear whether

[16] Cutoff of the predicted probabilities to define to which class an observation belongs, being the default value in most techniques 0.5.

[17] Refers to a numeric value that depicts the degree at which an instance pertains to class.

[18] Is the proportion of negative observations incorrectly classified as positive.

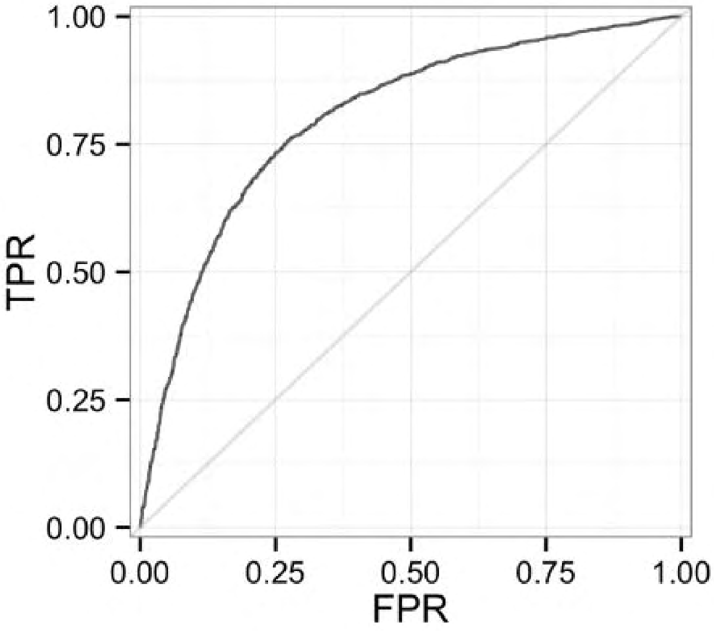


Figure 3.7: ROC plot.

the performance of one classifier dominates another in all the operating points. A solution is to reduce ROC performance to a single scalar value representing the expected performance. This value is known as the Area Under the Receiver Operating Characteristics (AUROC) curve, which values range in the interval [0, 1]. Therefore, a perfect classification would have an AUROC of 1 and an unuseful classifier (over the diagonal in the ROC plot) a value of 0.5. Both the ROC curve and AUROC measures how well the model is capable of distinguishing between classes (i.e., separability capacity). Moreover, AUROC is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [90].

In the context of class imbalance, ROC graphs may provide unreliable (i.e., optimistic) estimates of the models' performance, as change in the class distributions do not reflect in the TPR and FPR [89,91]. Saito and Rehmsmeier [92], suggest that the PR curve is more informative than the ROC Graph to evaluate binary classifiers in datasets with imbalanced classes, because they display the susceptibility of classifiers to the characteristics of the data. Different from ROC, the PR Graph plots the precision (PPV) of the classifier as a function of the recall (i.e.,Sensitivity), as depicted in Figure 3.8[19]. In other words, the PR Graph describes the behavior of precision at different

[19] Note that PR curve do not consider True Negative observations, hence the correct identification of the negative class must not play an important role in the study, in order to use this Graph as a model performance estimation.

degrees of recall. The Graph 3.8 shows that the PR curve has a negative slope, as precision decrease while recall increases. Then, unlike the ROC plot, the perfect classifier would be at the top right, representing the best trade-off between precision and recall. In addition, the baseline performance is not fixed like in the ROC plot but depends on the proportion of the positive class. It is also possible to summarize the PR Graph into a single value the Area Under the Precision-Recall (AUCPR) curve and can be interpreted as the expected precision when varying recall from 0 to 1.

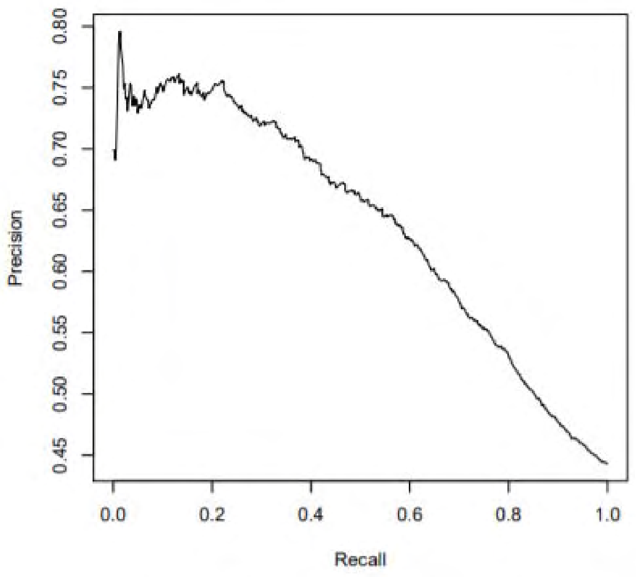


Figure 3.8: PR plot.

– *Probability-based metrics*

The other family of metrics is based on the predicted probabilities. However, instead of ranking them as the ROC and PR curves, the purpose is to quantify the uncertainty in a classifier's predictions [84]. Probability-based metrics measure not only when a model makes a mistake but also express to what extent they fail. Among these metrics are the Logarithmic Loss (LogLoss) and the Brier Score and are described below.

The LogLoss, also known as cross-entropy, is defined as the negative of the logarithm of the likelihood function [93] (see 3-8). This metric captures the extent to which predicted probabilities differ from the class label. Then, the Log loss value increases as the predicted probability diverges from the actual label, having a perfect classification value of 0.

On the other hand, the Brier Score [94], unlike the LogLoss (Measures the entire probability distribution), is focused on the positive class, being more suitable to estimate models performance with imbalanced classes. It is computed as the Mean Squared Error (MSE):

$$Brier = \frac{1}{N} \times \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{3-15}$$

where $\hat{y}_i$ is the predicted probability for the observation $i$, $y_i$ is the actual outcome of the class (1 for the positive class and 0 for the negative) and $N$ is the number of instance in the test set. The Brier Score, as a loss function, the lower its value is, the better the model's performance. Its values reflects the confidence in which the classifier asserts its prediction, i.e., if the classifier predicts the wrong class with high probability, it would be heavily penalized.

### 3.2.5 Evaluation

Once the Internal Cycle of Data Science ends, having selected the best models with optimized configurations, it is time to interpret and assess the results according to the data mining success criteria defined in the business understanding phase. This phase also relates to the Design Cycle in the DSR framework, where after the model building task, a scientific evaluation of the artifact is carried out. Tasks within this stage are, evaluate results, review the learning process, and determine the next steps.

#### a) Evaluate the results

This task aims to evaluate how well the model met the business objectives defined in Section 3.2.1. The evaluation also allows for discovering new information, further constrains and insights for future research directions. The outcome of this task is a report summarizing the results in terms of business success criteria.

#### b) Review the learning process

This review seeks to ensure that the learning process followed all the constraints and assumptions coming from the data characteristics and the modeling techniques. One crucial aspect is to check whether the attributes used as input are going to be available for futures deployments.

#### c) Determine the next steps

At this point, it is necessary to determine whether the project is finished or further iterations of the internal cycle of data science are required or even refine the data mining objective.

### 3.2.6 Deployment

The last stage, but not necessarily the end of the Data Science Cycle (note the cycle nature in Figure 3.2), aims to deploy the resulting model in a real context and receive the subsequent feedback from the field. This phase describes the interaction between the Design Cycle and the Relevance Cycle (denoted in Figure 3.1 as Field testing). Usually, it is the end-customer who is in charge of the deployment activities. Still, they need to understand all the model's characteristics in advance, for which a technical report is generated, listing all the requirements and adjustment necessaries.

The resulting feedback could feed into the earlier phases of Business understanding or any of the phases of the Internal Cycle of Data Science. In the case of Business understanding, this feedback is represented by the outer arrow that encompasses the entire cycle, being then a natural path to future works. On the other hand, feedback directed towards the three stages of the internal Cycle of Data Science may suggest modifications to refine the actual model or to correct any issue encountered during the deployment phase.

## 3.3 The Data Science Life Cycle within the DSR Cycles.

This Section aims to integrate the DSR's framework proposed by Hevner [1] with the work method described in the previous Section. The DSR approach can be extended to the Data Science field [95,96], and therefore integrate with the proposed work method, the Data Science Life Cycle (see Section 3.2) . The integration takes place through an iterative and evaluative scheme in order to diagnose, design, implement, and evolve data science artifacts.

The methodology to be followed, as outlined in this chapter, is summarized in Figure 3.9, where each phase of the work method is linked to a DSR cycle. The central cycle (Design) draws inputs (business understanding) from the contextual environment regarding problem identification and business requirements. It encompasses the stages of the internal cycle of data science and evaluation to design, develop, and improve the ML models in an iterative process by incorporating the field testing feedback. During this process, each time that the initial objective gets revised, a new design process (DSR cycle) is initiated[20]. Furthermore, the design cycle uses the current knowledge to support their theoretical research foundations and, at the same time, contributes with the resulting innovations.

[20] It is worth noticing that the possible design cycles results are not intended to be compared, instead they relate the process of improvement in the design process.

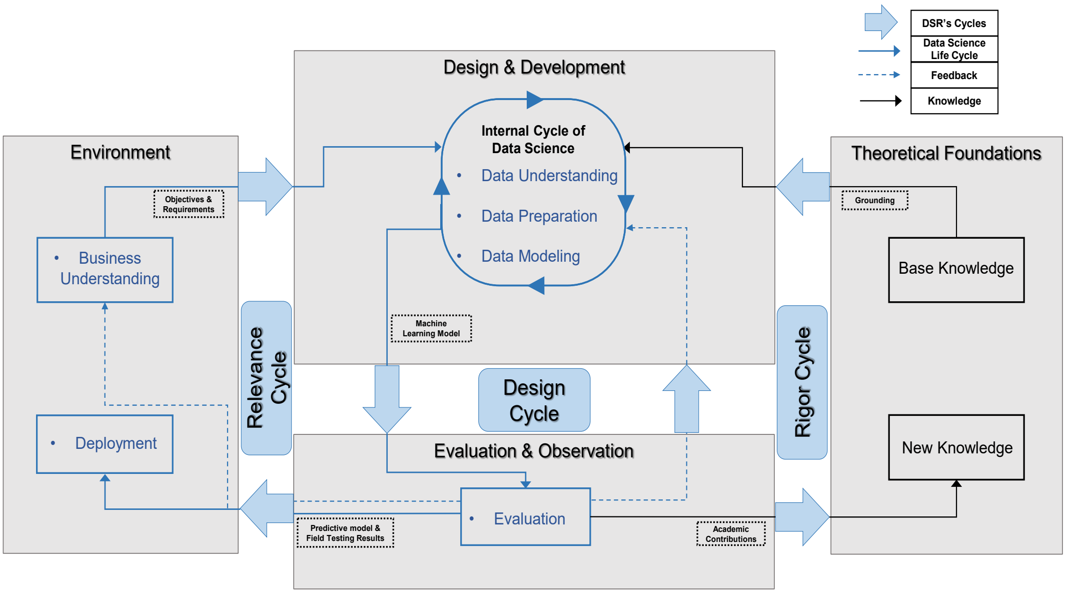


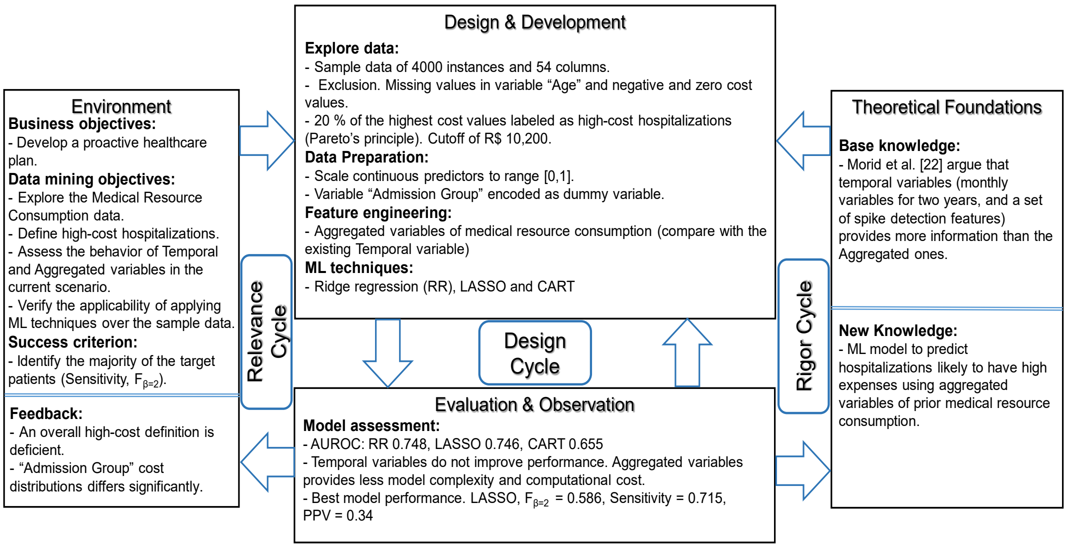
Figure 3.9: DSR and Data Science Life Cycles

# 4
# Application of a DSR methodology to predict high-cost hospitalizations

This Chapter describes a Data Science project proposed to address the problem of predicting high-cost hospitalizations in private hospital units, located all over Brazil. The project was developed in conjunction with a Health Consulting organization that servers as a broker from its clients to healthcare providers and health insurers. As explained in Chapter 3, the proposed process followed the integration of the Design Science Research (DSR) framework and the Data Science Life Cycle, as depicted in Figure 3.9. The application of the DSR framework consisted of two cycles, described in the following Sections.

## 4.1
## First DSR cycle: Data Exploration and Prediction of high-cost hospitalizations over sample data.

Figure 4.1 summarizes the activities and results obtained when instantiating the First cycle of DSR framework presented in the previous Chapter. The following subsections will describe each phase of this First cycle, in detail.



Environment

**Business objectives:**
- Develop a proactive healthcare plan.

**Data mining objectives:**
- Explore the Medical Resource Consumption data.
- Define high-cost hospitalizations.
- Assess the behavior of Temporal and Aggregated variables in the current scenario.
- Verify the applicability of applying ML techniques over the sample data.

**Success criterion:**
- Identify the majority of the target patients (Sensitivity, $F_{\beta=2}$).

**Feedback:**
- An overall high-cost definition is deficient.
- "Admission Group" cost distributions differs significantly.

Design & Development

**Explore data:**
- Sample data of 4000 instances and 54 columns.
- Exclusion. Missing values in variable "Age" and negative and zero cost values.
- 20 % of the highest cost values labeled as high-cost hospitalizations (Pareto's principle). Cutoff of R$ 10,200.

**Data Preparation:**
- Scale continuous predictors to range [0,1].
- Variable "Admission Group" encoded as dummy variable.

**Feature engineering:**
- Aggregated variables of medical resource consumption (compare with the existing Temporal variable)

**ML techniques:**
- Ridge regression (RR), LASSO and CART

Evaluation & Observation

**Model assessment:**
- AUROC: RR 0.748, LASSO 0.746, CART 0.655
- Temporal variables do not improve performance. Aggregated variables provides less model complexity and computational cost.
- Best model performance. LASSO, $F_{\beta=2}$ = 0.586, Sensitivity = 0.715, PPV = 0.34

Theoretical Foundations

**Base knowledge:**
- Morid et al. [22] argue that temporal variables (monthly variables for two years, and a set of spike detection features) provides more information than the Aggregated ones.

**New Knowledge:**
- ML model to predict hospitalizations likely to have high expenses using aggregated variables of prior medical resource consumption.

Figure 4.1: First Cycle of the DSR and Data Science Methodology

### 4.1.1 Business Understanding

From the organization perspective, there is a necessity to identify a group of patients to be included in a proactive health plan so as to reduce the cost spent to treat these patients. As stated in Chapter 1, a solution to this situation would be to analyze the expected cost of hospitalizations. In this context, the objectives of this First cycle are: (i) to explore the available sample data; (ii) to precisely define what is a high-cost hospitalization, and (iii) to apply some Machine Learning (ML) techniques so as to investigate their potential to build models that predict such high-cost hospitalizations. The organization expects to have a predictive model that captures most of the high-cost hospitalizations, controlling the amount of low-cost admissions misclassified. In other words, they seek a trade-off between a high quantity of correctly identified high-cost hospitalizations and a low quantity of misclassifications, with an emphasis on the first criterion. The success criterion can be translated as the maximization of the $F_\beta$ metric (see equation 3-13) with $\beta = 2$, thus giving more importance to Sensitivity. In addition, individual metrics such as the Sensitivity and Positive Predicted Value (PPV), together with the corresponding confusion matrix complements the performances evaluation process.

### 4.1.2 Data Understanding

At the first moment of the research project, the organization provided us with a dataset containing a representative sample of the whole population under their services, on which this First cycle was developed[1]. The dataset contains 4000 rows and 54 columns, where each row represents a patient hospitalization. Each column in the dataset is a variable, which can be grouped in (see table 4.1): demographic characteristics of the patients, past medical resource consumption variables and the variables related to the hospitalization itself, including the target variable "Hospitalization Cost"[2]. It is worth noticing that the variables related to the patient's past medical resource consumption were collected throughout the three years before the hospitalization in a temporal fashion. There were different time frequencies depending on the feature; for example, the number of exams and consultations were collected bi-annually (i.e., for each semester), image exams were collected monthly, while the rest of variables was collected annually.

[1] It is worth noticing that the use of a sample on this First cycle is due to the organization's data availability and is not a requisite of the proposed methodology.

[2] Appendix A presents the data dictionary with all variables names, data types, and detailed descriptions.

Table 4.1: Group of variables

| Patient Characteristics | Medical Resource Consumption | Patient Hospitalization |
|---|---|---|
| Gender | Ordinary exams | Admission Group |
| Age | Image exams | Admission Type |
| | Psychological consults | Flag Surgery |
| | Scheduled consults | Hospitalization Cost |
| | Emergency consults | |
| | Hemodialysis therapies | |
| | Pulsotherapies | |
| | Transplant | |
| | Chronic disease | |

A first analysis of the dataset revealed inconsistencies related to data quality that should be addressed. Figure 4.2 shows the first steps conducted to filter out data which did not meet quality specifications. First, blank values (missing values) in the "Age" variable were detected (which, according to the organization, could be due to newborn patients without a birth date registered or even an error coming from the health insurance). Then, since it was not possible to distinguish whether the missing value was related to a newborn on to an error, those observations were removed. The second issue was the existence of hospitalizations with negative or zero values in the "Hospitalization Cost" variable (this occurs when there is a refund from the hospital). Again, those observations were removed.

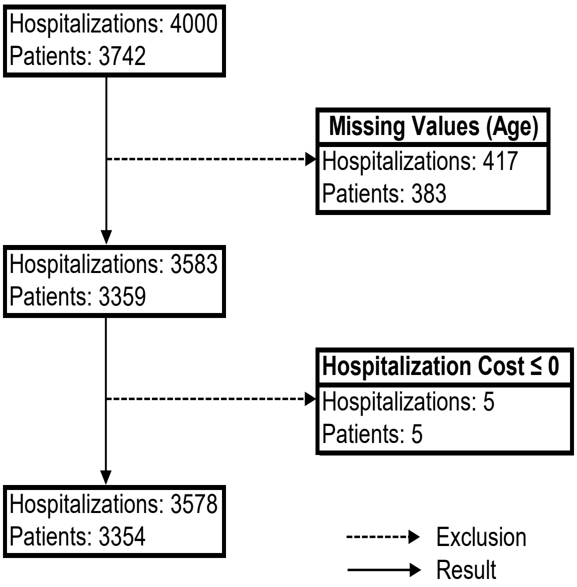


Figure 4.2: Quality Filter Procedure.

Moreover, "Hospitalization Cost" is a continuous variable indicating, as its name suggests, the monetary value associated with the patient's hospitalization expenses. In order to accomplish the business goals, it is necessary to first define what should be considered as a high-cost hospitalization. The following analysis uses the Pareto[3] principle [97] to create the binary outcome variable (1 indicates high-cost and 0 low-cost). This approach follows the definitions encountered in the literature review (see section 2.1), where the definition is made considering the whole population and hospitalizations cost belonging to the upper percentiles of the cost distribution are defined as high-cost hospitalizations. In Graph 4.3, the x-axis represents each row of the dataset grouped in 10 bins[4] (each bin roughly represents 10% of the dataset), the bars (left y-axis) represent the absolute "Hospitalization Cost" for each bin; finally, the plot line represents the cumulative percentage "Hospitalization Cost" (right y-axis).

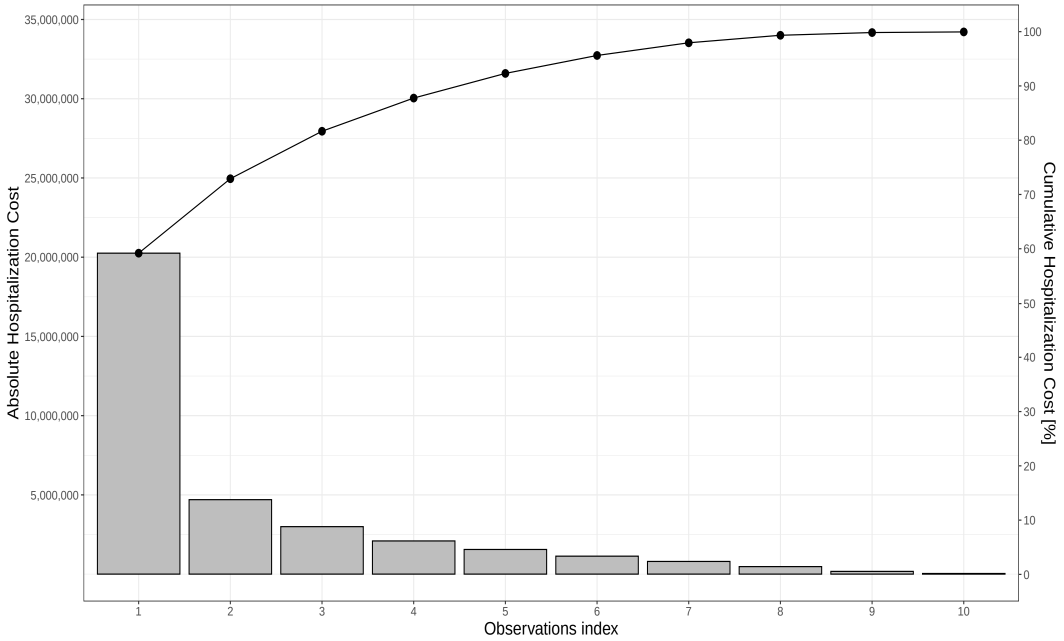


Figure 4.3: Pareto Graph for the Hospitalization cost.

As depicted in Graph 4.3, the application of the Pareto principle selects the first two bins (20% of rows), which account for approximately 80% of the total cost. Therefore, these 20% of the observations with the highest cost values were defined as high-cost hospitalizations, set in a new variable named "Hospitalization cost class". The lowest value labeled as high-cost was $R\$ \ 10,200$.

Once the dataset contains the desired information with the response variable defined according to the business requirement, the next step is to

[3] The Pareto principle states that, for many events, roughly 80% of the effects come from 20% of the causes.

[4] The observations were grouped for better visualization, each bin contains 350 rows except the last one which has 428 for a total of 3578 rows.

explore and analyze the set of variables. Tables 4.2 and 4.3 summarize each numeric and binary variable, respectively, regarding their overall distribution for each value of the response variable.

The numeric variables describe the medical resource consumption of patients in the past (except "Age"), and are characterized by having little variability, and most of the values are zero. This is a common behavior for this type of variable, as reported in related studies [98]. We observe extreme cases of variables that are completely zeroed, such as "Pulsotherapies" (in the three years) and "Transplant", which were then disconsidered in the study. Moreover, for some of the temporal variables ("Scheduled consult", "Emergency consults", "Ordinary exams" and "Image exams"), there is an increase in the median medical resource consumption on the dates closer to the hospital admission, which might suggest some positive trend in the resource usage followed by a more significant event, the hospitalization.

Table 4.2 also shows the variables' distribution within each group of the "Hospitalization cost class", which could throw the first hints in understanding the behavior of high and low-cost hospitalizations. However, the variables' distributions seem quite similar among the groups, with the most notable difference belonging to the number of "Ordinary exams" during the first semester before the hospitalization. This difference suggests that regarding the median value, patients with high-cost hospitalizations made more "Ordinary exams" than those with low-cost hospitalizations.[5]

The binary variables (see table 4.3) also favor the "No" label, indicating little use of resources. Another aspect of the data is that there are almost equal amounts of surgical and clinical hospitalizations, and most of them (73%) were admitted by appointment (scheduled). Furthermore, among the high-cost hospitalizations, 69% are surgeries, while clinic hospitalizations are mostly low-cost.

The "Admission Group" variable groups hospitalizations according to the type of procedure carried out during the in-hospital stay (e.g., digestive, musculoskeletal and urinary systems). Graph 4.4 shows the proportion of high-cost hospitalizations of the 12 most frequent admission groups[6] within the dataset. As depicted in the plot, there is an interesting difference in proportions among the groups, being significantly higher for hospitalizations related to procedures of the digestive (3rd), musculoskeletal (4th) and the urinary (10th) systems, where high-costs hospitalizations represent 31%, 41% and 44% of the total number of hospitalizations, respectively.

[5] Note that this is an exploratory analysis and the analysis is based on visual and numeric comparison, without applying any statistical test.

[6] Those 12 groups account for 78% of the observations, out of the 34 groups in total

Table 4.2: Descriptive statistics for numeric variables.

| **Variables (median [IQR])** | **Total** | **Hospitalization Cost Class** | |
|---|---|---|---|
| | | **High-cost** | **Low-cost** |
| Age | 34 [25 - 44] | 37 [29 - 49] | 33 [24 - 43] |
| Transplant_total | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Scheduled consult | | | |
| Consult_semester1 | 3 [1 - 6] | 4 [1 - 6] | 3 [1 - 6] |
| Consult_semester2 | 1 [0 - 3] | 1 [0 - 4] | 0 [0 - 3] |
| Consult_semester3 | 0 [0 - 2] | 0 [0 - 2] | 0 [0 - 1] |
| Consult_semester4 | 0 [0 - 0] | 0 [0 - 1] | 0 [0 - 0] |
| Consult_semester5 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Consult_semester6 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Emergency consults | | | |
| Emergency_year1 | 1 [0 - 3] | 1 [0 - 4] | 1 [0 - 3] |
| Emergency_year2 | 0 [0 - 1] | 0 [0 - 1] | 0 [0 - 1] |
| Emergency_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Ordinary exams | | | |
| Exam_semester1 | 7 [0 - 20] | 12 [1 - 28] | 6 [0 - 18] |
| Exam_semester2 | 0 [0 - 9] | 0 [0 - 15] | 0 [0 - 8] |
| Exam_semester3 | 0 [0 - 2] | 0 [0 - 4] | 0 [0 - 1] |
| Exam_semester4 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Exam_semester5 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Exam_semester6 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Hemodialysis therapy | | | |
| Hemodialysis_year1 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Hemodialysis_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Hemodialysis_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Image exams | | | |
| Image_month1 | 0 [0 - 1] | 0 [0 - 1] | 0 [0 - 1] |
| Image_month2 | 0 [0 - 1] | 0 [0 - 1] | 0 [0 - 1] |
| Image_month3 | 0 [0 - 0] | 0 [0 - 1] | 0 [0 - 0] |
| Pulsotherapies | | | |
| Pulsotherapies_year1 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Pulsotherapies_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Pulsotherapies_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Cardiovascular procedurec | | | |
| Cardiovascular_year1 | 0 [0 - 0] | 0 [0 - 1] | 0 [0 - 0] |
| Cardiovascular_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Cardiovascular_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Diabetes procedure | | | |
| Diabetes_year1 | 0 [0 - 0] | 0 [0 - 1] | 0 [0 - 0] |
| Diabetes_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Diabetes_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Chronic Obstructive procedure | | | |
| Obstructive_year1 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Obstructive_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Obstructive_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Neoplasm procedure | | | |
| Neoplasm_year1 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Neoplasm_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Neoplasm_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Chronic Musculoskeletal procedure | | | |
| Musculoskeletal_year1 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Musculoskeletal_year2 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |
| Musculoskeletal_year3 | 0 [0 - 0] | 0 [0 - 0] | 0 [0 - 0] |

Table 4.3: Descriptive statistics for binary variables.

| **Variables [count (column %)]** | **Total** | **Hospitalization Cost Class** | |
|---|---|---|---|
| | | **High-cost** | **Low-cost** |
| Total dataset | 3578 (100) | 715 (20) | 2863 (80) |
| Flag_Surgery | | | |
| Clinic | 1,706 (48) | 220 (31) | 1,486 (52) |
| Surgery | 1,872 (52) | 495 (69) | 1,377 (48) |
| Flag_Genetic | | | |
| No | 3,541 (99) | 701 (98) | 2,840 (99) |
| Yes | 37 (1) | 14 (2) | 23 (1) |
| Cardiovascular_flag | | | |
| No | 2,769 (77) | 456 (64) | 2,313 (81) |
| Yes | 809 (23) | 259 (36) | 550 (19) |
| Diabetes_flag | | | |
| No | 2,716 (76) | 440 (62) | 2,276 (79) |
| Yes | 862 (24) | 275 (38) | 587 (21) |
| Obstructive_flag | | | |
| No | 3,429 (96) | 650 (91) | 2,779 (97) |
| Yes | 149 (4) | 65 (9) | 84 (3) |
| Neoplasm_flag | | | |
| No | 3,509 (98) | 691 (97) | 2,818 (98) |
| Yes | 69 (2) | 24 (3) | 45 (2) |
| Musculoskeletal_flag | | | |
| No | 3,208 (90) | 602 (84) | 2,606 (91) |
| Yes | 370 (10) | 113 (16) | 257 (9) |
| Flag_psycho_year1 | | | |
| No | 3,392 (95) | 667 (93) | 2,725 (95) |
| Yes | 186 (5) | 48 (7) | 138 (5) |
| Flag_psycho_year2 | | | |
| No | 3,510 (98) | 700 (98) | 2,810 (98) |
| Yes | 68 (2) | 15 (2) | 53 (2) |
| Flag-psycho_year3 | | | |
| No | 3,558 (99) | 711 (99) | 2,847 (99) |
| Yes | 20 (1) | 4 (1) | 16 (1) |
| Gender | | | |
| Female | 2,070 (58) | 371 (52) | 1,699 (59) |
| Male | 1,508 (42) | 344 (48) | 1,164 (41) |
| Admission_Type | | | |
| Scheduled | 2,598 (73) | 525 (73) | 2,073 (72) |
| Emergency | 980 (27) | 190 (27) | 790 (28) |

The association between the variables was assessed through a correlation analysis. The heatmap in Figure 4.5 presents the correlation among pairs of variables according to the Spearman's rank correlation coefficient (see equation 3-1)[7]. The variables are not highly correlated to each other, in general. The most notable pairs of variables regarding the Spearman's coefficient are the temporal variables of "Scheduled consults" and "Ordinary exams", which are moderately correlated. The association among variables of the same type is related to a temporal effect that, in the context of this research, may indicate an increase or decrease in the usage of a specific medical resource. On the other hand, those types of exams and consults may have a medical association that the execution of one leads to the other (i.e., correlation coefficient between "Exam_semester6" and "Consult_semester6" equals 0.762).

[7] Appendix B presents the entire correlation matrix

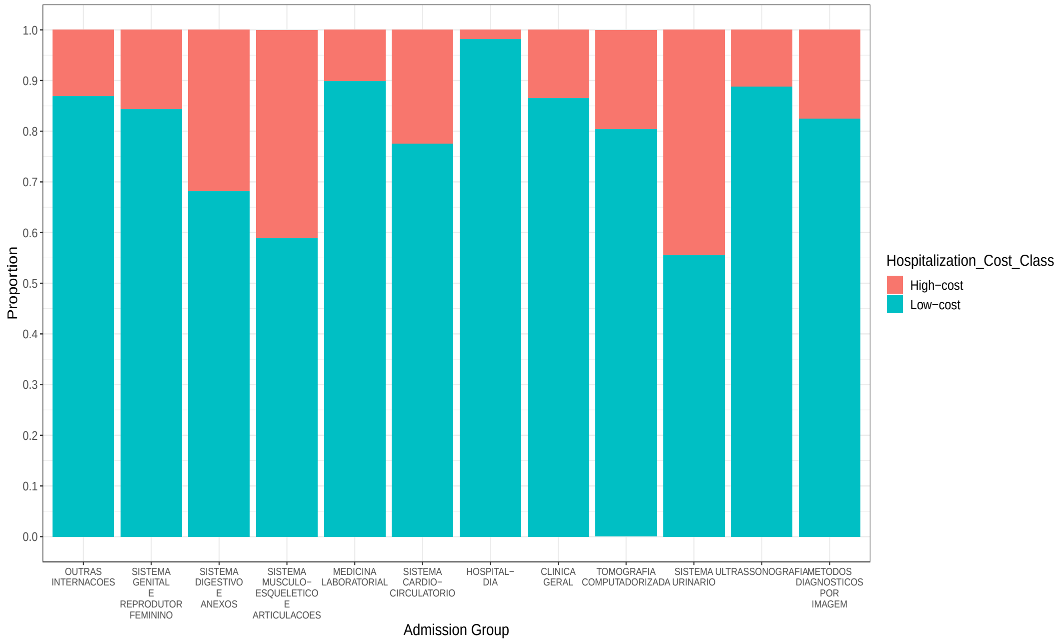


Figure 4.4: Proportion of high-cost hospitalizations for the 12 most frequent "Admission Group".

