UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO SÓCIO ECONÔMICO
PROGRAMA DE PÓS GRADUAÇÃO EM ECONOMIA

DOIS ENSAIOS DE APRENDIZADO DE MÁQUINA EM FINANÇAS: IDENTIFICAÇÃO DE VARIÁVEIS PARA PREVISÃO EM ALTA FREQUÊNCIA E COMPARAÇÃO DE MODELOS PREDITORES DE VOLATILIDADE

RAFAEL SILVA WAGNER

Florianópolis
2019

RAFAEL SILVA WAGNER

DOIS ENSAIOS DE APRENDIZADO DE MÁQUINA EM FINANÇAS: IDENTIFICAÇÃO DE VARIÁVEIS PARA PREVISÃO EM ALTA FREQUÊNCIA E COMPARAÇÃO DE MODELOS PREDITORES DE VOLATILIDADE

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Economia, no curso de Pós Graduação em Economia da Universidade Federal de Santa Catarina, Centro Sócio-Econômico.
Orientador: Prof. Dr. André Alves Portela Santos

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Wagner, Rafael Silva
Dois ensaios de aprendizado de máquina em finanças : Identificação de variáveis para previsão em alta frequência e comparação de modelos preditores de volatilidade / Rafael Silva Wagner ; orientador, André Alves Portela Santos, 2019.
172 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Sócio-Econômico, Programa de Pós-Graduação em Economia, Florianópolis, 2019.

Inclui referências.

1. Economia. 2. Aprendizado de Máquina. 3. Microestrutura de Mercado. 4. Volatilidade Realizada. I. Santos, André Alves Portela. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Economia. III. Título.

Rafael Silva Wagner

DOIS ENSAIOS DE APRENDIZADO DE MÁQUINA EM FINANÇAS: IDENTIFICAÇÃO DE VARIÁVEIS PARA PREVISÃO EM ALTA FREQUÊNCIA E COMPARAÇÃO DE MODELOS PREDITORES DE VOLATILIDADE

Esta Dissertação foi julgada adequada para a obtenção do Título de "Mestre em Economia", e aprovada em sua forma final pelo Programa de Pós-Graduação em Economia da Universidade Federal de Santa Catarina.

Florianópolis, 12 de Março 2019.

Prof. Dr. Jaylson Jair da Silveira
Coordenador
Universidade Federal de Santa Catarina

Banca Examinadora:

Prof. Dr. André Alves Portela Santos
Universidade Federal de Santa Catarina
Presidente
Orientador

Prof. Dr. Eraldo Sérgio Barbosa da Silva
Universidade Federal de Santa Catarina

Prof. Dr. Milton Biage
Universidade Federal de Santa Catarina

Prof. Dr. Leandro dos Santos Coelho
Pontifícia Universidade Católica do Paraná

RESUMO

Esta dissertação busca verificar a capacidade de algoritmos de aprendizado de máquina resolverem problema típicos em finanças. Inicialmente serão utilizadas técnicas de seleção de variáveis para identificar, dentro de um conjunto de variáveis que resumem a dinâmica de negociação do ativo financeiro, aquelas que mais impactam a direção dos movimentos dos retornos futuros. Posteriormente, será avaliada a capacidade de algoritmos de aprendizado de máquina realizarem melhores previsões de volatilidade realizada, quando comparados aos modelos tradicionais de econometria. Apesar de muitos resultados teóricos já terem atestado a importância da dinâmica de negociação para a evolução do preço dos ativos financeiros, há divergência sobre qual seria a correta especificação do processo. Um problema atual, dado o alto volume de informação que se tornou disponível, é saber verificar quais são as variáveis mais importantes para a previsão do sinal dos retornos. Nesta dissertação, serão utilizados dois modelos de conjuntos de árvores, *Boosting Trees* e *Random Forests*, para identificar as variáveis relevantes para a previsão da direção futura dos preços de ativos financeiros, permitindo verificar onde estão e quais são as variáveis mais importantes, isto é, que mais afetam o sinal dos retornos. Também será verificado, nesta dissertação,

a capacidade de modelos de aprendizado de máquina realizarem melhores previsões de volatilidade realizada, quando comparados aos modelos tradicionais da literatura econométrica. Será analisado se três modelos baseados em árvores, duas estruturas de penalização lineares e duas propostas de *prioris* bayesianas, são capazes de atingir melhor capacidade preditiva do que o principal modelo de previsão de volatilidade, o *Heterogeneous Auto-Regressive*. Além disso, também será verificado se modelos de combinação de previsões são capazes de gerar superior capacidade preditiva.

Palavras-chave: Aprendizado de Máquina. Microestrutura de Mercado. Volatilidade Realizada.

ABSTRACT

This work will evaluates the ability of machine learning algorithms to solve typical problems in finance. Initially, variable selection techniques are used to identify, within a set of variables that summarize the financial asset trading dynamics, those that most impact the direction of the movements of future returns. Afterwards, the ability of machine learning algorithms to perform better predictions of realized volatility is evaluated, when compared to the traditional models of econometrics. Although many theoretical results have already testified to the importance of trading dynamics for the evolution of price of financial assets, there is a great deal of divergence as to what would be the correct specification of the process. A current problem, given the high volume of information that has become available, is to know what are the most important variables for predicting the returns signal. In this work, two models of ensemble trees, Boosting Trees and Random Forests, will be used to identify the relevant variables for predicting the future direction of financial asset prices, allowing to verify where they are and which are the most important variables, that is, those that most affect the returns signal. We also asses the ability of machine learning models to perform better predictions of realized volatility when compared to traditional models of the econometric

literature. It is analyzed whether three tree-based models, two linear penalty structures and two Bayesian priori proposals, are able to achieve better predictive capacity than the main predictor of volatility, the Heterogeneous Auto-Regressive model. In addition, it is also considered whether forecasting combination techniques are capable of generating superior predictive ability.

Keywords: Machine Learning. Market Microstructure. Realized Volatility.

AGRADECIMENTOS

O curso de mestrado do Programa de Pós-Graduação em Economia mostrou-se como um programa de gigantesco crescimento acadêmico, pessoal e profissional na minha história. Inicialmente agradeço a todos que trabalham na gestão do Programa, os coordenadores Prof. Dr. Jaylson Jair da Silveira e Profa. Dra. Eva Yamila Catela, assim como as secretárias e estagiárias que auxiliavam os coordenadores com as demandas administrativas.

Agradeço também a Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES) por financiar o meu projeto de pesquisa ao longo de um ano e seis meses. Sem o apoio financeiro do ministério da educação eu com certeza não teria conseguido mergulhar de corpo e alma nos estudos.

Agradeço ao meu orientador Prof. Dr. André Alves Portela Santos por me guiar ao longo de um ano e meio. Inteligente, prestativo e perfeccionista, me tornei grande fã de seu trabalho, de sua postura profissional e de sua pessoa, impossível cogitar melhor orientação. Também agradeço a todos os professores que me auxiliaram no entendimento da matemática/estatística/computação. Posso citar como indispensáveis a minha formação as aulas do Prof. Dr. Francis Carlo Petterini, Prof. Dr. Guilherme Valle Moura e Prof. Dr. Milton Biage. O sucesso que tive em alcançar

o que me tornei começa nas aulas de meus professores, passa pelo guiar de meu orientador e se materializa neste texto.

Agradeço também minha amada companheira Caroline Medeiros Colombi, minha sempre presente mãe Luciana Dias da Silva, meu inspirador pai Roberto Machado Wagner e minha cara irmã Fernanda Silva Wagner. Também agradeço aos divertidos colegas que me acompanharam nessa jornada, meu colega de trabalho Mateus, meu companheiro de refeições Henrique, minha segunda professora Aishameriane e meu colega de sala de estudos Bruno. Finalmente nos tornamos mestres.

Por fim agradeço à Universidade Federal de Santa Catarina, quem convive comigo diariamente sabe que amo esta instituição. Frequento esta universidade ativamente há pelo menos oito anos e não consigo sonhar em melhor lugar pra ter passado esse tempo.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Métodos de aprendizado de máquina tornaram-se um ponto de grande interesse para diferentes campos do conhecimento. Diferentes técnicas vêm sendo utilizadas para detecção de *malwares* (UCCI; ANIELLO; BALDONI, 2018), detecção de sinal de wireless (KUMAR; AMGOTH; ANNAVARAPU, 2019), detecção de estresse mental (PANICKER; GAYATHRI, 2019), processamento de minerais (MCCOY; AURET, 2019) entre outras diversas aplicações.

Por meio de um grande grupo de dados x, funções matemáticas genéricas tentam conectar uma função dos dados $f(x)$ a um sinal y, a conexão é feita por meio de um algoritmo de otimização, que com base em um objetivo, como uma medida de qualidade preditiva, altera a estrutura da função. Esse procedimento é conhecido como aprendizado supervisionado, pois o sinal y supervisiona o algoritmo de otimização durante a estimação de uma função dos dados $f(x)$.

Uma vez estimada a função dos dados $f(x)$, é

possível analisar o que o algoritmo de otimização conseguiu aprender e aumentar o nosso próprio entendimento sobre o processo gerador de dados. Essa fonte de conhecimento é valiosa dada a incapacidade do ser humano ser plenamente criativo e conjecturar todas as possibilidades de modelagem. Na história da ciência, são conhecidos os famosos momentos *eureka* em que o cientista consegue entender algo até então não compreendido. Analisar aquilo que foi aprendido pelo algoritmo de otimização pode facilitar a que cheguemos a tais momentos, tanto a estrutura quanto as características da função permitem que o conjunto de conhecimento sobre o processo gerador de dados seja ampliado.

O primeiro capítulo deste trabalho se propõe a verificar quais das variáveis que descrevem a dinâmica de negociação dos ativos financeiros mais importam para os modelos que preveem o sinal dos retornos futuros.

Diferentes metodologias quantitativas foram desenvolvidas para se atingir regras para a realização de investimentos, notoriamente, três campos separados atingiram diferentes resultados: a primeira aplicou teorias de cálculo estocástico para solucionar problemas relacionados a opções financeiras; a segunda, avaliando correlações nos retornos, buscou imunizar portfólios e ocasionalmente obter retornos extraordinários na

ocorrência de desequilíbrios e, finalmente, a terceira – e mais importante para o presente trabalho – fez análises estruturais do sistema de negociação, buscando encontrar padrões de ineficiência no processo de apreçamento.

A terceira área comentada anteriormente é conhecida entre os financistas como *microestrutura de mercado*. Área de pesquisa em finanças voltada a estudar a estrutura do mercado financeiro. Após um demasiado esforço teórico, verificou-se que a estrutura do mercado poderia implicar ineficiências no sistema de negociação, tornando possível a previsibilidade futura do preço dos ativos.

Ainda que os resultados sejam empolgantes, pouco se pode afirmar sobre como realizar as previsões, embora praticamente não haja discordância sobre o fato de a previsibilidade existir. Tipicamente, isso se parece com um problema de identificabilidade, isto é, sabe-se que existe uma função dos dados e acredita-se que ela tenha impacto em um sinal futuro, mas não se sabe, ainda, como construir essa função.

No Capítulo 2, serão utilizados dois algoritmos de aprendizado de máquina para tentar identificar as fontes de informação desse sistema. Serão utilizados algoritmos da família conhecida como modelos baseados em árvores, cujas características que mais importam são

a alta versatilidade possível de ser colocada na função $f(x)$, quebrando o vínculo com funções de formatos já pensados, e o alto poder preditivo que as funções estimadas conseguem atingir.

Os algoritmos *Random Forest* e *Boosting Trees* permitem verificar quais foram as variáveis mais importantes na estimação do algoritmo aprendido, se o sinal supervisor y for a direção do movimento futuro do retorno financeiro, pode-se verificar quais são as variáveis que geram maior previsibilidade. Exemplificaremos a análise exploratória com o ativo *PETR4*, negociado na bolsa brasileira B3. Serão analisadas duas defasagens de mais de 30 variáveis que descrevem o sistema de negociação. Serão identificadas não somente quais as variáveis mais importantes, mas a qual período elas pertencem, isto é, se são do período anterior ou da defasagem superior. Além disso, buscaremos compreender qual a origem das variáveis, se são registros de negócios já realizados ou se são ofertas pretendentes.

Identificar essas variáveis, tidas como relevantes pelo algoritmo de previsão, pode permitir que futuros trabalhos teóricos foquem em como adicioná-las em novas propostas de modelos, dando um passo adiante, tentando modelar um relacionamento mais acurado para a dinâmica de curto prazo do preço dos ativos. No Capítulo 2 será

realizada essa investigação.

No capítulo seguinte, seguindo adiante, muda-se o perfil do objetivo de pesquisa. Com uma função estimada $\hat{f}(x)$, é possível utilizá-la para extrapolar o conjunto de dados original x com observações fora da amostra $\hat{x}$, ou seja, utilizar a função para fazer previsões $\hat{f}(\hat{x})$ sobre algo que não havia sido aprendido pelo algoritmo.

Nos modelos de previsão de volatilidade, a previsibilidade se destaca. Algo intrigante, para o leitor novo, é que, apesar de o sinal $sign(r_t)$ e o nível $|r_t|$ dos retornos serem previsíveis, sua média não o é $E[r_t] = E[sign(r_t) * |r_t|]$. Por isso, neste capítulo, cujo tema de interesse é a previsibilidade, o foco será encontrar um modelo "campeão" dentro de todos os modelos de volatilidade realizada; o *Heterogeneous Auto Regressive*.

Nessa fase, utilizaremos tanto modelos de *árvores de modelos*, indo desde árvores simples a modelos de conjuntos como *cubist* e *bagging*; regressões lineares com restrições de penalização e regressões bayesianas com prioris altamente informativas. Agora, a pergunta a ser respondida será mais objetiva: algumas das estruturas propostas têm melhor poder preditivo do que o modelo *benchmark* tradicional? E, se houver, o que esse modelo está aprendendo que o modelo base não consegue incorporar?

No Capitulo 3 será realizada essa análise. Serão utilizados 10 anos de registros de negociação de 10 ativos financeiros americanos para testar, empiricamente, qual especificação é a mais interessante, para, posteriormente, interpretá-la e verificar o que foi aprendido de novo.

Com base nessa sequência de duas aplicações de algoritmos de aprendizado de máquina em finanças pretende-se contribuir em dois principais campos:

- Identificar, dentro das variáveis que descrevem o sistema de negociação de um ativo financeiro, quais são aquelas que produzem maior capacidade preditiva da futura direção do preço dos ativos financeiros;
- Verificar se modelos de aprendizado de máquina, modelos penalizados e bayesianos ou uma combinação deles é capaz de gerar maior capacidade preditiva do que o *Heterogeneous Auto-Regressive*.

Sendo assim, após esta introdução, o Capítulo 2 dedica-se a identificar as fontes de informação casuística dentro do sistema de negociação. Na Seção 2.1 será introduzida a fundamentação necessária para entender o problema de previsão e os algoritmos a serem utilizados. Na seção 2.2, será apresentada a utilização dos algoritmos de aprendizado, avaliando os seus resultados.

No Capítulo 3 serão utilizados os modelos

preditivos para tentar ganhar do modelo *benchmark*. Na Seção 3.1 será introduzido o *background* necessário sobre volatilidade, e na Seção 3.2 serão apresentados os modelos para prevê-la. Já na Seção 3.3 será realizada uma simulação verificando se os modelos propostos atingem melhor capacidade preditiva. Por fim, no Capítulo 4 retoma-se a visão geral do trabalho, destacando as principais conclusões.

2 PREVISÃO DA DIREÇÃO DOS RETORNOS DE ALTA FREQUÊNCIA: QUAIS VARIÁVEIS IMPORTAM?

A dinâmica dos retornos de alta frequência é de interesse para diversos agentes. Órgãos governamentais, por exemplo, preocupam-se em garantir um ambiente estável de negociação; agentes financeiros, por outro lado, preocupam-se com movimentos abruptos que coloquem em risco seus patrimônios. Corretoras de investimento, ainda, tentam encontrar o melhor momento de executar suas ordens e, por fim, especuladores estão sempre atentos a novas oportunidades de negociações lucrativas.

A análise dos retornos de ativos financeiros se mostra, há tempos, uma tarefa desafiadora. O primeiro momento dos retornos de ativos financeiros demonstra ínfima previsibilidade, um fato já esperado pelas tradicionais teorias de finanças (MALKIEL; FAMA, 1970). A variância dos retornos, ao contrário da esperança, apresenta previsibilidade. Christoffersen e Diebold (2006) foram os primeiros a elucidar como a previsibilidade da

variância transforma-se em previsibilidade de sinal e, desde então, múltiplos esforços foram realizados para modelar a direção do movimento dos ativos financeiros.

Ao mesmo tempo em que se desenvolvia a literatura de previsão de sinal, um campo conhecido como microestrutura de mercado desenvolvia estudos analisando a dinâmica de negociação dos ativos financeiros. A disponibilidade de dados de negociação de altíssima frequência permitiu que novas pesquisas verificassem a relevância da dinâmica de negociação para a evolução dos preços. No entanto, essa grande quantidade de dados necessita de técnicas especiais para ser apropriadamente manipulada e analisada, forçando os pesquisadores a se afastarem da estatística clássica e se aproximarem da estatística computacional.

Trabalhos feitos por Dixon, Klabjan e Bang (2016), Tsantekidis et al. (2017) e Fletcher e Shawe-Taylor (2013) atestaram que essas técnicas conseguem aprender padrões na dinâmica de negociação que geram previsibilidade sobre a direção do movimento futuro dos ativos financeiros. Dos pontos da negociação avaliados constam o *spread*, número de ordens novas/canceladas/executadas, quantidade/volume de ordens, horário de negociação e outras. Apesar de a capacidade preditiva ter sido atestada, a verificação de em

qual parte da negociação está o poder preditivo é pouco trabalhada.

Propõe-se utilizar uma família de algoritmos que é especialmente capaz de realizar a identificação das variáveis de maior importância, aliando poder preditivo com compreensibilidade. Essa família é popularmente conhecida por *métodos baseados em árvores*. Breiman et al. (1984) foram os responsáveis por popularizar as árvores de decisão, caracterizadas por não presumirem qualquer estrutura nos dados, sendo preferíveis em situações em que não existe um bom conhecimento prévio para modelar o processo gerador de dados. Muitos *tree ensemble methods* (métodos de conjuntos de árvores) foram propostos na literatura para aumentar o poder preditivo das árvores de decisão. Este trabalho irá enfocar em dois dos mais populares métodos: *Boosting Trees* (SCHAPIRE, 1991) e *Random Forests* (BREIMAN, 2001). Ambos os modelos serão utilizados para detectar as variáveis da dinâmica de negociação que mais auxiliam a prever os retornos de alta frequência.

2.1 Referencial Teórico

2.1.1 Sinal dos Retornos

Para garantir a compreensibilidade, inicia-se definindo alguns dos principais termos utilizados pela

literatura de previsão de sinal. Seja Ω_t o conjunto de informação disponível em t e R_t, o retorno de um ativo no período t com a respectiva média e variância condicional $\mu_{t+1|t} = E[R_{t+1}|\Omega_t]$ e $\sigma^2_{t+1|t} = Var[R_{t+1}|\Omega_t]$. Diz-se que a série de retornos apresenta previsibilidade de média condicional se $\mu_{t+1|t}$ variar conjuntamente com Ω_t; a previsibilidade de variância condicional é definida da mesma forma. (CHRISTOFFERSEN et al., 2006)

Sendo $Pr[R_t > 0]$ a probabilidade de um ativo ter retorno positivo em t, caso $Pr[R_t > 0]$ exiba dependência condicional, isto é, se $Pr[R_{t+1} > 0|\Omega_t]$ variar conjuntamente com Ω_t, diz-se que a série de retornos apresenta previsibilidade de sinal.

2.1.1.1 Retornos Intradiários

Devido a recentes resultados da literatura, a análise se restringirá a janelas de tempo intradiárias. Nas janelas intradiárias existe o tradicional problema de escolher o horizonte de previsão. Contrariando uma primeira impressão em que se esperaria que dados mais granulares e detalhados acrescentassem informações mais detalhadas, verifica-se que frequências intermediárias apresentam padrões mais claros, devido, principalmente, a maior razão sinal/ruído delas.

Um grande volume de trabalhos realizados na área

de volatilidade realizada demonstrou extensivamente que o ruído da microestrutura de mercado confunde os sinais que deveriam estar sendo amplificados. Frequentemente, nessa literatura, verifica-se que os horizontes ótimos, apesar de curtos, não são levados ao limite; para uma recente comparação de diferentes metodologias nessa área, sugere-se o trabalho de Liu, Patton e Sheppard (2015).

Na alta frequência, situações nas quais não se vê qualquer mudança nos preços são comuns. Esse tipo de situação define um problema de previsão em 3 classes: os retornos futuros podem ser positivos, negativos ou nulos, sendo retorno nulo intuitivamente compreendido como $R_t = 0$. Problemas multi-classe desse tipo são descritos, tipicamente, pela função de distribuição de probabilidade multinomial.

Seguindo Lehmann e Casella (2006), a distribuição multinomial é caracterizada por uma sequência de n experimentos *i.i.d.* que podem levar a r diferentes classes com respectivas probabilidades $p_1, p_2, ..., p_r | \sum_{i=1}^{r} p_i = 1$. Definindo n_i como o número de experimentos que resultaram na classe i, a distribuição conjunta de todos os n experimentos é descrita pela distribuição multinomial $M(p_1, ..., p_r; n_1, ..., n_r)$:

$$P(X_1 = n_1, \ldots, X_r = n_r) = \frac{n!}{n_1! \ldots n_r!} p_1^{n_1} \ldots p_r^{n_r}, \tag{2.1}$$

sendo $\sum_{i=1}^{r} n_i = n$.

Pode-se derivar, também, essa equação de maneira intuitiva. Inicialmente, note que a sequência de n experimentos que resultam na classe i n_i vezes têm, pela hipótese de independência, probabilidade de ocorrência igual a $p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$. Como há $n!\,/(n_1!\,n_2!\dots n_r!\,)$ sequências de resultados, existirão $n!\,/(n_1!\,n_2!\dots n_r!\,)$ diferentes permutações da sequência n, na qual $n_1, n_2, ...n_i$ dessas permutações terão resultados iguais. Até o fim deste capítulo consideraremos que o problema é descrito pela densidade multinomial e que, no caso em análise, considerar-se-á que $r = 3$ (ROSS, 2013).

2.1.1.2 Previsibilidade de sinal

Christoffersen e Diebold (2006) foram os primeiros a dar um tratamento rigoroso para o relacionamento entre sinal e volatilidade. A grande inovação introduzida foi a elucidação de que a dependência de sinal não requer dependência de média, isto é, o fato estilizado de os retornos de ativos financeiros terem a média pouco previsível não contagia a previsibilidade de seu sinal.

Os autores demonstraram que, se a média dos retornos for diferente de zero $\mu \neq 0$, pode-se atingir a previsibilidade de seu sinal através da previsibilidade da variância. Exemplificando, supondo que $E[R_{t+1}] = \mu > 0$,

pode-se verificar que a probabilidade de um dado retorno ser positivo é:

$$\begin{aligned} Pr(R_{t+1} > 0) &= 1 - Pr_t(R_{t+1} < 0), \\ &= 1 - Pr\left(\frac{R_{t+1} - \mu}{\sigma_{t+1|t}} < \frac{-\mu}{\sigma_{t+1|t}}\right), \\ &= F\left(\frac{\mu}{\sigma_{t+1|t}}\right), \end{aligned} \tag{2.2}$$

em que F é a função de distribuição acumulada (cdf). Logo, a probabilidade de um retorno ser positivo é igual a cdf até a média padronizada.

Pode-se ver que, conforme a variância condicional muda, a probabilidade de um retorno ser positivo muda conjuntamente. Caso a média seja positiva, quanto menor a variância, maior será a probabilidade de um dado retorno ser positivo. Ainda, quanto mais a média se aproximar de zero, menor se torna a previsibilidade de sinal, especialmente se a distribuição de probabilidade dos retornos for simétrica. Em resumo, a variância interage com uma média diferente de zero produzindo o objetivo: a previsibilidade de sinal.

Posteriormente Christoffersen et al. (2006) demonstraram que esse resultado se estende a casos em que a média incondicional é zero, desde que a assimetria e a curtose apresentem previsibilidade. Para chegar a esse

resultado eles realizaram uma expansão de Gram-Charlier:

$$1 - F\left(\frac{\mu}{\sigma_{t+1|t}}\right) \approx 1 - \Phi\left(\frac{\mu}{\sigma_{t+1|t}}\right) \times \left[-1 + \frac{\gamma_{3,t+1|t}}{3!}\left(\frac{\mu^2_{t+1|t}}{\sigma^2_{t+1|t}} - 1\right) + \frac{\gamma_{4,t+1|t}}{4!}\left(\frac{-\mu^3_{t+1|t}}{\sigma^3_{t+1|t}} + \frac{3\mu_{t+1|t}}{\sigma_{t+1|t}}\right)\right],$$

em que Φ é a função de distribuição de probabilidade (pdf) e $\gamma_{3,t+1|t}$, $\gamma_{4,t+1|t}$ são respectivamente a assimetria e curtose condicionais. Para mais detalhes sobre essa expansão, veja o trabalho de Cramér (2016). Com essa expressão, mesmo nos casos em que a média se iguala a zero, ainda é possível termos previsibilidade de sinal, desde que haja previsibilidade de assimetria ou curtose.

2.1.2 Microestrutura de Mercado

Microestrutura de mercado é um campo de pesquisa em finanças que estuda a estrutura dos mercados. Seu objeto de pesquisa é, tipicamente, o próprio sistema de negociação, verificando como ele afeta a dinâmica de preços, ordens e volumes.

Como usual, deve-se introduzir o sistema de negociação que será analisado. Os *traders* vão às bolsas negociar ativos e informam sua oferta e

demanda aos outros participantes. Para isso eles utilizam, predominantemente, dois tipos de ordens: ordens a mercado e ordens limitadas. A diferença entre elas é trivial; em uma ordem limitada o *trader* não está disposto a negociar pelo preço que está sendo ofertado no momento; o participante sabe a quantidade desejada e seu preço limite exigido, que está além das atuais melhores ofertas. Nesse caso, o participante apregoa uma nova oferta que satisfaça o seu preço alvo.

Conforme a negociação evolui, novas ordens limitadas são apregoadas; o conjunto das ordens limitadas ativas, aquelas que ainda não foram canceladas ou negociadas, forma o livro de ofertas, uma lista de ordens ativas ordenadas por preço e hora de chegada; as ordens que chegam antes têm prioridade, assim como aquelas com preço mais próximo àquele que está sendo atualmente negociado. O maior preço de uma ordem de compra ativa é conhecido como *bid* e o menor preço de uma ordem de venda ativa é conhecido como *ask* e a diferença entre eles é conhecida como *bid-ask spread*.

Em uma ordem a mercado não há exigência de preço. O participante informa, apenas, a quantidade desejada e o sistema automaticamente designará para contrapartida as primeiras ordens disponíveis no livro de ofertas, sendo o preço do negócio aquele previamente

estabelecido nas ordens limitadas. Na rotina do mercado, o *bid* e *ask* apenas podem mudar de três maneiras: se todas as ordens limitadas ativas com o preço do *bid*/*ask* forem canceladas; caso todas elas sejam negociadas, ou, ainda, se previamente havia um "buraco" no livro (um preço entre o *bid* e o *ask* sem nenhuma ordem ativa) e esse buraco foi preenchido por uma oferta nesse preço. Pode-se ver que esses dois preços podem variar muitas vezes entre um negócio e outro.

2.1.2.1 Mercados e Conjuntos de Informação

Podemos dividir os mercados conforme duas características. A primeira é o nível de competição, ou seja, em mercados completamente competitivos os agentes não têm capacidade de influenciar os preços, enquanto que em mercados não perfeitamente competitivos, os agentes têm alguma influência sobre eles. A segunda característica para dividir os mercados é pelo tipo de informação. Mercados de informação heterogênea ocorrem quando todos os participantes, ou pelo menos a maior parte deles, recebem informações privadas correlacionadas ao preço do ativo, enquanto que mercados de informação assimétrica ocorrem quando um pequeno número de agentes têm acesso a um conjunto de informação diferente dos demais.

No caso de informação heterogênea têm-se pouca

diferenciação em relação aos resultados clássicos de microeconomia. Grossman et al. (1989) demonstram que os preços da negociação terão o papel de compartilhar a informação heterogênea para todos os agentes, mas, curiosamente esse conjunto de informação heterogêneo tende a não gerar novas negociações (MILGROM; STOKEY, 1982). Esse fenômeno ocorre em equilíbrios completamente reveladores, em que, através do preço e do volume negociado, a informação é compartilhada e não há beneficio em se ter um conjunto de informação diferente dos demais.

Apesar disso, trabalhos posteriores verificaram a possibilidade de dois conhecidos fatos estilizados que contrariam a intuição dos resultados clássicos. *Momentum* é a característica dos retornos que apresentam a primeira autocorrelação positiva (KIM; VERRECCHIA, 1991),(HE; WANG, 1995), já reversão à média implica que o excesso de retorno é devolvido após um número suficiente de períodos (CAMPBELL; GROSSMAN; WANG, 1993),(CONRAD; HAMEED; NIDEN, 1994). Mesmo em mercados competitivos, ambos os fenômenos podem ocorrer. Em geral, *momentum* seria gerado por negócios motivados por informações relacionadas ao preço fundamental do ativo, enquanto que a reversão à média seria gerada por negociações não motivadas por

informações ligadas ao preço fundamental dos ativos.

