Pontifícia Universidade Católica
do Rio de Janeiro

Thaís Tuyane de Azevedo

Métodos de Machine Learning aplicados à modelagem preditiva de Cancelamentos de Clientes para Seguros de Vida

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção de grau de Mestre pelo Programa de Pós-Graduação em Macroeconomia e Finanças do Departamento de Economia do Centro de Ciências Sociais da PUC-Rio.

Orientador: Prof. Diogo Abry Guillén

Rio de Janeiro
Junho de 2018

Pontifícia Universidade Católica
do Rio de Janeiro

Thaís Tuyane de Azevedo

Métodos de Machine Learning aplicados à modelagem preditiva de Cancelamentos de Clientes para Seguros de Vida

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção de grau de Mestre pelo Programa de Pós-Graduação em Macroeconomia e Finanças do Departamento de Economia do Centro de Ciências Sociais da PUC-Rio. Aprovado pela Comissão Examinadora abaixo assinada.

Prof. Pedro Carvalho Loureiro de Souza
Presidente
Departamento de Economia – PUC-Rio

Prof. Diogo Abry Guillén
Orientador
Itaú Asset Management

Prof. Flavio Erthal Abdenur
SLQ Soluções Quantitativas

Prof. Augusto Cesar Pinheiro da Silva
Vice-Decano Setorial de Pós-Graduação do
Centro de Ciências Sociais

Rio de Janeiro, 18 de junho de 2018

Thaís Tuyane de Azevedo

Graduou-se em Ciências Atuariais pela Universidade Federal do Rio de Janeiro (UFRJ) em 2011.

Ficha Catalográfica

Azevedo, Thaís Tuyane de

Métodos de machine learning aplicados à modelagem preditiva de cancelamento de clientes para seguros de vida / Thaís Tuyane de Azevedo ; orientador: Diogo Abry Guillén. – 2018.
46 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2018.
Inclui bibliografia

1. Economia – Teses. 2. Propensão a cancelamento. 3. Aprendizado de máquina. 4. Árvore de decisão. 5. Dados desbalanceados. 6. Seguro de vida. I. Guillén, Diogo Abry. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 330

Dedico este trabalho à minha mãe pelo exemplo de vida que é e ao meu marido pelo apoio incondicional em todos os momentos.

Agradecimentos

Agradeço primeiramente a Deus por me dar saúde e força para realizar os meus sonhos e por ter colocado pessoas tão especiais ao meu lado.

Ao meu marido, Hugo, meu profundo agradecimento pelas infinitas leituras e releituras deste trabalho, pelas críticas construtivas e por me manter calma em momentos de desespero, que não foram poucos. Sua confiança em mim me fez chegar ao final desta difícil, porém gratificante etapa. Obrigada por saber me fazer feliz!

À minha mãe, Solange, meu eterno apreço pelo apoio durante toda a minha vida, por sempre acreditar no meu potencial, por todas as noites de sono mal dormidas e por ter assumido com grande maestria as funções de mãe e pai. Obrigada por ser a melhor mãe do mundo!

Ao Prof. e orientador, Diogo Guillén, o meu sincero agradecimento pela orientação valiosa, confiança e por ter acreditado neste trabalho.

Aos meus avós maternos, Geraldo (in memoriam) e Maria (in memoriam), pelo amor incondicional e por fazerem parte da base da minha criação. Vocês fazem muita falta na minha vida.

Finalmente, gostaria de agradecer aos amigos e familiares que entenderam minha ausência em certos momentos para que este trabalho pudesse ser realizado.

Resumo

Azevedo, Thaís Tuyane de; Guillén, Diogo Abry. **Métodos de Machine Learning aplicados à modelagem preditiva de Cancelamentos de Clientes para Seguros de Vida.** Rio de Janeiro, 2018. 46p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

O objetivo deste estudo foi explorar o problema de churn em seguros de vida, no sentido de prever se o cliente irá cancelar o produto nos próximos 6 meses. Atualmente, métodos de machine learning vêm se popularizando para este tipo de análise, tornando-se uma alternativa ao tradicional método de modelagem da probabilidade de cancelamento através da regressão logística. Em geral, um dos desafios encontrados neste tipo de modelagem é que a proporção de clientes que cancelam o serviço é relativamente pequena. Para isso, este estudo recorreu a técnicas de balanceamento para tratar a base naturalmente desbalanceada – técnicas de undersampling, oversampling e diferentes combinações destas duas foram utilizadas e comparadas entre si. As bases foram utilizadas para treinar modelos de Bagging, Random Forest e Boosting, e seus resultados foram comparados entre si e também aos resultados obtidos através do modelo de Regressão Logística. Observamos que a técnica SMOTE-modificado para balanceamento da base, aplicada ao modelo de Bagging, foi a combinação que apresentou melhores resultados dentre as combinações exploradas.

Palavras-chave

Propensão a cancelamento; Aprendizado de máquina; Árvore de decisão; Bagging, Random Forest; Boosting; Dados desbalanceados; Seguro de vida; Under-sampling; Over-sampling; SMOTE.

Abstract

Azevedo, Thaís Tuyane de; Guillén, Diogo Abry (Advisor). **Machine Learning Methods applied to Predictive Models of Churn for Life Insurance**. Rio de Janeiro, 2017. 46p. Dissertação de Mestrado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

The purpose of this study is to explore the churn problem in life insurance, in the sense of predicting if the client will cancel the product in the next 6 months. Currently, machine learning methods are becoming popular in this type of analysis, turning it into an alternative to the traditional method of modeling the probability of cancellation through logistics regression. In general, one of the challenges found in this type of modelling is that the proportion of clients who cancelled the service is relatively small. For this, the study resorted to balancing techniques to treat the naturally unbalanced base – under-sampling and over-sampling techniques and different combinations of these two were used and compared among each other. The bases were used to train models of Bagging, Random Forest and Boosting, and its results were compared among each other and to the results obtained through the Logistics Regression model. We observed that the modified SMOTE technique to balance the base, applied to the Bagging model, was the combination that presented the best results among the explored combinations.

Keywords

Churn prediction; Machine learning; Bagging, Random Forest; Boosting; unbalanced data; Life insurance; Under-sampling; Over-sampling; SMOTE.

Sumário

1. Introdução 11

2. Fonte de Dados 14

3. Metodologia 16
3.1. Regressão Logística 16
3.2. Árvores de decisão 17
3.3. Bagging 18
3.4. Random Forests 19
3.5. Boosting 20
3.6. Técnicas de balanceamento 22
3.7. Métricas de seleção 23

4. Resultados 25
4.1. Balanceamento 25
4.2. Regressão Logística 27
4.3. Bagging 27
4.4. Random Forest 29
4.5. Boosting 31
4.6. Comparação de resultados 33

5. Conclusão 35

6. Referências bibliográficas 37

7. Anexos 39
A. Análise Exploratória 40
B. Resultados do Modelo Linear 43
C. Seleção de Variáveis 44
D. Resultados Specificity 46

Lista de gráficos

Gráfico 1 – Número de árvores por Bagging 28

Gráfico 2 – Importância das Variáveis - Bagging 29

Gráfico 3 – Número de árvores por Random Forest 30

Gráfico 4 – Importância das Variáveis - Random Forest 31

Gráfico 5 – Número de árvores por Boosting 31

Gráfico 6 – Importância das Variáveis - Boosting 32

Gráfico 7 – Propensão a Cancelamento por Parcela 33

Gráfico 8 – Propensão a Cancelamento por Idade no Início de Vigência 33

Gráfico 9 – Emissões e Cancelamentos por Competência 40

Gráfico 10 – Exposição por Idade Atual 41

Gráfico 11 – Exposição por UF 41

Gráfico 12 – Cancelamento por Parcela 42

Gráfico 13 – Cancelamento por Idade Atual 42

Lista de tabelas

Tabela 1 – Descrição das Variáveis 14

Tabela 2 – Matriz de Confusão 23

Tabela 3 – Resultados Sensitivity 34

Tabela 4 – Resultados Accuracy 34

Tabela 5 – Análise Descritiva 40

Tabela 6 – Resultados Linear Model 43

Tabela 7 – Resultado Stepwise Forward 46

Tabela 8 – Resultado Specificity 46

1.
Introdução

Com a expansão do mercado segurador e a facilidade de transição entre empresas, o ambiente competitivo incentiva os investimentos na melhoria do relacionamento com seus clientes, com o objetivo de retê-los através de um serviço mais adequado às suas necessidades.

O termo churn é utilizado quando se analisa a retenção de clientes para representar a situação onde o cliente deixa de fazer negócios com uma empresa ou serviço. Neste estudo, definimos churn como o cancelamento do serviço, o seguro de vida, por parte do cliente. Ele pode indicar que algo está errado na sua missão de reter e satisfazer o cliente, seja porque seu serviço não está atingindo as expectativas necessárias, seja porque o serviço se tornou dispensável, ou porque os valores cobrados estão muito altos.

É natural que haja situações em que o churn esteja relacionado a motivos sobre os quais a companhia não tem gestão. É possível, por exemplo, que o cliente tenha perdido poder aquisitivo e, por este motivo, tenha decidido cancelar o produto. Neste contexto, talvez não haja solução viável para a companhia evitar o churn.

No entanto, existem casos em que o cliente cancela por simplesmente achar que o produto não é mais vantajoso, logo uma ação da companhia poderia evitar este churn. Em outros casos, o cliente pode estar indo para a concorrência a procura de um preço menor, este também é um cenário em que a companhia teria a oportunidade de reter o cliente.

É importante lembrar que a identificação de um potencial churn não implica necessariamente que a empresa deva tomar alguma atitude em busca da retenção deste cliente. A empresa pode concluir que não há o que se fazer, ou mesmo verificar que determinadas retenções seriam mais onerosas do que simplesmente aceitar o cancelamento. No entanto, a previsão de churn é uma ferramenta valiosa que permitirá ao gestor desenvolver análises no sentido de reter com eficiência clientes valiosos para a empresa.

