

FUNDAÇÃO GETULIO VARGAS
ESCOLA BRASILEIRA DE ECONOMIA E FINANÇAS

PEDRO XAVIER SZYSZKA

RANDOM FOREST: UMA INVESTIGAÇÃO DA EFICIÊNCIA DE MERCADO DE
AÇÕES DA VALE

Rio de Janeiro

2018

PEDRO XAVIER SZYSZKA

**RANDOM FOREST: UMA INVESTIGAÇÃO DA EFICIÊNCIA DE MERCADO DE
AÇÕES DA VALE**

Dissertação para obtenção do grau de mestre apresentada
à Escola Brasileira de Economia e Finanças

Área de Concentração: Economia e Finanças Corporativas

Orientador: Rafael Chaves Santos

Rio de Janeiro

2018

Ficha catalográfica elaborada pela Biblioteca Mario Henrique Simonsen/FGV

Szyszk, Pedro Xavier

Random Forest : uma investigação da eficiência de mercado de ações
da Vale / Pedro Xavier Szyszk . – 2018.

45 f.

Dissertação (mestrado) - Fundação Getulio Vargas, Escola de Pós-
Graduação em Economia.

Orientador: Rafael Chaves Santos.

Inclui bibliografia.

1. Mercado de capitais – Modelos econométricos. 2. Teoria do mercado
eficiente. 3. Modelo de precificação de ativos. 4. Modelagem de dados. I.
Santos, Rafael Chaves . II. Fundação Getulio Vargas. Escola de Pós-Graduação
em Economia. III. Título.

CDD – 332.642

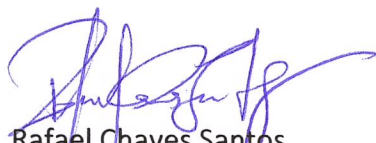
PEDRO XAVIER SZYSZKA

“RANDOM FOREST - UMA INVESTIGAÇÃO DA EFICIÊNCIA DE MERCADO DE AÇÕES DA VALE”.

Dissertação apresentado(a) ao Curso de Mestrado Profissional em Economia Empresarial e Finanças do(a) Escola de Pós-Graduação em Economia para obtenção do grau de Mestre(a) em Economia Empresarial e Finanças.

Data da defesa: 12/12/2018

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA



Rafael Chaves Santos
Orientador(a)



Daniela Kubudi Glasman



Axel André Simonsen

PEDRO XAVIER SZYSZKA

**RANDOM FOREST: UMA INVESTIGAÇÃO DA EFICIÊNCIA DE MERCADO DE
AÇÕES DA VALE**

Dissertação para obtenção do grau de mestre apresentada
à Escola Brasileira de Economia e Finanças
Área de concentração: Economia.

Aprovado em 12/12/2018.

Pela comissão organizadora:

Rafael Chaves Santos

Escola Brasileira de Economia e Finanças (EPGE)

Fundação Getúlio Vargas - RJ

Daniela Kubudi Glasman

Escola Brasileira de Economia e Finanças (EPGE)

Fundação Getúlio Vargas - RJ

Axel Simonsen

Escola Brasileira de Economia e Finanças (EPGE)

Fundação Getúlio Vargas - RJ

AGRADECIMENTOS

Gostaria de agradecer ao meu orientador, pelo direcionamento da pesquisa, análise, pensamento e interpretação. De forma precisa e objetiva as indicações foram cruciais para o desenvolvimento deste trabalho, desde a escolha do tema, metodologia, pesquisa, mudanças de direção e conclusão. No entanto, a maior contribuição vai além deste trabalho, com papel fundamental no aconselhamento e tutela na minha trajetória profissional em uma área completamente nova para mim. Por tudo, obrigado.

De forma alguma em menor importância, agradeço à minha “esposa” e eterna namorada por todo apoio, motivação e paciência nesta minha trajetória. Não poderia esperar companhia melhor para o caminho que trilhei, e não consigo me imaginar querendo outra companhia qualquer. De forma bem menos importante, mas enormemente significativa, devo a ela todo o mérito da formatação deste trabalho, em especial no meu convencimento em aderir a certas normas e formas textuais mais ortodoxas de escrita acadêmica. Certamente o trabalho estaria bem diferente sem sua presença e dedicação.

De forma mais ampla, gostaria de agradecer a todas as outras pessoas envolvidas neste processo, desde a secretaria, especialmente Gisele e Vitor, assim como todos os professores que demonstraram profissionalismo e comprometimento com a qualidade do curso assim como o aprendizado de seus alunos. Sou grato pelos debates e em especial ao diálogo e explicações, especialmente nas discussões minuciosas e dúvidas que, mesmo que demandassem pesquisa, nunca deixaram de ser sanadas.

Concluindo, meu último agradecimento fica para a Fundação Getúlio Vargas e à coordenação do curso, que mantém seu comprometimento com a excelência e a meta de aprimoramento contínuo, sempre buscando formas de apresentação do conteúdo assim como sua aderência ao conteúdo demandado pelo mercado.

“If you don't know, the thing to do is not to get scared, but to learn.”

Ayn Rand

RESUMO

O presente trabalho investigou sobre a aleatoriedade dos preços do ativo VALE3 negociado na Bolsa de Valores de São Paulo. Utilizamos a metodologia Random Forest para tentar prever sinais de retornos de 15min futuros. Através do modelo, conseguimos extrair caráter preditivo de sinais de retornos com 95% de significância estatística. Em especial, observamos a evolução de caráter preditivo com o aumento da janela de dados do modelo. Assim, conseguimos refutar a hipótese de passeio aleatório dos preços (*random walk*) para retornos de 15 minutos. Além disso, discutimos ao longo do texto as hipóteses utilizadas, assim como levantamos temas para estudos futuros acerca da hipótese de eficiência de mercado ainda sob a ótica da metodologia de aprendizado computacional, como aumento de base de dados e utilização de dados que carreguem informações acerca das expectativas dos agentes. Por fim, apresentamos brevemente os resultados obtidos do mesmo método aplicado a diferentes janelas de predição.

Palavras-chave: Random Forest. Random walk. Eficiência de mercado.

ABSTRACT

The present work investigated the randomness of the prices of the VALE3 asset traded on the São Paulo Stock Exchange. We used the Random Forest methodology to try to predict signs of 15min future returns. Through the model we were able to extract a predictive character of signals of returns with 95% of statistical significance. Particularly, we observe the evolution of a predictive character with an increase in the model data window. Thus, we were able to refute the random walk hypothesis for 15-minute returns. In addition, we have discussed the hypotheses used throughout the text, as well as raising topics for future studies about the hypothesis of market efficiency still from the perspective of computational learning methodology, such as increasing the database and using data that carries information about expectations of the agents. Finally, we briefly present the results obtained from the same method applied to different prediction windows.

Keywords: Random Forest. Random walk. Market efficiency.

LISTA DE ILUSTRAÇÕES

Gráfico 1:	Distribuição de R^2 ajustado de diferentes fatores parametrizados.	17
Gráfico 2:	Distribuição de R^2 da parametrização dos preços.....	18
Gráfico 3:	Exemplo de árvore decisória	19
Gráfico 4:	Ilustração de construção de árvores do modelo random forest	21
Gráfico 5:	Resultados para regressões em toda a amostra.	25
Gráfico 6:	Média móvel (1000obs) de acertos (base de dados >19.000).....	26
Gráfico 7:	Representação ilustrada de distribuições de probabilidades	27
Gráfico 8:	Distribuições via <i>probit</i> de acertos de sinal de retornos	28
Gráfico 9:	Distribuições via <i>logit</i> de acertos de sinal de retornos	29
Gráfico 10:	Resultados para regressões em toda a amostra.	39
Gráfico 11:	Média móvel (1000obs) de acertos por intervalo dados.....	39
Gráfico 12:	Distribuições via probit de acertos de sinal de retornos	39
Gráfico 13:	Distribuições via <i>logit</i> de acertos de sinal de retornos	39
Gráfico 14:	Resultados para regressões em toda a amostra.	41
Gráfico 15:	Média móvel (1000obs) de acertos por intervalo de dados.	41
Gráfico 16:	Distribuições via probit de acertos de sinal de retornos	41
Gráfico 17:	Distribuições via <i>logit</i> de acertos de sinal de retornos	41
Gráfico 18:	Resultados para regressões em toda a amostra.	42
Gráfico 19:	Média móvel (1000obs) de acertos por intervalo de dados.	42
Gráfico 20:	Distribuições via probit de acertos de sinal de retornos	42
Gráfico 21:	Distribuições via <i>logit</i> de acertos de sinal de retornos	42
Gráfico 22:	Resultados para regressões em toda a amostra.	43
Gráfico 23:	Média móvel (1000obs) de acertos por intervalo de dados.	43
Gráfico 24:	Distribuições via probit de acertos de sinal de retornos	43

Gráfico 25: Distribuições via <i>logit</i> de acertos de sinal de retornos	43
Gráfico 26: Resultados para regressões em toda a amostra.	44
Gráfico 27: Média móvel (1000obs) de acertos por intervalo de dados.	44
Gráfico 28: Distribuições via probit de acertos de sinal de retornos	44
Gráfico 29: Distribuições via <i>logit</i> de acertos de sinal de retornos	44
Gráfico 30: Resultados para regressões em toda a amostra.	45
Gráfico 31: Média móvel (1000obs) de acertos por intervalo de dados.	45
Gráfico 32: Distribuições via probit de acertos de sinal de retornos	45
Gráfico 33: Distribuições via <i>logit</i> de acertos de sinal de retornos	45

SUMÁRIO

1	INTRODUÇÃO	11
2	METODOLOGIA.....	13
2.1	DEFINIÇÃO DAS VARIÁVEIS.....	14
2.2	TRATAMENTO DOS DADOS	14
2.3	PARAMETRIZAÇÃO DAS VARIÁVEIS	16
2.4	APRESENTAÇÃO DO MODELO <i>RANDOM FOREST</i>	18
3	RESULTADOS	23
3.1	RESULTADOS DA PREVISÃO	23
3.2	PROBIT E LOGIT	26
3.3	TESTES DO MODELO SEM PARAMETRIZAÇÃO DOS DADOS.....	31
4	FUTUROS PONTOS DE INVESTIGAÇÃO.....	33
4.1	INFORMAÇÕES DE OFERTAS DE NEGOCIAÇÃO	33
4.2	ANÁLISE DA VOLATILIDADE IMPLÍCITA NAS OPÇÕES DO ATIVO OBJETO	34
4.3	ANÁLISE DE DIVERSOS ATIVOS SIMULTANEAMENTE.....	35
5	CONSIDERAÇÕES FINAIS	37
	REFERÊNCIAS.....	38
	ANEXO 1 – RESULTADOS DE PREVISÕES DE RETORNOS DE 30 MINUTOS.....	39
	ANEXO 2 – RESULTADOS DE PREVISÕES DE RETORNOS DE 1 HORA.....	40
	ANEXO 3 – RESULTADOS DE PREVISÕES DE RETORNOS DE 1 HORA.....	41
	ANEXO 4 – RESULTADOS DE PREVISÕES DE RETORNOS DE 2 HORAS.....	42
	ANEXO 5 – RESULTADOS DE PREVISÕES DE RETORNOS DE 3 HORAS.....	43
	ANEXO 6 – RESULTADOS DE PREVISÕES DE RETORNOS DE 4 HORAS.....	44
	ANEXO 7 – RESULTADOS DE PREVISÕES DE RETORNOS DE 5 HORAS.....	45

1 INTRODUÇÃO

A hipótese de que os preços seguem uma trajetória aleatória, imprevisível, é fundada na ideia de que as informações relevantes à precificação de um ativo são absorvidas imediatamente pelo mercado e que não seria possível antever estas variações. Daí os termos “*random walk*” e “passeio aleatório”, amplamente utilizados nas áreas de finanças e economia.

Existem diversas definições da hipótese de passeio aleatório, a mais forte definindo a distribuição de retornos como i.i.d. (“*independent and identical distribution*” ou, em português, distribuições idênticas e independentes), outra, um pouco menos rígida, definindo as distribuições dos retornos com média igual e não correlacionadas temporalmente.

