Projeto de Graduação

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Junho 19, 2021

# Projetando preços de commodities de metais utilizando métodos de Machine Learning em um ambiente rico em dados

Ingo Varejão Seckelmann

PUC RIO

www.ele.puc-rio.br

Projeto de Graduação

DEPARTAMENTO
DE ENGENHARIA
ELÉTRICA

# Projetando preços de commodities de metais utilizando métodos de Machine Learning em um ambiente rico em dados

**Estudante: Ingo Varejão Seckelmann**

**Orientador: Álvaro Veiga**

Trabalho apresentado como requerimento parcial para a conclusão da graduação de Engenharia Elétrica na Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil.

## Agradecimentos

Agradeço a todos que contribuíram com minha formação ao longo dos últimos anos, especialmente aos meus familiares e amigos, que estiveram presentes nos momentos mais difíceis desta trajetória. Aos meus pais, Tanit Varejão Seckelmann e Dirk Seckelmann, obrigado pelo investimento na minha educação, por todo apoio emocional e pelos bons exemplos. Ao meu irmão, Udo Seckelmann, agradeço o bom exemplo de persistência e foco no trabalho e na faculdade.

Um agradecimento especial ao meu orientador, Álvaro Veiga, que foi de vital importância para o desenvolvimento deste projeto. Por último, queria agradecer a todos os colegas e a todas as pessoas que conheci na PUC que tiveram alguma influência na minha formação como pessoa e na minha visão de mundo.

# Projetando preços de commodities de metais utilizando métodos de Machine Learning em um ambiente rico em dados

## Resumo

O projeto tem por objetivo aplicar métodos de Machine Learning (ML) a fim de obter ganhos de projeção quando comparados a modelos mais tradicionais. Os métodos de Machine Learning são capazes de operar com um número elevado de variáveis explicativas mesmo em uma situação em que há menos dados do que variáveis candidatas, o que não é possível, por exemplo, com modelos lineares. Um grande número de modelos foi testado especificamente para o mundo de commodities metálicas, como minério de ferro, aço e cobre, num horizonte de projeção mensal dentro de até 2 anos. Motivados por isso, reunimos um conjunto de dados com uma quantidade relevante de passado, dentre elas, instrumentos financeiros, taxa de câmbio e indicadores macroeconômicos, que foram classificadas em tabelas. As metodologias utilizadas em cada modelo também são explicitadas ao longo do relatório e, por último, discutimos as performances de projeção de cada modelo em horizontes distintos para o preço spot do cobre na London Metal Exchange (LME).

Como veremos, o LASSO e o ElasticNet, dentre todos os modelos, foram os que melhores performaram na grande parte do horizontes. Importante ressaltar que o LASSO obteve um ganho significativo entre os modelos que utilizamos como benchmark, como o Random Walk e o Autoregressive, assim como contra modelos que, tradicionalmente, são utilizados para estimar preço futuro de séries financeiras heterocedásticas, como o ARCH e o GARCH. Seu método de seleção de variáveis provou-se robusto em diferentes janelas para cada horizonte de projeção, apresentando medidas de aderências, como RMSE e MAE, mais estáveis e menores ao longo da janela de projeção.

**Palavras-chave: Projeção, Machine Learning, Commodities, Big Data, Forecasting, Commodities prices, Metals**

# Forecasting metal commodities in a data-rich environment using Machine Learning methods

## Abstract

This project aims at applying Machine Learning (ML) methods in order to obtain forecasting gains in comparison to traditional models. The Machine Learning methods work with a high-dimension of variables even when there are more variables than historical data available, which is not possible for linear models. A big number of models were tested with metal commodities, such as the iron ore, steel and copper, in a monthly forecasting horizon up to two years. Motivated by that, we gathered a set of time series with a considerable amount of observations, including financial instruments, exchange rates and macroeconomic indicators, which were classified and displayed in tables. The methodologies used in each model are also explained in this work. Finally, we discuss the performance of each model in different forecasting horizons for the copper spot price in the London Metal Exchange (LME).

As we will see, the LASSO and the Elastic Net models, in comparison to the others, presented the best performance among different forecasting horizons. Even further, the LASSO obtained a significant gain in comparison to the benchmark models Random Walk and Autoregressive, as well as the more traditional models for forecasting heteroskedastic financial time series, ARCH and GARCH. Its method of variable selection proved to be robust in different windows for each forecasting horizon, presenting smaller and more stable performance metrics, such as the RMSE, MAE and MAD along the forecasting windows.

**Keywords: Forecasting, Machine Learning, Commodities, Big Data, ML, Commodities prices, Metals**

## Lista de Figuras

1 Preço do Cobre . . . 3
2 Log-variação do Cobre . . . 4
3 Frequência de melhor performance nas janelas móveis: RMSE . . . 15
4 Frequência de melhor performance nas janelas móveis: MAE . . . 16
5 Frequência de melhor performance nas janelas móveis: MAD . . . 17
6 Frequência de melhor performance nas janelas móveis: MAPE . . . 18
7 Janelas móveis de métrica de performance RMSE . . . 19
8 Janelas móveis de métrica de performance MAE . . . 20
9 Janelas móveis de métrica de performance MAD . . . 21
10 Projeções dentro da janela de teste . . . 22
11 Erros de Projeção . . . 23
12 Seleção de Regressores - LASSO . . . 24
13 Tabela de metadados cheia - Grupos 1, 2 e 3 . . . 28
14 Tabela de metadados cheia - Grupos 4, 5 e 6 . . . 29

## Lista de Tabelas

1 Descrição das séries: Atividade . . . . . 4
2 Descrição das séries: Ações, Juros e Câmbio . . . . . 5
3 Descrição das séries: Commodities . . . . . 5
4 Descrição das séries: Comércio Exterior . . . . . 5
5 Descrição das séries: Preços . . . . . 5
6 Descrição das séries: Imobiliário . . . . . 6
7 Estatísticas dos resultados . . . . . 10
8 Contagem de erros mínimos e máximos por modelo . . . . . 10
9 Performance por horizonte: RMSE . . . . . 11
10 Performance por horizonte: MAE . . . . . 11
11 Performance por horizonte: MAD . . . . . 12
12 Performance por horizonte: MAPE . . . . . 12
13 Performance por horizonte: % de Erros relativos a 50% . . . . . 13

## Sumário

**Lista de Figuras** **iv**

**Lista de Tabelas** **v**

**1 Introdução** **1**
a Principais Pontos . . . 1
b Organização do Relatório . . . 2

**2 Dados** **3**

**3 Metodologia** **7**

**4 Resultados** **8**
a Resultado Geral . . . 9
b Resultados por horizonte . . . 11
c Frequência de menor erro por horizonte . . . 14
d Janelas móveis de aderência . . . 19
e Projeção dos modelos na janela de teste . . . 22
f Erros de Projeção por horizonte . . . 23
g Seleção de Regressores via LASSO . . . 24

**5 Conclusão** **25**

**6 Referências** **26**

**A Apêndice** **27**

**B Apêndice** **30**
a Modelos Benchmark: . . . 30
b Modelos lineares com penalização (Shrinkage): . . . 30
c Modelos ARCH: . . . 31
d Modelos Ensemble: . . . 31

# 1 Introdução

Para muitos países, em especial aqueles cujas principais atividades sejam a produção e a exportação de produtos primários, a flutuação e a imprevisibilidade de preços de commodities no mercado global continua sendo um dos principais fatores de risco para suas principais empresas e, na maioria das vezes, para suas próprias economias. Porém, o impacto não se restringe a apenas esses países, uma vez que toda a economia global pode ser afetada por um choque, seja de oferta, seja de demanda, de uma determinada commodity, como ocorrido nas crises de petróleo de 1973 e de 1979 devido a um choque de oferta do barril por fatores políticos no Oriente Médio. Por isso, a tentativa de projetar o preço desses materiais primários, e de reduzir o erro ao máximo, é de suma importância para a gestão de risco de toda cadeia produtiva global, especialmente dos mercados produtores e consumidores diretos.

Dessa forma, a prática sempre foi um grande foco de interesse de empresas, bancos centrais, economistas e estatísticos para o melhor planejamento financeiro e para a melhor formulação de políticas macroeconômicas. Entretanto, projetar preços de commodities é complicado, pois sua variabilidade é muito impactada por decisões políticas externas, que acarretam choques que os modelos não captam bem em geral. Ainda assim, torna-se relevante o acompanhamento das forças de mercado na determinação do preço desses materiais para que se auxilie no processo de tomada de decisão de diversas entidades.

Com o avanço do poder computacional e com a facilidade prática que a aplicação de modelos que requerem uma grande quantidade de dados se tornou, métodos de Machine Learning passaram a ser alternativas para se obter algum ganho sistemático na projeção de indicadores de interesse [1]. Portanto, a fim de explorar tal universo de modelagem, buscamos, a partir de um ambiente rico em dados que possam explicar o comportamento de commodities metálicas, como aço, cobre e minério de ferro, obter algum ganho de projeção num prazo de até 2 anos.

Assim, mediremos o erro médio de cada horizonte de projeção por modelo, utilizando RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAD (Median Absolute Deviation from Median) e MAPE (Mean absolute percentage error) como métricas de performance. Como uma proposta a metrificar a performance de cada modelo, também registramos a proporção de erros negativos e de erros positivos de cada um por horizonte de projeção, verificando como se ajustam a movimentos muito fortes de alta ou de baixa do metal de interesse, o cobre.

Nesse projeto, utilizamos diversos modelos: dos mais básicos, que chamamos de Benchmark, como o Random Walk (RW) e o Autoregressivo (AR); modelos tradicionais para projeção de indicadores que apresentam volatilidade heterogênea ao longo do tempo, como o ARCH (Autoregressive conditional heteroskedasticity) [2] e o GARCH (Generalized Autoregressive conditional heteroskedasticity); métodos lineares que buscam reduzir a variância do modelo de regressão linear simples quando os aplicamos no conjunto de teste, como o Least Absolute Shrinkage and Selection Operator (LASSO), Ridge e ElasticNet; modelos altamente não-lineares e não-paramétricos como o Random Forest (RF) [3], Bagging [4] e Boosting, que também são conhecidos como modelos de “ensemble”.

## a Principais Pontos

Sobre os resultados encontrados:

1. Mostramos que é possível melhorar as métricas de aderência, consistentemente, ao longo de todos os horizontes de projeção, contra outros modelos básicos que servem de benchmark, assim como modelos mais tradicionais que tratam por volatilidade de uma série financeira.
2. Modelos de Machine Learning lineares, em especial o LASSO, mostraram-se mais performáticos dentre todos os modelos testados neste trabalho.
3. Modelos não-lineares e não-paramétricos - RF e Bagging - vêm logo em seguida no ranking de performance após LASSO e ElasticNet.
4. As projeções resultantes de modelos mais comumente utilizados na literatura para estimar retornos de séries financeiras altamente voláteis [5] não apresentou ganhos significativos como os modelos de ML.
5. A seleção de regressores do LASSO para o curto prazo tende a ter mais séries de preços e juros, enquanto para o longo prazo tende a ter séries de autoregressivo para explicar melhor o movimento da log-variação do preço do cobre. Para as demais classes de variáveis, a taxa de participação de cada uma para cada horizonte é bem uniforme.

### b Organização do Relatório

O relatório baseia-se no paper sobre projeção da inflação americana em um ambiente rico em dados [1]. Portanto, seguimos a mesma estrutura de organização: na seção 2 falamos um pouco das séries temporais que usamos tanto para projetar quanto para explicar a variação; na seção 3 descrevemos a metodologia utilizada para projetar e medir a performance de cada modelo; na seção 4 falamos dos resultados a partir de algumas métricas de performance, comparando os modelos benchmark com os demais; finalmente, concluímos o relatório na seção 5.