Neste sentido, seria interessante que este gestor conseguisse dizer o custo da ação de retenção, bem como o valor que o cliente representa, de modo que se possa comparar em termos financeiros as alternativas que a companhia tem. A partir disso, uma função objetivo poderia ser utilizada em conjunto com os modelos de previsão para maximizar as decisões de reter ou não os clientes que irão cancelar. Este estudo não desenvolveu uma função objetivo, em contrapartida pretende apresentar as técnicas mais eficazes na previsão de um cancelamento.

A pouca interação entre clientes e empresas é uma característica importante e que dificulta consideravelmente nosso estudo. Ao contrário de empresas que prestam serviços de uso contínuo, e assim conseguem medir com maior facilidade o comportamento de seus clientes, para Seguradoras de Vida o contato restringe-se basicamente ao pagamento mensal. Operadoras de cartões de crédito e de planos de saúde, por exemplo, conseguem analisar os hábitos de utilização de seus clientes, enquanto no caso de seguradoras de vida as interações limitam-se ao pagamento das mensalidades.

A partir da revisão de diversos trabalhos existentes na literatura de estimação de churn (KIM; JUN; LEE, 2014; SU et al., 2011; OWCZARCZUK, 2010; BUCKINX; POEL, 2005; MOZER et al., 2000; EIBEN; KOUDIJS; SLISSER, 1998), assim como da literatura atual de métodos computacionais e estatísticos para reconhecimento de padrões (FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R., 2009), verifica-se como técnicas viáveis passíveis de avaliação:

- Bagging;
- Random Forest (BENOIT;POEL, 2012; CHEN; LIAW; BREIMAN, 2004; IDRIS;RIZWAN;KHAN, 2012);
- Boosting (LEMMENS;GUPTA, 2013).

Outro ponto que merece atenção e que também é objetivo deste estudo corresponde ao tratamento de bases desbalanceadas (JAPKOWICZ, 2000; LING; LI, 1998; DOMINGOS, 1999; CHAWLA; BOWYER; HALL; KEGELMEYER, 2002). A utilização destas técnicas de balanceamento da amostra se faz importante pois a base de dados utilizada neste estudo é desbalanceada, com apenas 12,2%

dos clientes classificados como churners. Nas literaturas analisadas pode-se destacar:

- under-sampling;
- over-sampling;
- combinação de under-sampling com over-sampling;
- SMOTE (Synthetic Minority Over-sampling Technique).

Embora seja possível observar nos papers utilizados como referência tanto comparativos entre modelos de machine learning quanto vantagens no uso de técnicas de balanceamento, raramente se observa a combinação dos dois recursos.

Este estudo emprega técnicas de Bagging, Random Forest e Boosting para prever quais clientes/parcelas emitidas são canceladas dentro de 6 meses. Utilizamos em todos os modelos a base desbalanceada bem como aplicamos técnicas de balanceamento como under-sampling, over-sampling, uma combinação das duas técnicas anteriores e uma proposta de SMOTE modificado para variáveis categóricas. Além disso, comparamos todos os resultados com o clássico modelo de regressão logística, técnica amplamente utilizada para a previsão de churn tanto na literatura quanto pela indústria de seguros.

Este estudo busca contribuir para a literatura acadêmica sobre problema de churn, tema que é pouco abordado em produtos de seguros, especialmente para produtos de seguro de vida. Neste trabalho mostramos a importância e a eficácia da utilização de técnicas de balanceamento em amostras desbalanceadas, bem como o ganho e a aplicabilidade do uso de modelos de machine learning.

Este trabalho está organizado da seguinte forma: no Capítulo 2 são descritas a fonte de dados e as variáveis utilizadas. O Capítulo 3 apresenta a descrição dos modelos e metodologias utilizados no estudo. Os parâmetros empregados e os resultados obtidos em cada modelo são apresentados no Capítulo 4. Por fim, o Capítulo final contém a conclusão do trabalho e apresenta possíveis evoluções do tema abordado para estudos futuros. Os apêndices apresentam a análise exploratória dos dados, os resultados do modelo linear, bem como o resultado da seleção de variáveis para o modelo logit.

2.
Fonte de Dados

Este estudo utiliza como fonte de dados os registros de uma grande seguradora do Brasil no ramo Vida. Foram utilizadas todas as emissões e cancelamentos feitos dentro do período de abril/2015 a outubro/2017. Nosso objetivo é prever a cada emissão de parcela a possibilidade do cliente cancelar ou não o produto dentro de um período de 6 meses. A base de dados utilizada no estudo totaliza 1.056.847 emissões de parcelas, das quais 129.435 foram canceladas dentro de um período de 6 meses a contar da data de emissão da parcela, ou seja 12,2% de churn.

A análise exploratória dos dados é apresentada no apêndice A. A base possui 11 variáveis independentes que são descritas na Tabela 1- Descrição de Variáveis abaixo.

Variável	Classe	Descrição
PARCELA	Discreta	Valor ordinal da parcela emitida desde o início da apólice (emissões mensais)
PREMIO	Contínua	Valor do prêmio emitido
BENEFICIO	Contínua	Valor do Capital Segurado
IDADE_VIG	Discreta	Idade na data da emissão
DIF_EMISSAO	Discreta	Diferença (em meses) entre a emissão e a data de pagamento ou diferença entre emissão e data de corte (caso a parcela não tenha sido paga)
UF	Categórica	Estados Brasileiros
SEXOID	Categórica	Gênero do segurado
ESTADOCIVILID	Categórica	Estado civil do segurado
DS_FORMACOBRANCA	Categórica	Forma de cobrança (boleto, cheque,...)
PAGAMENTO	Dummy	1 – caso tenha efetuado o pagamento; 0 – caso contrário.
MÊS_ING	Categórica	Mês em que o segurado ingressou no produto
CANCELAMENTO	Dummy	1 – caso tenha cancelado dentro de 6 meses após a emissão da parcela; 0 – caso contrário.

Tabela 1- Descrição de Variáveis.

A base completa foi dividida em dois grupos: emissões até outubro/2016 para treinamento dos modelos preditivos e emissões entre novembro/2016 a abril/2017 para o teste, isto é, in-sample e out-of-sample, respectivamente.

Algumas variáveis como estado civil e forma de pagamento podem sofrer alterações ao longo da vigência do contrato, para esses casos foram utilizados os últimos registros da companhia de acordo com a data-base de estudo.

A variável mês de ingresso foi incluída, pois sabemos da existência de programas de incentivos à venda feito aos corretores, desta forma a quantidade de novas vendas tende a possuir um efeito sazonal que pode vir a influenciar os futuros cancelamentos.

Como podemos perceber, a base utilizada contém um grande número de variáveis categóricas que prejudicam a utilização de metodologias comuns, dado que estas costumam valer-se de variáveis numéricas.

Nossa variável resposta é o Cancelamento, tendo em mente que erros do tipo II (assumir que o cliente irá cancelar quando na verdade ele se mantém ativo) são menos custosos para a empresa do que cometer erros do tipo I (assumir que o cliente não irá cancelar, quando ele cancela).

A partir da base de treino foram geradas outras 4 bases de treino, uma para cada tipo de balanceamento, sendo elas over-sampling, under-sampling, both e SMOTE-m. A base de treino possui 91.867 cancelados em um total amostras de 794.887.

Toda a programação foi gerada no software RStudio devido sua praticidade de aplicação e ampla utilização.

3.
Metodologia

O objetivo neste capítulo é descrever as técnicas e metodologias utilizadas no estudo. A primeira técnica abordada é a regressão logística, que é o recurso mais popular para modelagem de churn. Em seguida são apresentados os métodos alternativos: Bagging, Random Forest e Boosting. Por fim, são detalhadas as técnicas de amostragem e as métricas de seleção utilizadas neste estudo.

Em todos os modelos partimos de uma mesma amostra de treino para definir os coeficientes do modelo, todas as previsões foram analisadas fora da amostra (out-of-sample). As previsões serão apresentadas através de matrizes de confusão, onde nosso objetivo minimizar o erro tipo 1, considerando que o prejuízo seria maior ao deixar de identificar um cliente que faria o cancelamento do que ao prever equivocadamente um cancelamento.

3.1.
Regressão Logística

A regressão logística é amplamente utilizada como método de modelagem estatística de dados binários por ser de fácil aplicação e pela facilidade de interpretação dos parâmetros utilizados. Dentre suas vantagens estão a simplicidade para lidar com variáveis categóricas e a definição de um pequeno número de suposições.

O uso de regressão linear simples poderia gerar probabilidades menores que zero ($\mathrm{p}(churn|X) < 0$) ou maiores que 1 ($\mathrm{p}(churn|X) > 1$). Para evitar este problema devemos modelar $\mathrm{p}(churn|X)$ usando uma função que fornece apenas saídas entre 0 e 1 para todos os valores de x, tal qual a função logística descrita abaixo:

$$p(churn|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Para estimar o vetor de parâmetros $\beta = (\beta_0, \beta_1)$ é utilizado o método de máxima verossimilhança.

Após algumas manipulações chegamos à função logit conforme abaixo:

$$\log\left(\frac{p(churn|X)}{1-p(churn|X)}\right) = \beta_0 + \beta_1 X$$

O modelo logístico é bastante sensível à colinearidade e pode apresentar coeficientes superestimados para variáveis independentes com alta correlação. Outra desvantagem consiste na obrigatoriedade da utilização de métodos numéricos para obtenção da solução de máxima verossimilhança, dado que métodos numéricos nem sempre convergem.

3.2. Árvores de decisão

Os métodos baseados em árvore de decisão dividem o espaço amostral em um conjunto de retângulos de alta dimensão (ou caixas) e, em seguida, utilizam um modelo simples em cada um.