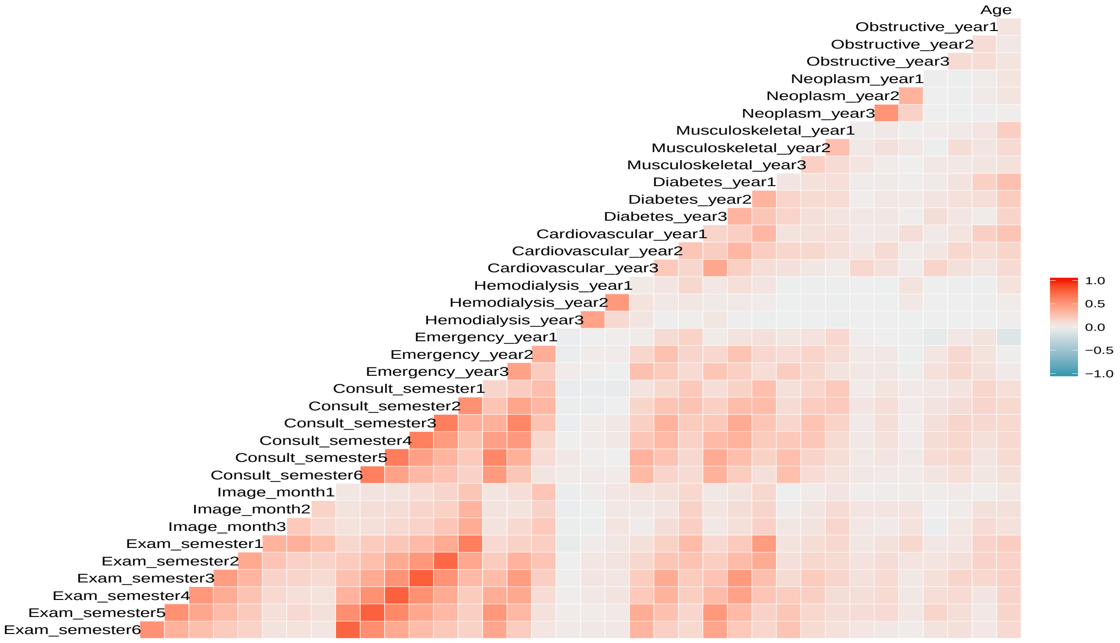


Figure 4.5: Heatmap of predictors' correlation.

### 4.1.3 Data Preparation

After the exploratory analysis the data is prepared according to the business and the ML technique requirements. As mentioned before, the rows with missing values or "Hospitalization Cost" values under zero were removed. Also, there are numeric variables that present extreme values indicating the presence of outliers; in agreement with the organization domain experts, those values were not removed since, in the context of this study, they are useful for the predictive task.

As explained in Subsection 4.1.2, the variables related to past medical resource consumption present different granularities. According to Modrid et al. [23], using temporal variables with a higher granularity (e.g., each individual exam taken in the past three years) could improve the ML model's performance, since this variables expose more information than using aggregated variables (e.g., the sum of exams in the past three years).Although the available variables do not present the degree of granularity showed in the referred study (monthly variables through two years), each set of temporal variables was aggregated into a variable that summarizes the whole period. Thus enabling further experiments comparing the "Temporal dataset" (i.e., the dataset comprising the variables with higher granularity) against the "Aggregated dataset" (i.e., the dataset comprising the aggregated variables). Table 4.4 describes the creation of the new aggregated variables. In general, counting variables were created by summing up numerical temporal variables (e.g., bi-annual count of elective consults), and binary variables were created by applying an OR operator to the corresponding binary temporal variables (e.g., Psychological consults). A particular case is the set of temporal variables related to Chronic diseases, which values, as described in the data dictionary (Appendix A), correspond to the sum of procedure complexities that the patient has taken each year for a specific disease (1 for simple procedures, 3 for medium-complex procedures, and 6 for complex procedures). This variable could lead to some misinterpretations, e.g., a value of 6 for a particular disease in a year could indicate six simple procedures, two medium-complex procedures, or even one complex procedure. Therefore, a binary variable was created, indicating whether there was at least one procedure (of any complexity) in any of the three years.

These new aggregated variables are a linear combination of the temporal ones, and then it is expected that they are highly correlated. Therefore both temporal and aggregated variables will not improve the model's performance if they are used together but increase the complexity and computational cost. Also, the main objective is to test whether temporal variables provides more

Table 4.4: Feature Engineering. Aggregated Medical Resource Consumption variables.

| **Medical Resource Consumption** | **Actual granularity** | **New aggregated variable name** | **New aggregated value** |
|---|---|---|---|
| Exams | bi-annual count | Exams total | sum of semester counts |
| Image exams | monthly (3 month) count | Image exams total | sum of monthly counts |
| Psychological consults | annual flag | Psychological consults flag | 1 : if any annual flag = 1;<br>0 : otherwise |
| Elective consults | bi-annual count | Elective consults total | sum of semester counts |
| Emergency consults | annual count | Emergency consults total | sum of annual counts |
| Hemodialysis | annual count | Hemodialysis total | sum of annual counts |
| Pulsotherapies | annual count | Pulsotherapies total | sum of annual counts |
| Chronic disease | annual cumulative sum | Chronic disease flag | 1 : if any annual cumulative sum > 0;<br>0 : otherwise |

information than aggregated. Hence, two data sets are created (temporal and aggregated) for each set of variables, and both datasets use the common variables of "Patients Characteristics" and "Patient Hospitalization" (see table 4.1).

The numeric variables in the dataset describe the patient's age, the number of particular events before the hospitalization, and the existence of a complex procedure for a "Chronic disease". Therefore, in order to avoid different scales among the dataset variables (which could further impact the predictive techniques), the predictors were equally scaled into the range [0, 1] using the equation[8]:

$$x'_i = \frac{x_i - min(X)}{max(X) - min(X)} \tag{4-1}$$

where $x'_i$ and $x_i$ are the scaled and original $i$ value of predictor $X$, respectively. Moreover, the variable "Admission Group" is categorical, and was transformed into a "dummy" variable as an input requirement for the predictive models. Then, as the "Admission Group" have 34 categories, 34 binary variables are created.

### 4.1.4 Data Modeling

Once the dataset is prepared according to the ML techniques and the business specifications, an experiment route is designed considering different scenarios. First, for the learning procedure, the available data was divided for training and validation, using 5-fold cross-validation with a stratified random sampling splitting approach. Then, six scenarios were considered, which includes the two sets of variables discussed in the previous Section (named "Temporal" and "Aggregated") and three ML techniques (see Appendix C for

[8] The scaling process was performed during the cross-validation procedure, aiming to provide reliable results.

a detailed description of the ML techniques settings). Table 4.5 summarizes those scenarios.

Table 4.5: Experimental Scenarios and results

| **Scenario** | **Machine Learning technique** | **Dataset** | **AUROC** | **F-score($\beta = 2$)** | **Sensitivity** | **PPV** | **Threshold** | **TP** | **FP** | **FN** | **TN** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ridge Regression | Temporal | 0.740 | 0.586 | 0.712 | 0.344 | 0.189 | 102 | 194 | 41 | 378 |
| 2 | | Aggregated | 0.748 | 0.586 | 0.709 | 0.346 | 0.185 | 101 | 192 | 42 | 381 |
| 3 | LASSO Regression | Temporal | 0.740 | 0.585 | 0.713 | 0.340 | 0.180 | 102 | 198 | 41 | 374 |
| 4 | | Aggregated | 0.746 | 0.586 | 0.715 | 0.340 | 0.179 | 102 | 198 | 41 | 374 |
| 5 | CART | Temporal | 0.666 | 0.497 | 0.571 | 0.328 | 0.166 | 82 | 170 | 61 | 402 |
| 6 | | Aggregated | 0.655 | 0.509 | 0.615 | 0.301 | 0.139 | 88 | 214 | 55 | 358 |

Each row of table 4.5 describes a scenario and their corresponding results[9]. The performance was evaluated first for each scenario from an overall perspective using AUROC, to select later the threshold[10] that minimizes the distance between the ROC curve and the top-left corner in the ROC Graph in Figure 4.6 (Sensitivity = Specificity =1). Then, for this optimum threshold, it is presented the corresponding confusion matrix with the derived metrics of interest ($F_{\beta=2}$, Sensitivity, and PPV). The first comparison seeks to define whether it is better to use the "Temporal" or "Aggregated" dataset comparing the error rates of each pair of models. The comparison is performed assessing the statistical difference between the two models errors[11]. For RR and LASSO, the p-values were higher than the significant threshold 0.05, indicating insufficient evidence to reject the null hypothesis ($H_0$) that the two models have the same error rates. This statistical interpretation indicates that the dataset with the temporal variables does not improve significantly the performance of the models. Therefore the dataset with aggregated variables is preferred, and thus the complexity of the models and future computational costs are reduced. However, was detected a statistical difference in the performance when using the CART technique, which indicates that the error rates are different, and the results in the contingency table indicate that CART with the "Temporal" dataset performs better. The behavior is also reflected in the confusion matrix (table 4.5), which indicates that the detected difference in the error rates relates to the FP. This result does not change the conclusion that the "Aggregated" data set is preferred, because the model with the "Temporal" variables only outperforms the one with the "Aggregate" variables with respect to the PPV, which is not the priority of the organization.

Among the ML techniques, CART had the poor performance regarding the AUROC, while LASSO and Ridge regression perform similarly. In this

[9] All values including the Confusion Matrix cells correspond to the average value across the CV iterations.

[10] The limit value above which the predicted probability of being of the positive class classifies an observation as such.

[11] Refer to Appendix C for all details of the learning process and the statistical test carried out for the models' evaluation.

context, LASSO regression is defined as the best model (using the "Aggregated" data set) over the Ridge, due to its ability to zero the coefficients of the variables that do not aggregate information and thus reduce the complexity of the model. According to the confusion matrix, the model is correctly predicting on average more than 71% of the truly high-cost hospitalizations (102 on average), which is the main concern of the organization. However, this comes to the cost of a 34% of precision (PPV), in other words, to accomplish a 71% of detection, the model wrongly classifies as high-cost the 34,6% (FPR) of the truly low-cost hospitalizations.

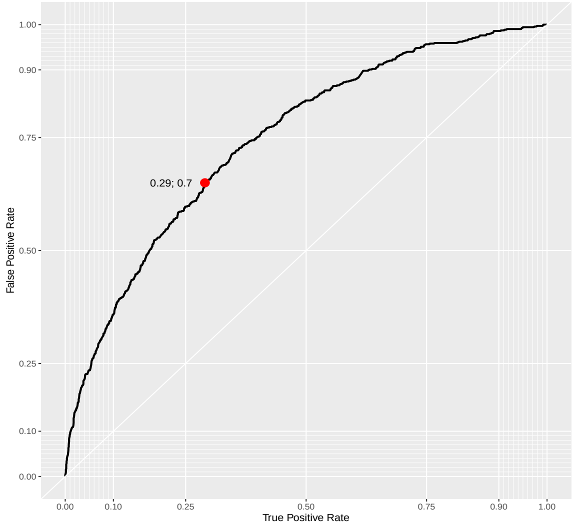


Figure 4.6: ROC curve for LASSO model using the "Aggregated" dataset. The red dot indicates the optimum threshold that generates the confusion matrix in table 4.5.

### 4.1.5 Evaluation

The evaluation phase defines whether the business goals are achieved, or further refinement of the design cycle is necessary before the artifact's deployment. First, we assess the relative contribution of each variable considered in the model to the final prediction. For that, the importance of each variable is extracted from the trained model (in the case of LASSO for classification, this importance refers to the magnitude[12] of the coefficients in the penalized logistic regression).

Figure 4.7 displays the top 20 variables, sorted in decreasing order of their coefficients magnitude. By analyzing the Graph, the organization domain

[12] Note that magnitude here refers to the size of the effect but do not relate if there is a positive or negative association with the log-odds. See Appendix C for a description of the coefficient.

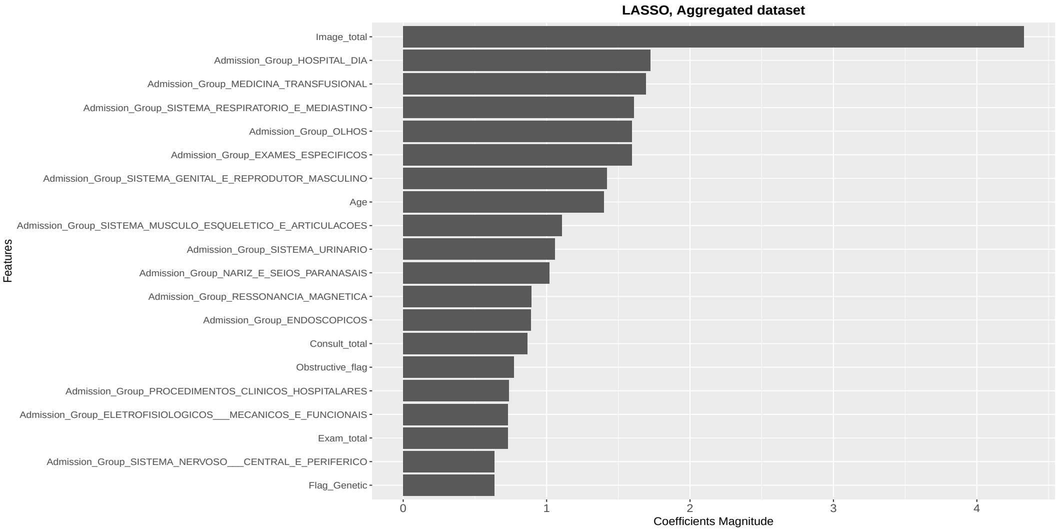


Figure 4.7: Variable importance for LASSO model using the "Aggregated" dataset. X-axis represents the absolute value of the variable coefficient.

expert noticed some interesting results that were not expected. First, the high association of the "HOSPITAL_DIA" group. From the domain expert perspective, this "Admission Group" has little significance since patients do not get admitted for more than one day and have a relatively low cost. The unexpected result also contrasts with the Graph 4.4, where this group of hospitalizations actually have a low proportion of high-cost, and also the three "Admission Group" pointed in Section 4.1.2 with the highest proportions have low importance in the fitted model. The discrepancies in the exploratory analysis results and the domain expert business knowledge with respect to the individual effect of the predictors are not conclusive since interactions among variables can occur, but it deserves further analysis in order to understand this behavior. To do that, the original cost variable ("Hospitalization Cost") is again analyzed, but this time by "Admission Group" instead of the whole dataset.

The Graph in Figure 4.8 exhibits the distribution of the "Hospitalization Cost" within each "Admission Group," and the blue dashed line represent the high-cost cutoff ($R\$ \ 10,200$) defined using Pareto's principle in Section 4.1.2. From the Graph, it turns out that hospitalizations for each "Admission Group" have different cost distribution, mainly the "HOSPITAL_DIA" group, which has the lowest median cost value (as pointed out by the domain expert). Then according to the high-cost definition, for this group just a few values are above the defined cut-off, which can explain why it receives a high coefficient in the fitted model. On the other hand, the groups with the larger proportions of high-cost hospitalizations are because the cutoff line is under their 75 percentile.

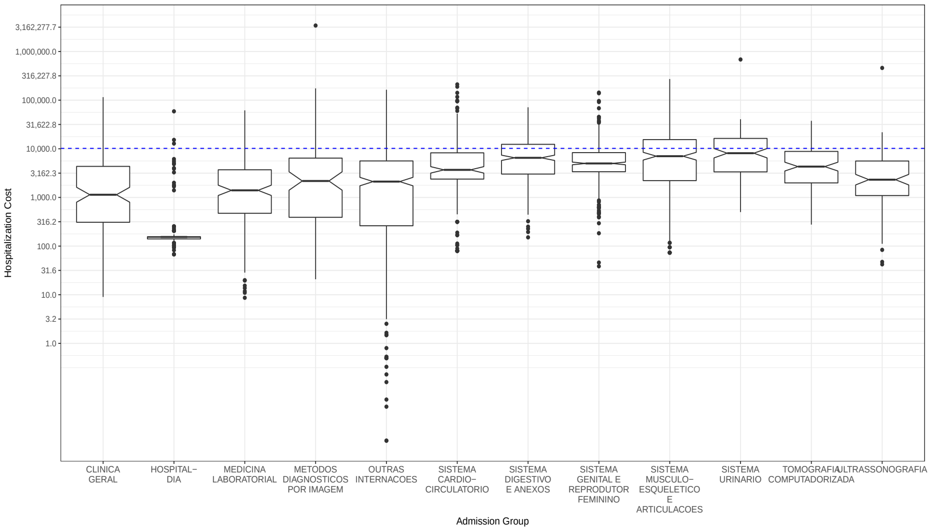


Figure 4.8: "Hospitalization Cost" for the 12 most frequent "Admission Groups". The y-axis is in logarithm (base 10) scale

The behavior observed in the previous analysis indicates that the definition of high-cost hospitalization following the Pareto's principle is not adequate in this context. The point is that the cost value of a hospitalization should not be used to define high-cost among different admission motives, which costs distributions differ depending on the nature of the procedures performed. This conclusion was reached in agreement with the organization, so it was decided to redefine the response variable, and that a new cycle should be initiated. Also, at this point in the study, the entire dataset comprising all the population attended by the organization is available and would be used in the following cycle.

## 4.2
## Second DSR cycle: Prediction of high-cost hospitalizations over the complete dataset

Figure 4.9 summarizes the activities and results obtained when instantiating the Second cycle of DSR framework presented in the previous Chapter. It is important to notice important differences when compared to the overview of the First cycle, presented in Figure 4.1:

- The entire data set with all the hospitalizations is now available[13], including new variables regarding past hospitalizations.
- Residual hospitalizations and "Admission Subgroup" out of the scope of the study were removed.

[13] The data set with all the hospitalizations was under development when the project started, thus a sample was provided to conduct the First DSR cycle.

- High-cost hospitalizations were defined at the "Admission Subgroup" level, applying an outlier detection approach.
- New variables were created for prior costs.
- Predictive models built for each "Admission Subgroup".

As in the First cycle, the following Subsections describe each phase of this Second DSR cycle, in detail.

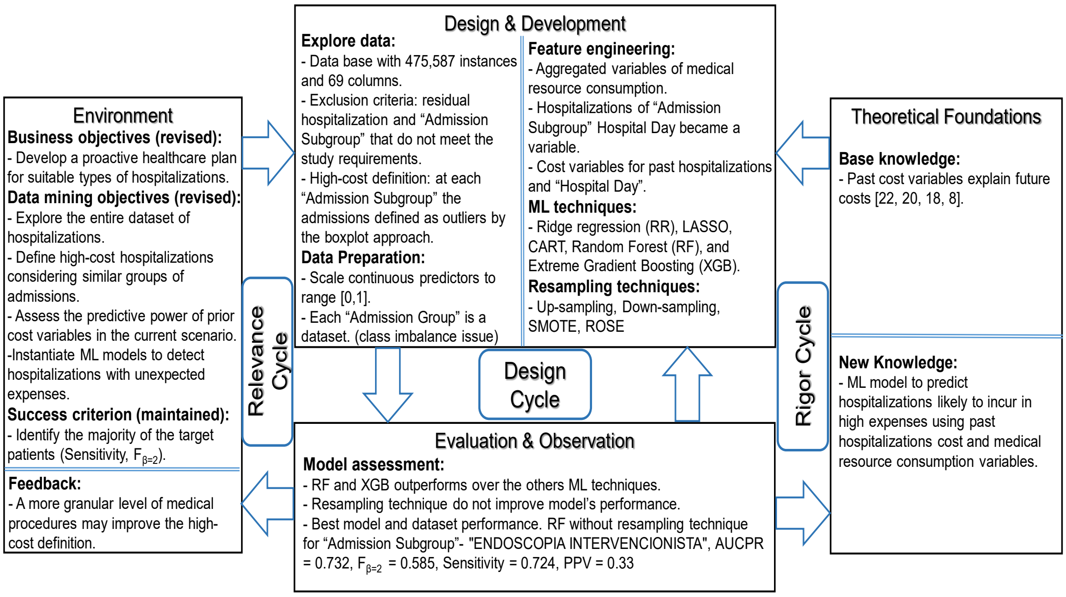

Environment
Business objectives (revised):
- Develop a proactive healthcare plan for suitable types of hospitalizations.
Data mining objectives (revised):
- Explore the entire dataset of hospitalizations.
- Define high-cost hospitalizations considering similar groups of admissions.
- Assess the predictive power of prior cost variables in the current scenario.
-Instantiate ML models to detect hospitalizations with unexpected expenses.
Success criterion (maintained):
- Identify the majority of the target patients (Sensitivity, $F_{\beta=2}$).
Feedback:
- A more granular level of medical procedures may improve the high-cost definition.
Relevance Cycle
Design & Development
Explore data:
- Data base with 475,587 instances and 69 columns.
- Exclusion criteria: residual hospitalization and “Admission Subgroup” that do not meet the study requirements.
- High-cost definition: at each “Admission Subgroup” the admissions defined as outliers by the boxplot approach.
Data Preparation:
- Scale continuous predictors to range [0,1].
- Each “Admission Group” is a dataset. (class imbalance issue)
Feature engineering:
- Aggregated variables of medical resource consumption.
- Hospitalizations of “Admission Subgroup” Hospital Day became a variable.
- Cost variables for past hospitalizations and “Hospital Day”.
ML techniques:
- Ridge regression (RR), LASSO, CART, Random Forest (RF), and Extreme Gradient Boosting (XGB).
Resampling techniques:
- Up-sampling, Down-sampling, SMOTE, ROSE
Design Cycle
Evaluation & Observation
Model assessment:
- RF and XGB outperforms over the others ML techniques.
- Resampling technique do not improve model’s performance.
- Best model and dataset performance. RF without resampling technique for “Admission Subgroup”- "ENDOSCOPIA INTERVENCIONISTA", AUCPR = 0.732, $F_{\beta=2}$ = 0.585, Sensitivity = 0.724, PPV = 0.33
Rigor Cycle
Theoretical Foundations
Base knowledge:
- Past cost variables explain future costs [22, 20, 18, 8].
New Knowledge:
- ML model to predict hospitalizations likely to incur in high expenses using past hospitalizations cost and medical resource consumption variables.

Figure 4.9: Second Cycle of the DSR and Data Science Methodology

## 4.2.1 Business Understanding

The development of this cycle, as stated at the end of Section 4.1.5, followed the need to redefine the concept of a "high-cost hospitalization", as well as to consider the complete dataset made available by the organization. In this context, the objectives of the Second cycle are: (i) to explore the new dataset, with the entire population of hospitalizations of the past 3 years; (ii) to evolve the precise definition of a high-cost hospitalization considering similar groups of admissions (that is, hospitalization with a high risk of incurring in unexpected high-cost, given its admission group), and (iii) to build ML models that successfully predict such high-cost hospitalizations. The organization expectation is to have predictive models that capture most of the high-cost hospitalizations (with a high risk of having an unexpected high-cost considering the hospitalizations of the same admission group), while at the same time reducing the number of low-cost hospitalizations that are misclassified as high-cost. As in the First cycle the success criterion is measured maximizing the $F_\beta$ metric (see equation 3-13) with $\beta = 2$. In addition, the Sensitivity and PPV, together with the corresponding confusion matrix complements the performances evaluation process.

The need for evolving the definition of a high-cost hospitalization arose from the conclusion, obtained in the First DSR cycle, that it is necessary to stop analyzing the cost in isolation. The magnitude of the cost value is not enough, since the cost distributions vary a lot among distinct admission groups. Instead, it would be more useful to predict which hospitalizations have an unexpected cost according to some reference value. The next Section will address this definition.

### 4.2.2 Data Understanding

The complete dataset containing the entire population has $475,587$ hospitalizations of a total of $269,976$ patients, and 69 features describing characteristics of the hospitalizations at the moment of the admission (e.g., "Admission Subgroup", "Flag Surgery", "Admission Type") and a set of medicals events in the past three years. As can be noticed, there are more features compared to the sample data used in the First DSR cycle. The new features are from the group of past medical resource consumption, and denote the number of physiotherapies consults and past hospitalizations for the same patient. In addition, there is a new variable that gathers related hospitalizations at a more granular level in relation to the "Admission Group", which is called "Admission Subgroup"[14]. This information is displayed in table 4.6 and more detailed in the dictionary of variables in Appendix A.

To start the analysis of the new dataset, the same data selection filters of the First DSR cycle were applied. Also, duplicated registries were encountered and removed (see Figure 4.10). Graph 4.11 illustrates the behavior of the hospitalization cost for each "Admission Subgroup". To do that, the box-plots were substituted by the hospitalization's instances represented as dots, and plotted with some random dispersal (jitter) to avoid overlapping. The idea is to observe if there are unknown patterns in the cost variable distribution.

Graph 4.11 plots the "Hospitalization Cost" of the 10 most frequent "Admission Subgroup". The distribution cost of each subgroup presented a set of observations in the lower extreme of their distributions, which are

[14] There is a total of 37 groups and 196 subgroups.

Table 4.6: Group of variables whole dataset

| Patient Characteristics | Medical Resource Consumption | Patient Hospitalization |
|---|---|---|
| Gender | Ordinary exams | Admission Group |
| Age | Image exams | Admission Type |
| | Psychological consults | Flag Surgery |
| | Scheduled consults | Hospitalization Cost |
| | Emergency consults | *Admission Subgroup |
| | Hemodialysis therapies | |
| | Pulsotherapies | |
| | Transplant | |
| | Chronic disease | |
| | *Physiotherapies consults | |
| | *Surgical hospitalizations | |
| | *Clinical hospitalizations | |

*New variables.

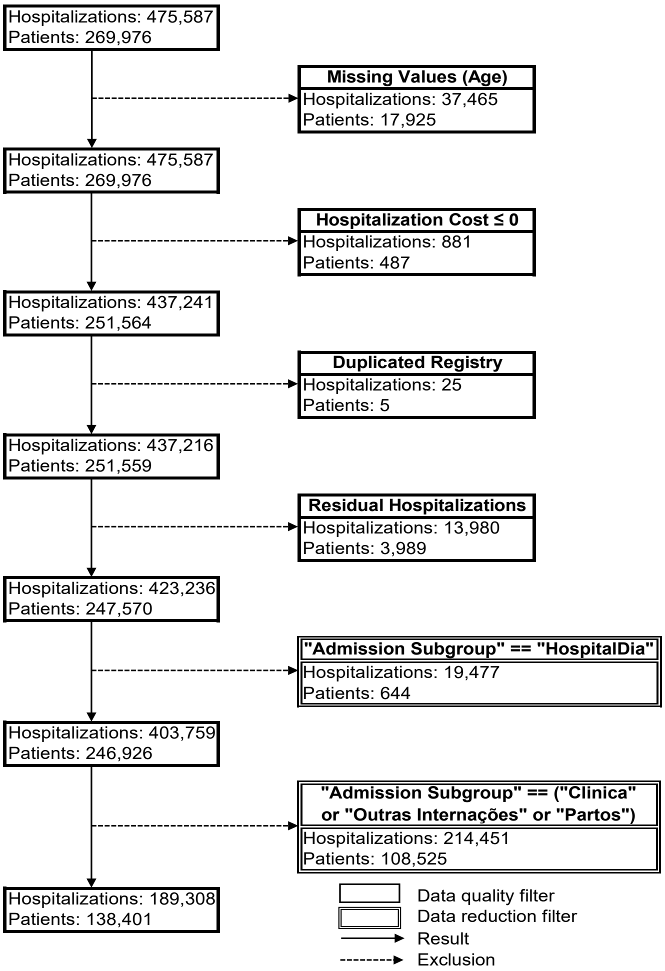


Figure 4.10: Data quality and reduction filters for the entire dataset.

significantly different from their median value (including values very close to zero). This behavior is more notable in subgroups with higher variability, such as "OUTRAS INTERNACOES" (where the median value is $R\$ \ 2,550$, and 25 % of the hospitalizations with the lowest costs are between $R\$ \ 0.01$ and $R\$ \ 304$). In a further analysis, the domain expert of the organization concluded that this is due to the existence of residual hospitalizations in the dataset. From the organization perspective, residual hospitalizations are delayed payments whose registries appear as individual hospitalizations while, in reality, they represent residual payments that could not be linked to the original hospitalization. In other words, they are part of some hospitalization that happened before, with no link to the original one. Then, in agreement with the organization domain expert, the residual hospitalizations within each "Admission Subgroup" were removed.

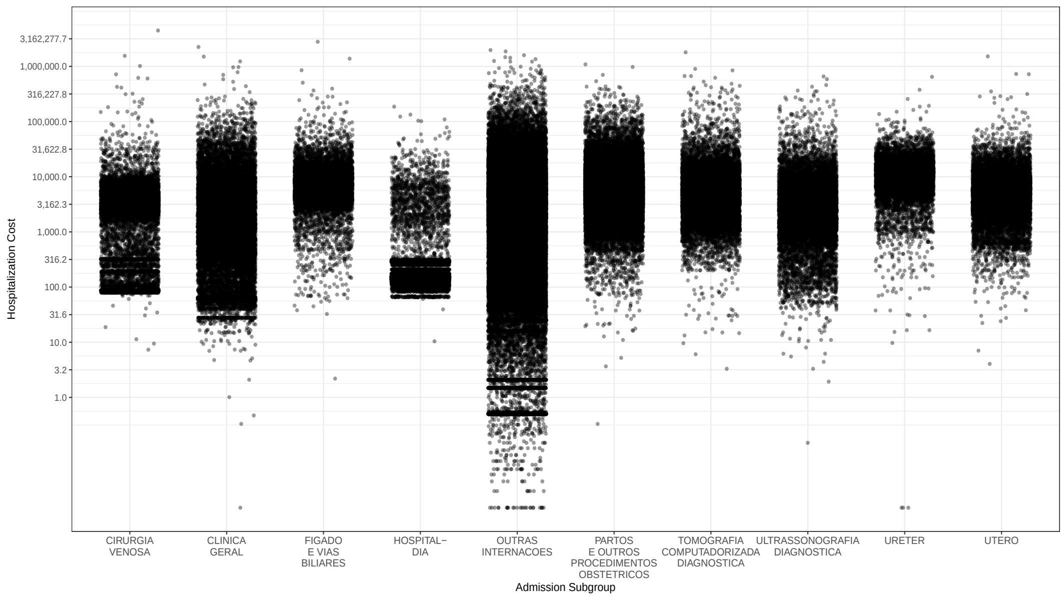


Figure 4.11: "Hospitalization Cost" for the 10 most frequent "Admission Subgroup". The y-axis is in logarithm (base 10) scale for better visualization.

In order to detect those residual hospitalizations, an outlier detection procedure was performed in the lower extreme of each subgroup distribution. The outlier definition followed the boxplot approach, where every point lower than the resulting difference between the 25 percentile and 1.5 times the IQR is identified as an outlier. One issue with this approach is that the distributions of cost are heavily right-skewed[15], thus most subgroups do not present outliers following the previous definition. Thus, we first apply a

[15] This is a characteristic of cost data in the healthcare context, just a minority of the population generate large expenses.

logarithmic transformation of the "Hospitalization Cost" variable (to decrease the skew) and then apply the outlier definition described above. Graph 4.12 depicts the resulting residual observations for the 10 most frequent "Admission Subgroup", which were then removed. Figure 4.10 shows the amount of removed data.

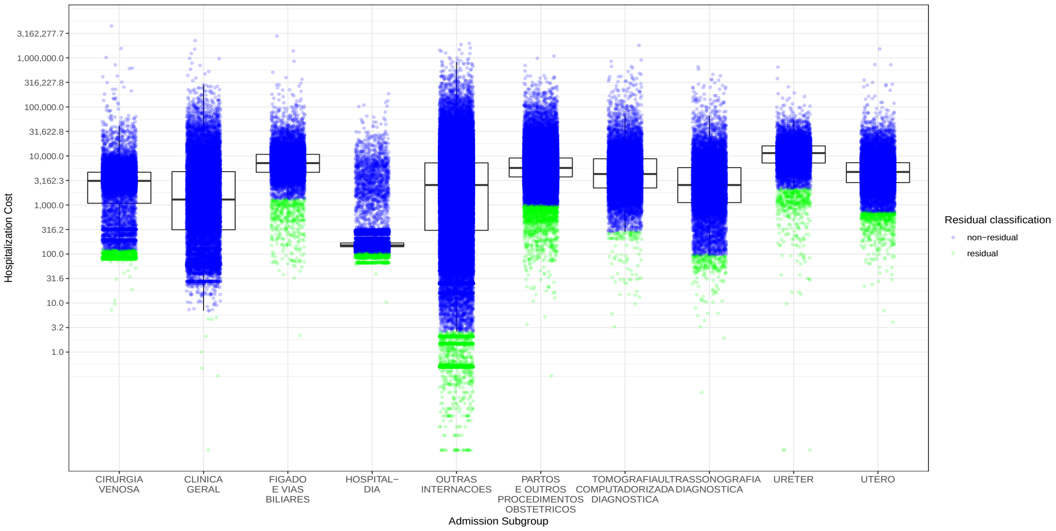


Figure 4.12: "Hospitalization Cost" for the 10 most frequent "Admission Subgroup". The green point are the residual hospitalizations. The y-axis is in logarithm (base 10) scale for better visualization.

At this point, the organization domain experts analyzed the "Admission Subgroup" categories, searching for inconsistencies with the research goals, and concluded that some of them should not be considered. First, the subgroups named "Clinica", "Outras Internações" and "Partos" should be discarded because they group very heterogeneous sets of procedures (note their variability in Graph 4.12). Second, the subgroup "HospitalDia" ("Daily Hospital") should be discarded as hospitalizations, for the purposes of the current research, since they typically refer to procedures that do not require the patient to be admitted for more than one day. However, they could be an important indicator for some possible hospitalization of the same patient in the future. Thus, the observations of this subgroup were removed and this event was transformed into a new variable, as described in the next Section. The subgroup removal procedure is one of the data reduction steps described in Figure 4.10.

#### 4.2.2.1
#### Defining the class variable

The task redefines the "Hospitalization cost class" variable, setting which hospitalizations should be considered as high- or low-cost for the data modeling phase. In this Second DSR cycle, a high-cost hospitalization is defined as the one having an unexpected much higher cost when compared to similar hospitalizations.

For this purpose, the "Admission Subgroup" variable was used as a grouping criterion, thus resulting in more homogeneous subsets of hospitalizations. Each subset has a distinct threshold (or cutoff) value for defining high-cost hospitalizations, set by applying an outlier detection procedure that is analogous to the one adopted to define residual hospitalizations, but this time looking in the upper extreme of the cost distribution. The rationale for this definition is that these outliers effectively represent hospitalizations which costs are unexpectedly high. Graph 4.13 illustrates the different cutoff values for each "Admission Subgroup" of the "Admission Group" named "SISTEMA DIGESTIVO E ANEXOS" ("Digestive System and Annexes").