## 2 Dados

A fim de melhor explicar as variações de preço da commodity metálica de interesse, selecionamos um conjunto de 59 séries temporais que se dividem em setores distintos da economia, como setor externo, financeiro, imobiliário e atividade, e que estão relacionadas de alguma forma com as forças de oferta e demanda do metal. Além disso, adicionamos 30 novas variáveis explicativas a partir da aplicação de lags em algumas séries, dentre elas a própria variável dependente. Portanto, no total, foram 89 séries utilizadas para compreender o comportamento do movimento no preço. Para o cobre, pegamos o preço do spot na London Metal Exchange (LME).

Usamos séries de diferentes frequências: diárias, mensais e trimestrais, e ajustamo-las de tal forma a convertê-las em mensais. Para as séries diárias de preço, aplicamos a média de cada mês respectivo a fim de evitar escolher um dia em que o preço tenha destoado muito (da média), seja para cima ou para baixo, por conta de uma volatilidade elevada. Enquanto para as séries trimestrais aplicamos uma interpolação linear simples entre os valores.

Nossa amostra de dados vai de janeiro de 2009 até dezembro de 2019 (132 pontos de observação), sendo o período de projeção de janeiro de 2018 a dezembro de 2019 (24 meses de projeção). Portanto, outras séries candidatas sem disponibilidade de dados durante esse intervalo não foram selecionadas. Também aplicamos uma transformação de log em algumas variáveis, geralmente nas não-estacionárias.

Seguem a seguir gráficos da variável dependente e tabelas das variáveis explicativas, indicando o tipo de transformação aplicada, a quantidade de lags usada como input nos modelos, o nome e uma breve descrição sobre unidade. A coluna "tcode"representa o tipo de transformação daquela variável no modelo. Caso igual a 1, nenhuma transformação é feita, caso seja 2, o log neperiano é aplicado, enquanto caso seja 3, aplicamos a log-variação, dada por $X_t = \ln P_t - \ln P_{t-1}$. Por último, colocamos tabelas de metadados mais completas no Apêndice A indicando de onde pegamos as séries, suas periodicidades e suas fontes oficiais.


Figura 1: Preço do Cobre

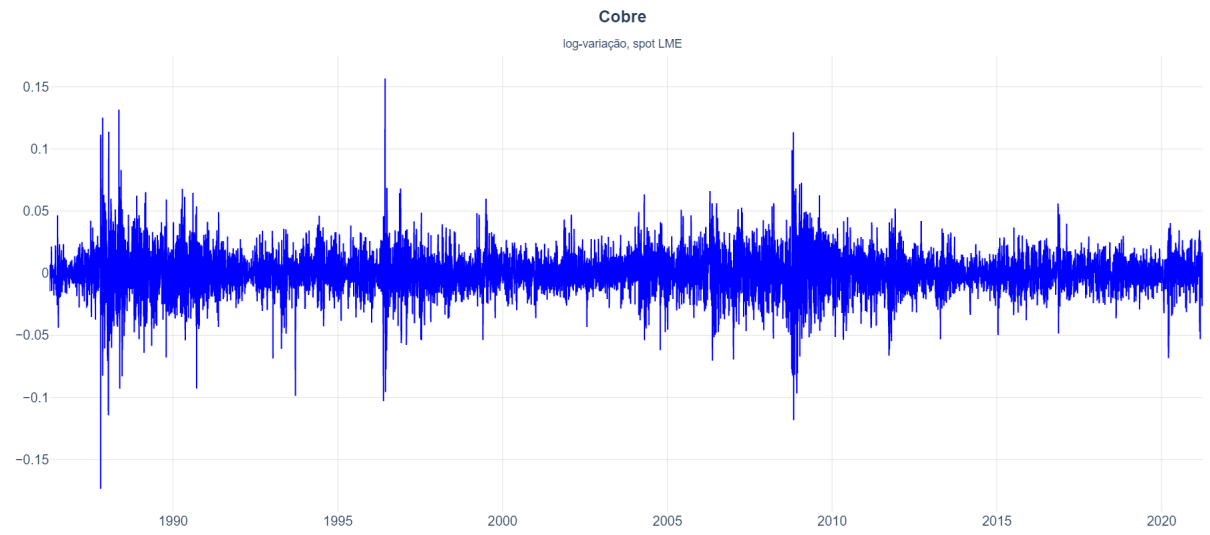



Figura 2: Log-variação do Cobre

**Grupo 1: Atividade**

| código | tcode | # lags | nome da série | descrição |
|---|---|---|---|---|
| ISMMFG | 1 | 0 | ISM Manufacturing | Diffusion Index, SA |
| ISMMFGPROD | 1 | 0 | ISM Manufacturing Report on Business Production | Diffusion Index, SA |
| ISMMFGNEWO | 1 | 0 | ISM Manufacturing Report on Business New Orders | Diffusion Index, SA |
| ISMMFGSUPL | 1 | 0 | ISM Manufacturing Report on Business Supplier Deliveries | Diffusion Index, SA |
| ISMMFGINV | 1 | 0 | ISM Manufacturing Report on Business Inventories | Diffusion Index, SA |
| CHINAGDP | 2 | 0 | China Real GDP | Constant Prices, Seasonally Adjusted |
| USGDP | 2 | 0 | US Real GDP | Billions of Chained 2012 Dollars, Constant Prices, SA |
| EAGDP | 2 | 0 | Euro Area Real GDP | Millions of Chained 2010 Euros, Constant Prices, SA |
| CHINDPROD | 2 | 2 | China Industrial Production | Seasonally Adjusted, Created from YoY Change series |
| USINDPROD | 2 | 2 | US Industrial Production: Total Index | Seasonally Adjusted |
| EAINDPROD | 2 | 2 | Euro Area Industrial Production Excluding Construction | Volume Index, Seasonally Adjusted |
| CPBINDPROD | 2 | 0 | World Industrial Production volume excluding construction | Volume Index SA |
| COPPERPRODCL | 2 | 1 | Chile Copper Production Total | Tons |
| USVHSALES | 2 | 0 | US Total Vehicle Sales | Millions of Units, SAAR |
| CHVEHICLESALES | 2 | 0 | China Automobile Sales | # of Cars |
| CHVEHICLEPROD | 2 | 0 | China Automobile Production | # of Cars |
| CHPMIMFGNBS | 1 | 0 | China NBS PMI Manufacturing | Diffusion Index, SA |
| CHPMIMFGPROD | 1 | 0 | China NBS PMI Manufacturing, Production | Diffusion Index, SA |
| CHPMIMFGNO | 1 | 0 | China NBS PMI Manufacturing, New Order | Diffusion Index, SA |
| CHPMIMFGNEO | 1 | 0 | China NBS PMI Manufacturing, New Export Order | Diffusion Index, SA |
| CHPMIMFGIMP | 1 | 0 | China NBS PMI Manufacturing, Import | Diffusion Index, SA |
| CHPMIMFGRMI | 1 | 0 | China NBS PMI Manufacturing, Raw Material Inventory | Diffusion Index, SA |

Tabela 1: Descrição das séries: Atividade

| Grupo 2: Ações, Juros e Câmbio | | | | |
|---|---|---|---|---|
| código | tcode | # lags | nome da série | descrição |
| BBDXY | 3 | 1 | Bloomberg Dollar Spot Index | Index |
| USDCNY | 3 | 1 | USDCNY Spot Exchange Rate - Price of 1 USD in CNY | Chinese renminbi (yuan) |
| USDCLP | 3 | 1 | USDCLP Spot Exchange Rate - Price of 1 USD in CLP | Chilean Peso |
| USDBRL | 3 | 1 | USDBRL Spot Exchange Rate - Price of 1 USD in BRL | Brazilian Real |
| USDAUD | 3 | 1 | USDAUD Spot Exchange Rate - Price of 1 USD in AUD | Australian Dollar |
| USDJPY | 3 | 1 | USDJPY Spot Exchange Rate - Price of 1 USD in JPY | Japanese Yen |
| USDKRW | 3 | 1 | USDKRW Spot Exchange Rate - Price of 1 USD in KRW | South Korean Won |
| US10Y | 1 | 0 | US Treasury Government Generic 10Y | US Treasury 10Y fixed rate |
| US5Y | 1 | 0 | US Treasury Government Generic 5Y | US Treasury 5Y fixed rate |
| FEDFUNDS | 1 | 0 | Effective Federal Funds Rate | FED funds effective rate |
| RIOTINTO | 3 | 2 | Rio Tinto ADR | Rio Tinto PLC operates as a mining company |
| VALE | 3 | 2 | Vale ADR | Vale S.A. produces and sells iron ore, copper and more |
| BHP | 3 | 2 | BHP Group ADR | BHP Group Limited operates as a mining company |

Tabela 2: Descrição das séries: Ações, Juros e Câmbio

| Grupo 3: Commodities | | | | |
|---|---|---|---|---|
| código | tcode | # lags | nome da série | descrição |
| COPPERLME | 3 | 2 | Copper Spot Price | Official Cash Offer Price of Copper in LME |
| IRONORE | 3 | 2 | China import Iron Ore Fines 62% FE spot | CFR Tianjin port, USD per metric ton |
| STEEL | 3 | 2 | Steel Prices, Hrc and Rebar mean | HRC and Rebar |
| OIL | 3 | 2 | Crude Oil Price Index | Average of U.K. Brent, Dubai and WTI |
| NATGAS | 3 | 2 | Natural Gas Price Index | European, Japanese, and American Natural Gas Price |

Tabela 3: Descrição das séries: Commodities

| Grupo 4: Comércio Exterior | | | | |
|---|---|---|---|---|
| código | tcode | # lags | nome da série | descrição |
| CONTFREIGHT | 2 | 1 | ISL Container Throughput Index | Volume Index SA |
| CPBTRADE | 2 | 0 | World Trade Volume | Volume Index SA |
| CHINTCTIMP | 2 | 0 | China Import Commodity Volume - Integrated Circuit | Billions, Units/Persons, NSA |
| CHINTCTEXP | 2 | 0 | China Export Commodity Volume - Integrated Circuit | Billions, Units/Persons, NSA |
| CHCOPPIMP | 2 | 0 | China Import Commodity Volume - Unwrought Copper and Copper Products | Thousands, Units/Persons, NSA |
| CHIOIMP | 2 | 0 | China Import Commodity Volume - Iron Ore and Concentrates | Millions, Units/Persons, NSA |
| CITITOTAUD | 1 | 0 | Citi Commodity Terms of Trade - Australia | Relative performance of comdty export and import prices |
| CITITOTBRL | 1 | 0 | Citi Coommodity Terms of Trade - Brazil | Relative performance of comdty export and import prices |
| CITITOTCNY | 1 | 0 | Citi Commodity Terms of Trade - China | Relative performance of comdty export and import prices |
| CITITOTCLP | 1 | 0 | Citi Commodity Terms of Trade - Chile | Relative performance of comdty export and import prices |
| CITITOTEUR | 1 | 0 | Citi Commodity Terms of Trade - Euro Area | Relative performance of comdty export and import prices |
| CITITOTUSD | 1 | 0 | Citi Commodity Terms of Trade - United States | Relative performance of comdty export and import prices |

Tabela 4: Descrição das séries: Comércio Exterior

| Grupo 5: Preços | | | | |
|---|---|---|---|---|
| código | tcode | # lags | nome da série | descrição |
| ISMMFGPRIC | 1 | 0 | ISM Manufacturing Report on Business Prices | Diffusion Index, SA |
| ISMMFGEXPT | 1 | 0 | ISM Manufacturing Report on Business Export Orders | Diffusion Index, SA |
| ISMMFGIMPT | 1 | 0 | ISM Manufacturing Report on Business Imports | Diffusion Index, SA |
| PPIFINGOODS | 3 | 0 | Producer Price Index (PPI): Final Demand: Finished Goods | Seasonally Adjusted |
| PPIINTMAT | 3 | 0 | PPI: Intermediate Demand: Processed Goods for Intermediate Demand | Seasonally Adjusted |
| PPICRUDEMAT | 3 | 0 | PPI: Intermediate Demand: Unprocessed Goods for Intermediate Demand | Seasonally Adjusted |
| PPIIRONSTEEL | 3 | 0 | PPI: Metals and Metal Products: Iron and Steel | Not Seasonally Adjusted |
| CHPMIMFGPURCHS | 1 | 0 | China NBS PMI Manufacturing, Purchasing Price Index | Diffusion Index, SA |