De forma geral, a hipótese de passeio aleatório para retornos intradiários (objeto deste estudo) define que os retornos devem ter média muito próxima de zero e serem não auto correlacionados temporalmente. Em nosso estudo, investigaremos se de fato não há previsibilidade de preços com análise técnica, ou seja, estudo do retorno acumulado intradiário até aquele instante. Buscaremos identificar padrões de comportamento passados de preços e outros dados de mercado do ativo VALE3 e tentar prever uma apreciação ou desvalorização do ativo 15 minutos à frente temporalmente.

Para tal estudo, utilizaremos o método matemático Random Forest, ou seja, um conjunto de árvores decisórias. Popularemos tais árvores com variáveis representativas de dados de mercado (preço, volatilidade, volume financeiro, volatilidade de volume financeiro e número de negociações) a título de conseguirmos com base em dados históricos identificarmos padrões de comportamento das variáveis que possam explicar o preço do ativo 15 minutos a frente temporalmente.

Assim, caso alcancemos sucesso em obter preditores estatisticamente significativos de retornos, declararemos que, para o ativo e janela de estudo escolhida, nós fomos capazes de refutar com certo grau de confiabilidade a hipótese de aleatoriedade das variações dos preços. Embora eventuais resultados neste sentido indiquem uma

potencial fraqueza da Teoria da Eficiência de Mercados, não podemos refutar tal teoria sem calcularmos os custos de transação e analisar de forma mais profunda potenciais fatores limitantes de uma implantação de estratégia de negociação utilizando estas informações.

2 METODOLOGIA

Buscando identificar caráter preditivo de retornos futuros das negociações intradiárias do ativo VALE3, alimentaremos o modelo *Random Forest* (utilizando a linguagem “R” de programação) com variáveis que, conforme o nosso entendimento e disponibilidade da informação, possam carregar informações explicativas para retornos futuros.

Como o modelo busca confirmar a hipótese de que o movimento intradiário do preço (e outras variáveis) pode conter informações acerca de retornos futuros, nós precisamos de um método facilitador para o modelo conseguir comparar milhares de dados e classificá-los. Para tal, parece fazer sentido a parametrização das curvas de preços e demais variáveis, de forma a reduzirmos o número de variáveis a serem comparadas pelo modelo com a menor perda possível de potencial informação contidas nas curvas das variáveis, assim como operadores fazem diariamente ao buscar padrões de comportamento que possam indicar uma “reversão”, “canal”, “agulhada” ou inúmeros outros formatos de curvas utilizados em análises grafistas para predição de retornos futuros.

A parametrização é realizada nas curvas intradiárias desde a abertura do mercado até o instante da observação a fim de alimentarmos o *Random Forest* com as variáveis de parametrização.

Mais à frente, no capítulo sobre a parametrização das variáveis ficará mais claro quais são as referidas variáveis de parametrização e como são obtidas. Com a base de dados formada e parametrizada, aplicamos o modelo de *Random Forest* com uma janela extensível de dados (começando em 2000 observações) procurando prever a direção do movimento futuro de preços em predições de retornos de 15min no futuro em dados “*out-of-the-bag*”, ou seja, fora de nosso conjunto de dados..

2.1 Definição das Variáveis

De forma a investigarmos a fundo a existência ou não de informações preditivas de retorno (e, conseqüentemente, de preço) é necessária uma base de dados detalhada e rica em informações. Para este fim, plataformas como *Bloomberg* e *Reuters* se mostraram insuficientes, visto que seus dados intradiários são disponibilizados apenas para curtas janelas temporais. Além disso, a fim de evitarmos pré-tratamento dos dados, foi preferida a análise a partir dos próprios registros de negociação (“negs”), sendo baixados os arquivos contendo todos os *ticks* de negociação via *site* da Bovespa¹.

Filtramos os arquivos a serem baixados, selecionando do diretório FTP (*File Transfer Protocol*) apenas aqueles que contém negociações efetivas do ativo objeto de lotes inteiros (não parciais), dispensando aqueles que são registros de ofertas de compra e venda e aqueles relativos à negociação de opções do ativo objeto. Baixados os arquivos de negociação, selecionamos apenas o ativo de interesse (VALE3).

Embora haja interesse de se utilizar os dados das ofertas de compra e venda do ativo, assim como na análise de informações contidas nas opções (em especial, volatilidade implícita do ativo objeto), não nos propusemos a trabalhar com os mesmos devido ao enorme tamanho dos arquivos e volume de dados, visto que o estudo foi feito em um computador pessoal e não em uma estação especial de trabalho dedicada e preparada para tal demanda de processamento.

2.2 Tratamento dos dados

Após a aquisição e segregação dos dados, realizamos o tratamento dos mesmos para a parametrização. Para os preços, consolidamos os dados em intervalos de 5 minutos, utilizando a última observação de cada intervalo. Após o tratamento dos dados de

¹ Disponível em: <<ftp://ftp.bmf.com.br/MarketData/Bovespa-Vista/>>.

preços, realizamos os cálculos para as outras variáveis escolhidas, todas em blocos de 5 minutos, sendo os cálculos de volume como o somatório (número de ações ou montante financeiro) negociado no intervalo e os de volatilidade como o desvio padrão da variável estudada (preço ou volume negociados).

Segundo Taylor (2005), 5 minutos é definido como o melhor intervalo intradiário para estimativa de curva de volatilidade. Intervalos muito curtos, embora possíveis, podem apresentar grandes distorções dada uma variação pontual (uma grande negociação ou um intervalo com poucas ou nenhuma negociação por qualquer motivo). Por outro lado, intervalos maiores, como 10, 20 ou 30 minutos representariam poucos pontos no dia para termos uma boa parametrização da volatilidade intradiária, e amortizariam grande parte da variação do preço, fazendo com que muita informação fosse perdida no processo.

Desta forma, optamos por intervalos de 5 minutos por entendermos que apresenta um bom grau de conservação das variações intradiárias sem estar extremamente sujeito a variações bruscas para ativos muito líquidos dadas movimentações pontuais. Não deixamos de observar, no entanto, que a otimização do intervalo escolhido, ou mesmo do modelo (por exemplo, a aplicação de um EWMA) é possível, mas não é o objetivo deste estudo.

Uma etapa importante do processo é a normalização das variáveis na primeira observação. Desta forma, não nos preocupamos com intercepto nos regressores, assim como com a fase nas análises espectrais, da mesma maneira que não teremos mais problemas com a diferença de nível dentre os dias analisados, facilitando o aprendizado computacional. Assim, daqui para frente, ao tratarmos preço, volatilidade e volume financeiro, na realidade estaremos analisando os ganhos acumulados intradiários destas variáveis, voltando a zero em todo início de novo dia.

O preço desta manipulação é perdermos a informação de nível das variáveis, que não nos parece ser de grande influência (desde que o valor de seu lote não inviabilize a liquidez). Assim, entenderemos as variáveis temporais como a variação percentual acumulada das mesmas ao longo do dia relativas à primeira observação, como demonstrado na fórmula a seguir:

$$X_t = \frac{x_t}{x_0} - 1$$

Caso seja válida a hipótese de que os níveis de preço, volatilidade e volume financeiro possam trazer informações explicativas para o retorno posterior, nosso modelo não será capaz de captar estes sinais.

Por outro lado, considerando os objetivos deste estudo, há vantagem em padronizarmos os inícios das curvas a fim de tornar as variáveis obtidas nas parametrizações mais comparáveis para o método de aprendizado computacional da metodologia de florestas randômicas, que será explicada no capítulo de apresentação do modelo.

2.3 Parametrização das variáveis

Posterior à normalização dos dados, é chegado o momento da parametrização dos mesmos. Para tal, foi escolhido o modelo de somatório de senóides, com o entendimento de que se adequa bem à maior gama possível de curvas.

A parametrização consiste em um somatório de senóides contendo as 7 principais frequências identificadas, assim como suas amplitudes correspondentes obtidas através de regressão computacional. Entendemos, então, nossas variáveis com o seguinte formato:

$$x_t = \sum_{k=1}^7 a_k \times \{\sin(2\pi \times \varphi_k \times t) + \cos(2\pi \times \varphi_k \times t)\} + \varepsilon_t$$

onde:

φ_k : frequência k

a_k : amplitude k

ε_t : erro da parametrização

Com os resultados das análises espectrais e as amplitudes obtidas através de regressões, analisamos se nossa capacidade explicativa das curvas (in-sample) é adequada. Para isto, analisaremos os resultados de R^2 das nossas regressões, a fim de entendermos nossa capacidade explicativa de variância da parametrização. Para amostras “in-sample”, obtivemos as seguintes distribuições de R^2 ajustado para o método descrito:

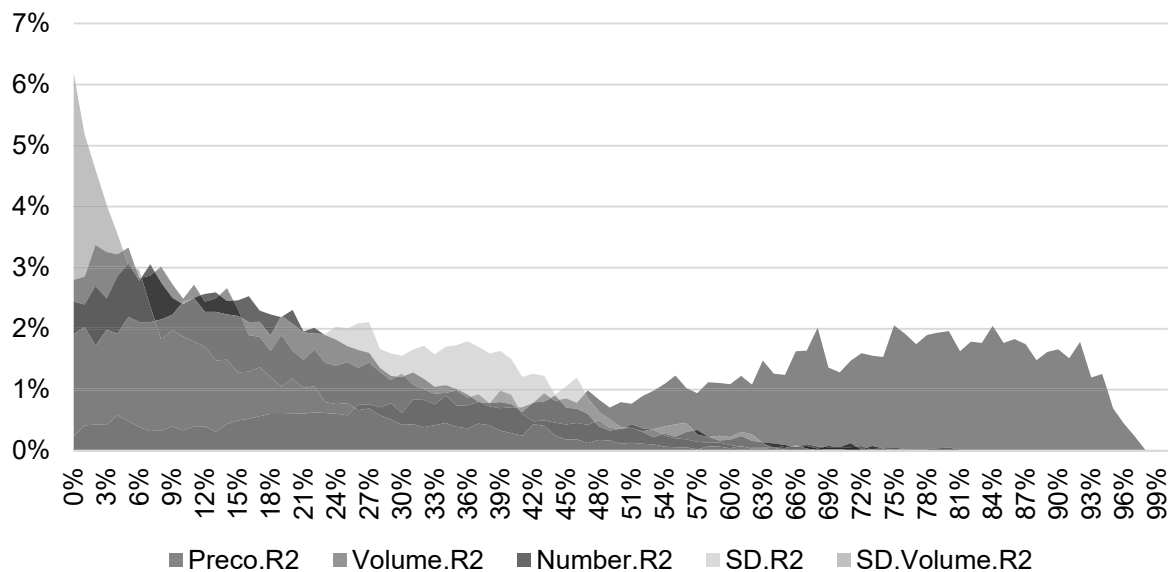


Gráfico 1: Distribuição de R^2 ajustado de diferentes fatores parametrizados.

Fonte: Autoria própria.

Observamos através deste gráfico uma parametrização satisfatória das curvas de preços, mas não das outras variáveis. A partir do gráfico a seguir, para a curva de preços, podemos apontar que temos R^2 ajustado superior a 50% em mais de dois terços da amostra (67,45%), metade de nossa amostra possui R^2 ajustado superior a 66% e mais de um quarto (25,66%) de nossa amostra possui R^2 ajustado igual ou superior a 80%.