Primeiro dividimos o espaço amostral em duas regiões (R_1 e R_2) e modelamos a resposta pela média de Y (número de cancelamentos) em cada região. Escolhemos a variável *j* e o ponto de divisão *s* para obter o melhor ajuste que minimiza o seguinte erro:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

Então, uma ou ambas as regiões são divididas em mais duas regiões. Este processo continua até que seja aplicada alguma regra de parada.

Seja,

$$\hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i\in R_m} I(y_i = k)$$

A proporção de observações da classe *k* no nó *m*, onde o nó *m* representa a região R_m com N_m observações. Desta forma, classificamos as observações no nó

m para a classe majoritária referente a este nó, dada por $k(m) = argmax_k \hat{p}_{mk}$. Para maiores informações veja Friedman, Hastie e Tibshirani (2009).

O maior problema das árvores de decisão é sua alta variabilidade dos resultados. Apesar de ser de fácil compreensão e de necessitar de pouca manipulação de dados, as árvores de decisão podem ser muito grandes e uma pequena mudança na base pode gerar ramificações completamente diferentes na árvore.

3.3. Bagging

O método de bootstrap foi introduzido por Efron (1979) e é usado frequentemente quando é moroso, ou até mesmo impraticável, calcular diretamente o desvio padrão de uma quantidade de interesse. Nele as observações são escolhidas de forma aleatória e as estimativas recalculadas.

Conforme descrito no item anterior, as árvores de decisão sofrem de alta variância. Isso significa que, se dividirmos os dados de treinamento em duas partes aleatoriamente, e definir uma árvore de decisão para ambas as metades, os resultados que obtemos podem ser bastante diferentes. Em contraste, um procedimento com baixa variação produzirá resultados semelhantes se aplicado repetidamente em conjuntos de dados distintos.

Bootstrap Aggregation, ou Bagging, é um procedimento para reduzir a variância de um método de aprendizagem estatística onde, dada uma amostra de treinamento, construímos B amostras bootstrap que são utilizadas para construir uma coleção de árvores de decisão. Cada árvore de decisão terá alta variância, porém baixo viés.

Em outras palavras, nós podemos produzir B amostras de bootstrap a partir de uma base de treino, construir um modelo de previsão separado para cada árvore e tirar a média dos resultados $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ para obter um modelo de aprendizagem estatística de baixa variação, dado por

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{b}(x).$$

Uma grande vantagem do bagging é que, mesmo para um B grande, não corremos o risco de overfitting. Este modelo aumenta a precisão sobre a previsão usando apenas uma única árvore, no entanto, conforme descrito em Friedman, Hastie e Tibshirani (2009), o modelo resultante pode ser obscuro e complexo de se interpretar.

3.4. Random Forests

O modelo Random Forests fornece uma melhoria em relação ao bagging por meio de um pequeno ajuste aleatório. Assim como no bagging, nós construímos um número de árvores de decisão a partir de amostras bootstrap. Mas a cada árvore é selecionada uma amostra aleatória de m possíveis preditores de p preditores totais.

De acordo com Friedman, Hastie e Tibshirani (2009), tipicamente escolhemos m, tal que

- $m \approx \sqrt{p}$, quando o algoritmo é usado para classificação;
- $m \approx {}^{p}/_{3}$, quando usado para regressão.

Além disso, eles recomendam um número mínimo de nós igual a 1 ou 5, para classificação ou regressão, respectivamente.

Logo, construir uma Random Forest utilizando $m = p$ equivale a simplesmente utilizar o modelo bagging.

Em outras palavras, na construção de um modelo Random Forest, em cada divisão da árvore, o algoritmo nem sequer permite considerar a maioria dos preditores disponíveis.

Suponha que haja um preditor muito forte no conjunto de dados, juntamente com um número de outros preditores moderadamente fortes. Neste

caso, a maioria ou todas as árvores usarão este forte preditor no primeiro nó, o que levará a árvores bastante similares entre si. Consequentemente, as previsões serão altamente correlacionadas. Infelizmente, a média de muitas variáveis altamente correlacionadas não leva a uma redução de variância tão grande quanto a média de muitas variáveis não correlacionadas. Em particular, isso significa que o bagging não levará a uma redução substancial da variância.

O modelo de Random Forest supera este problema forçando cada divisão a considerar apenas um subconjunto de preditores.

Uma característica do modelo Random Forest é o uso de amostras out-of-bag. Eis a citação de Friedman, Hastie e Tibshirani:

> *"For each observation $z_i = (x_i, y_i)$, construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which z_i did not appear."*

Ao contrário de muitos outros estimadores não-lineares, random forest pode ser construído em uma sequência, sendo realizada a validação cruzada ao longo do caminho. Uma vez que o erro out-of-bag for estabilizado, o treinamento pode ser encerrado.

3.5. Boosting

Assim como o bagging, o boosting é aplicado a muitos métodos de aprendizagem estatística para regressão ou classificação. Aqui restringimos nossa discussão ao contexto de árvores de decisão. Boosting funciona de forma semelhante ao bagging, exceto que as árvores crescem sequencialmente: cada árvore cresce usando informações das árvores passadas. Este método não envolve amostragem de boostrap, em vez disso cada árvore é ajustada em uma versão modificada da base original.

Algoritmo

1. Defina $\hat{f}(x) = 0$ e $r_i = y_i$ na base de treino
2. Para $b = 1, 2, \ldots, B$, repita:

a) Defina a árvore $\hat{f}^b$ com d divisões (d +1 nós terminais) para treinar a base (X, r).

b) Atualiza $\hat{f}$ adicionando uma versão encolhida da nova árvore:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

c) Atualize os resíduos,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

3. Gere o modelo Boosting,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

(James, Witten, Hastie e Tibshirani, 2013)

Ao contrário de incluir uma única grande árvore de decisão, o método de boosting utiliza um modelo de aprendizado lento. Dado o modelo atual, nós estabelecemos árvores de decisão para os resíduos do modelo. Ou seja, ajustamos uma árvore usando os resíduos atuais, em vez do resultado Y como resposta. Em seguida, adicionamos esta nova árvore de decisão na função para atualizar os resíduos. Cada uma dessas árvores pode ser bastante pequena, com poucos nós terminais, determinados pelo parâmetro d no algoritmo. Com regressões de pequenas árvores sobre os resíduos, aperfeiçoamos lentamente $\hat{f}$ em áreas onde o modelo não performa bem. O parâmetro de encolhimento λ retarda o processo, permitindo que mais árvores diferentes sejam criadas para atacar os resíduos.

Em geral, as abordagens de aprendizagem estatística que aprendem lentamente tendem a funcionar bem.

3.6. Técnicas de balanceamento

Conforme apresentado anteriormente, nossa base de dados é desbalanceada, ou seja, possui muito mais casos de uma classe do que de outra, portanto possui uma classe rara (cancelados). Como nosso objetivo de estudo é identificar os casos raros, é aconselhável que seja feito um balanceamento prévio da base de dados.

Neste estudo utilizaremos os métodos de under-sampling, over-sampling e SMOTE-modificado.

As técnicas de balanceamento visam mudar a distribuição dos dados de treinamento, de modo a aumentar a acurácia de seus modelos. Isto é alcançado com a eliminação de casos da classe majoritária (denominado na literatura como under-sampling) ou replicação de casos da classe minoritária (denominado over-sampling). Esta última técnica não aumenta a quantidade de informação, porém faz com que a classe menor possua mais peso na função de perda. Também é possível combinar as duas metodologias de tal forma que é feita a eliminação de alguns casos da classe majoritária e replicação de casos da classe minoritária (denominado neste estudo como both).

Weiss (2004) subdivide-os em duas vertentes: métodos básicos e métodos avançados de amostragem. Os métodos básicos de amostragem são métodos que não utilizam heurística na eliminação e na replicação de casos, ou seja, são métodos que visam balancear a distribuição de classes de forma aleatória.

Neste estudo também propusemos a utilização do método avançado denominado SMOTE. Chawla, Bowyer, Hall e Kegelmeyer (2002) apresentaram um modelo de amostragem no qual a classe minoritária é superestimada criando exemplos "sintéticos", em vez de uma amostragem aleatória com substituição. Para evitar que os exemplos sejam replicados, é feita uma interpolação entre dois valores próximos da classe minoritária, gerando assim valores numa vizinhança em torno desta classe. Conforme descrito anteriormente, nossa base de estudos possui algumas variáveis categóricas. É comum descartarem a utilização do modelo SMOTE nestes casos, porém sugerimos aqui uma adaptação do modelo

original utilizando uma distribuição multinominal para estas variáveis. No presente estudo denominaremos este modelo de SMOTE-modificado (SMOTE-m).

Ling and Li (1998) constataram que fazer over-sampling ou under-sampling gerava ganhos efetivos, no entanto a combinação dos dois não gerou um improvement significante no resultado. Domingos (1999) apontou que utilizar over-sampling da classe minoritária aumenta o risco de overfit, portanto aplicar under-sampling é mais indicado e mais utilizado. Por fim, Japkowicz (2000) também notou que fazer over-sampling ou under-sampling gerava ganhos efetivos, no entanto utilizar técnicas sofisticadas de amostragem não geravam nenhuma vantagem considerável sobre os métodos básicos.

3.7. Métricas de seleção

Para cada teste out-of-bag será gerada uma matriz de confusão semelhante a apresentada na Tabela 2 - Matriz de Confusão, onde as linhas representam os valores previstos pelo modelo e as colunas os valores reais.

		VALOR REAL	
		POSITIVO	**NEGATIVO**
VALOR PREVISTO	**POSITIVO**	TP (Verdadeiro Positivo)	FP (Falso Positivo)
	NEGATIVO	FN (Falso Negativo)	TN (Verdadeiro Negativo)

Tabela 2 - Matriz de Confusão.