Tabela 5: Descrição das séries: Preços

**Grupo 6: Imobiliário**

| código | tcode | # lags | nome da série | descrição |
|---|---|---|---|---|
| USHOUSESTART | 2 | 0 | US Housing Starts: New Privately Owned Housing Units Started | Thousands of Units, SAAR |
| USHOUSEPERMIT | 2 | 3 | US New Private Housing Units Authorized by Building Permits | Thousands of Units, SAAR |

Tabela 6: Descrição das séries: Imobiliário

## 3 Metodologia

Consideremos o modelo genérico a seguir, onde $h$ representa o horizonte de projeção em meses:

$$X_{t+h} = T_h(\boldsymbol{x_t}) + u_{t+h}, \qquad h = 1, ..., H, \qquad t = 1, ..., T \tag{1}$$

Onde:

$X_{t+h}$ = Delta do $\ln$ do Preço da commodity, $h$ meses à frente

$\boldsymbol{x_t}$ = Vetor de dimensão $n$ $\boldsymbol{x_t} = (x_{1t}, ..., x_{nt})'$, contendo as variáveis explicativas

$T_h$ = O modelo a ser aplicado utilizando-se todos os regressores, inclusive alguns lags dos próprios

$H$ = O horizonte máximo de projeção, no caso serão 24 meses

$T$ = Número máximo da janela de projeção, no caso serão 24 datas dentro da janela

$u_{t+h}$ = Erro de média zero

A projeção será dada por:

$$\hat{P}_{t+h} = \hat{T}_{h,t-w+1:t}(\boldsymbol{x_t}) \tag{2}$$

Onde:

$\hat{X}_{t+h}$ = Delta do $\ln$ do preço estimado, $h$ meses à frente, utilizando uma janela de $t - w + 1$ até $t$

$\hat{T}_{h,t-w+1:t}$ = Função estimada de determinado modelo, $h$ meses à frente, utilizando uma janela de $t - w + 1$ até $t$

$w$ = Tamanho da janela, que varia de acordo com: $w = 108 - h - 1$

As projeções são feitas numa janela móvel de tamanho fixo por horizonte de projeção. Por exemplo, o número de observações in-sample de uma projeção é dado por $w = 108 - h - 1$, onde $h$ é o número do mês que se está projetando para frente. Assim, para cada horizonte de projeção $h$, teremos várias janelas de variáveis explicativas que usaremos para projetar o preço do cobre dentro do período de janeiro de 2018 a dezembro de 2019.

Aplicamos essa mesma lógica de janelas móveis por horizonte a todos os modelos utilizados. Alguns não se utilizaram de variáveis explicativas distintas da variável que se busca explicar, como é o caso dos modelos benchmark, como o Random Walk e o Autoregressive, que aplicamos como base de comparação aos demais. Além dos modelos benchmark, consideramos modelos lineares com penalização paramétrica (LASSO, Ridge e Elastic Net), que nos permitem fazer seleção de variáveis que melhor explicam a variação do preço do cobre; modelos de *ensemble* (Bagging, Boosting e Random Forest), que criam múltiplas instâncias de um determinado modelo, geralmente árvore de decisão [6], e depois combina-las para produzir resultados mais precisos e com menor variância; e modelos mais tradicionais para a projeção de séries temporais financeiras que exibem uma volatilidade não-homogênea ao longo do tempo, como o ARCH e o GARCH [7].

Para mais detalhes de cada modelo implementado, verificar o Apêndice B.

## 4 Resultados

Nesta seção, tratarei de discutir os resultados encontrados com a tentativa de obter ganhos de projeção do preço do cobre. Comparamos os modelos usando distintas estatísticas, dentre elas: a Raiz do Erro Médio Quadrático (RMSE), Erro Médio Absoluto (MAE), Desvio Mediano Absoluto da Mediana (MAD) e Erro Percentual Absoluto Médio (MAPE). Além dessas, acrescentamos mais 1 tipo de estatística não-simétrica para verificar se algum modelo costuma ter erros mais positivos ou negativos. Portanto, essa estatística busca computar o % de erros positivos e o % de erros negativos de cada modelo, por horizonte.

Cada estatística pode ser descrita da seguinte forma:

$$\text{RMSE}_{m,h} = \sqrt{\frac{1}{T-T_0+1}\sum_{n=T_0}^{T}\hat{e}_{t,m,h}^2}$$

$$\text{MAE}_{m,h} = \frac{1}{T-T_0+1}\sum_{n=T_0}^{T}|\hat{e}_{t,m,h}|$$

$$\text{MAD}_{m,h} = \text{mediana}[|\hat{e}_{t,m,h} - \text{mediana}(\hat{e}_{t,m,h})|\,]$$

$$\text{MAPE}_{m,h} = \frac{100}{T-T_0+1}\sum_{n=T_0}^{T}\left|\frac{\hat{e}_{t,m,h}}{X_{t,m,h}}\right|$$

$$\%\ \text{erros}^{+} = \frac{100}{T-T_0+1}\sum_{n=T_0}^{T} I(\hat{e}_{t,m,h}), \text{ onde } I(\hat{e}_{t,m,h}) = \begin{cases} 1, & \text{se } \hat{e}_{t,m,h} > 0 \\ 0, & \text{caso contrário} \end{cases}$$

$$\%\ \text{erros}^{-} = \frac{100}{T-T_0+1}\sum_{n=T_0}^{T} I(\hat{e}_{t,m,h}), \text{ onde } I(\hat{e}_{t,m,h}) = \begin{cases} 1, & \text{se } \hat{e}_{t,m,h} < 0 \\ 0, & \text{caso contrário} \end{cases}$$

onde $\hat{e}_{t,m,h} = X_t - \hat{X}_{t,m,h}$ e $\hat{X}_{t,m,h}$ é a projeção para o mês $t$, feito pelo modelo $m$ com informação até $t-h$. As duas primeiras métricas de performance são as mais comumente utilizadas pela literatura, enquanto o MAPE apresenta uma medida relativa em termos percentuais do erro e o MAD é robusto a outliers por utilizar medianas. As medidas de % de erros positivos e negativos são para verificar se algum modelo apresenta, em média, assimetria de projeção.

A seguir, avaliaremos cada modelo por horizonte de projeção utilizando diversas métricas de aderência de nossos modelos. Como padrão, adotamos o uso de cada métrica relativa à resultante do Random Walk. Portanto, os valores do RW sempre serão iguais a 1 e todos os demais serão mostrados com relação a ele. Destacamos, em cada tabela, os maiores erros em vermelho, os menores em verde (e negrito) e os valores intermediários com um gradiente que apresenta o amarelo como cor intermediária da amplitude. Entretanto, vale ressaltar que, para a tabela de contagem de erros, a formatação condicional muda para a contagem de mínimos: modelos com um número grande de horizontes com erro mínimo é colorido de verde, indicando melhor performance que os demais modelos, para determinada métrica (na coluna), na maioria dos horizontes.

### a Resultado Geral

Primeiramente, falaremos do resultado geral dos modelos utilizados neste trabalho.

Começando pela tabela 7, registramos as médias e medianas de cada métrica de performance por modelo. Assim, veremos, nas primeiras 4 colunas, as médias do RMSE, MAE, MAD e MAPE, respectivamente, dos 24 horizontes de projeção, e, nas 4 colunas logo em seguida, a mediana dessas mesmas medidas. Percebe-se, destacados em verde e negrito, os modelos que tiveram a melhor performance média e mediana.

Em seguida, na tabela 8, apresentamos a contagem dos horizontes de projeção em que cada modelo obteve a métrica máxima e mínima. Um modelo apresentar a métrica máxima em determinado horizonte sinaliza sua má performance e contabiliza 1 em uma das 4 primeiras colunas, de acordo com a métrica, enquanto apresentar a métrica mínima sinaliza sua boa aderência ao valor observado e contabiliza em uma das 4 últimas colunas. Importante notar que, nessa tabela, cada coluna soma 24, representando os 24 meses de projeção a frente.

Assim, podemos pontuar os resultados encontrados:

1. Os modelos que usaram uma grande quantidade de preditores apresentaram um ganho de projeção sistemático sobre os modelos mais tradicionais e os benchmarks.
2. No geral, os modelos que obtiveram as menores métricas relativas ao Random Walk foram o LASSO e o ElasticNet, indicando superioridade de projeção com relação aos demais. A capacidade de seleção de variáveis de ambos os modelos permitiu obter ganhos de mais de 80% com relação ao RW, no caso do RMSE médio e mediano. O LASSO obteve as melhores medianas para o RMSE, MAE e MAD, e as melhores médias para o MAE e MAD, enquanto o ElasticNet mostrou-se ligeiramente superior ao LASSO na média do RMSE.
3. Além disso, podemos observar que o Bagging e o Random Forest vêm logo em seguida no ranking dessas estatísticas, mostrando que esses modelos não-paramétricos também melhoram sistematicamente a projeção da log-variação do cobre.
4. Os modelos mais tradicionais de projeção para séries financeiras heterocedásticas, como o ARCH e o GARCH [8], tiveram performance muito parecida do AR e não apresentaram a melhora esperada na projeção da log-variação do cobre.
5. Na tabela de contagem, o LASSO tem claro destaque na capacidade preditiva dentre todos os 24 horizontes de projeção, apresentando a maioria na contagem de métricas de performance com valor mínimo dentre todos os outros modelos, alternando muito pouco com o ElasticNet e, às vezes, com o Random Forest.
6. Vale ressaltar que a métrica MAPE, por estarmos projetando um valor que é a log-variação do preço do cobre, acaba tendo valores melhores (menores) para aqueles modelos que geram projeções mais estáveis e sem muita oscilação, mas não constantes (RW), como o AR, ARCH e GARCH, que obtiveram destaque nesse caso.
7. Entre os modelos benchmark, já se observou um ganho de performance do AR vs o RW.

Abaixo, a tabela de sumário mostra duas estatísticas para cada métrica de performance por modelo. Mostramos a média e a mediana de cada métrica em todos os horizontes de projeção, de 1 a 24 meses.

**Precisão de Projeção**

| | avg RMSE | avg MAE | avg MAD | avg MAPE | median RMSE | median MAE | median MAD | median MAPE |
|---|---|---|---|---|---|---|---|---|
| AR | 0,421 | 0,635 | 0,757 | 0,204 | 0,414 | 0,622 | 0,760 | **0,228** |
| ARCH | 0,422 | 0,632 | 0,742 | 0,215 | 0,415 | 0,619 | 0,742 | 0,243 |
| BAGGING | 0,280 | 0,521 | 0,511 | 0,573 | 0,275 | 0,507 | 0,514 | 0,667 |
| BOOSTING | 0,353 | 0,605 | 0,672 | 0,780 | 0,346 | 0,597 | 0,669 | 0,841 |
| ELASTICNET | **0,189** | 0,471 | 0,430 | 0,639 | 0,168 | 0,441 | 0,445 | 0,685 |
| GARCH | 0,419 | 0,629 | 0,739 | **0,204** | 0,412 | 0,616 | 0,741 | 0,229 |
| LASSO | 0,214 | **0,440** | **0,385** | **0,555** | **0,153** | **0,414** | **0,382** | 0,648 |
| RF | 0,274 | 0,517 | 0,539 | 0,586 | 0,267 | 0,504 | 0,533 | 0,655 |
| RIDGE | 0,409 | 0,552 | 0,504 | 0,649 | 0,268 | 0,518 | 0,498 | 0,711 |
| RW | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

Tabela 7: Estatísticas dos resultados

Abaixo, a tabela mostra mais duas estatísticas para cada métrica de performance por modelo. Em seguida, apresentamos o número de horizontes em que aquele modelo alcançou o erro relativo mais alto dentre os demais, e o número de horizontes em que aquele modelo alcançou o menor erro relativo dentre os demais.