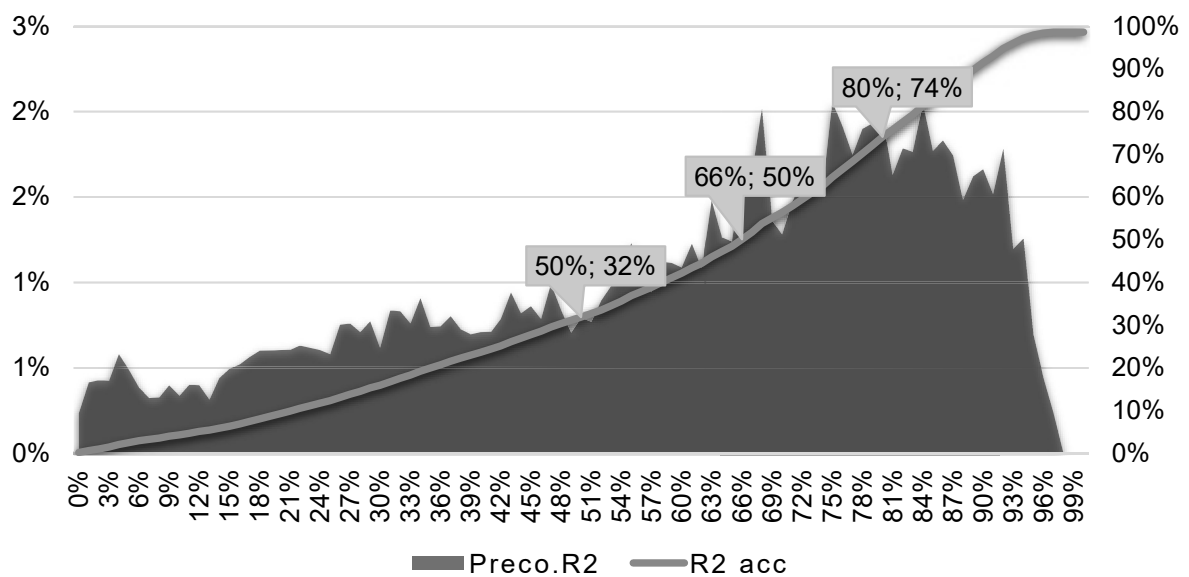


Gráfico 2: Distribuição de R^2 da parametrização dos preços

Fonte: Autoria própria.

Acreditamos, então, que estes 14 fatores (7 frequências e 7 amplitudes) contém significativa capacidade explicativa da variância de nossa amostra ao menos na curva da variável “Preços”. Adicional a estes 14 fatores para cada variável escolhida, adicionamos o próprio R^2 da regressão, tentando trazer à base de dados do modelo *Random Forest* um valor que indique o quão bem-sucedida é nossa parametrização da curva da variável até aquele momento. A forma como o modelo tratará as variáveis de parametrização e através dela retornará os resultados ficará mais clara no capítulo a seguir.

2.4 Apresentação do modelo *Random Forest*

Random forests são a combinação de árvores preditivas de forma que cada árvore dependa de valores de um vetor randômico independente e com mesma dimensão de todas as árvores (Breiman, 1999, p. 01, tradução livre, grifo nosso).

Antes de abordarmos a metodologia do modelo matemático *Random Forest*, vamos antes introduzir o conceito de árvores de decisão, crucial para nosso estudo.

A estrutura de uma árvore decisória consiste em uma série de perguntas/testes acerca de variáveis exógenas, criando as ramificações ou os “galhos” da árvore. Eventualmente, após todas as perguntas da árvore respondidas, chega-se a uma resposta. Um exemplo é a sátira a seguir:

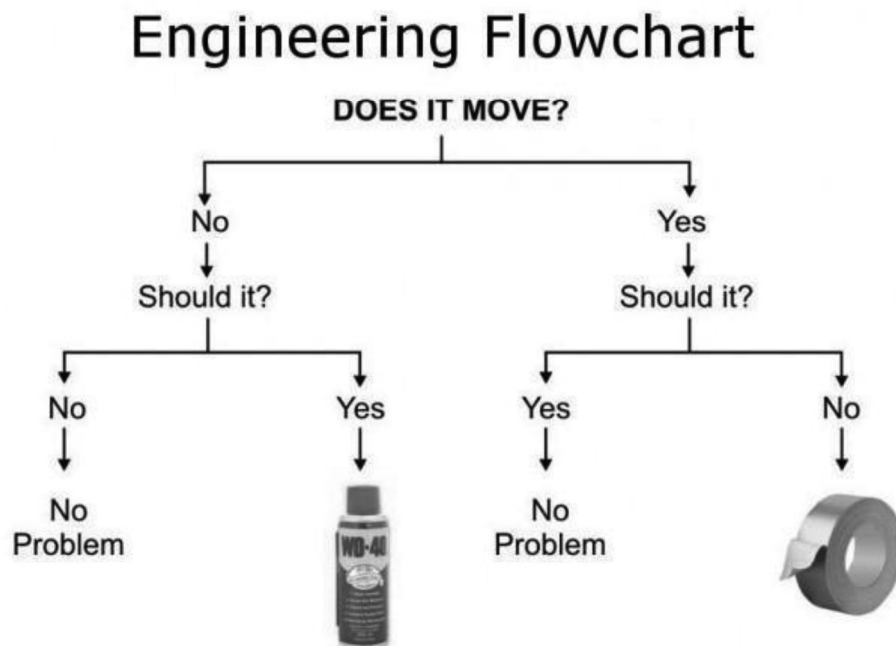


Gráfico 3: Exemplo de árvore decisória

Fonte: USA Naval Academy

(<https://www.usna.edu/Users/cs/aviv/classes/ic312/f16/units/04/unit.html>)

No exemplo acima, temos duas perguntas/testes: a primeira questionando se algo se move, e a segunda perguntando se o mesmo objeto deveria se mover. Caso as respostas sejam iguais (ambas positivas ou ambas negativas), a árvore retorna a resposta “No Problem”, ou seja, o objeto em questão está funcionando como esperado. No entanto, caso algo se mova quando não deve ou não se mova quando deve, a árvore apresenta duas diferentes soluções, fita adesiva e WD-40 (um lubrificante industrial conhecido), respectivamente. Ou seja, dependendo das variáveis exógenas (no exemplo, o fato de o objeto se mover ou não e se deveria se mover ou não) podemos chegar a uma conclusão sobre uma variável endógena (não há problemas, devo colar com fita ou devo aplicar WD-40).

No processamento do *Random Forest* (traduzindo, “florestas aleatórias”), diversas árvores de decisão são criadas com o intuito de responder ao mesmo problema,

porém com subconjuntos distintos de variáveis explicativas, selecionadas de forma randômica.

No caso de variáveis numéricas, este teste é realizado de forma que a divisão dos dados da amostra de aprendizado seja máxima para cada galho seguinte, ou seja, se tratando de uma variável numérica, como preços ou volatilidade, a pergunta será “É maior que um valor X?” ou “A variável se encontra no intervalo entre X e Y?” tal que X e Y são os valores que dividem a amostra da maneira mais igualitária possível entre o número de galhos a ser formado. O processo se repete até que todos os galhos da árvore tenham sido formados.

Uma floresta aleatória é um classificador que consiste em uma coleção de árvores classificadoras $\{h(x, \Theta_k), k = 1, \dots\}$ onde $\{\Theta_k\}$ são vetores independentes aleatórios identicamente distribuídos e cada árvore fornece um voto para a classe mais popular dado o "input" x . (Breiman, 2001, p. 06, tradução livre, grifo nosso).

Neste trabalho, o fenômeno a ser estudado pelo *Random Forest*, trata-se das variações intradiárias de 15 minutos de preços do ativo VALE3, buscando encontrar uma estratégia para prever como tais variáveis explicativas interferem nas oscilações de preço deste ativo.

De forma simplificada, o modelo conta com um conjunto grande de árvores de decisão, com cada uma, ao final de seus processos decisórios, contribuindo com um voto para uma classificação (neste caso, valorização ou desvalorização do preço do ativo) da variável endógena.

Desta maneira, dado qualquer subconjunto de variáveis exógenas acrescentadas futuramente ao modelo (dados “out-of-the-bag”), cada árvore terá uma classificação (valorizará ou depreciará) para o mesmo. Ou seja, ao final da formação de todas as árvores, cada árvore conseguirá classificar qualquer vetor de variáveis exógenas usado para prever a variável a ser explicada, de forma a apresentar um “voto” final de em qual categoria (apreciação ou depreciação) a variável endógena estará. A classificação com o maior percentual de votos será a vencedora e a previsão final do modelo.

A seguir podemos observar de forma gráfica o processo explicado:

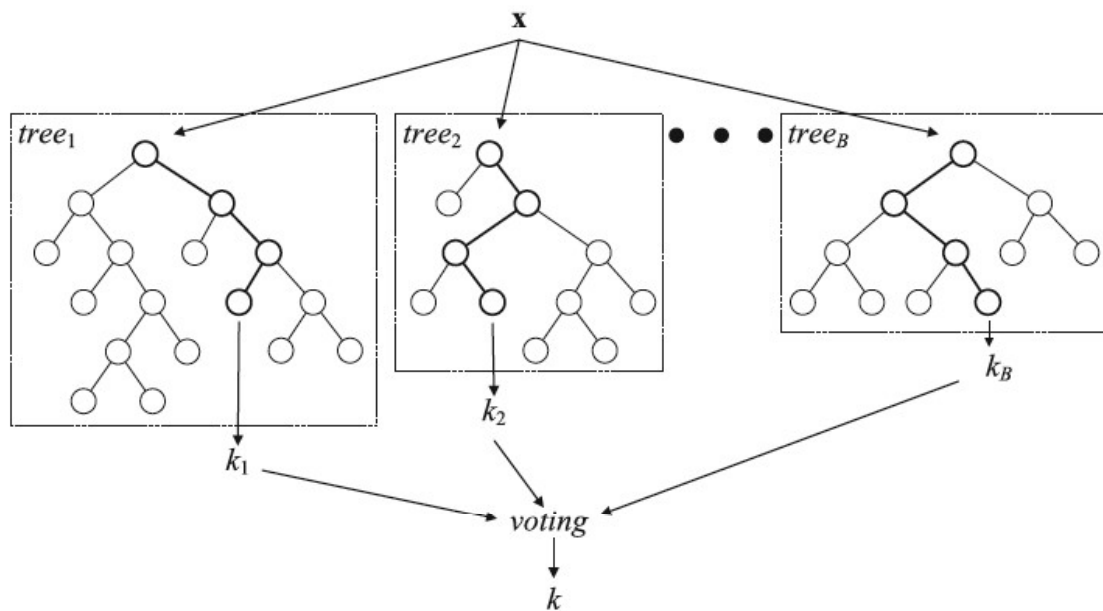


Gráfico 4: Ilustração de construção de árvores do modelo Random Forest

Fonte: VERIKAS, A.; GELZINIS, A. (2011).

Vale observarmos que não é interessante o estudo de uma árvore específica. A ideia é que sequências de testes (lembrando que as sequências são sorteadas aleatoriamente entre as árvores) acabem por criar uma quantidade de testes tal que, se houver alguma informação relevante à predição do evento estudado, ela se destaque e a árvore apresente um percentual de acertos elevado em suas predições.

Um detalhe ainda não abordado acerca do modelo é o intervalo de treino. Até agora, quando tratamos de amostra de aprendizado, estamos nos referenciando a um conjunto de dados separado pelo algoritmo *Random Forest* da base de dados histórica fornecida para a criação dos galhos de decisão, como explicado anteriormente. A amostra de aprendizado é formada apenas por uma parcela dos dados (“in-sample”) fornecidos. O restante dos dados (e não vou me alongar na proporção da divisão entre as duas amostras) se trata do intervalo de treino do modelo.

O intervalo de treino serve para, uma vez criadas as árvores, testá-las em um conjunto de dados não utilizados no aprendizado, de forma a identificarmos o percentual de acerto de cada árvore individual. Este passo será importante para a atribuição de pesos de votações de cada árvore a fim de um resultado geral do modelo, com árvores mais bem-sucedidas em acertos tendo maior peso nos votos e árvores pior sucedidas

tendo menor peso. Assim, teremos maior sensibilidade na escolha de classificação final (apreciação ou depreciação da VALE3) a árvores que performaram bem na fase de treino, e menor para aquelas que acertaram pouco, tendo assim menor ruído e maior acurácia na previsão.

Uma crítica que pode ser levantada a respeito da robustez do modelo é um problema comum quando trabalhamos com regressões e modelagens econométricas de forma geral: o excesso de variáveis explicativas levando a uma correlação entre as mesmas e a um “overfitting” do modelo.

Em seu artigo, Breiman (2001, p. 7) demonstra que não há o problema de “overfitting” na modelagem matemática. Segundo o mesmo, pela Lei Forte dos Grandes Números, podemos chegar à fórmula da probabilidade de erro como:

$$P_{X,Y}(P_{\Theta}(h(X, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(X, \Theta) = j) < 0).$$

Breiman (2001) Teorema 1.2, página 7.