Há, ainda, o caso mais interessante de informação assimétrica. Em mercados perfeitamente competitivos isso gera o problema da seleção adversa, e *traders* mal informados podem ser a contraparte de um negócio lucrativo de *traders* bem informados. A cautela gerada por essa possibilidade provoca uma diminuição do volume de negociação. Já no caso do mercado ser imperfeitamente competitivo, tem-se forte diminuição da razão sinal/ruído (KYLE, 1989).

Ainda, verifica-se que os volumes aumentam nos dias em que há presença de *traders* informados (MEULBROEK, 1992), mas não é possível diferenciar esse volume daquele gerado por *traders* não informados (CHAKRAVARTY; MCCONNELL, 1999). Apesar disso, grande esforço é feito tanto pelo meio acadêmico (EASLEY et al., 1996) quanto por órgãos oficiais (MITCHELL; NETTER, 1994) para criar sistemas de detecção de negócios motivados por informação privilegiada, visando inibir o poder que esse tipo de privilégio tem na dinâmica de negociação e na evolução do preço dos ativos (BENABOU; LAROQUE, 1992).

2.1.3 Prevendo os Retornos Dentro da Altíssima Frequência

Tanto as características típicas das séries dos ativos financeiros que trazem previsibilidade aos preços, como *momentum* e reversão à média, quanto a possibilidade de existir informação privilegiada dentro do fluxo de negócios ainda não incorporada pelos outros agentes tornam a análise da dinâmica de negociação capaz de captar sinais que se relacionam ao sinal dos retornos futuros.

Existem muitas abordagens para organizar, modelar e usar as informações para prever os retornos de alta frequência. Breiman (2001) descreve os dois caminhos que um estatístico tipicamente segue: o primeiro é criar uma lista de hipóteses sobre o processo gerador de dados e usá-las para construir e estimar modelos teóricos. O segundo usa algoritmos genéricos de alta versatilidade, sem realizar hipóteses específicas referentes ao processo. A alta versatilidade da segunda abordagem permite maior adaptação a diferentes conjuntos de dados.

Aprendizado de máquina está inserido nessa segunda abordagem. Suas técnicas são desenhadas para capturar relacionamentos não lineares em conjuntos de dados de alta dimensão, sem requerer hipóteses sobre o comportamento do verdadeiro processo gerador

de dados. Recentemente, muitos trabalhos atingiram bom desempenho utilizando modelos de aprendizado de máquina para prever o sinal dos retornos em alta frequência. E é um conjunto desta literatura, que combina os algoritmos e variáveis de microestrutura de mercado, que é mais importante para o presente trabalho.

Dixon (2017) utilizou a estrutura de redes neurais recorrentes, um tipo especial de rede neural desenvolvida para dados ordenados, para modelar o relacionamento entre as variáveis de microestrutura com os futuros sinais dos ativos. O conjunto de dados consiste dos registros do *E-mini S&P 500* ao longo de agosto de 2016. O autor comparou seu modelo com outras estruturas de redes neurais, assim como métodos estatísticos tradicionais e reportou um desempenho preditivo superior das redes neurais recorrentes em comparação aos demais modelos.

Também dentro do paradigma das redes neurais Tsantekidis et al. (2017) usaram redes neurais convolucionárias, estrutura muito comum em reconhecimento de imagens, para prever a direção da média entre o *bid-ask* de cinco ações finlandesas. Os autores utilizaram o volume nos 10 primeiros níveis do livro de ofertas defasados por 10 períodos como variáveis preditivas. Ao comparar o desempenho do algoritmo com redes neurais de multicamada e máquinas de vetor-suporte,

atestaram um desempenho superior para as redes neurais convolucionárias. Ainda, testaram diferentes horizontes de agregação temporal e reportaram que a melhor qualidade preditiva surgia nas menores frequências.

Também utilizando máquinas de vetor-suporte, outro algoritmo de aprendizado de máquina, Fletcher, Hussain e Shawe-Taylor (2010) modelaram a função entre o sinal dos retornos futuros e o volume dentro dos 6 primeiros níveis do livro de ofertas e verificaram que estruturas multi-*kernels* tinham desempenho preditivo superior aos demais modelos, verificando que as previsões atingiam performance preditiva interessante e geravam resultados positivos na simulação financeira.

Posteriormente Fletcher e Shawe-Taylor (2013) adicionaram dois novos conjuntos de variáveis. Inicialmente, adicionaram um conjunto de informações referentes ao histórico do preço negociado, incluindo máximo, mínimo, média e desvio-padrão do preço dentro do período; além disso, adicionaram as estimativas de modelos típicos de apreçamento de ativos, modelaram a evolução dos preços supondo processos Wiener e Poison e adicionaram as estimativas paramétricas como potenciais regressores. Com esse amplo conjunto de informações, construíram 165 *kernels*, que foram selecionados, posteriormente, visando identificar quais continham o maior valor preditivo. Os

autores verificaram que os *kernels* mais importantes eram aqueles que continham informações do novo conjunto de estatísticas descritivas referentes ao preço de negociação, mas verificaram que a combinação de múltiplos *kernels* provocava um significativo aumento na qualidade preditiva.

Han et al. (2015) estenderam o mesmo conjunto de variáveis que foi usado por Fletcher e Shawe-Taylor (2013) para prever o sinal do *spread* médio do $S\&P\ 500$. A inovação introduzida pelos autores foi a utilização de modelos baseados em árvores, que permitiram verificar que as informações mais importantes para realizar a previsão futura consistiam dos seis primeiros níveis do livro de ofertas. Os autores verificaram, ainda, que o algoritmo de floresta aleatória foi o que apresentou melhor capacidade preditiva.

Diversos trabalhos já verificaram a existência de previsibilidade do sinal futuro utilizando como variáveis preditivas estimadores de microestrutura de mercado. No entanto, pouco se aprofundou na identificação da importância das variáveis. Han et al. (2015) introduziram, recentemente, a utilização de algoritmos baseados em árvores ao problema e verificaram sua boa capacidade preditiva. Assim, esses algoritmos serão aproveitados para realizar um novo esforço de pesquisa, buscando não a sua capacidade preditiva, mas o entendimento dos

relacionamentos presentes dentro do sistema.

2.1.4 Importância das Variáveis

Capacidade preditiva interessa a todos que entram no mundo das previsões financeiras. É possível, e até mesmo provável, que a maior parte dos que começaram a pesquisar o tema buscavam realizar boas previsões. Apesar de esse esforço ser tentador, simulações computacionais tornam-se, facilmente, superajustadas aos dados e pode ser difícil diferenciar se a capacidade preditiva é decorrente de um modelo bem especificado ou simplesmente um produto de sequenciais simulações que findam em excesso de ajuste (WASSERSTEIN; LAZAR et al., 2016). Um objetivo de pesquisa igualmente intrigante é a análise do sistema de negociação e como as variáveis se relacionam dentro dele.

Prado (2018) apresenta as duas principais metodologias para pesquisa de importância de variáveis. Decaimento médio da impureza (DMI) é uma metodologia de identificação de variáveis, aplicável a modelos que permitem identificar o quanto cada variável colabora para a estimação do modelo. Posteriormente, será apresentado que modelos baseados em conjunto de árvores são estimados sequencialmente, e a cada novo teste o modelo aprende um novo elemento da função. Decaimento médio da impureza é a quantificação de quanto cada variável

contribuiu para diminuir o erro do estimador.

Decaimento médio da acurácia (DMA) é, possivelmente, a mais lenta das duas metodologias. Dado um classificador $f(x)$, permuta-se as variáveis contidas na matriz de regressores X e avalia-se a capacidade preditiva fora da amostra por meio de validação cruzada. O crescimento do erro quando uma variável é retirada é o quanto esta contribuiu para a performance preditiva do modelo. Esse simples procedimento pode ser aplicado a qualquer classificador.

Alguns cuidados devem ser tomados ao se avaliar os resultados de pesquisas de importância das variáveis. Inicialmente, o fenômeno conhecido como multicolinearidade pode atestar que variáveis extremamente relacionadas e que têm forte capacidade preditiva não são importantes para previsão. No DMI isso ocorre através da designação de um baixo *score* para a variável na pontuação geral. Já no DMA isso ocorre quando a retirada de uma variável importante não diminui o erro do modelo, pois outra variável extremamente correlacionada também captura o padrão de correlação, potencialmente concluindo que ambas as variáveis não têm importância, mesmo quando as duas têm alto poder preditivo.

A análise fora da amostra tem o benefício de poder concluir que não há variável alguma que tenha importância

enquanto a análise dentro da amostra sempre devolverá as variáveis que o modelo melhor conseguiu relacionar aos dados, mesmo que o relacionamento seja muito fraco.

Existem, ainda, outros métodos de análise de importância, como a estimação individual de um modelo para cada regressor com posterior avaliação de cada modelo fora da amostra, técnica que restringe os benefícios que se tem ao combinar múltiplos regressores. Outra possibilidade seria usar técnicas de ortogonalização para resumir as características em poucos vetores descritivos, metodologia principalmente conduzida através da análise de componentes principais. Para mais detalhes sobre essas e outras metodologias alternativas com enumeração de seus benefícios e malefícios, ver o trabalho de Prado (2018).

2.1.5 Aprendizado de máquina

A investigação da importância das variáveis será feita utilizando dois dos principais métodos de conjunto de árvores: *random forest* e *boosting trees*. Essa família de algoritmos tem duas características interessantes: a primeira é a utilização dela nos mais variados tipos de problemas de previsão, incluindo previsões financeiras de alta frequência; a segunda, e mais importante característica, é que o método de investigação de variáveis focará, principalmente, na estimação das árvores, permitindo que,

em um único modelo, encontre-se as melhores combinações de preditores e tenha-se um método objetivo para avaliar a importância das múltiplas variáveis.

2.1.5.1 Árvores de Regressão

Ambos os modelos foram construídos expandindo o conceito de árvores de classificação, e é importante definir o que são as árvores de classificação.

Conforme Mitchell (1997), pode-se dividir um modelo baseado em árvores em três estruturas: a raiz, os nós e as folhas. A raiz pode ser entendida como o ponto de onde todas as observações começam a ser analisadas, isto é, a estimação de uma árvore inicia com todas as observações na raiz, e a partir desse conjunto completo de dados criam-se dois subconjuntos diferentes não intersecionados, que quando unidos são iguais ao conjunto que estava inicialmente na raiz. Para realizar essa divisão, escolhe-se uma variável dentro do conjunto de regressores para ser a variável de corte; quando a variável é discreta - variáveis representadas por classes, por exemplo - a regra de corte é uma regra de igualdade. Já para variáveis contínuas, a regra de decisão é expressa, tipicamente, por meio de uma desigualdade.

O procedimento para encontrar a variável de corte e definir o ponto ou categoria de corte é conhecido como

split. Com os subgrupos de observações resultantes do primeiro *split*, criam-se novos subgrupos com novos *splits*. A sequência de *splits* gera uma partição no espaço de regressores, e o processo de gerar, sequencialmente, novos splits é conhecido como crescer ou construir a árvore. Após o processo de crescimento das árvores, o produto final será uma partição do espaço de regressores; cada célula da partição é conhecida como folha e, em cada folha, será alocada uma diferente regra de previsão, tipicamente uma constante. Então, quando uma nova observação for apresentada, esta fluirá através dos *splits*, e será possível verificar a qual partição a nova observação pertence. Após esse processo, prevê-se um valor para essa observação conforme o modelo estimado na região específica.

Um exemplo ilustrativo desse procedimento pode ser encontrado na Figura 1; no lado esquerdo, apresenta-se a estrutura de uma árvore de decisão. Partindo da raiz, no ponto superior verifica-se se $x_1 \leq t_1$. Caso seja verdade flui-se para a esquerda, caso contrário, flui-se para a direita. Sequencialmente, avalia-se as desigualdades até chegar-se às folhas R_1, R_2, R_3, R_4, R_5. A consequência sobre o espaço de regressores pode ser vista no quadro direito da figura. A função de previsão nesse exemplo pode ser

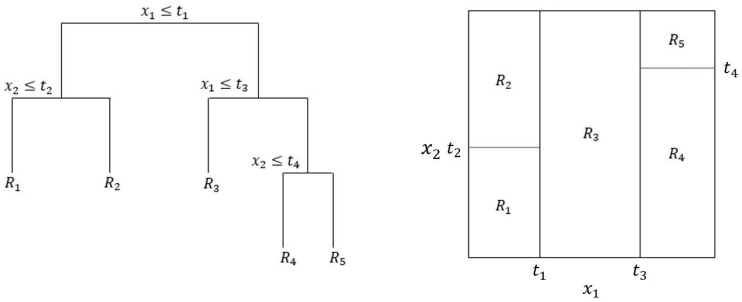


Figura 1 – Exemplo de Árvore de Decisão

A figura apresenta uma típica estrutura de árvore de decisão. Na esquerda tem-se a árvore propriamente dita; partindo do ponto superior, avalia-se se uma varíavel x_1 é menor ou igual a um valor limite t_1 e, caso seja, a variável passa para o lado esquerdo; caso contrário, passa para o lado direito, seguindo para subsequentes regras de separação. Os pontos no fim das regras são conhecidos como folhas R. Nas folhas, são alocadas diferentes regras preditivas. O quadro na direita apresenta o efeito que esse tipo de processo decisório tem no espaço de regressores. As sucessivas regras de decisão geram uma partição com as células tendo o formato de retângulos.

representada por uma função como: (2.3):

$$\hat{f}(X) = \sum_{m=1}^{M} w_m I\{(X1, X2) \in \mathbb{R}_m\}, \qquad (2.3)$$

onde w_m são as constantes utilizadas para previsão em cada célula R_m, e I é uma função indicadora que reporta em qual célula da partição a observação se encaixa.

Generalizando esse exemplo, os *splits* são encontrados por meio de um algoritmo baseado no gradiente. Considerando todos os pares de regressores

e regredidos (x_i, y_i) com $i = 1, 2, \ldots, N$ e $x_i = (x_{i1}, x_{i1}, ..., x_{ip})$, o processo de construção de árvores é feito buscando minimizar uma medida de erro. Tipicamente, as medidas de erro em problemas de classificação são:

$$\text{Erro de Má-Classificação} : 1/N_m \sum_{i \in R_M} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}, \quad (2.4)$$

$$\text{Índice de Gini} : \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \quad (2.5)$$

$$\text{Entropia Cruzada} : -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}. \quad (2.6)$$

Para efetivamente encontrar a variável e o ponto de divisão, considere começar da raiz e, recursivamente, dividir uma variável j no ponto s. As regiões geradas por essa divisão podem ser expressas por:

$$R_1(j, s) = \{X | x_j \leq s\}, \quad (2.7)$$

$$R_2(j, s) = \{X | x_j > s\}. \quad (2.8)$$

Os dois conjuntos definidos em (2.7) caracterizam um *split* baseado em uma regra do tipo "maior/menor do que". No exemplo, se a variável x_j é maior do que s, a observação flui para a região 2; caso contrário, flui para a região 1.

Para encontrar um *split* é necessário definir o seu objetivo. Em geral, define-se como objetivo minimizar um dos critérios de erro previamente expostos. Exemplificando

o problema com a entropia cruzada, o problema de minimização pode ser escrito como:

$$min_{j,s}[min_{w1} \sum_{x_i \in R_1(j,s)} \sum_{k=1}^{K} -(\hat{p}_{mk} log \hat{p}_{mk}) + \tag{2.9}$$

$$min_{w2} \sum_{x_i \in R_2(j,s)} \sum_{k=1}^{K} (-\hat{p}_{mk} log \hat{p}_{mk})], \tag{2.10}$$

em que K é o número de classes. O minimizador externo em (2.9) refere-se a qual das j variáveis será usada para fazer o *split* no ponto s. As duas minimizações internas escolhem quais constantes w_1 e w_2 devem ser alocadas em cada região para minimizar o erro.

2.1.5.2 Boosting Trees

Até Schapire (1991) publicar seu trabalho, a definição do tamanho que deveria ter a árvore de decisão era tomada através da filosofia de poda. Na ideia de poda, deixa-se que a árvore cresça um número exagerado de nós para depois retirar aqueles que não apresentam poder preditivo fora da amostra. Apesar da popularidade da técnica, ela gera um desempenho preditivo fraco, especialmente quando comparada a outros modelos de aprendizado de máquina como Máquinas de Vetor Suporte ou Redes Neurais.

A literatura de modelos conjuntos opta por combinar um conjunto de diferentes árvores em um único modelo para melhorar seu poder preditivo. A primeira proposta que será revisada é *boosting*, introduzida por Schapire (1991). Seguindo Ferreira e Figueiredo (2012), *boosting* é uma metodologia de conjuntos inspirada na ideia de que a combinação de aprendizes fracos, modelos com taxa de assertividade levemente maior do que a chance, podem gerar um aprendiz forte, isto é, um modelo com uma taxa de acerto superior a um limite preestabelecido. *Boosting* consiste em treinar aprendizes fracos sequenciais e posteriormente combiná-los para se gerar um aprendiz forte. A popularidade dessa técnica é explicada por ser mais fácil, computacionalmente, teinar muitos modelos simples ao invés de um modelo complexo.

Para desenvolver a intuição, defina $H_m : \mathscr{X} \rightarrow \{-1, +1\}$ como o *m*-ésimo aprendiz fraco binário, com $m = 1, ..., M$, sendo $x \in \mathscr{X}$ as variáveis utilizadas como regressores. Cada classificador representa um diferente jeito de combinar essas variáveis, e essas diferentes possibilidades serão simbolizadas por $H_1(x_1), ..., H_M(x_M)$. Aqui, é importante expor um conceito essencial de *boosting*. Quando se diz que *boosting* é a estimação de modelos sequenciais, isso significa que o modelo M depende do modelo $M - 1$. Em *boosting*, essa dependência é atingida

ao se ajustar a variável-alvo pelos erros cometidos pelo modelo prévio. A regra para fazer esse ajuste varia em cada aplicação, mas a filosofia se mantém.

Se cada classificador não estiver correlacionado aos demais, pode-se atingir uma boa qualidade preditiva considerando-se o voto majoritário dos diferentes modelos estimados. Se o conjunto de dados original for composto de M variáveis explicativas, usa-se para treinar cada modelo somente um subconjunto das M variáveis. Isso é feito para diminuir a covariância entre os modelos, permitindo que cada um capte uma diferente parte do verdadeiro processo gerador de dados. Assim, a combinação desses diferentes modelos atingirá uma taxa de erro menor do que qualquer modelo individual (SCHAPIRE; FREUND, 2012). Se forem utilizados os pesos $\alpha_1, ..., \alpha_M$ para cada aprendiz, pode-se representar o modelo $H : \mathscr{X} \rightarrow \{-1, +1\}$ como: $H(x) = sinal(\sum_{m=1}^{M} \alpha_m H_m(x))$, definindo o modelo final como uma combinação de M modelos individuais.

De acordo com Ferreira e Figueiredo (2012), *boosting* consiste em treinar uma sequência de modelos fracos. No presente trabalho, esses modelos serão árvores, e o algoritmo de estimação ajustará, sequencialmente, o *target* dos modelos sequenciais, dando mais importância às observações ainda não completamente explicadas. Pode-se, agora, juntar ambos os conceitos de *boosting* e árvores

seguindo a apresentação de Chen e Guestrin (2016). A implementação deles foi a efetivamente utilizada para a estimação deste trabalho devido a sua grande eficiência e agilidade.

Suponha que haja um conjunto de dados com n observações, m regressores x_i, $i = 1, \dots m$ e uma varíavel explicativa y_i, sendo este conjunto representado por $\mathscr{D} = \{(x_i, y_i)\}$ ($|\mathscr{D}| = n_i$, $x_i \in R^m$, $y_i \in \{0, 1, \dots, k\}$), em que k é o numero de classes.

Já se sabe como árvores funcionam: elas dividem o espaço de regressores em hiper-retângulos e designam uma constante real w_j para cada célula da partição, de forma que dado um vetor $x \in R_j$ $f(x) = w_j$. Pode-se representar as árvores como $T(x; \Theta) = \sum_{j=1}^{J} w_j I(x \in R_J)$, sendo que $\Theta = \{R_j, w_j\}_1^J$, é a coleção de parâmetros que contém os R_j hiper-retângulos e as w_j constantes usadas para prever nas J regiões. Usualmente, J é tratado como um hiperparâmetro previamente escolhido.

Para encontrar esses parâmetros, é necessário definir a função objetivo. Em aprendizado de máquina, a função tem, tipicamente, duas partes: uma medida de erro e um termo de regularização:

$$obj(\theta) = \mathscr{L}(\theta) + \Omega(\theta), \tag{2.11}$$

O primeiro termo $\mathscr{L}(\theta)$ é o responsável por medir o erro,

e $\Omega(\theta)$ é responsável por controlar o excesso de ajuste, regulando a complexidade da árvore.

Boosting é uma filosofia de modelagem que combina diferentes aprendizes fracos, buscando construir um único aprendiz forte. Pode-se representar a combinação de K árvores como: (2.12):

$$\hat{y}_i = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \;\; f_k \in \mathscr{F}, \tag{2.12}$$

em que $\mathscr{F} = \{f(x) = w_{q(x)}\}$ $(q : R^m \Rightarrow T, w \in R)$ restringe os modelos a serem árvores de decisão e, consequentemente, definem o modelo $\phi(x_i)$ como a soma de K árvores. Em outras palavras, para cada uma das k árvores construídas, será avaliado o valor a ser previsto de acordo com os regressores observados x_i, e o valor previsto final $\hat{y}_i$ será a soma dos $f_k(x_i)$ valores previstos em cada uma das k árvores. Assim, pode-se reescrever a função objetivo como:

$$Obj = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{2.13}$$

em que se substitui a função de erro global $\mathscr{L}(\theta)$ pela soma de i funções que medem o erro individualmente para cada previsão.

As árvores são encontradas minimizando-se a função objetivo. De acordo com Friedman, Hastie e

Tibshirani (2001), minimizar (2.13) com respeito a $f(x)$ pode ser visto como um problema de minimização numérica, no qual se busca:

$$\hat{\boldsymbol{f}} = argmin_{\boldsymbol{f}}\, Obj(\boldsymbol{f}).$$

Cada árvore passa a ser vista como um parâmetro a ser estimado e $f \in R^N$ são as aproximações desses parâmetros $f(x_i)$. Dados os n pontos disponíveis, consegue-se estimar essa sequência de árvores $\hat{f} = f(x_1), f(x_2), ..., f(x_n)$.

Analiticamente, esse problema não está bem-definido e é necessário um algoritmo iterativo de otimização numérica para resolver o problema, usualmente como uma soma de árvores. Em cada iteração t, adicionaremos uma árvore à sequência:

$$\begin{aligned}
&\hat{y}_i^{(0)} = 0,\\
&\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i),\\
&\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i),\\
&\dots\\
&\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i),
\end{aligned} \tag{2.14}$$

sendo $f_t(x_i)$ a função que minimiza a função objetivo em cada consecutivo passo da iteração. Podemos representar o objetivo de cada uma dessas t árvores como:

$$\mathscr{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{2.15}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda |||w||^2, \tag{2.16}$$

sendo l uma função perda convexa que mede a qualidade das previsões. No nosso problema de classificação, usualmente essa função é a *mlogloss* (equação 2.6).

O elemento de penalização (2.16) suaviza a função ajudando a evitar o excesso de ajuste. Quando se tenta minimizar (2.15) com mais árvores T ou folhas w, a adição proposta tem que gerar um queda significativa na função perda l para contrabalancear o crescimento de Ω. Esse mecanismo leva o algoritmo a minimizar a função objetivo através das mais simples estruturas.

Pode-se demonstrar que, por meio de uma expansão de Taylor de segunda ordem, a otimização pode ser reescrita em cada passo t como:

$$\sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t), \tag{2.17}$$

$$g_i = \partial_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{(t-1)}), \tag{2.18}$$

$$h_i = \partial^2_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{(t-1)}), \tag{2.19}$$

em que (2.17) se torna a função objetivo regularizada para a árvore subsequente, e (2.18) e (2.19) são, respectivamente, o gradiente e a hessiana da função

objetivo. Substituindo a função de regularização por (2.16), chega-se à equação final que define os pesos atribuídos a cada folha e a função objetivo regularizada:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{2.20}$$

$$Obj* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{2.21}$$

$$G_j = \sum_{i \in I_j} g_i, \qquad H_j = \sum_{i \in I_j} h_i. \tag{2.22}$$

Esse conjunto de equações gera uma série de informações sobre as árvores aprendidas. A equação (2.20) dá os valores a serem previstos em cada uma das j folhas. A equação (2.21) é uma função pontuação, que mede a qualidade da estrutura das árvores q. É importante ressaltar o fato de que ambas as equações (2.20) e (2.21) são diretamente calculáveis para qualquer função perda, com primeira e segunda derivadas bem definidas.

De acordo com Chen e Guestrin (2016), é, em geral, impossível enumerar todas as q possíveis estruturas, fazendo-se necessário implementar um algoritmo que inicia da raiz e, iterativamente, adiciona novos nós. Supondo que I_l e I_r são os conjuntos gerados após um *split*, a função perda da divisão se torna:

$$\mathscr{L}_{split} = \frac{1}{2}\left[\sum_{j=1}^{T}\frac{(\sum_{i\in I_L} g_i)^2}{H_j+\lambda} + \sum_{j=1}^{T}\frac{(\sum_{i\in I_R} g_i)^2}{H_j+\lambda} - \right. \quad (2.23)$$

$$\left.\sum_{j=1}^{T}\frac{(\sum_{i\in I} g_i)^2}{H_j+\lambda}\right] - \gamma \quad (2.24)$$

A equação (2.23) é a formula implementada para encontrar os candidatos a um *split*. A primeira soma dentro dos colchetes mede a qualidade do modelo usando as observações que fluíram para o galho da direita. A segunda soma mede a qualidade com as observações da esquerda, e a última soma mede a qualidade na folha antes de o *split* ser feito. A equação (2.23) da uma medida da contribuição que cada split teve para melhorar a qualidade preditiva.

Toda a simulação foi feita na linguagem R Core Team (2013), para a estimação do modelo *Boosting Trees* foi utilizado o pacote de Chen et al. (2018), sendo que para a definição do número de árvores foi utilizada a função de validação cruzada nativa do pacote, todos os demais hiper-parâmetros não foram otimizados sendo utilizados os *defaults* do pacote.

2.1.5.3 Random Forest

Random forest é outra metodologia de conjuntos de árvores. Como ela é uma extensão de *bagging*,

é importante apresentar esse conceito primeiro. A palavra *bagging* é a conjunção das palavras *bootstrap* e *aggregation*. *Bootstrap* é uma famosa técnica de inferência utilizada para avaliar a precisão em um parâmetro. Aqueles acostumados com estatística clássica sabem que, usualmente, inferência é feita após serem realizadas algumas hipóteses a respeito da distribuição de probabilidade de algum parâmetro. *Booststrap* faz inferência de uma maneira diferente. Supondo um conjunto de treinamento com N observações, amostrar-se-á, iterativamente, com reposição, $b < N$ observações em cada rodada B do *bootstrap*. Com cada uma das B amostras, avalia-se a mudança no parâmetro de interesse, como, por exemplo, uma previsão pontual $\hat{y}_i$. As diferentes B amostras gerarão diferentes previsões $\hat{y}_i$ e, pela metodologia *Bootstrap* usaremos a $Var(\hat{y}_i)$ como a estimativa da variância dessa previsão (DAVISON; HINKLEY, 1997).

Bagging é inspirado nesse conceito. Suponha que se queira regredir y_i em m variáveis $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{m,i})$, com $i = 1, \ldots, N$, usando o modelo $\hat{f}(x)$. Novamente, será selecionada uma amostra de tamanho b do conjunto de dados B vezes, e se estimará B modelos $\hat{f}^{*B}(x)$ para cada amostra. Diferentemente da explicação anterior, não há interesse especial na variância

das previsões, mas em suas médias. *Bagging* é uma metodologia de combinação de previsões que agrega B previsões e usa a média delas para gerar uma previsão final:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{g=1}^{B}\hat{f}^{*g}(x). \tag{2.25}$$

Essa estimativa empírica converge para a verdadeira função conforme $B \to \infty$.

Os modelos que mais se beneficiam desse procedimento são aqueles com baixo viés e alta variância, isto é, grandes árvores de decisão. Árvores são conhecidas pela alta variância de seus parâmetros, e a adição ou exclusão de uma única observação pode fazer um *split* ser preferido a outro, gerando, potencialmente, modificações em todos os *splits* subsequentes. Mas, grandes árvores - aquelas com vários nós - têm, usualmente, baixo viés, característica que, nos estágios iniciais dos modelos de árvores, causou muitos problemas de excesso de ajuste.

Suponha que haja B árvores de classificação estimadas $\hat{T}_B(x)$, sendo cada uma estimada com uma diferente amostra aleatoriamente selecionada. O vetor resposta dessas árvores de dimensão $K-1$ fica povoado com zeros, contendo apenas o valor 1 na célula k, que corresponde à classe prevista. Se todas as células

forem iguais a zero, prevê-se a classe base K. *Bagging* constrói a árvore conjunta $\hat{f}_{bag}(x)$ como um vetor $[p_1(x), p_2(x), \ldots, p_K(x)]$, em que $p_k(x)$ é a proporção de modelos que preveem a classe k. O classificador combinado preverá a classe que tem o maior numero de votos das B árvores $\hat{T}_{bag}(x) = argmax_k \hat{f}_{bag}(x)$. Esse simples procedimento melhora, consideravelmente, o desempenho preditivo das árvores.