**Contagem de mínimos e máximos**

| | # max RMSE | # max MAE | # max MAD | # max MAPE % | # min RMSE | # min MAE | # min MAD | # min MAPE % |
|---|---|---|---|---|---|---|---|---|
| AR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| ARCH | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BAGGING | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| BOOSTING | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 |
| ELASTICNET | 0 | 0 | 0 | 2 | 3 | 2 | 5 | 0 |
| GARCH | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| LASSO | 0 | 0 | 0 | 0 | 21 | 20 | 14 | 0 |
| RF | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| RIDGE | 1 | 0 | 0 | 2 | 0 | 0 | 3 | 0 |
| RW | 23 | 24 | 24 | 13 | 0 | 0 | 0 | 0 |

Tabela 8: Contagem de erros mínimos e máximos por modelo

### b Resultados por horizonte

As tabelas a seguir mostram a performance relativa de cada modelo ao Random Walk. A performance foi medida a partir dos resultados das aplicações de cada modelo em 24 janelas móveis por horizonte de projeção, para o período de janeiro de 2018 a dezembro de 2019. Como veremos, o LASSO apresenta, nos horizontes de projeção dispostos abaixo, os menores valores de cada métrica, revezando algumas vezes com o ElasticNet e com o Ridge, no caso do MAD para horizontes maiores.

**Copper Spot Price 2018-2019**

**Horizonte de Projeção**

| RMSE | 1 | 3 | 6 | 12 | 24 |
|---|---|---|---|---|---|
| AR | 0,591 | 0,426 | 0,392 | 0,392 | 0,311 |
| ARCH | 0,575 | 0,415 | 0,386 | 0,393 | 0,319 |
| BAGGING | 0,339 | 0,230 | 0,246 | 0,271 | 0,239 |
| BOOSTING | 0,463 | 0,353 | 0,361 | 0,321 | 0,247 |
| ELASTICNET | 0,187 | 0,160 | 0,158 | 0,146 | 0,222 |
| GARCH | 0,573 | 0,414 | 0,384 | 0,390 | 0,316 |
| LASSO | **0,186** | **0,153** | **0,145** | **0,135** | **0,163** |
| RF | 0,328 | 0,254 | 0,235 | 0,250 | 0,213 |
| RIDGE | 0,356 | 0,221 | 0,174 | 0,263 | 2,622 |
| RW | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

Tabela 9: Performance por horizonte: RMSE

**Copper Spot Price 2018-2019**

**Horizonte de Projeção**

| MAE | 1 | 3 | 6 | 12 | 24 |
|---|---|---|---|---|---|
| AR | 0,685 | 0,625 | 0,600 | 0,650 | 0,531 |
| ARCH | 0,682 | 0,614 | 0,590 | 0,643 | 0,536 |
| BAGGING | 0,536 | 0,466 | 0,480 | 0,549 | 0,478 |
| BOOSTING | 0,576 | 0,573 | 0,603 | 0,631 | 0,472 |
| ELASTICNET | 0,401 | 0,402 | **0,406** | 0,419 | 0,508 |
| GARCH | 0,678 | 0,611 | 0,588 | 0,639 | 0,533 |
| LASSO | **0,398** | **0,389** | 0,410 | **0,397** | **0,431** |
| RF | 0,526 | 0,523 | 0,482 | 0,539 | 0,435 |
| RIDGE | 0,563 | 0,461 | 0,426 | 0,522 | 0,975 |
| RW | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

Tabela 10: Performance por horizonte: MAE

**Copper Spot Price 2018-2019**

**Horizonte de Projeção**

| MAD | 1 | 3 | 6 | 12 | 24 |
|---|---|---|---|---|---|
| AR | 0,667 | 0,765 | 0,731 | 0,773 | 0,630 |
| ARCH | 0,716 | 0,715 | 0,708 | 0,729 | 0,639 |
| BAGGING | 0,443 | 0,500 | 0,527 | 0,529 | 0,506 |
| BOOSTING | 0,467 | 0,520 | 0,679 | 0,750 | 0,449 |
| ELASTICNET | 0,327 | 0,435 | 0,467 | 0,471 | 0,483 |
| GARCH | 0,715 | 0,715 | 0,703 | 0,720 | 0,636 |
| LASSO | **0,306** | **0,373** | 0,466 | 0,425 | **0,359** |
| RF | 0,523 | 0,549 | 0,568 | 0,581 | 0,450 |
| RIDGE | 0,561 | 0,508 | **0,421** | **0,407** | 0,597 |
| RW | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

Tabela 11: Performance por horizonte: MAD

**Copper Spot Price 2018-2019**

**Horizonte de Projeção**

| MAPE | 1 | 3 | 6 | 12 | 24 |
|---|---|---|---|---|---|
| AR | 0,367 | 0,292 | 0,341 | 0,231 | **0,159** |
| ARCH | 0,306 | 0,289 | 0,358 | 0,248 | 0,182 |
| BAGGING | 0,678 | 0,574 | 0,852 | 0,754 | 0,507 |
| BOOSTING | 0,539 | 0,757 | 1,034 | 1,055 | 0,542 |
| ELASTICNET | 0,478 | 0,501 | 0,955 | 0,789 | 0,759 |
| GARCH | **0,291** | **0,275** | **0,340** | **0,224** | 0,173 |
| LASSO | 0,563 | 0,575 | 1,015 | 0,666 | 0,674 |
| RF | 0,641 | 0,717 | 0,982 | 0,675 | 0,430 |
| RIDGE | 0,662 | 0,606 | 1,469 | 0,931 | 0,902 |
| RW | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

Tabela 12: Performance por horizonte: MAPE

O % de erros é relativo a 50% em termos absolutos

**Copper Spot Price 2018-2019**

**Horizonte de Projeção**

| Desvio do % de erros a 50% | 1 | 3 | 6 | 12 | 24 |
|---|---|---|---|---|---|
| AR | 0% | 0% | 0% | 0% | 4% |
| ARCH | 8% | 8% | 8% | 8% | 8% |
| BAGGING | 8% | 0% | 8% | 4% | 8% |
| BOOSTING | 13% | 8% | 4% | 4% | 4% |
| ELASTICNET | 8% | 4% | 13% | 0% | 25% |
| GARCH | 8% | 8% | 8% | 8% | 8% |
| LASSO | 4% | 13% | 4% | 8% | 29% |
| RF | 4% | 4% | 4% | 4% | 4% |
| RIDGE | 0% | 4% | 4% | 8% | 13% |
| RW | 0% | 4% | 8% | 8% | 4% |

Tabela 13: Performance por horizonte: % de Erros relativos a 50%

### c Frequência de menor erro por horizonte

Nesta avaliação de performance, estamos calculando cada métrica de aderência numa janela de 3 meses, dado um horizonte de projeção, dentro da nossa janela de projeção de 2 anos (2018-2019). Em seguida, registramos a frequência com que cada modelo demonstrou a melhor performance nessas janelas móveis, para cada um dos 24 horizontes.

A partir das tabelas abaixo, percebe-se uma frequência maior do LASSO e do ElasticNet em vários horizontes em termos de melhor performance nas métricas RMSE, MAE e MAD. Especialmente na métrica MAD, o ElasticNet encontra-se mais presente em termos de frequência de mínimo valor. Ao mesmo tempo, há alternância em diferentes horizontes com o Random Forest, Bagging e Ridge para distintas métricas.

**Horizonte de Projeção**

| RMSE | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 5% | 9% | 5% | 9% | 5% | 0% | 5% | 0% | 0% | 0% | 0% | 0% |
| ARCH | 5% | 5% | 5% | 0% | 0% | 5% | 5% | 5% | 5% | 5% | 5% | 9% |
| BAGGING | 18% | 5% | 9% | 23% | 9% | 27% | 9% | 14% | 14% | 14% | 9% | 5% |
| BOOSTING | 14% | 14% | 14% | 9% | 14% | 14% | 14% | 14% | 5% | 14% | 0% | 0% |
| ELASTICNET | 18% | 23% | 27% | 23% | 14% | 14% | 14% | 18% | 9% | 23% | 23% | 18% |
| GARCH | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 5% | 0% |
| LASSO | 32% | 27% | 27% | 32% | 23% | 9% | 9% | 18% | 36% | 14% | 27% | 45% |
| RF | 5% | 14% | 14% | 0% | 14% | 9% | 23% | 9% | 14% | 23% | 14% | 5% |
| RIDGE | 5% | 5% | 0% | 5% | 23% | 23% | 23% | 23% | 14% | 9% | 14% | 18% |
| RW | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 5% | 0% | 5% | 0% |

| | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 5% | 5% | 5% | 5% | 5% | 5% | 0% | 0% | 5% | 5% | 5% | 5% |
| ARCH | 5% | 9% | 9% | 5% | 5% | 5% | 5% | 9% | 5% | 5% | 5% | 5% |
| BAGGING | 18% | 0% | 27% | 18% | 23% | 18% | 9% | 14% | 9% | 14% | 5% | 5% |
| BOOSTING | 0% | 5% | 5% | 5% | 14% | 14% | 0% | 18% | 9% | 14% | 18% | 18% |
| ELASTICNET | 9% | 0% | 5% | 5% | 18% | 14% | 36% | 9% | 14% | 14% | 18% | 18% |
| GARCH | 5% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| LASSO | 50% | 41% | 27% | 36% | 9% | 27% | 14% | 32% | 36% | 27% | 18% | 9% |
| RF | 0% | 14% | 14% | 14% | 9% | 14% | 14% | 9% | 14% | 5% | 18% | 32% |
| RIDGE | 5% | 27% | 9% | 5% | 5% | 0% | 23% | 0% | 9% | 18% | 14% | 9% |
| RW | 5% | 0% | 0% | 9% | 14% | 5% | 0% | 9% | 0% | 0% | 0% | 0% |