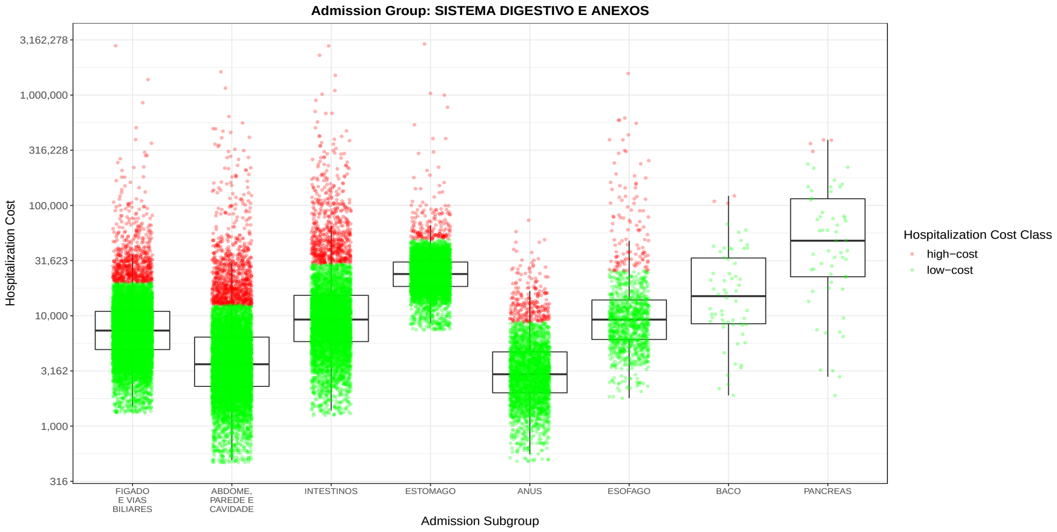


Figure 4.13: "Hospitalization cost class" definition for each "Admission Subgroup" of the "Admission Group" named "SISTEMA DIGESTIVO E ANEXOS".

The described approach for defining high-cost hospitalizations generates new challenges to the project that should be addressed in the preparation phase. For example, as depicted in Graph 4.13, the same cost value could represent both high-cost and low-cost hospitalizations for different subgroups. Moreover, the definition of high-cost hospitalizations as outliers significantly reduces the number of observations belonging to the high-cost category, thus creating a class imbalance issue, as pictured in the two subgroups to the right of Graph 4.13 (see Appendix D table D for a detailed description of each "Admission Subgroup").

### 4.2.3 Data Preparation

The preparation phase in this Second DSR cycle followed most tasks performed in the First cycle. All observations meeting the exclusion criteria described in the previous sections were removed, and the presence of outliers in this context refers to values that could have valuable information, thus were not removed. Also, according to the conclusion reached in the First cycle that temporal variables do not improve the model's performance, only aggregated variables are considered in the Second cycle.

Table 4.7: Feature Engineering. Aggregated Medical Resource Consumption variables.

| **Medical Resource Consumption** | **Actual granularity** | **New aggregated variable name** | **New aggregated value** |
|---|---|---|---|
| Physiotherapy | annual count | Physiotherapy total | sum of annual counts |
| Surgery hospitalization | bi-annual count | Surgery hospitalization total | sum of semester counts |
| Surgery hospitalization cost | bi-annual sum | Surgery hospitalization cost total | sum of semester costs |
| Clinic hospitalization | bi-annual count | Clinic hospitalization total | sum of semester counts |
| Clinic hospitalization cost | bi-annual sum | Clinic hospitalization cost total | sum of semester costs |
| Hospital-Dia | bi-annual count | Hospital-Dia total | sum of semester counts |
| Hospital-Dia cost | bi-annual sum | Hospital-Dia cost total | sum of semester costs |

In addition to creating the aggregated variables as described in Table 4.4, a new variable was created to represent the number of "Day-Hospital" hospitalizations that occurred in the past. Also, for each variable describing past hospitalizations of a specific type ("Surgical", "Clinical", and "Day-Hospital") a new variable was created to denote the cost incurred for its corresponding type, as an attempt to break ties among admissions with the same amount of past hospitalizations[16] (Table 4.7 complements the variable aggregations previously described in Table 4.4). Moreover, as in the First cycle, all predictors were scaled into a range $[0, 1]$.

On the other hand, to address the challenges caused by the redefinition of a high-cost hospitalization (described in Subsection 4.2.2), the predictive models will be built at the "Admission Subgroup" level, thus each subgroup will be treated as a different dataset. This decision, in agreement with the organization domain expert, also addresses the need for prioritizing groups of hospitalizations by their frequency, their cost variability or their average cost. Figure 4.14 displays the frequency and average cost for all "Admission Subgroup" with more than 800 hospitalizations. The 47 most frequent subgroups, which account for 88% of the hospitalizations, will be the object of study.[17]

[16] Previous related works argue that past costs variables are valuable predictors of futures cost (see Section 2.2).

[17] Appendix D, Table D, provides a characterization of the "Admission Subgroup".

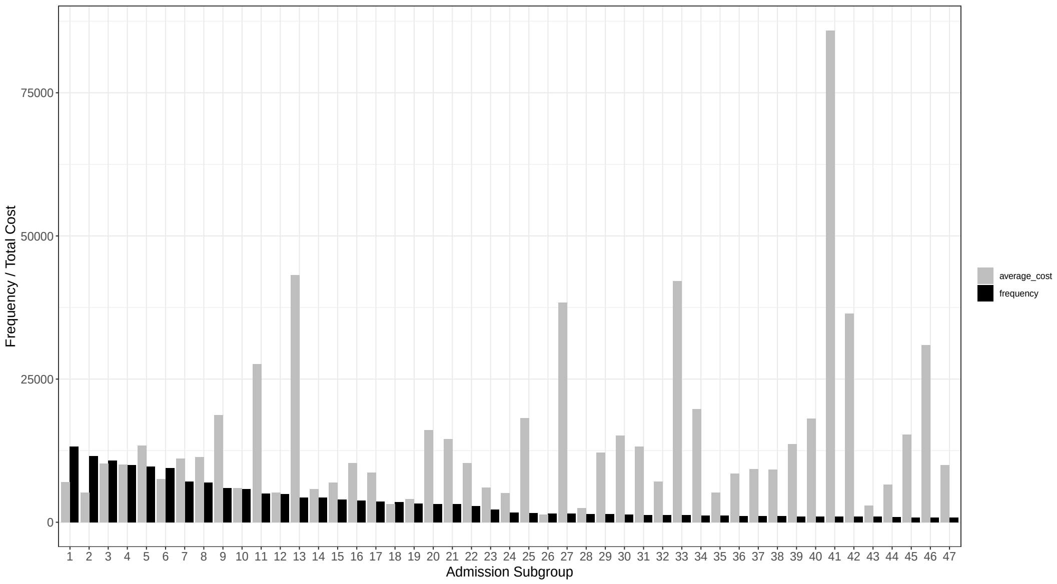


Figure 4.14: "Admission Subgroup" frequency and average cost. The numbers in the x-axis represent each "Admission Subgroup" (see Appendix D for a detailed description).

## 4.2.4 Data Modeling

The learning procedure for all the models in this Second DSR cycle also applied a 5-fold cross validation methodology approach with a stratified random sampling splitting criterion, as in the First DSR cycle. The experiment design comprised a wider variety of scenarios, including other ML techniques and also resampling techniques to treat the class imbalance issue (which is more evident at the "Admission Subgroup" level). It is worth noting that both the resampling technique and data scaling were performed during the CV procedure to the training folds, in order to avoid bias in the subsequent performance metrics. Table 4.8 summarizes these scenarios and Appendix E provides detailed information about the hyperparameters settings for each ML technique.

Model assessment was different from that of the First design cycle, because of the great number of scenarios conducted in this experiment[18]. In summary, the experiment results evaluated: (i) the overall performance of the ML techniques on all subgroups; (ii) the performance of the best ML techniques for all the subgroups under study; (iii) the predictors contribution to the final prediction. The performance results are presented in the following subsections.

[18] Note that for the 47 "Admission Subgroup", 5 ML techniques were trained, each one testing the 4 resampling technique plus the scenario without any, and a set of specific parameters for each ML technique (details in Appendix E) for a total of 1175 scenarios.

Table 4.8: Experimental Scenarios

| Machine Learning technique | Resampling Technique |
|---|---|
| Ridge Regression | Up-sampling |
| Lasso Regression | Down-sampling |
| CART | SMOTE |
| Random Forest | ROSE |
| Extreme Gradient Boosting | |

#### 4.2.4.1 Evaluating ML Techniques Performance among all Admission Subgroup

First, the performance of each ML technique on all datasets was analyzed. Figure 4.15 shows the results from this analysis, summarizing the performance of each ML technique for the best scenario on each dataset. For each combination of "Admission Subgroup" and ML technique, the "best scenario" was defined by the resampling method and specific set of hyperparameters that produced the highest AUCPR. This metric provides and overall measurement of the model performance, and may be interpreted as the average Precision (PPV) when varying the Sensitivity from 0 to 1 at different probabilities thresholds.

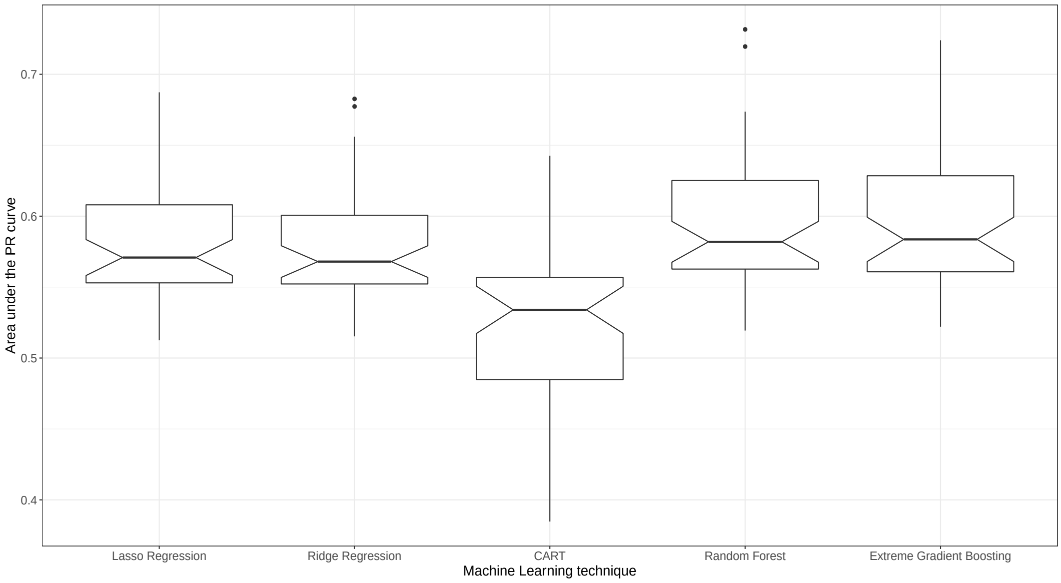


Figure 4.15: Overall ML technique performance across dataset.

The boxplot in Figure 4.15 illustrates that the RF and XGB models had a better overall performance, showing a right-skewed distribution with median

AUCPR of 0.582 and 0.584, respectively[19], and maximum values reaching 0.732 in the case of RF. In contrast, the CART technique presented the lowest median AUCPR (0.534), with a left-skewed distribution and a higher variability among all datasets. It is worth noticing that, in order to compare the magnitude of the models' performance considering AUCPR, the baseline model (i.e., one that predicts both classes randomly) performance depends on the positive (high-cost) class proportion in the dataset. Since the analysis intends to relate an overall performance, a median high-cost proportion of 0.07 with a median frequency of 1638 hospitalizations among all datasets is considered as a baseline model.

### 4.2.4.2 Evaluating ML Techniques Performance for each Admission Subgroup

The next analysis aimed to compare the results of the best ML techniques. For that, the stripchart[20] in Figure 4.16 depicts the AUCPR of the best 5 (for each resampling technique) RF and XGB models, for each "Admission Subgroup"[21]. Grey dots represent the results of XGB, while black dots depict the RF results. Also, the average values and ranges are pictured with a horizontal and a vertical line, respectively.

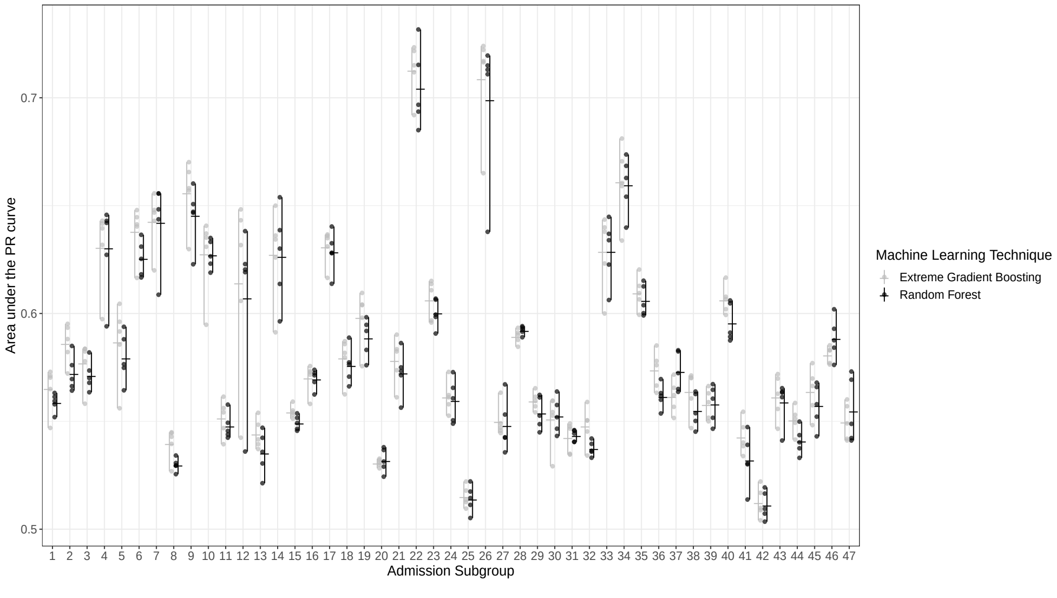


Figure 4.16: Overall "Admission Subgroup" performance using RF and XGB.

[19] The Notch displays the confidence interval around the median calculated as $median \pm 1.58 \times IQR/\sqrt{n}$, where $n$ is the number of samples. Although, not an statistical test non overlapping notch suggest strong evidence that the median values are significantly different.

[20] Unidimensional scatterplot.

[21] The corresponding subgroup name for each number in the x-axis is provided in Appendix D.

Graph 4.16 evidences a great variability in the results for each subgroup, showing different behaviors of the resampling techniques. The lower median (AUCPR = 0.561) performance corresponds to the ROSE technique, and the highest values were obtained using either SMOTE or no resampling technique, with a median performance of 0.573 and 0.576, respectively. There also seems to be no clear winner among the two ML techniques, since they both outperform each other in different datasets.

Moreover, two "Admission Subgroup" stand out over the rest, both presenting AUCPR over 0.70. The "Admission Subgroup" number 22 (named "ENDOSCOPIA INTERVENCIONISTA") has a higher AUCPR value of 0.732 (reached for the RF technique without any resampling technique), which is a considerable improvement compared with a baseline model for this subgroup of 0.113 (2803 observations). On the other hand, in the "Admission Subgroup" number 26 (named "APARELHOS GESSADOS") the best performance corresponds to the XGB technique using the SMOTE resampling procedure, achieving an AUCPR of 0.724 with a ratio of 0.174 high-cost in 1562 hospitalizations.

Although these two subgroups distinguished by the performance obtained, as depicted in Figure 4.14, their average costs are not that interesting for the business goals, mainly for the subgroup 26, which has one of the lowest frequencies and average costs. On the other hand, the "Admission Subgroup" number 33 (named "ACESSOS VASCULARES") presents an elevated average cost. Thus, despite not having a prominent performance, subgroup 33 is a candidate of interest for the study. Its best performance was achieved with XGB and the SMOTE resampling techniques. The AUCPR reached 0.644 and is compared with a baseline model with a performance equal to the subgroup high-cost proportion of 0.106 (1244 observations). The performance metrics and the confusion matrix for subgroups numbers 22 and 33 are provided in Table 4.9.

An interesting conclusion for these subgroups is that both RF and XGB techniques achieve high performances without any resampling technique to deal with the class imbalance problem. This behavior highlights the ability of these two ML algorithms, but also suggests that the techniques used for the imbalance issue were not the most appropriate in this context. Therefore, others could be tested to further refine the models.

In Table 4.9, the highest values for AUCPR are highlighted, and among them stands out the results for the subgroup 22 where the value for $F_{\beta=2}$ almost reach 0.60, with which the model is correctly predicting more than 70 % of the high-cost hospitalizations with around 30 % of precision. These performance rates, and the absolute values depicted in their corresponding confusion matrix,

Table 4.9: Summary of performance of best models for "Admission Subgroup" 22 and 33.

| Dataset | Scenario | AUCPR | Threshold | $F_{\beta=2}$ | Sensitivity | PPV | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|---|---|---|
| **22** | RF | | | | | | | | | |
| | none | **0.732** | 0.159 | 0.585 | 0.724 | 0.330 | 45 | 92 | 17 | 406 |
| | up | 0.715 | 0.383 | 0.590 | 0.724 | 0.339 | 45 | 89 | 17 | 410 |
| | down | 0.697 | 0.519 | 0.585 | 0.750 | 0.310 | 47 | 107 | 16 | 392 |
| | SMOTE | 0.685 | 0.337 | 0.572 | 0.702 | 0.328 | 44 | 91 | 19 | 407 |
| | ROSE | 0.694 | 0.528 | 0.544 | 0.606 | 0.387 | 38 | 84 | 25 | 414 |
| | XGB | | | | | | | | | |
| | none | 0.722 | 0.116 | 0.587 | 0.708 | 0.348 | 44 | 86 | 18 | 413 |
| | up | **0.723** | 0.436 | 0.576 | 0.737 | 0.307 | 46 | 104 | 16 | 394 |
| | down | 0.712 | 0.506 | 0.568 | 0.737 | 0.296 | 46 | 109 | 16 | 389 |
| | SMOTE | 0.715 | 0.170 | 0.579 | 0.769 | 0.291 | 48 | 117 | 14 | 381 |
| | ROSE | 0.692 | 0.945 | 0.563 | 0.771 | 0.271 | 48 | 284 | 14 | 215 |
| **33** | RF | | | | | | | | | |
| | none | 0.637 | 0.183 | 0.464 | 0.552 | 0.283 | 15 | 38 | 12 | 185 |
| | up | 0.634 | 0.358 | 0.486 | 0.751 | 0.202 | 20 | 79 | 7 | 143 |
| | down | 0.623 | 0.471 | 0.480 | 0.733 | 0.202 | 19 | 77 | 7 | 145 |
| | SMOTE | **0.645** | 0.310 | 0.483 | 0.636 | 0.246 | 17 | 53 | 10 | 170 |
| | ROSE | 0.606 | 0.344 | 0.415 | 0.916 | 0.130 | 24 | 166 | 2 | 57 |
| | XGB | | | | | | | | | |
| | none | 0.640 | 0.089 | 0.478 | 0.660 | 0.228 | 17 | 60 | 9 | 163 |
| | up | 0.638 | 0.416 | 0.466 | 0.636 | 0.225 | 17 | 58 | 10 | 164 |
| | down | 0.623 | 0.507 | 0.449 | 0.681 | 0.190 | 18 | 77 | 8 | 146 |
| | SMOTE | **0.644** | 0.197 | 0.494 | 0.697 | 0.228 | 18 | 63 | 8 | 159 |
| | ROSE | 0.600 | 0.001 | 0.432 | 0.632 | 0.191 | 17 | 89 | 10 | 133 |

are within the expected results from the organization perspective to be able to carry out the proactive health plan.

### 4.2.4.3 Evaluating Variable Importance for the best ML Models

In addition, once the model is trained, another possible approach for understanding the high-cost hospitalizations is to analyze the predictors' ability (importance) to explain the response variable. In Graph 4.17 the predictors' importance are ranked for the best results of the two datasets selecting the scenario with the highest $F_{\beta=2}$.

For both RF and XGB, the variable's importance relate to the Impurity Gain (according to the Gini Index) obtained when splitting a tree node on a specific variable. It is interesting that, for both subgroups, the "Admission Type" variable (Scheduled or Emergency) appears in the first positions. This behavior may be caused because the proportions of the "Hospitalizations Cost Class" labels are dominant within the two classes of this variable. In fact, for subgroup 22, emergency admissions have a higher proportion of high-cost hospitalizations (almost 50%), while the majority (93%) of scheduled hospitalizations are of low-cost. It is also worth noting that the new features created in Section 4.2.3, which are related to the past hospitalizations (Surgical, Clinical and Day-Hospital) and their respective costs, provide valuable information for

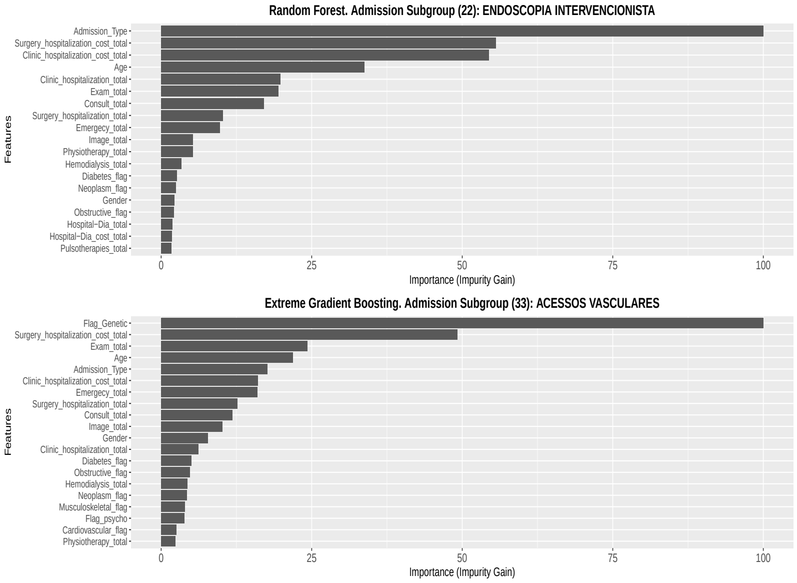


Figure 4.17: Variable importance. Top Graph, "Admission Subgroup" 22 with RF without resampling technique. Bottom Graph, "Admission Subgroup" 33 using XGB and SMOTE.

the learning process.

### 4.2.5 Evaluation

The model evaluation phase delimits the frontier between the Design and Relevance cycles (see Figure 3.9). It designs the point, in the project pathway, where the model design and development are assessed from the organization perspective as the testing field to decide whether additional model refinements are needed. For that, further analyses of the resulting models performance are conducted, this time trying to understand particular cases where the errors are being made, and also their impact.

#### 4.2.5.1 Visualizing the predictions

First, the best performance scenario for subgroup 22 is considered (RF with none resampling technique). For that, Figure 4.18 displays the classification predictions for each testing fold during the CV iterations. The colored dots differentiate the classes predicted by the model, and the blue dashed line denotes the high-cost cutoff value (which divides the observations in their real classes). Then, for example, a red dot (predicted as a high-cost hospitalization) above the blue line indicates a true positive (TP); a red dot below the blue line indicates a false positive (FP); a green dot (predicted as a low-cost hos-

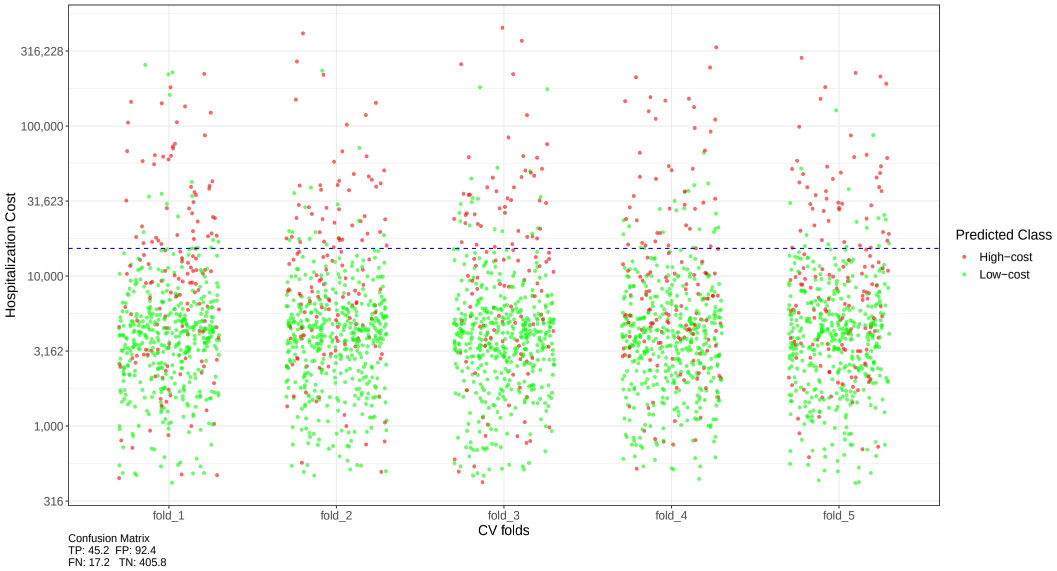


Figure 4.18: Classification predictions for the testing folds during the CV procedure of "Admission Subgroup" 22. The blue dashed line represent the high-cost boundary for this subgroup.

pitalization) below the blue line indicates a true negative (TN); and a green dot above the blue line indicates a false negative (FN). Figure 4.18 effectivelly shows each cell of the Confusion Matrix in contrast with the continuous variable "Hospitalization Cost", making hits and misclassifications explicit. It is worth noting that the Confusion Matrix values are calculated as the average of the classification's results across the CV folds, allowing to test the model using all observations once. In other words, as depicted in Figure 4.18, the CV procedure permits to test the trained model on different patterns of data, increasing the confidence of the performance estimates. In this subgroup, the majority of the high-cost hospitalizations are correctly classified, except for the first fold where, among the five highest costs, just two were predicted as such. Also, an unexpected behavior was observed in the lower extreme of the cost distribution for all folds, where high-cost classifications are wrongly made.

On the other hand, the best performance scenario for subgroup 33 is depicted in Figure 4.19 (XGB using SMOTE). Similar behaviors in the errors are also reflected in this subgroup, although it is good to note that apparently fewer mistakes are made in the area where the hospitalizations have the highest cost. It also highlights the correct detection of the two most expensive hospitalizations, the largest being over 3 million. Then, those observations would be further analyzed, providing some characteristics of the most important predictors (see Figure 4.17).

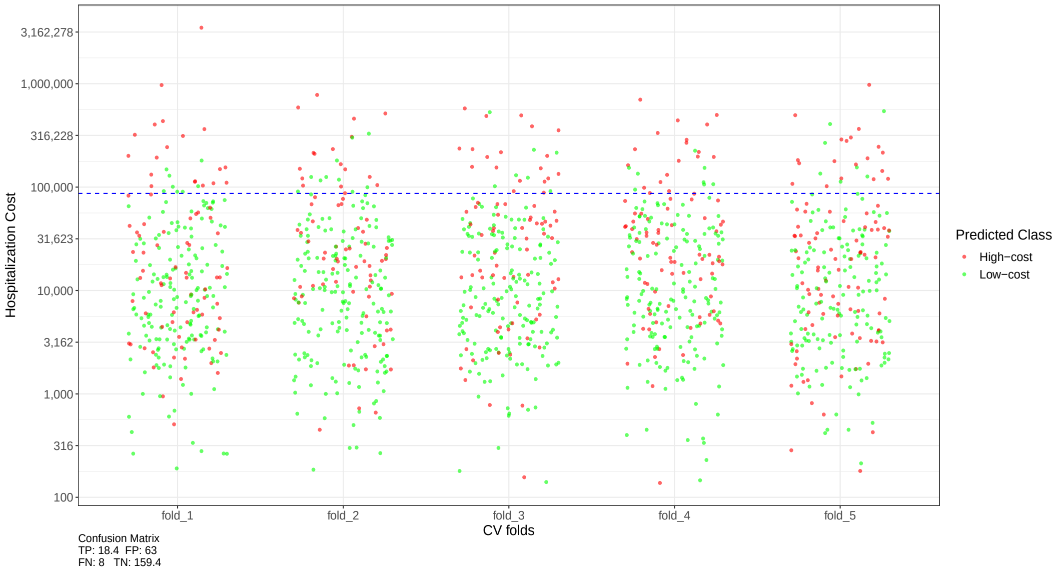


Figure 4.19: Classification predictions for the testing folds during the CV procedure of "Admission Subgroup" 33. The blue dashed line represent the high-cost boundary for this subgroup.

## 4.2.5.2 Analyzing Individual Cases

Table 4.10 summarizes the analysis for the FN with the highest cost value (which is referred to as "FN-high") and for the FP with the lowest cost value ("FP-low"). It shows the distribution (median and IQR for continuous variables and count and proportion % for binary) of the four most important variables for each model considering the group of observations correctly predicted as low-cost (TN) and as high-cost (TP). It is expected that the value of FN-high should correspond to the TN distribution, while the value of FP-low should be within the TP distribution.[22]

For subgroup 22, the analysis of the FN-high and FP-low cases is presented as follows. With regard to the "Admission_Type (Emergency)" variable, both FN-high and FP-low were "Scheduled" admissions (i.e., not "Emergency"). The observations that the model correctly classifies as high-cost (TP) are half scheduled and half emergency admissions, while all low-cost hospitalizations correctly classified (TN) were "Scheduled". For the "Age" variable, it is interesting to notice that FN-high was a 92 years-old patient, staying out of both correctly classified distributions (TP and TN).

With regard to the "Clinic_hospitalizations_cost_total" variable (which refers to the total cost of past clinical hospitalizations), FP-low indeed pre-

[22] This behavior does not necessarily occur with this analysis because the variables can be used on different occasions and several times in tree-based models.

Table 4.10: Predictors' characteristics for extreme misclassifications in both subgroups.

**Subgroup 22: "ENDOSCOPIA INTERVENCIONISTA"**

| **Fold** | **Variable** | **median [IQR] count [%]** |
|---|---|---|
| Fold_1 | Admission_Type (Emergency) | |
| | FN-high | Scheduled |
| | TP | 23 [50%] |
| | TN | 0 [0%] |
| Fold_3 | FP-low | Scheduled |
| | TP | 24 [53%] |
| | TN | 0 [0%] |
| Fold_1 | Surgery_hospitalization_cost_total | |
| | FN-high | 0 |
| | TP | 11,861 [0-148,996] |
| | TN | 0 [0-2,360] |
| Fold_3 | FP-low | 0 |
| | TP | 23,563 [0-62,813] |
| | TN | 0 [0-2,675] |
| Fold_1 | Clinic_hospitalization_cost_total | |
| | FN-high | 0 |
| | TP | 8,858 [0-65,534] |
| | TN | 0 [0-0] |
| Fold_3 | FP-low | 1,380 |
| | TP | 0 [0-29,580] |
| | TN | 0 [0-0] |
| Fold_1 | Age | |
| | FN-high | 92 |
| | TP | 42 [22-64] |
| | TN | 53 [44-59] |
| Fold_3 | FP-low | 35 |
| | TP | 53 [40-59] |
| | TN | 51 [41-59] |

**Subgroup 33: "ACESSOS VASCULARES"** [b]

| **Fold** | **Variable** | **median [IQR] count [%]** |
|---|---|---|
| Fold_5 | Flag_Genetic (Yes) | |
| | FN-high | No |
| | TP | 2 [10.5%] |
| | TN | 3 [2.1%] |
| Fold_4 | FP-low | No |
| | TP | 3 [16.7%] |
| | TN | 1 [0.6%] |
| Fold_5 | Surgery_hospitalization_cost_total | |
| | FN-high | 2063.36 |
| | TP | 146,506 [0-811,758] |
| | TN | 6,617 [0-27,727] |
| Fold_4 | FP-low | 203189 |
| | TP | 44,482 [2,126-288,360] |
| | TN | 5,676 [0-27,810] |
| Fold_5 | Exam_total | |
| | FN-high | 67 |
| | TP | 115 [46-254] |
| | TN | 43 [17-88] |
| Fold_4 | FP-low | 20 |
| | TP | 100 [30-226] |
| | TN | 40 [9-110] |
| Fold_5 | Age | |
| | FN-high | 65 |
| | TP | 48 [34-67] |
| | TN | 46 [35-57] |
| Fold_4 | FP-low | 38 |
| | TP | 47 [34-58] |
| | TN | 45 [30-56] |

sented some of the characteristics of the TP observations since its value of $1,380$ is in the TP range of $[0 - 29,580]$ (comprising 75 % of the high-cost observations), while 75 % of the TN observations had no cost of past clinical hospitalizations (value $= 0$). Therefore, it makes sense for the model to classify FP-low as a high-cost hospitalization, even though this is not correct.

For subgroup 33, the analysis of the FN-high and FP-low cases is more inconclusive, since TP and TN distributions are quite similar for most variables. One reason that justifies the model prediction relates to the total cost of past surgical hospitalization ("Surgery_hospitalization_cost_total" variable) of FP-low, which is valued $(203,189)$, thus fitting in the TP distribution (with 75 % ranging in $[2,126 - 288,360]$), rather than on the TN distribution (with 75 % ranging in $[0 - 27,727]$).

In summary, this analysis of individual cases provided some insights into the characteristics that differentiate the extreme misclassified observations. The FP-low cases seem to present characteristics of a high-cost hospitalization. A possible explanation is that it represents a residual hospitalization, since the method used to detect and eventually remove those hospitalizations was an approximation. This should be further investigated by the company domain experts and, for future model refinements, it is worth reviewing this aspect. In the case of FN-high, the issue seems to be not a confusion among classes, but rather behaviors that differed from both classes, therefore remaining as uncovered patterns. Further analysis should analyze those observations separately in order to understand their characteristics and even the high-cost definition.

#### 4.2.5.3 Analyzing from an Economic Perspective

From a business perspective, it is interesting to analyze the results of the classification model from an economic point of view. In theory, the patients selected as having a risk of incurring unexpected high-costs would require dedicated attention in order to reduce the forecasted high expenses and, of course, improve the health care service and thus patient's health quality. In this context, a crucial analysis would be to assess the amount of cost reduction this proactive plan would potentially achieved, thus providing an economic evaluation of the model's performance. However, at the current phase of the project, it was not possible to address both the expected hospitalization cost reduction and the expenses caused by the proactive attention. Despite this, a naive proposal for a possible evaluation from an economic perspective was developed, and its implementation proposed as future works.