A variância da média $\tilde{\sigma}_B^2$ de B estimativas independentes, e igualmente distribuídas, decresce de acordo com B conforme $\tilde{\sigma}_B^2 = \sigma^2/B$, onde σ é a variância do modelo. Mas, caso as estimativas não sejam independentes, essa quantidade aumenta proporcionalmente com a correlação ρ entre os modelos, $\tilde{\sigma}^2 = \rho\sigma^2 + [(1-\rho^2)\sigma^2]/B$. Em *Bagging*, o procedimento de *bootstrap* gera B amostras do mesmo processo gerador de dados, e *Random Forests* foca em diminuir a variância das previsões, diminuindo a correlação entre os modelos. Ao construir as árvores de classificação, em cada *split*. Muda-se o processo gerador de dados através da alternância das variáveis candidatas. Em cada *split*, serão selecionadas, aleatoriamente, $r < m$ variáveis elegíveis para o *split*. Esse procedimento de aleatoriarização tenta descorrelacionar ainda mais os B modelos. Após o crescimento das B árvores $\{T(x; \Theta_g)\}_1^B$, o modelo *random forest* $\hat{T}_{rf}$ é

encontrado pegando o voto majoritário entre todos os modelos, similar ao procedimento de *bootstrap*:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b) \tag{2.26}$$

$$\hat{T}_{rf}(x) = argmax_k \hat{f}_{rf}(x) \tag{2.27}$$

Esse procedimento, aparentemente simples, tem grande impacto na qualidade preditiva. *Random forest* é uma metodologia moderna para realizar previsões acuradas, e tem inúmeras aplicações famosas em diferentes problemas de previsão. O número e a profundidade das árvores são, tipicamente, encontrados por validação cruzada (validação cruzada será apresentada na seção 2.1.5.5). Uma característica importante de *random forest* é que, comumente, a adição de novas árvores não aumenta o erro de previsão. Quando se olha para a fórmula da variância, pode-se ver que, conforme $B \to \infty$, a variância sempre decrescerá, convergindo para $\rho\sigma^2$ e, como comentado anteriormente, o procedimento de aleatoriarização trabalha para diminuir ρ. A regra de parada ocorre, usualmente, quando a adição de uma nova árvore apenas aumenta a complexidade do modelo, sem diminuir a variância das previsões.

Para a estimação do modelo *Random Forests* foi utilizada a biblioteca de LeDell et al. (2019). Todos os

hiperparâmetros utilizados foram os nativos do pacote, a função de importância de variáveis nativa do pacote foi utilizada para mensuração da importância das variáveis.

2.1.5.4 Importância das Variáveis

Uma das características mais interessantes dos modelos baseados em árvores é a possibilidade de unir alta capacidade preditiva com compreensibilidade. Foi visto que a equação (2.23) gera a diferença entre o nível de erro antes e após um *split* ser feito. Agora, chama-se esse acréscimo de $\hat{i}$. Ao somar todos os acréscimos de qualidade $\hat{i}_t$ ocorridos nos $t = 1, \ldots, K-1$ *splits*, pode-se resumir a importância que dada variável teve na estimação do modelo por:

$$\mathscr{I}_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2, \tag{2.28}$$

em que $\mathscr{I}_l^2(T)$ é a importância que uma dada variável teve na estimação da árvore T. A extensão para um modelo de conjunto de árvores é feita de maneira intuitiva. Se tiver M árvores, a importância que uma dada variável tem para a estimação do modelo completo é:

$$\mathscr{I}_l^2 = \frac{1}{M} \mathscr{I}_l^2(T_m). \tag{2.29}$$

A verificação do nível de importância que uma dada variável tem na estimação do modelo será feita com

base na contribuição da variável para a redução do nível de erro calculado durante o processo de estimação. Esse tipo de medida é altamente influenciada pelos hiperparâmetros escolhidos para realizar a regularização, como pode ser verificado nas equações (2.20) e (2.23).

Além disso, esses valores carecem de intuição. Um procedimento comum para facilitar a comparação é escalar o valor máximo como um número b e transformar os outros valores nas respectivas proporções em relação ao valor inicialmente estimado. Valores comuns para b são 1 ou 100.

2.1.5.5 Validação Cruzada

Métodos de aprendizado de máquina facilmente se tornam superajustados aos dados. A alta capacidade de generalização da função $f(x)$ se mostra uma faca de dois gumes: a função se mostra versátil o suficiente para capturar correlações altamente não lineares em espaços multidimensionais, mas, não há como afirmar se esse complexo padrão é de fato gerador de causalidade, ou se é mero acaso.

Para solucionar esse problema se desenvolveu a literatura de validação de modelos, sendo que um de seus principais produtos foi a concepção da metodologia de validação cruzada. Essa metodologia tenta estimar o

erro $Err = E[L(Y, \hat{f}(X))]$ que o modelo estimado $\hat{f}(x)$ teria com um conjunto de dados $\{X, Y\}$ fora da amostra utilizada na estimação. Na sua concepção mais simples, suponha que peguemos o nosso conjunto de dados de tamanho N e dividamos esse conjunto em K subconjuntos de igual tamanho: valores comuns para K são $\{5, 10, N\}$. Estima-se o modelo $\hat{f}(x)$ com $K - 1$ partes da amostra e posteriormente utiliza-se esse modelo estimado para realizar previsões sobre as observações não utilizadas na estimação.

Sucessivamente se retira um bloco $K = 1, 2, \ldots, k$ da amostra e coloca-se aquele que estava de fora para o conjunto de estimação, reestima-se o modelo e utiliza-se dele para realizar novas previsões no conjunto de dados que ficou fora da amostra na respectiva rodada. Esse procedimento está exemplificado na Figura 2:

Conforme Tibshirani (1996), podemos definir uma função $\mathscr{K} : \{1, \ldots, N\} \longmapsto \{1, \ldots, K\}$ que indica a qual dos K conjuntos cada observação pertence. Denote por $\hat{f}^{-\mathscr{K}(i)}(x)$ a função estimada sem as observações pertencentes ao k-ésimo conjunto. A estimativa de erro de validação cruzada é:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\mathscr{K}(i)}(x_i^{\mathscr{K}})). \tag{2.30}$$

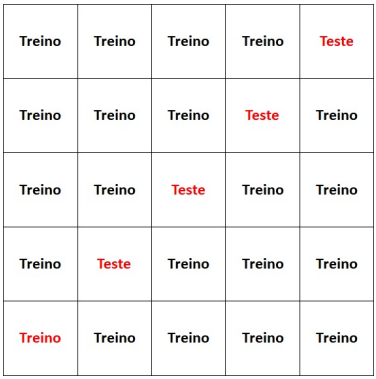


Figura 2 – Ilustração da Validação Cruzada
A figura ilustra o procedimento de separar uma amostra em K diferentes subamostras e utilizar apenas uma parte dessas subamostras para treinar um modelo, posteriormente utilizando esse modelo para realizar previsões para as observações que ficaram fora do grupo de treino.

onde $L(.,.)$ é uma função perda, medidora da qualidade das previsões.

Muitos modelos de aprendizado de máquina têm em sua especificação um hiperparâmetro. Hiperparâmetros são parâmetros não estimáveis e que servem como uma ferramenta de controle que o pesquisador tem sobre a versatilidade do modelo. Como esses parâmetros não são estimados, e quase sempre carecemos de uma boa estimativa a priori sobre eles, se faz necessário analisar múltiplas possibilidades para verificar aquela que tem o melhor comportamento com o conjunto de dados que está sendo analisado. Isso é feito muitas vezes por meio de

validação cruzada.

Seja o modelo estimado por validação cruzada $\hat{f}^{-\mathscr{K}(i)}(x, \alpha_i)$ indexado pelo hiperparâmetro de interesse α_i. Conforme varia-se α_i, o valor de (2.31) também variará:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\mathscr{K}(i)}(x_i^{\mathscr{K}}, \alpha_i)). \qquad (2.31)$$

O hiperparâmetro final escolhido será aquele que atender a alguma medida de qualidade definida pelo pesquisador, por exemplo, o hiperparâmetro que gerar o menor valor para (2.31). Devido a problemas de excesso de ajuste, Tibshirani (1996) recomenda que não se use o valor que atinge o menor valor para a função (2.31), mas que se estime o desvio-padrão do erro no ponto de mínimo e some-se esse desvio-padrão ao valor do erro. O autor recomenda que se utilize o hiperparâmetro mais parcimonioso que gera o valor dessa soma na função erro, geralmente produzindo um modelo mais parcimonioso.

Em séries de tempo existe um detalhe adicional na separação dos conjuntos de teste e treinamento. Tipicamente o produto final de um modelo preditivo de séries de tempo é um modelo de extrapolação, e não interpolação, ou seja, queremos prever o futuro e não o presente ou o passado. Utilizar uma amostra com informações do passado e do futuro para explicar

o presente torna a estimativa de erro potencialmente subviesada: é mais difícil extrapolar do que interpolar. Por isso, desenvolveram-se procedimentos específicos para validação cruzada de séries de tempo, como a utilização de janelas expansivas e janelas rolantes.

Em janelas expansivas separa-se uma fatia dos dados originais para servir de conjunto de validação da primeira janela. Esse conjunto será composto pelas últimas $N - k$ observações da amostra. O conjunto de dados que será utilizado para a estimação corresponde as k primeiras observações, ordenadas pelo tempo. Com esse conjunto de estimação, estima-se um modelo que é posteriormente utilizado para realizar as previsões sobre as últimas $N - k$ observações. Em seguida, adiciona-se a k-ésima observação ao conjunto de treinamento e a retira do conjunto de teste; reestima-se um novo modelo e utiliza-se dele para prever as últimas $N - k - 1$ observações. Os erros de previsão dos sucessivos $N - k$ modelos estimados servirão de subsidio para estimar-se o erro de previsão fora da amostra. Esse procedimento é conhecido como validação por meio de janela expansiva, pois o conjunto de treinamento vai se expandindo conforme as previsões vão se realizando. O procedimento está ilustrado na Figura 3:

Uma variante dessa abordagem é a utilização de janelas móveis ao invés de janelas expansivas. Na utilização

de janelas móveis o conjunto de treinamento se mantém do mesmo tamanho ao longo de toda a validação. Para isso, em toda a rodada da validação ao mesmo tempo em que adiciona-se uma observação ao conjunto de treinamento também retira-se a observação mais antiga do mesmo, sendo o modelo reestimado sucessivas vezes com uma amostra de mesmo tamanho. Esse procedimento está ilustrado na figura 4.

O principal benefício da utilização de janelas móveis ocorre quando temos um número de observações grande. Quando o número de observações do primeiro conjunto de treinamento já é exagerado, pequenos são os benefícios em se adicionar mais informação. Ainda, o fato do conjunto de treinamento se manter o mesmo nos permite verificar se houve períodos de maior previsibilidade na amostra. Na utilização de janelas expansivas, a aparente melhoria de um modelo pode se dever à expansão do conjunto de treinamento e consequentemente melhor estimação dos parâmetros, e não devido a características específicas do período. Neste trabalho utilizaremos o conceito de janelas móveis para realizar as sucessivas estimações e previsões.

Todos os modelos que serão utilizados neste capítulo contém hiperparâmetros que controlam a sua versatilidade. Para a definição do melhor hiperparâmetro

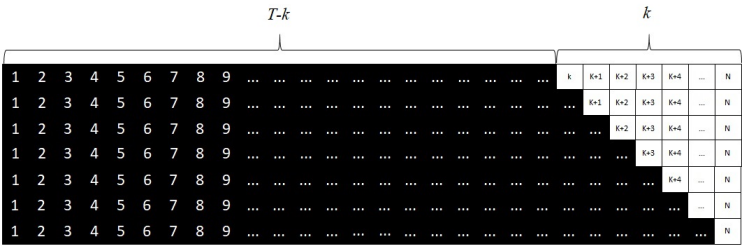


Figura 3 – Ilustração de Validação Cruzada com Janelas Expansivas

A figura ilustra a separação da amostra em conjunto de treinamento e de teste. As caixas pretas mostram as observações utilizadas para reestimar o modelo; as caixas brancas ilustram a amostra selecionada para se avaliar o modelo.

será separada uma amostra com os $T - k$ dias iniciais, e se utilizará da técnica de validação cruzada clássica para se escolher os hiperparâmetros. Uma vez definidos os hiperparâmetros será utilizada a técnica de validação cruzada com janelas expansivas para se estimar o erro fora da amostra de cada função. Sendo assim, realizaremos a escolha dos hiperparâmetros com a amostra original de tamanho $T-k$ e utilizaremos aqueles escolhidos para todas as sucessivas k estimações, visando verificar a qualidade preditiva da função.

2.1.6 Dados

Este trabalho se dedicará a analisar a dinâmica do ativo PETR4. O filtro inicial para essa escolha foi baseado

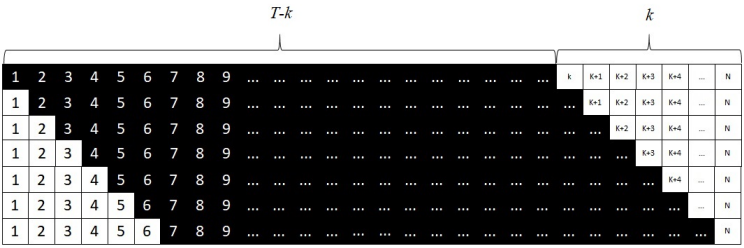


Figura 4 – Ilustração de Validação Cruzada com Janelas Móveis

A figura ilustra a separação da amostra em conjunto de treinamento e de teste. As caixas pretas mostram as observações utilizadas para se estimar o modelo; as caixas brancas ilustram a amostra selecionada para se avaliar o modelo.

na disponibilização dos arquivos de alta frequência pela bolsa brasileira B3.SA. Perlin e Ramos (2016) criaram um sistema que torna simples o tratamento e manipulação destes dados, o que contribuiu sobremaneira para a realização deste trabalho.

A escolha específica pelo ativo Petrobrás se deu por dois motivos. Dado o volume de dados que apenas um dia de negociação gera, analisar múltiplos ativos torna-se inviável. O segundo motivo é o interesse específico pela dinâmica de negociação da companhia. Recentemente, a mesma esteve em grande destaque na mídia brasileira e internacional, devido a consecutivos escândalos de corrupção e, após esse período, a gestão interna de empresa passou a anunciar ações de reestruturação

interna e relevante política de desinvestimentos. A grande sequência de fatos relevantes torna a dinâmica desse ativo especialmente interessante, tanto pela possibilidade de alto volume de informações privilegiadas como por consecutivos deslocamentos de manada dos *traders* de varejo.

A base de dados inicia em 01/12/2015 e vai até 12/06/2018. Com esse conjunto de dados, construiu-se as variáveis descritas na Tabela 1.

Inicialmente, será feita uma análise da qualidade preditiva do modelo em múltiplas frequências, para depois verificar a importância das variáveis. Isso se mostrou necessário, pois, não há uma conclusão na literatura de qual é o melhor horizonte para se fazer a previsão. Para realizar a análise da melhor frequência construiu-se 60 bases de dados com variáveis agregadas em intervalos que vão de 1 a 60 minutos e, a cada base de dados consecutiva, a agregação aumenta em 1 minuto.

Para adicionar a estrutura de série de tempo, trabalhou-se com variáveis defasadas, além de adicionar, como variável explicativa, há quanto tempo o pregão estava aberto. Devido a estrutura da bolsa brasileira, um dia típico de negociação tem duração de 7 horas, já que a bolsa abre às 10:00 e fecha às 17:00, optou-se por adicionar apenas 2 defasagens das variáveis explicativas à matriz de regressores, de maneira que, nas maiores frequências,

Tabela 1 – Variáveis

Horário	Último Preço
Retorno	Sinal do Retorno
Volatilidade do Retornos**	Número de Negócios*
Quantidade de Ações Negociadas*	Volume *
Preço Ponderado por Volume*	Número de Novas ordens*
Número de Ordens Atualizadas *	Número de Ordens Canceladas*
Preço Máximo das Ofertas *	Preço Mínimo das Ofertas*
Preço Ponderado das Ofertas*	

* Apesar de estarem listadas como se fossem apenas uma variável, criou-se duas colunas para cada variável: uma para ordens de compra e outra para ordens de venda;

** Como estimativa de volatilidade se utilizou uma adaptação do estimador de volatilidade realizada de Andersen et al. (2001). Dentro de cada base de dados se utilizou como estimativa de volatilidade de cada observação, o somatório dos 10 retornos quadrados igualmente espaçados que compuseram aquela observação. Em todas as observações se verificou quais foram os dez retornos, igualmente espaçados, que ocorreram dentro do período de tempo daquela observação, tomou-se seus quadrados e utilizou-se a soma deles como uma estimativa de volatilidade.

ainda haverá 4 horas de negociação, pelo menos, a serem previstas.

Em vez de amostrar os dados uniformemente, optou-se por amostrá-los a cada um minuto,

independentemente da agregação temporal das variáveis. Logo, em todas as bases de dados as previsões serão feitas a cada um minuto, ainda que o horizonte de previsão mude. O horizonte iniciará em 1 minuto e avançará até 60, evoluindo sempre de 1 em 1 minuto. Por exemplo, na base de dados de 30 minutos, a primeira previsão será feita verificando o sinal da variação do preço entre 11:30 e 12:00; a segunda previsão será feita as 11:31, prevendo o sinal do retorno até 12:01, e assim sucessivamente.

A janela de previsão de todas as frequências será restrita para o espaço da maior, de 60 minutos. Isso será feito para que todas as diferentes bases sejam comparadas durante o mesmo período do dia e passem a ter o mesmo número de previsões. Caso isso não fosse feito, as frequências menores iriam ter substancial vantagem aumentando o número de previsões.

Após avaliar em qual frequência há a melhor capacidade preditiva, optar-se-á por aquelas que apresentarem melhores propriedades para realizar a investigação da importância das variáveis.

2.1.7 Comparação de Frequências

Ambos os modelos, *boosting trees* e *random forest* serão estimados utilizando os conjuntos de dados descritos na Seção 2.1.6. Como já comentado, serão

avaliados os impactos que diferentes frequências de amostragem têm sobre o desempenho das previsões, visando encontrar uma boa frequência para fazer a análise de variáveis.

Ao longo da análise do desempenho estatístico das previsões, será realizada a comparação dos modelos com base em duas estatísticas descritivas. A primeira avaliará o desempenho global do modelo, e a segunda avaliará o desempenho do modelo em prever e classificar as classes extremas.

Uma previsão positiva é uma previsão na qual se afirma que uma observação pertence a uma classe, e uma previsão negativa é aquela na qual se afirma que uma observação não pertence a uma classe. Pode-se definir VP (verdadeiro positivo) como o número de vezes que se fez uma previsão positiva e se acertou essa previsão. De maneira similar, pode-se definir FN (falso negativo) como o número de vezes no qual se fez uma previsão negativa e se errou a previsão. De maneira análoga, define-se FP (falso positivo) e VN (verdadeiro negativo). As medidas de qualidade das previsões são:

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}, \tag{2.32}$$

$$ACC_{p,n} = \frac{TP_{p,n} + TN_{p,n}}{TP_{p,n} + FP_{p,n} + FN_{p,n} + TN_{p,n}}, \quad (2.33)$$

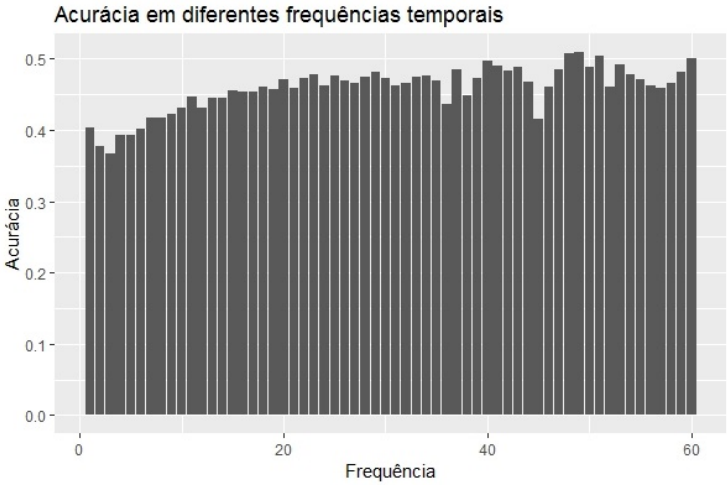
A equação (2.32) é a definição de acurácia, que pode ser entendida como uma descrição dos erros sistemáticos do modelo, ou como uma combinação de dois tipos de erro: o erro aleatório, no qual se erra por mudanças desconhecidas e imprevisíveis no processo gerador de dados, e o erro sistemático, devido a imperfeições na própria estimação do modelo.

A equação (2.33) é uma versão modificada da acurácia chamada de acurácia modificada. Nela, serão excluídas todas as previsões feitas para a classe nula. Desse modo, o intuito é clarear a capacidade do modelo em prever movimentos direcionais, excluindo todas as previsões de estabilidade, o que permitirá avaliar apenas as previsões realizadas para as classes positiva e negativa.

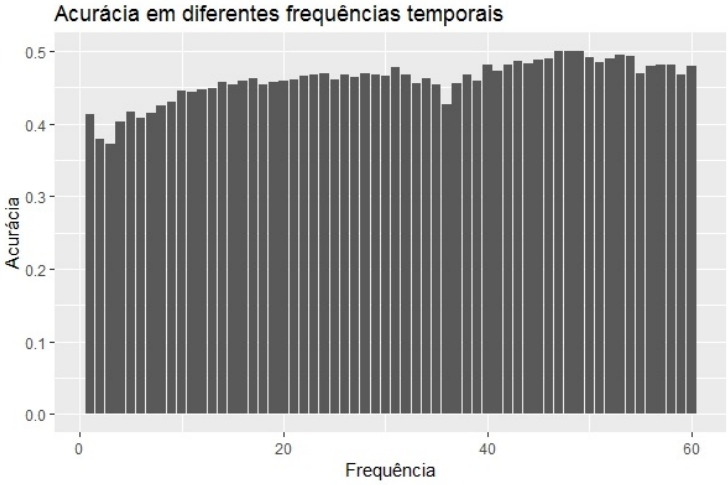
2.2 Resultados

2.2.1 Análise de Frequência

Inicialmente será verificado o impacto que o horizonte de previsão tem na capacidade preditiva de ambos os modelos. A Figura 5 apresenta a variação da acurácia das previsões nos dois modelos analisados.



(a) *Boosting Trees*



(b) *Random Forests*

Figura 5 – Acurácia das Previsões

Evolução da acurácia dos dois modelos quando avaliados prevendo o sinal dos retornos um passo a frente. A figura superior apresenta a acurácia do modelo *boosting trees* ao prever o sinal dos retornos um passo a frente. O gráfico inferior apresenta a mesma informação para o modelo *random forest*.

A Figura 5 mostra que, ao se expandir o horizonte de previsão, conjuntamente à agregação das variáveis, tem-se uma evolução gradual da qualidade preditiva dos modelos. O ápice dos dois modelos não ocorre nos extremos, mas um pouco antes, próximo à frequência de 50 minutos.

Já a Figura 6 apresenta o desempenho preditivo de ambos os modelos sob a ótica da acurácia modificada. Como já exposto, a modificação exclui as previsões feitas para a classe nula, evidenciado a capacidade preditiva exclusivamente de movimento direcionais, isto é, movimentos em que o preço sobe ou desce. É possível verificar na figura que o ápice preditivo de ambos os modelos ocorre na menor frequência possível, atingindo ainda um valor diferenciado na frequência de dois minutos.

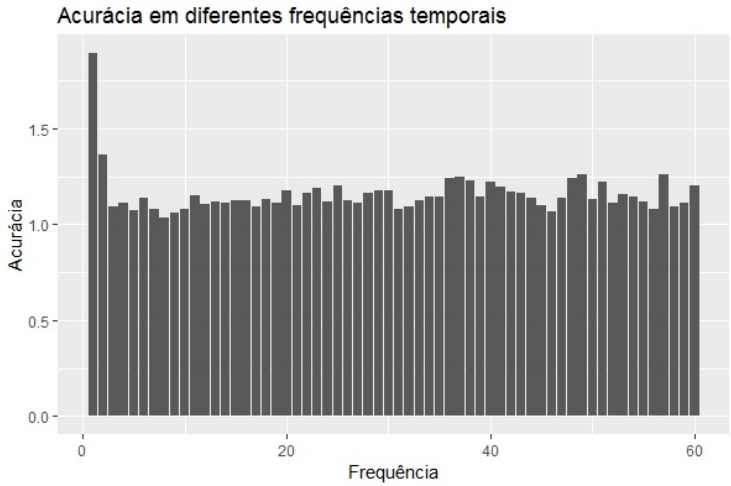
Essas duas informações, aparentemente contraditórias, são facilmente conciliáveis. Conforme o horizonte temporal aumenta, diminuem os casos de retornos nulos. Conjuntamente a isso, comum em modelos frequentistas, as árvores passam a adotar um menor número de previsões de retorno nulo. Isso faz com que o problema de previsão se distancie gradativamente de um problema de previsão em três classes, relativamente homogêneas, para um problema altamente concentrado em duas classes. Essa mutação do problema é o que faz os

modelos terem, aparentemente, um melhor desempenho na acurácia global tendo uma capacidade preditiva próxima a 50%, mas, conforme isso ocorre, o nível de acerto ingênuo de ambos os modelos também muda. Quando se começa com um problema de previsão em 3 classes, o *benchmarking* ingênuo é que as previsões têm que acertar em mais de 33% dos casos; já quando se tem um problema com duas classes majoritárias, busca-se uma qualidade preditiva maior do que 50%.

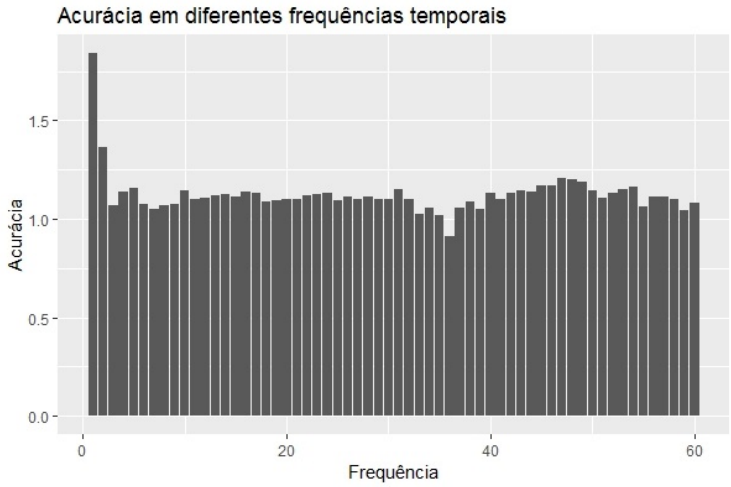
A acurácia modificada vem para qualificar essa aparente melhoria. Apesar de os modelos, em geral, errarem menos, não têm uma consistente melhora da acurácia modificada, indicando que a melhor capacidade preditiva pouco se relaciona com o aprendizado de padrões de previsão nas classes extremas, mas ao aprendizado que a classe intermediária passa a ser relativamente rara. Dada à dualidade de sinais, optou-se por passar para a próxima etapa, a etapa de análise de variáveis, com as duas frequências extremas, verificando as variáveis agregadas em intervalos de 1 minuto e em intervalos de 60 minutos.

2.2.2 Importância das Variáveis

A primeira pergunta que precisa ser respondida é: quais variáveis dentre aquelas apresentadas na Tabela 1 são as mais relevantes para a estimação dos modelos baseados



(a) *Boosting Trees*



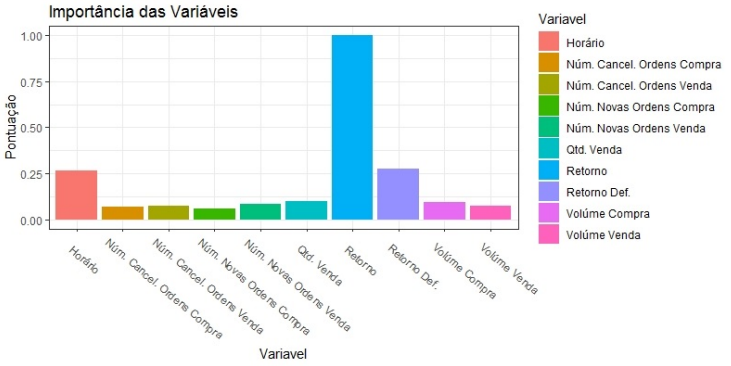
(b) *Random Forests*

Figura 6 – Acurácia Modificada das Previsões
A figura apresenta a diferença da acurácia dos dois modelos quando avaliados prevendo o sinal dos retornos um passo a frente. O modelo acerta mais do que erra quando o valor verificado é maior que 1. A figura superior apresenta a acurácia do modelo *Boosting Trees* ao prever o sinal dos retornos um passo a frente, a figura inferior apresenta a informação para o modelo *Random Forests*.