Figura 3: Frequência de melhor performance nas janelas móveis: RMSE

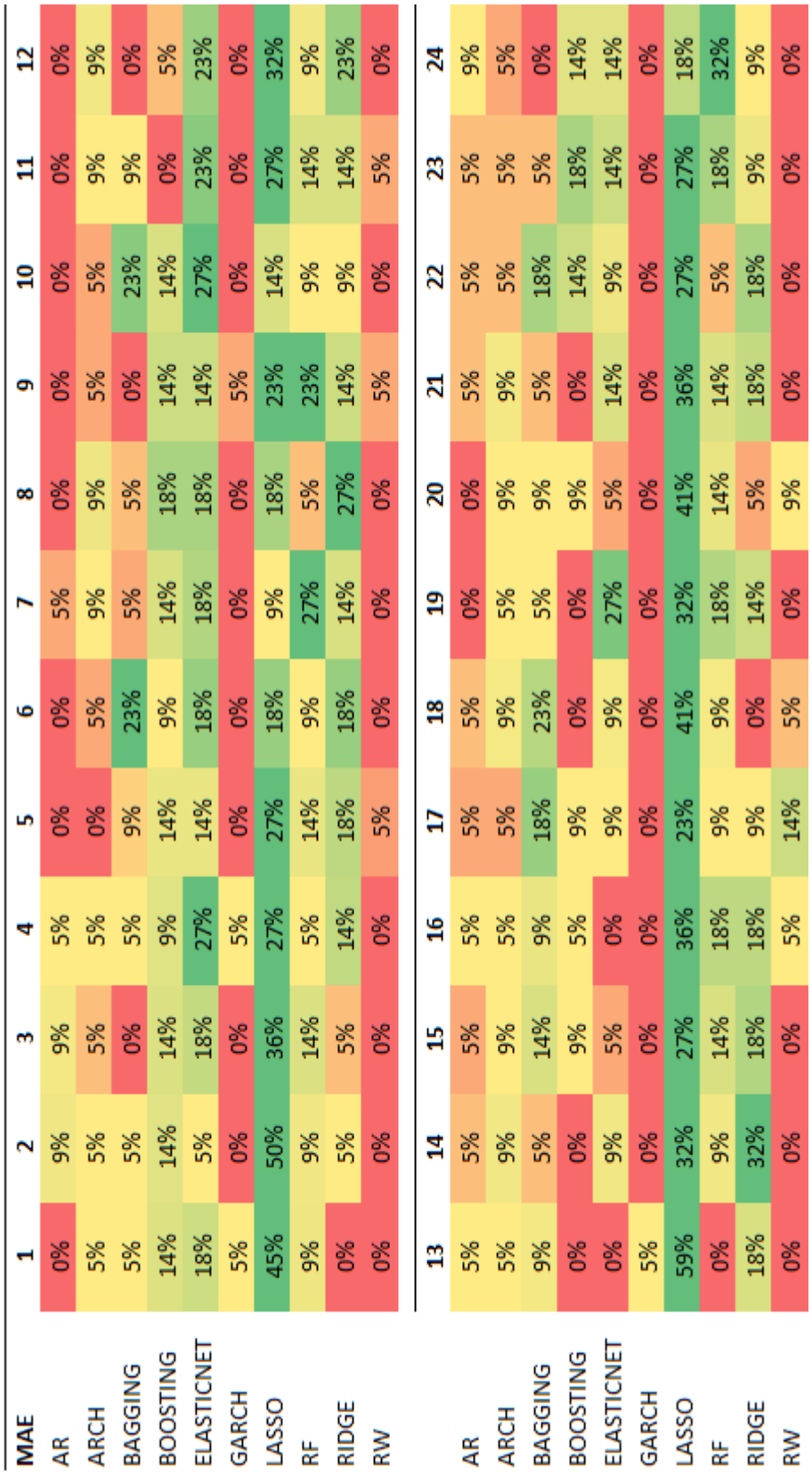



Figura 4: Frequência de melhor performance nas janelas móveis: MAE

**Horizonte de Projeção**

| MAD | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 5% | 5% | 0% | 0% | 0% | 0% | 0% | 0% | 5% | 5% | 5% | 9% |
| ARCH | 0% | 0% | 0% | 0% | 5% | 9% | 5% | 5% | 5% | 0% | 5% | 0% |
| BAGGING | 5% | 14% | 9% | 5% | 18% | 18% | 18% | 9% | 9% | 5% | 18% | 5% |
| BOOSTING | 14% | 14% | 9% | 27% | 0% | 5% | 5% | 14% | 9% | 9% | 5% | 18% |
| ELASTICNET | 18% | 18% | 18% | 27% | 9% | 45% | 32% | 23% | 32% | 18% | 23% | 27% |
| GARCH | 5% | 0% | 5% | 5% | 0% | 0% | 0% | 0% | 0% | 0% | 5% | 0% |
| LASSO | 18% | 9% | 14% | 14% | 23% | 5% | 23% | 18% | 9% | 32% | 5% | 14% |
| RF | 14% | 23% | 14% | 9% | 9% | 9% | 9% | 14% | 14% | 9% | 9% | 9% |
| RIDGE | 5% | 14% | 32% | 14% | 27% | 5% | 9% | 5% | 18% | 14% | 18% | 14% |
| RW | 18% | 5% | 0% | 0% | 9% | 5% | 0% | 14% | 0% | 9% | 9% | 5% |

| | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 5% | 5% | 5% | 5% | 5% | 0% | 0% | 0% | 5% | 5% | 0% | 0% |
| ARCH | 0% | 5% | 0% | 0% | 5% | 14% | 0% | 0% | 0% | 5% | 0% | 5% |
| BAGGING | 14% | 9% | 5% | 5% | 18% | 5% | 5% | 5% | 5% | 9% | 5% | 5% |
| BOOSTING | 18% | 5% | 5% | 14% | 5% | 0% | 14% | 9% | 23% | 5% | 32% | 14% |
| ELASTICNET | 14% | 14% | 14% | 32% | 27% | 23% | 32% | 18% | 27% | 18% | 32% | 36% |
| GARCH | 0% | 0% | 9% | 0% | 0% | 0% | 5% | 0% | 0% | 0% | 5% | 0% |
| LASSO | 5% | 27% | 23% | 0% | 14% | 14% | 14% | 36% | 14% | 27% | 0% | 14% |
| RF | 14% | 5% | 14% | 5% | 5% | 14% | 23% | 0% | 23% | 5% | 9% | 9% |
| RIDGE | 23% | 14% | 23% | 14% | 5% | 23% | 9% | 27% | 0% | 18% | 9% | 5% |
| RW | 9% | 18% | 5% | 27% | 18% | 9% | 0% | 5% | 5% | 9% | 9% | 14% |

Figura 5: Frequência de melhor performance nas janelas móveis: MAD

**Horizonte de Projeção**

| MAPE | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 5% | 23% | 14% | 23% | 23% | 18% | 23% | 23% | 14% | 18% | 23% | 14% |
| ARCH | 0% | 0% | 0% | 0% | 5% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| BAGGING | 0% | 0% | 5% | 14% | 5% | 0% | 5% | 5% | 0% | 5% | 5% | 5% |
| BOOSTING | 5% | 5% | 5% | 5% | 5% | 14% | 5% | 9% | 9% | 0% | 0% | 5% |
| ELASTICNET | 23% | 14% | 23% | 18% | 9% | 14% | 9% | 9% | 5% | 9% | 18% | 5% |
| GARCH | 32% | 23% | 27% | 18% | 5% | 32% | 27% | 14% | 27% | 27% | 27% | 23% |
| LASSO | 18% | 23% | 18% | 18% | 14% | 9% | 5% | 5% | 9% | 18% | 14% | 27% |
| RF | 9% | 5% | 9% | 5% | 14% | 5% | 14% | 9% | 9% | 0% | 5% | 0% |
| RIDGE | 9% | 9% | 0% | 0% | 9% | 9% | 14% | 27% | 18% | 18% | 9% | 23% |
| RW | 0% | 0% | 0% | 0% | 14% | 0% | 0% | 0% | 9% | 5% | 0% | 0% |

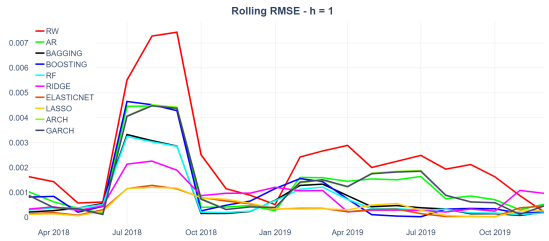
| | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 14% | 18% | 14% | 41% | 27% | 36% | 18% | 41% | 41% | 14% | 27% | 27% |
| ARCH | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| BAGGING | 0% | 5% | 5% | 0% | 9% | 9% | 0% | 5% | 5% | 0% | 0% | 0% |
| BOOSTING | 5% | 0% | 0% | 5% | 5% | 5% | 5% | 0% | 0% | 14% | 9% | 14% |
| ELASTICNET | 9% | 5% | 5% | 0% | 0% | 5% | 9% | 5% | 5% | 5% | 5% | 5% |
| GARCH | 14% | 23% | 23% | 14% | 18% | 18% | 14% | 0% | 0% | 27% | 23% | 23% |
| LASSO | 41% | 27% | 18% | 14% | 18% | 14% | 36% | 32% | 36% | 18% | 18% | 18% |
| RF | 5% | 0% | 5% | 5% | 5% | 5% | 9% | 9% | 0% | 14% | 0% | 9% |
| RIDGE | 14% | 23% | 32% | 18% | 9% | 0% | 9% | 9% | 9% | 5% | 18% | 5% |
| RW | 0% | 0% | 0% | 5% | 9% | 9% | 0% | 0% | 5% | 5% | 0% | 0% |

Figura 6: Frequência de melhor performance nas janelas móveis: MAPE

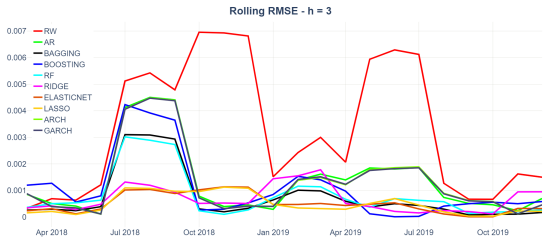
### d Janelas móveis de aderência

Os gráficos com os valores das métricas de aderência móveis (de 3 meses) anteriores encontram-se abaixo para alguns horizontes. Como padrão, adotamos os horizontes de 1, 3, 6, 12 e 24 meses à frente para cada métrica. Percebe-se que as métricas do LASSO e do ElasticNet se mantém estáveis ao longo de toda a janela de projeção, independente do horizonte, motivo pelo qual apresentam as maiores frequências de melhor performance para a maioria dos horizontes. Vale ressaltar que o Ridge apresenta um pico nas medidas de RMSE e MAE no horizonte de 24 meses à frente e, por isso, colocamos um outro gráfico logo ao lado sem a distorção causada por esse modelo.

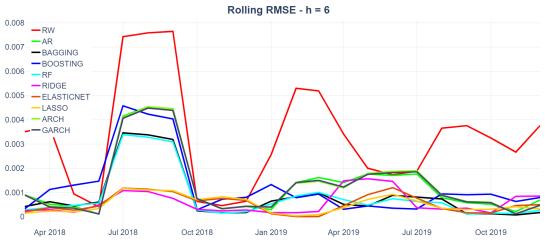
É possível observar, também, que o Random Walk, ARCH, GARCH e os modelos de ensemble apresentam picos de erros em meados de 2018, um dos motivos pelo qual performam, na média, abaixo dos modelos lineares de ML.