The objective is to measure the model's mistakes, not in number of patients, but considering costs. In terms of the cells of the Confusion Matrix, and having prior knowledge of the expected cost reduction (Cr) and the healthcare plan costs (Cp) for each subgroup, the following procedure was designed. Let $A$ be the set of hospitalizations (i) for each "Admission Subgroup", then $\forall i \in A$:

$$\begin{array}{ll}
\text{If } X_i = TP, \text{ then} & C^i_{final} = C^i_{real} - C_r + C_p; \\
\text{If } X_i = TN, \text{ then} & C^i_{final} = C^i_{real}; \\
\text{If } X_i = FP, \text{ then} & C^i_{final} = C^i_{real} + C_p; \\
\text{If } X_i = FN, \text{ then} & C^i_{final} = C^i_{real} + (C^i_{real} - C_r + C_p);
\end{array}$$

where $X_i$ is the class predicted by the ML model for hospitalization $i$, and $C^i_{final}$ the expected cost after the model is used for predictions and the proactive plan successfully carry out. Then, the overall cost reduction for a specific "Admission Subgroup" equals $\sum_i C^i_{final}$ and is compared with the total cost without applying the proactive plan $\sum_i C^i_{real}$. Note that if the predicted class is a FN the referred $C^i_{final}$ is an analogy of the opportunity cost for not detecting a high-cost patient, represented by the term $(C^i_{real} - C_r + C_p)$.

In addition to the proposed analysis, it is possible to still provide an economic evaluation of the models' performance using the available information. Table 4.11 presents for both subgroups, the cost (average across the five folds) proportion for the correctly identified (TP-cost) high-cost hospitalization. In both cases, almost 80% of the total cost related to the high-cost hospitalizations were identified. In other words, from table 4.9 the best results for the subgroup 22 shows that the 72.4% (Sensitivity) of the high-cost admission were correctly predicted, which represents the 78.7% of the total cost of those hospitalizations. Then it is possible to argue that the ML model is predicting the majority of the positive class and the ones with the highest cost. The models' achievements and the conclusions reached from the economic perspective corroborates the accomplishment of the organization's expectations.

Table 4.11: Cost analysis of models' performance for "Admission Subgroups" numbers 22 and 33.

| **Admission Subgroup** | **TP total cost** | **TP + FN total cost** | **TP %** | **FP total cost** | **TP + FP total cost** | **FP %** |
|---|---|---|---|---|---|---|
| 22 | 2,756,476 | 3,504,420 | 78.7 | 533,043 | 3,289,519 | 16.2 |
| 36 | 5,269,644 | 6,622,624 | 79.6 | 1,352,491 | 6,622,624 | 20.4 |

The last economic analysis closes the evaluation phase of the developed models, with significant results meeting the proposed business goals. The advances made in the labeling of the "Hospitalizations Cost Class" provide an improvement to the ML technique performance, and also allowed a better understanding of the problem and context of the study. It showed that the inherent characteristics of each hospitalization's group related to medical concerns are crucial when modeling cost and medical resource consumption variables. Also, the priority analysis demonstrated that there are subgroups in which cost values are naturally a few times bigger than others, but a small cost variability does not make them suitable for building a predictive model.

With the related scenario and the clear improvements obtained, the organization got involved in a project to create one more variable describing a level even more specific to group hospitalizations. This variable is specifically related to the medical procedure performed within the hospitalization, and it turns out that for the same "Admission Subgroup", there may be different procedures. As the construction of the variable is an ongoing project, this feature was not used in this thesis project. However, the organization provides a sample of this variable ("Admission Procedure" from now on) for one subgroup, in order to investigate whether the cost variability within it is significant, as was corroborated in the case of "Admission Group", see Figure 4.8. The following plot illustrates this analysis for the "Admission Subgroup" named "FARINGE" (pharynx).

The Graph illustrates a similar behavior when analyzing the subgroups in Section 4.2.2, showing a relevant cost variability among procedures of the same subgroup. Actually, with the high-cost definition at the subgroup level, for the second and the second to last (left to right) procedure, there are

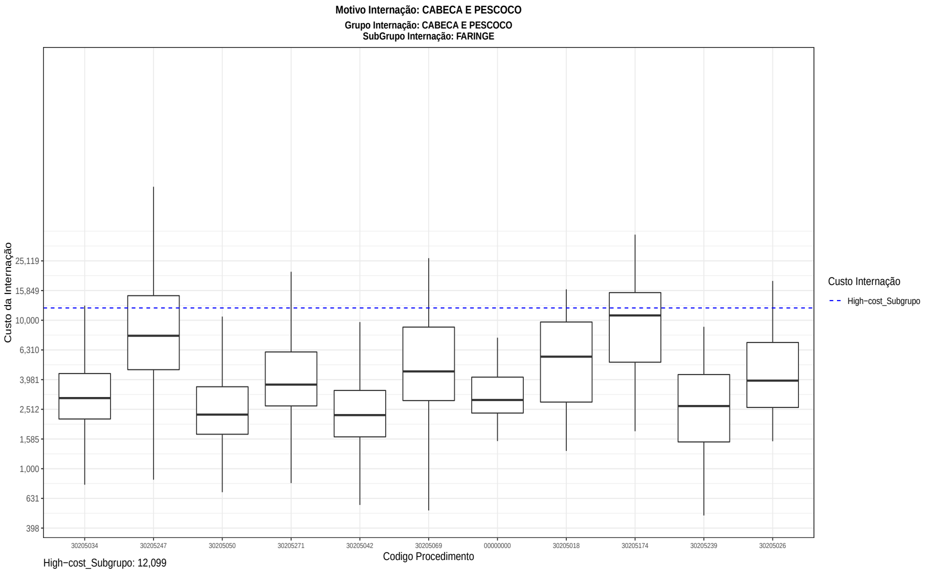


Figure 4.20: Cost distribution for each "Admission Procedure" within the *FARINGE* subgroup (only procedures with more than 10 hospitalizations are shown). The blue dashed line represents the high-cost boundary defined at subgroup level.

hospitalizations under the 75 percentile that were labeled as high-cost, which contradicts the actual definitions for detecting unexpected admissions. With this behavior, it is evident that further analyses are necessary even to consider a new labeling strategy or a data splitting criterion for predictive model building. As in the First cycle, these decisions imply the beginning of a new cycle, but this time would be part of the suggested future works, which also give continuity to the ongoing project with the organization.

# 5
# Conclusions

This thesis proposed a solution to identify high-cost hospitalizations, which may integrate the proactive plan of a Health Consulting organization. The proposed approach was developed as a Data Science project and model cost variables as indicators of the patients' hospitalization risk, their severity, and their medical resources consumption. In this context, this study built predictive models using statistical and ML techniques to detect high-cost hospitalizations learned from a historical database of patient's admissions in the past 3 years, comprising also their records of medical resource consumption. Methodologically, the development of this research was guided by the integration of a Data Science Life cycle with the DSR methodology. Two cycles were conducted for the design and development of the predictive models.

During the First design cycle, important insights into the characteristics of the problem and the data were found. The response variable "Hospitalization Cost Class" was defined to label procedures with unexpectedly high expenses, considering the characteristics of the procedures performed during the in-hospital admission. Therefore, this definition should be done at a grouping level, in which the medical characteristics of all the procedures are similar. This high-cost definition prevented bias in selecting naturally expensive procedures as such. During the Second design cycle, the definition of a high-cost hospitalization was improved, which led to enhanced predictive models. The final definition considered the "Admission Subgroup" variable as the finest granular level available in the dataset, and an outlier detection procedure was applied to label as high-cost hospitalizations all the unexpectedly high valued observations within each subgroup distribution.

On the other hand, predictors describing the medical resource consumption were aggregated summarizing their temporal behavior. The "aggregated" set of features reached similar results when compared to the "temporal" set, while also leading to less complex predictive models with improved interpretability. In addition, variables related to previous hospitalizations cost provided additional valuable information to predict future high-cost hospitalizations.

Regarding the performance of the predictive models, better results were achieved when splitting the population into admission subgroups, ranked by their frequency and average cost. In summary, the Random Forest (RF) and the Extreme Gradient Boosting (XGB) techniques produced models with higher performance, reaching an AUCPR of 0.732 and 0.644 when assessed over a dataset with a 0.113 and 0.106 proportion of the high-cost class, respectively. This performance was achieved for "Admission Subgroups" "ENDOSCOPIA INTERVENCIONISTA" (Interventional Endoscopy) and "ACESSOS VASCULARES" (Vascular Accesses), respectively. The predictive models using RF detected 72,4% of the high-cost hospitalizations with a 33% of Precision, meeting the business expectations since the majority of the unexpected expenses were identified, even paying an acceptable cost for misclassifications. Moreover, in terms of the cost value, the correctly predicted high-cost hospitalizations represent 78.7% of all the real ones. In contrast, hospitalizations misclassified as high-cost represent 16.2% of the total cost of all those that were classified as high-cost.

Further analysis of the misclassifications made by the models evidenced the existence of False Positive observations in the lower extreme of the cost distribution, which may indicate residual hospitalizations. Also, identified False Negative observations in the upper extreme correspond to hospitalizations with characteristics that are very different from the existing patterns present in the learning dataset.

## 5.1 Suggestions for future works

This thesis was developed as part of a project with a Health Consulting company, and accomplished the First phase of the planned research. Hence, in the following, future works related to further improvement of the whole process are provided, which implementation requires that a new DSR cycle be initiated.

First, as concluded at the end of the Second cycle and with the new variable of "Admission Procedures" being collected by the organization, it would be possible and necessary to assess whether the definition of a high-cost hospitalization needs refinement. Furthermore, as residual hospitalizations cause possible misclassifications in the lower extreme of the cost distribution, the definition of those could be improved, searching for a minimum expected cost value for each type of procedure, and then using it as a threshold for filtering the lower observations.

Regarding the modeling phase, further improvement can be made. The resampling techniques used to overcome the class imbalance issue did not improve the performance of ML techniques, which highlights the ability of the predictive models in this scenario but also indicates that these techniques may not be the most appropriate in this context. A racing method should be conducted to select the most appropriate strategy for the given unbalanced task, as proposed by DalPozzolo et al. [99]. Moreover, the ML techniques employed can be further improved, expanding the search space of the hyperparameter tunning process. This may become time-consuming, requiring great computational resources. Hence, instead of using a grid or random search strategy, Kuhn [100] proposed a futility analysis during the CV procedure to reduce the training time by adaptively resampling candidate values and the clearly sub-optimal ones are discarded.

The economic analysis conducted to assess the predictive model performance in this context did not account for the missed high-cost hospitalizations (i.e., FN), which are of interest to the organization. The method proposed in section 4.2.5.3 could be implemented to address this issue. This procedure uses two values that were not available at the time of this project, the expected hospitalization cost reduction, and expenses caused by the proactive attention. However, its purpose is to assess the amount of cost reduced by the pretended proactive plan and also has an economic evaluation of the model's performance accounting for all hits and misclassifications.

# Bibliography

[1] HEVNER, A. R.. **A Three Cycle View of Design Science Research A Three Cycle View of Design Science Research**. Scandinavian Journal of Information Systems, 19(2):87–92, 2007.

[2] SHEARER, C.. **The CRISP-DM model: The New Blueprint for Data Mining**. Journal of Data Warehousing, 5(4):13–22, 2000.

[3] BATARSEH, F. A.; LATIF, E. A.. **Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare**. Big Data Research, 4(October):13–24, 2016.

[4] CARVALHO, J. V.; ROCHA, Á.; VASCONCELOS, J. ; ABREU, A.. **A health data analytics maturity model for hospitals information systems**. International Journal of Information Management, 46:278–285, 2019.

[5] FOSSO WAMBA, S.; AKTER, S.; EDWARDS, A.; CHOPIN, G. ; GNANZOU, D.. **How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study**. International Journal of Production Economics, 165:234–246, 2015.

[6] MANS, R. S.; VAN DER AALST, W. M. P. ; VANWERSCH, R. J. B.. **Process Mining in Healthcare**. SpringerBriefs in Business Process Management. Springer International Publishing, Cham, 2015.

[7] GALETSI, P.; KATSALIAKI, K. ; KUMAR, S.. **Big data analytics in health sector: Theoretical framework, techniques and prospects**. International Journal of Information Management, 50:206–216, feb 2020.

[8] PANICACCI, S.; DONATI, M.; FANUCCI, L.; BELLIN, I.; PROFILI, F. ; FRANCESCONI, P.. **Population Health Management Exploiting Machine Learning Algorithms to Identify High-Risk Patients**. In: 2018 IEEE 31ST INTERNATIONAL SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS (CBMS), p. 298–303. IEEE, 2018.

[9] KIM, Y. J.; PARK, H.. **Improving Prediction of High-Cost Health Care Users with Medical Check-Up Data**. Big Data, 7(3):163–175, sep 2019.

[10] DUNCAN, I.; LOGINOV, M. ; LUDKOVSKI, M.. **Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs**. North American Actuarial Journal, 20(1):65–87, jan 2016.

[11] BERTSIMAS, D.; BJARNADÓTTIR, M. V.; KANE, M. A.; KRYDER, J. C.; PANDEY, R.; VEMPALA, S. ; WANG, G.. **Algorithmic Prediction of Health-Care Costs**. Operations Research, 56(6):1382–1392, dec 2008.

[12] CARROLL, A. E.. **The High Costs of Unnecessary Care**. JAMA, 318(18):1748–1749, nov 2017.

[13] LYU, H.; XU, T.; BROTMAN, D.; MAYER-BLACKWELL, B.; COOPER, M.; DANIEL, M.; WICK, E. C.; SAINI, V.; BROWNLEE, S. ; MAKARY, M. A.. **Overtreatment in the United States**. PLOS ONE, 12(9), sep 2017.

[14] GUO, X.; GANDY, W.; COBERLEY, C.; POPE, J.; RULA, E. ; WELLS, A.. **Predicting Health Care Cost Transitions Using a Multidimensional Adaptive Prediction Process**. Population Health Management, 18(4):290–299, aug 2015.

[15] LAHIRI, B.; AGARWAL, N.. **Predicting Healthcare Expenditure Increase for an Individual from Medicare Data**. p. 8, 2014.

[16] MORID, M. A.; KAWAMOTO, K.; AULT, T.; DORIUS, J. ; ABDELRAHMAN, S.. **Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation**. AMIAnAnnual Symposium proceedings, 2017:1312–1321, 2017.

[17] WAMMES, J. J. G.; VAN DER WEES, P. J.; TANKE, M. A. C.; WESTERT, G. P. ; JEURISSEN, P. P. T.. **Systematic review of high-cost patients' characteristics and healthcare utilisation**. BMJ Open, 8(9):e023113, sep 2018.

[18] YANG, C.; DELCHER, C.; SHENKMAN, E. ; RANKA, S.. **Machine learning approaches for predicting high cost high need patient expenditures in health care**. BioMedical Engineering OnLine, 17(S1):131, nov 2018.

[19] SUSHMITA, S.; NEWMAN, S.; MARQUARDT, J.; RAM, P.; PRASAD, V.; COCK, M. D. ; TEREDESAI, A.. **Population Cost Prediction on Public Healthcare Datasets**. In: PROCEEDINGS OF THE 5TH INTERNATIONAL CONFERENCE ON DIGITAL HEALTH 2015 - DH '15, p. 87–94, New York, New York, USA, 2015. ACM Press.

[20] KUO, R. N.; DONG, Y. H.; LIU, J. P.; CHANG, C. H.; SHAU, W. Y. ; LAI, M. S.. **Predicting healthcare utilization using a pharmacy-based metric with the WHO's anatomic therapeutic chemical algorithm**. Medical Care, 49(11):1031–1039, 2011.

[21] DUNCAN, I.; LOGINOV, M. ; LUDKOVSKI, M.. **Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs**. North American Actuarial Journal, 20(1):65–87, jan 2016.

[22] FREES, E. W.; JIN, X. ; LIN, X.. **Actuarial Applications of Multivariate Two-Part Regression Models**. Annals of Actuarial Science, 7(2):258–287, 2013.

[23] MORID, M. A.; SHENG, O. R. L.; KAWAMOTO, K.; AULT, T.; DORIUS, J. ; ABDELRAHMAN, S.. **Healthcare cost prediction: Leveraging fine-grain temporal patterns**. In: JOURNAL OF BIOMEDICAL INFORMATICS, volumen 91, p. 103113. mar 2019.

[24] AMINIKHANGHAHI, S.; COOK, D. J.. **A survey of methods for time series change point detection**. Knowledge and Information Systems, 51(2):339–367, may 2017.

[25] TAMANG, S.; MILSTEIN, A.; SØRENSEN, H. T.; PEDERSEN, L.; MACKEY, L.; BETTERTON, J. R.; JANSON, L. ; SHAH, N.. **Predicting patient 'cost blooms' in Denmark: A longitudinal population-based study**. BMJ Open, 7(1), 2017.

[26] CHANG, H. Y.; BOYD, C. M.; LEFF, B.; LEMKE, K. W.; BODYCOMBE, D. P. ; WEINER, J. P.. **Identifying Consistent High-cost Users in a Health Plan: Comparison of Alternative Prediction Models**. Medical Care, 54(9):852–859, sep 2016.

[27] ROBST, J.. **Developing Models to Predict Persistent High-Cost Cases in Florida Medicaid**. Population Health Management, 18(6):467–476, dec 2015.

[28] BOSCARDIN, C. K.; GONZALES, R.; BRADLEY, K. L. ; RAVEN, M. C.. **Predicting cost of care using self-reported health status data.** BMC Health Services Research, 15(1):406, sep 2015.

[29] IZAD SHENAS, S. A.; RAAHEMI, B.; HOSSEIN TEKIEH, M. ; KUZIEMSKY, C.. **Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes.** Computers in Biology and Medicine, 53:9–18, oct 2014.

[30] FLEISHMAN, J. A.; COHEN, J. W.. **Using information on clinical conditions to predict high-cost patients.** Health Services Research, 45(2):532–552, apr 2010.

[31] COHEN, S. B.; EZZATI-RICE, T. ; YU, W.. **The utility of extended longitudinal profiles in predicting future health care expenditures.** Medical Care, 44(5 SUPPL.):I45–53, may 2006.

[32] CRAWFORD, A. G.; FUHR, J. P.; CLARKE, J. ; HUBBS, B.. **Comparative effectiveness of total population versus disease-specific neural network models in predicting medical costs.** Disease Management, 8(5):277–287, oct 2005.

[33] DRESCH, A.; LACERDA, D. P. ; ANTUNES, J. A. V.. **Design science research: A method for science and technology advancement.** Springer International Publishing, jan 2015.

[34] VAN AKEN, J. E.. **Management Research Based on the Paradigm of the Design Sciences: The Quest for Field-Tested and Grounded Technological Rules.** Journal of Management Studies, 41(2):219–246, feb 2004.

[35] DENYER, D.; TRANFIELD, D. ; VAN AKEN, J. E.. **Developing design propositions through research synthesis.** Organization Studies, 29(3):393–413, mar 2008.

[36] CHALMERS, A. F. A. F.. **What is this thing called science?** Hackett Pub, 1999.

[37] HEVNER, A. R.; MARCH, S. T.; PARK, J. ; RAM, S.. **Design science in information systems research.** MIS Quarterly: Management Information Systems, 28(1):75–105, 2004.

[38] FAYYAD, U.; PIATETSKY-SHAPIRO, G. ; SMYTH, P.. **From data mining to knowledge discovery in databases.** AI Magazine, 17(3):37–53, 1996.

[39] SAS INSTITUTE INC. **SAS Help Center: Introduction to SEMMA**, 2019.

[40] NISBET, R.; MINER, G. ; YALE, K.. **Handbook of statistical analysis and data mining applications**. Elsevier Inc., nov 2018.

[41] SHCHERBAKOV, M.; SHCHERBAKOVA, N.; BREBELS, A.; JANOVSKY, T. ; KAMAEV, V.. **Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development**. In: COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE, volumen 466 CCIS, p. 708–716. Springer Verlag, 2014.

[42] SPEARMAN, C.. **The Proof and Measurement of Association between Two Things**. The American Journal of Psychology, 15(1):72, jan 1904.

[43] MUKAKA, M. M.. **Statistics corner: A guide to appropriate use of correlation coefficient in medical research**. Malawi Medical Journal, 24(3):69–71, 2012.

[44] RICE, J. A.. **Mathematical Statistics and Data Analysis, Third Edition**. Belmont, CA: Duxbury Press, third edition, 2007.

[45] CRAMÉR, H.. **Mathematical methods of statitics**. Princeton University Press, 1946.

[46] ALLEN, M.. **Cramér's V**. In: THE SAGE ENCYCLOPEDIA OF COMMUNICATION RESEARCH METHODS. SAGE Publications, Inc, jul 2017.

[47] HAWKINS, D. M.; BASAK, S. C. ; MILLS, D.. **Assessing model fit by cross-validation**. In: JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, volumen 43, p. 579–586, 2003.

[48] MOLINARO, A. M.; SIMON, R. ; PFEIFFER, R. M.. **Prediction error estimation: A comparison of resampling methods**. Bioinformatics, 21(15):3301–3307, aug 2005.

[49] KIM, J. H.. **Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap**. Computational Statistics and Data Analysis, 53(11):3735–3745, sep 2009.

[50] XU, Q. S.; LIANG, Y. Z.. **Monte Carlo cross validation**. Chemometrics and Intelligent Laboratory Systems, 56(1):1–11, apr 2001.

[51] EFRON, B.. **Estimating the error rate of a prediction rule: Improvement on cross-validation**. Journal of the American Statistical Association, 78(382):316–331, 1983.

[52] KUHN, M.; JOHNSON, K.. **Applied Predictive Modeling**. Springer New York, New York, NY, 2013.

[53] JAMES, G.; WITTEN, D.; HASTIE, T. ; TIBSHIRANI, R.. **An Introduction to Statistical Learning: With Applications in R**. Springer Publishing Company, Incorporated, 2014.

[54] TING, K. M.. **An instance-weighting method to induce cost-sensitive trees**. IEEE Transactions on Knowledge and Data Engineering, 14(3):659–665, may 2002.

[55] WEISS, G. M.; PROVOST, F.. **The Effect of Class Distribution on Classifier Learning**. Training, 2001.

[56] MCCARTHY, K.; ZABAR, B. ; WEISS, G.. **Does cost-sensitive learning beat sampling for classifying rare classes?** In: PROCEEDINGS OF THE 1ST INTERNATIONAL WORKSHOP ON UTILITY-BASED DATA MINING, UBDM '05, p. 69–77, 2005.

[57] CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O. ; KEGELMEYER, W. P.. **SMOTE: Synthetic minority over-sampling technique**. Journal of Artificial Intelligence Research, 16:321–357, jan 2002.

[58] MENARDI, G.; TORELLI, N.. **Training and assessing classification rules with imbalanced data**. Data Mining and Knowledge Discovery, 28(1):92–122, 2014.

[59] TANTITHAMTHAVORN, C.; HASSAN, A. E. ; MATSUMOTO, K.. **The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models**. IEEE Transactions on Software Engineering, 2018.

[60] BUREZ, J.; VAN DEN POEL, D.. **Handling class imbalance in customer churn prediction**. Expert Systems with Applications, 36(3 PART 1):4626–4636, 2009.

[61] JEATRAKUL, P.; WONG, K. W. ; FUNG, C. C.. **Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm**. In: LECTURE NOTES IN COMPUTER SCIENCE, volumen 6444 LNCS, p. 152–159. 2010.

[62] HASTIE, T.; TIBSHIRANI, R. ; FRIEDMAN, J.. **The Elements of Statistical Learning**, volumen 27 de **Springer Series in Statistics**. Springer New York, New York, NY, 2009.

[63] DOBSON, A. J.; BARNETT, A. G.. **An Introduction to Generalized Linear Models**. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America, fourth edition, 2018.

[64] ROSSI, R. J.. **Mathematical Statistics : An Introduction to Likelihood Based Inference**. 2018.

[65] ZOU, H.; HASTIE, T.. **Regularization and variable selection via the elastic net**. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 67(2):301–320, 2005.

[66] CESSIE, S. L.; HOUWELINGEN, J. C. V.. **Ridge Estimators in Logistic Regression**. Applied Statistics, 41(1):191, 1992.

[67] TIBSHIRANI, R.. **Regression Shrinkage and Selection Via the Lasso**. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[68] BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. ; STONE, C.. **Classification and Regression Trees**. Chapman and Hall/CRC, New York, NY, first edit edition, 1984.

[69] STROBL, C.; MALLEY, J. ; TUTZ, G.. **An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests**. Psychological Methods, 14(4):323–348, dec 2009.

[70] BREIMAN, L.. **Bagging Predictors**. Technical report, 1996.

[71] BREIMAN, L.. **Random forests**. Machine Learning, 45(1):5–32, 2001.

[72] MARTÍNEZ MUÑOZ, G.; SUÁREZ, A.. **Out-of-bag estimation of the optimal sample size in bagging**. Pattern Recognition, 43(1):143–152, jan 2010.

[73] JANITZA, S.; BINDER, H. ; BOULESTEIX, A. L.. **Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications**. Biometrical Journal, 58(3):447–473, may 2016.

[74] PROBST, P.; BOULESTEIX, A. L.. **To tune or not to tune the number of trees in random forest**. Journal of Machine Learning Research, 18:1–8, 2018.

[75] HOTHORN, T.; HORNIK, K. ; ZEILEIS, A.. **Unbiased recursive partitioning: A conditional inference framework**. Journal of Computational and Graphical Statistics, 15(3):651–674, 2006.

[76] GEURTS, P.; ERNST, D. ; WEHENKEL, L.. **Extremely randomized trees**. Machine Learning, 63(1):3–42, apr 2006.

[77] PROBST, P.; WRIGHT, M. N. ; BOULESTEIX, A. L.. **Hyperparameters and tuning strategies for random forest**, may 2019.

[78] WITTEK, P.. **Boosting**. In: QUANTUM MACHINE LEARNING, p. 89–95. Elsevier, 2014.

[79] FREUND, Y.; SCHAPIRE, R. E.. **A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting**. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[80] FRIEDMAN, J. H.. **Greedy function approximation: A gradient boosting machine**. Annals of Statistics, 29(5):1189–1232, oct 2001.

[81] CHEN, T.; GUESTRIN, C.. **XGBoost: A scalable tree boosting system**. In: PROCEEDINGS OF THE ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, volumen 13-17-Augu, p. 785–794. Association for Computing Machinery, aug 2016.

[82] MCNEMAR, Q.. **Note on the sampling error of the difference between correlated proportions or percentages**. Psychometrika, 12(2):153–157, jun 1947.

[83] DIETTERICH, T. G.. **Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms**. Neural Computation, 10(7):1895–1923, oct 1998.

[84] FERRI, C.; HERNÁNDEZ-ORALLO, J. ; MODROIU, R.. **An experimental comparison of performance measures for classification**. Pattern Recognition Letters, 30(1):27–38, jan 2009.

[85] KUBAT, M.; MATWIN, S.. **Addressing the Curse of Imbalanced Training Sets: One Sided Selection**. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, volumen 97, p. 179–186, 1997.

[86] BAEZA YATES, R. A.; RIBEIRO NETO, B.. **Modern Information Retrieval**. ACM Press, New York, 1999.

[87] CHICCO, D.; JURMAN, G.. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation**. BMC genomics, 21(1):6, jan 2020.

[88] FAWCETT, T.. **An introduction to ROC analysis**. Pattern Recognition Letters, 27(8):861–874, 2006.

[89] DAVIS, J.; GOADRICH, M.. **The relationship between precision-recall and ROC curves**. In: ACM INTERNATIONAL CONFERENCE PROCEEDING SERIES, volumen 148, p. 233–240, 2006.

[90] FERNÁNDEZ, A.; GARCÍA, S.; GALAR, M.; PRATI, R. C.; KRAWCZYK, B. ; HERRERA, F.. **Learning from Imbalanced Data Sets**. Springer International Publishing, 2018.

[91] HANCZAR, B.; HUA, J.; SIMA, C.; WEINSTEIN, J.; BITTNER, M. ; DOUGHERTY, E. R.. **Small-sample precision of ROC-related estimates**. Bioinformatics, 26(6):822–830, mar 2010.

[92] SAITO, T.; REHMSMEIER, M.. **The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets**. PLoS ONE, 10(3), mar 2015.

[93] MURPHY, K. P.. **Machine Learning A Probabilistic Perspective**. MIT Press Journals, London, 2012.

[94] BRIER, G. W.. **Verification Forecast Expressed in Terms of Probability**. Monthly Weather Review, 78(1):1–3, jan 1950.

[95] MULLARKEY, M. T.; HEVNER, A. R.; GRANDON GILL, T. ; DUTTA, K.. **Citizen Data Scientist: A Design Science Research Method for the Conduct of Data Science Projects**. In: LECTURE NOTES IN COMPUTER SCIENCE, volumen 11491 LNCS, p. 191–205. Springer Verlag, 2019.

[96] BRAHMA, A.; CHATTERJEE, S. ; LI, Y.. **Designing a Machine Learning Model to Predict Cardiovascular Disease Without Any Blood Test**. In: LECTURE NOTES IN COMPUTER SCIENCE, volumen 11491 LNCS, p. 125–139. Springer Verlag, 2019.

[97] NEWMAN, M. E.. **Power laws, Pareto distributions and Zipf's law**. Contemporary Physics, 46(5):323–351, sep 2005.

[98] NEELON, B.; O'MALLEY, A. J.. **Two-Part Models for Zero-Modified Count and Semicontinuous Data**. In: HEALTH SERVICES EVALUATION, chapter 28, p. 695–716. Springer, New York, NY, 2019.

[99] DAL POZZOLO, A.; CAELEN, O.; WATERSCHOOT, S. ; BONTEMPI, G.. **Racing for Unbalanced Methods Selection**. In: LECTURE NOTES IN COMPUTER SCIENCE, volumen 8206 LNCS, p. 24–31. 2013.

[100] KUHN, M.. **Futility Analysis in the Cross-Validation of Machine Learning Models**. may 2014.