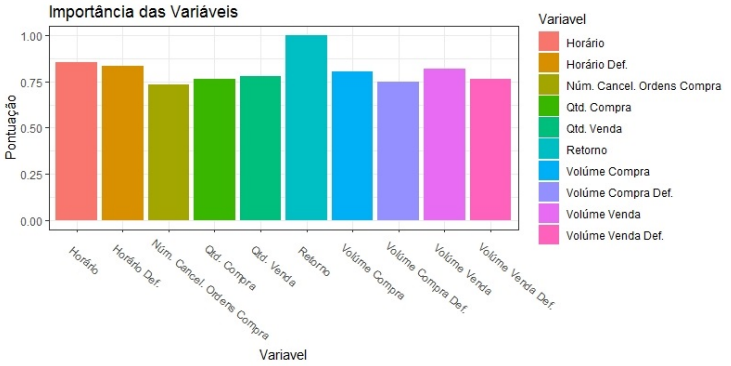
em árvores? Como já comentado, foram incorporadas aos modelos duas defasagens de todas as variáveis descritas na tabela. Além disso, muitas delas, conforme sinalizado, foram separadas entre variáveis que ocorreram em ordens de compra ou de venda. Dado esse grande número de variáveis, o foco da análise foi identificar as dez mais importantes variáveis.

A Figura 7 mostra as 10 variáveis mais importantes para ambos os modelos, quando são feitas previsões 1 minuto à frente, os valores reportados foram reescalados para que a variável mais importante tivesse importância igual a um, os valores originais foram calculados com base no ganho de informação. Pode-se ver um comportamento muito diferente dos dois modelos: enquanto *boosting trees* apresenta uma variável com extrema distinção das demais, *random forest* coloca uma importância muito similar para todas as variáveis.

Pode-se ver que, para ambos os modelos, a variável que mais se destaca é o retorno do último período, sendo que em *boosting trees* a primeira defasagem do retorno e o horário da negociação também tem elevada importância. Para *random forest* o horário da negociação também aparece com grande relevância, mas outras variáveis como volume de negociação e numero de ordens não parecem ter importância muito diferente.



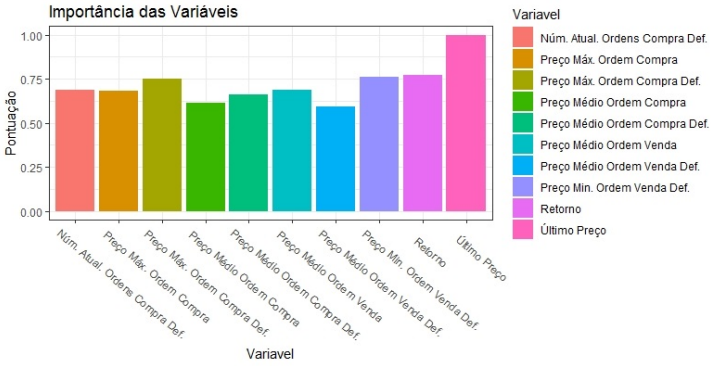
(a) *Boosting Trees*



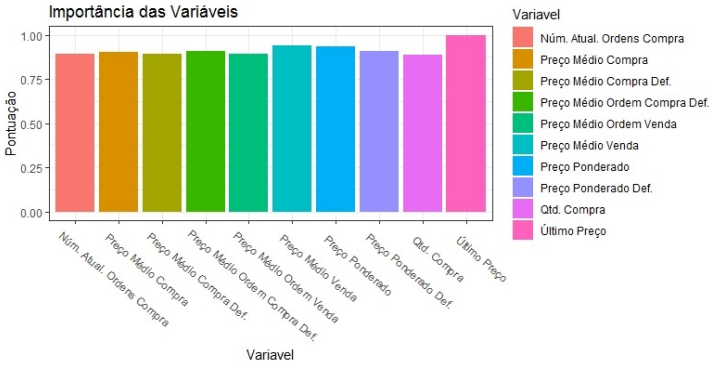
(b) *Random Forests*

Figura 7 – Importância das Variáveis - Horizonte de 1 Minuto

10 variáveis mais importantes para as previsões com horizonte de 1 minuto à frente. O quadro superior apresenta as 10 variáveis mais importantes para o modelo *boosting trees* e o quadro inferior apresenta as 10 variáveis mais importantes para o modelo *random forest*.



(a) *Boosting Trees*



(b) *Random Forests*

Figura 8 – Importância de Variáveis - Horizonte de 60 Minutos

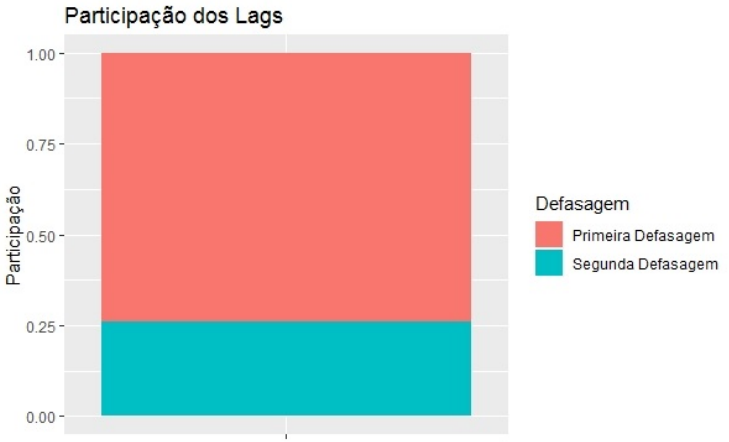
10 variáveis mais importantes para as previsões com horizonte de 60 minutos à frente. O quadro superior apresenta as 10 variáveis mais importantes para o modelo *boosting trees* e o quadro inferior apresenta as 10 variáveis mais importantes para o modelo *random forest*.

Ao ampliar o horizonte de previsão para 60 minutos à frente, a diferença da importância entre as variáveis cai drasticamente. A Figura 8 mostra que a variável mais importante para ambos os modelos é o último preço negociado, mas, agora, há pouca diferença em relação às demais.

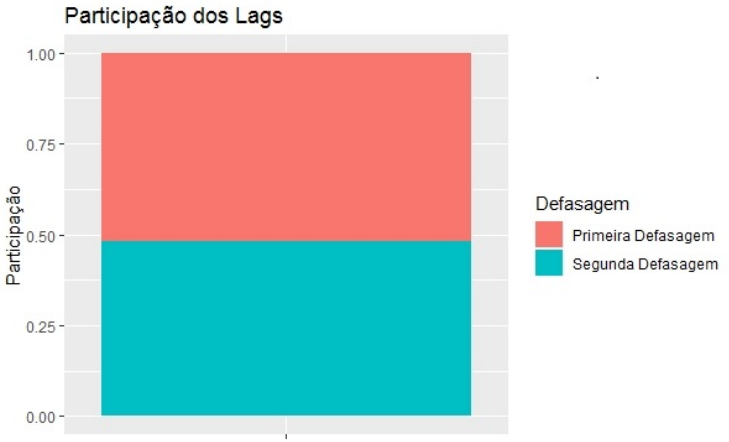
Dado o grande volume de variáveis em análise, pode ser mais interessante verificar qual característica das variáveis é mais relevante para comparar a origem dessa informação. Por exemplo, qual é a relevância de se adicionar uma defasagem adicional das variáveis explicativas para a estrutura do modelo? A Figura 9 apresenta um comparativo da relevância dos dois grupos de defasagens. Os valores foram padronizados para que a soma da importância de todas as variáveis fosse igual a um.

É possível ver que, novamente, o modelo *Boosting Trees* realiza maior discretização da importância das variáveis. Quase 75% da importância das variáveis vêm da Primeira Defasagem. Já o modelo *random forest* aloca participação muito similar para ambas as defasagens. A Figura 10 apresenta a mesma informação para o modelo que realiza previsões 60 minutos à frente.

A Figura 10 demonstra que, quando aumentamos o horizonte de previsão, a participação da variáveis se confunde. Em ambos os modelos, a importância das



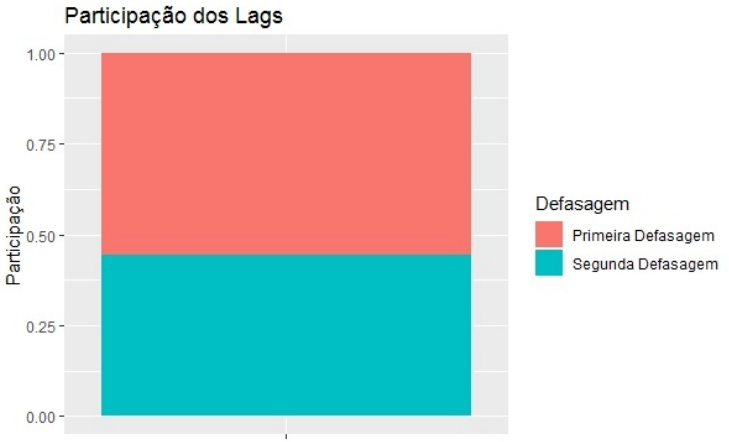
(a) *Boosting Trees*



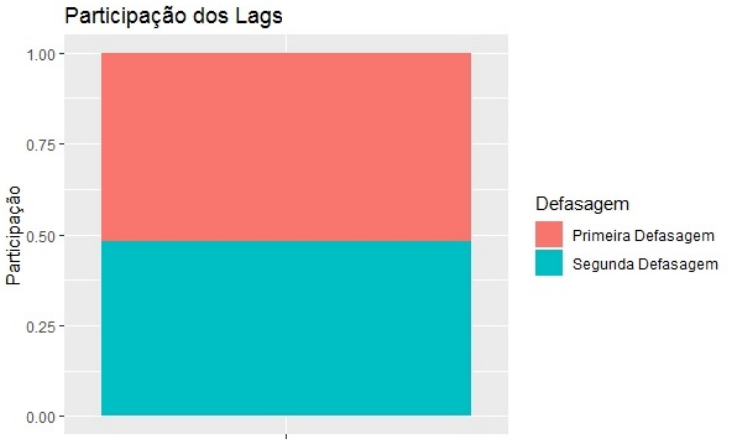
(b) *Random Forests*

Figura 9 – Importância de Variáveis - Horizonte de 1 Minuto

Variáveis mais importantes para as previsões com horizonte de 1 minuto à frente, segregadas pela ordem de defasagem. O quadro superior apresenta a relevância para o modelo *boosting trees* e o quadro inferior apresenta a relevância para o modelo *random forest*.



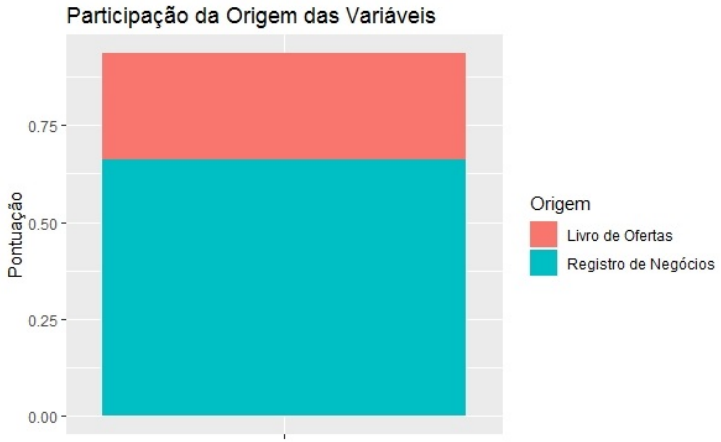
(a) *Boosting Trees*



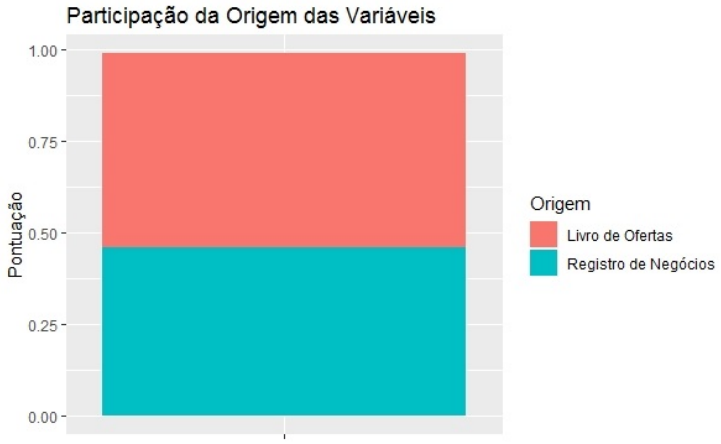
(b) *Random Forests*

Figura 10 – Importância das Variáveis - Horizonte de 60 Minutos

Variáveis mais importantes para as previsões com horizonte de 60 minutos à frente, segregadas pela sua defasagem. O quadro superior apresenta a relevância para o modelo *boosting trees* e o quadro inferior apresenta a relevância para o modelo *random forest*.



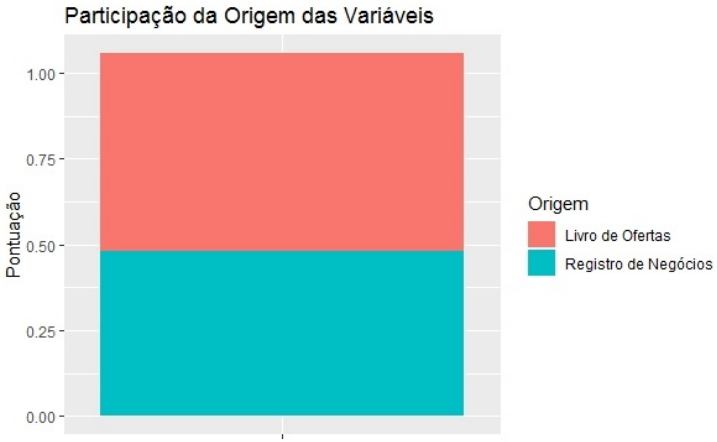
(a) *Boosting Trees*



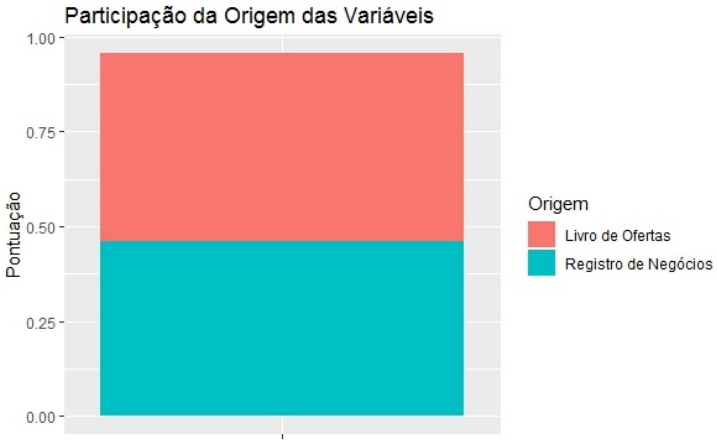
(b) *Random Forests*

Figura 11 – Importância das Variáveis - Horizonte de 1 Minuto

Variáveis mais importantes para as previsões com horizonte de 1 minuto à frente, segregadas pela sua origem. O quadro superior apresenta a relevância para o modelo *boosting trees* e o quadro inferior apresenta a relevância para o modelo *random forests*.



(a) *Boosting Trees*



(b) *Random Forests*

Figura 12 – Importância das Variáveis - Horizonte de 60 Minutos

Variáveis mais importantes para as previsões com horizonte de 60 minutos à frente, segregadas pela sua origem. O quadro superior apresenta a relevância para o modelo *boosting trees* e o quadro inferior apresenta a relevância para o modelo *random forest*.

variáveis da primeira e da segunda defasagem é similar.

Outra análise interessante seria comparar as variáveis pela origem da informação. Durante a negociação, há o apregoamento, atualização e cancelamento de ofertas que não serão, necessariamente, negociadas. O livro de ofertas, a qualquer tempo, informa os *potenciais* negócios a serem realizados, enquanto que o registro de negócios apresenta informações de negócios efetivamente concretizados, isto é, que realmente geraram volume financeiro. As informações de potenciais negócios afetam o livro de ofertas e os negócios efetivamente realizados afetam o registro de negócios. A Figura 11 apresenta a importância das variáveis conforme esse local de impacto.

Pode-se observar que o modelo *boosting trees* consegue, novamente, realizar uma maior separação de qual característica das variáveis é mais importante, apontando que as variáveis mais importantes para estimação do modelo vêm do registro de negócios, com mais de 60% de toda a importância. De forma contrária, o modelo *Random Forests* não realiza grande separação da importância das variáveis. A Figura 12 apresenta a mesma análise para os modelos que realizam previsões 60 minutos a frente.

É possível ver, novamente, que, na medida em que o horizonte de previsão aumenta, os modelos se tornam muito menos capazes de apontar a origem das variáveis

mais importantes, com ambas às fontes de informação tendo uma contribuição extremamente similar.

2.3 Conclusões

Analisamos com os dois principais modelos de conjuntos de árvores, a relação entre a dinâmica de negociação do ativo PETR4 e o sinal de seus retornos de alta frequência. Inicialmente, verificou-se se existia um horizonte de previsão ótimo, e conclui-se que, globalmente, os modelos tinham uma capacidade preditiva melhor na medida em que se aumentava o horizonte de previsão, mas a previsão de movimentos direcionais tinha melhor qualidade nas menores frequências.

Posteriormente verificou-se a importância das variáveis nesses dois extremos, e conclui-se que, nas frequências curtas, a variável que mais auxiliava a prever o sinal do retorno futuro era o retorno passado, embora não tenha havido concordância entre os modelos de quais seriam as outras variáveis relevantes. Passando para previsões feitas 60 minutos à frente, o último preço negociado no período anterior passou a ser a variável mais importante para prever os retornos futuros, mas, novamente, não houve concordância de quais seriam as outras variáveis importantes.

Ao agrupar as variáveis por ordem de defasagem,

pode-se observar que o único modelo que consegue diferenciar a relevância delas é *boosting trees*, ao prever o sinal dos retornos 1 minuto à frente. Já *random forest* não consegue encontrar grande diferença entre as duas defasagens. Nas previsões 60 minutos à frente, nenhum dos modelos consegue diferenciar a importância dos conjuntos de informação.

E, por fim, ao agrupar as variáveis pelo local onde são apresentadas, contrapondo o livro de ofertas ao registro de negócios, verificou-se que, novamente, o único modelo que aponta uma grande diferença de importância é o *boosting trees*, ao prever o sinal dos retornos 1 minuto à frente.

3 PREVISÃO DE VOLATILIDADE REALIZADA UTILIZANDO APRENDIZADO DE MÁQUINA, MÉTODOS BAYESIANOS E COMBINAÇÃO DE PREVISÕES.

Nas últimas décadas, um dos grandes sucessos da econometria financeira foi o desenvolvimento de modelos acurados para previsão da volatilidade dos retornos de ativos financeiros. O foco na volatilidade é devido, parcialmente, a baixa capacidade de se modelar e prever a média dos retornos, resultado já esperado pela hipótese dos mercados eficientes (FAMA, 1970), enquanto que a variância, uma *proxy* para volatilidade, apresenta alta previsibilidade.

A compreensão da importância em se prever a volatilidade dos retornos data, pelo menos, do trabalho de Markowitz (1952), mas, foi somente 30 anos depois, quando Engle (1982) apresentou o modelo ARCH (*Autoregressive Conditional Heteroskedaticity*), que se formou a base teórica dos atuais modelos de variância condicional. O modelo ARCH, e suas

generalizações como o GARCH (*Generalized Autoregressive Conditional Heteroskedasticity*) (BOLLERSLEV, 1986), tentam modelar a variância com base nos retornos realizados, utilizando os quadrados dos resíduos para estimá-la.

Quase vinte anos depois, em uma sequência de artigos, Andersen et al. (2000) e Andersen et al. (2001) demonstraram que, na disponibilidade de dados intradiários, poderiam ser estimadas *proxys* consistentes para a volatilidade. Essas *proxys* ficaram conhecidas como estimadores de volatilidade realizada. Desde então, dois campos ganharam força dentro da econometria financeira: a pesquisa visando encontrar a melhor *proxy*, gerando uma literatura com numerosos estimadores de volatilidade, e os trabalhos tentando modelar e prever essas *proxys*, área em que se encaixa este trabalho.

A literatura de modelos de volatilidade preocupa-se em criar modelos com duas características: a capacidade de imitar as propriedades vistas nas séries reais, também conhecidas como fatos estilizados, incluindo aglomeração, memória longa, não linearidade e mudança no tempo (TSAY, 2005); e, conjuntamente, a capacidade de fazer as melhores previsões possíveis.

Entre os modelos de volatilidade realizada, um bom competidor é o HAR (*Heterogeneous Auto-*

Regressive) (CORSI, 2009), que é uma estrutura parcimoniosa capaz de captar as mesmas propriedades do modelo ARFIMA (GRANGER; JOYEUX, 1980), incluindo memória longa e excesso de curtose. HAR também é conhecido pela sua boa capacidade preditiva, além de ter como base uma teoria econômica, a hipótese dos mercados heterogêneos (MÜLLER et al., 1993).

O presente trabalho comparará a qualidade preditiva do HAR com modelos que alteram a dinâmica da memória. Para isso, serão utilizadas duas *priori* Bayesianas altamente informativas e regressões lineares com termos de penalização. Isso permitirá investigar se, ao violar as restrições de igualdade do HAR, é possível atingir melhoria no desempenho preditivo.

Ao mesmo tempo, verificará se métodos baseados em árvores, que inserem a possibilidade de não linearidade na dinâmica da volatilidade, conseguem gerar melhorias na qualidade preditiva. Devido à recente popularidade de conjuntos de árvores em competições de aprendizado de máquina, o foco se dará em três diferentes estruturas: *model trees*, *cubist* e *bagging*.

Por fim, será verificado se duas metodologias de combinação de previsões conseguem produzir estimativas da volatilidade futura mais acuradas do que qualquer uma das previsões individuais. A primeira proposta

será o tradicional método de regressão como em Granger e Ramanathan (1984), e a segunda será uma versão modificada da ponderação de modelos dinâmica (RAFTERY; KÁRNỲ; ETTLER, 2010) proposta por Caldeira, Moura e Santos (2018).

A comparação será feita por meio de uma simulação contendo dez ativos americanos, com a série iniciando em 12/2003 e avançando até 12/2014, para a estimação das séries de volatilidade realizada foram utilizadas amostras minuto à minuto. Será mensurada a qualidade preditiva por duas diferentes medidas de erro, MSE e QLIKE (PATTON, 2011) a serem comparadas por meio de dois testes de comparação de previsão: para comparação par a par será usado o teste para habilidade preditiva condicional (CPA) (GIACOMINI; WHITE, 2006) e para comparação em grupo será utilizado o conjunto de modelos de confiança (MCS) (HANSEN; LUNDE; NASON, 2011).

3.1 Volatilidade Realizada

Merton (1980) foi o primeiro a propor *proxies* para a volatilidade com dados de alta frequência, mas foi somente quando Andersen et al. (2000) e Andersen et al. (2001) apresentaram um estimador com boas propriedades estatísticas que a utilização de estimadores de volatilidade

baseados em dados de alta frequência se disseminou (BUCCI, 2017). Suponha que ao longo do dia t, os preços logaritimicos $p(t)$ de um ativo sigam um processo de Wiener, com média e variância contidas em um processo de difusão em tempo contínuo:

$$dp(t) = \mu(t)dt + \sigma(t)dW(t). \tag{3.1}$$

A equação (3.1) descreve a trajetória de um $semimartingale$ ao longo do intervalo $[0, \tau]$ com $0 \leq s \leq \tau \leq T$, em que $\mu(t)$ é o *drift*; $\sigma(t)$ é a volatilidade instântanea do processo, estritamente positiva e quadrado integrável (i.e. $E[\int_0^t \sigma_s^2 ds] < \infty$), e $W(t)$ é um movimento *browniano*. Por fim, suponha que $\sigma(t)$ é ortogonal em relação a $W(t)$, de maneira que não existe efeito alavancagem.

Uma maneira comum de medir a variabilidade do processo de difusão em teoria de integração estocástica é a variação quadrática, definida como:

$$QV_t = \int_{t-h}^{t} \sigma^2(s)ds, \tag{3.2}$$

Como a variação quadrática é gerada somente pelos choques de um *martingale* local, ela corresponde à variância integrada:

$$IV_t = \int_{t-h}^{t} \sigma^2(s)ds = QV_t. \tag{3.3}$$

O estimador de volatilidade realizada amostra p_t ao longo da sessão de negociação em intervalos regulares como $1,\ 5,\ 30, \ldots$ minutos. Suponha que os preços no dia t foram amostrados em frequências regulares com $m+1$ pontos $0, 1, \ldots, m$ sendo $p_{i,t}$ a *i*-ésima observação do preço logarítimico do dia t. O estimador de volatilidade realizada pode ser definido como:

$$RV_t^m = \sum_{i=1}^{m} (p_{i,t} - p_{i-1,t})^2 = \sum_{i=1}^{m} r_{i,t}^2, \qquad (3.4)$$

que é um estimador consistente para a variância integrada, i.e. $RV_t^m \xrightarrow{p} IV_t$ conforme $m \to \infty$.

Um grande esforço tem sido feito na tentativa de ampliar a modelagem original incorporando saltos, ruídos de micro-estrutura e amostragem *tick by tick*. (BUCCI, 2017). Apesar de toda a discussão sobre qual seria o melhor estimador, este trabalho usará o tradicional estimador de volatilidade realizada, com retornos amostrados em intervalos de 5 minutos. De acordo com Liu, Patton e Sheppard (2015), na maioria das medidas de qualidade há pouca melhoria que pode ser atingida em todos os outros estimadores analisados.

3.2 Modelos de Previsão

3.2.1 Abordagens Tradicionais

3.2.1.1 Heterogeneous Auto-Regressive (HAR)

O estimador de volatilidade realizada, que mede a volatilidade dentro de um dia, nos permite registrar a sua evolução conforme o tempo passa. Uma vez que existe esse registro, um segundo interesse natural é modelar a sua evolução. Para realizar a modelagem, trata-se as estimativas de volatilidade realizada como dados observados, permitindo que a série seja modelada diretamente.

Modelos de volatilidade tentam mimetizar os fatos estilizados das séries empíricas, e uma propriedade bem conhecida das séries é a sua memória longa, exemplificada na Figura 13. Dentre os modelos derivados da filosofia GARCH, os modelos derivados da estrutura FIGARCH (BAILLIE; BOLLERSLEV; MIKKELSEN, 1996) estão entre aqueles que captam de maneira mais eficiente o comportamento de memória longa. No entanto, este trabalho optará por modelar a série diretamente, desconsiderando possíveis erros de medição.

Granger (1980) foi o primeiro a demonstrar que um processo de memória longa pode ser mimetizado por *infinitos* processos de memória curta sobrepondo-

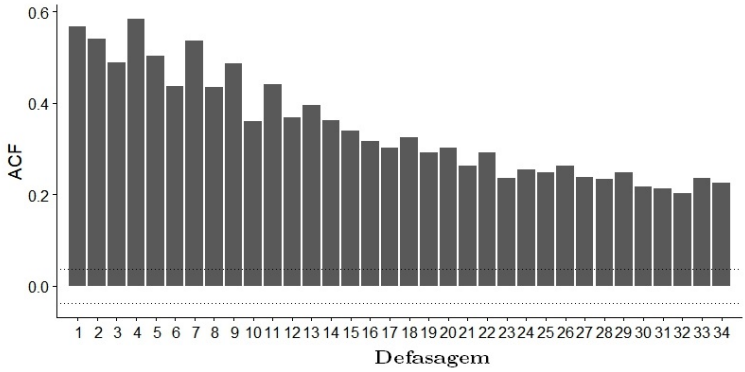


Figura 13 – Função de Autocorrelação - Volatilidade Realizada

Estimativas da função de autocorrelação da volatilidade realizada. A amostra é composta pelos registros da AAPL com a série iniciando em 12/2003 e indo até 12/2014. As linhas pontilhadas apresentam os níveis de significância.

se uns aos outros. Posteriormente, LeBaron et al. (2001) demonstraram que esse comportamento pode ser aproximado por *finitos* processos de memória curta. Em outras palavras, talvez não haja um processo de memória longa, mas alguns diferentes processos de memória curta atuando em escalas de tempo diferentes. Com base nisso, Corsi (2009) propôs o *Heterogeneous Auto-Regressive* (HAR) como uma sobreposição de três processos de memória curta atuando em diferentes frequências.

De acordo com a hipótese dos mercados heterogêneos, os participantes do mercado têm diferentes necessidades, tipos de estratégia e, entre

outras características, diferentes horizontes de previsão (DACOROGNA et al., 1998). Isso inspirou Corsi (2009) a propor que a dinâmica da volatilidade fosse conduzida por três processos autorregressivos atuando em diferentes frequências. Considere a média logarítimica da volatilidade definida como:

$$logRV_t^{(n)} = \frac{1}{n}\sum_{j=1}^{n} logRV_{t-j+1}, \tag{3.5}$$

aonde n representa três diferentes escalas de tempo $n = \{d, s, m\} = \{1, 5, 22\}$, e.g. diária, semanal e mensal.

Poderíamos, também, modelar diretamente a volatilidade realizada RV_t, mas Gonçalves e Meddahi (2011) argumentam que é preferível modelar o seu logaritmo. Assim, para economia de notação, será utilizado $\sigma_t^{(n)}$ no lugar de $logRV_t^{(n)}$.