Ou seja, desde que a distribuição de Θ seja independente e idêntica, o modelo não deverá apresentar distorções conforme utilizamos mais árvores ou uma base de dados (X) elevada, apenas aumentando a probabilidade de erro conforme é aumentado o número de opções de classificações (J), que em nossa pesquisa será 2.

3 RESULTADOS

A seguir apresentaremos os resultados obtidos através das predições, assim como diferentes visões e métricas para a mensuração do eventual sucesso ou fracasso dos modelos em suas diferentes bases de dados. Realizaremos testes para tentarmos encontrar fatores que indiquem aprendizado do modelo, com objetivo final do melhor entendimento do algoritmo, suas limitações e em última instância da própria eficiência de mercado. Os resultados não significativos estatisticamente estão com fonte itálica em cor cinza, enquanto que aqueles que satisfazem significância estatística de 95% estão com fonte normal em cor preta nas tabelas.

3.1 Resultados da Previsão

Fazemos abaixo um demonstrativo do percentual de previsão da estratégia para retornos diferentes de zero. Em outras palavras, os dados abaixo representam, para cada frequência analisada e para cada espaço amostral, qual a proporção de acertos da estratégia quando o retorno realizado não foi nulo.

Os resultados abaixo apresentam a proporção de vezes que o modelo previu corretamente a valorização ou desvalorização do ativo dada a amostra de retornos não nulos. As linhas marcadas como “Confiança (95%)” definem o limite inferior do intervalo de confiança (com significância estatística de 95%) para a média apresentada. Se assumirmos a hipótese de passeio aleatório dos preços, não deverá ser possível atingir, com significância estatística, um percentual de acertos diferente de 50%. Caso nossa modelagem supere esta marca, significa que o histórico das variáveis tratadas carrega caráter explicativo para retornos futuro, assim como a extração destes sinais se deu, ao menos parcialmente, de forma bem-sucedida.

Vale observarmos na tabela a seguir que a classificação dos resultados por amostra de dados significa que, dada a metodologia de janela expansível, os resultados apresentados contaram com janelas de dados compreendidos naquele intervalo, sempre prevendo retornos “*out-of-the-sample*”. Ou seja, uma janela de 100 a 200 quer

dizer que a primeira previsão contou com uma janela de dados de 100 observações, a segunda 101 e assim por diante até a centésima observação, que contou com uma amostra de dados de 200 observações. Traçando um paralelo, segue a mesma lógica que uma “*rolling regression*”, para os familiarizados.

Tabela 1: Percentual de acertos do modelo

<i>Amostra de dados:</i>	<i>2000 a 21099</i>	<i>19099 a 21099</i>	<i>2099 a 21099</i>
<i>Média (Preços)</i>	50.05%	50.15%	48.90%
<i>Confiança (95%)</i>	49.34%	47.96%	45.80%
<i>Média</i>	51.05%	53.85%	54.10%
<i>Confiança (95%)</i>	50.34%	51.66%	51.01%

Fonte: Autoria própria.

Podemos observar pelas tabelas que a proporção de acertos fora da amostra tende a aumentar conforme rodamos previsões apenas em janelas de dados maiores. Entendemos assim que, conforme avançamos no tempo, nossa previsão se torna cada vez mais acurada dado o aumento da janela de dados (que é expansível, e não fixa), e podemos esperar retornos mais expressivos.

Para reforçar esta hipótese, observamos abaixo a evolução da proporção de acertos conforme aumentamos a janela de dados. O caso analisado primeiramente é o da previsão de retornos de 15 minutos utilizando todos os dados para alimentação do modelo (não somente os preços). Nos anexos apresentamos graficamente os retornos para outros intervalos de previsão.

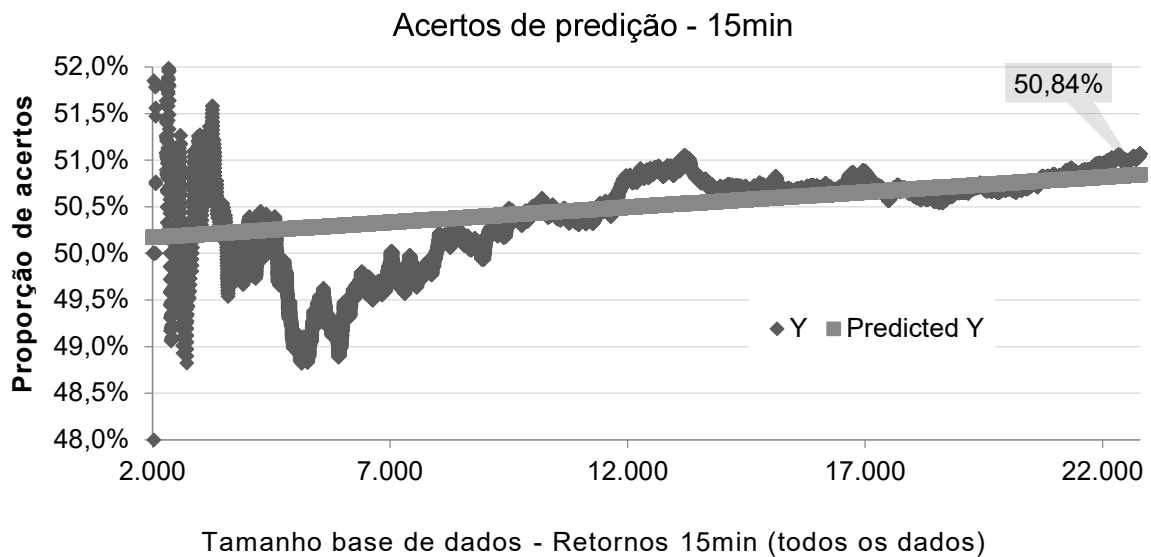


Gráfico 5: Resultados para regressões em toda a amostra.

Fonte: Autoria própria.

Como podemos observar no gráfico acima, temos um ganho médio de previsibilidade a cada amostra. Caso fosse verdadeira a hipótese de passeio aleatório dos preços, passado um pequeno período de ruído, a média de acertos do modelo convergiria para 50% e oscilaria pouco em torno desta marca, com amplitude decrescente dado o maior número de dados. Não é o que observamos nos nossos resultados.

Se rodarmos a regressão de ganho de previsibilidade por número de elementos no intervalo de treino percebemos um múltiplo positivo e relevante para todos os níveis de significância estatística, como apresentado abaixo.

Tabela 2: Resultados do percentual de acertos por intervalo de dados

	<i>Coeficiente</i>	<i>Desvio Padrão</i>	<i>p-valor</i>
<i>X Variable</i>	3,24172E-07	1,29476E-08	2,5056E-136

Fonte: Autoria própria.

Podemos também observar, pela inclinação final da curva, que o modelo apresenta ganho significativo de previsibilidade no final da janela de teste, assim como mostravam nossos resultados regredidos nas últimas 2000 e 1000 observações. A seguir vamos observar com mais atenção estes intervalos amostrais, analisando a média móvel das últimas 1000 observações, a título de evitarmos muito ruído nas observações.

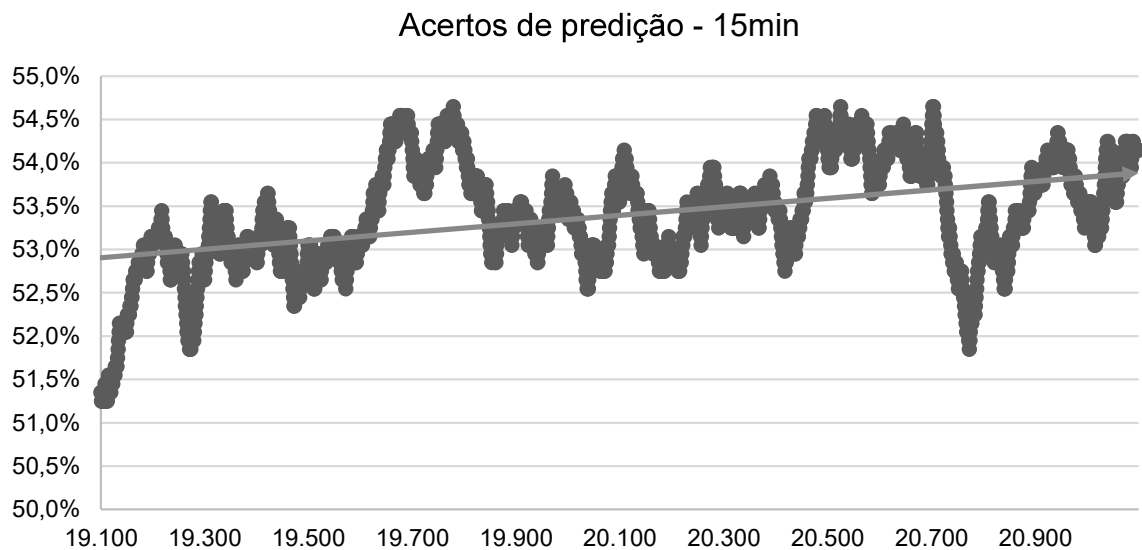


Gráfico 6: Média móvel (1000obs) de acertos (base de dados >19.000).

Fonte: Autoria própria.

De forma geral, obtivemos maior caráter preditivo ainda significativo ao analisar de forma separada os últimos 2000 elementos do resultado. Este ganho de previsibilidade pode ser dado pelo mencionado fato de utilizarmos janelas cada vez maiores de dados para o modelo, de forma que a última previsão têm como base de treino todo o histórico realizado até aquele dia, enquanto que a primeira previsão tem apenas as observações do intervalo mínimo de treino, definido como 2000 observações.

3.2 Probit e Logit

Ao tratarmos de previsão de modelos binários (acerto ou falha), nos vem à mente o estudo da probabilidade de acerto do resultado, e o teste do aumento ou diminuição da mesma conforme variação na amostra.

Neste capítulo trataremos duas possíveis distribuições probabilísticas que se adequariam ao nosso modelo de previsão, cada uma com suas peculiaridades e hipóteses de melhor adaptação ao modelo.

O primeiro modelo tratado será o Probit, que considera a curva normal para distribuição de probabilidades. Isto equivale a dizer que há um ponto, diferente de zero em que o ganho incremental de previsibilidade é máximo, ou seja, o modelo terá ganho marginal maior em certo ponto, com decrescimento de ganho marginal de probabilidade de acerto conforme nos afastamos do “intervalo ideal”, com redução de probabilidade de forma simétrica.

De maneira distinta da distribuição normal, a distribuição log-normal não pressupõe simetria em sua distribuição com relação ao ponto máximo. Sendo assim, o modelo adequa melhor à ideia de que os ganhos de previsibilidade nos primeiros dados inseridos serão superiores aos posteriores após um intervalo breve de observações, enquanto que preservando uma cauda direita mais pesada com relação à distribuição normal.

Assim, o modelo “logit” se adequará melhor caso nossa distribuição não tenha um declínio de caráter preditivo incremental tão elevado quanto uma distribuição normal sugeriria. O gráfico abaixo representa de forma meramente ilustrativa a forma das duas distribuições de probabilidades acompanhadas de suas curvas de probabilidades acumuladas.