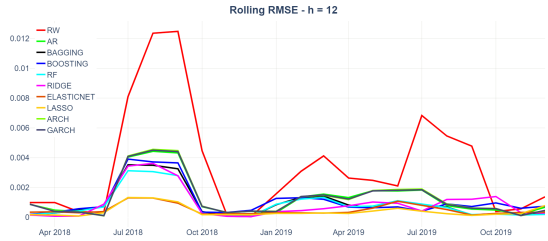

(a) Rolling RMSE - h = 1



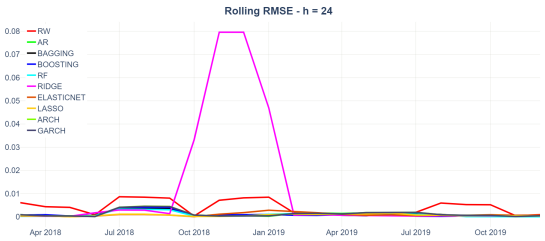

(b) Rolling RMSE - h = 3



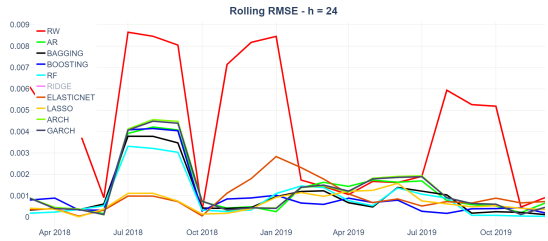

(c) Rolling RMSE - h = 6




(d) Rolling RMSE - h = 12

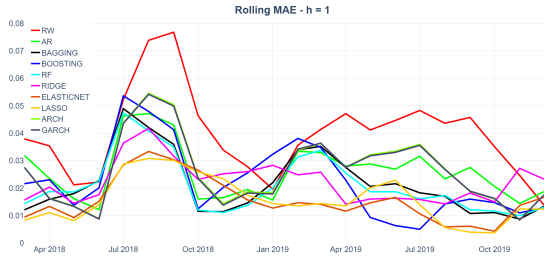



(e) Rolling RMSE - h = 24

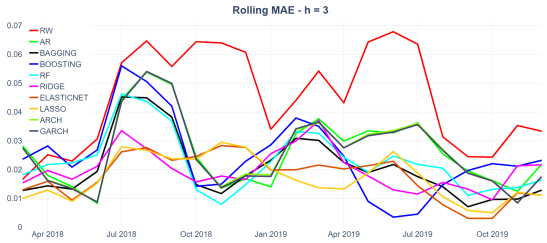



(f) Rolling RMSE - h = 24 ex Ridge

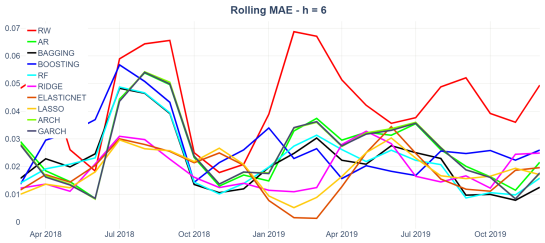
Figura 7: Janelas móveis de métrica de performance RMSE



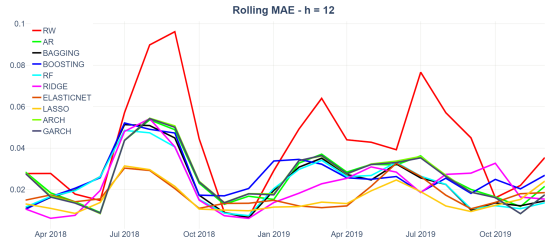

(a) Rolling MAE - h = 1



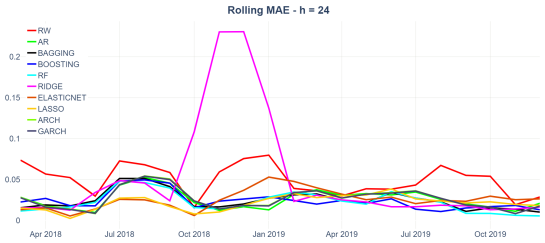

(b) Rolling MAE - h = 3



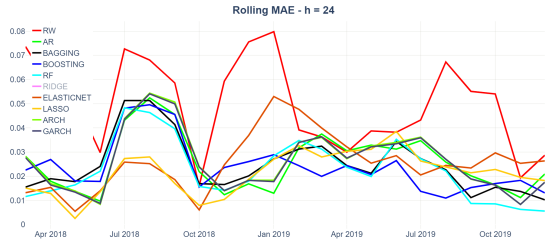

(c) Rolling MAE - h = 6




(d) Rolling MAE - h = 12

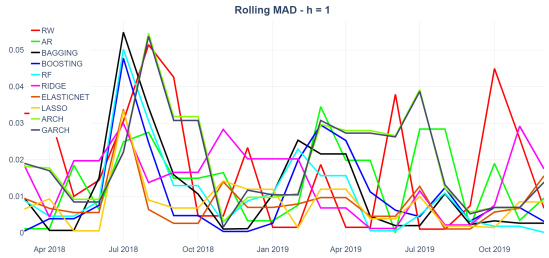



(e) Rolling MAE - h = 24

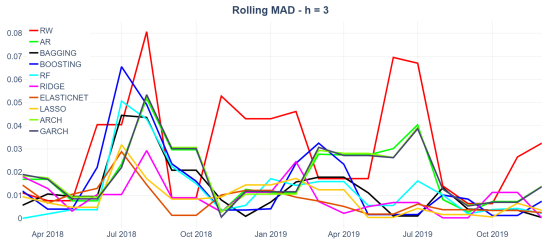



(f) Rolling MAE - h = 24 ex Ridge

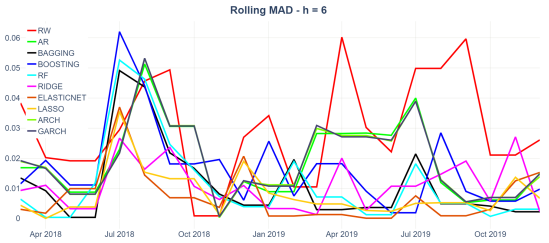
Figura 8: Janelas móveis de métrica de performance MAE



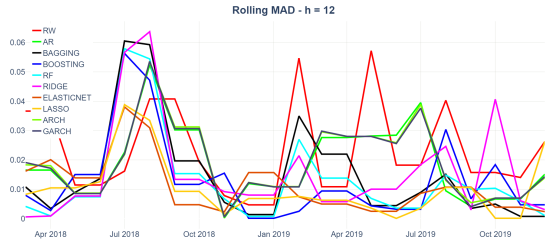

(a) Rolling MAD - h = 1



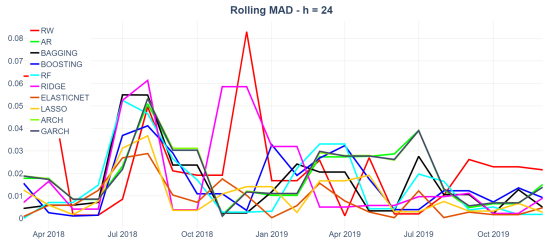

(b) Rolling MAD - h = 3




(c) Rolling MAD - h = 6




(d) Rolling MAD - h = 12




(e) Rolling MAD - h = 24

Figura 9: Janelas móveis de métrica de performance MAD

### e Projeção dos modelos na janela de teste

Projeção de cada modelo, $h$ meses à frente, dentro da janela de teste, de janeiro de 2018 até dezembro de 2019. Vale ressaltar que, como temos uma janela de 24 meses e um horizonte de projeção máximo também de 24 meses, as projeções no gráfico resultam da aplicação de observações que vão até dezembro de 2017 em cada modelo. Por exemplo, a projeção de 1 mês à frente seria com dados de janeiro de 2009 até dezembro de 2017 para a data janeiro de 2018, enquanto que a projeção de 24 meses seria com dados de dezembro de 2010 até dezembro de 2017 para a data dezembro de 2019.

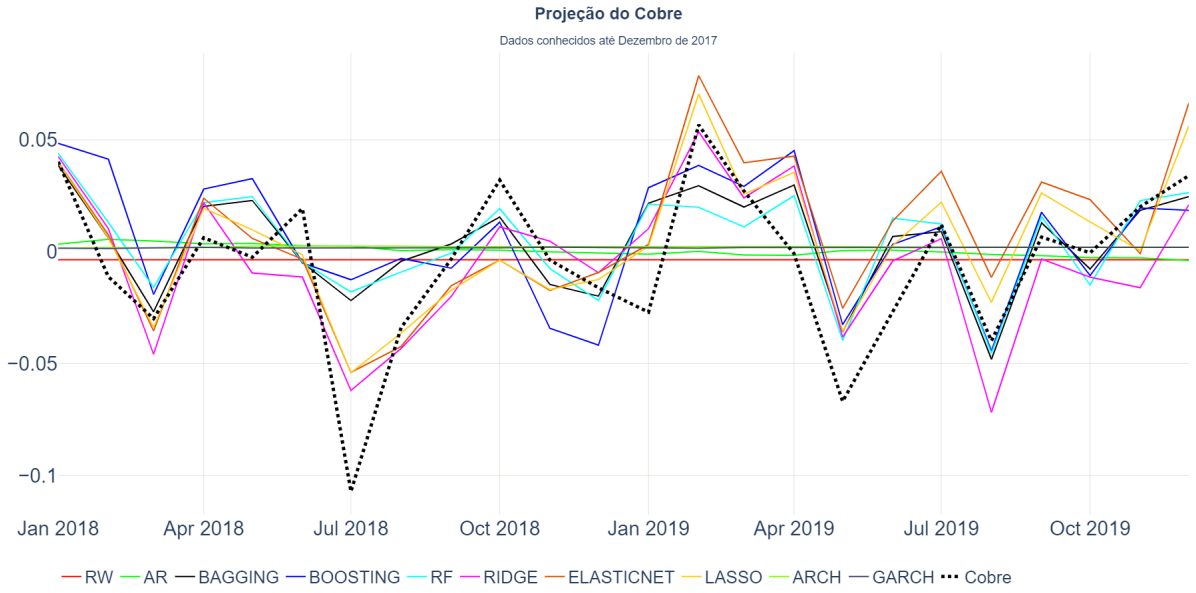
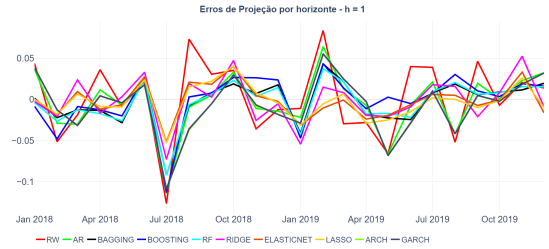


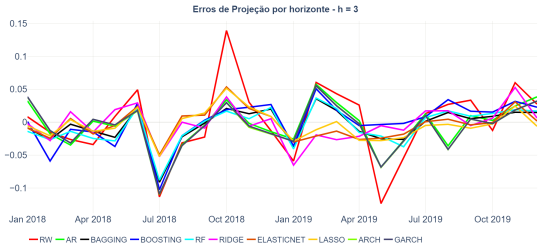

Figura 10: Projeções dentro da janela de teste

## f Erros de Projeção por horizonte

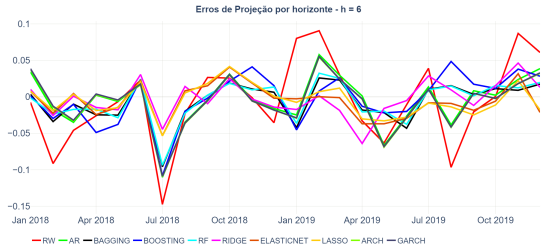
Os gráficos abaixo usam os mesmos horizontes de projeção, porém apresentam os erros de projeção de cada modelo. Percebe-se, novamente, os erros do LASSO e do ElasticNet mais estáveis e próximos a zero com relação aos demais modelos, o que justifica uma performance média superior.



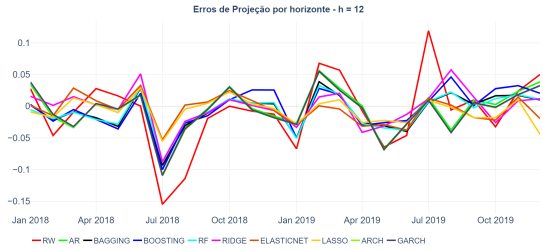

(a) h = 1



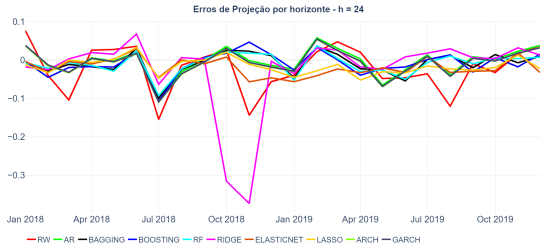

(b) h = 3




(c) h = 6




(d) h = 12




(e) h = 24

Figura 11: Erros de Projeção

### g Seleção de Regressores via LASSO

Participação de cada classe de regressor em cada horizonte de projeção, em todas as 24 janelas móveis. Percebe-se uma seleção bem distribuída entre todas as classes de regressores, com participações ligeiramente superiores das classes AR, câmbio, comércio externo e atividades.