Supõe-se que os modelos parciais para a volatilidade não observada em cada escala de tempo d, s, m, sejam função da última observação na mesma escala e da esperança da próxima realização na frequência superior. Por meio de substituições recursivas, isso pode ser representado por meio de uma única equação:

$$\sigma_{t+1d}^{(d)} = c + \beta^{(d)}\sigma_t^{(d)} + \beta^{(s)}\sigma_t^{(s)} + \beta^{(m)}\sigma_t^{(m)} + \omega_{t+1d}, \tag{3.6}$$

em que ω_{t+1d} é uma variável aleatória normalmente distribuída.

Desde que o modelo foi proposto, múltiplas extensões foram apresentadas na literatura, mas poucos trabalhos tentaram testar e analisar suas hipóteses centrais. Apresentamos que o HAR é uma regressão linear aonde a volatilidade de um dia é regredida na volatilidade do dia anterior, na média da volatilidade dos últimos cinco dias e na média da volatilidade dos últimos 22 dias. Essa técnica de regressão com base em médias pode ser reescrita como um modelo autorregressivo de ordem 22, $AR(22)$:

$$\sigma_{t+1d}^{(d)} = c + \sum_{i=1}^{22} \beta^{(i)} \sigma_{t-(i-1)d}^{(d)} + \omega_{t+1}, \tag{3.7}$$

no qual valem as seguintes restrições de igualdade:

$$\beta^{(i)} = \begin{cases} \beta^{(d)} + \frac{1}{5}\beta^{(w)} + \frac{1}{22}\beta^{(m)}, & \text{para } i = 1; \\ \frac{1}{5}\beta^{(w)} + \frac{1}{22}\beta^{(m)}, & \text{para } i = 2, ..., 5; \\ \frac{1}{22}\beta^{(m)} & \text{para } i = 6, ..., 22. \end{cases} \tag{3.8}$$

Essa especificação atesta que a contribuição para a volatilidade de amanhã, advinda da segunda à quinta defasagem, é a mesma, assim como a contribuição da sexta à vigésima-segunda defasagem.

Pode-se arguir que essa hipótese não é razoável, já que é possível verificar na Figura 13 que não parece existir uma especial queda de memória na quinta ou na vigésima-segunda defasagem, e que a queda se assemelha

muito mais a um decaimento exponencial do que a um decaimento de escada. É possível verificar, também, que a trucagem na defasagem 22 não é compatível com o longo decaimento da memória ou, ainda, com o comportamento verificado no mundo real, no qual parte dos grandes fundos de investimento têm horizontes de investimento além de 22 dias.

3.2.2 Métodos inspirados em aprendizado de máquina e aprendizado estatístico

Este trabalho buscará verificar se algumas técnicas de aprendizado, oriundas de técnicas computacionais e estatísticas, são capazes de gerar modelos preditivos mais acurados que o *HAR*. Em exercícios de previsão, é comum definir-se um modelo como o limiar mínimo a ser atingido pela nova estrutura proposta e, às vezes, o modelo base é escolhido pela sua simplicidade, nesse caso, o modelo *HAR* foi escolhido pela sua aceitação. A estrutura simples faz dele um bom modelo de comparação, pois as novas estruturas passarão a incorporar estruturas mais ricas e complexas.

3.2.2.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Em aprendizado estatístico, existem dois tradicionais desafios: a acurácia preditiva e a identificabilidade. Modelos de regressão penalizados conseguem atender a ambos os objetivos. Apesar de LASSO (TIBSHIRANI, 1996) ter ganho a sua fama por resolver problemas de identificação, o modelo estimado também pode ser usado para melhorar a estrutura preditiva.

Viu-se na Seção 3.2.1.1 que o autocorrelograma dos ativos financeiros põe em dúvida as restrições de igualdade do HAR apresentadas na equação (3.8). Esse questionamento foi feito, inicialmente, por Craioveanu, Hillebrand et al. (2010), que verificaram que alterar o número de defasagens ou adicionar ou excluir restrições não era capaz de gerar modelos de melhor qualidade preditiva.

Mais recentemente, em uma sequência de artigos, Audrino, Camponovo e Roth (2015) e Audrino, Huang e Okhrin (2016), usando o LASSO Adaptativo de Zou (2006) com os novos resultados de inferência de Audrino e Camponovo (2013), tentaram recuperar a estrutura de restrições por meio de informações provenientes dos dados. O LASSO adaptativo pode ser visto como uma extensão do LASSO de Tibshirani (1996). Neste trabalho, será testada

se a especificação tradicional é capaz de atingir melhor desempenho preditivo, visto que Audrino e Knaus (2016) já atestaram que o LASSO adaptativo não é.

Tomando como base as equações (3.7) e (3.8), o LASSO pode ser visto como uma regressão linear restrita, em que os parâmetros são encontrados ao se resolver o problema:

$$\beta^{LASSO} = \underset{\beta}{argmin} \sum_{n=1}^{N} (\sigma_{n,t+1d}^{(d)} - c - \sum_{i=1}^{22} \beta^{(i)} \sigma_{n,t-(i-1)d}^{(d)})^2, \tag{3.9}$$

$$sujeito \quad a \quad \sum_{i=1}^{p} |\beta_i| \leq t, \tag{3.10}$$

que pode ser reescrito como um lagrangeano:

$$\begin{aligned} \beta^{LASSO} = \underset{\beta}{argmin} \Bigg[& \frac{1}{2} \sum_{n=1}^{N} (\sigma_{n,t+1d}^{(d)} - c - \\ & \sum_{i=1}^{22} \beta^{(i)} \sigma_{n,t-(i-1)d}^{(d)})^2 + \\ & \lambda \sum_{i=1}^{22} |\beta_i| \Bigg]. \end{aligned} \tag{3.11}$$

Se t for maior ou igual do que a soma das estimativas de mínimos quadrados ordinários (MQO), ou seja se $t_0 \geq \sum_1^p |\hat{\beta}_i^{MQO}|$, as estimativas de LASSO serão iguais as de MQO, $\beta_i^{LASSO} = \beta_i^{mqo}$. Na prática, t é encontrado via

validação cruzada, conceito apresentado na Seção 2.1.5.5. Nem sempre o melhor parâmetro escolhido por meio de validação cruzada é aquele cujo modelo atingiu o menor valor para a função perda. Tibshirani (1996) observa que o melhor parâmetro se encontra um pouco antes do ponto de mínimo. O autor defende que se estime o desvio padrão do erro gerado pelo modelo que teve a melhor capacidade, some-se o valor do erro ao desvio-padrão, e encontre-se o modelo mais parcimonioso que atingiu performance igual a essa soma. Essa regra para escolha de parâmetros é conhecida como a regra de um desvio padrão, se escolhe o modelo mais parcimonioso que atingiu esse valor: o valor do menor erro somado ao desvio padrão do menor erro.

Uma das propriedades mais famosas da estimação via mínimos quadrados ordinários é a sua consistência. Quando se restringe que os parâmetros estimados respeitem o espaço de otimização do LASSO, os parâmetros se tornam viesados, ou seja, são diferentes dos parâmetros consistentes de MQO. Apesar de o LASSO gerar estimativas viesadas, a performance preditiva é muitas vezes melhorada devido a menor variância dos parâmetros, a possibilidade de forçar alguns parâmetros a zero diminui a correlação entre eles e consequentemente a sua estimação se torna mais precisa.

3.2.2.2 *Adaptative LASSO (ADALASSO)*

Quando Zou (2006) propôs o *Adaptative* LASSO (ADALASSO) como uma metodologia de seleção de variáveis, havia a preocupação de o LASSO apenas respeitar as propriedades de oráculo, propriedade definida como a capacidade de um modelo apresentar desempenho tão bom quanto o modelo original, sobre o guarda-chuva de hipóteses demasiadamente restritivas.

A solução foi encontrada com o ADALASSO, uma versão ponderada do LASSO original que redefine o problema de minimização para:

$$\beta^{ADALASSO} = \underset{\beta}{argmin}\Bigg[\frac{1}{2}\sum_{n=1}^{N}(\sigma_{n,t+1d}^{(d)} - c - \sum_{i=1}^{22}\beta^{(i)}\sigma_{n,t-(i-1)d}^{(d)})^2 + \lambda\sum_{i=1}^{22}\lambda_i|\beta_i|\Bigg\}. \quad (3.12)$$

em que cada parâmetro tem seu próprio peso λ_i. A estimação é feita em duas etapas; primeiro, encontra-se estimativas consistentes para β, em geral as estimativas de MQO. Posteriormente, escolhe-se λ para calcular os pesos $\lambda_i = 1/|\hat{\beta}^{mqo}|^{\lambda}$. A estimativa para λ é encontrada por validação cruzada. Bühlmann e Geer (2011) trazem detalhes computacionais sobre como implementar esse algoritmo de estimação.

3.2.2.3 Árvores

Audrino e Corsi (2010) foram os primeiros a propor o uso de métodos baseados em árvores para prever volatilidade realizada. Neste trabalho, será utilizada uma versão modificada da estrutura regular de árvores conhecida popularmente por árvore de modelos (WANG; WITTEN, 1997), uma generalização do modelo de Quinlan et al. (1992).

Por simplicidade, suponha que se queira prever a volatilidade do próximo dia $\sigma_{t+1d}^{(d)}$ com base nas médias das últimas 22 medições de volatilidade. Uma árvore pode ser vista como um conjunto de proposições *se-então* que culminam em uma regra de previsão. Uma *árvore de modelos* pode ser expressa por uma estrutura similar, mas, agora, a regra de previsão tem a estrutura de um modelo linear. O algoritmo a seguir expressa um exemplo de árvore de modelo:

se $\sigma_t^{(d)} < t_1$ então:
$\hat{\sigma}_{t+1d}^{(d)} = c_1 + \sum_{i=1}^{22} \beta_1^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$
caso contrário
se $\sigma_{t-1}^{(d)} > t_2$
$\hat{\sigma}_{t+1d}^{(d)} = c_2 + \sum_{i=1}^{22} \beta_2^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$
caso contrário
$\hat{\sigma}_{t+1d}^{(d)} = c_3 + \sum_{i=1}^{22} \beta_3^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$

Essa árvore pode ser compreendida seguindo o

seguinte raciocínio: se a volatilidade de hoje $\sigma_t^{(d)}$ for menor do que um valor limite t_1, então se usará o primeiro modelo linear $c_1 + \sum_{i=1}^{22} \beta_1^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$ para prever a volatilidade de amanhã $\hat{\sigma}_{t+1d}^{(d)}$, mas, se a volatilidade de hoje for maior do que t_1, é necessário verificar se $\sigma_{t-1}^{(d)}$ é maior do que um valor t_2. Caso isso seja verdade, utiliza-se o segundo modelo para prever a volatilidade de amanhã; caso contrário, utiliza-se um terceiro.

Esse simples modelo pode ser expandido ampliando-se o número de proposições condicionais do tipo *se então* gerando um grande número de *splits*, com respectivos valores-limite $t_1, t_2, \ldots$ Assim, quando uma nova observação surgir, a previsão será feita simplesmente seguindo a sequência de testes até que se chegue no que se chama uma folha ou nó terminal. Para cada respectivo nó terminal, existe um diferente modelo linear. No exemplo anterior, os três modelos lineares.

Adicionando um pouco de rigor, estima-se uma partição $\{\mathscr{B}_b\}_{1,\ldots,B}$ com B células no espaço de preditores relevantes, que, para simplificar, simbolizaremos por $\mathbf{\Sigma}$. Esse espaço de preditores contém a amostra de regressores $\Sigma_{n\times i}$ correspondente à coleção de todas as $i = 1, \ldots, 22$ médias de volatilidade realizadas defasadas $\sigma_{n,t-(i-1)d}^{(d)}$, em que $n = 1, \ldots, N$ é o índice da amostra. Cada célula de $\mathscr{B}_b$ conterá uma diferente subamostra Σ_b e um

diferente modelo linear $\mathscr{M}_b(\Sigma_b, \beta_b)$ com os respectivos parâmetros β_b. Denota-se esse modelo segmentado como uma árvore $T_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau_{\mathscr{B}})$, em que $\{\beta_b\}_{b=1,\ldots,B}$ é a coleção completa de parâmetros dos modelos lineares e $\tau_{\mathscr{B}} = \{\tau_1, \tau_2, \ldots, \tau_B\}$ são os limites que definem as B partições das células.

Para estimar os limites dos *splits*, inicia-se encontrando o desvio-padrão dos vetores resposta $sd(\sigma^{(d)}_{t+1d})$. Para cada defasagem i, testa-se diferentes valores limites t^*, que sempre definem dois subconjuntos candidatos $\mathbf{\Sigma}^R \cup \mathbf{\Sigma}^L = \mathbf{\Sigma}$ em que uma observação $\Sigma^n \in \mathbf{\Sigma}^R$ se $\sigma^{(d)}_{n,t-(i-1)d} < t^*$ e $\Sigma^n \in \mathbf{\Sigma}^L$ se $\sigma^{(d)}_{n,t-(i-1)d} \geq t^*$.

Para todos os limites e respectivos subconjuntos gerados, a *redução do desvio-padrão* (SDR) é calculada como:

$$SDR = sd(\sigma^{(d)}_{t+1d}) - \sum_{b=R,L} \frac{\#\sigma^{(d)}_{b,t+1d}}{\#\sigma^{(d)}_{t+1d}} * sd(\sigma^{(d)}_{b,t+1d}) \quad (3.13)$$

em que $\#$ denota o número de elementos em cada célula da partição. A defasagem i^* e limite t^* que atingirem o maior SDR são escolhidas para o *split* τ_b, e os respectivos subconjuntos candidatos começam a definir a partição $\mathscr{B}_1 = \mathbf{\Sigma}^{\boldsymbol{R}}$ e $\mathscr{B}_2 = \mathbf{\Sigma}^{\boldsymbol{L}}$. Um modelo linear é estimado em ambos os subconjuntos selecionados $\mathscr{B}_1, \mathscr{B}_2$ e são usados para prever a futura volatilidade realizada $\hat{\sigma}^{(d)}_{\Sigma_1,t+1d} = \mathscr{M}_1(\Sigma_1, \beta_1) = c_1 + \beta_1 \sigma^{(d)}_{\Sigma_1,t-(i^*-1)d}$.

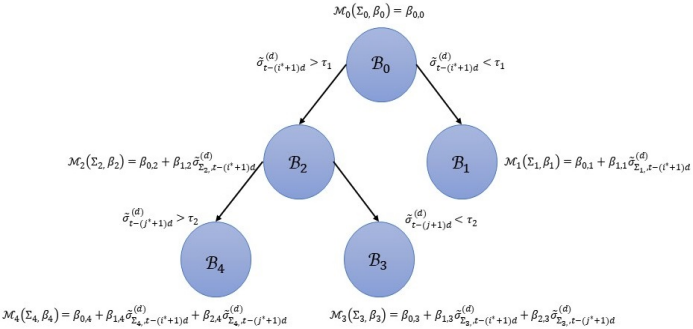


Figura 14 – Exemplo de Árvore de Modelos
Os círculos azuis representam as partições criadas pelo algoritmo; as inequações nos lados da flechas representam os limites das partições; e as equações lineares representam os modelos utilizados para a previsão na respectiva partição.

Realiza-se esse processo, iterativamente, em cada subconjunto gerado, mas, após o primeiro *split*, o erro padrão a ser usado no SDR se torna o erro-padrão do modelo linear. Em cada sucessivo *split*, o modelo linear é ampliado ao se adicionar a defasagem j^* ao modelo de previsão, $\hat{\sigma}^{(d)}_{\Sigma_b, t+1d} = \mathscr{M}_b(\Sigma_b, \beta_b) = c_b + \beta_b^{\mathscr{S}}\Sigma_b^{\mathscr{S}}$ em que $\mathscr{S}$ indica a defasagem usada nos *splits*. As Figuras 14 e 15 exemplificam os dois primeiros *splits* com o correspondente efeito no espaço de regressores. Esse processo de particionamento continua até que haja 4 elementos, ou menos, em uma folha, ou até que as variáveis que chegaram àqueles nós variem apenas um pouco, em geral, menos do que 5% da variação original.

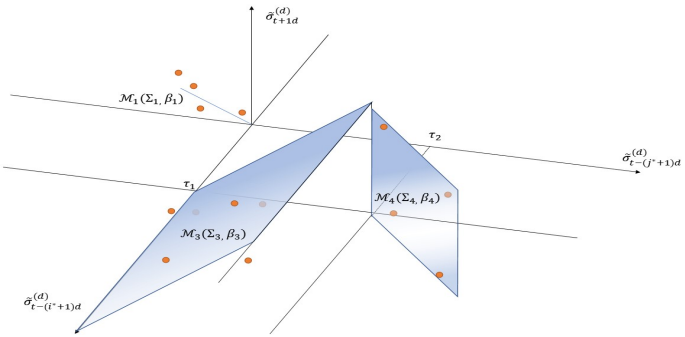


Figura 15 – Exemplo de Árvore de Modelos
Os espelhos azuis representam os modelos lineares multivariados estimados. Os pontos em laranja ilustram a amostra. No eixo vertical, tem-se a volatilidade realizada um período à frente e nos dois eixos horizontais têm-se as defasagens selecionadas.

Após a árvore ser estimada $T_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$, começa um procedimento de seleção de variáveis. Seja v o número de parâmetros em β_b, utilizados para realizar a previsão na célula b, e seja $e_b = |\hat{\sigma}^{(d)}_{\Sigma_b, t+1d} - \sigma^{(d)}_{\Sigma_b, t+1d}|$ um vetor contendo o módulo dos erros de previsão realizados na célula b de dimensão $n_b^* \times 1$, em que n_b^* é o número de pontos em b. Descarta-se os parâmetros um a um de β_b gerando o subconjunto β_b^* de dimensão $(v-1) \times 1$. Assim, calcula-se o erro desse modelo restrito e_b^* e a correspondente *Taxa de Erro Real (AER)*:

$$AER = \frac{n_b^* + p}{n_b^* - p} \sum_{n=1}^{n_b^*} (e^*_{n,b}) \tag{3.14}$$

se $AER < \sum_{n=1}^{n_b^*} (e_{n,b})$ substitui-se $\mathscr{M}_b(\Sigma_b, \beta_b)$ por

$\mathscr{M}_b(\Sigma_b, \beta_b^*)$. Esse procedimento é realizado em todas as células e em todos os nós internos do modelo de maneira iterativa.

Em seguida é iniciado um procedimento de suavização. Durante o processo de estimação da árvore, muitos modelos, que agora serão utilizados com fins de suavização, foram estimados. Seja $\mathscr{M}_p(\Sigma_p, \beta_p)$ o modelo de um nó superior, ou nó pai, e seja $\mathscr{M}_c(\Sigma_c, \beta_c)$ o modelo de um nó inferior, ou nó filho. Note que o nó filho é uma célula da partição do nó pai. Todo nó pai p tem dois nós filhos $c1, c2$ que formam uma partição do nó pai $\mathbf{\Sigma}_p = \mathbf{\Sigma}_{c1} \cup \mathbf{\Sigma}_{c2}$.

Sejam $\hat{\sigma}_{p,t+1d}^{(d)}$ e $\hat{\sigma}_{c1,t+1d}^{(d)}$ respectivamente, as previsões do nó pai e do nó filho para uma observação, e n_p^*, n_{c1}^* o número de observações que chegou nesses nós. A previsão suavizada $\hat{\sigma}_{s,t+1d}^{(d)}$ pode ser definida como:

$$\hat{\sigma}_{s,t+1d}^{(d)} = \frac{c\hat{\sigma}_{p,t+1d}^{(d)} + n_{c1}^* \hat{\sigma}_{c1,t+1d}^{(d)}}{n_{c1}^* + c} \tag{3.15}$$

em que c é um parâmetro de regularização que controla o grau de suavização. Conforme c cresce, as previsões se tornam mais suaves e as partições em $T_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau_{\mathscr{B}})$ se tornam imprecisas. Esse procedimento de suavização interage ao longo de todos os *splits* superiores avançando até a raiz da árvore.

Por fim, ocorre um procedimento de poda. Seja $e_f = |\hat{\sigma}^{(d)}_{\Sigma_f,t+1d} - \sigma^{(d)}_{\Sigma_f,t+1d}|$ o erro do nó pai, e e_{c1}, e_{c2} os correspondentes erros dos nós filhos, caso $e_f < e_{c1} + e_{c2}$, o nó será podado e ambos os nós filhos $\mathscr{B}_{c1}$, $\mathscr{B}_{c2}$ são excluídos da árvore, e o nó pai $\mathscr{B}_f$ torna-se uma folha. Esse processo interage ao longo de toda árvore.

3.2.2.4 Bagging

Modelos baseados em árvores são populares pela sua compreensibilidade, mas, em exercícios de previsão, abre-se mão de compreensibilidade por poder preditivo. Breiman (1984) foi o primeiro a propor que a agregação de múltiplas árvores poderia gerar melhor performance preditiva do que qualquer uma delas sozinha.

Em estatística clássica, encara-se, recorrentemente, o problema de haver parâmetros viesados. Em aprendizado estatístico, isso não é diferente. As árvores de modelos $T_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$ são uma função da amostra Σ e, como a amostra afeta as estimativas de $\{\beta_b\}_{\mathscr{B}}$ e τ, a incerteza dentro da amostra se traduz em incerteza para os parâmetros, levando a baixa precisão das previsões. *Bagging* surge como uma solução para esse problema, sendo uma metodologia que busca minimizar a incerteza paramétrica.

Em cada rodada $bag = 1, \ldots, O$, amostra-se

aleatoriamente uma fatia Σ_{bag} da amostra Σ_N com tamanho $o < N$ e estima-se um novo modelo de árvores com os respectivos parâmetros $T_{\mathscr{B}}^{bag}(\Sigma_{bag}, \{\beta_b\}_{\mathscr{B}}, \tau)$. Esse processo gera O diferentes árvores que são, então, agregadas para realizar as previsões $BT(\Sigma, O)$:

$$BT(\Sigma, O) = \frac{1}{O} \sum_{bag=1}^{O} T_{\mathscr{B}}^{bag}(\Sigma_{bag}, \{\beta_b\}_{\mathscr{B}}, \tau) \qquad (3.16)$$

Pode-se demonstrar que essa agregação das previsões faz com que a previsão do modelo convirja para a verdadeira previsão de *bagging*, conforme $O \to \infty$. Em outras palavras, supondo que as observações $\Sigma^* = \{\Sigma_1^*, \Sigma_2^*, \ldots, \Sigma_N^*\}$ venham da mesma distribuição desconhecida $\Sigma_n^* \sim \hat{\mathscr{P}}$, a estimativa de *bagging* (3.16) passa a ser vista como uma estimativa de Monte Carlo, convergindo para a verdadeira estimativa $E_{\hat{\mathscr{P}}}$, conforme $O \to \infty$, divergindo, apenas, caso a função seja não linear ou uma função adaptativa dos dados (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

As previsões fora da amostra são feitas com (3.16) e o número de árvores e seus tamanhos são encontrados por validação cruzada, conceito apresentado na Seção 2.1.5.5. *Bagging* é um dos métodos de aprendizado estatístico que se torna, mais dificilmente, superajustado aos dados. A convergência das previsões ocorre conforme o número de

árvores aumenta. Usualmente, o número de árvores não é encontrado devido a problemas de excesso de ajuste, mas, devido à quantidade limitada de informação disponível dentro do conjunto de dados, torna-se inútil, em algum momento, estimar novas árvores. *Bagging* atua reduzindo a variância das previsões, especialmente em modelos de alta variância, como árvores.

3.2.2.5 Métodos Baseados em Regras

Uma árvore pode ser vista como um conjunto de testes sequenciais, e cada caminho ao longo da árvore gera algo chamado de regra. Pode-se reescrever o modelo exemplificado na Seção 3.2.2.3 por meio de regras como:

se $\sigma_t^{(d)} < t_1$ então $\hat{\sigma}_{t+1d}^{(d)} = c_1 + \sum_{i=1}^{22} \beta_1^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$,
se $\sigma_t^{(d)} > t_1$ e $\sigma_{t-1}^{(d)} > t_2$ então $\hat{\sigma}_{t+1d}^{(d)} = c_2 + \sum_{i=1}^{22} \beta_2^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$,
se $\sigma_t^{(d)} > t_1$ e $\sigma_{t-1}^{(d)} < t_2$ então $\hat{\sigma}_{t+1d}^{(d)} = c_3 + \sum_{i=1}^{22} \beta_3^{(i)} \sigma_{n,t-(i-1)d}^{(d)}$.

Em que cada linha define uma regra diferente. Regras são populares pela sua simplicidade, bem como pela aparente independência de cada modelo, permitindo a adição de uma nova regra sem mudar qualquer uma das regras prévias. Ao invés de utilizar o conceito de regras para criar novas regras, pode-se usar esse conceito para podar o modelo. A

remoção de um caminho inteiro é possível e não implica a necessidade de alterar o restante da estrutura da árvore (QUINLAN, 1993). Para ver mais a respeito do conceito de regras Witten et al. (2016) é uma referência atual.

O foco será em um modelo específico, baseado em regras, chamado *cubist*, que corresponde à união de vários artigos antigos, que inspiraram as árvores de modelos, e que foi lançado ao público somente em 2011. Até então, o algoritmo estava disponível apenas em *softwares* comerciais. A metodologia por trás do algoritmo de crescimento das regras é similar à metodologia das árvores de modelos, e as maiores diferenças são: o meio pelo qual é realizada a combinação de modelos, a adição de um procedimento similar a *boosting* e um procedimento adicional de suavização para previsões fora da amostra (KUHN; JOHNSON, 2013).

No processo de construção das B células da partição $\{\mathscr{B}_b\}_{1,\ldots,B}$ foram estimados modelos lineares em cada *split* $\mathscr{M}_b(\Sigma_b, \beta_b)$. Após estimar a estrutura completa é possível transformá-la no formato de regras.

Em cada caminho da árvore existe um conjunto de modelos e, se um caminho completo é definido por $d-1$ *splits*, um caminho da raiz até a folha contém d modelos, incluindo o modelo na folha. Quando o *cubist* muda a estrutura de árvores para uma estrutura de regras,

o algoritmo combina todos os modelos que foram definidos nos nós interiores. Sejam $\mathscr{M}_f(\Sigma_f, \beta_f)$ e $\mathscr{M}_c(\Sigma_c, \beta_f)$ os modelos dos nós pai e filho e $e_f = \hat{\sigma}^{(d)}_{\Sigma_c, t+1d} - \sigma^{(d)}_{\Sigma_c, t+1d}$ e $e_c = \hat{\sigma}^{(d)}_{\Sigma_c, t+1d} - \sigma^{(d)}_{\Sigma_c, t+1d}$ os erros cometidos pelo nó pai e filho ao prever as observações que foram alocadas no nó filho Σ_c. O modelo gerado pela combinação desses dois modelos, ao passar de uma árvore para uma regra, é definido iterativamente conforme a seguinte regra de combinação pai-filho:

$$\mathscr{M}_r(\Sigma_f, \beta_r) = a\mathscr{M}_c(\Sigma_c, \beta_c) + (1-a)\mathscr{M}_f(\Sigma_f, \beta_f), \quad (3.17)$$

$$a = \frac{Var(e_f) - Cov(e_c, e_f)}{Var(e_f - e_c)}. \quad (3.18)$$

em que $Var(.)$ e $Cov(.,.)$ denotam a variância e a covariância dos erros, e o subíndice r denota o modelo definido ao transformar os nós pai e filho em uma regra. Esse processo é feito iterativamente, iniciando da folha até à raiz. Se os modelos tiverem os mesmos erros, a covariância e o denominador aumentam, colocando mais peso no nó pai, que foi estimado com um maior conjunto de observações.

Após transformar cada caminho na árvore $T_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$ em um conjunto de regras $R_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$ com os respectivos parâmetros, inicia-se um procedimento de poda. Cada caminho na

árvore define uma respectiva regra individual $r_b(\Sigma_b, \beta_b, \tau_b)$. Usando (3.14), cada condição dentro de τ_b é testada e, se a exclusão do específico elemento dentro da regra melhorar a taxa de acerto do modelo, ela será definitivamente excluída. Esse procedimento se inicia pelas condições que foram criadas na definição das folhas até que se chegue nas condições definidas na raiz, enfatizando a importância das condições feitas com um maior conjunto de dados.