As categorias usadas na análise são: Verdadeiro Positivo (TP), Falso Positivo (FP), Verdadeiro Negativo (TN) e Falso Negativo (FN). Essas quantidades são utilizadas para calcular as métricas de performance usadas no trabalho. Neste estudo utilizaremos métricas básicas para avaliar o desempenho de modelos preditivos, tais como:

$$Verdadeiro\ Positivo\ ou\ Sensitivity\ (Acc^{+}) = \frac{TP}{TP + FN}$$

$$Verdadeiro\ Negativo\ ou\ Specificity\ (Acc^{-}) = \frac{TN}{TN + FP}$$

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + FP + TN + FN}$$

Em qualquer modelo de classificação, existe uma penalidade entre aumentar Acc^{+} ou Acc^{-}, ou seja, quando buscamos aumentar o valor de uma dessas métricas teremos como consequência a queda da outra. Neste estudo buscaremos um equilíbrio entre ambas, no entanto nosso foco principal será buscar o maior Acc^{+}, dado que cometer erros deste tipo, supor que o cliente continua ativo quando cancela, gera perda de receita futura.

Breiman (2004), propôs duas medidas de importância de atributos para utilização com árvores de decisão, neste estudo utilizaremos a medida denominada Importância de Gini (IG) que é baseada na soma dos decréscimos do índice de Gini em todos os nós rotulados pelo atributo.

4.
Resultados

Este capítulo é subdividido em seções onde apresentamos as técnicas de balanceamento e os resultados de cada modelo, bem como os parâmetros utilizados para cada teste. Por fim, apresentamos um consolidado de todos os resultados encontrados no estudo. Vale ressaltar que todos os testes apresentados neste estudo foram aplicados à amostra out-of-sample.

4.1.
Balanceamento

Descrevemos aqui, os parâmetros e a forma como utilizamos cada técnica de balanceamento já apresentadas anteriormente.

Over-sampling

Foi feita uma amostragem com reposição da base de cancelados até que a base completa de treino atingisse 1 milhão de registros. O resultado foi uma base com 296.980 observações da classe minoritária e 703.020 observações da classe majoritária. Desta forma aumentamos a representatividade de churners de 12,2% para 29,7%.

Optamos por não replicar a classe minoritária até igualar ao nº de amostra da classe majoritária para evitar problemas de overfiting, conforme descrito em Domingos (1999). Sendo assim o desbalanceamento nesta base é suavizado, mas não eliminado.

Under-sampling

Foi feita uma amostra sem reposição da classe majoritária até que a base completa possuísse 200mil registros. Sendo assim reduzimos a classe majoritária para 108.133 amostras e, consequentemente, os churners passaram a corresponder a 45,9% desta nova base.

Nesse exemplo optamos por eliminar o desbalanceamento mesmo que isto gerasse uma perda de informação da classe de não-churners. Nos resultados

descritos a seguir verificamos que esta perda não foi relevante, ou seja, o ganho com o balanceamento foi muito maior do que uma possível ausência de informação.

Both

Criamos uma nova amostra com 400mil observações, onde a probabilidade de seleção de cada classe foi de 50%, consideramos amostragem aleatória com reposição. O resultado foi uma base com 200.166 churners e 199.834 não-churners.

Como esta técnica também possui o objetivo de aumentar a classe minoritária, optamos por utilizar uma amostra que fosse o dobro da amostra de under-sampling, que não considera este aumento.

SMOTE-m

Esta técnica contém 3 diferentes bases, onde a primeira corresponde a um under-sampling da classe majoritária para 180mil observações. A segunda equivale a inclusão de todos os registros da classe minoritária (91.867 churners) contidas na base original. A terceira trata-se da geração de dados sintéticos da classe minoritária de tal forma que a base completa possuísse o mesmo número de observações de churners e não-churners.

Para a geração dos dados sintéticos primeiro selecionamos aleatoriamente um dado já existente, depois simulamos um valor na vizinha da amostra selecionada para as variáveis quantitativas, semelhante ao descrito em Chawla, Bowyer, Hall e Kegelmeyer (2002). Já para as variáveis qualitativas propusemos um tratamento diferenciado, simulamos uma multinomial utilizando como probabilidades a frequência de cada item.

Esta proposta de técnica de balanceamento enriqueceu consideravelmente o nosso teste, pois com ela tivemos resultados bastante positivos nos modelos que serão apresentados a seguir. Vale ressaltar que entre os papers utilizados como referência, embora seja possível observar tanto a consideração de variáveis qualitativas quanto a utilização de técnicas de geração de dados sintéticos, não se

observa a combinação dos dois recursos. A proposta para este trabalho é fazer uso desta combinação.

4.2. Regressão Logística

Como apresentado em outros estudos, tal como Nie et al. (2011), a inclusão ou exclusão de uma variável gera impacto significativo para o modelo. Neste estudo utilizamos a técnica stepwise para definir a combinação de variáveis consideradas nos modelos de regressão logística. A importância de cada variável foi definida a partir do critério de AIC.

Por ser de fácil execução este modelo é comumente utilizado em seguradoras para definir a probabilidade de churn.

O resultado do teste indicou que nenhuma variável deveria ser excluída, pois o menor AIC encontrado foi o que utilizou todas as disponíveis. O resultado da seleção de variáveis pode ser visto no apêndice 0.

Ao analisar apenas os resultados deste modelo, já é possível verificar ganho significativo ao utilizar bases balanceadas. Todas as quatro bases balanceadas geraram resultados superiores à base PADRAO, com destaque para a base BOTH que obteve um Acc^{+} de 64,86%, representando um aumento de aproximadamente 23% em relação à base PADRAO.

Este resultado já indica que devemos, sempre que possível, evitar o uso de bases desbalanceadas.

4.3. Bagging

Um dos parâmetros que influencia no modelo de bagging é a quantidade de árvores selecionadas para a geração de cada modelo. Inicialmente fizemos testes com a base de under-sampling gerando resultados com o nº de árvores variando entre 10 e 500. Pelo gráfico abaixo, onde o eixo horizontal representa o número de árvores utilizadas e o eixo vertical corresponde à métrica Acc^{+}, podemos verificar que os testes de bagging convergem muito rapidamente, por

este motivo decidimos utilizar uma amostra de 100 árvores para gerar os modelos de bagging.

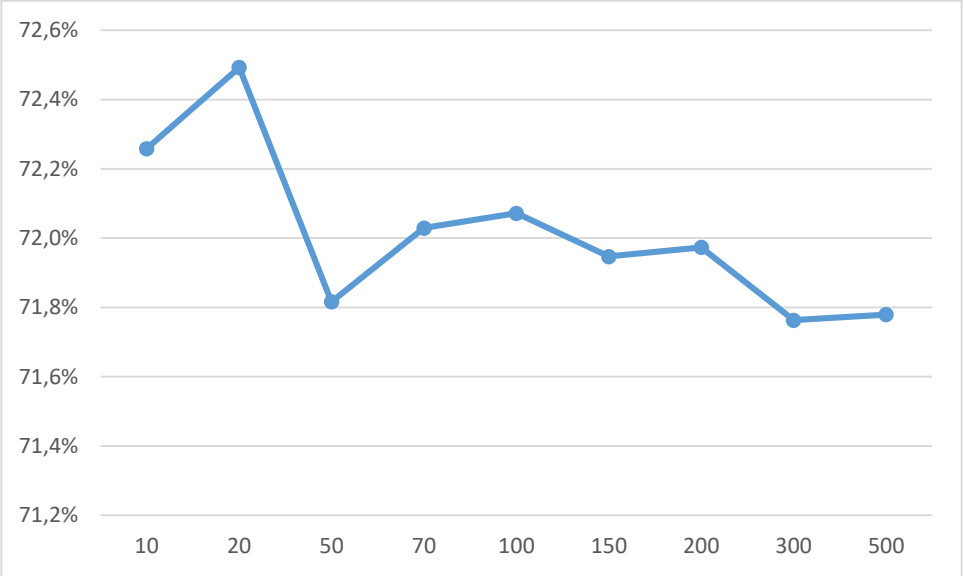

Gráfico 1 – Número de árvores por Bagging.

Outro parâmetro que deve ser definido é o número total de possíveis preditores que serão considerados em cada árvore. Para o modelo bagging por padrão se utiliza o número total de preditores disponíveis, $m = p$, conforme descrito em Tibshirani. Neste caso, para nossa base foi utilizado $m = 11$.

Verificamos que o modelo gerado a partir da base SMOTE-m foi o que produziu melhor resultado, tendo elevado o Acc^+ de 53,60% (na base PADRAO) para 73,52%, o que representa um crescimento de aproximadamente 37%. Isto mostra que utilizar dados sintéticos aplicando nossa sugestão de simulação para variáveis categóricas, pode gerar bastante valor para um modelo preditivo.

Pelos resultados também pudemos verificar quais variáveis são mais relevantes para este modelo. Como o modelo gerado a partir da base SMOTE foi o que retornou maior Acc^+, apresentamos abaixo a relevância das variáveis consideradas neste modelo tomando por base a Importância de Gini. Como já esperado, a variável pagamento é a que possui maior importância na definição de possíveis churners, pois o primeiro indício de um possível cancelamento ocorre quando o cliente se torna inadimplente. Porém também possuímos outras variáveis importantes como o valor do prêmio pago e a parcela emitida, onde

clientes novos (parcelas iniciais) tendem a cancelar muito mais do que clientes antigos.

Gráfico 2 – Importância das Variáveis – Bagging.

4.4. Random Forest

Assim como para o modelo bagging, fizemos testes com a base de under-sampling gerando resultados com o nº de árvores variando entre 10 e 500. O número de árvores selecionados para nossos testes foi 300.