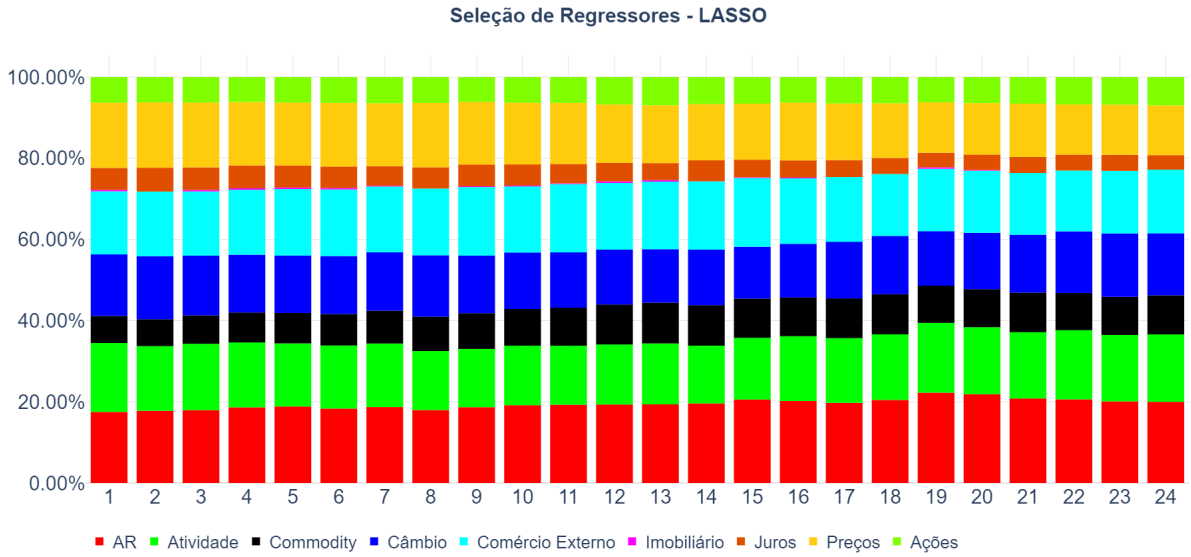



Figura 12: Seleção de Regressores - LASSO

## 5 Conclusão

A partir dos estudos feitos e dos resultados obtidos com a aplicação de modelos de Machine Learning mais recentes, mostramos que é possível obter ganhos de projeção do preço do cobre. Em especial com os modelos LASSO e ElasticNet, obtivemos as projeções com a maior precisão, com ganhos de até 80% em certas métricas de aderência. Além desses, modelos como Random Forest, Bagging e Ridge também apresentaram resultados superiores aos benchmarks. Dessa forma, destacamos que, em um ambiente rico em dados e com os modelos desenvolvidos nos últimos anos, é possível melhorar as previsões do preço do cobre ou até indicadores financeiros e econômicos [1]. Portanto, seria interessante avaliar, em eventuais pesquisas, se esse ganho sistêmico de projeção também seria obtido para outras commodities ou outras variáveis macroeconômicas e financeiras de diferentes países.

Dentre todos os modelos testados, aquele que apresentou os melhores resultados sistemáticos foi o LASSO, provavelmente por conta de sua capacidade de seleção de variáveis, penalizando aquelas que pouco explicam a log-variação do preço do cobre. O modelo obteve destaque tanto em métricas de janela móvel, quanto em avaliações médias e medianas para diferentes horizontes, mostrando tanto sua estabilidade de projeção, ao longo dos horizontes e das janelas, quanto anulando qualquer justificativa de uma eventualidade de projeção fora da curva que explique a melhor performance no geral.

Em suma, o LASSO foi o modelo que, ao longo de todos os horizontes de projeção, apresentou as melhores métricas de aderência, tanto em janelas móveis quanto na média de todos os horizontes. Sua seleção de variáveis, como podemos ver na figura 12, foi bem distribuída entre as classes de regressores, com seleção ligeiramente superior dos autoregressivos (AR), câmbio, comércio externo e atividades.

Por fim, espera-se, nos próximos anos, uma nova gama de estudos que explora e desenvolve modelos de ML com o fim de obter ganhos sistêmicos de projeção em variáveis macroeconômicas e financeiras, especialmente na área de modelos não-paramétricos que nos permite trabalhar com um universo de dados ainda superior aos modelos lineares com penalização.

## 6 Referências

[1] M. Cunha Medeiros, G. Vasconcelos, A. Veiga, and E. Zilberman, "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods." *Journal of Economic Literature*, (April 30, 2019). [Online]. Available: https://ssrn.com/abstract=3155480

[2] T. Bollerslev, R. F. Engle, and D. B. Nelson, "Chapter 49 arch models," ser. Handbook of Econometrics. Elsevier, 1994, vol. 4, pp. 2959–3038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1573441205800182

[3] L. . Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324

[4] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] T. Bollerslev, "A conditionally heteroskedastic time series model for speculative prices and rates of return," *The Review of Economics and Statistics*, vol. 69, no. 3, pp. 542–547, 1987. [Online]. Available: http://www.jstor.org/stable/1925546

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, no, Ed. Belmont, CA: Wadsworth International Group, 1984.

[7] S. Boutouria and F. Abid, "Modeling copper prices," *SSRN Electronic Journal*, 05 2010.

[8] R. Engle, "Garch 101: The use of arch/garch models in applied econometrics," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 157–168, December 2001. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/jep.15.4.157

[9] H. Zou and T. Hastie, "Zou h, hastie t. regularization and variable selection via the elastic net. j r statist soc b. 2005;67(2):301-20," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301 – 320, 04 2005.

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451

[11] J. Friedman, "Stochastic gradient boosting," *Computational Statistics  Data Analysis*, vol. 38, pp. 367–378, 02 2002.

# A Apêndice

**Grupo 1: Atividade**

| código | tcode | # lags | nome da série | descrição | região | fonte oficial | fonte de download | periodicidade | código fonte de download |
|---|---|---|---|---|---|---|---|---|---|
| ISMMFG | 1 | 0 | ISM Manufacturing | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMPMI Index |
| ISMMFGPROD | 1 | 0 | ISM Manufacturing Report on Business Production | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMPROD Index |
| ISMMFGNEWO | 1 | 0 | ISM Manufacturing Report on Business New Orders | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMNEWO Index |
| ISMMFGSUPL | 1 | 0 | ISM Manufacturing Report on Business Supplier Deliveries | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMSUPL Index |
| ISMMFGINV | 1 | 0 | ISM Manufacturing Report on Business Inventories | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMINV Index |
| CHINAGDP | 2 | 0 | China Real GDP | Constant Prices, Seasonally Adjusted | China | National Bureau of Statistics of China | Reuters - Datastream | QUARTER | CHCGDP..D |
| USGDP | 2 | 0 | US Real GDP | Billions of Chained 2012 Dollars, Constant Prices, SA | United States | U.S. Bureau of Economic Analysis | FRED | QUARTER | GDPC1 |
| EAGDP | 2 | 0 | Euro Area Real GDP | Millions of Chained 2010 Euros, Constant Prices, SA | Euro Area | Eurostat | FRED | QUARTER | CLVMEURSCAB1GQEA19 |
| CHINDPROD | 2 | 2 | China Industrial Production | Seasonally Adjusted, Created from YoY Change series | China | National Bureau of Statistics of China | Created | QUARTER | |
| USINDPROD | 2 | 2 | US Industrial Production: Total Index | Seasonally Adjusted | United States | Board of Governors of the Federal Reserve System (US) | FRED | MONTH | INDPRO |
| EAINDPROD | 2 | 2 | Euro Area Industrial Production Excluding Construction | Volume Index, Seasonally Adjusted | Euro Area | Eurostat | Datastream | MONTH | EKIPTOT.G |
| CPBINDPROD | 2 | 0 | World Industrial Production volume excluding construction | Volume Index SA | Global | CPB Netherlands Bureau for Economic Policy Analysis | CPB | MONTH | ipz_w1_qnmi_sp |
| COPPERPRODCL | 2 | 1 | Chile Copper Production Total | Tons | Chile | INE Chile | Bloomberg | MONTH | CLPCTOTL Index |
| USVHSALES | 2 | 0 | US Total Vehicle Sales | Millions of Units, SAAR | United States | U.S. Bureau of Economic Analysis | FRED | MONTH | TOTALSA |
| CHVEHICLESALES | 2 | 0 | China Automobile Sales | # of Cars | China | China Association of Automobile Manufacturers | CEIC | MONTH | 8474301 (CRAAAFC) |
| CHVEHICLEPROD | 2 | 0 | China Automobile Production | # of Cars | China | China Association of Automobile Manufacturers | CEIC | MONTH | 8466901 (CRAAACG) |
| CHPMIMFGNBS | 1 | 0 | China NBS PMI Manufacturing | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69851501 (CBAWLX) |
| CHPMIMFGPROD | 1 | 0 | China NBS PMI Manufacturing, Production | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69851701 (CBAWLZ) |
| CHPMIMFGNO | 1 | 0 | China NBS PMI Manufacturing, New Order | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69851601 (CBAWLY) |
| CHPMIMFGNEO | 1 | 0 | China NBS PMI Manufacturing, New Export Order | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69852101 (CBAWMD) |
| CHPMIMFGIMP | 1 | 0 | China NBS PMI Manufacturing, Import | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69852501 (CBAWMH) |
| CHPMIMFGRMI | 1 | 0 | China NBS PMI Manufacturing, Raw Material Inventory | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69852001 (CBAWMC) |

**Grupo 2: Ações, Juros e Câmbio**

| código | tcode | # lags | nome da série | descrição | região | fonte oficial | fonte de download | periodicidade | código fonte de download |
|---|---|---|---|---|---|---|---|---|---|
| BBDXY | 3 | 1 | Bloomberg Dollar Spot Index | Index | United States | Bloomberg | Bloomberg | DAY | BBDXY Index |
| USDCNY | 3 | 1 | USDCNY Spot Exchange Rate - Price of 1 USD in CNY | Chinese renminbi (yuan) | China | Bloomberg | Bloomberg | DAY | USDCNY Curncy |
| USDCLP | 3 | 1 | USDCLP Spot Exchange Rate - Price of 1 USD in CLP | Chilean Peso | Chile | Bloomberg | Bloomberg | DAY | USDCLP Curncy |
| USDBRL | 3 | 1 | USDBRL Spot Exchange Rate - Price of 1 USD in BRL | Brazilian Real | Brazil | Bloomberg | Bloomberg | DAY | USDBRL Curncy |
| USDAUD | 3 | 1 | USDAUD Spot Exchange Rate - Price of 1 USD in AUD | Australian Dollar | Australia | Bloomberg | Bloomberg | DAY | USDAUD Curncy |
| USDJPY | 3 | 1 | USDJPY Spot Exchange Rate - Price of 1 USD in JPY | Japanese Yen | Japan | Bloomberg | Bloomberg | DAY | USDJPY Curncy |
| USDKRW | 3 | 1 | USDKRW Spot Exchange Rate - Price of 1 USD in KRW | South Korean Won | South Korea | Bloomberg | Bloomberg | DAY | USDKRW Curncy |
| US10Y | 1 | 0 | US Treasury Government Generic 10Y | US Treasury 10Y fixed rate | United States | Bloomberg | Bloomberg | DAY | USGG10YR Index |
| US5Y | 1 | 0 | US Treasury Government Generic 5Y | US Treasury 5Y fixed rate | United States | Bloomberg | Bloomberg | DAY | USGG5YR Index |
| FEDFUNDS | 1 | 0 | Effective Federal Funds Rate | FED funds effective rate | United States | Bloomberg | Bloomberg | DAY | FEDL01 Index |
| RIOTINTO | 3 | 2 | Rio Tinto ADR | Rio Tinto PLC operates as a mining company | Australia | Bloomberg | Bloomberg | DAY | RIO US Equity |
| VALE | 3 | 2 | Vale ADR | Vale S.A. produces and sells iron ore, copper and more | Brazil | Bloomberg | Bloomberg | DAY | VALE US Equity |
| BHP | 3 | 2 | BHP Group ADR | BHP Group Limited operates as a mining company | Australia | Bloomberg | Bloomberg | DAY | BHP US Equity |