Posteriormente, é iniciado um procedimento de conjuntos que se parece muito com o *boosting*. Suponha que se tenha feito o primeiro conjunto completo de regras $R^{(1)}_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$, e se faça previsões com elas $\hat{\sigma}^{(d)(1)}_{\Sigma,t+1d}$ para todas as observações no conjunto de dados Σ. Sejam $e^{(1)}_R = \hat{\sigma}^{(d)(1)}_{\Sigma,t+1d} - \sigma^{(d)}_{\Sigma,t+1d}$ os erros deste modelo, o algoritmo do *cubist* ajusta o vetor de erros para que um novo conjunto de regras possa ser estimado $\sigma^{(d)(2)}_{\Sigma,t+1d} = \sigma^{(d)}_{\Sigma,t+1d} - e^{(1)}_R$, e posteriormente utilizado para estimar um novo conjunto de regras $R^{(2)}_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$ com o mesmo conjunto de variáveis Σ. Esse processo de estimação, avaliação e ajuste é feito interativamente $g = 1, 2, \ldots, G$ vezes com o vetor-resposta se ajustando:

$$\sigma^{(d)(g+1)}_{\Sigma,t+1d} = \sigma^{(d)}_{\Sigma,t+1d} - e^{(g)}_R, \tag{3.19}$$

$$e^{(g)}_R = \hat{\sigma}^{(d)(g)}_{\Sigma,t+1d} - \sigma^{(d)}_{\Sigma,t+1d}, \tag{3.20}$$

onde $\hat{\sigma}^{(d)(g)}_{\Sigma,t+1d}$ são as previsões feitas pelo g-ésimo modelo de regras $R^{(g)}_{\mathscr{B}}(\Sigma, \{\beta_b\}_{\mathscr{B}}, \tau)$. Esse processo é repetido iterativamente, na esperança de que as observações que foram sub/sobre previstas sejam corrigidas no futuro modelo com maiores/menores previsões. Uma vez que o conjunto completo de regras foi criado, uma nova previsão fora da amostra pode ser realizada $\hat{\sigma}^{(d)*}_{t+1d}$ por meio de uma nova amostra de regressores $\hat{\Sigma}^*$ com a ponderação de todos os modelos:

$$\hat{\sigma}^* = \frac{1}{G}\sum_{g=1}^{G} R^{(g)}_{\mathscr{B}}(\Sigma_*, \{\beta_b\}_{\mathscr{B}}, \tau). \tag{3.21}$$

O *cubist* também tem um procedimento adicional para suavizar as previsões fora da amostra. Suponha que um conjunto inédito de observações $\Sigma^* = \{\sigma^{(d)*}_t, \sigma^{(d)*}_{t-1d}, \sigma^{(d)*}_{t-2d}, \ldots, \sigma^{(d)*}_{t-22d}\}$ sejam apresentadas, e seja K uma constante previamente definida; para todas as observações fora da amostra Σ, faça $D = \{D_1, D_2, \ldots, D_N\}$ um vetor contendo a distância de Manhattan do novo conjunto de pontos e da amostra utilizada durante o treinamento:

$$D_n = \sum_{i=1}^{22} |\sigma^{(d)*}_{n,t-(i-1)d} - \sigma^{(d)}_{n,t-(i-1)d}|. \tag{3.22}$$

As K mais próximas previsões fora da amostra $\hat{\sigma}^{(d)}_{k,t+1d}$, e os valores observados $\sigma^{(d)}_{k,t+1d}$ com as respectivas distâncias

D^k no sentido de (3.22), serão utilizadas para suavizar as previsões fora da amostra $\hat{\sigma}^{(d)*}_{t+1d}$, como em:

$$\hat{\sigma}^{(d)s}_{t+1d} = \sum_{k=1}^{K} w_k[\sigma^{(d)}_{k,t+1d} + (\hat{\sigma}^{(d)*}_{t+1d} - \hat{\sigma}^{(d)}_{k,t+1d})] \qquad (3.23)$$

$$w_c = \frac{1}{D^k + 0.5} \qquad (3.24)$$

Essa versão suavizada $\hat{\sigma}^{(d)s}_{t+1d}$ é aquela a ser usada definitivamente para prever as observações fora da amostra. Para maiores detalhes e exemplos a respeito, ver o trabalho de Kuhn e Johnson (2013).

3.2.3 Métodos Bayesianos

Há muitas maneiras de se regularizar um modelo. Uma abordagem que será desenvolvida é a regularização informativa por meio de *prioris* Bayesianas. Para desenvolver o conceito, o primeiro passo é revisar inferência Bayesiana no modelo de regressão linear clássico.

3.2.3.1 O Modelo

Seja o vetor resposta $\sigma^{(d)}_{t+1d}$ a ser regredido em uma matriz de regressores representada por Σ, sendo a primeira coluna dessa matriz de regressores preenchida apenas pelo número 1. Uma primeira possibilidade seria

modelar esse relacionamento através do modelo linear clássico:

$$\sigma_{t+1d}^{(d)} = \Sigma\beta + \epsilon, \quad \epsilon \sim N(0_N, \mathbb{I}h^{-1}), \tag{3.25}$$

em que $\mathbb{I}$ é uma matriz identidade.

Pode-se demonstrar que uma variável aleatória normal, adicionada de uma constante, gera uma outra variável aleatória normal com uma mudança de nível mas de mesma variância, e.g. $\sigma_{t+1d}^{(d)} = \Sigma\beta + \epsilon \sim N(\Sigma\beta, \mathbb{I}h^{-1})$, implicando, assim, a seguinte função de verossimilhança:

$$p(\sigma_{t+1d}^{(d)}|\beta, h) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \exp\left[-\frac{h}{2}(\sigma_{t+1d}^{(d)} - \Sigma\beta)'(\sigma_{t+1d}^{(d)} - \Sigma\beta)\right], \tag{3.26}$$

Visando facilitar futuras conjugações, pode-se reescrever (3.26) com as quantidades de mínimos quadrados ordinários (MQO):

$$\nu = N - k, \tag{3.27}$$

$$\hat{\beta} = (\Sigma'\Sigma)^{-1}(\Sigma'\sigma_{t+1d}^{(d)}), \tag{3.28}$$

$$s^2 = \frac{(\sigma_{t+1d}^{(d)} - \Sigma\hat{\beta})'(\sigma_{t+1d}^{(d)} - \Sigma\hat{\beta})}{\nu}. \tag{3.29}$$

resultando em uma equação que se assemelha à função de densidade normal-Gamma:

$$p(\sigma_{t+1d}^{(d)}|\beta,h) = \left[\frac{1}{(2\pi)^{\frac{N}{2}}}h^{\frac{k}{2}}\exp\left[-\frac{h}{2}(\beta-\hat{\beta})'\Sigma'\Sigma(\beta-\hat{\beta})\right]\right] \left[h^{\frac{\nu}{2}}\exp\left[-\frac{h}{2}\frac{\nu}{s^{-2}}\right]\right]. \tag{3.30}$$

Essa densidade, para $\sigma_{t+1d}^{(d)}$, não apresenta, explicitamente, a variável aleatória, mas $\sigma_{t+1d}^{(d)}$ afeta essa função por dois canais diferentes, de maneira implícita. A primeira parte, que se parece com a densidade normal para $\beta|h$, e a segunda parte, que se parece com uma densidade Gamma para h. De fato, essa função tem o formato de uma densidade normal-Gamma marginalizada $p(\beta,h) = p(\beta|h)p(h)$, reorganizada por meio de uma densidade independente para h e uma densidade condicional para β.

É interessante filtrar as *prioris* possíveis para aquelas que geram *posteriores* em forma fechada. Um tipo especial de *prioris* é a conjugada natural, que tem duas propriedades: tem a mesma forma funcional da verossimilhança, e gera uma posteriori que também tem a mesma forma funcional. Essa simplificação na estrutura tornará a inferência mais simples e rápida.

A densidade normal-Gamma têm conjugada natural, assim, a expressão do conhecimento a *priori* se

dará através dos hiperparâmetros da função de densidade normal-Gamma, decomposta na primeira densidade condicional para $\beta|h$, e a segunda função de densidade para h como em (3.31):

$$p(\beta,h) = \frac{h^{\frac{k}{2}}}{(2\pi)^{\frac{k}{2}}}|V|^{-\frac{1}{2}}\exp\left[-\frac{h}{2}(\beta-\beta)'V^{-1}(\beta-\beta)\right] \cdot \left[\left(\frac{2s^{-2}}{\nu}\right)^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)\right]^{-1} h^{\frac{\nu-2}{2}}\exp\left[-\frac{h}{2}\frac{\nu}{s^{-2}}\right]. \tag{3.31}$$

Essa densidade tem quatro hiperparâmetros $\underline{\beta}, \underline{V}, \underline{s}, \underline{v}$ que podem ser escolhidos para melhor representar o conhecimento a *priori* sobre o verdadeiro processo gerador de dados. Gary (2003) mostra que a posteriori gerada por (3.30) e (3.31) é uma densidade normal-Gamma como em (3.33):

$$p(\beta,h|y) = \frac{|\bar{V}|^{-\frac{1}{2}}}{\left(\frac{2\bar{s}^{-2}}{\bar{\nu}}\right)^{\frac{\bar{\nu}}{2}}\Gamma\left(\frac{\bar{\nu}}{2}\right)(2\pi)^{\frac{k}{2}}}h^{\frac{\bar{\nu}+k-2}{2}} \tag{3.32}$$

$$\exp\left[-\frac{h}{2}\left(\bar{\nu}\bar{s}^2+(\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right)\right] \tag{3.33}$$

em que:

$$\overline{V} = (\underline{V}^{-1}+\Sigma'\Sigma)^{-1}, \quad \overline{\beta} = \overline{V}(\underline{V}^{-1}\underline{\beta}+X'X\hat{\beta}), \quad \overline{v} = \underline{v}+N, \tag{3.34}$$

$$\overline{vs}^2 = \underline{vs}^2 + vs^2 + (\hat{\beta} - \underline{\beta})'[\underline{V} + (X'X)^{-1}]^{-1}(\hat{\beta} - \underline{\beta}). \quad (3.35)$$

Pode-se ver que os parâmetros da *posteriori* são influenciados tanto pelas estimativas da verossimilhança, quanto pelos hiperparâmetros da *priori*. Esse resultado será usado para passar o conhecimento externo para a densidade a posteriori. Os momentos correspondentes dessa densidade são:

$$E[\beta|\sigma^{(d)}_{t+1d}] = \overline{\beta}, \quad Var[\beta|\sigma^{(d)}_{t+1d}] = \frac{\overline{vs}^2}{\overline{v} - 2}\overline{V}, \quad (3.36)$$

$$E[h|\sigma^{(d)}_{t+1d}] = \overline{s}^{-2}, \quad Var[h|\sigma^{(d)}_{t+1d}] = \frac{2\overline{s}^{-4}}{\overline{v}}. \quad (3.37)$$

Após recuperar esse *background* Bayesiano, deve-se pensar no conhecimento a *priori*. O procedimento de inferência Bayesiana será usado para regularizar uma regressão da futura volatilidade realizada σ^2_t nas 22 médias das últimas j volatilidades realizadas σ^2_{t-j}, $j = 1, 2, \ldots, 22$. O primeiro problema a ser evitado, ao se inserir vinte e duas defasagens na regressão linear, é o da multicolinearidade. Volatilidade, por sua natureza, tem uma memória forte e, ao se trabalhar com médias de volatilidade defasadas, isso pode ser ainda mais significativo. Assim, apesar de se ter as informações de todas as 22 defasagens diretamente, usá-las

como regressores decrescerá, significativamente, a precisão das previsões. Em adição a isso, conforme mencionado anteriormente, nem todas as defasagens devem ter a mesma importância.

3.2.3.2 Primeira Priori: Decaimento Exponencial

A proposta da primeira *priori* é focar somente no primeiro objetivo, isto é, na redução da multicolinearidade. Para reduzir o efeito que uma defasagem tem em toda a *priori*, se poderia, inicialmente, passar dentro da *priori* que a média dos parâmetros é igual à zero:

$$\underline{\beta} = 0_{23x1}, \tag{3.38}$$

No entanto, multicolinearidade não é um problema da média, mas da variância. Assim, implicar que o primeiro momento dos parâmetros seja zero não só não diminuiria a sua variância, como induziria a *priori* a fazer previsões com pouca qualidade que todas as observações são iguais a 0 por exemplo, algo diferente do resultado pretendido para a posteriori, o que levaria a uma falta de credibilidade para essas previsões por meio de *prioris*.

Primeiramente, o intercepto não gerará multicolinearidade, então não há qualquer problema em deixá-lo ser diferente de zero. Assim, pode-se colocar um valor de qualquer magnitude na matriz

de escala, permitindo que o parâmetro se ajuste de acordo com os dados. Como, em geral, a volatilidade de amanhã é próxima da volatilidade de hoje, o intercepto provavelmente será próximo de zero, mas a ideia é alocar um valor de escala grande para ele $\underline{V}_{1,1} = 1$. Ainda tentando reduzir a multicolinearidade, todos os betas fora da diagonal principal serão igualados a 0, $\underline{V}_{i,j} = 0$ para todo $i \neq j$.

Com exceção do intercepto, todos os termos na diagonal principal utilizarão um padrão de decaimento exponencial, em que $V_{i,i} = \Gamma_i^{\Psi}$, $i = 1, 2, \ldots, 22$. Esse vetor Γ é um *grid* igualmente espaçado, que depende dos hiperparâmetros γ e Ψ que são escolhidos, conjuntamente, por validação cruzada. A estrutura de Γ e os dois *grids* de procura serão:

$$\Gamma = Grid[1/\gamma, 1/(22 * \gamma)], \quad \gamma = Grid[1, 10], \quad (3.39)$$

$$\Psi = 10^{\psi}, \quad \psi = Grid[0, 0.5], \quad (3.40)$$

em que $Grid[\rho_1, \rho_2]$ é um *grid* igualmente espaçado, que inicia em ρ_1 e vai até ρ_2. O primeiro *grid* de procura para γ define uma sequência de valores para Γ que será a sequência usada, efetivamente, na matriz de escala. A *priori* final para a matriz de escala será uma matriz diagonal dos elementos concatenados $\underline{V} = diag(1, \Gamma)$.

Também é necessário falar sobre a *priori* para $\underline{v}$ e $\underline{s}$. Os valores foram escolhidos tentando tornar a *priori* plausível para o conjunto de dados. A variação total dentro de uma série é, usualmente, menor do que 8%, os menores valores verificados tipicamente são próximos a 2% e os maiores são tipicamente próximos a 10%. O HAR consegue explicar, em geral, pelo menos 50% da variação total. Assim, serão colocados valores para $\underline{v}$ e $\underline{s}$ de maneira a gerar $E[h] = 8$ e $Var[h] = 4$. Sabe-se que se chegará, de acordo com a função Gamma, suficientemente perto de uma posteriori com média de 4 em poucos desvios-padrões, mas, caso o modelo não se ajuste bem aos dados, ainda se estará perto o suficiente de nenhum poder explicativo, representada por uma média maior da posteriori para h. Efetivamente os valores usados foram $\underline{s} = 8^{-1/2}$ e $\underline{v} = 32$.

Na Figura 16, estão representados quatro *box-plots* de 1000 pontos amostrados aleatoriamente, de quatro possíveis densidades dentro do espaço de procura que acabou de ser apresentado. Os hiperparâmetros escolhidos serão encontrados por validação cruzada e os parâmetros que gerarem os menores MSE serão utilizados para a previsão fora da amostra, sem que haja modificação durante o processo de estimação da janela móvel.

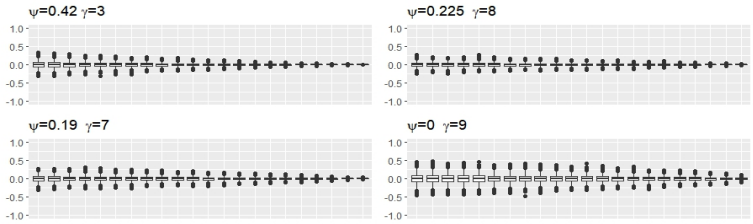


Figura 16 – Diferentes comportamentos que a primeira *priori* pode apresentar.
Foram amostrados 1000 pontos de 4 *prioris* aleatoriamente escolhidas para representar a primeira proposta.

3.2.3.3 Segunda Priori- HAR

A segunda *priori* tentará passar a teoria por trás do HAR. Ao tomar o limite dos hiperparâmetros, o HAR aparecerá como um caso especial. A *priori* acreditará que somente as defasagens $1, 5, 22$ são importantes e, para isso, colocará os seus valores esperados para aqueles sabidamente comuns nas estimativas do HAR. Como a *priori* acredita que toda a informação importante está dentro dessas defasagens, pode-se colocar a média dos outros parâmetros para 0. Os valores usados são:

$$\underline{\beta}_{HAR} = 0.2, \quad \beta_{\neq HAR} = 0, \tag{3.41}$$

em que o β é respectivamente o intercepto e as três médias de defasagens do *HAR* original são $HAR = \{1, 5, 22\}$.

Mas, a ideia de colocar os outros parâmetros para 0 dentro da *priori*, ocorre porque se quer evitar os

problema típicos da multicolinearidade. Na prática, não se quer que alguns desses parâmetros sejam 0 e, então, será necessário testar diferentes convicções sobre o HAR. Em um limite haverá o AR(22) e, no outro, o HAR(1,5,22). Caso colocássemos a variância dos parâmetros de 2 à 4 e de 6 a 22 igual a 0, a posteriori iria em direção ao HAR (1,5,22) pois apenas os parâmetros 1, 5 e 22 seriam diferentes de zero na posteriori. No caso contrário, caso colocássemos a precisão igual a 0, a posteriori iria em direção ao AR(22).

Novamente, todos os valores fora da diagonal principal serão zero $\underline{V}_{i,j} = 0 \; i \neq j$, e se permitirá que os parâmetros do HAR fiquem significativamente livres. Os valores efetivamente utilizados foram $\underline{V}_{0,1,5,22} = 0.2$. Para todas as outras médias de defasagens, o nível de crença será uma constante real $V_{d,d} = \omega_1$ para $d = 2, 3, 4$ e $V_{s,s} = \omega_2$ para $s = 6, 7, \ldots, 22$.

Essa estrutura equaliza a força da crença de que as defasagens 2 a 4 são iguais a 0, assim como as defasagens indo de 6 a 22. No limite, quando $\omega_1 \to \infty$ e $\omega_2 \to \infty$, têm-se o HAR original. No outro extremo, há o AR(22). Esses dois hiperparâmetros ω_1, ω_2 serão escolhidos por meio de validação cruzada buscando minimizar a MSE, e o espaço de busca está apresentado em (3.42):

$$\omega_{1,2} = Grid[1e^{-1}, 1e^{-15}]. \tag{3.42}$$

Os valores para $\underline{v}$ e $\underline{s}$, serão os mesmos já utilizados na *priori* anterior, respectivamente $\underline{s} = 8^{-1/2}$ e $\underline{v} = 32$. A Figura 17 ilustra quatro possíveis comportamentos que poderiam ser gerados ao se amostrar dessa *priori*. Pode-se ver que há duas partes de ajuste independentes, permitindo colocar uma *priori* solta para o primeiro grupo de defasagens e uma *priori* apertada para o segundo grupo.

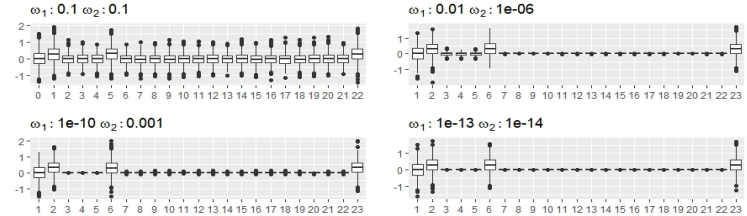


Figura 17 – Diferentes comportamentos que a segunda *priori* é capaz de representar.
Foram amostrados 1000 pontos de quatro *prioris* aleatoriamente escolhidas para representar a flexibilidade da segunda proposta.

3.2.3.4 Densidade Preditiva

Uma vez que as duas *prioris* foram apresentadas e se sabe o formato da posteriori, Koop (2003) mostra que se pode usar esse procedimento de inferência para prever os novos valores para $\hat{\sigma}^{(d)}_{t+1d}$ com base em novos dados observados Σ^*. A densidade preditiva é descrita por uma distribuição *t* de student $\hat{\sigma}^{(d)}_{t+1d}|\sigma^{(d)}_{t+1d} \sim t(\overline{\beta}\Sigma^*, \overline{s}^2\{1+$

$\overline{V}\Sigma^{*2}\}, \overline{v})$, com os correspondentes momentos:

$$E[\hat{\sigma}^{(d)}_{t+1d}|\sigma^{(d)}_{t+1d}] = \Sigma^*\overline{\beta}, \tag{3.43}$$

$$Var[\hat{\sigma}^{(d)}_{t+1d}|\sigma^{(d)}_{t+1d}] = \overline{s}^2[\mathbb{I}_T + \Sigma^*\overline{V}\Sigma^{*'}]\frac{\overline{v}}{\overline{v}-2}. \tag{3.44}$$

Como há equações em forma fechada para a densidade preditiva, é possível realizar as previsões apenas por uma estatística descritiva e, neste trabalho, será utilizada a média $\Sigma^*\overline{\beta}$.

3.2.4 Combinação de Previsões

3.2.4.1 Método da Regressão

Modelar um processo gerador de dados nem sempre é um processo fácil. Encontrar a especificação correta pode levar anos, décadas ou, ainda, ser uma jornada sem fim. Mas, é possível imaginar um cenário, muito comum em economia, em que múltiplas especificações foram propostas e cada especificação captura uma diferente parte da dinâmica do processo gerador de dados, e poderia surgir o questionamento de que se a união desses modelos mal especificados obteria um desempenho preditivo superior a dos modelos individuais.

Com essa ideia inspiradora, chega-se ao mundo de combinação de previsões presente na teoria econométrica,

pelo menos, desde os trabalhos de Nelson (1972) e Cooper e Nelson (1975). O método da regressão é um dos métodos mais simples para se combinar previsões. Com a disponibilidade de $n = 1, 2, ...N$ pontos estimados $\hat{\sigma}^{(d)}_{n,t+1d}$ dentro da amostra, pode-se estimar um modelo linear do tipo:

$$\sigma^{(d)}_{t+1d} = \beta_0 + \sum_{i=1}^{n} \beta_i \hat{\sigma}^{(d)}_{n,t+1d}. \tag{3.45}$$

Espera-se que, se a previsão do primeiro modelo $\hat{\sigma}^{(d)}_{1,t+1d}$ for feita com base no verdadeiro processo gerador de dados, a previsão dos demais modelos não conterá nenhuma nova informação referente a volatilidade esperada que já não esteja embutida na estimação do primeiro modelo, sendo nula a contribuição dos demais modelos para a previsão final $\hat{\beta}_i = 0 \ \forall \ i \neq 1$. Mas, no caso mais realístico de má-especificação do modelo, Granger e Ramanathan (1984) demonstram que a combinação das previsões através da regressão pode gerar previsões mais acuradas do que qualquer previsão individual.

Essa foi uma das primeiras metodologias de combinação de previsões e, hoje em dia, ela é conhecida apenas por *método da regressão*. Essa abordagem foi, primeiramente, usada por Patton e Sheppard (2009) em problemas de previsão conectados ao HAR, mas, diferentemente do trabalho deles, que tentou combinar diferentes estimativas de volatilidade, agora serão

combinados diferentes modelos.

A estimação será feita através da minimização de duas diferentes funções perda: a MSE e a QLIKE, a serem apresentadas na Seção 3.2.5.1. Além disso, será testado, também, dois conjuntos de restrições na otimização; primeiramente a restrição será $\beta_i > 0$ e, posteriormente, $\beta_i > 0$ e $\sum_{i=1}^{n} \beta_i = 1$, para $i = 1, 2, \ldots, 22$, excluindo o intercepto de ambas as restrições. Apesar de se saber que esse tipo de restrição pode ser inconveniente (DIEBOLD, 1988), verifica-se que restringir o modelo ajuda a evitar instabilidade numérica na presença de *outliers*. Para uma revisão completa sobre o assunto, veja o trabalho de Diebold (2017).

3.2.4.2 Dynamic Model Averaging (DMA)

Alguns avanços têm sido apresentados na literatura buscando especificações mais robustas. Uma proposta relevante foi a ponderação dinâmica de modelos ou *Dynamic Model Averaging* (DMA) de Raftery, Kárnỳ e Ettler (2010). Suponha que se tenha um vetor σ_t com uma correspondente matriz de regressores Σ_t contendo defasagens do regredido. Suponha, também, que o vetor alvo seja uma combinação linear das defasagens, em que os parâmetros da combinação evoluem ao longo do tempo seguindo um processo autorregressivo. Pode-se reescrever

esse modelo no formato típico de espaço de estados:

$$\sigma_t^{(d)} = \Sigma_t^T \theta_t + \epsilon_t, \tag{3.46}$$

$$\theta_t = \theta_{t-1} + \delta_t, \tag{3.47}$$

com $\epsilon_t \overset{iid}{\sim} N(0, V)$ e $\delta_t \overset{ind}{\sim} N(0, W_t)$. A equação (3.46) é popularmente conhecida como equação de medida, e a equação (3.47) é conhecida como a equação de transição.

Considere o caso em que haja múltiplos modelos $M_1, ..., M_k$ que podem ser expressos por meio de (3.46) e (3.47), com $\{\theta_t^k, \Sigma_t^k\}$ diferente para cada um dos k modelos. Simboliza-se por $L_t = k$ o processo governado pelo M_k-modelo durante o período t. Pode-se reescrever as equações (3.46) e (3.47) como:

$$\sigma_t^{(d)} | L_t = k \sim N(\Sigma_t^{(k)T} \theta_t^{(k)}, V^{(k)}), \tag{3.48}$$

$$\theta_t^{(k)} | L_t = k \sim N(\theta_{t-1d}^{(k)}, W_t^{(k)}). \tag{3.49}$$

Nesse tipo de modelagem, é necessário especificar a matriz de transição dos estados $Q = (q_{kl})$ de dimensão kXk, em que $q_{kl} = P[L_t = l | L_{t-1d} = k]$ é a probabilidade de o processo ser governado pelo modelo M_l no próximo período, uma vez que hoje ele é governado pelo modelo M_k. Mas, Raftery, Kárnỳ e Ettler (2010) sugerem uma

aproximação diferente para essa cadeia de Markov. Eles propõem que a transição seja feita por uma regra de atualização para $\theta_t^{(k)}$. A metodologia é feita em duas etapas: a primeira é prever o indicador de modelo L_t usando a *equação de previsão dos modelos*:

$$\pi_{t|t-1d,k} = \frac{\pi_{t-1d|t-1d,k}^{\alpha}}{\sum_{l=1}^{K} \pi_{t-1d|t-1d,k}^{\alpha}}, \tag{3.50}$$

em que α é um parâmetro de esquecimento, tipicamente menor do que 1 e $\pi_{t|t-1d,k}$ são as probabilidades de que em t se tenha L_k, uma vez que se está em $t - 1d$. Após isso, padroniza-se as probabilidades usando a *equação de atualização dos modelos*:

$$\pi_{t|t,k} = \frac{\omega_{tk}}{\sum_{l=1}^{k} \omega_{t,l}}, \tag{3.51}$$

com $\omega_{t,l} = \pi_{t|t-1d,l} f_l(\sigma_t^{(d)}|\mathbf{\Sigma}_t)$ e $f_l(\sigma_t^{(d)}|\mathbf{\Sigma}_t)$ sendo a densidade normal $N(\Sigma^{(l)T}\hat{\theta}_{t-1d}^{(l)}, V^{(l)} + \Sigma_t^{(l)T} R_t^{(l)} \Sigma_t^{(l)})$ avaliada em σ_t. A previsão um passo à frente é calculada por $\hat{\sigma}_t^{(d),DMA} = \sum_{k=1}^{K} \pi_{t|t-1d,k} \hat{\sigma}_t^{(d),(k)}$.

Inspirados nessa abordagem, Caldeira, Moura e Santos (2018) propuseram uma atualização dos pesos baseada em critérios econômicos. No lugar de atualizar os pesos com base em medidas estatísticas, propuseram que os pesos sejam atualizados pelos retornos que os portfólios tiveram. Definindo $rx_{i,t}^{net}$ como o retorno líquido de um

portfólio, a equação (3.51) é modificada para:

$$\omega_{t|t,i} = \frac{\omega_{t|t-1d,i}(1 + rx_{i,t}^{net})}{\sum_{i=1}^{M} \omega_{t|t-1d,i}(1 + rx_{i,t}^{net})}, \tag{3.52}$$

em que, em vez de utilizar uma medida estatística para avaliar a performance preditiva, usa-se uma medida econômica, isto é, os retornos líquidos. O modelo que teve maior retorno no período anterior ganhará um peso maior para o próximo período. A taxa de esquecimento do modelo α fica, tipicamente, entre 0.95 e 0.99.

A segunda metodologia para combinação de previsões será inspirada em Caldeira, Moura e Santos (2018). Em cada período, serão combinados todos os N modelos gerando uma previsão $\hat{\sigma}_t^2 = \sum_{i=1}^{N} w_i h_{t,i}$. No primeiro período, os pesos serão distribuídos igualitariamente para todos os modelos $w_i = 1/n$, e os pesos serão atualizados levando-se em conta a sua performance relativa:

$$\pi_t = \frac{w_{i,t-1} * L(\sigma_{t-1}^2, \hat{\sigma}_t^2)^{-1}}{\sum_1^n (w_{i,t-1} * L(\sigma_{t-1}^2, \hat{\sigma}_t^2)^{-1})}, \tag{3.53}$$

$$w_t = \frac{\pi_t^{\omega}}{\sum_1^n \pi_{t,k}^{\omega}}, \tag{3.54}$$

em que $L(.,.)$ é uma função perda a ser apresentada na Seção 3.2.5 e ω é um parâmetro de esquecimento que controla a velocidade da atualização. No presente trabalho,

serão testadas três diferentes opções de esquecimento $\gamma = \{0.95, 0.8, 0.5\}$. A equação (3.53) torna os pesos para o próximo período proporcionais à performance preditiva do período anterior e quanto maior for o valor da função perda L menor será o peso do modelo. A equação (3.54) aplica uma taxa de esquecimento que suaviza a regra de atualização.

3.2.5 Metodologia para Avaliar as Previsões de Volatilidade Realizada.

Métodos de comparação de previsões são cruciais em problemas de volatilidade realizada, pois, apesar da existência de diversos estimadores de volatilidade realizada terem sido apresentados na literatura, não há como se descobrir qual é a verdadeira volatilidade realizada. Para aqueles interessados em uma revisão completa do tema, ver Violante e Laurent (2012).