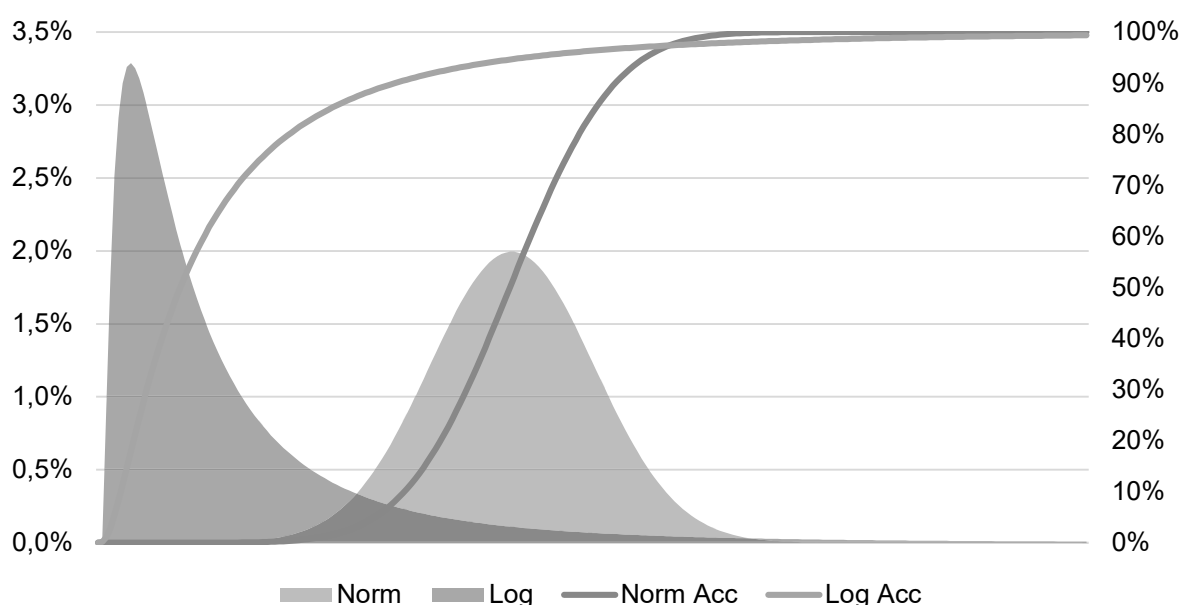


Gráfico 7: Representação ilustrada de distribuições de probabilidades

Fonte: Autoria própria.

Primeira a ser tratada, a modelagem via “Probit” considera para a regressão uma distribuição acumulada normal, na qual o regressor será um múltiplo das variáveis explicativas escolhidas que, quando calculadas no modelo, tentarão explicar o quão sensível a probabilidade de ocorrência do evento estudado é em relação àquelas variáveis. Abaixo, podemos observar os resultados (demonstrados de forma gráfica) da regressão da probabilidade de acerto do modelo com o número de elementos na base de dados.

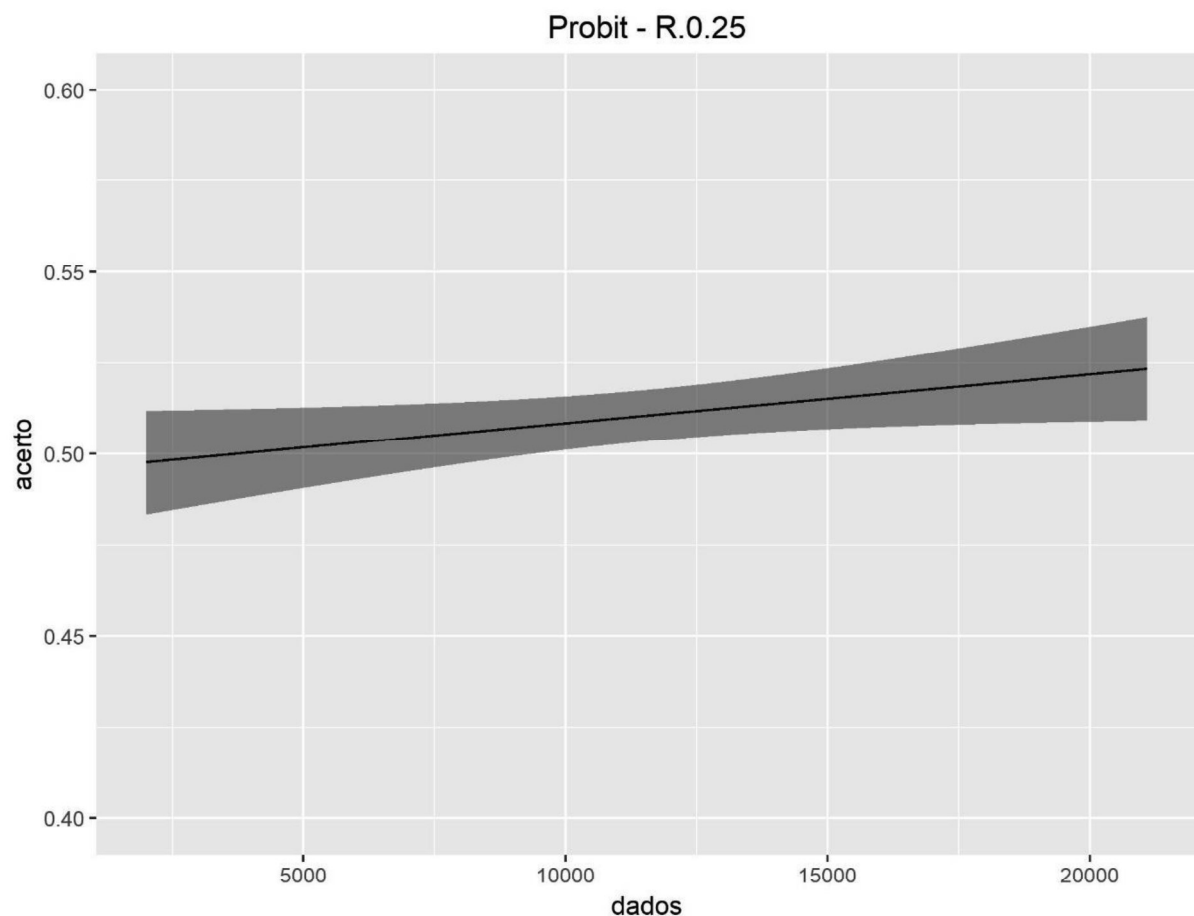


Gráfico 8: Distribuições via *Probit* de acertos de sinal de retornos

Fonte: Autoria própria.

A linha negra representa a probabilidade esperada de acerto dado o valor utilizado de intervalo de dados em cada janela de predição e a área sombreada representa o intervalo de confiança com 95% de significância estatística das previsões. À exceção das previsões de 30 minutos, podemos observar uma clara evolução de caráter preditivo com o aumento da amostra de dados, indicando o aperfeiçoamento da

previsão do algoritmo conforme mais dados são fornecidos, de forma simples, podemos dizer que o modelo está “aprendendo”.

Além disso, como podemos observar no gráfico, temos zonas de intervalo de confiança completamente acima de 50% a partir de aproximadamente 10.000 observações no intervalo de dados. Isto mostra que existem intervalos em que refutamos com 95% de confiabilidade estatística através do modelo Probit a hipótese de passeio aleatório dos preços.

A seguir, analisamos os resultados via método de modelagem log-normal:

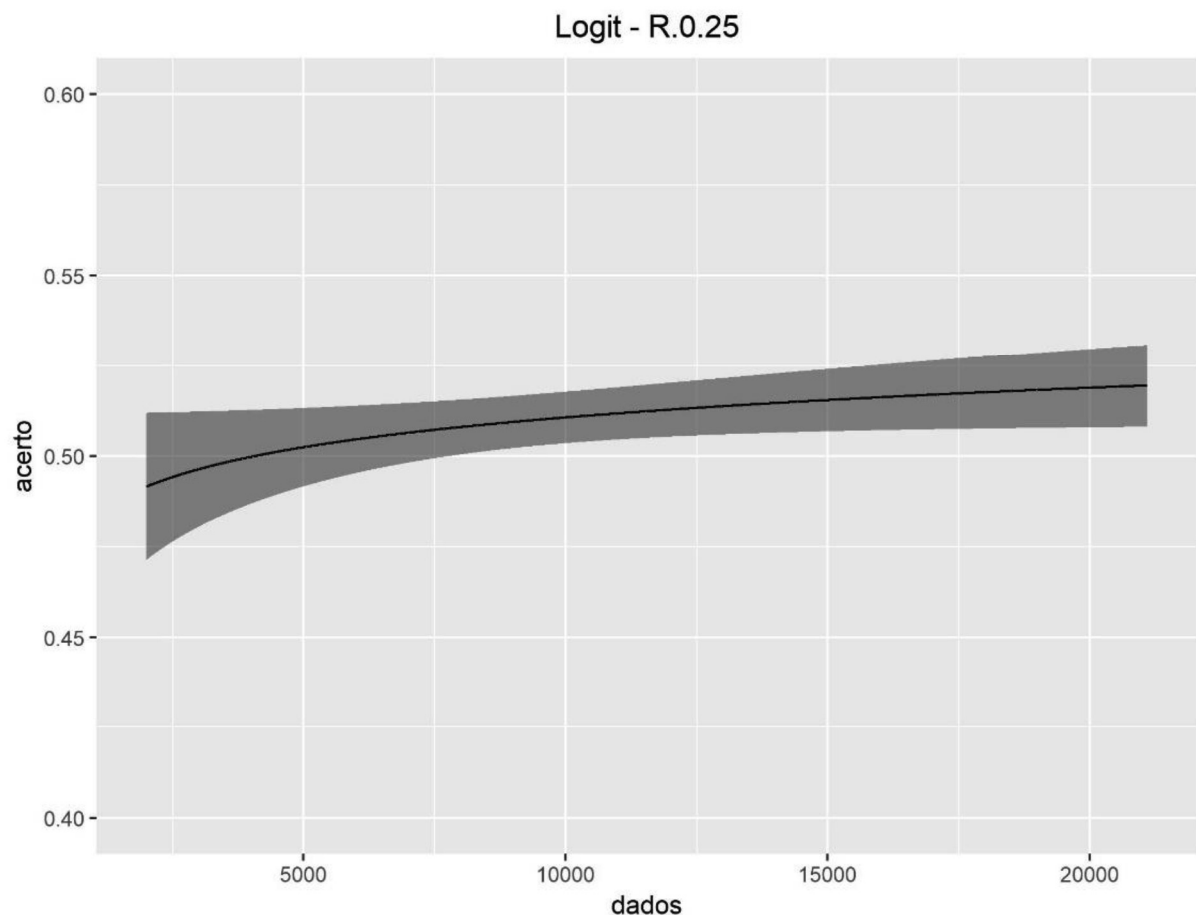


Gráfico 9: Distribuições via *Logit* de acertos de sinal de retornos

Fonte: Autoria própria.

Novamente, podemos observar intervalos de confiança fora da linha de 50%, representando assim intervalos em que refutamos a hipótese nula de passeio aleatório dos preços.

Além disso, é válido observarmos maior concavidade, que acaba por trazer um formato mais páreo ao esperado no sentido de ganho marginal preditivo decrescente com o aumento da base de dados. No entanto, não observamos intervalos de confiança nitidamente menores a ponto de destacarmos este modelo como claramente superior ao probit.

Abaixo observaremos os resultados das regressões por ambos os modelos, com os desvios padrões e os respectivos p-valores indicando a significância com que refutamos (caso refutemos) a hipótese de o regressor ser nulo. Vale ressaltarmos que por se tratarem de distribuições distintas, não podemos comparar os regressores diretamente, no sentido de compararmos a “sensibilidade” ao aumento da base de dados.

Tabela 3: Coeficientes, desvio padrão e teste de hipótese Probit e Logit

	<i>Coef</i>	<i>Std Dev</i>	<i>Pr(Coef≤0)</i>
<i>Probit</i>	0,003	0,002	2,01%
<i>Logit</i>	0,047	0,024	2,52%

Fonte: Autoria própria.

Como mencionado, obtivemos intervalos de confiança acima de 50% para ambas as modelagens probabilísticas, sem clara distinção na adequabilidade de uma ou outra para a real curva de probabilidade. Uma explicação para tais resultados é a que ao estarmos na cauda direita das distribuições de probabilidades de acerto, pode não haver muita discrepância nas localidades das distribuições, sendo assim indiferente a utilização de uma ou outra modelagem.

Obviamente não podemos esperar o mesmo ganho por amostra de dados conforme aumentarmos sua dimensão indefinidamente, pelo simples fato que, se fosse verdade, uma base de dados suficientemente grande nos garantiria um acerto, o que é absurdo. Temos então a abertura de uma nova frente de estudos, que não será abordada neste trabalho, que é a análise da concavidade do ganho de previsibilidade dado o incremento amostral para bases de dados substancialmente grandes.

3.3 Testes do modelo sem parametrização dos dados

A seguir realizaremos os mesmos testes sem o pré-processamento elaborado dos dados. Utilizaremos como variáveis exógenas preditivas de retornos uma janela móvel de preços (normalizados e fora de nível) de dimensão 25 e espaçados 5 minutos entre si, ou seja, o modelo tentará prever o retorno futuro de 15 minutos considerando o histórico de preços de 2h e 5 minutos.