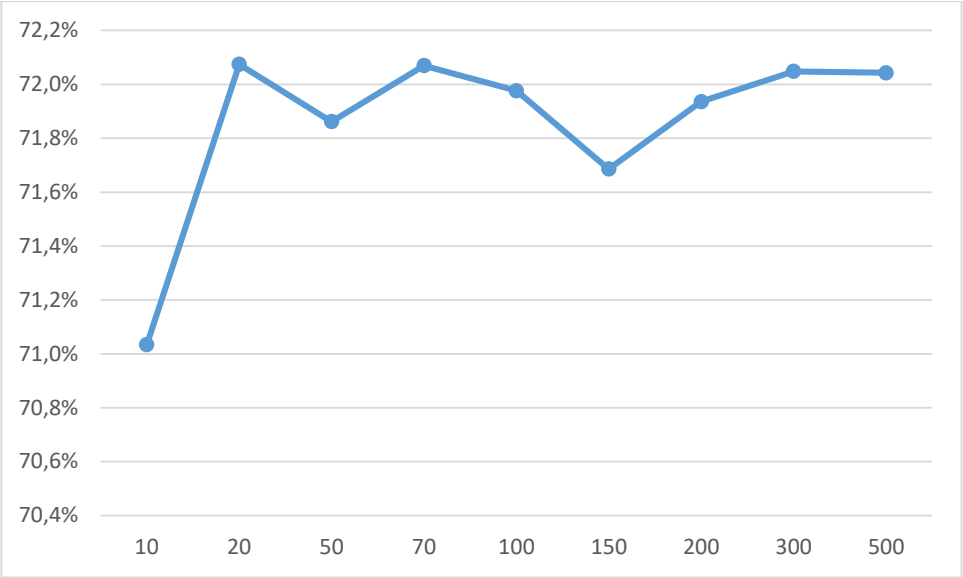

Gráfico 3 – Número de árvores por Random Forest.

Com relação ao o número total de possíveis preditores considerados em cada árvore, seguimos o critério descrito no capítulo de Metodologias, definido $m \approx \sqrt{p}$. Partindo da nossa base de dados temos que $m = 4$.

Todas as bases balanceadas geraram um resultado melhor do que a base PADRAO, com destaque para o modelo gerado a partir da base UNDER que aumentou o Acc^+ de 46,73% para 72,05% (crescimento de 54% em relação à base PADRAO).

Como base gerada por under-sampling obteve Acc^+ superior às outras, apresentamos o gráfico de Importância de Gini baseado nos valores gerados por esta base. Podemos verificar uma grande diferença entre os modelos bagging e random forest com relação à distribuição da importância das variáveis. Enquanto no modelo bagging a variável pagamento é muito superior às outras, no modelo random forest esta importância é amenizada e outras variáveis também passam a ter maior influência, tais como dif_emissao e prêmio. A variável pagamento fica menos destoante no modelo random forest porque estamos limitando o número de possíveis variáveis em cada árvore, ou seja, temos várias árvores que não estão utilizando esta variável.

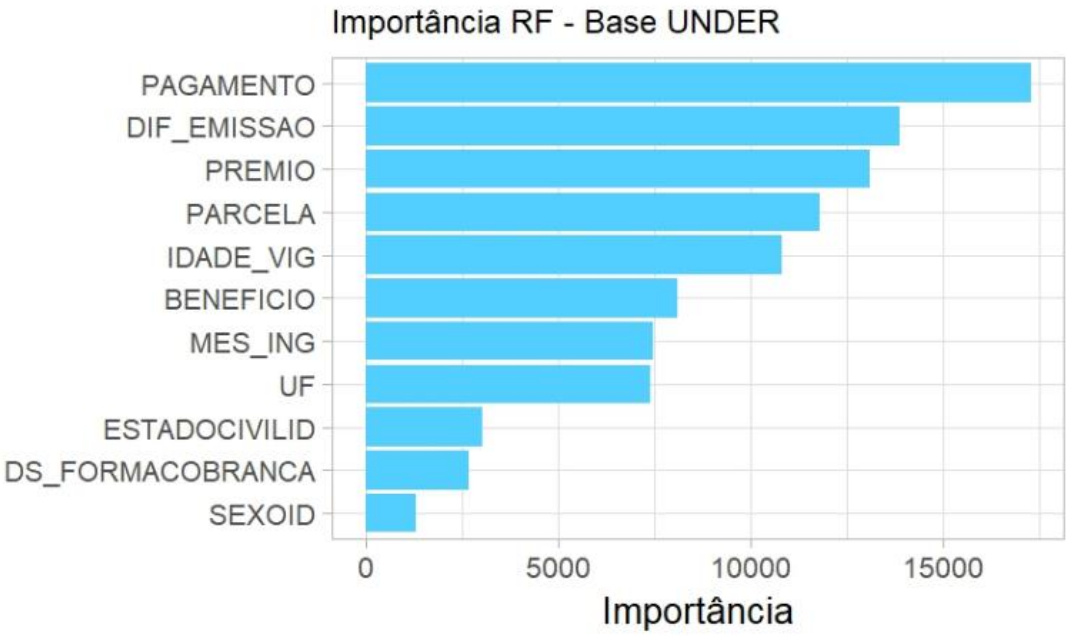

Gráfico 4 – Importância das Variáveis - Random Forest.

4.5. Boosting

Assim como nos modelos anteriores, usamos a base de under-sampling para simular o tamanho ótimo para utilizar como número de árvores do modelo. Por sabermos que o modelo de Boosting trabalha com aprendizado lento, nosso teste para definir o número de árvores simulou valores entre 100 e 5.000, onde o tamanho escolhido foi 1.000.

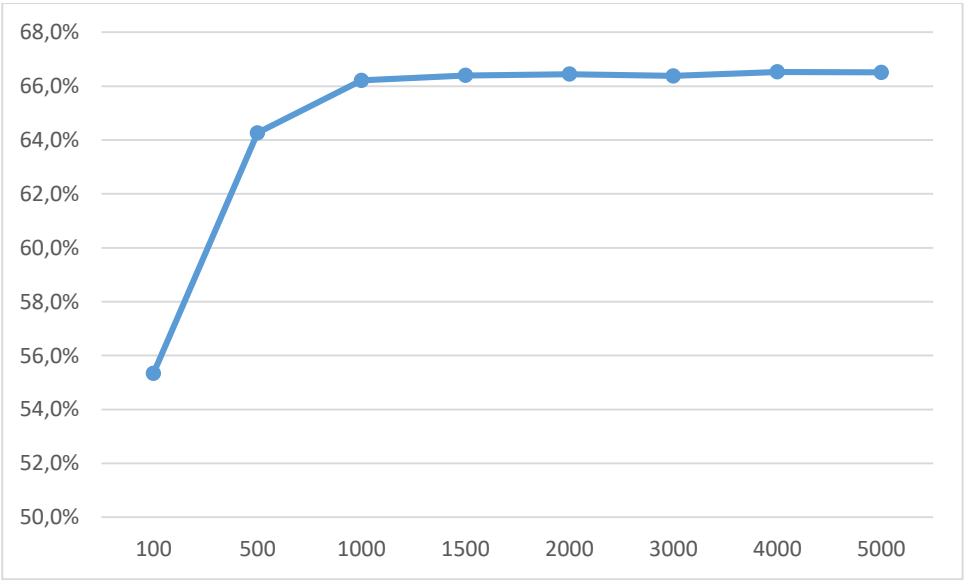

Gráfico 5 – Número de árvores por Boosting.

Semelhante ao Random Forest, utilizamos o número de interações igual a 4 e, por padrão, utilizamos shrinkage igual a 0,01.

Assim como em todos os modelos anteriores, as bases balanceadas apresentaram resultados melhores do que a base desbalanceada. Aqui devemos destacar a base BOTH que gerou um Acc^{+} igual a 72,64%, correspondente a um ganho de 41% em relação à base desbalanceada.

A seguir são presentados com mais detalhes os resultados obtidos pelo modelo de boosting a partir da base BOTH. Neste modelo vemos que a variável pagamento volta a aumentar seu grau de importância, em contrapartida este modelo reduz a importância da variável dif_emissao e aumenta a representatividade do item parcela, tornando este o segundo mais importante.

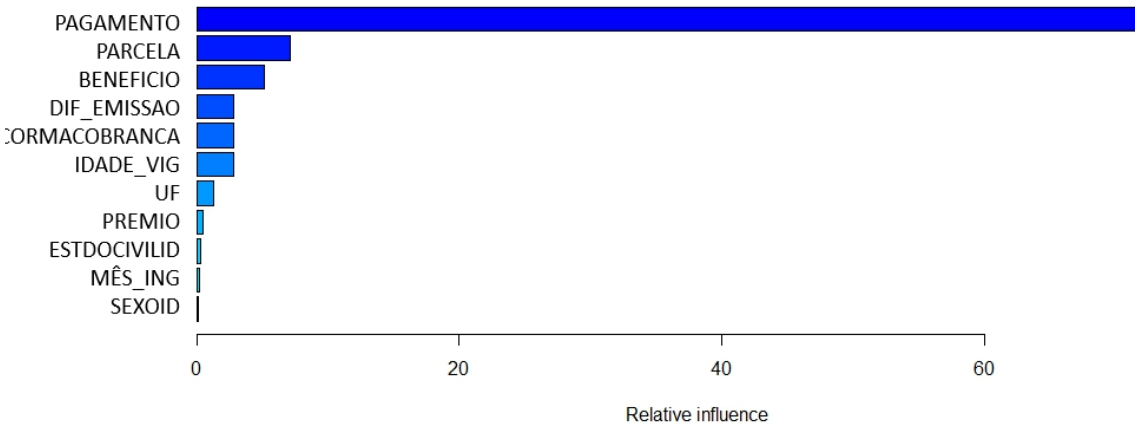

Gráfico 6 – Importância das Variáveis – Boosting.