# A
# Dictionary of variables

| Category | Name | Data type | Description |
|---|---|---|---|
| Patient Characteristics | Gender | binary (logical) | Number identifying the gender of a patient [if 0: Female, if 1: Male] |
| Patient Characteristics | Age | number (interger) | Age in the hospitalization date |
| Patient Hospitalization | Admission Group | character (varchar) | Hospitalization Group according to CBHPM * |
| Patient Hospitalization | Admission Subgroup | character (varchar) | Hospitalization Subgroup according to CBHPM * |
| Patient Hospitalization | Admission Type | binary (logical) | The way a patient was admitted [if 0: Scheduled, if 1: Emergency] |
| Patient Hospitalization | Flag Surgery | binary (logical) | Whether a surgery was performed during the hospitalization [if 0: Clinical, if 1: Surgery] |
| Patient Hospitalization | Hospitalization cost | number (float) | Hospitalization total cost |
| Patient Hospitalization | Hospitalization cost class | binary (logical) | High-cost definition [if 0: low-cost, if 1:high-cost] |
| Medical Resource Consumption | Flag Genetic | binary (logical) | Existence of Cytogenetic or Molecular Genetics examas (CBHPM) 3 years before the hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Exam semester6 | number (interger) | Number of exam claims (common and special) performed from the 30th to the 36th month before hospitalization |
| Medical Resource Consumption | Exam semester5 | number (interger) | Number of exam claims (common and special) performed from the 24th to the 30th month before hospitalization |
| Medical Resource Consumption | Exam semester4 | number (interger) | Number of exam claims (common and special) performed from the 18th to the 24th month before hospitalization |
| Medical Resource Consumption | Exam semester3 | number (interger) | Number of exam claims (common and special) performed from the 12th to the 18th month before hospitalization |
| Medical Resource Consumption | Exam semester2 | number (interger) | Number of exam claims (common and special) performed from the 6th to the 12th month before hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Exam semester1 | number (interger) | Number of exam claims (common and special) performed from the 1st to the 6th month before hospitalization |
| Medical Resource Consumption | Exam total | number (interger) | Total number of exam claims (common and special) 3 years before the hospitalization |
| Medical Resource Consumption | Image month3 | number (interger) | Number of imaging tests performed in the 3rd month before hospitalization |
| Medical Resource Consumption | Image month2 | number (interger) | Number of imaging tests performed in the 2nd month before hospitalization |
| Medical Resource Consumption | Image month1 | number (interger) | Number of imaging tests performed in the month of hospitalization |
| Medical Resource Consumption | Image total | number (interger) | Total number of imaging exams performed 3 months before the intention |
| Medical Resource Consumption | Flag psycho year3 | binary (logical) | Existence of claims from psychological consultations carried out in the 3rd year before hospitalization |
| Medical Resource Consumption | Flag psycho year2 | binary (logical) | Existence of claims from psychological consultations carried out in the 2nd year before hospitalization |
| Medical Resource Consumption | Flag psycho year1 | binary (logical) | Existence of claims from psychological consultations carried out in the 1st year before hospitalization |
| Medical Resource Consumption | Flag psycho | binary (logical) | Existence of claims for psychological consultations carried out in the 3rd year before hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Consult semester6 | number (interger) | Number of claims for elective consultations carried out from the 30th to the 36th month before hospitalization |
| Medical Resource Consumption | Consult semester5 | number (interger) | Number of claims for elective consultations carried out from the 24th to the 30th month before hospitalization |
| Medical Resource Consumption | Consult semester4 | number (interger) | Number of claims for elective consultations carried out from the 18th to the 24th month before hospitalization |
| Medical Resource Consumption | Consult semester3 | number (interger) | Number of claims for elective consultations carried out from the 12th to the 18th month before hospitalization |
| Medical Resource Consumption | Consult semester2 | number (interger) | Number of claims for elective consultations carried out from the 6th to the 12th month before hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Consult semester1 | number (interger) | Number of claims for elective consultations carried out from the 1st to the 6th month before hospitalization |
| Medical Resource Consumption | Consult total | number (interger) | Total number of elective consultation claims made 3 years before the hospitalization |
| Medical Resource Consumption | Emergency year3 | number (interger) | Number of the hospitalizations to the Emergency Room in the 3rd year before the hospitalization |
| Medical Resource Consumption | Emergency year2 | number (interger) | Number of the hospitalizations to the Emergency Room in the 2nd year before the hospitalization |
| Medical Resource Consumption | Emergency year1 | number (interger) | Number of the hospitalizations to the Emergency Room in the 1st year before the hospitalization |
| Medical Resource Consumption | Emergecy total | number (interger) | Total number of Emergency Room entries 3 years before the hospitalization |
| Medical Resource Consumption | Hemodialysis year3 | number (interger) | Number of hemodialysis performed in the 3rd year before hospitalization |
| Medical Resource Consumption | Hemodialysis year2 | number (interger) | Number of hemodialysis performed in the 2nd year before hospitalization |
| Medical Resource Consumption | Hemodialysis year1 | number (interger) | Number of hemodialysis performed in the 1st year before hospitalization |
| Medical Resource Consumption | Hemodialysis total | number (interger) | Total number of hemodialysis performed 3 years before the hospitalization |
| Medical Resource Consumption | Pulsotherapies year3 | number (interger) | Number of pulsetherapies performed in the 3rd year before hospitalization |
| Medical Resource Consumption | Pulsotherapies year2 | number (interger) | Number of pulsetherapies performed in the 2nd year before hospitalization |
| Medical Resource Consumption | Pulsotherapies year1 | number (interger) | Number of pulsetherapies performed in the 1st year before hospitalization |
| Medical Resource Consumption | Pulsotherapies total | number (interger) | Total number of pulsetherapies performed 3 years before hospitalization |
| Medical Resource Consumption | Transplant total | number (interger) | Total number of transplants performed before hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Cardiovascular year3 | number (interger) | Cardiovascular Sentinel Counter performed in the 3rd year before the hospitalization |
| Medical Resource Consumption | Cardiovascular year2 | number (interger) | Cardiovascular Sentinel Counter performed in the 2nd year before the hospitalization |
| Medical Resource Consumption | Cardiovascular year1 | number (interger) | Cardiovascular Sentinel Counter performed in the 1st year before the hospitalization |
| Medical Resource Consumption | Cardiovascular flag | binary (logical) | Existence of Cardiovascular Sentinel performed up to 3rd year before hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Diabetes year3 | number (interger) | Diabetes Sentinel Counter performed in the 3rd year before the hospitalization |
| Medical Resource Consumption | Diabetes year2 | number (interger) | Diabetes Sentinel Counter performed in the 2nd year before the hospitalization |
| Medical Resource Consumption | Diabetes year1 | number (interger) | Diabetes Sentinel Counter performed in the 1st year before the hospitalization |
| Medical Resource Consumption | Diabetes flag | binary (logical) | Existence of Diabetes Sentinel performed up to 3rd year before hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Musculoskeletal year3 | number (interger) | Musculoskeletal Sentinel Counter performed in the 3rd year before the hospitalization |
| Medical Resource Consumption | Musculoskeletal year2 | number (interger) | Musculoskeletal Sentinel Counter performed in the 2nd year before the hospitalization |
| Medical Resource Consumption | Musculoskeletal year1 | number (interger) | Musculoskeletal Sentinel Counter performed in the 1st year before the hospitalization |
| Medical Resource Consumption | Musculoskeletal flag | binary (logical) | Existence of Musculoskeletal Sentinel performed up to 3rd year before hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Neoplasm year3 | number (interger) | Neoplasm Sentinel Counter performed in the 3rd year before the hospitalization |
| Medical Resource Consumption | Neoplasm year2 | number (interger) | Neoplasm Sentinel Counter performed in the 2nd year before the hospitalization |
| Medical Resource Consumption | Neoplasm year1 | number (interger) | Neoplasm Sentinel Counter performed in the 1st year before the hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Neoplasm flag | binary (logical) | Existence of Neoplasm Sentinel performed up to 3rd year before hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Obstructive year3 | number (interger) | Obstructive Sentinel Counter performed in the 3rd year before the hospitalization |
| Medical Resource Consumption | Obstructive year2 | number (interger) | Obstructive Sentinel Counter performed in the 2nd year before the hospitalization |
| Medical Resource Consumption | Obstructive year1 | number (interger) | Obstructive Sentinel Counter performed in the 1st year before the hospitalization |
| Medical Resource Consumption | Obstructive flag | binary (logical) | Existence of Obstructive Sentinel performed up to 3rd year before hospitalization [if 0: No, if 1: Yes] |
| Medical Resource Consumption | Physiotherapy year3 | number (interger) | Number of Physiotherapy sessions performed in the 3rd year before hospitalization |
| Medical Resource Consumption | Physiotherapy year2 | number (interger) | Number of Physiotherapy sessions performed in the 2nd year before hospitalization |
| Medical Resource Consumption | Physiotherapy year1 | number (interger) | Number of Physiotherapy sessions performed in the 1st year before hospitalization |
| Medical Resource Consumption | Physiotherapy total | number (interger) | Total number of Physiotherapy sessions performed in the 3 years before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization semester6 | number (interger) | Number of surgical the hospitalizations performed from the 30th to the 36th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization semester5 | number (interger) | Number of surgical the hospitalizations performed from the 24th to the 30th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization semester4 | number (interger) | Number of surgical the hospitalizations performed from the 18th to the 24th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization semester3 | number (interger) | Number of surgical the hospitalizations performed from the 12th to the 18th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization semester2 | number (interger) | Number of surgical the hospitalizations performed from the 6th to the 12th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization semester1 | number (interger) | Number of surgical the hospitalizations performed from the 1st to the 6th month before the hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Surgery hospitalization total | number (interger) | Sum of the number of surgical the hospitalizations in the last 3 years before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost semester6 | number (interger) | Cost of surgical the hospitalizations performed from the 30th to the 36th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost semester5 | number (interger) | Cost of surgical the hospitalizations performed from the 24th to the 30th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost semester4 | number (interger) | Cost of surgical the hospitalizations performed from the 18th to the 24th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost semester3 | number (interger) | Cost of surgical the hospitalizations performed from the 12th to the 18th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost semester2 | number (interger) | Cost of surgical the hospitalizations performed from the 6th to the 12th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost semester1 | number (interger) | Cost of surgical the hospitalizations performed from the 1st to the 6th month before the hospitalization |
| Medical Resource Consumption | Surgery hospitalization cost total | number (interger) | Sum of the cost of surgical the hospitalizations in the last 3 years before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization semester6 | number (interger) | Number of clinical the hospitalization performed from the 30th to the 36th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization semester5 | number (interger) | Number of clinical the hospitalization performed from the 24th to the 30th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization semester4 | number (interger) | Number of clinical the hospitalization performed from the 18th to the 24th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization semester3 | number (interger) | Number of clinical the hospitalization performed from the 12th to the 18th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization semester2 | number (interger) | Number of clinical the hospitalization performed from the 6th to the 12th month before the hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Clinic hospitalization semester1 | number (interger) | Number of clinical the hospitalization performed from the 1st to the 6th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization total | number (interger) | Sum of the number of clinical hospitalizations in the last 3 years before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost semester6 | number (interger) | Cost of clinical the hospitalizations performed from the 30th to the 36th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost semester5 | number (interger) | Cost of clinical the hospitalizations performed from the 24th to the 30th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost semester4 | number (interger) | Cost of clinical the hospitalizations performed from the 18th to the 24th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost semester3 | number (interger) | Cost of clinical the hospitalizations performed from the 12th to the 18th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost semester2 | number (interger) | Cost of clinical the hospitalizations performed from the 6th to the 12th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost semester1 | number (interger) | Cost of clinical the hospitalizations performed from the 1st to the 6th month before the hospitalization |
| Medical Resource Consumption | Clinic hospitalization cost total | number (interger) | Sum of the cost of clinical hospitalizations in the last 3 years before the hospitalization |
| Medical Resource Consumption | Day-Hospital semestre6 | number (interger) | Number of Day-Hospital the hospitalizations performed from the 30th to the 36th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital semestre5 | number (interger) | Number of Day-Hospital the hospitalizations performed from the 24th to the 30th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital semestre4 | number (interger) | Number of Day-Hospital the hospitalizations performed from the 18th to the 24th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital semestre3 | number (interger) | Number of Day-Hospital the hospitalizations performed from the 12th to the 18th month before the hospitalization |

| | | | |
|---|---|---|---|
| Medical Resource Consumption | Day-Hospital semestre2 | number (interger) | Number of Day-Hospital the hospitalizations performed from the 6th to the 12th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital semestre1 | number (interger) | Number of Day-Hospital the hospitalizations performed from the 1st to the 6th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital total | number (interger) | Sum of the number of Day-Hospital hospitalizations in the last 3 years before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost semester6 | number (interger) | Cost of Day-Hospital the hospitalizations performed from the 30th to the 36th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost semester5 | number (interger) | Cost of Day-Hospital the hospitalizations performed from the 24th to the 30th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost semester4 | number (interger) | Cost of Day-Hospital the hospitalizations performed from the 18th to the 24th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost semester3 | number (interger) | Cost of Day-Hospital the hospitalizations performed from the 12th to the 18th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost semester2 | number (interger) | Cost of Day-Hospital the hospitalizations performed from the 6th to the 12th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost semester1 | number (interger) | Cost of Day-Hospital the hospitalizations performed from the 1st to the 6th month before the hospitalization |
| Medical Resource Consumption | Day-Hospital cost total | number (interger) | Sum of the cost of Day-Hospital hospitalizations in the last 3 years before the hospitalization |

Table A.1: Dictionary of Variables.

# B
# Correlation Matrix

Table B.1: Spearman's rank correlation coefficient for numeric variables.

| | Exam semester6 | Exam semester5 | Exam semester4 | Exam semester3 | Exam semester2 | Exam semester1 | Image month3 | Image month2 | Image month1 | Consult semester6 | Consult semester5 | Consult semester4 | Consult semester3 | Consult semester2 | Consult semester1 | Emergency year3 | Emergency year2 | Emergency year1 | Hemodialysis year3 | Hemodialysis year2 | Hemodialysis year1 | Cardiovascular year3 | Cardiovascular year2 | Cardiovascular year1 | Diabetes year3 | Diabetes year2 | Diabetes year1 | Musculoskeletal year3 | Musculoskeletal year2 | Musculoskeletal year1 | Neoplasm year3 | Neoplasm year2 | Neoplasm year1 | Obstructive year3 | Obstructive year2 | Obstructive year1 | Age |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Exam semester6** | - | 0.52 | 0.35 | 0.27 | 0.20 | 0.16 | 0.07 | 0.07 | 0.04 | 0.76 | 0.54 | 0.40 | 0.29 | 0.22 | 0.16 | 0.41 | 0.20 | 0.05 | 0.03 | 0.04 | 0.03 | 0.33 | 0.18 | 0.12 | 0.42 | 0.22 | 0.14 | 0.24 | 0.11 | 0.08 | 0.05 | 0.04 | 0.02 | 0.06 | 0.08 | 0.04 | 0.12 |
| **Exam semester5** | 0.52 | - | 0.53 | 0.41 | 0.30 | 0.21 | 0.09 | 0.12 | 0.08 | 0.54 | 0.77 | 0.58 | 0.40 | 0.31 | 0.19 | 0.49 | 0.31 | 0.07 | 0.02 | 0.02 | 0.03 | 0.38 | 0.27 | 0.14 | 0.49 | 0.30 | 0.17 | 0.23 | 0.13 | 0.10 | 0.10 | 0.11 | 0.04 | 0.14 | 0.09 | 0.04 | 0.14 |
| **Exam semester4** | 0.35 | 0.53 | - | 0.49 | 0.38 | 0.25 | 0.12 | 0.09 | 0.07 | 0.34 | 0.53 | 0.77 | 0.53 | 0.39 | 0.20 | 0.37 | 0.40 | 0.10 | 0.01 | 0.03 | 0.06 | 0.21 | 0.34 | 0.19 | 0.29 | 0.44 | 0.23 | 0.19 | 0.18 | 0.10 | 0.09 | 0.11 | 0.04 | 0.09 | 0.12 | 0.05 | 0.16 |
| **Exam semester3** | 0.27 | 0.41 | 0.49 | - | 0.46 | 0.33 | 0.15 | 0.15 | 0.12 | 0.27 | 0.39 | 0.52 | 0.78 | 0.52 | 0.30 | 0.29 | 0.46 | 0.18 | 0.01 | 0.05 | 0.05 | 0.19 | 0.39 | 0.20 | 0.25 | 0.48 | 0.27 | 0.11 | 0.20 | 0.14 | 0.08 | 0.12 | 0.04 | 0.09 | 0.15 | 0.13 | 0.17 |
| **Exam semester2** | 0.20 | 0.30 | 0.38 | 0.46 | - | 0.39 | 0.25 | 0.17 | 0.15 | 0.20 | 0.27 | 0.39 | 0.50 | 0.73 | 0.40 | 0.20 | 0.34 | 0.25 | 0.00 | 0.05 | 0.06 | 0.13 | 0.24 | 0.25 | 0.18 | 0.29 | 0.39 | 0.08 | 0.13 | 0.12 | 0.07 | 0.10 | 0.06 | 0.06 | 0.08 | 0.15 | 0.16 |
| **Exam semester1** | 0.16 | 0.21 | 0.25 | 0.33 | 0.39 | - | 0.35 | 0.36 | 0.27 | 0.13 | 0.20 | 0.23 | 0.30 | 0.39 | 0.63 | 0.12 | 0.16 | 0.18 | -0.04 | 0.02 | 0.05 | 0.08 | 0.17 | 0.30 | 0.12 | 0.21 | 0.47 | 0.07 | 0.10 | 0.13 | 0.05 | 0.07 | 0.13 | 0.04 | 0.05 | 0.15 | 0.19 |
| **Image month3** | 0.07 | 0.09 | 0.12 | 0.15 | 0.25 | 0.35 | - | 0.21 | 0.12 | 0.07 | 0.08 | 0.12 | 0.16 | 0.24 | 0.38 | 0.07 | 0.12 | 0.22 | -0.02 | -0.01 | 0.06 | 0.02 | 0.10 | 0.18 | 0.03 | 0.09 | 0.17 | 0.05 | 0.07 | 0.14 | 0.04 | 0.03 | 0.07 | -0.02 | 0.04 | 0.11 | 0.08 |
| **Image month2** | 0.07 | 0.12 | 0.09 | 0.15 | 0.17 | 0.36 | 0.21 | - | 0.16 | 0.07 | 0.09 | 0.10 | 0.14 | 0.17 | 0.33 | 0.07 | 0.07 | 0.16 | -0.02 | 0.00 | 0.03 | 0.03 | 0.05 | 0.16 | 0.06 | 0.10 | 0.15 | 0.02 | 0.05 | 0.11 | 0.06 | 0.06 | 0.06 | 0.01 | 0.02 | 0.08 | 0.08 |
| **Image month1** | 0.04 | 0.08 | 0.07 | 0.12 | 0.15 | 0.27 | 0.12 | 0.16 | - | 0.05 | 0.06 | 0.07 | 0.10 | 0.14 | 0.23 | 0.06 | 0.09 | 0.24 | -0.02 | 0.02 | 0.05 | 0.06 | 0.09 | 0.13 | 0.04 | 0.06 | 0.14 | 0.00 | 0.02 | 0.05 | 0.01 | 0.02 | 0.02 | 0.02 | 0.03 | 0.01 | 0.04 |
| **Consult semester6** | 0.76 | 0.54 | 0.34 | 0.27 | 0.20 | 0.13 | 0.07 | 0.07 | 0.05 | - | 0.62 | 0.43 | 0.31 | 0.25 | 0.17 | 0.48 | 0.23 | 0.05 | 0.02 | 0.03 | 0.03 | 0.30 | 0.15 | 0.11 | 0.35 | 0.20 | 0.09 | 0.27 | 0.10 | 0.07 | 0.04 | 0.03 | 0.02 | 0.05 | 0.09 | 0.05 | 0.08 |
| **Consult semester5** | 0.54 | 0.77 | 0.53 | 0.39 | 0.27 | 0.20 | 0.08 | 0.09 | 0.06 | 0.62 | - | 0.63 | 0.45 | 0.33 | 0.21 | 0.58 | 0.36 | 0.10 | 0.03 | 0.01 | 0.00 | 0.34 | 0.25 | 0.13 | 0.39 | 0.28 | 0.16 | 0.27 | 0.15 | 0.10 | 0.03 | 0.06 | 0.03 | 0.11 | 0.13 | 0.06 | 0.11 |
| **Consult semester4** | 0.40 | 0.58 | 0.77 | 0.52 | 0.39 | 0.23 | 0.12 | 0.10 | 0.07 | 0.43 | 0.63 | - | 0.62 | 0.47 | 0.26 | 0.45 | 0.48 | 0.13 | 0.00 | 0.02 | 0.04 | 0.23 | 0.31 | 0.17 | 0.30 | 0.34 | 0.20 | 0.22 | 0.22 | 0.11 | 0.04 | 0.08 | 0.02 | 0.10 | 0.13 | 0.09 | 0.12 |
| **Consult semester3** | 0.29 | 0.40 | 0.53 | 0.78 | 0.50 | 0.30 | 0.16 | 0.14 | 0.10 | 0.31 | 0.45 | 0.62 | - | 0.63 | 0.35 | 0.36 | 0.58 | 0.25 | -0.02 | 0.02 | 0.04 | 0.17 | 0.35 | 0.20 | 0.22 | 0.38 | 0.24 | 0.14 | 0.25 | 0.16 | 0.03 | 0.09 | 0.01 | 0.08 | 0.14 | 0.13 | 0.12 |
| **Consult semester2** | 0.22 | 0.31 | 0.39 | 0.52 | 0.73 | 0.39 | 0.24 | 0.17 | 0.14 | 0.25 | 0.33 | 0.47 | 0.63 | - | 0.53 | 0.25 | 0.42 | 0.32 | -0.02 | -0.02 | 0.00 | 0.15 | 0.25 | 0.25 | 0.17 | 0.28 | 0.29 | 0.11 | 0.19 | 0.21 | 0.05 | 0.09 | 0.03 | 0.06 | 0.10 | 0.14 | 0.12 |
| **Consult semester1** | 0.16 | 0.19 | 0.20 | 0.30 | 0.40 | 0.63 | 0.38 | 0.33 | 0.23 | 0.17 | 0.21 | 0.26 | 0.35 | 0.53 | - | 0.14 | 0.20 | 0.27 | -0.03 | -0.03 | -0.04 | 0.07 | 0.13 | 0.22 | 0.09 | 0.16 | 0.27 | 0.09 | 0.14 | 0.21 | 0.02 | 0.06 | 0.03 | 0.04 | 0.06 | 0.14 | 0.09 |
| **Emergency year3** | 0.41 | 0.49 | 0.37 | 0.29 | 0.20 | 0.12 | 0.07 | 0.07 | 0.06 | 0.48 | 0.58 | 0.45 | 0.36 | 0.25 | 0.14 | - | 0.44 | 0.20 | 0.02 | 0.01 | 0.00 | 0.27 | 0.20 | 0.11 | 0.23 | 0.17 | 0.10 | 0.19 | 0.13 | 0.06 | 0.03 | 0.03 | 0.01 | 0.08 | 0.13 | 0.08 | 0.03 |
| **Emergency year2** | 0.20 | 0.31 | 0.40 | 0.46 | 0.34 | 0.16 | 0.12 | 0.07 | 0.09 | 0.23 | 0.36 | 0.48 | 0.58 | 0.42 | 0.20 | 0.44 | - | 0.37 | -0.02 | 0.02 | 0.02 | 0.14 | 0.26 | 0.14 | 0.13 | 0.24 | 0.13 | 0.11 | 0.15 | 0.11 | 0.03 | 0.04 | -0.01 | 0.05 | 0.13 | 0.07 | 0.01 |
| **Emergency year1** | 0.05 | 0.07 | 0.10 | 0.18 | 0.25 | 0.18 | 0.22 | 0.16 | 0.24 | 0.05 | 0.10 | 0.13 | 0.25 | 0.32 | 0.27 | 0.20 | 0.37 | - | -0.02 | 0.00 | 0.01 | 0.02 | 0.11 | 0.16 | 0.02 | 0.06 | 0.08 | 0.05 | 0.07 | 0.13 | 0.01 | 0.00 | 0.01 | -0.05 | 0.04 | 0.06 | -0.10 |
| **Hemodialysis year3** | 0.03 | 0.02 | 0.01 | 0.01 | 0.00 | -0.04 | -0.02 | -0.02 | -0.02 | 0.02 | 0.03 | 0.00 | -0.02 | -0.02 | -0.03 | 0.02 | -0.02 | -0.02 | - | 0.44 | 0.12 | 0.05 | -0.01 | 0.01 | 0.05 | -0.01 | -0.01 | 0.00 | 0.00 | -0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.01 | 0.01 |
| **Hemodialysis year2** | 0.04 | 0.02 | 0.03 | 0.05 | 0.05 | 0.02 | -0.01 | 0.00 | 0.02 | 0.03 | 0.01 | 0.02 | 0.02 | -0.02 | -0.03 | 0.01 | 0.02 | 0.00 | 0.44 | - | 0.48 | 0.07 | 0.04 | 0.05 | 0.03 | 0.03 | 0.03 | -0.01 | -0.01 | -0.01 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | -0.01 | 0.02 |
| **Hemodialysis year1** | 0.03 | 0.03 | 0.06 | 0.05 | 0.06 | 0.05 | 0.06 | 0.03 | 0.05 | 0.03 | 0.00 | 0.04 | 0.04 | 0.00 | -0.04 | 0.00 | 0.02 | 0.01 | 0.12 | 0.48 | - | 0.03 | 0.05 | 0.13 | 0.04 | 0.09 | 0.06 | -0.01 | -0.01 | 0.00 | 0.00 | -0.01 | 0.07 | 0.00 | -0.01 | -0.01 | 0.07 |
| **Cardiovascular year3** | 0.33 | 0.38 | 0.21 | 0.19 | 0.13 | 0.08 | 0.02 | 0.03 | 0.06 | 0.30 | 0.34 | 0.23 | 0.17 | 0.15 | 0.07 | 0.27 | 0.14 | 0.02 | 0.05 | 0.07 | 0.03 | - | 0.22 | 0.14 | 0.40 | 0.17 | 0.10 | 0.08 | 0.04 | 0.02 | 0.13 | 0.08 | 0.02 | 0.15 | 0.08 | 0.05 | 0.11 |
| **Cardiovascular year2** | 0.18 | 0.27 | 0.34 | 0.39 | 0.24 | 0.17 | 0.10 | 0.05 | 0.09 | 0.15 | 0.25 | 0.31 | 0.35 | 0.25 | 0.13 | 0.20 | 0.26 | 0.11 | -0.01 | 0.04 | 0.05 | 0.22 | - | 0.24 | 0.19 | 0.32 | 0.19 | 0.14 | 0.13 | 0.09 | 0.05 | 0.12 | 0.02 | 0.04 | 0.13 | 0.10 | 0.15 |
| **Cardiovascular year1** | 0.12 | 0.14 | 0.19 | 0.20 | 0.25 | 0.30 | 0.18 | 0.16 | 0.13 | 0.11 | 0.13 | 0.17 | 0.20 | 0.25 | 0.22 | 0.11 | 0.14 | 0.16 | 0.01 | 0.05 | 0.13 | 0.14 | 0.24 | - | 0.15 | 0.18 | 0.31 | 0.06 | 0.08 | 0.09 | 0.03 | 0.04 | 0.09 | 0.03 | 0.06 | 0.18 | 0.25 |
| **Diabetes year3** | 0.42 | 0.49 | 0.29 | 0.25 | 0.18 | 0.12 | 0.03 | 0.06 | 0.04 | 0.35 | 0.39 | 0.30 | 0.22 | 0.17 | 0.09 | 0.23 | 0.13 | 0.02 | 0.05 | 0.03 | 0.04 | 0.40 | 0.19 | 0.15 | - | 0.33 | 0.23 | 0.15 | 0.08 | 0.06 | 0.03 | 0.05 | 0.01 | 0.10 | 0.05 | 0.02 | 0.15 |
| **Diabetes year2** | 0.22 | 0.30 | 0.44 | 0.48 | 0.29 | 0.21 | 0.09 | 0.10 | 0.06 | 0.20 | 0.28 | 0.34 | 0.38 | 0.28 | 0.16 | 0.17 | 0.24 | 0.06 | -0.01 | 0.03 | 0.09 | 0.17 | 0.32 | 0.18 | 0.33 | - | 0.33 | 0.15 | 0.11 | 0.10 | 0.01 | 0.04 | 0.03 | 0.05 | 0.08 | 0.09 | 0.19 |
| **Diabetes year1** | 0.14 | 0.17 | 0.23 | 0.27 | 0.39 | 0.47 | 0.17 | 0.15 | 0.14 | 0.09 | 0.16 | 0.20 | 0.24 | 0.29 | 0.27 | 0.10 | 0.13 | 0.08 | -0.01 | 0.03 | 0.06 | 0.10 | 0.19 | 0.31 | 0.23 | 0.33 | - | 0.05 | 0.08 | 0.09 | 0.02 | 0.03 | 0.01 | 0.03 | 0.06 | 0.17 | 0.27 |
| **Musculoskeletal year3** | 0.24 | 0.23 | 0.19 | 0.11 | 0.08 | 0.07 | 0.05 | 0.02 | 0.00 | 0.27 | 0.27 | 0.22 | 0.14 | 0.11 | 0.09 | 0.19 | 0.11 | 0.05 | 0.00 | -0.01 | -0.01 | 0.08 | 0.14 | 0.06 | 0.15 | 0.15 | 0.05 | - | 0.18 | 0.11 | 0.06 | 0.02 | 0.00 | 0.03 | 0.04 | 0.05 | 0.08 |
| **Musculoskeletal year2** | 0.11 | 0.13 | 0.18 | 0.20 | 0.13 | 0.10 | 0.07 | 0.05 | 0.02 | 0.10 | 0.15 | 0.22 | 0.25 | 0.19 | 0.14 | 0.13 | 0.15 | 0.07 | 0.00 | -0.01 | -0.01 | 0.04 | 0.13 | 0.08 | 0.08 | 0.11 | 0.08 | 0.18 | - | 0.27 | 0.04 | 0.08 | 0.04 | -0.01 | 0.10 | 0.05 | 0.11 |
| **Musculoskeletal year1** | 0.08 | 0.10 | 0.10 | 0.14 | 0.12 | 0.13 | 0.14 | 0.11 | 0.05 | 0.07 | 0.10 | 0.11 | 0.16 | 0.21 | 0.21 | 0.06 | 0.11 | 0.13 | -0.01 | -0.01 | 0.00 | 0.02 | 0.09 | 0.09 | 0.06 | 0.10 | 0.09 | 0.11 | 0.27 | - | 0.02 | 0.03 | 0.01 | 0.02 | 0.03 | 0.06 | 0.18 |
| **Neoplasm year3** | 0.05 | 0.10 | 0.09 | 0.08 | 0.07 | 0.05 | 0.04 | 0.06 | 0.01 | 0.04 | 0.03 | 0.04 | 0.03 | 0.05 | 0.02 | 0.03 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.13 | 0.05 | 0.03 | 0.03 | 0.01 | 0.02 | 0.06 | 0.04 | 0.02 | - | 0.51 | 0.17 | 0.00 | 0.00 | -0.01 | 0.01 |
| **Neoplasm year2** | 0.04 | 0.11 | 0.11 | 0.12 | 0.10 | 0.07 | 0.03 | 0.06 | 0.02 | 0.03 | 0.06 | 0.08 | 0.09 | 0.09 | 0.06 | 0.03 | 0.04 | 0.00 | 0.00 | 0.00 | -0.01 | 0.08 | 0.12 | 0.04 | 0.05 | 0.04 | 0.03 | 0.02 | 0.08 | 0.03 | 0.51 | - | 0.35 | 0.00 | -0.01 | 0.03 | 0.06 |
| **Neoplasm year1** | 0.02 | 0.04 | 0.04 | 0.04 | 0.06 | 0.13 | 0.07 | 0.06 | 0.02 | 0.02 | 0.03 | 0.02 | 0.01 | 0.03 | 0.03 | 0.01 | -0.01 | 0.01 | 0.00 | 0.04 | 0.07 | 0.02 | 0.02 | 0.09 | 0.01 | 0.03 | 0.01 | 0.00 | 0.04 | 0.01 | 0.17 | 0.35 | - | -0.01 | -0.01 | 0.02 | 0.06 |
| **Obstructive year3** | 0.06 | 0.14 | 0.09 | 0.09 | 0.06 | 0.04 | -0.02 | 0.01 | 0.02 | 0.05 | 0.11 | 0.10 | 0.08 | 0.06 | 0.04 | 0.08 | 0.05 | -0.05 | 0.00 | 0.00 | 0.00 | 0.15 | 0.04 | 0.03 | 0.10 | 0.05 | 0.03 | 0.03 | -0.01 | 0.02 | 0.00 | 0.00 | -0.01 | - | 0.11 | 0.10 | 0.06 |
| **Obstructive year2** | 0.08 | 0.09 | 0.12 | 0.15 | 0.08 | 0.05 | 0.04 | 0.02 | 0.03 | 0.09 | 0.13 | 0.13 | 0.14 | 0.10 | 0.06 | 0.13 | 0.13 | 0.04 | 0.00 | 0.00 | -0.01 | 0.08 | 0.13 | 0.06 | 0.05 | 0.08 | 0.06 | 0.04 | 0.10 | 0.03 | 0.00 | -0.01 | -0.01 | 0.11 | - | 0.10 | 0.04 |
| **Obstructive year1** | 0.04 | 0.04 | 0.05 | 0.13 | 0.15 | 0.15 | 0.11 | 0.08 | 0.01 | 0.05 | 0.06 | 0.09 | 0.13 | 0.14 | 0.14 | 0.08 | 0.07 | 0.06 | -0.01 | -0.01 | -0.01 | 0.05 | 0.10 | 0.18 | 0.02 | 0.09 | 0.17 | 0.05 | 0.05 | 0.06 | -0.01 | 0.03 | 0.02 | 0.10 | 0.10 | - | 0.06 |
| **Age** | 0.12 | 0.14 | 0.16 | 0.17 | 0.16 | 0.19 | 0.08 | 0.08 | 0.04 | 0.08 | 0.11 | 0.12 | 0.12 | 0.12 | 0.09 | 0.03 | 0.01 | -0.10 | 0.01 | 0.02 | 0.07 | 0.11 | 0.15 | 0.25 | 0.15 | 0.19 | 0.27 | 0.08 | 0.11 | 0.18 | 0.01 | 0.06 | 0.06 | 0.06 | 0.04 | 0.06 | - |

Table B.2: Cramer's correlation coeficient for categorical variables.

| | Gender | Flag Genetic | Admission Type | Flag Surgery | Flag psycho year3 | Flag psycho year2 | Flag psycho year1 | Admission Group |
|---|---|---|---|---|---|---|---|---|
| **Gender** | - | 0.02 | 0.00 | 0.01 | 0.01 | 0.01 | 0.05 | 0.38 |
| **Flag Genetic** | 0.02 | - | 0.04 | 0.02 | 0.01 | 0.03 | 0.04 | 0.12 |
| **Admission Type** | 0.00 | 0.04 | - | 0.17 | 0.00 | 0.05 | 0.07 | 0.43 |
| **Flag Surgery** | 0.01 | 0.02 | 0.17 | - | 0.02 | 0.01 | 0.08 | 0.97 |
| **Flag psycho year3** | 0.01 | 0.01 | 0.00 | 0.02 | - | 0.29 | 0.15 | 0.11 |
| **Flag psycho year2** | 0.01 | 0.03 | 0.05 | 0.01 | 0.29 | - | 0.39 | 0.14 |
| **Flag psycho year1** | 0.05 | 0.04 | 0.07 | 0.08 | 0.15 | 0.39 | - | 0.29 |
| **Admission Group** | 0.38 | 0.12 | 0.43 | 0.97 | 0.11 | 0.14 | 0.29 | - |

# C
# Learning analysis of cycle 1

All the ML technique were implemented in the R programming language using, the "caret" package. Also, create table with package and the use. Table C.1 depicts the ML techniques used and their corresponding hyperparameters settings.

Table C.1: Machine Learning techniques hyperparameters for cylce 1.

| Machine Learning technique | Hyperparameters* | Value |
|---|---|---|
| Ridge Regression | Mixing Percentage | 0 |
| | Regularization Parameter | default |
| LASSO Regression | Mixing Percentage | 1 |
| | Regularization Parameter | default |
| CART | Complexity Parameter | default |

*Hyperparameters according to the "glmnet" and "rpart" packages of the R programming language

Figures C.1, C.2 and C.3 show the optimization of the hyperparameters during the CV procedure for RR, LASSO and CART, respectively.