Desta vez, não haverá a parametrização da curva intradiária, assim como não utilizaremos os dados de todo o dia (a não ser o preço de abertura, que normalizará a curva de preços diária) nem outros dados de negociação, como volume e volatilidade.

O objetivo deste teste é perceber se o modelo consegue identificar padrões e prever retornos com maior, menor ou nenhuma eficácia comparada aos resultados obtidos com a base de dados completa e parametrizada. A ideia é investigarmos as limitações do algoritmo e entendermos que tipo de base de dados se adequa melhor ao modelo, visando o aperfeiçoamento do código e técnicas de previsão de retornos via *Random Forest* para estudos futuros.

Se pretendemos testar a robustez do modelo, desafiaremos o algoritmo a encontrar correlações entre movimentos de preços e retorno futuro sem a “ajuda” da parametrização e inserção de dados derivados da negociação. Caso encontremos resultados estatisticamente significativos e com maior caráter preditivo que o testado anteriormente significará que nosso tratamento de dados para alimentação do modelo acaba por atrapalhar seus cálculos e devemos priorizar a inserção de dados “crus” e deixar que o próprio algoritmo identifique padrões sem nossa ajuda. Caso o modelo falhe em apresentar resultados superiores aos anteriormente apresentados, interpretamos que o modelo se beneficia com informações mais completas e trabalhadas para suas correlações nas árvores de escolha e que é interessante a investigação de novos dados derivados das negociações que possam ser inseridos em sua base de dados.

A seguir, então, observamos os resultados dos cálculos sem o tratamento das variáveis:

Tabela 4: Resultados do modelo sem parametrização dos dados

<i>Amostra de dados:</i>	<i>2000 a 21099</i>	<i>19099 a 21099</i>	<i>2099 a 21099</i>
<i>Média (Preços)</i>	46.67%	47.75%	48.60%
<i>Confiança (95%)</i>	46.00%	45.56%	45.50%

Fonte: Autoria própria.

O modelo com a base de dados reduzida e sem pré-processamento performa de forma distintamente pior.

Pelos resultados das previsões, podemos inferir que o modelo se beneficia de mais dados, e dados que carreguem mais informações. Ao realizamos os cálculos somente com observações dos preços (normalizados pelo ao preço de abertura) a cada 5 minutos, acabamos por dispensar de nossa base dados informações que possam ter relevância na análise de parâmetros, como retornos mais antigos que 2h e 5 minutos, volatilidade, número de negociações, volume financeiro e outras informações quaisquer dispensadas.

A conclusão que chegamos com este exercício é que, ao que parece, não há prejuízo observado em um tratamento dos dados e inserção de novas variáveis de mercado.

4 FUTUROS PONTOS DE INVESTIGAÇÃO

A partir dos resultados, nos indagamos de que forma poderíamos trazer mais informações para o modelo e de quais formas poderíamos aumentar a capacidade preditiva da análise.

Algumas possibilidades de investigação são apresentadas a seguir, abrindo diálogo para discussões a respeito da quantidade e qualidade de fator explicativo de retorno que estes parâmetros carregam, assim como sua facilidade de acesso e praticidade operacional. Não temos, no entanto, a intenção de nos aprofundarmos nas discussões a seguir, mas apenas de levantarmos ideias para pontos de partida de novos estudos buscando o aperfeiçoamento do algoritmo e a compreensão de funcionamento dos mercados.

Vale ressaltar que os pontos destacados abaixo não são restritos à previsão de retornos intradiários das ações da VALE negociadas na Bolsa de Valores de São Paulo, objeto de estudo da pesquisa, mas também à investigação da hipótese de passeio aleatório dos preços de forma geral e, em última análise, à própria hipótese de eficiência dos mercados.

4.1 Informações de ofertas de negociação

Para o estudo realizado, observamos apenas informações de negociações realizadas, ou seja, quando algum preço de oferta de compra foi igual ou superior a um preço de oferta de venda. No entanto, a maior parte das ofertas registradas na bolsa não é realizada, assim, é possível imaginarmos que estas ofertas carreguem informações acerca das expectativas de variação de preços dos agentes do mercado.

Exemplificando tal situação, Paital e Sharma (2016) encontraram evidências de que o “*bid-ask spread*” (a diferença entre preços de ofertas de compra e venda) tem uma

correlação positiva com retornos futuros, o que foi interpretado por eles como uma indicação de captação não homogênea de informação pelo mercado.

Dada a quantidade de dados que estes arquivos representam e a capacidade computacional disponível, não incluímos tais dados nas análises, porém permanece a indagação se há a possibilidade de extração de caráter preditivo de retornos baseado na expectativa dos agentes via suas ofertas de negociação. Talvez métricas simples de distribuição de preços e volume ofertados como média, mediana, desvio padrão, curtose, médias móveis e/ou outras diversas possam servir como indicadores de como os agentes esperam o comportamento da curva de preços.

Para estudos futuros que disponham de melhores recursos computacionais e base de dados ampla, seria ideal um aprofundamento na análise das ofertas de negociação, tanto de compra como de venda, dado que o preço realizado de negociação se dá, necessariamente, neste intervalo.

4.2 Análise da volatilidade implícita nas opções do ativo objeto

No mercado financeiro, para todo ativo com alta liquidez existe uma gama de produtos derivados do mesmo, chamados genericamente de “derivativos”, onde o próprio ativo passa a ser referenciado como “ativo objeto”. A partir dos dados de negociações dos derivativos, conseguimos extrair informações acerca da expectativa dos agentes para o ativo objeto, o que não conseguiríamos somente com a análise de negociações do mesmo.

O derivativo mais popular para ações são as opções, que podem ser de compra ou de venda. Estes produtos dão ao seu detentor a possibilidade de compra ou venda do ativo objeto a um certo preço em uma certa data ou intervalo de datas, a depender do tipo de opção. Pela análise dos preços das opções e dos ativos objeto podemos extrair a chamada “volatilidade implícita”, ou seja, a volatilidade esperada pelo mercado para tal ativo objeto no prazo da opção, caso admitamos uma distribuição normal de probabilidades de retornos para o mesmo.

De maneira simplificada, volatilidades implícitas carregam a informação do quão incerto é o preço futuro do ativo. Quando as opções são negociadas a volatilidades implícitas maiores, significa que o mercado está precificando um futuro mais incerto. De maneira inversa, quão menores forem as volatilidades implícitas maior é a confiança do mercado de que o preço estará próximo do valor esperado.

À priori, na hipótese de passeio aleatório dos preços, a volatilidade implícita deve ser a mesma para opções de compra ou venda e não deve carregar informações acerca do valor esperado do ativo objeto em seu vencimento. No entanto, estudos anteriores (GIOT, 2003, CHEN; CHUNG; TSAI, 2016 e BALI, 2009) indicam que pode haver informação preditiva de retornos nas curvas de volatilidades implícitas dos ativos por eles analisados. Bali, 2009 cita que seu estudo conseguiu evidências de que há uma correlação positiva entre retornos esperados pelos agentes e a diferença entre a volatilidade implícita das opções de compra (call) e de venda (put).

Desta maneira, a inclusão de volatilidade implícita do ativo objeto (neste caso, o ativo VALE3 negociado na Bolsa de Valores de São Paulo) na base de dados do algoritmo nos traz informações acerca da expectativa do mercado sobre a volatilidade dos preços, mas também pode carregar informações quanto às expectativas dos mesmos para o próprio preço do ativo no período, podendo neste caso carregar informações preditivas a serem exploradas pelo algoritmo.

4.3 Análise de diversos ativos simultaneamente

Como último ponto de investigação, podemos levantar a possibilidade da utilização de dados de mercado de diferentes ativos em nossa base de dados, desde que satisfeita a condição de alta liquidez. A ideia é, uma vez que visamos predição de retornos independente de fatores intrínsecos de setores e empresas, podemos utilizar para o aprendizado do modelo o mesmo conjunto de variáveis exógenas para outros ativos com o retorno de seus respectivos ativos.

Assim, uma vez que observamos aumento de caráter preditivo com incremento de base de dados, esta técnica pode aumentar significativamente nossas análises sem estarmos sujeitos à crítica da inserção de dados distantes temporalmente.

5 CONSIDERAÇÕES FINAIS

Através do modelo de Random Forest e parametrização de variáveis extraídas de dados de mercado para a Vale conseguimos obter caráter preditivo estatisticamente significativo para previsões de sinais de retornos intradiários de 15 minutos.

Buscamos entender melhor o funcionamento do aprendizado do modelo, buscando sinais de melhora preditiva com aumento de base de dados. Fomos bem-sucedidos neste ponto, ao encontramos sinais de melhora preditiva de sinais de retorno coincidente com o aumento de base de dados do modelo, como fica claro principalmente ao analisarmos a probabilidade de acerto através das regressões de “*probit*” e “*logit*”.

Posterior à análise dos resultados, rodamos o mesmo modelo com variáveis sem tratamento ou parametrização, a fim de identificarmos limitações do algoritmo e potencialidades de melhora. Pelos resultados, entende-se que o algoritmo aumenta significativamente em caráter preditivo conforme aumentamos a amostra de aprendizado e a qualidade dos dados.

Ao final das análises quantitativas, apresentamos novas ideias de pontos de investigação, a fim de aprofundar o estudo na hipótese de passeio aleatório dos preços e em última instância a própria hipótese de eficiência de mercado.

Os anexos mostram os resultados do mesmo modelo aplicado a retornos de 30 minutos, 1 hora, 2 horas, 3 horas, 4 horas e 5 horas do ativo estudado.

REFERÊNCIAS

- ALI, A.; SILAS, B. **Analysis of the bid-ask spread and its implication for portfolio returns**. 2016. Mestrado (Master of Science in Business). Nord University, Bodø, Noruega, June, 2016.
- BALI, T. G.; HOVAKIMIAN, A. **Volatility Spreads and Expected Stock Returns**. *Management Science*, v. 55, n. 11, p. 1797-1812, 2009. Disponível em: <<https://dx.doi.org/10.2139/ssrn.1029197>>. Acesso em: 08 de dezembro de 2018.
- BREIMAN, L. **Random Forests**. 1999. Statistics Department, University of California Berkeley, Setember, 1999.
- CHEN, T.-F.; CHUNG, S.-L.; TSAI, W.-C. **Option-Implied Equity Risk and the Cross Section of Stock Returns**. *Financial Analysts Journal*, CFA Institute, v. 72, n. 6, 2016. Disponível em: <<https://doi.org/10.2469/faj.v72.n6.2>>. Acesso em: 08 de dezembro de 2018.
- GAÁL, M.; MORIONDO, M.; BINDI, M. **Modelling the impact of climate change on the hungarian wine regions using Random Forest**. *Applied Ecology And Environmental Research*, Budapest, Hungary, v. 10, n. 2, p. 121-140, 2012.
- GIOT, P. **On the relationships between implied volatility indices and stock index returns**. *The Journal of Portfolio Management*, v. 3, n. 3, p. 92-100, April 2005.
- KANDASWAMY, K. K.; CHOU, K.-C.; MARTINETZ, T. MOLLER S., SUGANTHAN P.N., SRIDHARAN S., PUGALENTI G. AFP-Pred: **A random forest approach for predicting antifreeze proteins from sequence-derived properties**. *Journal of Theoretical Biology*, v. 270, n. 1, p. 56-62, November, 2010. DOI: 10.1016/j.jtbi.2010.10.037.
- PAITAL, R. R.; SHARMA, K. N. **Do trading volume and bid-ask spread contain information to predict stock returns? Intraday evidence from India**, Hyderabad, March, 2016.
- STOCK, J. H. **Introduction to Econometrics**. 3a. ed. Pearson Education, Inc, 2014.
- TAYLOR, S. J. **Asset Price Dynamics, Volatility and Prediction**. Princenton University Press, 2005.
- VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. **Mining data with random forests: A survey and results of new tests**, 2011

ANEXO 1 – RESULTADOS DE PREVISÕES DE RETORNOS DE 30 MINUTOS

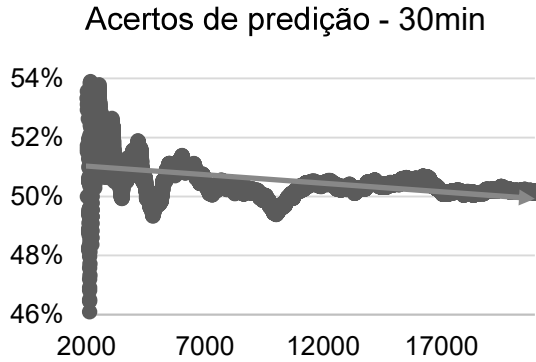


Gráfico 10: Resultados para regressões em toda a amostra.
Fonte: Autoria própria.