Enquanto os outros modelos priorizaram o item prêmio, o modelo de boosting priorizou a variável benefício. Dado que ambas possuem certo grau de dependência (benefícios maiores necessitam de prêmios maiores), é natural que ao considerar uma como importante a outra perca poder.

Corroborando a análise exploratória dos dados exposta nas referências, no subitem **Erro! Fonte de referência não encontrada.**, verificamos que de fato inscrições recentes possuem uma maior propensão ao cancelamento, que talvez seja gerada por uma maior insegurança e/ou desconfiança contida nos novos

clientes. O mesmo também pode ser visto com relação à idade do segurado, segurados jovens tendem a cancelar mais do que segurados mais velhos.

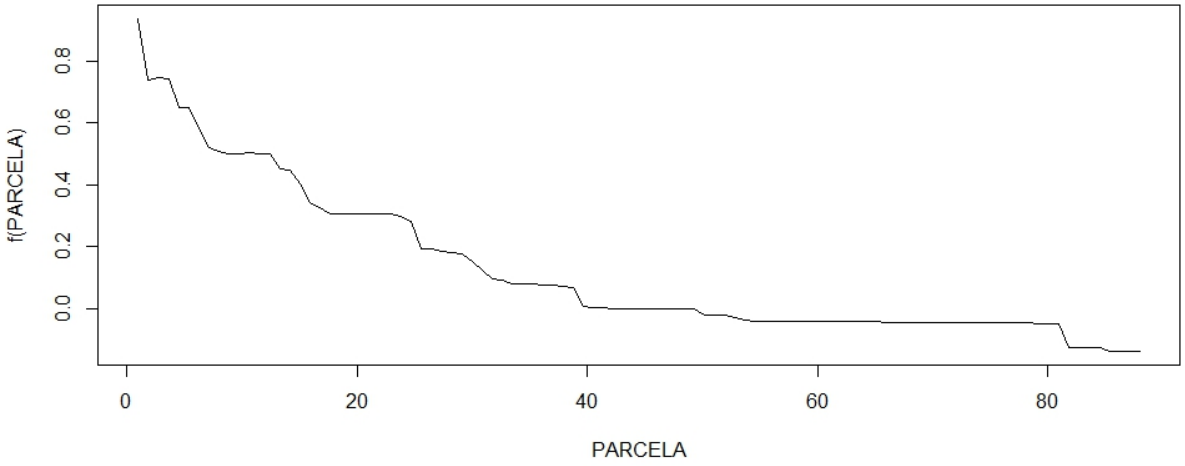

Gráfico 7 – Propensão a Cancelamento por Parcela.

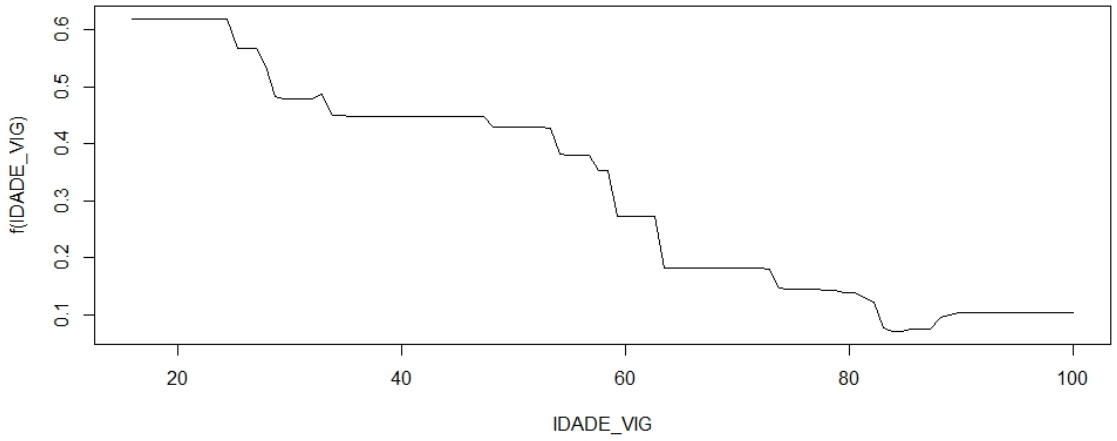

Gráfico 8 – Propensão a Cancelamento por Idade no Início de Vigência.

4.6. Comparação de resultados

Como podemos ver nos resultados abaixo muitos testes melhoram significativamente o Acc^{+}, em contrapartida temos uma ligeira queda do Acc. É importante observar que a combinação do modelo Random Forest com a base balanceada por over-sampling possibilitou um aumento tanto da métrica Sensitivity quanto Accuracy. Além disso, em todos os modelos obtivemos resultados melhores nas bases balanceadas do que na base padrão, o que corrobora

a ideia de que utilizar bases desbalanceadas é prejudicial para modelos preditivos de eventos raros. Inclusive, se comparássemos os modelos apenas através da base padrão, sem balanceamento, chegaríamos na conclusão de que aplicar Random Forest ou Boosting seria prejudicial para nossa estimação.

Acc$^+$					
MODELO	**PADRAO**	**OVER**	**UNDER**	**BOTH**	**SMOTE**
LOGIT	52,8%	53,2%	64,8%	64,9%	61,1%
BAGGING	53,6%	58,9%	72,1%	72,1%	73,5%
RF	46,7%	56,0%	72,0%	70,7%	70,7%
BOOSTING	51,3%	54,7%	66,2%	72,6%	65,0%

Tabela 3 – Resultados Sensitivity.

Acc					
MODELO	**PADRAO**	**OVER**	**UNDER**	**BOTH**	**SMOTE**
LOGIT	91,1%	90,8%	83,7%	83,6%	86,1%
BAGGING	90,8%	90,8%	85,2%	76,4%	83,1%
RF	90,8%	91,3%	86,2%	79,2%	79,2%
BOOSTING	91,2%	90,8%	83,1%	77,2%	83,6%

Tabela 4 – Resultados Accuracy.

Os estudos que tomamos por base e que estão descritos nas referências comparam os modelos citados anteriormente ou exemplificam ganhos de se usar bases balanceadas, mas nenhum faz as duas comparações simultaneamente como fizemos aqui. Este é um dos diferenciais deste trabalho.

5. Conclusão

Este trabalho é uma grande contribuição para a literatura, que é pouco desenvolvida quando observados estudos de previsão de churn no segmento de seguros, especialmente no de seguro de vida. A baixa frequência de cancelamentos e a falta de interação entre cliente e empresa durante a vigência do contrato são duas dificuldades presentes neste segmento e que foram trabalhadas neste estudo. Buscamos tornar a previsão mais precisa através do uso de técnicas de balanceamento (over-sampling, under-sampling e SMOTE) em conjunto com a utilização de métodos baseados em árvores de decisão (Bagging, Random Forest e Boosting).

As métricas sensitivity, specificity e accuracy foram utilizadas para avaliação e comparação das técnicas testadas neste projeto. Para cada modelo foi selecionada a base (balanceada ou não) que gerou maior sensitivity (Acc^+).

Comparando a base desbalanceada (denominada PADRAO) com as técnicas de balanceamento, observamos melhoras significativas nos resultados ao se utilizar under-sampling, a combinação de under-sampling com over-sampling e SMOTE-modificado. A única técnica que não gerou ganhos expressivos foi fazer apenas over-sampling. A utilização das bases balanceadas gerou uma melhora no resultado que variou entre 15,7% a 54,2%, de acordo com a técnica utilizada.

Na comparação entre o modelo de regressão logística (LOGIT) e as técnicas de machine learning exploradas neste estudo, estas últimas geraram melhorias que variaram entre 2,2% e 20,4%. O destaque foi a combinação da técnica SMOTE-modificado com o modelo Bagging, que gerou o maior Acc^+ (73,5%).

Em trabalhos futuros, a inclusão de variáveis que representem o ambiente macroeconômico pode resultar em ganhos significativos nas previsões, ao captar possíveis situações de perda da capacidade financeira dos clientes.

Outra potencial melhoria nos modelos poderia ser observada ao considerar informações sobre o perfil do corretor de seguros responsável pela venda do

serviço, visto que diferentes tendências ao cancelamento podem ser inferidas através da identificação do canal de venda.

Além disso, este estudo priorizou a minimização do erro tipo 1, ou seja, as comparações entre os modelos se restringiram apenas a evitar de deixar prever um cancelamento. No entanto, é possível que em determinados casos seja preferível uma pequena perda no Acc^{+} se, desta forma, seja possível evitar a indicação equivocada de um cancelamento. A depender do equilíbrio desejado entre os dois tipos de erro, futuros estudos podem explorar diferentes funções de perda ao comparar os resultados de cada modelo.

6.
Referências bibliográficas

BENOIT, D. F.; POEL, D. Van den. **Improving customer retention in financial services using kinship network information**. Expert Systems with Applications, Elsevier, v. 39, n. 13, p. 11435–11442, 2012

BUCKINX, W.; POEL, D. Van den. **Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting**. European Journal of Operational Research, Elsevier, v. 164, n. 1, p. 252–268, 2005.

CHAWLA, N.; BOWYER, K.; HALL, L.; KEGELMEYER,W. SMOTE: **Synthetic Minority Over-sampling Technique**, 2002

CHEN, C.; LIAW, A.; BREIMAN, L. Using random forest to learn imbalanced data. [S.l.], july 2004. Disponível em: <xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html>.

DOMINGOS, P. MetaCost: **A General Method for Making Classifiers Cost-Sensitive**. 1999. Disponível em: < https://homes.cs.washington.edu/~pedrod/papers/kdd99.pdf>.