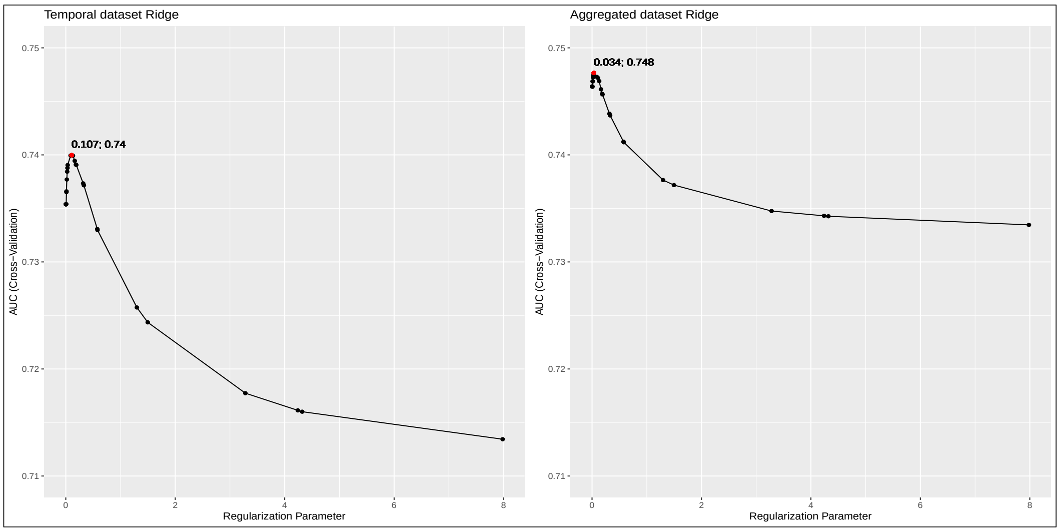


Figure C.1: Hyperparameter tuning for RR and both dataset "Temporal" and "Aggregated"

## C.1
## Variables Coefficient for RR and LASSO

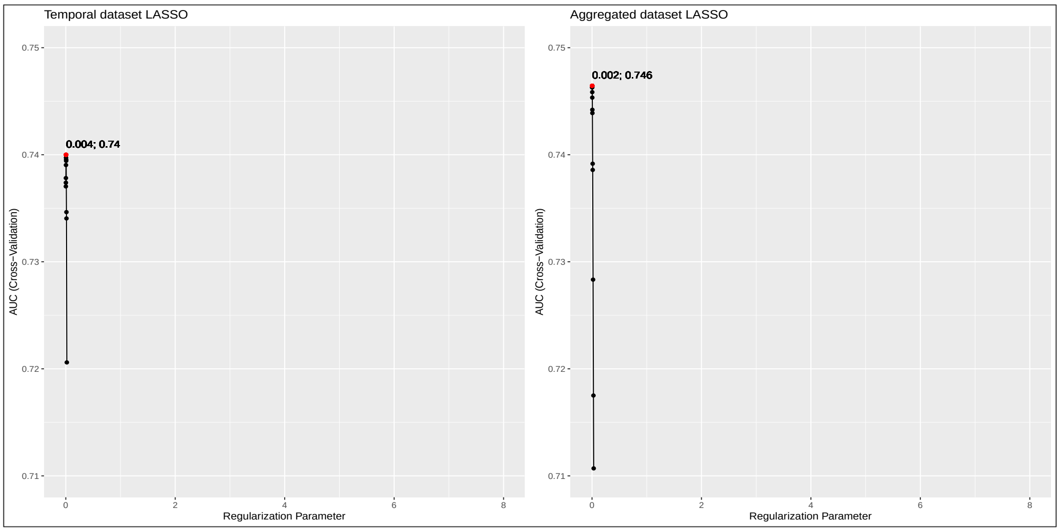


Figure C.2: Hyperparameter tuning for LASSO and both dataset "Temporal" and "Aggregated"

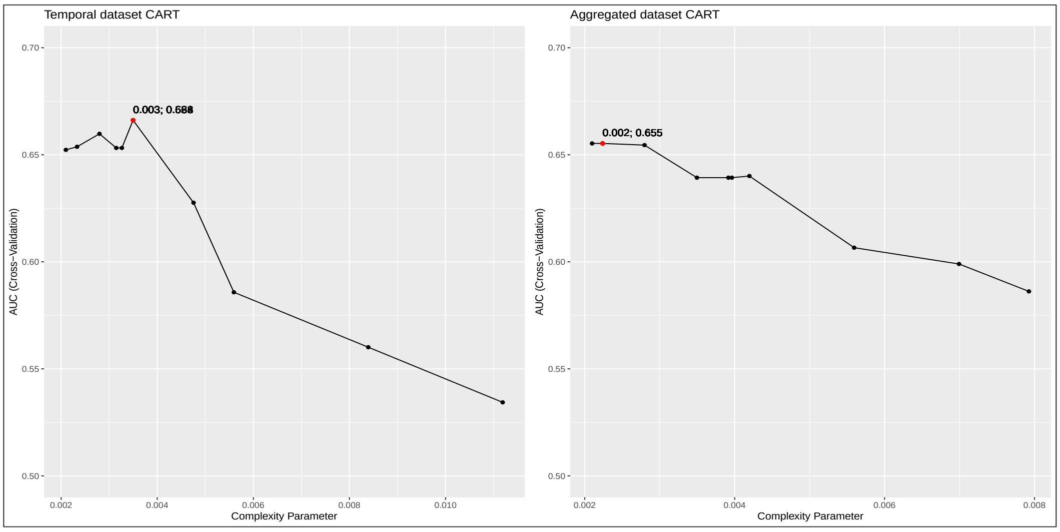


Figure C.3: Hyperparameter tuning for CART and both dataset "Temporal" and "Aggregated"

Table C.2: Variable coefficients, RR for "Temporal dataset"

| **Variables** | **Coefficients** |
|---|---|
| Intercept | 2.36 |
| Gender | -0.15 |
| Admission Type | -0.06 |
| Exam semester6 | 0.08 |
| Exam semester5 | -0.18 |
| Exam semester4 | -0.44 |
| Exam semester3 | -0.26 |
| Exam semester2 | -0.33 |
| Exam semester1 | -1.05 |
| Image month3 | -1.41 |
| Image month2 | -1.08 |
| Image month1 | -1.90 |
| Flag psycho year3 | 0.26 |
| Flag psycho year2 | 0.08 |
| Flag psycho year1 | -0.28 |
| Consult semester6 | -0.93 |
| Consult semester5 | 0.20 |
| Consult semester4 | 0.18 |
| Consult semester3 | 0.06 |
| Consult semester2 | -0.17 |
| Consult semester1 | 0.19 |
| Emergency year3 | 0.24 |
| Emergency year2 | -0.06 |
| Emergency year1 | -0.39 |
| Hemodialysis year3 | -2.01 |
| Hemodialysis year2 | -1.41 |
| Hemodialysis year1 | 0.35 |
| Cardiovascular year3 | 0.29 |
| Cardiovascular year2 | -0.31 |
| Cardiovascular year1 | -4.42 |
| Diabetes year3 | -0.61 |
| Diabetes year2 | 0.06 |
| Diabetes year1 | -1.85 |
| Musculoskeletal year3 | 0.46 |
| Musculoskeletal year2 | 0.69 |
| Musculoskeletal year1 | 0.12 |
| Neoplasm year3 | 1.14 |
| Neoplasm year2 | -0.93 |
| Neoplasm year1 | -1.46 |
| Obstructive year3 | -1.02 |
| Obstructive year2 | -1.06 |
| Obstructive year1 | -2.23 |
| Age | -0.95 |
| Admission Group CABECA E PESCOCO | -0.21 |

| | |
|---|---|
| Admission Group CLINICA GERAL | 0.08 |
| Admission Group ELETROFISIOLOGICOS MECANICOS E FUNCIONAIS | 0.47 |
| Admission Group ENDOSCOPICOS | 0.54 |
| Admission Group EXAMES ESPECIFICOS | 0.78 |
| Admission Group HOSPITAL DIA | 0.70 |
| Admission Group MEDICINA LABORATORIAL | 0.25 |
| Admission Group MEDICINA NUCLEAR | 0.68 |
| Admission Group MEDICINA TRANSFUSIONAL | -1.13 |
| Admission Group METODOS DIAGNOSTICOS POR IMAGEM | -0.03 |
| Admission Group NARIZ E SEIOS PARANASAIS | -0.64 |
| Admission Group OLHOS | 0.76 |
| Admission Group ORELHA | -0.19 |
| Admission Group OUTRAS INTERNACOES | 0.14 |
| Admission Group OUTROS | 0.61 |
| Admission Group OUTROS PROCEDIMENTOS INVASIVOS | -0.11 |
| Admission Group PAREDE TORACICA | -0.07 |
| Admission Group PELE E TECIDO CELULAR SUBCUTANEO ANEXOS | 0.22 |
| Admission Group PROCEDIMENTOS CLINICOS AMBULATORIAIS | -0.23 |
| Admission Group PROCEDIMENTOS CLINICOS HOSPITALARES | -0.35 |
| Admission Group RESSONANCIA MAGNETICA | 0.58 |
| Admission Group SISTEMA CARDIO CIRCULATORIO | 0.04 |
| Admission Group SISTEMA DIGESTIVO E ANEXOS | -0.38 |
| Admission Group SISTEMA GENITAL E REPRODUTOR FEMININO | 0.21 |
| Admission Group SISTEMA GENITAL E REPRODUTOR MASCULINO | 0.67 |
| Admission Group SISTEMA MUSCULO ESQUELETICO E ARTICULACOES | -0.70 |
| Admission Group SISTEMA NERVOSO CENTRAL E PERIFERICO | -0.41 |
| Admission Group SISTEMA RESPIRATORIO E MEDIASTINO | -1.15 |
| Admission Group SISTEMA URINARIO | -0.71 |
| Admission Group TESTES PARA DIAGNOSTICOS | 0.69 |
| Admission Group TOMOGRAFIA COMPUTADORIZADA | -0.11 |
| Admission Group TRATAMENTOS ESPECIAIS | 0.78 |
| Admission Group ULTRASSONOGRAFIA | 0.31 |
| Flag Surgery | -0.30 |

Table C.3: Variable coefficients, RR for "Aggregated dataset"

| **Variables** | **Coefficients** |
|---|---|
| Intercept | 2.65 |
| Flag Genetic | -0.58 |

| | |
|---|---|
| Cardiovascular flag | -0.32 |
| Diabetes flag | -0.46 |
| Musculoskeletal flag | -0.14 |
| Neoplasm flag | -0.30 |
| Obstructive flag | -0.71 |
| Flag Surgery | -0.35 |
| Exam total | -0.86 |
| Consult total | 0.72 |
| Image total | -3.73 |
| Emergecy total | -0.37 |
| Hemodialysis total | -0.52 |
| Age | -1.17 |
| Admission Type | -0.12 |
| Gender | -0.20 |
| Admission Group CABECA E PESCOCO | -0.35 |
| Admission Group CLINICA GERAL | 0.07 |
| Admission Group ELETROFISIOLOGICOS MECANICOS E FUNCIONAIS | 0.78 |
| Admission Group ENDOSCOPICOS | 0.77 |
| Admission Group EXAMES ESPECIFICOS | 1.30 |
| Admission Group HOSPITAL DIA | 1.08 |
| Admission Group MEDICINA LABORATORIAL | 0.32 |
| Admission Group MEDICINA NUCLEAR | 1.21 |
| Admission Group MEDICINA TRANSFUSIONAL | -1.49 |
| Admission Group METODOS DIAGNOSTICOS POR IMAGEM | -0.06 |
| Admission Group NARIZ E SEIOS PARANASAIS | -0.84 |
| Admission Group OLHOS | 1.21 |
| Admission Group ORELHA | -0.40 |
| Admission Group OUTRAS INTERNACOES | 0.13 |
| Admission Group OUTROS | 0.99 |
| Admission Group OUTROS PROCEDIMENTOS INVASIVOS | -0.18 |
| Admission Group PAREDE TORACICA | -0.09 |
| Admission Group PELE E TECIDO CELULAR SUBCUTANEO ANEXOS | 0.34 |
| Admission Group PROCEDIMENTOS CLINICOS AMBULATORIAIS | -0.44 |
| Admission Group PROCEDIMENTOS CLINICOS HOSPITALARES | -0.60 |
| Admission Group RESSONANCIA MAGNETICA | 0.86 |
| Admission Group SISTEMA CARDIO CIRCULATORIO | 0.04 |
| Admission Group SISTEMA DIGESTIVO E ANEXOS | -0.49 |
| Admission Group SISTEMA GENITAL E REPRODUTOR FEMININO | 0.23 |
| Admission Group SISTEMA GENITAL E REPRODUTOR MASCULINO | 1.08 |

| | |
|---|---|
| Admission Group SISTEMA MUSCULO ESQUELETICO E ARTICULACOES | -0.91 |
| Admission Group SISTEMA NERVOSO CENTRAL E PERIFERICO | -0.54 |
| Admission Group SISTEMA RESPIRATORIO E MEDIASTINO | -1.42 |
| Admission Group SISTEMA URINARIO | -0.90 |
| Admission Group TESTES PARA DIAGNOSTICOS | 1.12 |
| Admission Group TOMOGRAFIA COMPUTADORIZADA | -0.14 |
| Admission Group TRATAMENTOS ESPECIAIS | 1.29 |
| Admission Group ULTRASSONOGRAFIA | 0.40 |

Table C.4: Variable coefficients, LASSO for "Temporal dataset"

| **Variables** | **Coefficients** |
|---|---|
| Intercept | 2.79 |
| Gender | -0.17 |
| Admission Type | -0.07 |
| Exam semester6 | 0.00 |
| Exam semester5 | 0.00 |
| Exam semester4 | -0.15 |
| Exam semester3 | 0.00 |
| Exam semester2 | -0.21 |
| Exam semester1 | -1.16 |
| Image month3 | -1.47 |
| Image month2 | -0.90 |
| Image month1 | -2.40 |
| Flag psycho year3 | 0.16 |
| Flag psycho year2 | 0.00 |
| Flag psycho year1 | -0.47 |
| Consult semester6 | -0.92 |
| Consult semester5 | 0.00 |
| Consult semester4 | 0.00 |
| Consult semester3 | 0.00 |
| Consult semester2 | 0.00 |
| Consult semester1 | 0.21 |
| Emergency year3 | 0.00 |
| Emergency year2 | 0.00 |
| Emergency year1 | -0.41 |
| Hemodialysis year3 | -2.12 |
| Hemodialysis year2 | -1.37 |
| Hemodialysis year1 | 0.00 |
| Cardiovascular year3 | 0.00 |
| Cardiovascular year2 | 0.00 |
| Cardiovascular year1 | -7.30 |
| Diabetes year3 | -0.17 |
| Diabetes year2 | 0.00 |

| | |
|---|---|
| Diabetes year1 | -2.21 |
| Musculoskeletal year3 | 0.00 |
| Musculoskeletal year2 | 0.70 |
| Musculoskeletal year1 | 0.00 |
| Neoplasm year3 | 0.06 |
| Neoplasm year2 | -0.04 |
| Neoplasm year1 | -1.81 |
| Obstructive year3 | -0.67 |
| Obstructive year2 | -0.38 |
| Obstructive year1 | -2.67 |
| Age | -1.53 |
| Admission Group CABECA E PESCOCO | -0.35 |
| Admission Group CLINICA GERAL | 0.00 |
| Admission Group ELETROFISIOLOGICOS MECANICOS E FUNCIONAIS | 0.14 |
| Admission Group ENDOSCOPICOS | 0.69 |
| Admission Group EXAMES ESPECIFICOS | 0.79 |
| Admission Group HOSPITAL DIA | 1.47 |
| Admission Group MEDICINA LABORATORIAL | 0.14 |
| Admission Group MEDICINA NUCLEAR | 0.00 |
| Admission Group MEDICINA TRANSFUSIONAL | -1.53 |
| Admission Group METODOS DIAGNOSTICOS POR IMAGEM | 0.00 |
| Admission Group NARIZ E SEIOS PARANASAIS | -0.93 |
| Admission Group OLHOS | 1.32 |
| Admission Group ORELHA | 0.00 |
| Admission Group OUTRAS INTERNACOES | 0.00 |
| Admission Group OUTROS | 0.00 |
| Admission Group OUTROS PROCEDIMENTOS INVASIVOS | -0.11 |
| Admission Group PAREDE TORACICA | 0.00 |
| Admission Group PELE E TECIDO CELULAR SUBCUTANEO ANEXOS | 0.04 |
| Admission Group PROCEDIMENTOS CLINICOS AMBULATORIAIS | -0.28 |
| Admission Group PROCEDIMENTOS CLINICOS HOSPITALARES | -0.59 |
| Admission Group RESSONANCIA MAGNETICA | 0.49 |
| Admission Group SISTEMA CARDIO CIRCULATORIO | 0.00 |
| Admission Group SISTEMA DIGESTIVO E ANEXOS | -0.57 |
| Admission Group SISTEMA GENITAL E REPRODUTOR FEMININO | 0.15 |
| Admission Group SISTEMA GENITAL E REPRODUTOR MASCULINO | 1.02 |
| Admission Group SISTEMA MUSCULO ESQUELETICO E ARTICULACOES | -1.04 |
| Admission Group SISTEMA NERVOSO CENTRAL E PERIFERICO | -0.50 |

| | |
|---|---|
| Admission Group SISTEMA RESPIRATORIO E MEDIASTINO | -1.51 |
| Admission Group SISTEMA URINARIO | -1.00 |
| Admission Group TESTES PARA DIAGNOSTICOS | 0.00 |
| Admission Group TOMOGRAFIA COMPUTADORIZADA | -0.19 |
| Admission Group TRATAMENTOS ESPECIAIS | 0.00 |
| Admission Group ULTRASSONOGRAFIA | 0.23 |
| Flag Surgery | -0.39 |

Table C.5: Variable coefficients, LASSO for "Aggregated dataset"

| **Variables** | **Coefficients** |
|---|---|
| Intercept | 2.86 |
| Flag Genetic | -0.64 |
| Cardiovascular flag | -0.34 |
| Diabetes flag | -0.53 |
| Musculoskeletal flag | -0.08 |
| Neoplasm flag | -0.25 |
| Obstructive flag | -0.77 |
| Flag Surgery | -0.34 |
| Exam total | -0.73 |
| Consult total | 0.87 |
| Image total | -4.33 |
| Emergecy total | -0.41 |
| Hemodialysis total | -0.51 |
| Age | -1.40 |
| Admission Type | -0.13 |
| Gender | -0.22 |
| Admission Group CABECA E PESCOCO | -0.47 |
| Admission Group CLINICA GERAL | 0.00 |
| Admission Group ELETROFISIOLOGICOS MECANICOS E FUNCIONAIS | 0.73 |
| Admission Group ENDOSCOPICOS | 0.89 |
| Admission Group EXAMES ESPECIFICOS | 1.59 |
| Admission Group HOSPITAL DIA | 1.72 |
| Admission Group MEDICINA LABORATORIAL | 0.27 |
| Admission Group MEDICINA NUCLEAR | 0.41 |
| Admission Group MEDICINA TRANSFUSIONAL | -1.69 |
| Admission Group METODOS DIAGNOSTICOS POR IMAGEM | -0.06 |
| Admission Group NARIZ E SEIOS PARANASAIS | -1.02 |
| Admission Group OLHOS | 1.59 |
| Admission Group ORELHA | -0.42 |
| Admission Group OUTRAS INTERNACOES | 0.05 |
| Admission Group OUTROS | 0.00 |
| Admission Group OUTROS PROCEDIMENTOS INVASIVOS | -0.23 |
| Admission Group PAREDE TORACICA | -0.05 |

| | |
|---|---|
| Admission Group PELE E TECIDO CELULAR SUBCUTANEO ANEXOS | 0.23 |
| Admission Group PROCEDIMENTOS CLINICOS AMBULATORIAIS | -0.54 |
| Admission Group PROCEDIMENTOS CLINICOS HOSPITALARES | -0.74 |
| Admission Group RESSONANCIA MAGNETICA | 0.89 |
| Admission Group SISTEMA CARDIO CIRCULATORIO | 0.00 |
| Admission Group SISTEMA DIGESTIVO E ANEXOS | -0.62 |
| Admission Group SISTEMA GENITAL E REPRODUTOR FEMININO | 0.16 |
| Admission Group SISTEMA GENITAL E REPRODUTOR MASCULINO | 1.42 |
| Admission Group SISTEMA MUSCULO ESQUELETICO E ARTICULACOES | -1.11 |
| Admission Group SISTEMA NERVOSO CENTRAL E PERIFERICO | -0.64 |
| Admission Group SISTEMA RESPIRATORIO E MEDIASTINO | -1.61 |
| Admission Group SISTEMA URINARIO | -1.06 |
| Admission Group TESTES PARA DIAGNOSTICOS | 0.00 |
| Admission Group TOMOGRAFIA COMPUTADORIZADA | -0.17 |
| Admission Group TRATAMENTOS ESPECIAIS | 0.19 |
| Admission Group ULTRASSONOGRAFIA | 0.37 |

## C.2 Statistical test to compare predictive models.

Table C.6: Mc Nemar's test to compare homogenity in the errors among ML techniques (first cycle).

| **A** | **B** | **Statistic** $\chi^2$ | **p-value** | **Correct A & Correct B** | **Incorrect A & Correct B** | **Incorrect A & Incorrect B** | **Correct A & Incorrect B** |
|---|---|---|---|---|---|---|---|
| RR temporal | RR aggregated | 0.3 | 0.575 | 2,277 | 134 | 1,043 | 124 |
| RR temporal | LASSO temporal | 2.1 | 0.146 | 2,315 | 67 | 1,110 | 86 |
| RR temporal | LASSO aggregated | 1.0 | 0.315 | 2,249 | 134 | 1,043 | 152 |
| RR temporal | CART temporal | 0.3 | 0.587 | 1,919 | 500 | 677 | 482 |
| RR temporal | CART aggregated | 33.8 | 0.000 | 1,894 | 337 | 840 | 507 |
| RR aggregated | LASSO temporal | 2.7 | 0.101 | 2,251 | 131 | 1,036 | 160 |
| RR aggregated | LASSO aggregated | 11.0 | 0.001 | 2,364 | 19 | 1,148 | 47 |
| RR aggregated | CART temporal | 0.1 | 0.819 | 1,948 | 471 | 696 | 463 |
| RR aggregated | CART aggregated | 38.2 | 0.000 | 1,902 | 329 | 838 | 509 |
| LASSO temporal | LASSO aggregated | 0.0 | 1.000 | 2,236 | 147 | 1,049 | 146 |
| LASSO temporal | CART temporal | 1.3 | 0.251 | 1,908 | 511 | 685 | 474 |
| LASSO temporal | CART aggregated | 27.3 | 0.000 | 1,894 | 337 | 859 | 488 |
| LASSO aggregated | CART temporal | 1.3 | 0.251 | 1,936 | 483 | 712 | 447 |
| LASSO aggregated | CART aggregated | 26.8 | 0.000 | 1,882 | 349 | 846 | 501 |
| CART temporal | CART aggregated | 45.3 | 0.000 | 1,939 | 292 | 867 | 480 |

# D
# Detailed description of the data set by subgroups

Table D.1: "Admission Subgroup" number and correponding names, Total frequency, High-cost frequency and proportion.

| Number | "Admission Subgroup" description | Frequency | High-cost frequency | High-cost (%) |
|---|---|---|---|---|
| 1 | genital\|sistema genital e reprodutor feminino\|utero | 13,234 | 1,069 | 8.1 |
| 2 | cardiovascular\|sistema cardio-circulatorio\|cirurgia venosa | 11,591 | 621 | 5.4 |
| 3 | digestivo\|sistema digestivo e anexos\|figado e vias biliares | 10,823 | 649 | 6.0 |
| 4 | outros\|outros procedimentos invasivos\|bloqueios anestesicos de nervos e estimulos neurovasculares | 9,969 | 1,206 | 12.1 |
| 5 | urinario\|sistema urinario\|ureter | 9,750 | 397 | 4.1 |
| 6 | digestivo\|sistema digestivo e anexos\|abdome, parede e cavidade | 9,506 | 941 | 9.9 |
| 7 | pele\|pele e tecido celular subcutaneo / anexos\|procedimentos | 7,157 | 733 | 10.2 |
| 8 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|procedimentos videoartroscopicos de joelho | 6,958 | 294 | 4.2 |
| 9 | digestivo\|sistema digestivo e anexos\|intestinos | 5,954 | 647 | 10.9 |
| 10 | cabeca e pescoco\|cabeca e pescoco\|faringe | 5,821 | 614 | 10.5 |
| 11 | digestivo\|sistema digestivo e anexos\|estomago | 5,037 | 173 | 3.4 |
| 12 | cabeca e pescoco\|olhos\|cristalino | 4,936 | 310 | 6.3 |
| 13 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|coluna vertebral | 4,346 | 184 | 4.2 |
| 14 | urinario\|sistema urinario\|bexiga | 4,308 | 463 | 10.7 |
| 15 | cabeca e pescoco\|nariz e seios paranasais\|nariz | 4,014 | 378 | 9.4 |

|  |  |  |  |  |
|---|---|---|---|---|
| 16 | sistema nervoso\|sistema nervoso - central e periferico\|nervos perifericos | 3,814 | 360 | 9.4 |
| 17 | mamas\|parede toracica\|mamas | 3,623 | 380 | 10.5 |
| 18 | genital\|sistema genital e reprodutor masculino\|penis | 3,526 | 323 | 9.2 |
| 19 | digestivo\|sistema digestivo e anexos\|anus | 3,236 | 237 | 7.3 |
| 20 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|procedimentos videoartroscopicos de ombro | 3,183 | 127 | 4.0 |
| 21 | cabeca e pescoco\|nariz e seios paranasais\|seios paranasais | 3,171 | 245 | 7.7 |
| 22 | digestivo\|endoscopicos\|endoscopia intervencionista | 2,803 | 312 | 11.1 |
| 23 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|tendoes, bursas e sinovias | 2,264 | 236 | 10.4 |
| 24 | genital\|sistema genital e reprodutor masculino\|testiculo | 1,698 | 137 | 8.1 |
| 25 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|joelho | 1,638 | 62 | 3.8 |
| 26 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|aparelhos gessados | 1,562 | 272 | 17.4 |
| 27 | cardiovascular\|sistema cardio-circulatorio\|hemodinamica - cardiologia intervencionista (procedimentos terapeuticos) | 1,501 | 86 | 5.7 |
| 28 | genital\|sistema genital e reprodutor masculino\|cordao espermatico | 1,457 | 87 | 6.0 |
| 29 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|pe | 1,450 | 110 | 7.6 |
| 30 | cardiovascular\|sistema cardio-circulatorio\|hemodinamica - cardiologia intervencionista (procedimentos diagnosticos) | 1,364 | 110 | 8.1 |
| 31 | cabeca e pescoco\|cabeca e pescoco\|tireoide | 1,281 | 62 | 4.8 |
| 32 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|mao | 1,271 | 107 | 8.4 |
| 33 | cardiovascular\|sistema cardio-circulatorio\|acessos vasculares | 1,244 | 132 | 10.6 |
| 34 | digestivo\|sistema digestivo e anexos\|esofago | 1,172 | 116 | 9.9 |

| | | | | |
|---|---|---|---|---|
| 35 | cabeca e pescoco\|olhos\|cornea | 1,143 | 112 | 9.8 |
| 36 | genital\|sistema genital e reprodutor feminino\|ovarios | 1,124 | 61 | 5.4 |
| 37 | genital\|sistema genital e reprodutor feminino\|cavidade e paredes pelvicas | 1,098 | 66 | 6.0 |
| 38 | cabeca e pescoco\|orelha\|orelha media | 1,097 | 118 | 10.8 |
| 39 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|tornozelo | 1,025 | 71 | 6.9 |
| 40 | urinario\|sistema urinario\|rim, bacinete e supra-renal | 1,003 | 64 | 6.4 |
| 41 | sistema nervoso\|sistema nervoso - central e periferico\|encefalo | 984 | 58 | 5.9 |
| 42 | cabeca e pescoco\|cabeca e pescoco\|cirurgia reparadora e funcional da face | 969 | 57 | 5.9 |
| 43 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|retirada de material de sintese | 963 | 79 | 8.2 |
| 44 | genital\|sistema genital e reprodutor feminino\|tubas | 928 | 49 | 5.3 |
| 45 | genital\|sistema genital e reprodutor masculino\|prostata e vesiculas seminais | 839 | 58 | 6.9 |
| 46 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|articulacao coxo-femoral | 838 | 45 | 5.4 |
| 47 | osteomuscular\|sistema musculo-esqueletico e articulacoes\|antebraco | 800 | 47 | 5.9 |

# E
# Learning analysis of cycle 2

Table E.1: Machine Learning techniques hyperparameters for cylce 2.

| **Machine Learning technique** | **Caret method** | **Hyperparameters** | **Value** |
|---|---|---|---|
| Ridge Regression | "glmnet" | Mixing Percentage | 1 |
| | | Regularization Parameter | default |
| LASSO Regression | "glmnet" | Mixing Percentage | 0 |
| | | Regularization Parameter | default |
| CART | "rpart" | Complexity Parameter | default |
| Random Forest (random search) | "ranger" | Number of Tree | 500, 1000 |
| | | Number of candidate variables at each split | Sample with replacement [1 - total_variables] |
| | | Node split criterion | Sample with replacement ["gini_index", "extra_tree"] |
| | | Minimum size of terminal nodes | Sample with replacement [1 - min(20, total_rows)] |
| Extreme Gradient Boosting (grid search) | "xgbTree" | Maximum depth of a tree (max_depth) | [1 - 5] |
| | | the number of decision trees (nrounds) | [50, 150, 250, 350, 450] |
| | | Learning rate (eta) | [0.3, 0.4] |
| | | Minimum loss reduction to split node (gamma) | 0 (default) |
| | | Subsample ratio of columns (colsample_bytree | 0.6 (default) |
| | | Minimum sum of instance weight to split node | 1 (default) |
| | | Subsample ratio of instances | [0.5, 0.8, 1] |

Table E.2: Performance results for all scenarios tested in "Asmission Subgroup" 22 and 33.

| Admissino Subgroup | Imbalance technique | ML technique | ML parameter | AUROC | AUCPR | Optimum threshold | TP | FP | FN | TN | F2 | TNR | PPV | TPR | NPV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | none | ridge | - | 0.817 | 0.677 | 0.108 | 49 | 151 | 13 | 348 | 0.550 | 0.792 | 0.248 | 0.698 | 0.964 |
| 22 | up | ridge | - | 0.817 | 0.671 | 0.487 | 49 | 140 | 13 | 358 | 0.561 | 0.788 | 0.261 | 0.719 | 0.964 |
| 22 | down | ridge | - | 0.809 | 0.665 | 0.488 | 48 | 142 | 15 | 356 | 0.545 | 0.766 | 0.253 | 0.715 | 0.961 |
| 22 | smote | ridge | - | 0.783 | 0.647 | 0.414 | 40 | 88 | 23 | 410 | 0.526 | 0.635 | 0.312 | 0.823 | 0.947 |
| 22 | rose | ridge | - | 0.196 | 0.424 | 0.363 | 62 | 498 | 0 | 0 | 0.384 | 0.997 | 0.111 | 0.000 | 0.500 |
| 22 | none | lasso | - | 0.799 | 0.678 | 0.087 | 41 | 90 | 21 | 408 | 0.543 | 0.660 | 0.318 | 0.819 | 0.951 |
| 22 | up | lasso | - | 0.802 | 0.676 | 0.404 | 47 | 140 | 16 | 359 | 0.537 | 0.750 | 0.252 | 0.720 | 0.958 |
| 22 | down | lasso | - | 0.798 | 0.675 | 0.405 | 47 | 144 | 16 | 355 | 0.535 | 0.750 | 0.250 | 0.712 | 0.958 |
| 22 | smote | lasso | - | 0.787 | 0.655 | 0.370 | 41 | 105 | 21 | 393 | 0.522 | 0.660 | 0.283 | 0.789 | 0.949 |
| 22 | rose | lasso | - | 0.217 | 0.424 | 0.091 | 62 | 498 | 0 | 0 | 0.384 | 0.997 | 0.111 | 0.000 | 0.000 |
| 22 | none | random forest | 500 | 0.848 | 0.732 | 0.159 | 45 | 92 | 17 | 406 | 0.585 | 0.724 | 0.330 | 0.815 | 0.959 |
| 22 | up | random forest | 500 | 0.847 | 0.715 | 0.383 | 45 | 89 | 17 | 410 | 0.590 | 0.724 | 0.339 | 0.822 | 0.960 |
| 22 | down | random forest | 500 | 0.834 | 0.697 | 0.519 | 47 | 107 | 16 | 392 | 0.585 | 0.750 | 0.310 | 0.786 | 0.962 |
| 22 | smote | random forest | 500 | 0.833 | 0.685 | 0.337 | 44 | 91 | 19 | 407 | 0.572 | 0.702 | 0.328 | 0.817 | 0.956 |
| 22 | rose | random forest | 500 | 0.817 | 0.694 | 0.528 | 38 | 84 | 25 | 414 | 0.544 | 0.606 | 0.387 | 0.831 | 0.947 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | none | xgbTree | - | 0.846 | 0.722 | 0.116 | 44 | 86 | 18 | 413 | 0.587 | 0.708 | 0.348 | 0.828 | 0.958 |
| 22 | up | xgbTree | - | 0.837 | 0.723 | 0.436 | 46 | 104 | 16 | 394 | 0.576 | 0.737 | 0.307 | 0.791 | 0.960 |
| 22 | down | xgbTree | - | 0.833 | 0.712 | 0.506 | 46 | 109 | 16 | 389 | 0.568 | 0.737 | 0.296 | 0.781 | 0.960 |
| 22 | smote | xgbTree | - | 0.832 | 0.715 | 0.170 | 48 | 117 | 14 | 381 | 0.579 | 0.769 | 0.291 | 0.765 | 0.964 |
| 22 | rose | xgbTree | - | 0.816 | 0.692 | 0.945 | 48 | 284 | 14 | 215 | 0.563 | 0.771 | 0.271 | 0.431 | 0.870 |
| 22 | none | rpart | information | 0.774 | 0.643 | 0.098 | 37 | 65 | 25 | 433 | 0.532 | 0.600 | 0.367 | 0.870 | 0.946 |
| 22 | up | rpart | gini | 0.784 | 0.619 | 0.643 | 42 | 79 | 20 | 419 | 0.570 | 0.676 | 0.350 | 0.842 | 0.954 |
| 22 | down | rpart | gini | 0.789 | 0.516 | 0.700 | 41 | 95 | 22 | 403 | 0.536 | 0.654 | 0.312 | 0.809 | 0.949 |
| 22 | smote | rpart | information | 0.764 | 0.599 | 0.182 | 37 | 77 | 25 | 421 | 0.510 | 0.593 | 0.326 | 0.846 | 0.943 |
| 22 | rose | rpart | information | 0.547 | 0.049 | 0.833 | 55 | 404 | 8 | 94 | 0.497 | 0.879 | 0.181 | 0.188 | 0.925 |
| 33 | none | ridge | - | 0.697 | 0.605 | 0.095 | 19 | 100 | 7 | 122 | 0.422 | 0.719 | 0.159 | 0.549 | 0.943 |
| 33 | up | ridge | - | 0.693 | 0.604 | 0.468 | 19 | 100 | 8 | 123 | 0.415 | 0.704 | 0.157 | 0.551 | 0.941 |
| 33 | down | ridge | - | 0.703 | 0.607 | 0.436 | 20 | 100 | 7 | 123 | 0.440 | 0.750 | 0.165 | 0.551 | 0.950 |
| 33 | smote | ridge | - | 0.688 | 0.598 | 0.370 | 17 | 72 | 10 | 150 | 0.431 | 0.636 | 0.189 | 0.674 | 0.940 |
| 33 | rose | ridge | - | 0.314 | 0.447 | 0.434 | 26 | 222 | 0 | 0 | 0.370 | 0.992 | 0.105 | 0.000 | 0.000 |
| 33 | none | lasso | - | 0.715 | 0.605 | 0.082 | 20 | 102 | 6 | 120 | 0.448 | 0.772 | 0.167 | 0.540 | 0.952 |
| 33 | up | lasso | - | 0.720 | 0.613 | 0.429 | 20 | 97 | 7 | 125 | 0.446 | 0.750 | 0.170 | 0.564 | 0.951 |
| 33 | down | lasso | - | 0.704 | 0.606 | 0.433 | 19 | 88 | 7 | 134 | 0.447 | 0.720 | 0.178 | 0.603 | 0.948 |
| 33 | smote | lasso | - | 0.688 | 0.594 | 0.326 | 18 | 87 | 9 | 136 | 0.424 | 0.673 | 0.171 | 0.610 | 0.941 |
| 33 | rose | lasso | - | 0.357 | 0.452 | 0.261 | 26 | 222 | 0 | 0 | 0.370 | 0.992 | 0.105 | 0.001 | 0.500 |
| 33 | none | random forest | 500 | 0.747 | 0.637 | 0.183 | 15 | 38 | 12 | 185 | 0.464 | 0.552 | 0.283 | 0.831 | 0.940 |
| 33 | up | random forest | 500 | 0.764 | 0.634 | 0.358 | 20 | 79 | 7 | 143 | 0.486 | 0.751 | 0.202 | 0.644 | 0.956 |
| 33 | down | random forest | 500 | 0.753 | 0.623 | 0.471 | 19 | 77 | 7 | 145 | 0.480 | 0.733 | 0.202 | 0.653 | 0.955 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | smote | random forest | 500 | 0.754 | 0.645 | 0.310 | 17 | 53 | 10 | 170 | 0.483 | 0.636 | 0.246 | 0.763 | 0.947 |
| 33 | rose | random forest | 500 | 0.684 | 0.606 | 0.344 | 24 | 166 | 2 | 57 | 0.415 | 0.916 | 0.130 | 0.254 | 0.970 |
| 33 | none | xgbTree | - | 0.727 | 0.640 | 0.089 | 17 | 60 | 9 | 163 | 0.478 | 0.660 | 0.228 | 0.732 | 0.948 |
| 33 | up | xgbTree | - | 0.733 | 0.638 | 0.416 | 17 | 58 | 10 | 164 | 0.466 | 0.636 | 0.225 | 0.737 | 0.945 |
| 33 | down | xgbTree | - | 0.732 | 0.623 | 0.507 | 18 | 77 | 8 | 146 | 0.449 | 0.681 | 0.190 | 0.655 | 0.946 |
| 33 | smote | xgbTree | - | 0.757 | 0.644 | 0.197 | 18 | 63 | 8 | 159 | 0.494 | 0.697 | 0.228 | 0.717 | 0.952 |
| 33 | rose | xgbTree | - | 0.691 | 0.600 | 0.001 | 17 | 89 | 10 | 133 | 0.432 | 0.632 | 0.191 | 0.599 | 0.935 |
| 33 | none | rpart | information | 0.631 | 0.520 | 0.067 | 13 | 48 | 14 | 174 | 0.384 | 0.477 | 0.216 | 0.784 | 0.927 |
| 33 | up | rpart | gini | 0.740 | 0.566 | 0.270 | 19 | 83 | 7 | 139 | 0.469 | 0.733 | 0.192 | 0.627 | 0.954 |
| 33 | down | rpart | information | 0.682 | 0.449 | 0.286 | 20 | 105 | 6 | 118 | 0.450 | 0.765 | 0.170 | 0.529 | 0.953 |
| 33 | smote | rpart | information | 0.696 | 0.550 | 0.196 | 15 | 58 | 11 | 164 | 0.424 | 0.566 | 0.212 | 0.739 | 0.936 |
| 33 | rose | rpart | gini | 0.602 | 0.344 | 0.000 | 26 | 222 | 0 | 0 | 0.370 | 0.993 | 0.105 | 0.001 | 0.500 |