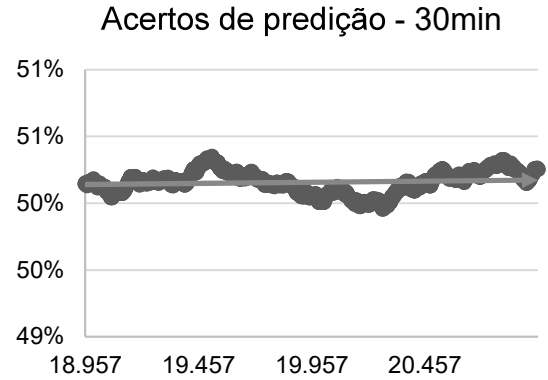


Gráfico 11: Média móvel (1000obs) de acertos por intervalo dados.
Fonte: Autoria própria.

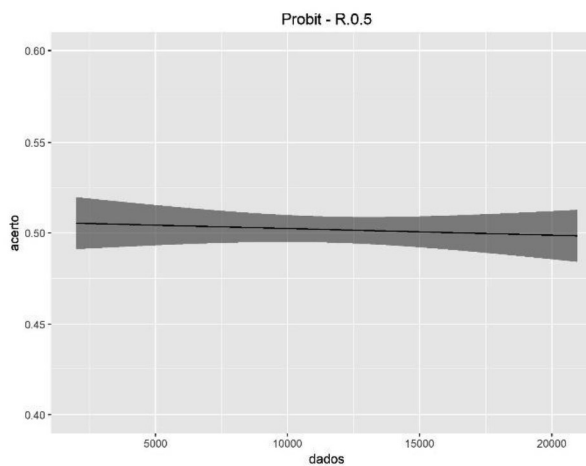


Gráfico 12: Distribuições via Probit de acertos de sinal de retornos
Fonte: Autoria própria.

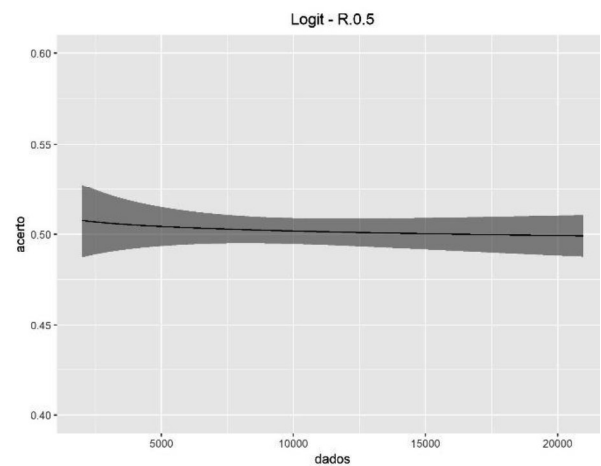


Gráfico 13: Distribuições via Logit de acertos de sinal de retornos
Fonte: Autoria própria.

ANEXO 2 – RESULTADOS DE PREVISÕES DE RETORNOS DE 1 HORA

Tabela 5: Resultados das previsões de retornos 3 horas

<i>Amostra de dados:</i>	<i>2000 a 19267</i>	<i>17268 a 19267</i>	<i>18267 a 19267</i>
<i>Média (Preços)</i>	50.98%	51.90%	52.80%
<i>Confiança (95%)</i>	50.60%	49.71%	49.70%
<i>Média</i>	50.03%	55.40%	57.70%
<i>Confiança (95%)</i>	49.29%	53.22%	54.64%

Fonte: Autoria própria.

ANEXO 3 – RESULTADOS DE PREVISÕES DE RETORNOS DE 1 HORA

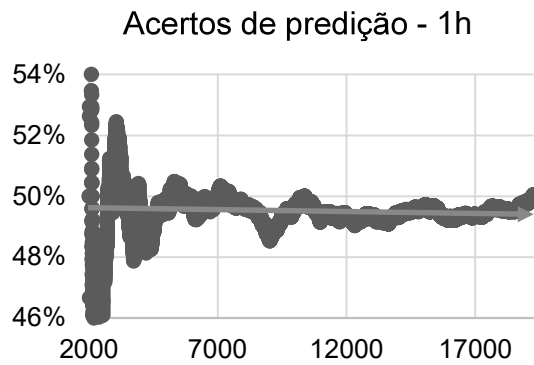


Gráfico 14: Resultados para regressões em toda a amostra.
Fonte: Autoria própria.

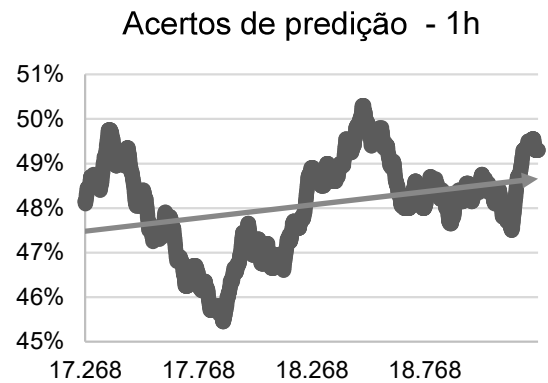


Gráfico 15: Média móvel (1000obs) de acertos por intervalo de dados.
Fonte: Autoria própria.

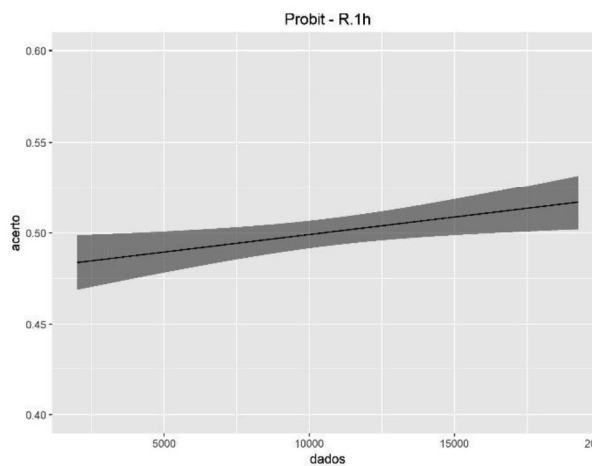


Gráfico 16: Distribuições via Probit de acertos de sinal de retornos
Fonte: Autoria própria.

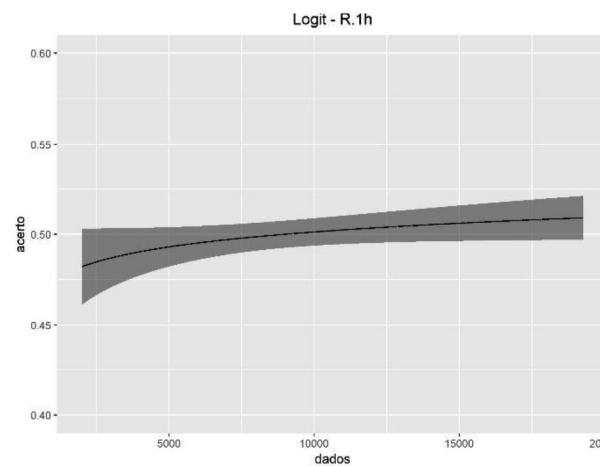


Gráfico 17: Distribuições via Logit de acertos de sinal de retornos
Fonte: Autoria própria.

ANEXO 4 – RESULTADOS DE PREVISÕES DE RETORNOS DE 2 HORAS

Tabela 6: Resultados das previsões de retornos 3 horas

Amostra de aprendizado:	2000 a 15549	13550 a 15549	14550 a 15549
Média (Preços)	50.89%	52.61%	51.71%
Confiança (95%)	50.05%	51.06%	49.92%
Média	50.83%	54.45%	57.50%
Confiança (95%)	49.99%	52.27%	54.43%

Tabela 7: Resultados das previsões de retornos 2 horas

Fonte: Autoria própria.

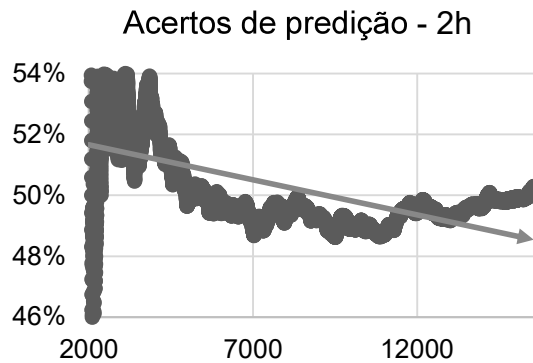


Gráfico 18: Resultados para regressões em toda a amostra.
Fonte: Autoria própria.

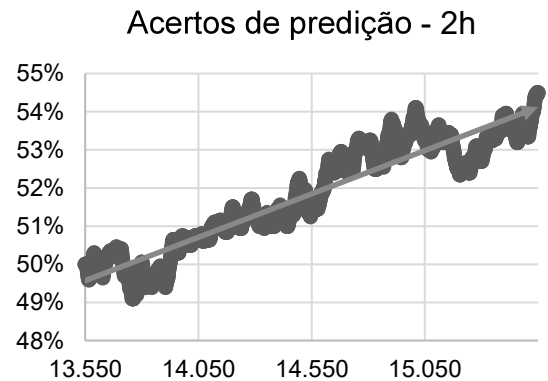


Gráfico 19: Média móvel (1000obs) de acertos por intervalo de dados.
Fonte: Autoria própria.

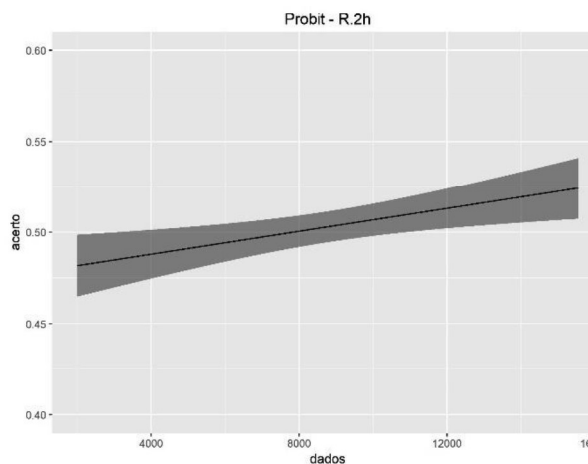


Gráfico 20: Distribuições via Probit de acertos de sinal de retornos
Fonte: Autoria própria.

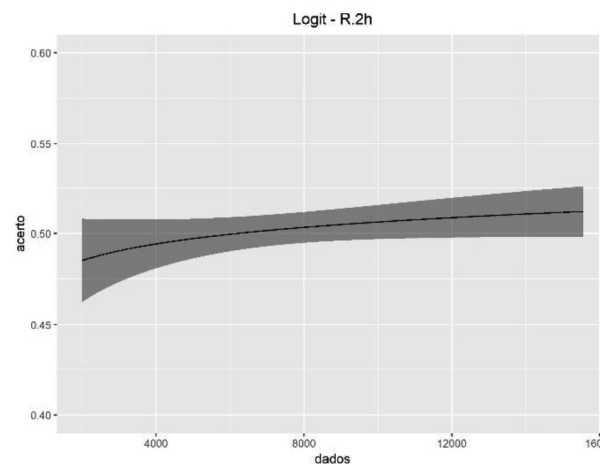


Gráfico 21: Distribuições via Logit de acertos de sinal de retornos
Fonte: Autoria própria.

ANEXO 5 – RESULTADOS DE PREVISÕES DE RETORNOS DE 3 HORAS

Tabela 8: Resultados das previsões de retornos 3 horas

<i>Amostra de aprendizado:</i>	<i>2000 a 11694</i>	<i>9695 a 11694</i>	<i>10695 a 11694</i>
<i>Média (Preços)</i>	51.51%	53.75%	53.10%
<i>Confiança (95%)</i>	50.52%	51.56%	50.00%
<i>Média</i>	52.09%	57.10%	58.20%
<i>Confiança (95%)</i>	51.10%	54.93%	55.14%

Fonte: Autoria própria.