EFRON, B. **Bootstrap Methods: Another Look at the Jackknife**. 1979. Disponível em:<https://www.jstor.org/stable/2958830?seq=1#page_scan_tab_contents>

EIBEN, A.; KOUDIJS, A.; SLISSER, F. **Genetic modelling of customer retention**. In: SPRINGER. European Conference on Genetic Programming. [S.l.], 1998. p. 178–186.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics Springer, Berlin, 2009.

IDRIS, A.; RIZWAN, M.; KHAN A. **Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies**. 2012. Disponível em: < http://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1490814>.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **Na Introduction to Statistical Learning**. 2013

JAPKOWICZ, N. **Learning from Imbalanced Data Sets: A Comparison of Various Strategies**. AAAI Technical Report, 2000. Disponível em: < http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-003.pdf>.

KIM, K.; JUN, C.-H.; LEE, J. **Improved churn prediction in telecommunication industry by analyzing a large network**. Expert Systems with Applications, Elsevier, v. 41, n. 15, p. 6575–6584, 2014.

LEMMENS, A. ;GUPTA, S. **Managing Churn to Maximize Profits**. 2013. Disponível em: <https://www.hbs.edu/faculty/Publication%20Files/14-020_3553a2f4-8c7b-44e6-9711-f75dd56f624e.pdf>.

LING, C.; LI, C. **Data Mining for Direct Marketing: Problems and Solutions**. AAAI Technical Report, 1998. Disponível em: < https://www.aaai.org/Papers/KDD/1998/KDD98-011.pdf>.

MOZER, M. C. et al. **Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry**. IEEE Transactions on neural networks, IEEE, v. 11, n. 3, p. 690–696, 2000

NIE, G. et al. **Credit card churn forecasting by logistic regression and decision tree**. Expert Systems with Applications, Elsevier, v. 38, n. 12, p. 15273–15285, 2011.

OWCZARCZUK, M. **Churn models for prepaid customers in the cellular telecommunication industry using large data marts**. Expert Systems with Applications, Elsevier, v. 37, n. 6, p. 4710–4712, 2010.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004.

SU, J. et al. **Customer retention predictive modeling in HealthCare Insurance Industry**. Jacksonville, Florida, 2011.

WEISS, G. M. **Mining with rarity: a unifying framework**. ACM Sigkdd Explorations Newsletter, ACM, v. 6, n. 1, p. 7–19, 2004.

ANEXOS

A. Análise Exploratória

Realizamos uma análise exploratória para entender as características da base que está sendo trabalhada, bem como para entender o comportamento das variáveis em relação ao cancelamento.

PARCELA	PREMIO	BENEFICIO	IDADE_VIG	UF
Min. : 1.00	Min. : 15.00	Min. : 3000	Min. : 16.00	RJ :288733
1st Qu.:11.00	1st Qu.: 46.91	1st Qu.: 15539	1st Qu.: 47.00	SP :168519
Median :28.00	Median : 109.75	Median : 29536	Median : 65.00	MG :105011
Mean :33.11	Mean : 203.94	Mean : 55102	Mean : 61.42	PR : 72823
3rd Qu.:54.00	3rd Qu.: 241.34	3rd Qu.: 60000	3rd Qu.: 76.00	RS : 70464
Max. :94.00	Max. : 817.73	Max. : 1000000	Max. :101.00	SC : 63150
				(Other):288147

SEXOID	ESTADOCIVILID	DS_FORMACOBRANCA	PAGAMENTO	MES_ING
F:599585	-1: 7946	BOLETO :192561	0: 85840	8 :102054
M:457262	1 :178398	CARTÃO DE CRÉDITO: 180	1:971007	9 : 98741
	2 :516327	CONVÊNIO COBRANÇA: 9224		3 : 96286
	3 : 56652	DÉBITO CONTA :582075		7 : 95077
	4 :280454	DESCONTO EM FOLHA:272807		11 : 94637
	5 : 17059			10 : 94463
	6 : 11			(Other):475589

Tabela 5 –Análise Descritiva.

O gráfico abaixo ilustra as parcelas emitidas e canceladas por mês. Podemos verificar um crescimento nas emissões durante o ao de 2015 e uma estabilidade nos períodos seguintes, com exceção do mês de dez/2016. A queda de emissão desse mês foi ocasionada por uma falha sistêmica corrigida imediatamente no mês seguinte. Como já seria previsto, o cancelamento possui a mesma estabilidade da emissão, porém numa escala menor.

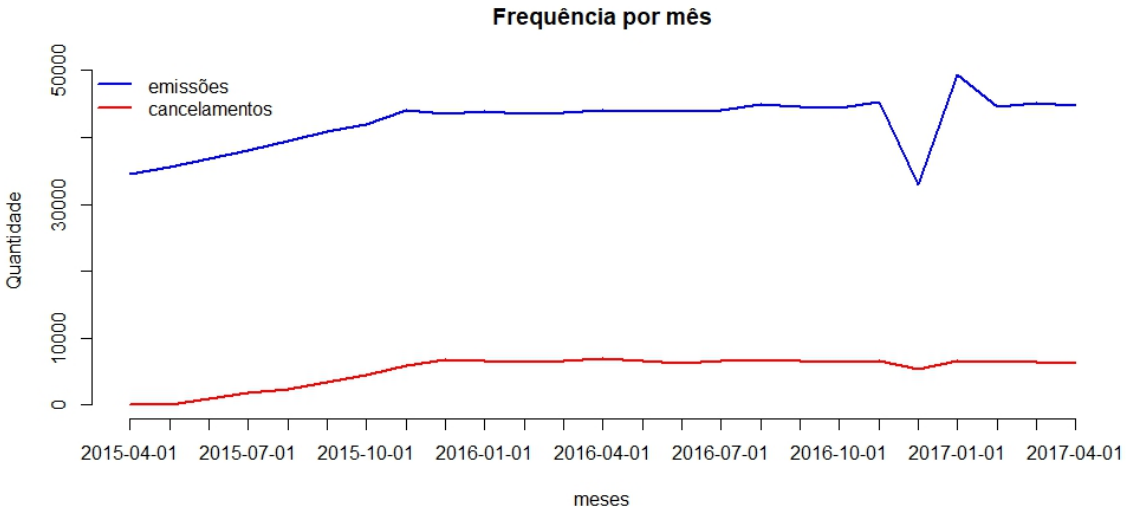

Gráfico 9 – Emissões e Cancelamentos por Competência.

O gráfico a seguir indica a frequência por idade atual do cliente, podemos observar que a distribuição dos clientes é mais distribuída entre as idades de 35 a 85 anos, com uma concentração maior nas faixas de 65 a 85 anos.

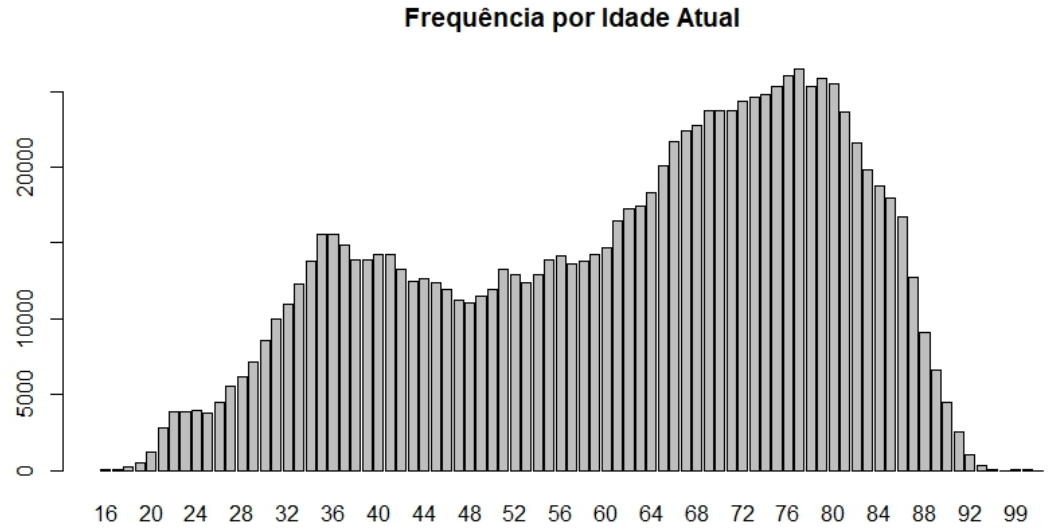

Gráfico 10 – Exposição por Idade Atual.

Devido à forte concentração por UF, exposta no gráfico abaixo, pela análise exploratória já poderíamos esperar que esta variável não fosse tão significativa ao modelo.

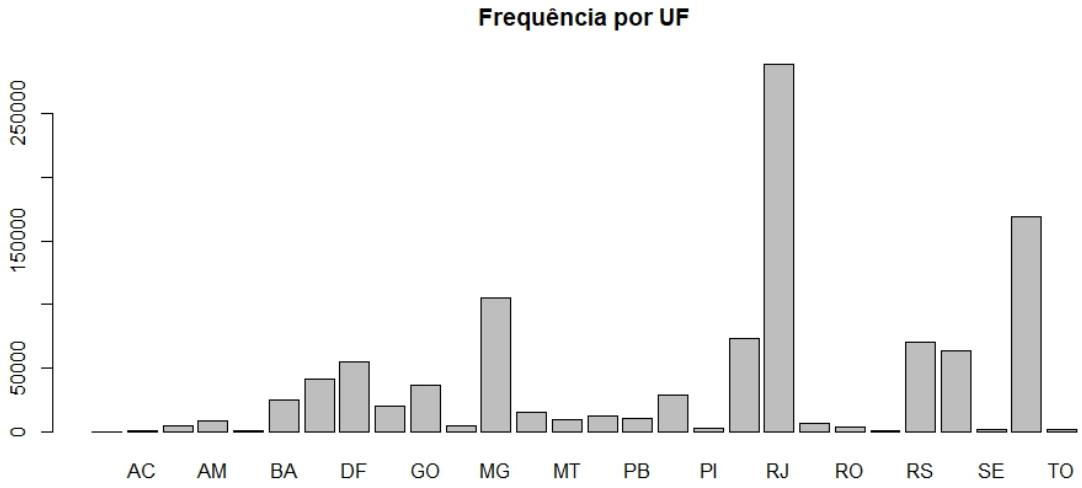

Gráfico 11 – Exposição Cancelamento por UF.