3.2.5.1 Funções Perda

Andersen, Bollerslev e Meddahi (2005) mostraram que, na existência de erros de estimação, algumas funções perda poderiam, assintoticamente, apontar que um modelo que não é, de fato, o melhor, fosse eleito como se fosse. Hansen e Lunde (2006) apresentaram as condições suficientes que uma medida de erro deve cumprir para que

o modelo apontado como o melhor efetivamente seja. Seja $h_{k,t}$ a previsão do modelo k no período t; σ_t^2 a verdadeira variância; e $\hat{\sigma}_t^2$ a sua proxy. Uma função perda não-viesada condicionalmente pode ser definida como:

$$E[L(\sigma_t^2, h_{k,t})] \leq E[L(\sigma_t^2, h_{j,t})] \leftrightarrow$$
$$E[L(\hat{\sigma}_t^2, h_{k,t})] \leq E[L(\hat{\sigma}_t^2, h_{j,t})].$$

Hansen e Lunde (2006) mostram que uma condição suficiente para que essa condicional dupla seja verdadeira é que:

$$\frac{\partial^2 L(\sigma_t^2, h_t)}{(\partial \sigma_t^2)^2} \text{ exista e não dependa de } h_t. \tag{3.55}$$

Patton (2011) desenvolveu uma família de funções capazes de selecionar a verdadeira ordenação dos modelos. A família tem um hiperparâmetro ξ que controla o grau de homogeneidade da função. A família completa de funções é definida como:

$$L(\hat{\sigma}_t^2, h_t) = \begin{cases} \frac{1}{(\xi-1)\xi}(\hat{\sigma}_t^{2\xi} - h_t^{\xi}) - \frac{1}{\xi-1} h_t^{\xi-1}(\hat{\sigma}_t^2 - h_t), & \text{para } \xi \notin (0,1); \\ h_t - \hat{\sigma}_t^2 + \hat{\sigma}_t^2 log\frac{\hat{\sigma}_t^2}{h_t}, & \text{para } \xi = 1; \\ \frac{\hat{\sigma}_t^2}{h_t} - log\frac{\hat{\sigma}_t^2}{h_t} - 1, & \text{para } \xi = 0; \end{cases} \tag{3.56}$$

O hiperparâmetro ξ tem importante propriedade de controlar o grau de simetria da função perda. Se $\xi = 2$, a função perda será simétrica; se $\xi < 2$, ela penalizará mais

fortemente previsões com erros negativos; e se $\xi > 2$ ela penalizará mais fortemente previsões com erros positivos. Este trabalho usará dois casos especiais dessa família de funções:

$$\text{MSE: } (\xi = 2) \quad L(\hat{\sigma}_t^2, h_t) = (\hat{\sigma}_t^2 - h_t)^2, \tag{3.57}$$

$$\text{QLIKE: } (\xi = 0) \quad L(\hat{\sigma}_t^2, h_t) = \frac{\hat{\sigma}_t^2}{h_t} - log\frac{\hat{\sigma}_t^2}{h_t} - 1. \tag{3.58}$$

A equação (3.57) é a conhecida função erro quadrado médio, que pode ser derivada da densidade gaussiana e essa é a única função simétrica da família. A segunda função (3.58) é conhecida como QLIKE e é amplamente utilizada para avaliar previsões financeiras. Essa medida de erro, como dito anteriormente, penaliza de maneira mais grave previsões abaixo do valor realizado do que previsões acima do valor realizado.

3.2.5.2 Teste para Habilidade Preditiva Condicional (CPA)

Comparação em pares é um procedimento em que se compara duas séries de previsões. Esse teste será utilizado neste trabalho para encontrar os modelos que conseguem apresentar desempenho diretamente melhor do que o HAR. Múltiplos testes existem para realizar a comparação em pares, mas, desde que Giacomini e White

(2006) apresentaram o teste para habilidade preditiva condicional (*Conditional Predictive Ability*) (CPA), este se tornou o principal teste na literatura.

O teste CPA generaliza o teste proposto por Diebold e Mariano (2002) de muitas formas: permitindo que modelos aninhados e não aninhados sejam comparados, incorporando erros de medida nas variáveis e criando um ambiente que permite não somente diferenciar modelos que tiveram a melhor performance no passado, mas inferir quais são os modelos que terão a melhor performance em uma data específica no futuro, dada a sua performance passada.

Sendo $L(\sigma_t^2, h_{k,t})$ a função perda do modelo k, $d_t = L(\sigma_t^2, h_{k,t}) - L(\sigma_t^2, h_{j,t})$ a diferença entre duas funções perdas, τ o horizonte de previsão e $\mathscr{F}_t$ o conjunto de informação disponível durante o período t, pode-se definir a hipótese nula como:

$$\boldsymbol{H}_0 : \quad E[L(\sigma_{t+\tau}^2, h_{k,t+\tau}) - L(\sigma_{t+\tau}^2, h_{j,t+\tau})|\mathscr{F}_t] \equiv \tag{3.59}$$

$$E[\Delta L_{t+\tau|}|\mathscr{F}_t] = 0 \tag{3.60}$$

Pela hipótese nula, pode-se reescrever a diferença das sequências de previsões $\Delta L_{m,y+1}$ como uma sequência de diferenças em martingale, reescrevendo (3.60) como $E[\delta_t d_{\mathscr{T},t}] = 0$ para todas as funções mensuráveis δ_t, sendo

δ_t a função teste. De acordo com a teoria assintótica, pode-se escrever a estatística teste para as previsões um passo à frente como:

$$T_{n,m}^h = n(n^{-1}\sum_{t=m}^{T-1} h_t \Delta L_{m,t+1})'\hat{\Omega}_n^{-1}(n^{-1}\sum_{t=m}^{T-1} h_t \Delta L_{m,t+1}) \tag{3.61}$$

$$= n\overline{Z}'_{m,n}\hat{\Omega}_n^{-1}\overline{Z}_{m,n}, \tag{3.62}$$

$$T_{n,m}^h \xrightarrow{p} \chi_q^2 \quad conforme \quad n \to \infty, \tag{3.63}$$

em que $\overline{Z}_{m,n} \equiv n^{-1}\sum_{t=m}^{T-1} Z_{m,t+1}$, $Z_{m,t+1} \equiv h_t \Delta L_{m,t+1}$ e $\hat{\Omega}_n \equiv n^{-1}\sum_{t=m}^{t-1} Z_{m,t+1}Z'_{m,t+1}$, sendo $H_{A,h}$: $E[\overline{Z}'_{m,n}\overline{Z}_{m,n} \geq \delta > 0]$ a hipótese alternativa.

Caso a hipótese nula seja rejeitada, os autores propõem comparar os coeficiente da regressão de $\Delta L_{m,t+\tau}$ em h_t sobre todo o período fora da amostra $t = m, ..., T - \tau$. Após definir um limite c, aponta-se que o modelo j é superior se $h_{j,t}\hat{\alpha}_n > c$, e o modelo k, se $h_{k,t}\hat{\alpha}_n < c$. em que $\overline{Z}_{m,n} \equiv n^{-1}\sum_{t=m}^{T-1} Z_{m,t+1}$, $Z_{m,t+1} \equiv h_t \Delta L_{m,t+1}$ e $\hat{\Omega}_n \equiv n^{-1}\sum_{t=m}^{t-1} Z_{m,t+1}Z'_{m,t+1}$.

Esse teste também é extensível para previsões multiperíodos. Ele será utilizado para comparar previsões feitas cinco passos a frente. Para mais a respeito desse procedimento, ver Giacomini e White (2006).

3.2.5.3 Conjunto de Modelos de Confiaça (MCS)

A comparação múltipla é outra abordagem para realizar comparações de modelos. Muda-se a abordagem de comparar a performance entre dois modelos para comparar a performance entre vários modelos. O teste de comparação em grupo mais popular é o conjunto de modelos de confiança (*Model Confidence Set*, MCS)(HANSEN; LUNDE; NASON, 2011).

Seguindo Violante e Laurent, defina-se o conjunto inicial de modelos com capacidade superior por M^0. Este conjunto inicial contém todos os modelos dentro do conjunto de modelos de previsão. Definindo a performance relativa de cada modelo por $d_{k,j,t} = L(\sigma_t, h_{k,t}) - L(\sigma_t, h_{j,t})$ para todo $k \neq j \in M^0$, a hipótese nula se torna:

$$H^0 : \quad E[d_{k,j,t}] = 0 \quad \forall \quad k, j \in M^0 \tag{3.64}$$

Se a hipótese nula for rejeitada em um nível escolhido de confiança α, então o pior modelo será removido pela regra da eliminação, e o teste é repetido até que haja a não rejeição da hipótese nula. Ao se manter o nível de confiança α inalterado em cada iteração, cria-se um conjunto $(1 - \alpha)$ de confiança $M^* = \{k \in M_0 : E(d_{k,j,t}) \leq 0 \ \ \forall j \in M^0\}$. As duas estatísticas de teste

utilizadas são:

$$MCS_{T_Q} = T(T^{-1}\sum_{t=1}^{T}\iota_{\dagger}'\mathbf{L}_t)'\hat{\Omega}^{+}(T^{-1}\sum_{t=1}^{T}\iota_{\dagger}'\mathbf{L}_t) \rightarrow \chi_q^2 \tag{3.65}$$

$$MCS_{T_F} = \frac{T-q}{q(T-1)}MCS_{T_Q} \rightarrow F_{q,t-q} \tag{3.66}$$

em que L_t é o vetor de performance amostral; $\iota_{\dagger}$ é o complemento ortogonal de um vetor m-dimensional de 1s; m é o número de modelos remanescentes; $\hat{\Omega}^{+}$ é a pseudo inversa More-Penrose de $\hat{\Omega}$, um estimador consistente de Ω com $q = rank(\hat{\Omega})$. Na prática, é dificil estimar $\hat{\Omega}$ e um procedimento de *bootstrap* é usado como aproximação. Para mais detalhes, consultar Hansen, Lunde e Nason (2011).

Se a hipótese nula for rejeitada, segue-se a regra recomendada e será excluído o modelo que tiver o maior excesso de perda relativo, $\epsilon_M = argmax_{k\in M}t_k$, em que $t_k = \sqrt{T}\overline{d}_k/\sqrt{\hat{\omega}_k^D}$, $k = 1,\ldots,m$,$k \neq j$, e $\overline{d}_k = m^{-1}\sum_{j\in M}\overline{d}_{k,j}$, $\overline{d}_{k,j} = T^{-1}\sum_{t=1}^{T}d_{k,j,t}$. O teste acaba quando a hipótese nula não for mais rejeitada, ou quando houver apenas um modelo restante.

3.3 Análise Empírica

3.3.1 Dados

A base de dados consiste do registro de fechamento dos preços a cada 5 minutos durante 2767 dias, iniciando em 17/12/2003 e avançando até 16/12/2014 para 10 empresas americanas: Apple (AAPL) , eBay (EBAY), Goldman Sachs Group (GS), Microsoft (MSFT), Procter Gamble (PG), Abbott Laboratories (ABT), Bank of America (BAC), Chevron (CVX), General Electric (GE), Caterpillar (CAT). A volatilidade realizada foi estimada com retornos amostrados a cada 5 minutos, gerando dez séries de volatilidade realizada com 2767 observações. Utilizou-se o pacote de Boudt, Cornelissen e Payseur (2018) para mensurar a volatilidade realizada. A Figura 18 apresenta a série de volatilidade realizada e a Tabela 2 apresenta as estatísticas descritivas da amostra.

Pode-se ver que, devido ao baixo valor da volatilidade realizada, a sua média costuma ser menor que um milésimo. Todas as séries têm um desvio-padrão igual ou maior que a média, a assimetria costuma ser positiva e todas as séries apresentaram excesso de curtose, evidenciando a alta presença de valores extremos. É possível verificar na Figura 18 a evolução da volatilidade ao longo do tempo.

	Média	Desv. Pad.	Assim.	Kurt.
AAPL	0.0003	0.0005	9.03	135.01
EBAY	0.0003	0.0003	5.48	45.88
GS	0.0003	0.0012	16.42	403.57
MSFT	0.0001	0.0002	7.40	77.29
PG	0.0001	0.0001	9.02	105.04
ABT	0.0002	0.0006	8.35	99.02
BAC	0.0005	0.0017	7.43	72.99
CVX	0.0001	0.0004	10.77	172.47
GE	0.0002	0.0007	9.42	133.94
CAT	0.0002	0.0004	6.06	48.83

Tabela 2 – Estatísticas Descritivas da Amostra
Estatísticas descritivas da volatilidade realizada estimada com base no estimador de volatilidade realizada de (ANDERSEN; BOLLERSLEV; MEDDAHI, 2011), tendo como base retornos amostrados a cada 5 minutos durante 2767 dias indo de 17/12/2003 até 16/12/2014.

3.3.2 Detalhes da implementação dos modelos

Com o conjunto de dados descrito na Seção 3.3.1, dividiu-se a amostra em um conjunto de estimação, consistindo dos 80% dias iniciais disponíveis, e um conjunto de teste consistindo dos últimos 20%.

Tamanho do conjunto de treinamento:	2213 Dias
Tamanho do conjunto de teste:	554 Dias

Tabela 3 – Tamanho dos Conjunto de Treinamento e Teste.

Todos os modelos e combinações apresentados

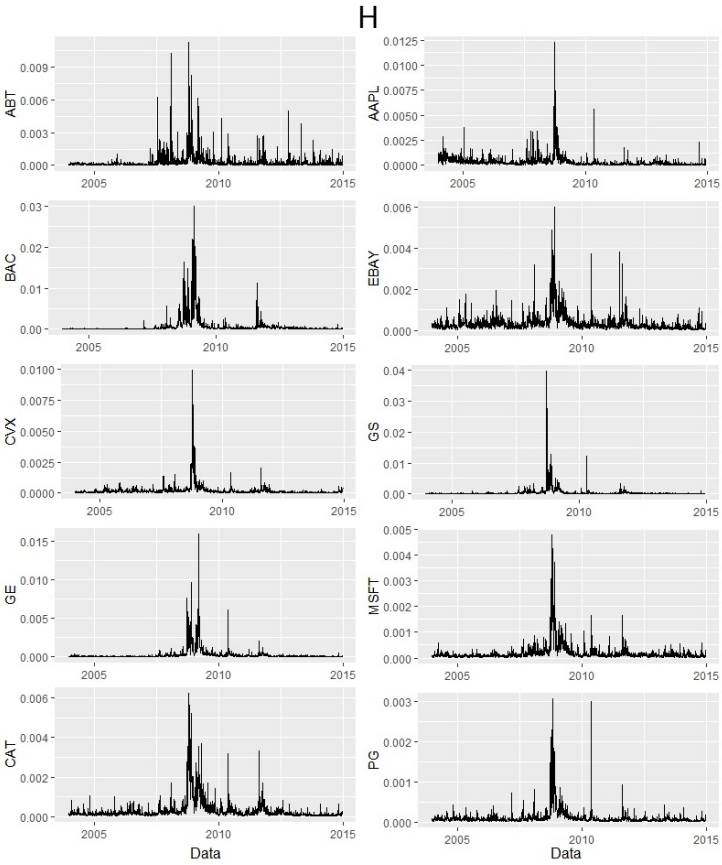


Figura 18 – Volatilidade Realizada
Evolução histórica da volatilidade realizada. A série inicia em 17/12/2003 e termina em 16/12/2014.

na Seção 3.2 foram estimados sequencialmente, com o conjunto de estimação sendo modificado por meio de uma janela móvel, sempre mantendo 2213 observações para realizar a estimação dos 554 consecutivos modelos. Para cada modelo estimado, foram salvas as previsões de um e cinco dias à frente. Cada um dos modelos se referirá às respectivas seções, previamente apresentadas: árvores (3.2.2.3), *bagging* (3.2.2.4), *cubist* (3.2.2.5), LASSO (3.2.2.1), ADALASSO (3.2.2.1), *Bayes-Exp* (3.2.3.2), *Bayes-HAR* (3.2.3.3), DMA95/DMA80/DMA50 (3.2.4.2), sendo os números indicativos da taxa de esquecimento de cada uma das combinações e Reg+/Reg+1 (3.2.4.1) sendo a metodologia de regressão com a restrição de que todos os parâmetros sejam positivos e que sejam positivos e somem um.

Toda a simulação foi feita na linguagem R Core Team (2013), sendo utilizados para a estimação dos modelos de aprendizado de máquina os pacotes referenciados logo ao lado do nome do modelo: LASSO (FRIEDMAN; HASTIE; TIBSHIRANI, 2010), ADALASSO (KRAEMER; SCHAEFER; BOULESTEIX, 2009), Árvores (*Model-Trees*) e Bagging (HORNIK; BUCHTA; ZEILEIS, 2009) e Cubist (KUHN; QUINLAN, 2018); todos os demais modelos e técnicas de combinação não utilizaram pacotes. Também agradecemos à Bernardi (2017) por

desenvolverem um pacote para a implementação do teste MCS.

3.3.3 Simulação

As Tabelas 4, 5, 6 e 7 apresentam o resultado da simulação. Nas linhas temos os modelos testados e nas colunas os ativos avaliados. Cada número na tabela apresenta o valor da função perda. Nas Tabelas 4, 5 temos os valores para previsões realizadas um dia à frente, e nas Tabelas 6 e 7 temos os valores para previsões realizadas 5 dias à frente. Quando um modelo pertencer ao conjunto de modelos de confiança de um ativo, o valor da função perda aparecerá **com** negrito, caso contrário aparecerá **sem** negrito. Caso o modelo apresente desempenho significativamente melhor que o HAR, seguindo o teste de habilidada preditiva condicional, será apresentado o sinal positivo ($+$) ao lado do valor da função perda, caso ele apresente desempenho significativamente pior aparecerá o sinal negativo (-), caso o modelo não pareça ter performance diferente do HAR não será apresentado nenhum símbolo ao lado da função perda.

A tabela 4 apresenta o erro, avaliado pela função perda MSE, das previsões de volatilidade com horizonte de um dia à frente. Em comparação ao modelo base - o HAR - somente o LASSO conseguiu realizar previsões

significativamente mais acuradas. As duas metodologias de combinação baseadas em regressão Reg+/Reg+1 atingiram a pior performance, sendo piores do que o HAR em 7 ativos. HAR/LASSO/Bayes-Exp/DMA50/DMA80 estiveram contidos dentro do MCS em todos os ativos.

A Tabela 5 apresenta o erro, avaliado pela função perda QLIKE, das previsões de volatilidade com horizonte de um dia à frente. Alguns modelos apresentam desempenho significativamente melhor do que o HAR: LASSO parece ser melhor do que o HAR em 9 ativos e Bayes-Exp/Reg+/Reg+1/DMA50/DMA80 parecem ter um desempenho melhor em 3 ativos. Árvores e *cubist* parecem não ser melhores em nenhum momento, chegando, eventualmente, a ser piores do que o HAR. Apenas dois modelos estiveram sempre dentro do MCS: o LASSO e Reg+1.

As Tabelas 6 e 7 apresentam uma simulação similar, com a única alteração sendo o horizonte de previsão, passando de um para cinco dias à frente. A tabela 6 apresenta a performance dos modelos avaliada pela MSE. LASSO parece ter melhor desempenho do que o HAR em cinco ativos, enquanto ADALASSO e DMA80 parecem ter melhor performance em 3 ativos. As combinações de regressão parecem ter desempenho pior do que o HAR em 6 ativos. LASSO/ADALASSO/*Bagging*/DMA50 estiveram

	MSE									
	AAPL	EBAY	GS	MSFT	PG	ABT	BAC	CVX	GE	CAT
HAR	**2,172**	**1,166**	**0,460**	**0,455**	**0,128**	**14,367**	**0,980**	**0,191**	**0,186**	**0,449**
LASSO	**2,162**	**1,142**	**0,452**	**0,444(+)**	**0,127**	**14,474**	**0,969**	**0,193**	**0,184**	**0,443**
ADALASSO	**2,171**	**1,165**	**0,461**	**0,447**	**0,127**	**14,408**	1,058(-)	**0,188**	**0,183**	**0,449**
Árvores	**2,180**	**1,152**	**0,456**	**0,461**	0,128	**14,379**	**0,980**	**0,189**	**0,186**	**0,454**
Bagging	**2,192**	**1,151**	**0,461**	**0,456**	**0,128**	**14,425**	**0,978**	**0,187**	0,189	**0,449**
Cubist	2,193	**1,151**	**0,459**	**0,463**	0,129	14,53(-)	**0,983**	**0,190**	0,188	**0,456**
Bayes HAR	**2,176**	**1,164**	**0,456**	**0,456**	0,129	**14,360**	**0,980**	**0,191**	**0,186**	**0,450**
Bayes Exp	**2,180**	**1,160**	**0,457**	**0,450**	**0,128**	**14,354**	**0,988**	**0,190**	**0,186**	**0,449**
Reg+	2,221	1,189	0,762(-)	**0,461**	0,132	**14,233**	1,065(-)	**0,192**	0,205(-)	0,456
Reg+1	2,320	1,264	1,038(-)	**0,466**	0,14(-)	**14,071**	2,307(-)	0,247(-)	0,388(-)	0,570(-)
DMA50	**2,175**	**1,153**	**0,454**	**0,451**	**0,127**	**14,410**	**0,979**	**0,189**	**0,185**	**0,446**
DMA80	**2,180**	**1,153**	**0,453**	**0,450**	**0,127**	**14,443**	**0,979**	**0,189**	**0,185**	**0,446**
DMA95	2,187	**1,156**	**0,457**	**0,454**	**0,127**	**14,434**	**0,986**	**0,190**	**0,187**	**0,450**

Tabela 4 – Previsões um dia a frente

O nível de significância do teste para habilidade preditiva condicional (CPA) é de 10%; para o conjunto de modelos de confiança (MCS) é 15%. Quando um modelo **pertencer** ao MCS, as medidas de erro serão reportadas em **negrito**. Quando um modelo tiver desempenho significativamente melhor/pior do que o HAR no CPA será colocado um sinal (+)/(-) ao lado da estimativa do erro.

	QLIKE									
	AAPL	EBAY	GS	MSFT	PG	ABT	BAC	CVX	GE	CAT
HAR	24,76	18,85	11,67	15,13	13,13	57,06	14,75	**11,62**	**11,07**	**14,18**
LASSO	**23,01(+)**	**17,15**	**10,93(+)**	**14,31(+)**	**12,60(+)**	**54,02(+)**	**14,00(+)**	**11,38**	**10,67(+)**	**13,41(+)**
ADALASSO	24,17	**18,013**	11,72	14,68(+)	**12,89**	55,93	15,64(-)	**11,47**	**10,96**	**13,86(+)**
Árvores	24,43	**18,61**	**11,26**	15,25	13,2	57,04	14,79	11,81	**11,13**	**14,55(-)**
Bagging	24,78	**18,41**	**11,11(+)**	15,04	13,06	58,51	14,65	**11,57**	11,21	**14,17**
Cubist	25,23	**19,06**	11,64	15,57	13,27	60,56(-)	15,26(-)	11,79	11,39(-)	**14,98(-)**
Bayes HAR	24,87(-)	19,08	11,49(+)	15,23	13,27(-)	57,10	14,79	11,73(-)	**11,09**	**14,28(-)**
Bayes Exp	24,84	18,64	**11,16(+)**	14,83(+)	**13,01**	56,69	14,74(+)	**11,61**	11,16	**14,00**
Reg+	**22,9**	**18,63**	**11,07(+)**	14,52(+)	**12,69**	**53,87**	**13,99(+)**	**11,45**	**10,91**	13,82
Reg+1	**22,45**	**17,11**	**11,43**	**14,24(+)**	**12,55**	**51,23(+)**	**13,74(+)**	**11,56**	**10,83**	**13,77**
DMA50	24,44	**18,40**	**11,23(+)**	14,86(+)	**12,88(+)**	56,93	14,67	**11,51**	**11,049**	**14,061**
DMA80	24,54	**18,54**	**11,21(+)**	14,82(+)	**12,89(+)**	57,47	14,70	**11,58**	**11,05**	**14,18**
DMA95	24,74	**18,72**	**11,26(+)**	14,95(+)	**12,98**	58,29	14,93	11,73	**11,11**	**14,40**

Tabela 5 – Previsões um dia a frente

O nível de significância do teste para habilidade preditiva condicional (CPA) é de 10%; para o conjunto de modelos de confiança (MCS) é 15%. Quando um modelo **pertencer** ao MCS, as medidas de erro serão reportadas em **negrito**. Quando um modelo tiver desempenho significativamente melhor/pior do que o HAR no CPA será colocado um sinal (+)/(-) ao lado da estimativa do erro.

dentro do MCS em todas as oportunidades.

A tabela 7 apresenta a performance dos modelos quando avaliada pela QLIKE. Novamente LASSO tem melhor desempenho do que o HAR em todos os ativos, mas, Reg+1 e DMA50, não são melhores em apenas um ativo. O único modelo que não parece ser melhor do que o HAR em nenhum momento é o *cubist*, pior do que o HAR em 7 ativos.

Em resumo, o único modelo que parece ter performance superior ou igual ao HAR, e sempre ficar dentro do MCS, é o LASSO. Outros modelos que nunca tem desempenho pior do que o HAR são DMA50/DMA80.

3.3.4 LASSO

Há, apenas, um modelo que não parece ser inferior ao HAR e, ao mesmo tempo, está quase sempre dentro do MCS. Nas tabelas 8 e 9, estão apresentadas as estimativas do LASSO para a primeira estimação da janela móvel.

Na tabela 8, tem-se as estimativas de LASSO para modelos que preveem um passo à frente. Pode-se ver que, em todos os ativos, as primeiras duas defasagens não são nulas. A maioria dos ativos seleciona algumas defasagens entre a terceira e sétima ordem e há uma forte concentração ao redor da defasagem dez, seguida de um "grande buraco" até que se chegue à defasagem 22. Ao se

	MSE									
	AAPL	EBAY	GS	MSFT	PG	ABT	BAC	CVX	GE	CAT
HAR	**2,465**	1,262	**0,614**	**0,573**	**0,156**	15,446	**1,166**	**0,305**	**0,270**	**0,511**
LASSO	**2,431**	**1,25(+)**	**0,602**	**0,552(+)**	**0,152(+)**	**15,205(+)**	**1,14(+)**	**0,305**	**0,263**	**0,511**
ADALASSO	**2,443(+)**	**1,24(+)**	**0,607**	**0,568(+)**	**0,154**	**15,298**	**1,164**	**0,308**	**0,267**	**0,516**
Árvores	**2,485**	**1,237**	**0,598**	0,593	**0,158**	15,455	**1,151**	**0,311**	**0,290**	**0,513**
Bagging	**2,462**	**1,247**	**0,605**	**0,570**	**0,152**	**15,379**	**1,182**	**0,314**	**0,301**	**0,521**
Cubist	**2,498**	**1,238**	**0,602**	**0,569**	**0,156**	15,469	**1,183(-)**	**0,309**	**0,290**	**0,518**
Bayes HAR	2,48(-)	**1,255(+)**	**0,609**	**0,578(-)**	**0,158(-)**	15,391	**1,162**	**0,306**	**0,270**	**0,511**
Bayes Exp	**2,462**	**1,246**	**0,606**	**0,564**	**0,156**	15,405	**1,141**	**0,304**	**0,272**	**0,508**
Reg+	2,915(-)	1,423(-)	1,187(-)	**0,588**	0,171	15,95(-)	1,533(-)	**0,340**	0,367(-)	0,584
Reg+1	2,798(-)	1,445(-)	1,667(-)	**0,549**	**0,168**	**14,881**	3,907(-)	**0,361**	0,579(-)	0,725(-)
DMA50	**2,455**	**1,237(+)**	**0,596(+)**	**0,559**	**0,154**	**15,359**	**1,148**	**0,302**	**0,271**	**0,509**
DMA80	**2,463**	**1,238(+)**	**0,597(+)**	**0,559**	**0,154**	15,388	**1,146(+)**	**0,301**	**0,268**	**0,509**
DMA95	**2,475**	**1,249**	**0,605**	**0,568**	**0,156**	15,420	**1,148**	**0,302**	**0,269**	**0,509**

Tabela 6 – Previsões cinco dias a frente

O nível de significância do teste para habilidade preditiva condicional (CPA) é de 10%; para o conjunto de modelos de confiança (MCS) é 15%. Quando um modelo **pertencer** ao MCS, as medidas de erro serão reportadas em **negrito**. Quando um modelo tiver desempenho significativamente melhor/pior do que o HAR no CPA será colocado um sinal (+)/(-) ao lado da estimativa do erro.