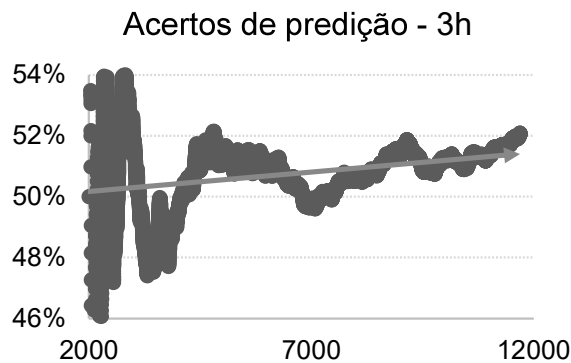


Gráfico 22: Resultados para regressões em toda a amostra.
Fonte: Autoria própria.

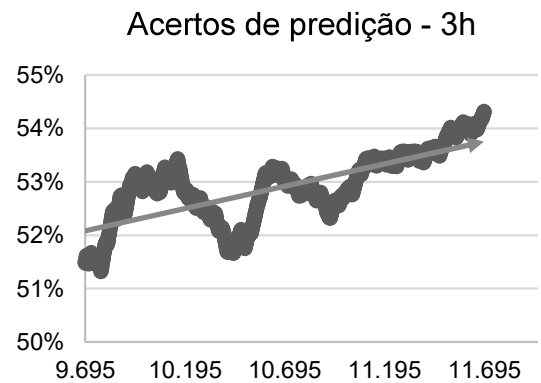


Gráfico 23: Média móvel (1000obs) de acertos por intervalo de dados.
Fonte: Autoria própria.

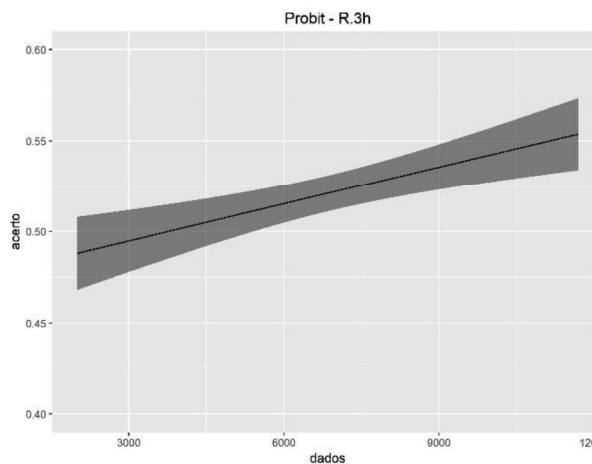


Gráfico 24: Distribuições via Probit de acertos de sinal de retornos
Fonte: Autoria própria.

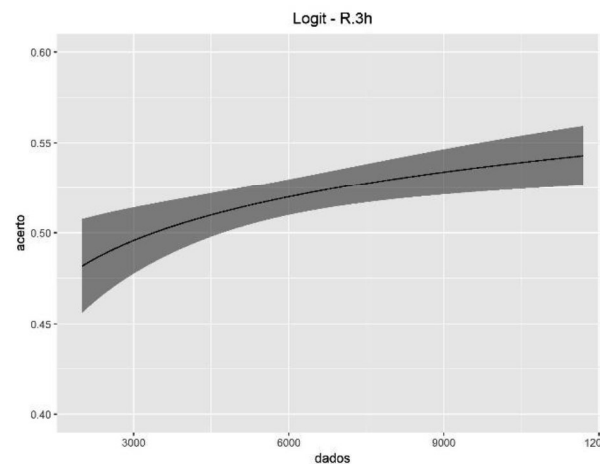


Gráfico 25: Distribuições via Logit de acertos de sinal de retornos
Fonte: Autoria própria.

ANEXO 6 – RESULTADOS DE PREVISÕES DE RETORNOS DE 4 HORAS

Tabela 9: Resultados das previsões de retornos 4 horas

Amostra de dados:	2000 a 7830	5829 a 7830	6829 a 7830
Média (Preços)	50.88%	51.15%	49.20%
Confiança (95%)	49.60%	48.96%	46.10%
Média	52.27%	54.65%	55.50%
Confiança (95%)	50.99%	52.47%	52.41%

Fonte: Autoria própria.

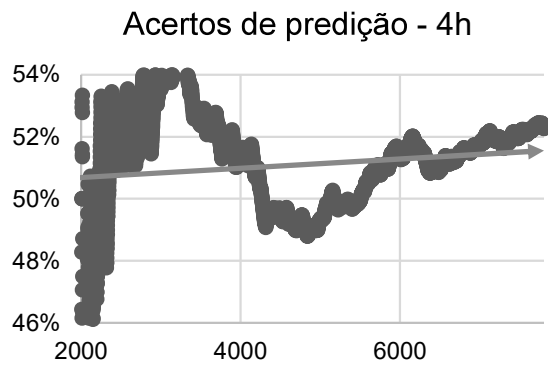


Gráfico 26: Resultados para regressões em toda a amostra.
Fonte: Autoria própria.

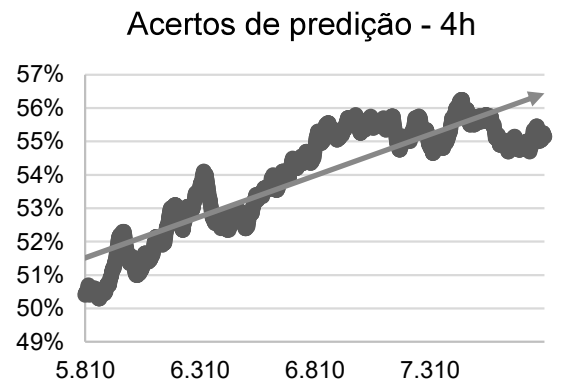


Gráfico 27: Média móvel (1000obs) de acertos por intervalo de dados.
Fonte: Autoria própria.

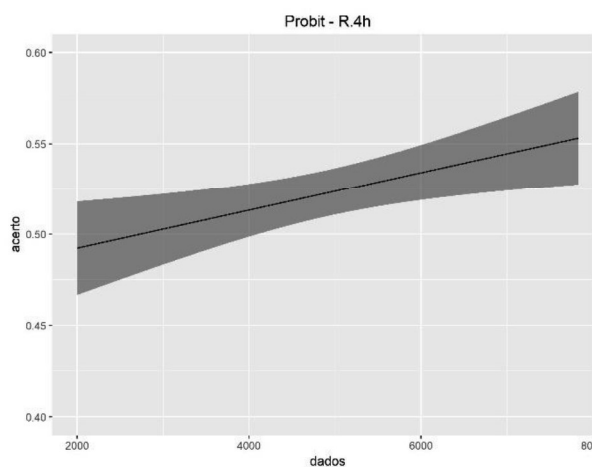


Gráfico 28: Distribuições via Probit de acertos de sinal de retornos
Fonte: Autoria própria

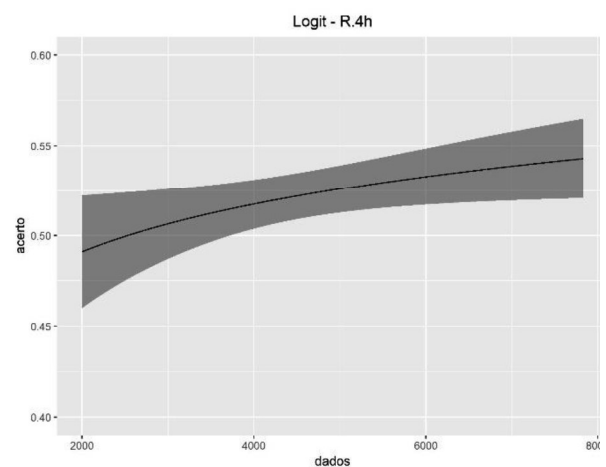


Gráfico 29: Distribuições via Logit de acertos de sinal de retornos
Fonte: Autoria própria.

ANEXO 7 – RESULTADOS DE PREVISÕES DE RETORNOS DE 5 HORAS

Tabela 10: Resultados das previsões de retornos 5 horas

Amostra de dados:	2000 a 3952	2953 a 3952
Média (Preços)	51.35%	51.30%
Confiança (95%)	49.14%	48.20%
Média	52.74%	54.90%
Confiança (95%)	50.52%	51.81%

Fonte: Autoria própria.

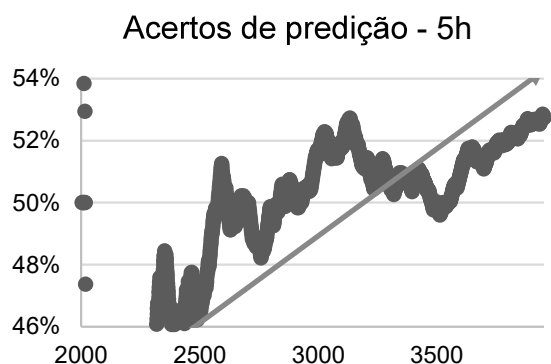


Gráfico 30: Resultados para regressões em toda a amostra.
Fonte: Autoria própria.

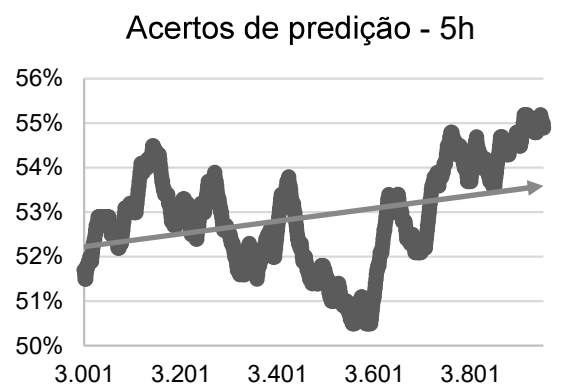


Gráfico 31: Média móvel (1000obs) de acertos por intervalo de dados.
Fonte: Autoria própria.

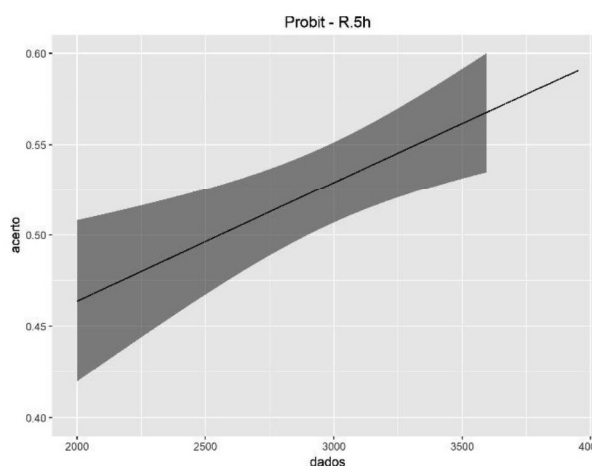


Gráfico 32: Distribuições via Probit de acertos de sinal de retornos
Fonte: Autoria própria.

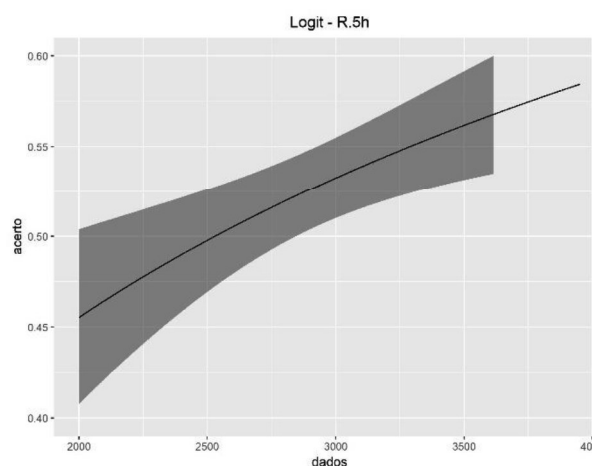


Gráfico 33: Distribuições via Logit de acertos de sinal de retornos
Fonte: Autoria própria.