**Grupo 3: Commodities**

| código | tcode | # lags | nome da série | descrição | região | fonte oficial | fonte de download | periodicidade | código fonte de download |
|---|---|---|---|---|---|---|---|---|---|
| COPPERLME | 3 | 2 | Copper Spot Price | Official Cash Offer Price of Copper in LME | | London Metal Exchange | Bloomberg | DAY | LOCADY LME Comdty |
| IRONORE | 3 | 2 | China import Iron Ore Fines 62% FE spot | CFR Tianjin port, USD per metric ton | | IMF | IMF | MONTH | |
| STEEL | 3 | 2 | Steel Prices, Hrc and Rebar mean | HRC and Rebar | China | Antaike | Bloomberg | DAY | CDSPDRAV Index, HRDCTIJI Index |
| OIL | 3 | 2 | Crude Oil Price Index | Average of U.K. Brent, Dubai and WTI | | IMF | IMF | MONTH | |
| NATGAS | 3 | 2 | Natural Gas Price Index | European, Japanese, and American Natural Gas Price | | IMF | IMF | MONTH | |

Figura 13: Tabela de metadados cheia - Grupos 1, 2 e 3

**Grupo 4: Comércio Exterior**

| código | tcode | # lags | nome da série | descrição | região | fonte oficial | fonte de download | periodicidade | código fonte de download |
|---|---|---|---|---|---|---|---|---|---|
| CONTFREIGHT | 2 | 1 | ISL Container Throughput Index | Volume Index SA | Global | ISL - The Institute of Shipping Economics and Logistics | Reuters - Datastream | MONTH | BDCTISWAG |
| CPBTRADE | 2 | 0 | World Trade Volume | Volume Index SA | Global | CPB Netherlands Bureau for Economic Policy Analysis | CPB | MONTH | tgz_w1_qnmi_sn |
| CHINTCTIMP | 2 | 0 | China Import Commodity Volume - Integrated Circuit | Billions, Units/Persons, NSA | China | China Customs General Administration | Bloomberg | MONTH | CNIVINCT Index |
| CHINTCTEXP | 2 | 0 | China Export Commodity Volume - Integrated Circuit | Billions, Units/Persons, NSA | China | China Customs General Administration | Bloomberg | MONTH | CNEVINCT Index |
| CHCOPPIMP | 2 | 0 | China Import Commodity Volume - Unwrought Copper and Copper Products | Thousands, Units/Persons, NSA | China | China Customs General Administration | Bloomberg | MONTH | CNIVCOPP Index |
| CHIOIMP | 2 | 0 | China Import Commodity Volume - Iron Ore and Concentrates | Millions, Units/Persons, NSA | China | China Customs General Administration | Bloomberg | MONTH | CNMVIROM Index |
| CITITOTAUD | 1 | 0 | Citi Commodity Terms of Trade - Australia | Relative performance of comdty export and import prices | Australia | Citi | Bloomberg | DAY | CTOTAUD Index |
| CITITOTBRL | 1 | 0 | Citi Coommodity Terms of Trade - Brazil | Relative performance of comdty export and import prices | Brazil | Citi | Bloomberg | DAY | CTOTBRL Index |
| CITITOTCNY | 1 | 0 | Citi Commodity Terms of Trade - China | Relative performance of comdty export and import prices | China | Citi | Bloomberg | DAY | CTOTCNY Index |
| CITITOTCLP | 1 | 0 | Citi Commodity Terms of Trade - Chile | Relative performance of comdty export and import prices | Chile | Citi | Bloomberg | DAY | CTOTCLP Index |
| CITITOTEUR | 1 | 0 | Citi Commodity Terms of Trade - Euro Area | Relative performance of comdty export and import prices | Euro Area | Citi | Bloomberg | DAY | CTOTEUR Index |
| CITITOTUSD | 1 | 0 | Citi Commodity Terms of Trade - United States | Relative performance of comdty export and import prices | United States | Citi | Bloomberg | DAY | CTOTUSD Index |

**Grupo 5: Preços**

| código | tcode | # lags | nome da série | descrição | região | fonte oficial | fonte de download | periodicidade | código fonte de download |
|---|---|---|---|---|---|---|---|---|---|
| ISMMFGPRIC | 1 | 0 | ISM Manufacturing Report on Business Prices | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMPRIC Index |
| ISMMFGEXPT | 1 | 0 | ISM Manufacturing Report on Business Export Orders | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMEXPT Index |
| ISMMFGIMPT | 1 | 0 | ISM Manufacturing Report on Business Imports | Diffusion Index, SA | United States | Institute for Supply Management (ISM) | Bloomberg | MONTH | NAPMIMPT Index |
| PPIFINGOODS | 3 | 0 | Producer Price Index (PPI): Final Demand: Finished Goods | Seasonally Adjusted | United States | U.S. Bureau of Labor Statistics | FRED | MONTH | WPSFD49207 |
| PPIINTMAT | 3 | 0 | PPI: Intermediate Demand: Processed Goods for Intermediate Demand | Seasonally Adjusted | United States | U.S. Bureau of Labor Statistics | FRED | MONTH | WPSID61 |
| PPICRUDEMAT | 3 | 0 | PPI: Intermediate Demand: Unprocessed Goods for Intermediate Demand | Seasonally Adjusted | United States | U.S. Bureau of Labor Statistics | FRED | MONTH | WPSID62 |
| PPIIRONSTEEL | 3 | 0 | PPI: Metals and Metal Products: Iron and Steel | Not Seasonally Adjusted | United States | U.S. Bureau of Labor Statistics | FRED | MONTH | WPU101 |
| CHPMIMFGPURCHS | 1 | 0 | China NBS PMI Manufacturing, Purchasing Price Index | Diffusion Index, SA | China | National Bureau of Statistics of China | CEIC | MONTH | 69852401 (CBAWMG) |

**Grupo 6: Imobiliário**

| código | tcode | # lags | nome da série | descrição | região | fonte oficial | fonte de download | periodicidade | código fonte de download |
|---|---|---|---|---|---|---|---|---|---|
| USHOUSESTART | 2 | 0 | US Housing Starts: New Privately Owned Housing Units Started | Thousands of Units, SAAR | United States | Census; HUD | FRED | MONTH | HOUST |
| USHOUSEPERMIT | 2 | 3 | US New Private Housing Units Authorized by Building Permits | Thousands of Units, SAAR | United States | Census; HUD | FRED | MONTH | PERMIT |

Figura 14: Tabela de metadados cheia - Grupos 4, 5 e 6

# B Apêndice

## a Modelos Benchmark:

**Random Walk**
O mais básico modelo que estamos utilizando para efeito comparativo é o Random Walk (RW), no qual, para todos os $h$ horizontes, temos:

$$\hat{X}_{t+h} = X_t \tag{3}$$

**Autoregressivo**
O segundo modelo benchmark é o Autoregressivo (AR) de ordem p, onde p é determinado pelo Critério de Informação Bayesiano (BIC) e os parâmetros estimados por MQO (Mínimos Quadrados Ordinários), ou OLS (Ordinary Least Squares), em inglês. A equação de projeção é dada por:

$$\hat{X}_{t+h} = \hat{\phi}_{0,h} + \hat{\phi}_{1,h} X_t + ... + \hat{\phi}_{p,h} X_{t-p+1} \tag{4}$$

Vale ressaltar que cada horizonte apresenta um valor ótimo $p$ diferente.

## b Modelos lineares com penalização (Shrinkage):

Estimaremos modelos lineares em que $T_h(\boldsymbol{x_t}) = \boldsymbol{\beta}_h' \boldsymbol{x_t}$, onde

$$\hat{\boldsymbol{\beta}_h} = arg\min_{\boldsymbol{\beta}} [\sum_{t=1}^{T-h} (y_{t+h} - \boldsymbol{\beta}_h' \boldsymbol{x_t})^2 + \lambda \sum_{i=1}^{n} p(\beta_i; \alpha)] \tag{5}$$

$p(\beta_i; \alpha)$ sendo a função de penalização, $\lambda$ um parâmetro definido a partir da tentativa de vários valores e avaliados via cross-validation utilizando um k-fold em que k=10. Utilizaremos apenas 3 modelos de penalização nesse trabalho: LASSO, Ridge e ElasticNet.

**LASSO**
O LASSO é muito parecido com o Ridge, porém penaliza a norma $l_1$ dos coeficientes ou a soma de seus valores absolutos:

$$\lambda \sum_{i=1}^{n} p(\beta_i; \alpha) = \lambda \sum_{i=1}^{n} |\beta_i| \tag{6}$$

O LASSO encolhe variáveis irrelevantes para zero e tem boas propriedades em seleção de variáveis e em ajuste.

**Ridge**
O Ridge usa uma penalização quadrática ou uma penalidade $l_2$. Foi um dos primeiros métodos capazes de lidar com uma grande quantidade de variáveis numa regressão múltipla (Hoerl & Kennard 1970b,a).

$$\lambda \sum_{i=1}^{n} p(\beta_i; \alpha) = \lambda \sum_{i=1}^{n} \beta_i^2 \tag{7}$$

A regressão de Ridge encolhe o coeficiente de variáveis menos relevantes no modelo a próximo de zero. Dada a geometria da penalidade, diferente do LASSO, os coeficientes não vão exatamente para zero.

**ElasticNet**
O ElasticNet é uma generalização que inclui o LASSO e o Ridge simultaneamente. Ele é uma combinação convexa das normas $l_1$ e $l_2$ (Zou Hastie 2005) [9]. Assim como o LASSO, o ElasticNet também faz seleção de variáveis. Entretanto, como apresenta uma ponderação com a penalização do Ridge, ele tende a selecionar mais variáveis para o mesmo valor de $\lambda$. Assim, teremos a seguinte penalização:

$$\lambda \sum_{i=1}^{n} p(\beta_i; \alpha) = \alpha\lambda \sum_{i=1}^{n} \beta_i^2 + (1-\alpha)\lambda \sum_{i=1}^{n} |\beta_i| \tag{8}$$

onde $\alpha \in [0, 1]$.

### c Modelos ARCH:

**ARCH**
O ARCH, em geral, é usado para estimar o retorno de um ativo. Por conta da volatilidade de qualquer série financeira, o modelo é baseado em duas premissas: a primeira de que alta volatilidade aparece em clusters e, assim, a variação de um ativo é dependente de valores anteriores; a segunda de que a distribuição da variação é explicada pelo termo quadrático de valores anteriores. Assim, a variância condicional pode ser descrita como a função de $p$ valores de variações anteriores da seguinte forma:

$$\sigma^2_{t+h} = \alpha_{0,h} + \alpha_{1,h} X^2_t + ... + \alpha_{p,h} X^2_{t-p} \tag{9}$$

Assim, o modelo ARCH(p) fica:

$$X_{t+h} = \epsilon_{t+h} \sigma_{t+h} \qquad \epsilon_{t+h} \sim N(0, 1), . \tag{10}$$

**GARCH**
Já no GARCH, acrescentamos termos anteriores da própria série na variância condicional:

$$\sigma^2_{t+h} = \alpha_{0,h} + \sum_{i=1}^{p} \alpha_{i,h} X^2_{t-i} + \sum_{j=1}^{q} \beta_{j,h} \sigma^2_{t-j} \tag{11}$$

E o modelo GARCH(p, q) fica:

$$X_{t+h} = \epsilon_{t+h} \sigma_{t+h} \qquad \epsilon_{t+h} \sim N(0, 1), . \tag{12}$$

### d Modelos Ensemble:

Para esses modelos, utilizamos como referência os algoritmos da biblioteca sklearn da linguagem Python, que utiliza as seguintes referências para aplicar os modelos para projeção: Boosting [10] [11], Bagging [4], Random Forest [3]