Table E.3: Mc Nemar's test to compare homogenity in the errors among ML techniques (second cycle).

| **Admission Sub-group** | **A** | **B** | **Statistic** $\chi^2$ | **p-value** | **Correct A & Correct B** | **Incorrect A & Correct B** | **Incorrect A & Incorrect B** | **Correct A & Incorrect B** |
|---|---|---|---|---|---|---|---|---|
| 22 | RR none | RR up | 12.7 | 0.0004 | 1,912 | 124 | 694 | 73 |
| 22 | RR none | RR down | 4.8 | 0.0287 | 1,875 | 146 | 672 | 110 |
| 22 | RR none | RR smote | 139.6 | 0.0000 | 1,872 | 375 | 443 | 113 |
| 22 | RR none | RR rose | 1548.8 | 0.0000 | 246 | 66 | 752 | 1,739 |
| 22 | RR none | LASSO none | 183.1 | 0.0000 | 1,930 | 317 | 501 | 55 |
| 22 | RR none | LASSO up | 4.7 | 0.0302 | 1,827 | 200 | 618 | 158 |
| 22 | RR none | LASSO down | 1.3 | 0.2603 | 1,822 | 185 | 633 | 163 |
| 22 | RR none | LASSO smote | 76.7 | 0.0000 | 1,853 | 319 | 499 | 132 |
| 22 | RR none | LASSO rose | 1551.5 | 0.0000 | 246 | 65 | 753 | 1,739 |
| 22 | RR none | RF none | 150.8 | 0.0000 | 1,880 | 375 | 443 | 105 |
| 22 | RR none | RF up | 138.9 | 0.0000 | 1,831 | 443 | 375 | 154 |
| 22 | RR none | RF down | 93.7 | 0.0000 | 1,862 | 330 | 488 | 123 |
| 22 | RR none | RF smote | 145.7 | 0.0000 | 1,873 | 381 | 437 | 112 |
| 22 | RR none | RF rose | 129.4 | 0.0000 | 1,834 | 425 | 393 | 151 |
| 22 | RR none | XGB none | 166.0 | 0.0000 | 1,867 | 417 | 401 | 118 |
| 22 | RR none | XGB up | 81.3 | 0.0000 | 1,811 | 389 | 429 | 174 |
| 22 | RR none | XGB down | 59.7 | 0.0000 | 1,781 | 394 | 424 | 204 |
| 22 | RR none | XGB smote | 40.3 | 0.0000 | 1,748 | 398 | 420 | 237 |
| 22 | RR none | XGB rose | 397.7 | 0.0000 | 1,083 | 230 | 588 | 902 |
| 22 | RR none | CART none | 211.8 | 0.0000 | 1,851 | 502 | 316 | 134 |
| 22 | RR none | CART up | 182.9 | 0.0000 | 1,863 | 445 | 373 | 122 |
| 22 | RR none | CART down | 86.5 | 0.0000 | 1,786 | 434 | 384 | 199 |
| 22 | RR none | CART smote | 140.4 | 0.0000 | 1,805 | 487 | 331 | 180 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | RR none | CART rose | 982.2 | 0.0000 | 580 | 163 | 655 | 1,405 |
| 22 | RR up | RR down | 0.8 | 0.3591 | 1,912 | 109 | 658 | 124 |
| 22 | RR up | RR smote | 96.5 | 0.0000 | 1,913 | 334 | 433 | 123 |
| 22 | RR up | RR rose | 1597.8 | 0.0000 | 245 | 67 | 700 | 1,791 |
| 22 | RR up | LASSO none | 117.0 | 0.0000 | 1,953 | 294 | 473 | 83 |
| 22 | RR up | LASSO up | 0.2 | 0.6611 | 1,865 | 162 | 605 | 171 |
| 22 | RR up | LASSO down | 2.4 | 0.1181 | 1,861 | 146 | 621 | 175 |
| 22 | RR up | LASSO smote | 43.2 | 0.0000 | 1,893 | 279 | 488 | 143 |
| 22 | RR up | LASSO rose | 1600.5 | 0.0000 | 245 | 66 | 701 | 1,791 |
| 22 | RR up | RF none | 102.6 | 0.0000 | 1,914 | 341 | 426 | 122 |
| 22 | RR up | RF up | 104.4 | 0.0000 | 1,886 | 388 | 379 | 150 |
| 22 | RR up | RF down | 54.4 | 0.0000 | 1,893 | 299 | 468 | 143 |
| 22 | RR up | RF smote | 96.5 | 0.0000 | 1,901 | 353 | 414 | 135 |
| 22 | RR up | RF rose | 87.5 | 0.0000 | 1,866 | 393 | 374 | 170 |
| 22 | RR up | XGB none | 123.5 | 0.0000 | 1,913 | 371 | 396 | 123 |
| 22 | RR up | XGB up | 52.3 | 0.0000 | 1,864 | 336 | 431 | 172 |
| 22 | RR up | XGB down | 34.6 | 0.0000 | 1,830 | 345 | 422 | 206 |
| 22 | RR up | XGB smote | 19.9 | 0.0000 | 1,793 | 353 | 414 | 243 |
| 22 | RR up | XGB rose | 463.4 | 0.0000 | 1,112 | 201 | 566 | 924 |
| 22 | RR up | CART none | 167.8 | 0.0000 | 1,897 | 456 | 311 | 139 |
| 22 | RR up | CART up | 136.5 | 0.0000 | 1,903 | 405 | 362 | 133 |
| 22 | RR up | CART down | 57.0 | 0.0000 | 1,834 | 386 | 381 | 202 |
| 22 | RR up | CART smote | 100.3 | 0.0000 | 1,840 | 452 | 315 | 196 |
| 22 | RR up | CART rose | 1045.2 | 0.0000 | 591 | 152 | 615 | 1,445 |
| 22 | RR down | RR smote | 109.6 | 0.0000 | 1,903 | 344 | 438 | 118 |
| 22 | RR down | RR rose | 1571.0 | 0.0000 | 238 | 74 | 708 | 1,783 |
| 22 | RR down | LASSO none | 121.7 | 0.0000 | 1,926 | 321 | 461 | 95 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | RR down | LASSO up | 0.1 | 0.8120 | 1,803 | 224 | 558 | 218 |
| 22 | RR down | LASSO down | 0.6 | 0.4254 | 1,881 | 126 | 656 | 140 |
| 22 | RR down | LASSO smote | 47.6 | 0.0000 | 1,860 | 312 | 470 | 161 |
| 22 | RR down | LASSO rose | 1573.6 | 0.0000 | 238 | 73 | 709 | 1,783 |
| 22 | RR down | RF none | 119.1 | 0.0000 | 1,910 | 345 | 437 | 111 |
| 22 | RR down | RF up | 117.4 | 0.0000 | 1,877 | 397 | 385 | 144 |
| 22 | RR down | RF down | 64.9 | 0.0000 | 1,884 | 308 | 474 | 137 |
| 22 | RR down | RF smote | 109.2 | 0.0000 | 1,891 | 363 | 419 | 130 |
| 22 | RR down | RF rose | 101.4 | 0.0000 | 1,863 | 396 | 386 | 158 |
| 22 | RR down | XGB none | 135.4 | 0.0000 | 1,899 | 385 | 397 | 122 |
| 22 | RR down | XGB up | 59.9 | 0.0000 | 1,846 | 354 | 428 | 175 |
| 22 | RR down | XGB down | 42.4 | 0.0000 | 1,822 | 353 | 429 | 199 |
| 22 | RR down | XGB smote | 26.6 | 0.0000 | 1,794 | 352 | 430 | 227 |
| 22 | RR down | XGB rose | 423.6 | 0.0000 | 1,077 | 236 | 546 | 944 |
| 22 | RR down | CART none | 188.2 | 0.0000 | 1,896 | 457 | 325 | 125 |
| 22 | RR down | CART up | 153.5 | 0.0000 | 1,898 | 410 | 372 | 123 |
| 22 | RR down | CART down | 63.3 | 0.0000 | 1,811 | 409 | 373 | 210 |
| 22 | RR down | CART smote | 111.3 | 0.0000 | 1,829 | 463 | 319 | 192 |
| 22 | RR down | CART rose | 990.7 | 0.0000 | 559 | 184 | 598 | 1,462 |
| 22 | RR smote | RR rose | 1727.6 | 0.0000 | 197 | 115 | 441 | 2,050 |
| 22 | RR smote | LASSO none | 0.0 | 1.0000 | 2,069 | 178 | 378 | 178 |
| 22 | RR smote | LASSO up | 109.5 | 0.0000 | 1,918 | 109 | 447 | 329 |
| 22 | RR smote | LASSO down | 121.0 | 0.0000 | 1,891 | 116 | 440 | 356 |
| 22 | RR smote | LASSO smote | 28.1 | 0.0000 | 2,112 | 60 | 496 | 135 |
| 22 | RR smote | LASSO rose | 1730.2 | 0.0000 | 197 | 114 | 442 | 2,050 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | RR smote | RF none | 0.1 | 0.7202 | 2,060 | 195 | 361 | 187 |
| 22 | RR smote | RF up | 1.6 | 0.2040 | 2,051 | 223 | 333 | 196 |
| 22 | RR smote | RF down | 7.1 | 0.0077 | 2,014 | 178 | 378 | 233 |
| 22 | RR smote | RF smote | 0.1 | 0.7345 | 2,094 | 160 | 396 | 153 |
| 22 | RR smote | RF rose | 0.3 | 0.5823 | 2,053 | 206 | 350 | 194 |
| 22 | RR smote | XGB none | 3.2 | 0.0722 | 2,065 | 219 | 337 | 182 |
| 22 | RR smote | XGB up | 4.9 | 0.0274 | 2,006 | 194 | 362 | 241 |
| 22 | RR smote | XGB down | 10.4 | 0.0012 | 1,969 | 206 | 350 | 278 |
| 22 | RR smote | XGB smote | 18.4 | 0.0000 | 1,925 | 221 | 335 | 322 |
| 22 | RR smote | XGB rose | 682.2 | 0.0000 | 1,142 | 171 | 385 | 1,105 |
| 22 | RR smote | CART none | 27.6 | 0.0000 | 2,100 | 253 | 303 | 147 |
| 22 | RR smote | CART up | 8.9 | 0.0029 | 2,075 | 233 | 323 | 172 |
| 22 | RR smote | CART down | 1.4 | 0.2397 | 1,989 | 231 | 325 | 258 |
| 22 | RR smote | CART smote | 4.1 | 0.0417 | 2,036 | 256 | 300 | 211 |
| 22 | RR smote | CART rose | 1250.8 | 0.0000 | 592 | 151 | 405 | 1,655 |
| 22 | RR rose | LASSO none | 1742.1 | 0.0000 | 206 | 2,041 | 450 | 106 |
| 22 | RR rose | LASSO up | 1570.2 | 0.0000 | 234 | 1,793 | 698 | 78 |
| 22 | RR rose | LASSO down | 1550.3 | 0.0000 | 234 | 1,773 | 718 | 78 |
| 22 | RR rose | LASSO smote | 1667.9 | 0.0000 | 206 | 1,966 | 525 | 106 |
| 22 | RR rose | LASSO rose | 0.0 | 1.0000 | 311 | - | 2,491 | 1 |
| 22 | RR rose | RF none | 1781.5 | 0.0000 | 225 | 2,030 | 461 | 87 |
| 22 | RR rose | RF up | 1802.0 | 0.0000 | 226 | 2,048 | 443 | 86 |
| 22 | RR rose | RF down | 1734.1 | 0.0000 | 234 | 1,958 | 533 | 78 |
| 22 | RR rose | RF smote | 1770.4 | 0.0000 | 219 | 2,035 | 456 | 93 |
| 22 | RR rose | RF rose | 1725.2 | 0.0000 | 188 | 2,071 | 420 | 124 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | RR rose | XGB none | 1803.5 | 0.0000 | 221 | 2,063 | 428 | 91 |
| 22 | RR rose | XGB up | 1735.3 | 0.0000 | 230 | 1,970 | 521 | 82 |
| 22 | RR rose | XGB down | 1710.4 | 0.0000 | 230 | 1,945 | 546 | 82 |
| 22 | RR rose | XGB smote | 1698.6 | 0.0000 | 240 | 1,906 | 585 | 72 |
| 22 | RR rose | XGB rose | 871.8 | 0.0000 | 239 | 1,074 | 1,417 | 73 |
| 22 | RR rose | CART none | 1816.5 | 0.0000 | 187 | 2,166 | 325 | 125 |
| 22 | RR rose | CART up | 1810.7 | 0.0000 | 211 | 2,097 | 394 | 101 |
| 22 | RR rose | CART down | 1712.2 | 0.0000 | 204 | 2,016 | 475 | 108 |
| 22 | RR rose | CART smote | 1753.1 | 0.0000 | 185 | 2,107 | 384 | 127 |
| 22 | RR rose | CART rose | 363.3 | 0.0000 | 273 | 470 | 2,021 | 39 |
| 22 | LASSO none | LASSO up | 134.7 | 0.0000 | 1,959 | 68 | 488 | 288 |
| 22 | LASSO none | LASSO down | 146.5 | 0.0000 | 1,932 | 75 | 481 | 315 |
| 22 | LASSO none | LASSO smote | 15.7 | 0.0001 | 2,035 | 137 | 419 | 212 |
| 22 | LASSO none | LASSO rose | 1743.1 | 0.0000 | 205 | 106 | 450 | 2,042 |
| 22 | LASSO none | RF none | 0.1 | 0.7129 | 2,070 | 185 | 371 | 177 |
| 22 | LASSO none | RF up | 1.6 | 0.2029 | 2,052 | 222 | 334 | 195 |
| 22 | LASSO none | RF down | 8.6 | 0.0035 | 2,049 | 143 | 413 | 198 |
| 22 | LASSO none | RF smote | 0.1 | 0.7467 | 2,078 | 176 | 380 | 169 |
| 22 | LASSO none | RF rose | 0.3 | 0.5842 | 2,051 | 208 | 348 | 196 |
| 22 | LASSO none | XGB none | 3.9 | 0.0472 | 2,101 | 183 | 373 | 146 |
| 22 | LASSO none | XGB up | 4.8 | 0.0289 | 2,002 | 198 | 358 | 245 |
| 22 | LASSO none | XGB down | 11.0 | 0.0009 | 1,981 | 194 | 362 | 266 |
| 22 | LASSO none | XGB smote | 18.4 | 0.0000 | 1,925 | 221 | 335 | 322 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | LASSO none | XGB rose | 693.1 | 0.0000 | 1,152 | 161 | 395 | 1,095 |
| 22 | LASSO none | CART none | 28.4 | 0.0000 | 2,106 | 247 | 309 | 141 |
| 22 | LASSO none | CART up | 10.2 | 0.0014 | 2,101 | 207 | 349 | 146 |
| 22 | LASSO none | CART down | 1.5 | 0.2249 | 2,004 | 216 | 340 | 243 |
| 22 | LASSO none | CART smote | 4.2 | 0.0413 | 2,037 | 255 | 301 | 210 |
| 22 | LASSO none | CART rose | 1287.9 | 0.0000 | 618 | 125 | 431 | 1,629 |
| 22 | LASSO up | LASSO down | 1.0 | 0.3098 | 1,842 | 165 | 611 | 185 |
| 22 | LASSO up | LASSO smote | 62.6 | 0.0000 | 1,934 | 238 | 538 | 93 |
| 22 | LASSO up | LASSO rose | 1571.2 | 0.0000 | 233 | 78 | 698 | 1,794 |
| 22 | LASSO up | RF none | 102.6 | 0.0000 | 1,890 | 365 | 411 | 137 |
| 22 | LASSO up | RF up | 110.6 | 0.0000 | 1,877 | 397 | 379 | 150 |
| 22 | LASSO up | RF down | 61.5 | 0.0000 | 1,891 | 301 | 475 | 136 |
| 22 | LASSO up | RF smote | 115.8 | 0.0000 | 1,920 | 334 | 442 | 107 |
| 22 | LASSO up | RF rose | 90.1 | 0.0000 | 1,847 | 412 | 364 | 180 |
| 22 | LASSO up | XGB none | 133.5 | 0.0000 | 1,910 | 374 | 402 | 117 |
| 22 | LASSO up | XGB up | 57.0 | 0.0000 | 1,854 | 346 | 430 | 173 |
| 22 | LASSO up | XGB down | 38.2 | 0.0000 | 1,818 | 357 | 419 | 209 |
| 22 | LASSO up | XGB smote | 23.2 | 0.0000 | 1,787 | 359 | 417 | 240 |
| 22 | LASSO up | XGB rose | 446.7 | 0.0000 | 1,101 | 212 | 564 | 926 |
| 22 | LASSO up | CART none | 180.9 | 0.0000 | 1,898 | 455 | 321 | 129 |
| 22 | LASSO up | CART up | 145.5 | 0.0000 | 1,898 | 410 | 366 | 129 |
| 22 | LASSO up | CART down | 63.4 | 0.0000 | 1,833 | 387 | 389 | 194 |
| 22 | LASSO up | CART smote | 116.4 | 0.0000 | 1,860 | 432 | 344 | 167 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | LASSO up | CART rose | 1032.7 | 0.0000 | 588 | 155 | 621 | 1,439 |
| 22 | LASSO down | LASSO smote | 61.5 | 0.0000 | 1,871 | 301 | 495 | 136 |
| 22 | LASSO down | LASSO rose | 1551.3 | 0.0000 | 233 | 78 | 718 | 1,774 |
| 22 | LASSO down | RF none | 120.6 | 0.0000 | 1,878 | 377 | 419 | 129 |
| 22 | LASSO down | RF up | 123.5 | 0.0000 | 1,854 | 420 | 376 | 153 |
| 22 | LASSO down | RF down | 73.1 | 0.0000 | 1,868 | 324 | 472 | 139 |
| 22 | LASSO down | RF smote | 119.4 | 0.0000 | 1,877 | 377 | 419 | 130 |
| 22 | LASSO down | RF rose | 113.3 | 0.0000 | 1,855 | 404 | 392 | 152 |
| 22 | LASSO down | XGB none | 144.5 | 0.0000 | 1,882 | 402 | 394 | 125 |
| 22 | LASSO down | XGB up | 65.9 | 0.0000 | 1,824 | 376 | 420 | 183 |
| 22 | LASSO down | XGB down | 47.1 | 0.0000 | 1,795 | 380 | 416 | 212 |
| 22 | LASSO down | XGB smote | 31.7 | 0.0000 | 1,776 | 370 | 426 | 231 |
| 22 | LASSO down | XGB rose | 420.5 | 0.0000 | 1,089 | 224 | 572 | 918 |
| 22 | LASSO down | CART none | 201.7 | 0.0000 | 1,885 | 468 | 328 | 122 |
| 22 | LASSO down | CART up | 160.4 | 0.0000 | 1,877 | 431 | 365 | 130 |
| 22 | LASSO down | CART down | 70.8 | 0.0000 | 1,796 | 424 | 372 | 211 |
| 22 | LASSO down | CART smote | 124.3 | 0.0000 | 1,825 | 467 | 329 | 182 |
| 22 | LASSO down | CART rose | 989.6 | 0.0000 | 569 | 174 | 622 | 1,438 |
| 22 | LASSO smote | LASSO rose | 1668.9 | 0.0000 | 205 | 106 | 525 | 1,967 |
| 22 | LASSO smote | RF none | 15.9 | 0.0001 | 2,002 | 253 | 378 | 170 |
| 22 | LASSO smote | RF up | 22.0 | 0.0000 | 1,991 | 283 | 348 | 181 |
| 22 | LASSO smote | RF down | 0.9 | 0.3385 | 1,985 | 207 | 424 | 187 |
| 22 | LASSO smote | RF smote | 20.1 | 0.0000 | 2,050 | 204 | 427 | 122 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | LASSO smote | RF rose | 16.3 | 0.0001 | 1,988 | 271 | 360 | 184 |
| 22 | LASSO smote | XGB none | 28.7 | 0.0000 | 2,013 | 271 | 360 | 159 |
| 22 | LASSO smote | XGB up | 1.6 | 0.2031 | 1,961 | 239 | 392 | 211 |
| 22 | LASSO smote | XGB down | 0.0 | 0.9284 | 1,926 | 249 | 382 | 246 |
| 22 | LASSO smote | XGB smote | 1.1 | 0.2855 | 1,885 | 261 | 370 | 287 |
| 22 | LASSO smote | XGB rose | 599.0 | 0.0000 | 1,128 | 185 | 446 | 1,044 |
| 22 | LASSO smote | CART none | 68.8 | 0.0000 | 2,027 | 326 | 305 | 145 |
| 22 | LASSO smote | CART up | 40.7 | 0.0000 | 2,016 | 292 | 339 | 156 |
| 22 | LASSO smote | CART down | 4.3 | 0.0378 | 1,940 | 280 | 351 | 232 |
| 22 | LASSO smote | CART smote | 30.3 | 0.0000 | 1,998 | 294 | 337 | 174 |
| 22 | LASSO smote | CART rose | 1172.6 | 0.0000 | 588 | 155 | 476 | 1,584 |
| 22 | LASSO rose | RF none | 1784.1 | 0.0000 | 225 | 2,030 | 462 | 86 |
| 22 | LASSO rose | RF up | 1803.0 | 0.0000 | 225 | 2,049 | 443 | 86 |
| 22 | LASSO rose | RF down | 1735.1 | 0.0000 | 233 | 1,959 | 533 | 78 |
| 22 | LASSO rose | RF smote | 1771.4 | 0.0000 | 218 | 2,036 | 456 | 93 |
| 22 | LASSO rose | RF rose | 1727.8 | 0.0000 | 188 | 2,071 | 421 | 123 |
| 22 | LASSO rose | XGB none | 1804.5 | 0.0000 | 220 | 2,064 | 428 | 91 |
| 22 | LASSO rose | XGB up | 1736.3 | 0.0000 | 229 | 1,971 | 521 | 82 |
| 22 | LASSO rose | XGB down | 1711.4 | 0.0000 | 229 | 1,946 | 546 | 82 |
| 22 | LASSO rose | XGB smote | 1699.6 | 0.0000 | 239 | 1,907 | 585 | 72 |
| 22 | LASSO rose | XGB rose | 874.3 | 0.0000 | 239 | 1,074 | 1,418 | 72 |
| 22 | LASSO rose | CART none | 1817.5 | 0.0000 | 186 | 2,167 | 325 | 125 |
| 22 | LASSO rose | CART up | 1811.7 | 0.0000 | 210 | 2,098 | 394 | 101 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | LASSO rose | CART down | 1713.2 | 0.0000 | 203 | 2,017 | 475 | 108 |
| 22 | LASSO rose | CART smote | 1754.1 | 0.0000 | 184 | 2,108 | 384 | 127 |
| 22 | LASSO rose | CART rose | 365.7 | 0.0000 | 273 | 470 | 2,022 | 38 |
| 22 | RF none | RF up | 1.2 | 0.2724 | 2,130 | 144 | 404 | 125 |
| 22 | RF none | RF down | 14.7 | 0.0001 | 2,093 | 99 | 449 | 162 |
| 22 | RF none | RF smote | 0.0 | 1.0000 | 2,134 | 120 | 428 | 121 |
| 22 | RF none | RF rose | 0.0 | 0.8831 | 2,049 | 210 | 338 | 206 |
| 22 | RF none | XGB none | 3.2 | 0.0725 | 2,148 | 136 | 412 | 107 |
| 22 | RF none | XGB up | 9.5 | 0.0021 | 2,074 | 126 | 422 | 181 |
| 22 | RF none | XGB down | 19.3 | 0.0000 | 2,053 | 122 | 426 | 202 |
| 22 | RF none | XGB smote | 27.4 | 0.0000 | 1,988 | 158 | 390 | 267 |
| 22 | RF none | XGB rose | 688.6 | 0.0000 | 1,141 | 172 | 376 | 1,114 |
| 22 | RF none | CART none | 26.9 | 0.0000 | 2,129 | 224 | 324 | 126 |
| 22 | RF none | CART up | 9.1 | 0.0025 | 2,133 | 175 | 373 | 122 |
| 22 | RF none | CART down | 2.8 | 0.0935 | 2,032 | 188 | 360 | 223 |
| 22 | RF none | CART smote | 2.9 | 0.0872 | 2,052 | 240 | 308 | 203 |
| 22 | RF none | CART rose | 1287.0 | 0.0000 | 612 | 131 | 417 | 1,643 |
| 22 | RF up | RF down | 21.6 | 0.0000 | 2,081 | 111 | 418 | 193 |
| 22 | RF up | RF smote | 1.2 | 0.2758 | 2,112 | 142 | 387 | 162 |
| 22 | RF up | RF rose | 0.4 | 0.5135 | 2,037 | 222 | 307 | 237 |
| 22 | RF up | XGB none | 0.4 | 0.5422 | 2,170 | 114 | 415 | 104 |
| 22 | RF up | XGB up | 24.4 | 0.0000 | 2,128 | 72 | 457 | 146 |
| 22 | RF up | XGB down | 32.8 | 0.0000 | 2,078 | 97 | 432 | 196 |
| 22 | RF up | XGB smote | 40.9 | 0.0000 | 2,013 | 133 | 396 | 261 |
| 22 | RF up | XGB rose | 726.2 | 0.0000 | 1,159 | 154 | 375 | 1,115 |
| 22 | RF up | CART none | 22.3 | 0.0000 | 2,177 | 176 | 353 | 97 |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| 22 | RF up | CART up | 4.2 | 0.0407 | 2,161 | 147 | 382 | 113 |
| 22 | RF up | CART down | 7.8 | 0.0052 | 2,067 | 153 | 376 | 207 |
| 22 | RF up | CART smote | 0.7 | 0.3881 | 2,089 | 203 | 326 | 185 |
| 22 | RF up | CART rose | 1333.8 | 0.0000 | 631 | 112 | 417 | 1,643 |
| 22 | RF down | RF smote | 11.9 | 0.0006 | 2,066 | 188 | 423 | 126 |
| 22 | RF down | RF rose | 10.1 | 0.0015 | 2,009 | 250 | 361 | 183 |
| 22 | RF down | XGB none | 30.9 | 0.0000 | 2,104 | 180 | 431 | 88 |
| 22 | RF down | XGB up | 0.1 | 0.7042 | 2,026 | 174 | 437 | 166 |
| 22 | RF down | XGB down | 0.8 | 0.3763 | 2,020 | 155 | 456 | 172 |
| 22 | RF down | XGB smote | 4.4 | 0.0351 | 1,941 | 205 | 406 | 251 |
| 22 | RF down | XGB rose | 656.1 | 0.0000 | 1,165 | 148 | 463 | 1,027 |
| 22 | RF down | CART none | 62.9 | 0.0000 | 2,069 | 284 | 327 | 123 |
| 22 | RF down | CART up | 39.1 | 0.0000 | 2,081 | 227 | 384 | 111 |
| 22 | RF down | CART down | 1.6 | 0.2061 | 1,978 | 242 | 369 | 214 |
| 22 | RF down | CART smote | 20.4 | 0.0000 | 2,002 | 290 | 321 | 190 |
| 22 | RF down | CART rose | 1251.8 | 0.0000 | 630 | 113 | 498 | 1,562 |
| 22 | RF smote | RF rose | 0.0 | 0.8413 | 2,057 | 202 | 347 | 197 |
| 22 | RF smote | XGB none | 2.9 | 0.0875 | 2,125 | 159 | 390 | 129 |
| 22 | RF smote | XGB up | 8.0 | 0.0046 | 2,052 | 148 | 401 | 202 |
| 22 | RF smote | XGB down | 15.2 | 0.0001 | 2,015 | 160 | 389 | 239 |
| 22 | RF smote | XGB smote | 26.6 | 0.0000 | 1,985 | 161 | 388 | 269 |
| 22 | RF smote | XGB rose | 707.4 | 0.0000 | 1,159 | 154 | 395 | 1,095 |
| 22 | RF smote | CART none | 26.6 | 0.0000 | 2,123 | 230 | 319 | 131 |
| 22 | RF smote | CART up | 8.7 | 0.0032 | 2,119 | 189 | 360 | 135 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | RF smote | CART down | 2.6 | 0.1065 | 2,028 | 192 | 357 | 226 |
| 22 | RF smote | CART smote | 3.5 | 0.0603 | 2,079 | 213 | 336 | 175 |
| 22 | RF smote | CART rose | 1303.7 | 0.0000 | 624 | 119 | 430 | 1,630 |
| 22 | RF rose | XGB none | 1.4 | 0.2319 | 2,070 | 214 | 330 | 189 |
| 22 | RF rose | XGB up | 6.6 | 0.0104 | 1,973 | 227 | 317 | 286 |
| 22 | RF rose | XGB down | 12.9 | 0.0003 | 1,949 | 226 | 318 | 310 |
| 22 | RF rose | XGB smote | 21.1 | 0.0000 | 1,905 | 241 | 303 | 354 |
| 22 | RF rose | XGB rose | 759.4 | 0.0000 | 1,198 | 115 | 429 | 1,061 |
| 22 | RF rose | CART none | 20.4 | 0.0000 | 2,094 | 259 | 285 | 165 |
| 22 | RF rose | CART up | 5.4 | 0.0202 | 2,070 | 238 | 306 | 189 |
| 22 | RF rose | CART down | 2.5 | 0.1137 | 1,951 | 269 | 275 | 308 |
| 22 | RF rose | CART smote | 2.0 | 0.1609 | 2,015 | 277 | 267 | 244 |
| 22 | RF rose | CART rose | 1332.9 | 0.0000 | 640 | 103 | 441 | 1,619 |
| 22 | XGB none | XGB up | 29.4 | 0.0000 | 2,125 | 75 | 444 | 159 |
| 22 | XGB none | XGB down | 42.7 | 0.0000 | 2,093 | 82 | 437 | 191 |
| 22 | XGB none | XGB smote | 43.9 | 0.0000 | 2,001 | 145 | 374 | 283 |
| 22 | XGB none | XGB rose | 757.0 | 0.0000 | 1,177 | 136 | 383 | 1,107 |
| 22 | XGB none | CART none | 14.5 | 0.0001 | 2,159 | 194 | 325 | 125 |
| 22 | XGB none | CART up | 1.8 | 0.1813 | 2,148 | 160 | 359 | 136 |
| 22 | XGB none | CART down | 10.0 | 0.0016 | 2,053 | 167 | 352 | 231 |
| 22 | XGB none | CART smote | 0.1 | 0.7327 | 2,078 | 214 | 305 | 206 |
| 22 | XGB none | CART rose | 1348.3 | 0.0000 | 634 | 109 | 410 | 1,650 |
| 22 | XGB up | XGB down | 2.0 | 0.1551 | 2,045 | 130 | 473 | 155 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | XGB up | XGB smote | 7.8 | 0.0052 | 1,993 | 153 | 450 | 207 |
| 22 | XGB up | XGB rose | 629.5 | 0.0000 | 1,133 | 180 | 423 | 1,067 |
| 22 | XGB up | CART none | 65.5 | 0.0000 | 2,100 | 253 | 350 | 100 |
| 22 | XGB up | CART up | 33.9 | 0.0000 | 2,085 | 223 | 380 | 115 |
| 22 | XGB up | CART down | 1.0 | 0.3272 | 2,022 | 198 | 405 | 178 |
| 22 | XGB up | CART smote | 19.2 | 0.0000 | 2,030 | 262 | 341 | 170 |
| 22 | XGB up | CART rose | 1237.6 | 0.0000 | 615 | 128 | 475 | 1,585 |
| 22 | XGB down | XGB smote | 1.9 | 0.1703 | 1,952 | 194 | 434 | 223 |
| 22 | XGB down | XGB rose | 603.7 | 0.0000 | 1,130 | 183 | 445 | 1,045 |
| 22 | XGB down | CART none | 78.3 | 0.0000 | 2,064 | 289 | 339 | 111 |
| 22 | XGB down | CART up | 47.0 | 0.0000 | 2,056 | 252 | 376 | 119 |
| 22 | XGB down | CART down | 4.3 | 0.0374 | 1,974 | 246 | 382 | 201 |
| 22 | XGB down | CART smote | 27.6 | 0.0000 | 1,990 | 302 | 326 | 185 |
| 22 | XGB down | CART rose | 1193.3 | 0.0000 | 601 | 142 | 486 | 1,574 |
| 22 | XGB smote | XGB rose | 542.1 | 0.0000 | 1,091 | 222 | 435 | 1,055 |
| 22 | XGB smote | CART none | 85.0 | 0.0000 | 2,000 | 353 | 304 | 146 |