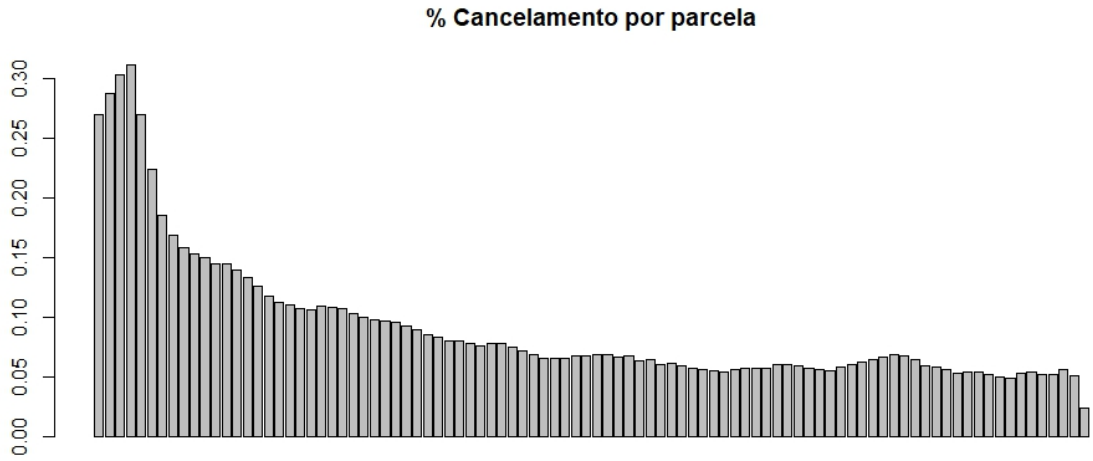

Gráfico 12 – Cancelamento por Parcela

Pelo gráfico acima fica evidente que clientes novos tendem a cancelar mais do que clientes antigos. O mesmo resultado vale quando comparamos jovens com clientes de idade mais avançada, gráfico abaixo.

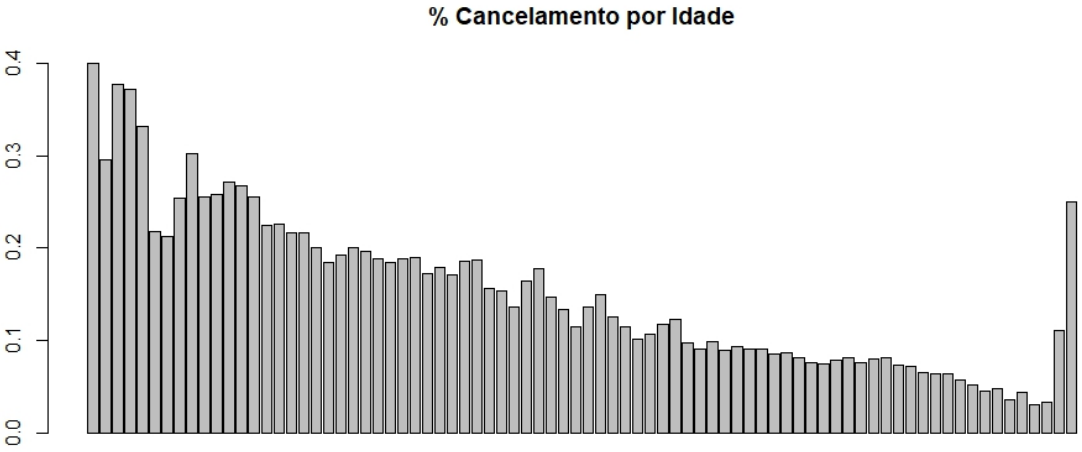

Gráfico 13 – Cancelamento por Idade Atual.

B. Resultados do Modelo Linear

Abaixo seguem os resultados do modelo de regressão linear para cada uma das bases disponíveis. Por ser mais simples que o modelo de regressão logística e por não incluir uma possível aleatoriedade dos dados, ele apresentou uma dificuldade ainda maior em prever o verdadeiro positivo (Acc^{+})das bases analisadas.

Modelo Linear					
MODELO	**PADRAO**	**OVER**	**UNDER**	**BOTH**	**SMOTE**
Acc^{+}	52,9%	52,9%	54,3%	58,0%	57,4%
Acc^{-}	97,5%	97,4%	95,9%	92,8%	93,3%
Acc	91,1%	91,0%	89,9%	87,8%	88,1%

Tabela 6 – Resultados Linear Model.

C. Seleção de Variáveis

Abaixo podemos ver o resultado do teste stepwise forward.

Start: AIC= 569156,2
CANCELAMENTO ~ 1

	Df	Deviance	AIC
+PAGAMENTO	1	409391	409395
+DIF_EMISSAO	1	442131	442135
+PARCELA	1	540702	540706
+IDADE_VIG	1	549360	549364
+DS_FORMA COBRANCA	4	550040	550050
+ESTADOCIVILID	6	561832	561846
+UF	27	562412	562468
+BENEFICIO	1	565390	565394
+PREMIO	1	566643	566647
+SEXOID	1	567226	567230
+MES_ING	11	568844	568868
<none>		569154	569156

Step: AIC= 409395,4
CANCELAMENTO ~ PAGAMENTO

	Df	Deviance	AIC
+IDADE_VIG	1	400200	400206
+PARCELA	1	400459	400465
+DS_FORMA COBRANCA	4	402152	402164
+ESTADOCIVILID	6	405418	405434
+UF	27	406275	406333
+BENEFICIO	1	407063	407069
+DIF_EMISSAO	1	407945	407951
+SEXOID	1	408149	408155
+PREMIO	1	408726	408732
+MES_ING	11	409254	409280
<none>		409391	409395

Step: AIC= 400206,1
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG

	Df	Deviance	AIC
+PARCELA	1	395457	395465
+DS_FORMA COBRANCA	4	396578	396592
+UF	27	398916	398976
+DIF_EMISSAO	1	399050	399058
+BENEFICIO	1	399446	399454
+ESTADOCIVILID	6	399829	399847
+SEXOID	1	399996	400004
+PREMIO	1	400058	400066
+MES_ING	11	400077	400105
<none>		400200	400206

Step: AIC= 395465,1
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA

	Df	Deviance	AIC
+DS_FORMA COBRANCA	4	392543	392559
+DIF_EMISSAO	1	394207	394217
+BENEFICIO	1	394478	394488
+UF	27	394479	394541
+ESTADOCIVILID	6	395047	395067
+PREMIO	1	395168	395178
+SEXOID	1	395267	395277
+MES_ING	11	395379	395409
<none>		395457	395465

Step: AIC= 392559
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA

	Df	Deviance	AIC
+DIF_EMISSAO	1	391503	391521
+BENEFICIO	1	391647	391665
+UF	27	391865	391935
+ESTADOCIVILID	6	392257	392285
+PREMIO	1	392280	392298
+SEXOID	1	392429	392447
+MES_ING	11	392474	392512
<none>		392543	392559

Step: AIC= 391520,9
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO

	Df	Deviance	AIC
+BENEFICIO	1	390803	390823
+UF	27	390819	390891
+ESTADOCIVILID	6	391237	391267
+PREMIO	1	391258	391278
+SEXOID	1	391405	391425
+MES_ING	11	391412	391452
<none>		391503	391521

Step: AIC= 390823
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO + BENEFICIO

	Df	Deviance	AIC
+UF	27	390147	390221
+ESTADOCIVILID	6	390524	390556
+MES_ING	11	390710	390752
+SEXOID	1	390730	390752
+PREMIO	1	390751	390773
<none>		390803	390823

Step: AIC= 390220,7
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO + BENEFICIO + UF

	Df	Deviance	AIC
+ESTADOCIVILID	6	389867	389953
+SEXOID	1	390066	390142
+MES_ING	11	390055	390151
+PREMIO	1	390105	390181
<none>		390147	390221

Step: AIC= 389953,3
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO + BENEFICIO + UF + ESTADOCIVILID

	Df	Deviance	AIC
+MES_ING	11	389775	389883
+SEXOID	1	389828	389916
+PREMIO	1	389828	389916
<none>		389867	389953

Step: AIC= 389882,7
CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO + BENEFICIO + UF + ESTADOCIVILID + MES_ING

	Df	Deviance	AIC
+SEXOID	1	389734	389844
+PREMIO	1	389736	389846
<none>		389775	389883

Step: AIC= 389844,2 CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO + BENEFICIO + UF + ESTADOCIVILID + MES_ING + SEXOID			
	Df	**Deviance**	**AIC**
+PREMIO	1	389700	389812
<none>		389734	389844

Step: AIC= 389812,2 CANCELAMENTO ~ PAGAMENTO + IDADE_VIG + PARCELA + DS_FORMACOBRANCA + DIF_EMISSAO + BENEFICIO + UF + ESTADOCIVILID + MES_ING + SEXOID + PREMIO

Tabela 7 – Resultado Stepwise Forward.

D. Resultados Specificity

			Acc⁻		
MODELO	**PADRAO**	**OVER**	**UNDER**	**BOTH**	**SMOTE**
LOGIT	97,5%	97,1%	86,9%	86,8%	90,3%
BAGGING	97,0%	96,1%	87,4%	77,1%	84,7%
RF	98,1%	97,2%	88,6%	80,6%	80,7%
BOOSTING	97,9%	96,9%	86,0%	78,0%	86,7%

Tabela 8 – Resultado Specificity.