| 22 | XGB smote | CART up | 55.9 | 0.0000 | 1,995 | 313 | 344 | 151 |
| 22 | XGB smote | CART down | 10.3 | 0.0013 | 1,924 | 296 | 361 | 222 |
| 22 | XGB smote | CART smote | 40.3 | 0.0000 | 1,958 | 334 | 323 | 188 |
| 22 | XGB smote | CART rose | 1158.3 | 0.0000 | 596 | 147 | 510 | 1,550 |
| 22 | XGB rose | CART none | 759.2 | 0.0000 | 1,122 | 1,231 | 259 | 191 |
| 22 | XGB rose | CART up | 732.4 | 0.0000 | 1,136 | 1,172 | 318 | 177 |
| 22 | XGB rose | CART down | 634.8 | 0.0000 | 1,120 | 1,100 | 390 | 193 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22 | XGB rose | CART smote | 682.7 | 0.0000 | 1,102 | 1,190 | 300 | 211 |
| 22 | XGB rose | CART rose | 507.5 | 0.0000 | 709 | 34 | 1,456 | 604 |
| 22 | CART none | CART up | 7.7 | 0.0055 | 2,205 | 103 | 347 | 148 |
| 22 | CART none | CART down | 51.1 | 0.0000 | 2,116 | 104 | 346 | 237 |
| 22 | CART none | CART smote | 10.3 | 0.0014 | 2,147 | 145 | 305 | 206 |
| 22 | CART none | CART rose | 1352.6 | 0.0000 | 591 | 152 | 298 | 1,762 |
| 22 | CART up | CART down | 18.4 | 0.0000 | 2,058 | 162 | 333 | 250 |
| 22 | CART up | CART smote | 0.6 | 0.4416 | 2,110 | 182 | 313 | 198 |
| 22 | CART up | CART rose | 1327.2 | 0.0000 | 604 | 139 | 356 | 1,704 |
| 22 | CART down | CART smote | 10.7 | 0.0011 | 2,021 | 271 | 312 | 199 |
| 22 | CART down | CART rose | 1244.2 | 0.0000 | 606 | 137 | 446 | 1,614 |
| 22 | CART smote | CART rose | 1275.3 | 0.0000 | 578 | 165 | 346 | 1,714 |
| 33 | RR up | RR down | 0.2 | 0.6831 | 634 | 78 | 460 | 72 |
| 33 | RR up | RR smote | 59.3 | 0.0000 | 634 | 200 | 338 | 72 |
| 33 | RR up | RR rose | 504.6 | 0.0000 | 92 | 39 | 499 | 614 |
| 33 | RR up | LASSO none | 0.1 | 0.8052 | 630 | 72 | 466 | 76 |
| 33 | RR up | LASSO up | 2.3 | 0.1306 | 637 | 89 | 449 | 69 |
| 33 | RR up | LASSO down | 19.6 | 0.0000 | 647 | 119 | 419 | 59 |
| 33 | RR up | LASSO smote | 12.5 | 0.0004 | 592 | 175 | 363 | 114 |
| 33 | RR up | LASSO rose | 502.0 | 0.0000 | 92 | 40 | 498 | 614 |
| 33 | RR up | RF none | 202.7 | 0.0000 | 644 | 353 | 185 | 62 |
| 33 | RR up | RF up | 30.3 | 0.0000 | 568 | 247 | 291 | 138 |
| 33 | RR up | RF down | 35.9 | 0.0000 | 577 | 246 | 292 | 129 |
| 33 | RR up | RF smote | 138.3 | 0.0000 | 636 | 296 | 242 | 70 |
| 33 | RR up | RF rose | 237.2 | 0.0000 | 364 | 40 | 498 | 342 |
| 33 | RR up | XGB none | 84.2 | 0.0000 | 580 | 321 | 217 | 126 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | RR up | XGB up | 85.5 | 0.0000 | 578 | 326 | 212 | 128 |
| 33 | RR up | XGB down | 27.9 | 0.0000 | 541 | 277 | 261 | 165 |
| 33 | RR up | XGB smote | 73.4 | 0.0000 | 572 | 317 | 221 | 134 |
| 33 | RR up | XGB rose | 4.9 | 0.0276 | 529 | 222 | 316 | 177 |
| 33 | RR up | CART none | 110.4 | 0.0000 | 585 | 350 | 188 | 121 |
| 33 | RR up | CART up | 16.3 | 0.0001 | 518 | 276 | 262 | 188 |
| 33 | RR up | CART down | 0.5 | 0.4883 | 431 | 258 | 280 | 275 |
| 33 | RR up | CART smote | 86.6 | 0.0000 | 593 | 304 | 234 | 113 |
| 33 | RR up | CART rose | 503.6 | 0.0000 | 93 | 39 | 499 | 613 |
| 33 | RR down | RR smote | 52.7 | 0.0000 | 634 | 200 | 332 | 78 |
| 33 | RR down | RR rose | 519.9 | 0.0000 | 98 | 33 | 499 | 614 |
| 33 | RR down | LASSO none | 0.6 | 0.4564 | 634 | 68 | 464 | 78 |
| 33 | RR down | LASSO up | 1.1 | 0.2885 | 644 | 82 | 450 | 68 |
| 33 | RR down | LASSO down | 28.7 | 0.0000 | 690 | 76 | 456 | 22 |
| 33 | RR down | LASSO smote | 11.6 | 0.0007 | 614 | 153 | 379 | 98 |
| 33 | RR down | LASSO rose | 517.3 | 0.0000 | 98 | 34 | 498 | 614 |
| 33 | RR down | RF none | 184.6 | 0.0000 | 636 | 361 | 171 | 76 |
| 33 | RR down | RF up | 29.0 | 0.0000 | 584 | 231 | 301 | 128 |
| 33 | RR down | RF down | 33.3 | 0.0000 | 586 | 237 | 295 | 126 |
| 33 | RR down | RF smote | 124.3 | 0.0000 | 629 | 303 | 229 | 83 |
| 33 | RR down | RF rose | 233.3 | 0.0000 | 356 | 48 | 484 | 356 |
| 33 | RR down | XGB none | 83.2 | 0.0000 | 594 | 307 | 225 | 118 |
| 33 | RR down | XGB up | 82.2 | 0.0000 | 586 | 318 | 214 | 126 |
| 33 | RR down | XGB down | 24.9 | 0.0000 | 544 | 274 | 258 | 168 |
| 33 | RR down | XGB smote | 70.9 | 0.0000 | 582 | 307 | 225 | 130 |
| 33 | RR down | XGB rose | 3.3 | 0.0704 | 511 | 240 | 292 | 201 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | RR down | CART none | 104.2 | 0.0000 | 587 | 348 | 184 | 125 |
| 33 | RR down | CART up | 14.6 | 0.0001 | 529 | 265 | 267 | 183 |
| 33 | RR down | CART down | 1.0 | 0.3285 | 447 | 242 | 290 | 265 |
| 33 | RR down | CART smote | 79.7 | 0.0000 | 592 | 305 | 227 | 120 |
| 33 | RR down | CART rose | 517.3 | 0.0000 | 98 | 34 | 498 | 614 |
| 33 | RR smote | RR rose | 616.8 | 0.0000 | 83 | 48 | 362 | 751 |
| 33 | RR smote | LASSO none | 58.4 | 0.0000 | 621 | 81 | 329 | 213 |
| 33 | RR smote | LASSO up | 40.3 | 0.0000 | 638 | 88 | 322 | 196 |
| 33 | RR smote | LASSO down | 18.0 | 0.0000 | 675 | 91 | 319 | 159 |
| 33 | RR smote | LASSO smote | 39.2 | 0.0000 | 745 | 22 | 388 | 89 |
| 33 | RR smote | LASSO rose | 614.3 | 0.0000 | 83 | 49 | 361 | 751 |
| 33 | RR smote | RF none | 72.7 | 0.0000 | 735 | 262 | 148 | 99 |
| 33 | RR smote | RF up | 0.8 | 0.3627 | 629 | 186 | 224 | 205 |
| 33 | RR smote | RF down | 0.3 | 0.6112 | 635 | 188 | 222 | 199 |
| 33 | RR smote | RF smote | 28.5 | 0.0000 | 718 | 214 | 196 | 116 |
| 33 | RR smote | RF rose | 337.1 | 0.0000 | 346 | 58 | 352 | 488 |
| 33 | RR smote | XGB none | 11.3 | 0.0008 | 675 | 226 | 184 | 159 |
| 33 | RR smote | XGB up | 12.0 | 0.0005 | 670 | 234 | 176 | 164 |
| 33 | RR smote | XGB down | 0.5 | 0.4814 | 599 | 219 | 191 | 235 |
| 33 | RR smote | XGB smote | 7.4 | 0.0066 | 664 | 225 | 185 | 170 |
| 33 | RR smote | XGB rose | 16.5 | 0.0000 | 589 | 162 | 248 | 245 |
| 33 | RR smote | CART none | 25.2 | 0.0000 | 686 | 249 | 161 | 148 |
| 33 | RR smote | CART up | 3.2 | 0.0745 | 575 | 219 | 191 | 259 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | RR smote | CART down | 35.4 | 0.0000 | 469 | 220 | 190 | 365 |
| 33 | RR smote | CART smote | 10.6 | 0.0011 | 685 | 212 | 198 | 149 |
| 33 | RR smote | CART rose | 617.3 | 0.0000 | 85 | 47 | 363 | 749 |
| 33 | RR rose | LASSO none | 514.9 | 0.0000 | 101 | 601 | 512 | 30 |
| 33 | RR rose | LASSO up | 533.8 | 0.0000 | 98 | 628 | 485 | 33 |
| 33 | RR rose | LASSO down | 566.9 | 0.0000 | 94 | 672 | 441 | 37 |
| 33 | RR rose | LASSO smote | 558.5 | 0.0000 | 88 | 679 | 434 | 43 |
| 33 | RR rose | LASSO rose | 0.0 | 1.0000 | 130 | 2 | 1,111 | 1 |
| 33 | RR rose | RF none | 760.4 | 0.0000 | 72 | 925 | 188 | 59 |
| 33 | RR rose | RF up | 622.0 | 0.0000 | 98 | 717 | 396 | 33 |
| 33 | RR rose | RF down | 626.6 | 0.0000 | 96 | 727 | 386 | 35 |
| 33 | RR rose | RF smote | 713.5 | 0.0000 | 83 | 849 | 264 | 48 |
| 33 | RR rose | RF rose | 250.8 | 0.0000 | 120 | 284 | 829 | 11 |
| 33 | RR rose | XGB none | 687.6 | 0.0000 | 86 | 815 | 298 | 45 |
| 33 | RR rose | XGB up | 685.8 | 0.0000 | 83 | 821 | 292 | 48 |
| 33 | RR rose | XGB down | 610.4 | 0.0000 | 89 | 729 | 384 | 42 |
| 33 | RR rose | XGB smote | 683.8 | 0.0000 | 91 | 798 | 315 | 40 |
| 33 | RR rose | XGB rose | 535.1 | 0.0000 | 83 | 668 | 445 | 48 |
| 33 | RR rose | CART none | 686.0 | 0.0000 | 63 | 872 | 241 | 68 |
| 33 | RR rose | CART up | 597.9 | 0.0000 | 96 | 698 | 415 | 35 |
| 33 | RR rose | CART down | 500.4 | 0.0000 | 100 | 589 | 524 | 31 |
| 33 | RR rose | CART smote | 665.0 | 0.0000 | 74 | 823 | 290 | 57 |
| 33 | RR rose | CART rose | 0.0 | 1.0000 | 130 | 2 | 1,111 | 1 |
| 33 | LASSO none | LASSO up | 8.3 | 0.0040 | 682 | 44 | 498 | 20 |
| 33 | LASSO none | LASSO down | 30.5 | 0.0000 | 669 | 97 | 445 | 33 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | LASSO none | LASSO smote | 15.3 | 0.0001 | 601 | 166 | 376 | 101 |
| 33 | LASSO none | LASSO rose | 512.3 | 0.0000 | 101 | 31 | 511 | 601 |
| 33 | LASSO none | RF none | 213.4 | 0.0000 | 647 | 350 | 192 | 55 |
| 33 | LASSO none | RF up | 36.8 | 0.0000 | 588 | 227 | 315 | 114 |
| 33 | LASSO none | RF down | 45.1 | 0.0000 | 603 | 220 | 322 | 99 |
| 33 | LASSO none | RF smote | 153.3 | 0.0000 | 646 | 286 | 256 | 56 |
| 33 | LASSO none | RF rose | 250.6 | 0.0000 | 377 | 27 | 515 | 325 |
| 33 | LASSO none | XGB none | 94.9 | 0.0000 | 595 | 306 | 236 | 107 |
| 33 | LASSO none | XGB up | 92.2 | 0.0000 | 584 | 320 | 222 | 118 |
| 33 | LASSO none | XGB down | 32.3 | 0.0000 | 555 | 263 | 279 | 147 |
| 33 | LASSO none | XGB smote | 84.2 | 0.0000 | 590 | 299 | 243 | 112 |
| 33 | LASSO none | XGB rose | 5.6 | 0.0179 | 521 | 230 | 312 | 181 |
| 33 | LASSO none | CART none | 117.8 | 0.0000 | 590 | 345 | 197 | 112 |
| 33 | LASSO none | CART up | 19.3 | 0.0000 | 534 | 260 | 282 | 168 |
| 33 | LASSO none | CART down | 0.3 | 0.5851 | 454 | 235 | 307 | 248 |
| 33 | LASSO none | CART smote | 97.8 | 0.0000 | 607 | 290 | 252 | 95 |
| 33 | LASSO none | CART rose | 512.3 | 0.0000 | 101 | 31 | 511 | 601 |
| 33 | LASSO up | LASSO down | 13.1 | 0.0003 | 688 | 78 | 440 | 38 |
| 33 | LASSO up | LASSO smote | 6.2 | 0.0126 | 618 | 149 | 369 | 108 |
| 33 | LASSO up | LASSO rose | 531.2 | 0.0000 | 98 | 34 | 484 | 628 |
| 33 | LASSO up | RF none | 183.6 | 0.0000 | 663 | 334 | 184 | 63 |
| 33 | LASSO up | RF up | 23.0 | 0.0000 | 602 | 213 | 305 | 124 |
| 33 | LASSO up | RF down | 28.2 | 0.0000 | 611 | 212 | 306 | 115 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | LASSO up | RF smote | 125.1 | 0.0000 | 661 | 271 | 247 | 65 |
| 33 | LASSO up | RF rose | 266.9 | 0.0000 | 372 | 32 | 486 | 354 |
| 33 | LASSO up | XGB none | 77.4 | 0.0000 | 618 | 283 | 235 | 108 |
| 33 | LASSO up | XGB up | 75.3 | 0.0000 | 607 | 297 | 221 | 119 |
| 33 | LASSO up | XGB down | 20.6 | 0.0000 | 571 | 247 | 271 | 155 |
| 33 | LASSO up | XGB smote | 67.1 | 0.0000 | 612 | 277 | 241 | 114 |
| 33 | LASSO up | XGB rose | 1.4 | 0.2388 | 531 | 220 | 298 | 195 |
| 33 | LASSO up | CART none | 97.2 | 0.0000 | 608 | 327 | 191 | 118 |
| 33 | LASSO up | CART up | 10.8 | 0.0010 | 553 | 241 | 277 | 173 |
| 33 | LASSO up | CART down | 2.7 | 0.1014 | 466 | 223 | 295 | 260 |
| 33 | LASSO up | CART smote | 77.9 | 0.0000 | 626 | 271 | 247 | 100 |
| 33 | LASSO up | CART rose | 531.2 | 0.0000 | 98 | 34 | 484 | 628 |
| 33 | LASSO down | LASSO smote | 0.0 | 1.0000 | 651 | 116 | 362 | 115 |
| 33 | LASSO down | LASSO rose | 564.4 | 0.0000 | 94 | 38 | 440 | 672 |
| 33 | LASSO down | RF none | 137.4 | 0.0000 | 689 | 308 | 170 | 77 |
| 33 | LASSO down | RF up | 7.0 | 0.0081 | 626 | 189 | 289 | 140 |
| 33 | LASSO down | RF down | 9.6 | 0.0020 | 631 | 192 | 286 | 135 |
| 33 | LASSO down | RF smote | 79.1 | 0.0000 | 677 | 255 | 223 | 89 |
| 33 | LASSO down | RF rose | 294.8 | 0.0000 | 364 | 40 | 438 | 402 |
| 33 | LASSO down | XGB none | 47.4 | 0.0000 | 644 | 257 | 221 | 122 |
| 33 | LASSO down | XGB up | 47.4 | 0.0000 | 637 | 267 | 211 | 129 |
| 33 | LASSO down | XGB down | 6.3 | 0.0118 | 587 | 231 | 247 | 179 |
| 33 | LASSO down | XGB smote | 38.3 | 0.0000 | 633 | 256 | 222 | 133 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | LASSO down | XGB rose | 0.5 | 0.4930 | 550 | 201 | 277 | 216 |
| 33 | LASSO down | CART none | 67.0 | 0.0000 | 640 | 295 | 183 | 126 |
| 33 | LASSO down | CART up | 1.8 | 0.1770 | 580 | 214 | 264 | 186 |
| 33 | LASSO down | CART down | 11.4 | 0.0007 | 474 | 215 | 263 | 292 |
| 33 | LASSO down | CART smote | 45.3 | 0.0000 | 645 | 252 | 226 | 121 |
| 33 | LASSO down | CART rose | 564.4 | 0.0000 | 94 | 38 | 440 | 672 |
| 33 | LASSO smote | LASSO rose | 556.0 | 0.0000 | 88 | 44 | 433 | 679 |
| 33 | LASSO smote | RF none | 127.3 | 0.0000 | 676 | 321 | 156 | 91 |
| 33 | LASSO smote | RF up | 5.8 | 0.0159 | 601 | 214 | 263 | 166 |
| 33 | LASSO smote | RF down | 7.7 | 0.0056 | 598 | 225 | 252 | 169 |
| 33 | LASSO smote | RF smote | 73.7 | 0.0000 | 667 | 265 | 212 | 100 |
| 33 | LASSO smote | RF rose | 268.0 | 0.0000 | 341 | 63 | 414 | 426 |
| 33 | LASSO smote | XGB none | 42.3 | 0.0000 | 625 | 276 | 201 | 142 |
| 33 | LASSO smote | XGB up | 43.5 | 0.0000 | 623 | 281 | 196 | 144 |
| 33 | LASSO smote | XGB down | 5.4 | 0.0204 | 560 | 258 | 219 | 207 |
| 33 | LASSO smote | XGB smote | 36.6 | 0.0000 | 628 | 261 | 216 | 139 |
| 33 | LASSO smote | XGB rose | 0.5 | 0.4663 | 547 | 204 | 273 | 220 |
| 33 | LASSO smote | CART none | 62.0 | 0.0000 | 626 | 309 | 168 | 141 |
| 33 | LASSO smote | CART up | 1.5 | 0.2279 | 548 | 246 | 231 | 219 |
| 33 | LASSO smote | CART down | 10.7 | 0.0010 | 452 | 237 | 240 | 315 |
| 33 | LASSO smote | CART smote | 45.0 | 0.0000 | 647 | 250 | 227 | 120 |
| 33 | LASSO smote | CART rose | 557.5 | 0.0000 | 89 | 43 | 434 | 678 |
| 33 | LASSO rose | RF none | 757.9 | 0.0000 | 72 | 925 | 187 | 60 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | LASSO rose | RF up | 619.3 | 0.0000 | 98 | 717 | 395 | 34 |
| 33 | LASSO rose | RF down | 624.0 | 0.0000 | 96 | 727 | 385 | 36 |
| 33 | LASSO rose | RF smote | 710.9 | 0.0000 | 83 | 849 | 263 | 49 |
| 33 | LASSO rose | RF rose | 248.1 | 0.0000 | 120 | 284 | 828 | 12 |
| 33 | LASSO rose | XGB none | 685.0 | 0.0000 | 86 | 815 | 297 | 46 |
| 33 | LASSO rose | XGB up | 683.3 | 0.0000 | 83 | 821 | 291 | 49 |
| 33 | LASSO rose | XGB down | 607.8 | 0.0000 | 89 | 729 | 383 | 43 |
| 33 | LASSO rose | XGB smote | 681.2 | 0.0000 | 91 | 798 | 314 | 41 |
| 33 | LASSO rose | XGB rose | 532.7 | 0.0000 | 83 | 668 | 444 | 49 |
| 33 | LASSO rose | CART none | 682.1 | 0.0000 | 62 | 873 | 239 | 70 |
| 33 | LASSO rose | CART up | 595.3 | 0.0000 | 96 | 698 | 414 | 36 |
| 33 | LASSO rose | CART down | 497.8 | 0.0000 | 100 | 589 | 523 | 32 |
| 33 | LASSO rose | CART smote | 664.0 | 0.0000 | 75 | 822 | 290 | 57 |
| 33 | LASSO rose | CART rose | 0.0 | 1.0000 | 130 | 2 | 1,110 | 2 |
| 33 | RF none | RF up | 107.8 | 0.0000 | 754 | 61 | 186 | 243 |
| 33 | RF none | RF down | 95.3 | 0.0000 | 753 | 70 | 177 | 244 |
| 33 | RF none | RF smote | 22.6 | 0.0000 | 874 | 58 | 189 | 123 |
| 33 | RF none | RF rose | 502.8 | 0.0000 | 352 | 52 | 195 | 645 |
| 33 | RF none | XGB none | 31.6 | 0.0000 | 806 | 95 | 152 | 191 |
| 33 | RF none | XGB up | 26.7 | 0.0000 | 792 | 112 | 135 | 205 |
| 33 | RF none | XGB down | 95.1 | 0.0000 | 741 | 77 | 170 | 256 |
| 33 | RF none | XGB smote | 36.9 | 0.0000 | 788 | 101 | 146 | 209 |
| 33 | RF none | XGB rose | 150.8 | 0.0000 | 675 | 76 | 171 | 322 |
| 33 | RF none | CART none | 15.5 | 0.0001 | 846 | 89 | 158 | 151 |
| 33 | RF none | CART up | 117.6 | 0.0000 | 722 | 72 | 175 | 275 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | RF none | CART down | 211.3 | 0.0000 | 620 | 69 | 178 | 377 |
| 33 | RF none | CART smote | 31.2 | 0.0000 | 790 | 107 | 140 | 207 |
| 33 | RF none | CART rose | 759.4 | 0.0000 | 73 | 59 | 188 | 924 |
| 33 | RF up | RF down | 0.2 | 0.6735 | 681 | 142 | 287 | 134 |
| 33 | RF up | RF smote | 51.6 | 0.0000 | 743 | 189 | 240 | 72 |
| 33 | RF up | RF rose | 301.8 | 0.0000 | 331 | 73 | 356 | 484 |
| 33 | RF up | XGB none | 32.5 | 0.0000 | 747 | 154 | 275 | 68 |
| 33 | RF up | XGB up | 35.4 | 0.0000 | 750 | 154 | 275 | 65 |
| 33 | RF up | XGB down | 0.0 | 0.9040 | 679 | 139 | 290 | 136 |
| 33 | RF up | XGB smote | 21.5 | 0.0000 | 728 | 161 | 268 | 87 |
| 33 | RF up | XGB rose | 8.6 | 0.0033 | 553 | 198 | 231 | 262 |
| 33 | RF up | CART none | 45.1 | 0.0000 | 718 | 217 | 212 | 97 |
| 33 | RF up | CART up | 1.3 | 0.2506 | 653 | 141 | 288 | 162 |
| 33 | RF up | CART down | 40.5 | 0.0000 | 559 | 130 | 299 | 256 |
| 33 | RF up | CART smote | 19.9 | 0.0000 | 691 | 206 | 223 | 124 |
| 33 | RF up | CART rose | 619.3 | 0.0000 | 98 | 34 | 395 | 717 |
| 33 | RF down | RF smote | 44.0 | 0.0000 | 745 | 187 | 234 | 78 |
| 33 | RF down | RF rose | 338.0 | 0.0000 | 355 | 49 | 372 | 468 |
| 33 | RF down | XGB none | 18.9 | 0.0000 | 705 | 196 | 225 | 118 |
| 33 | RF down | XGB up | 19.3 | 0.0000 | 698 | 206 | 215 | 125 |
| 33 | RF down | XGB down | 0.0 | 0.8323 | 642 | 176 | 245 | 181 |
| 33 | RF down | XGB smote | 12.6 | 0.0004 | 689 | 200 | 221 | 134 |
| 33 | RF down | XGB rose | 14.0 | 0.0002 | 607 | 144 | 277 | 216 |
| 33 | RF down | CART none | 35.2 | 0.0000 | 704 | 231 | 190 | 119 |
| 33 | RF down | CART up | 2.5 | 0.1147 | 651 | 143 | 278 | 172 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | RF down | CART down | 43.8 | 0.0000 | 554 | 135 | 286 | 269 |
| 33 | RF down | CART smote | 16.3 | 0.0001 | 697 | 200 | 221 | 126 |
| 33 | RF down | CART rose | 624.0 | 0.0000 | 96 | 36 | 385 | 727 |
| 33 | RF smote | RF rose | 445.1 | 0.0000 | 356 | 48 | 264 | 576 |
| 33 | RF smote | XGB none | 3.4 | 0.0664 | 783 | 118 | 194 | 149 |
| 33 | RF smote | XGB up | 2.5 | 0.1129 | 773 | 131 | 181 | 159 |
| 33 | RF smote | XGB down | 37.8 | 0.0000 | 706 | 112 | 200 | 226 |
| 33 | RF smote | XGB smote | 6.6 | 0.0104 | 776 | 113 | 199 | 156 |
| 33 | RF smote | XGB rose | 81.6 | 0.0000 | 643 | 108 | 204 | 289 |
| 33 | RF smote | CART none | 0.0 | 0.8983 | 811 | 124 | 188 | 121 |
| 33 | RF smote | CART up | 54.9 | 0.0000 | 692 | 102 | 210 | 240 |
| 33 | RF smote | CART down | 128.1 | 0.0000 | 582 | 107 | 205 | 350 |
| 33 | RF smote | CART smote | 4.2 | 0.0403 | 777 | 120 | 192 | 155 |
| 33 | RF smote | CART rose | 710.9 | 0.0000 | 83 | 49 | 263 | 849 |
| 33 | RF rose | XGB none | 375.6 | 0.0000 | 325 | 576 | 264 | 79 |
| 33 | RF rose | XGB up | 370.5 | 0.0000 | 318 | 586 | 254 | 86 |
| 33 | RF rose | XGB down | 294.1 | 0.0000 | 321 | 497 | 343 | 83 |
| 33 | RF rose | XGB smote | 379.7 | 0.0000 | 338 | 551 | 289 | 66 |
| 33 | RF rose | XGB rose | 270.2 | 0.0000 | 356 | 395 | 445 | 48 |
| 33 | RF rose | CART none | 391.8 | 0.0000 | 311 | 624 | 216 | 93 |
| 33 | RF rose | CART up | 265.5 | 0.0000 | 314 | 480 | 360 | 90 |
| 33 | RF rose | CART down | 146.4 | 0.0000 | 271 | 418 | 422 | 133 |
| 33 | RF rose | CART smote | 375.3 | 0.0000 | 328 | 569 | 271 | 76 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | RF rose | CART rose | 248.1 | 0.0000 | 120 | 12 | 828 | 284 |
| 33 | XGB none | XGB up | 0.0 | 0.8699 | 828 | 76 | 267 | 73 |
| 33 | XGB none | XGB down | 22.6 | 0.0000 | 711 | 107 | 236 | 190 |
| 33 | XGB none | XGB smote | 0.6 | 0.4412 | 793 | 96 | 247 | 108 |
| 33 | XGB none | XGB rose | 47.0 | 0.0000 | 590 | 161 | 182 | 311 |
| 33 | XGB none | CART none | 4.4 | 0.0354 | 795 | 140 | 203 | 106 |
| 33 | XGB none | CART up | 35.2 | 0.0000 | 688 | 106 | 237 | 213 |
| 33 | XGB none | CART down | 110.7 | 0.0000 | 594 | 95 | 248 | 307 |
| 33 | XGB none | CART smote | 0.0 | 0.8616 | 751 | 146 | 197 | 150 |
| 33 | XGB none | CART rose | 686.6 | 0.0000 | 87 | 45 | 298 | 814 |
| 33 | XGB up | XGB down | 23.8 | 0.0000 | 709 | 109 | 231 | 195 |
| 33 | XGB up | XGB smote | 1.0 | 0.3234 | 796 | 93 | 247 | 108 |
| 33 | XGB up | XGB rose | 47.2 | 0.0000 | 583 | 168 | 172 | 321 |
| 33 | XGB up | CART none | 3.3 | 0.0684 | 784 | 151 | 189 | 120 |
| 33 | XGB up | CART up | 36.7 | 0.0000 | 687 | 107 | 233 | 217 |
| 33 | XGB up | CART down | 105.8 | 0.0000 | 580 | 109 | 231 | 324 |
| 33 | XGB up | CART smote | 0.1 | 0.7259 | 754 | 143 | 197 | 150 |
| 33 | XGB up | CART rose | 684.8 | 0.0000 | 84 | 48 | 292 | 820 |
| 33 | XGB down | XGB smote | 15.3 | 0.0001 | 693 | 196 | 230 | 125 |
| 33 | XGB down | XGB rose | 9.2 | 0.0024 | 548 | 203 | 223 | 270 |
| 33 | XGB down | CART none | 38.6 | 0.0000 | 702 | 233 | 193 | 116 |
| 33 | XGB down | CART up | 1.5 | 0.2267 | 625 | 169 | 257 | 193 |
| 33 | XGB down | CART down | 39.9 | 0.0000 | 548 | 141 | 285 | 270 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 33 | XGB down | CART smote | 15.6 | 0.0001 | 663 | 234 | 192 | 155 |
| 33 | XGB down | CART rose | 609.4 | 0.0000 | 90 | 42 | 384 | 728 |
| 33 | XGB smote | XGB rose | 38.3 | 0.0000 | 575 | 176 | 179 | 314 |
| 33 | XGB smote | CART none | 7.1 | 0.0078 | 769 | 166 | 189 | 120 |
| 33 | XGB smote | CART up | 26.5 | 0.0000 | 675 | 119 | 236 | 214 |
| 33 | XGB smote | CART down | 93.8 | 0.0000 | 578 | 111 | 244 | 311 |
| 33 | XGB smote | CART smote | 0.2 | 0.6831 | 746 | 151 | 204 | 143 |
| 33 | XGB smote | CART rose | 682.8 | 0.0000 | 92 | 40 | 315 | 797 |
| 33 | XGB rose | CART none | 69.5 | 0.0000 | 602 | 333 | 160 | 149 |
| 33 | XGB rose | CART up | 4.2 | 0.0416 | 560 | 234 | 259 | 191 |
| 33 | XGB rose | CART down | 8.1 | 0.0045 | 490 | 199 | 294 | 261 |
| 33 | XGB rose | CART smote | 44.9 | 0.0000 | 590 | 307 | 186 | 161 |
| 33 | XGB rose | CART rose | 532.7 | 0.0000 | 83 | 49 | 444 | 668 |
| 33 | CART none | CART up | 74.5 | 0.0000 | 733 | 61 | 248 | 202 |
| 33 | CART none | CART down | 128.3 | 0.0000 | 578 | 111 | 198 | 357 |
| 33 | CART none | CART smote | 4.4 | 0.0356 | 761 | 136 | 173 | 174 |
| 33 | CART none | CART rose | 683.5 | 0.0000 | 63 | 69 | 240 | 872 |
| 33 | CART up | CART down | 30.5 | 0.0000 | 564 | 125 | 325 | 230 |
| 33 | CART up | CART smote | 30.5 | 0.0000 | 675 | 222 | 228 | 119 |
| 33 | CART up | CART rose | 596.9 | 0.0000 | 97 | 35 | 415 | 697 |
| 33 | CART down | CART smote | 94.8 | 0.0000 | 567 | 330 | 225 | 122 |
| 33 | CART down | CART rose | 499.4 | 0.0000 | 101 | 31 | 524 | 588 |
| 33 | CART smote | CART rose | 664.0 | 0.0000 | 75 | 57 | 290 | 822 |