	QLIKE									
	AAPL	EBAY	GS	MSFT	PG	ABT	BAC	CVX	GE	CAT
HAR	**0,306**	0,204	**0,154**	**0,205**	**0,178**	**0,757**	**0,193**	**0,180**	**0,158**	**0,169**
LASSO	**0,284(+)**	**0,191(+)**	**0,148(+)**	**0,188(+)**	**0,167(+)**	**0,661(+)**	**0,179(+)**	**0,177(+)**	**0,148(+)**	**0,163(+)**
ADALASSO	**0,3(+)**	**0,197**	**0,152(+)**	**0,199**	**0,174(+)**	**0,711(+)**	**0,189(+)**	**0,178(+)**	**0,156**	**0,169(+)**
Árvores	**0,308**	**0,198**	**0,152(+)**	**0,203(+)**	0,180	**0,743(+)**	**0,190**	**0,185**	**0,162(-)**	**0,171**
Bagging	**0,316**	**0,200**	**0,152(+)**	**0,193(+)**	**0,171(+)**	**0,734**	**0,206(-)**	**0,172(+)**	**0,165(-)**	**0,171**
Cubist	**0,305**	**0,201**	**0,155(-)**	**0,197**	0,181(-)	0,792(-)	0,213(-)	**0,181(-)**	**0,166(-)**	**0,171(-)**
Bayes HAR	0,312(-)	0,205	**0,154(-)**	0,21(-)	0,184(-)	**0,751(+)**	**0,193**	0,183(-)	**0,158(-)**	**0,171(-)**
Bayes Exp	**0,306**	**0,2(+)**	**0,152(+)**	**0,2(+)**	**0,178(+)**	**0,746(+)**	**0,188(+)**	**0,18(-)**	**0,156**	**0,166(+)**
Reg+	0,358	0,290	**0,177**	**0,217(-)**	**0,188**	**0,808**	**0,186**	**0,193(-)**	**0,157(+)**	0,209(-)
Reg+1	**0,275**	**0,185(+)**	**0,124(+)**	**0,172(+)**	**0,147(+)**	**0,621(+)**	**0,176(+)**	**0,137(+)**	**0,135(+)**	**0,149(+)**
DMA50	**0,297**	**0,198(+)**	**0,149(+)**	**0,195(+)**	**0,173(+)**	**0,731(+)**	**0,188(+)**	**0,173(+)**	**0,155(+)**	**0,166(+)**
DMA80	**0,300**	**0,200**	**0,149(+)**	**0,196(+)**	**0,173(+)**	**0,741(+)**	**0,187(+)**	**0,173(+)**	**0,153(+)**	**0,167**
DMA95	**0,306**	**0,205**	**0,152**	**0,2(+)**	**0,177**	0,758	**0,19(+)**	**0,175(+)**	**0,156(+)**	**0,168(+)**

Tabela 7 – Previsões cinco dias a frente

O nível de significância do teste para habilidade preditiva condicional (CPA) é de 10%; para o conjunto de modelos de confiança (MCS) é 15%. Quando um modelo **pertencer** ao MCS, as medidas de erro serão reportadas em **negrito**. Quando um modelo tiver desempenho significativamente melhor/pior do que o HAR no CPA será colocado um sinal (+)/(-) ao lado da estimativa do erro.

olhar para as estimativas dos modelos que preveem cinco passos à frente, tabela 9, vê-se um resultado similar, porém com uma diferença: agora há uma maior predominância da cor preta.

Ao comparar as tabelas com a teoria que embasa o HAR, percebe-se que LASSO seleciona muito mais defasagens do que aquelas sugeridas pelo HAR. A segunda e vigésima segunda defasagens foram selecionadas quase sempre para serem diferentes de zero, algo não esperado pelo HAR. Já em previsões cinco dias a frente, a quinta defasagem indicada como importante pelo HAR foi selecionada apenas 2 vezes para ser diferente de zero, e com valores muito pequenos para as estimativas.

3.3.5 Ponderação de Modelos Dinâmica (DMA)

Um segundo bom competidor na simulação é a ponderação de modelos dinâmica (DMA), especialmente nas menores taxas de esquecimento, 0.80 e 0.50. A ponderação de modelos dinâmica atualiza os pesos atribuidos a cada modelo da combinação todos os períodos a posteriori. Conforme os modelos prevejam, é verificado o nível de acerto de cada um deles, e dado esse nível de acerto o peso a ser dado para cada modelo na próxima combinação é atualizado. A Figura 19 apresenta a evolução dos pesos ao longo da janela de teste.

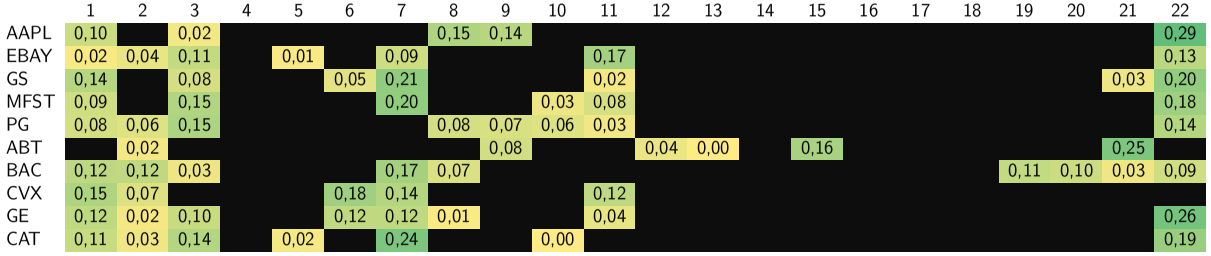
Tabela 8 – Estimativas do LASSO para a primeira rodada da janela de estimação - Previsões um dia a frente.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
AAPL	0,32	0,04		0,15						0,01	0,08	0,09										0,12
EBAY	0,21	0,15		0,06		0,01	0,13				0,03				0,15							
GS	0,32	0,15		0,04			0,08			0,10	0,01			0,02	0,03							0,05
MFST	0,22	0,03	0,05	0,12	0,13		0,02			0,04	0,12				0,01							0,07
PG	0,25	0,09	0,06		0,09	0,11						0,12	0,01	0,00	0,02				0,02	0,02		0,00
ABT	0,16	0,03	0,06		0,02	0,00		0,02					0,06			0,01	0,00		0,20			0,08
BAC	0,39	0,09	0,11			0,05					0,15											0,11
CVX	0,34	0,01			0,31					0,14												
GE	0,25	0,09	0,09	0,17						0,15					0,01						0,03	0,08
CAT	0,30	0,07	0,03		0,20		0,01				0,16											0,06

Estimativas LASSO

A tabela apresenta as estimativas dos parâmetros do Lasso na primeira janela de estimação da janela móvel. O hiperparâmetro de penalização foi encontrado por validação cruzada e foi respeitada a regra de um desvio-padrão. Caixas pretas indicam que um parâmetro é numericamente igual a zero, e as cores na escala verde-amarela ilustram a grandeza dos parâmtros: quanto mais verde, maior o parâmetro, e quanto mais amarelo, menor o parâmetro.

Tabela 9 – Estimativas do LASSO para a primeira rodada da janela de estimação - Previsões cinco dias a frente.



Estimativas LASSO

A tabela apresenta as estimativas dos parâmetros do Lasso na primeira janela de estimação da janela móvel. O hiperparâmetro de penalização foi encontrado por validação cruzada e foi respeitada a regra de um desvio-padrão. Caixas pretas indicam que um parâmetro é numericamente igual a zero, e as cores na escala verde-amarela ilustram a grandeza dos parâmetros: quanto mais verde, maior o parâmetro, e quanto mais amarelo, menor o parâmetro.

Ao comparar os três gráficos, pode-se ver que as combinações com menores taxas de esquecimento tendem a não concentrar os pesos em apenas um modelo. Efetivamente, quando se olha a evolução dos pesos no gráfico superior, pode-se ver que os pesos parecem ficar distribuídos de maneira mais igualitária, enquanto que, ao olhar o gráfico inferior, vê-se o cenário oposto. Os modelos que realizam boas previsões são beneficiados com um alto parâmetro para a previsão no próximo período, e continuam com essa elevada importância até que outro modelo tenha uma previsão que se destaque das demais.

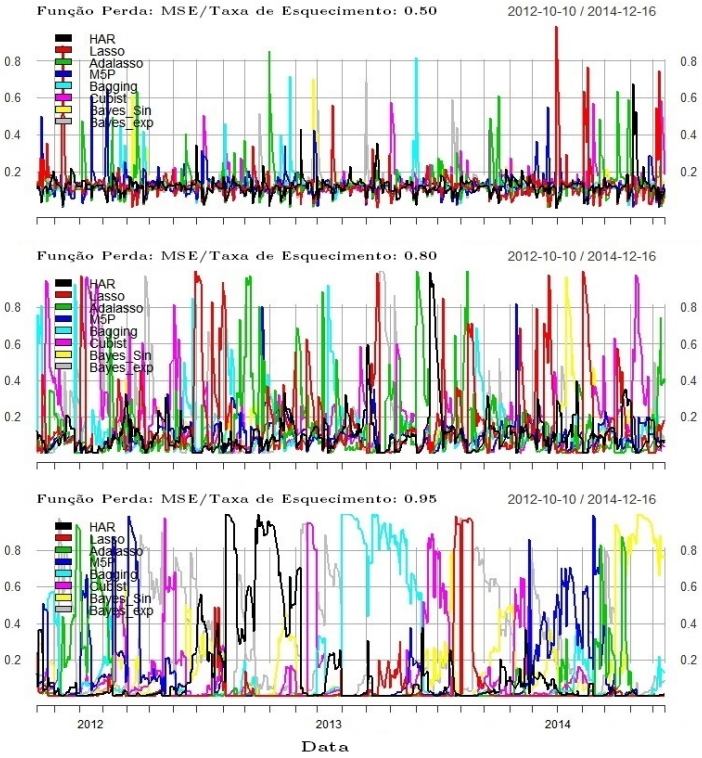


Figura 19 – Evolução dos Pesos na Poderação Dinâmica de Modelos

A ponderação de modelos dinâmica aloca pesos que variam no tempo para os diferentes modelos preditores. A variação desses pesos ocorre conforme eles vão realizando previsões fora da amostra, verifica-se o nível de acerto de cada modelo e atualiza-se o peso de cada modelo no próximo período. Para regularizar a velocidade de atualização pode ser escolhido um hiperparâmetro conhecido como taxa de esquecimento. A figura acima apresenta os pesos atribuídos a cada modelo quando temos uma taxa de esquecimento de 0.50, 0.80 e 0.95 respectivamente.

3.4 Conclusões do Capítulo

Desde o seu lançamento, o HAR se tornou um dos modelos mais presentes em trabalhos empirícos de previsão de volatilidade realizada. Sua capacidade de capturar a memória de longo prazo, em uma estrutura parcimoniosa, fez dele um grande sucesso. Neste trabalho, verificou-se que diferentes especificaçõess de modelos poderiam melhorar o seu desempenho preditivo.

Inicialmente, verificou-se se penalizações tipo LASSO ou ADALASSO poderiam gerar melhores previsões e, especialmente, a versão tradicional de LASSO parece ser capaz. O segundo conjunto de modelos é derivado de modelos de árvores. Aparentemente, a adição dessa estrutura não linear não melhora a qualidade preditiva. O terceiro conjunto de modelos era composto por duas regressões lineares Bayesianas com *prioris* informativas, mas, novamente, esse grupo de modelos não apresentou melhor performance preditiva em comparação ao HAR.

O único modelo que individualmente apresentou desempenho preditivo superior ao HAR foi o LASSO. Verificou-se que esse modelo seleciona mais médias de defasagens, além daquelas típicas do HAR, 1,5 e 22. Verifica-se que há um grande número de médias selecionadas especialmente das primeiras defasagens como

a média dos últimos dois, três, quatro dias. Quanto as médias de um maior número de defasagens, como 10, 15 ou 20, verifica-se que o LASSO seleciona menos dessas médias maiores para a estimação do modelo em comparação ao primeiro grupo que vai das defasagens de 1 à 5, mas sempre seleciona algumas dessa médias maiores.

Finalmente, combinaram-se todos os modelos apresentados e, de fato, as combinações utilizando o método da regressão não conseguiram demonstrar desempenho consistentemente superior, mas as combinações geradas pela ponderação de modelos dinâmica produzem previsões que nunca são piores do que o HAR, e recorrentemente são melhores. Verificou-se que combinações que alocam pesos mais igualitários para os modelos combinados geram previsões melhores do que aqueles que se concentram apenas no modelo que teve a recente melhor capacidade preditiva.

4 CONCLUSÕES

Este trabalho tinha dois principais objetivos. Primeiramente, utilizar conjuntos de árvores para aprender as fontes de informação, dentro da dinâmica de negociação de ativos financeiros, que geram previsibilidade sobre a futura direção do preço dos ativos. Na sequência, buscou-se verificar se diferentes metodologias preditivas conseguem superar um modelo tradicional de previsão de volatilidade, bem como verificar quais as características daqueles que conseguem.

Quanto ao primeiro objetivo, utilizou-se os algoritmos *Boosting Trees* e *Random Forests* para estimar a função que conecta o conjunto de dados fornecidos à direção futura do preço do ativo. Após a estimação, verificou-se quais foram as variáveis mais relevantes para a estimação da função.

Verificou-se, também, que a maior previsibilidade do sinal futuro ocorre na menor frequência possível. Neste trabalho, a menor frequência avaliada foi a de 1 minuto. Conclui-se que a variável mais importante para a futura

previsibilidade do sinal dos retornos um minuto a frente é o retorno defasado em um período. Além disso, outra variável que se destacou foi o horário do negócio, demonstrando o impacto que a sazonalidade tem na capacidade preditiva.

Quanto aos grupos de variáveis que mais importam, verificou-se algo esperado: o maior conjunto de informação relevante para a previsibilidade futura vêm do período imediatamente anterior, mas, a defasagem superior não é insignificante, sendo especialmente valorizada pelo algoritmo *Random Forests*. Quanto à origem dessa informação, alguns sinais conflituosos: enquanto o algoritmo *Boosting Trees* aponta que as informações advindas do registro de negócio são as mais importantes, o outro algoritmo aponta que as variáveis vindas do livro de ofertas são as mais relevantes.

Quando se avaliou os modelos estimados para prever retornos 60 minutos à frente, o sinal foi ainda mais nebuloso. Apesar de ambos os modelos apontarem que o último preço era a variável mais relevante na estimação, a diferença da importância dessa variável frente as demais foi menor. Quando se investigou quais as características dos conjuntos de variáveis mais importantes, as estimativas sempre ficaram muito próximas de um empate. Existe pouca diferença tanto quando se avalia a defasagem mais importante ou de que grupo essas variáveis vêm.

No capítulo seguinte, buscou-se verificar se três grupos de modelos - modelos simples e de conjuntos de árvores de modelos, penalizações de regressões lineares e regressões bayesianas com *prioris* informativas - eram capazes de gerar capacidade preditiva superior ao principal modelo de previsão de volatilidade realizada, o HAR.

Verificou-se que o único modelo que consegue, consistentemente, bater as previsões do HAR ao prever a volatilidade realizada 1 e 5 minutos a frente foi a regressão linear com termo de penalização Lasso. Esse modelo nunca tem desempenho inferior, e recorrentemente, tem desempenho superior ao HAR. Verificou-se que a estrutura de penalização anula algumas restrições de igualdade que são impostas pelo HAR, permitindo maior flexibilidade na estimação.

Outra estrutura que se mostrou interessante foi a combinação de todos os modelos propostos através da metologia *Dynamic Model Averaging*. Essa metodologia se sobressaiu a metodologia de combinação baseada em regressão linear. Verificou-se que os modelos com maior capacidade preditiva alocavam uma proporção igualitária entre os modelos disponíveis, aproveitando-se, apenas, em curtos espaços de tempo de desempenhos extraordinários de um modelo específico. A não ênfase em um único modelo, aliada a capacidade de aproveitar curtos períodos

de bom ajuste, permitiu que essa estrutura também fosse competitiva em relação ao modelo *benchmarking*.

Assim, verificou-se que modelos de aprendizado de máquina podem ajudar os economistas em duas das tarefas mais comuns em seu trabalho. Inicialmente, modelos de aprendizado de máquina podem ser utilizados para gerar novos insights sobre as variáveis relevantes para a modelagem e, na sequência, conclui-se que técnicas de penalização de regressões lineares e modelos de combinação de previsões podem gerar previsões mais acuradas do que os métodos econométricos tradicionais.

Como sugestão de trabalhos futuros apontamos alguns caminhos: analisar a diversidade e acurácia dos modelos antes de combiná-los, realizar análise de correlação cruzada das entradas com as saídas dos modelos, analisar o impacto que os hiperparâmetros não otimizados têm na performance dos modelos, comparar o comportamento preditivo saindo de um passo a frente para n passos, testar outras metodologias de modelos de aprendizado de máquina e testar métodos de regressão robustos.

REFERÊNCIAS

ANDERSEN, T. G. et al. Exchange rate returns standardized by realized volatility are (nearly) gaussian. **Multinational Finance Journal**, v. 4, n. 3&4, p. 159–179, 2000.

ANDERSEN, T. G. et al. The distribution of realized stock return volatility. **Journal of Financial Economics**, Elsevier, v. 61, n. 1, p. 43–76, 2001.

ANDERSEN, T. G.; BOLLERSLEV, T.; MEDDAHI, N. Correcting the errors: Volatility forecast evaluation using high-frequency data and realized volatilities. **Econometrica**, Wiley Online Library, v. 73, n. 1, p. 279–296, 2005.

ANDERSEN, T. G.; BOLLERSLEV, T.; MEDDAHI, N. Realized volatility forecasting and market microstructure noise. **Journal of Econometrics**, Elsevier, v. 160, n. 1, p. 220–234, 2011.

AUDRINO, F.; CAMPONOVO, L. Oracle properties and finite sample inference of the adaptive lasso for time series regression models. **arXiv preprint arXiv:1312.1473**, 2013.

AUDRINO, F.; CAMPONOVO, L.; ROTH, C. Testing the lag structure of assets' realized volatility dynamics. 2015.

AUDRINO, F.; CORSI, F. Modeling tick-by-tick realized correlations. **Computational Statistics & Data Analysis**, Elsevier, v. 54, n. 11, p. 2372–2382, 2010.

AUDRINO, F.; HUANG, C.; OKHRIN, O. Flexible har model for realized volatility. University of St. Gallen, 2016.

AUDRINO, F.; KNAUS, S. D. Lassoing the har model: A model selection perspective on realized volatility dynamics. **Econometric Reviews**, Taylor & Francis, v. 35, n. 8-10, p. 1485–1521, 2016.

BAILLIE, R. T.; BOLLERSLEV, T.; MIKKELSEN, H. O. Fractionally integrated generalized autoregressive conditional heteroskedasticity. **Journal of Econometrics**, Elsevier, v. 74, n. 1, p. 3–30, 1996.

BENABOU, R.; LAROQUE, G. Using privileged information to manipulate markets: Insiders, gurus, and credibility. **Quarterly Journal of Economics**, MIT Press, v. 107, n. 3, p. 921–958, 1992.

BERNARDI, L. C. . M. **MCS: Model Confidence Set Procedure**. [S.l.], 2017. R package version 0.1.3. Disponível em: <https://CRAN.R-project.org/package=MCS>.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. **Journal of Econometrics**, Elsevier, v. 31, n. 3, p. 307–327, 1986.

BOUDT, K.; CORNELISSEN, J.; PAYSEUR, S. **highfrequency: Tools for Highfrequency Data Analysis**. [S.l.], 2018. R package version 0.5.3. Disponível em: <https://CRAN.R-project.org/package=highfrequency>.

BREIMAN, F. Olshen, and stone. **Classification and Regression trees**, 1984.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. **Classification and regression trees**. [S.l.]: CRC press, 1984.

BUCCI, A. Forecasting realized volatility: a review. 2017.

BÜHLMANN, P.; VAN DE GEER, S. **Statistics for high-dimensional data: methods, theory and applications**. [S.l.]: Springer Science & Business Media, 2011.

CALDEIRA, J. F.; MOURA, G. V.; SANTOS, A. A. Yield curve forecast combinations based on bond portfolio performance. **Journal of Forecasting**, Wiley Online Library, v. 37, n. 1, p. 64–82, 2018.

CAMPBELL, J. Y.; GROSSMAN, S. J.; WANG, J. Trading volume and serial correlation in stock returns. **Quarterly Journal of Economics**, MIT Press, v. 108, n. 4, p. 905–939, 1993.

CHAKRAVARTY, S.; MCCONNELL, J. J. Does insider trading really move stock prices? **Journal of Financial**

and Quantitative Analysis, Cambridge University Press, v. 34, n. 2, p. 191–209, 1999.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. p. 785–794, 2016.

CHEN, T. et al. **xgboost: Extreme Gradient Boosting**. [S.l.], 2018. R package version 0.71.2. Disponível em: <https://CRAN.R-project.org/package=xgboost>.

CHRISTOFFERSEN, P. et al. Direction-of-change forecasts based on conditional variance, skewness and kurtosis dynamics: international evidence. 2006.

CHRISTOFFERSEN, P. F.; DIEBOLD, F. X. Financial asset returns, direction-of-change forecasting, and volatility dynamics. **Management Science**, INFORMS, v. 52, n. 8, p. 1273–1287, 2006.

CONRAD, J. S.; HAMEED, A.; NIDEN, C. Volume and autocovariances in short-horizon individual security returns. **Journal of Finance**, Wiley Online Library, v. 49, n. 4, p. 1305–1329, 1994.

COOPER, J. P.; NELSON, C. R. The ex ante prediction performance of the st. louis and frb-mit-penn econometric models and some results on composite predictors. **Journal of Money, Credit and Banking**, JSTOR, v. 7, n. 1, p. 1–32, 1975.

CORSI, F. A simple approximate long-memory model of realized volatility. **Journal of Financial Econometrics**, Oxford University Press, v. 7, n. 2, p. 174–196, 2009.

CRAIOVEANU, M.; HILLEBRAND, E. et al. Why it is ok to use the har-rv (1, 5, 21) model. **Unpublished manuscript**, 2010.

CRAMÉR, H. **Mathematical Methods of Statistics (PMS-9)**. [S.l.]: Princeton university press, 2016. v. 9.

DACOROGNA, M. et al. Modelling short-term volatility with garch and harch models. 1998.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. [S.l.]: Cambridge university press, 1997. v. 1.

DIEBOLD, F. X. Serial correlation and the combination of forecasts. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 6, n. 1, p. 105–111, 1988.

DIEBOLD, F. X. Forecasting. Department of Economics, University of Pennsylvania, 2017.

DIEBOLD, F. X.; MARIANO, R. S. Comparing predictive accuracy. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 20, n. 1, p. 134–144, 2002.

DIXON, M. Sequence classification of the limit order book using recurrent neural networks. **Journal of Computational Science**, Elsevier, 2017.

DIXON, M. F.; KLABJAN, D.; BANG, J. H. Classification-based financial markets prediction using deep neural networks. 2016.

EASLEY, D. et al. Liquidity, information, and infrequently traded stocks. **Journal of Finance**, Wiley Online Library, v. 51, n. 4, p. 1405–1436, 1996.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. **Econometrica**, JSTOR, p. 987–1007, 1982.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **Journal of Finance**, JSTOR, v. 25, n. 2, p. 383–417, 1970.

FERREIRA, A. J.; FIGUEIREDO, M. A. Boosting algorithms: A review of methods, theory, and applications. In: **Ensemble Machine Learning**. [S.l.]: Springer, 2012. p. 35–85.

FLETCHER, T.; HUSSAIN, Z.; SHAWE-TAYLOR, J. Multiple kernel learning on the limit order book. In: **Proceedings of the First Workshop on Applications of Pattern Analysis**. [S.l.: s.n.], 2010. p. 167–174.

FLETCHER, T.; SHAWE-TAYLOR, J. Multiple kernel learning with fisher kernels for high frequency currency prediction. **Computational Economics**, Springer, v. 42, n. 2, p. 217–240, 2013.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. v. 1.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical**

Software, v. 33, n. 1, p. 1–22, 2010. Disponível em: <http://www.jstatsoft.org/v33/i01/>.

GARY, K. Bayesian econometrics. **J Wiley and Sons. Sussex, England**, 2003.

GIACOMINI, R.; WHITE, H. Tests of conditional forecast accuracy. **Econometrica**, v. 74, p. 1545–1578, 2006.

GONÇALVES, S.; MEDDAHI, N. Box–cox transforms for realized volatility. **Journal of Econometrics**, Elsevier, v. 160, n. 1, p. 129–144, 2011.

GRANGER, C. W. Long memory relationships and the aggregation of dynamic models. **Journal of Econometrics**, Elsevier, v. 14, n. 2, p. 227–238, 1980.

GRANGER, C. W.; JOYEUX, R. An introduction to long-memory time series models and fractional differencing. **Journal of Time Series Analysis**, Wiley Online Library, v. 1, n. 1, p. 15–29, 1980.

GRANGER, C. W.; RAMANATHAN, R. Improved methods of combining forecasts. **Journal of Forecasting**, Wiley Online Library, v. 3, n. 2, p. 197–204, 1984.

GROSSMAN, S. et al. The informational role of prices. **MIT Press Books**, the MIT Press, v. 1, 1989.

HAN, J. et al. Machine learning techniques for price change forecast using the limit order book data. **Machine learning**, 2015.

HANSEN, P. R.; LUNDE, A. Consistent ranking of volatility models. **Journal of Econometrics**, Elsevier, v. 131, n. 1-2, p. 97–121, 2006.

HANSEN, P. R.; LUNDE, A.; NASON, J. M. The model confidence set. **Econometrica**, Wiley Online Library, v. 79, n. 2, p. 453–497, 2011.

HE, H.; WANG, J. Differential information and dynamic behavior of stock trading volume. **The Review of Financial Studies**, Oxford University Press, v. 8, n. 4, p. 919–972, 1995.

HORNIK, K.; BUCHTA, C.; ZEILEIS, A. Open-source machine learning: R meets Weka. **Computational Statistics**, v. 24, n. 2, p. 225–232, 2009.

KIM, O.; VERRECCHIA, R. E. Trading volume and price reactions to public announcements. **Journal of accounting research**, JSTOR, p. 302–321, 1991.

KOOP, G. **Bayesian Econometrics. 2003**. [S.l.]: John Wiley, Chichester, 2003.

KRAEMER, N.; SCHAEFER, J.; BOULESTEIX, A.-L. Regularized estimation of large-scale gene regulatory networks using gaussian graphical models. **BMC Bioinformatics**, v. 10, n. 384, 2009.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. [S.l.]: Springer, 2013. v. 26.

KUHN, M.; QUINLAN, R. **Cubist: Rule- And Instance-Based Regression Modeling**. [S.l.],

2018. R package version 0.2.2. Disponível em: <https://CRAN.R-project.org/package=Cubist>.

KUMAR, D. P.; AMGOTH, T.; ANNAVARAPU, C. S. R. Machine learning algorithms for wireless sensor networks: A survey. **Information Fusion**, Elsevier, v. 49, p. 1–25, 2019.

KYLE, A. S. Informed speculation with imperfect competition. **Review of Economic Studies**, Wiley-Blackwell, v. 56, n. 3, p. 317–355, 1989.

LEBARON, B. et al. Stochastic volatility as a simple generator of apparent financial power laws and long memory. **Quantitative Finance**, Taylor & Francis, v. 1, n. 6, p. 621–631, 2001.

LEDELL, E. et al. **h2o: R Interface for 'H2O'**. [S.l.], 2019. R package version 3.22.1.1. Disponível em: <https://CRAN.R-project.org/package=h2o>.

LEHMANN, E. L.; CASELLA, G. **Theory of point estimation**. [S.l.]: Springer Science & Business Media, 2006.

LIU, L. Y.; PATTON, A. J.; SHEPPARD, K. Does anything beat 5-minute rv? a comparison of realized measures across multiple asset classes. **Journal of Econometrics**, Elsevier, v. 187, n. 1, p. 293–311, 2015.

MALKIEL, B. G.; FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **Journal of Finance**, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970.

MARKOWITZ, H. Portfolio selection. **Journal of finance**, Wiley Online Library, v. 7, n. 1, p. 77–91, 1952.

MCCOY, J.; AURET, L. Machine learning applications in minerals processing: A review. **Minerals Engineering**, Elsevier, v. 132, p. 95–109, 2019.

MERTON, R. C. On estimating the expected return on the market: An exploratory investigation. **Journal of Financial Economics**, Elsevier, v. 8, n. 4, p. 323–361, 1980.

MEULBROEK, L. K. An empirical analysis of illegal insider trading. **Journal of Finance**, Wiley Online Library, v. 47, n. 5, p. 1661–1699, 1992.

MILGROM, P.; STOKEY, N. Information, trade and common knowledge. **Journal of Economic Theory**, Elsevier, v. 26, n. 1, p. 17–27, 1982.

MITCHELL, M. L.; NETTER, J. M. The role of financial economics in securities fraud cases: Applications at the securities and exchange commission. **The Business Lawyer**, JSTOR, p. 545–590, 1994.

MITCHELL, T. M. Machine learning. 1997. **Burr Ridge, IL: McGraw Hill**, v. 45, n. 37, p. 870–877, 1997.

MÜLLER, U. A. et al. Fractals and intrinsic time: A challenge to econometricians. **Unpublished manuscript, Olsen & Associates, Zürich**, 1993.

NELSON, C. R. The prediction performance of the frb-mit-penn model of the us economy. **The American**

Economic Review, JSTOR, v. 62, n. 5, p. 902–917, 1972.

PANICKER, S. S.; GAYATHRI, P. A survey of machine learning techniques in physiology based mental stress detection systems. **Biocybernetics and Biomedical Engineering**, Elsevier, 2019.

PATTON, A. J. Volatility forecast comparison using imperfect volatility proxies. **Journal of Econometrics**, Elsevier, v. 160, n. 1, p. 246–256, 2011.

PATTON, A. J.; SHEPPARD, K. Optimal combinations of realised volatility estimators. **International Journal of Forecasting**, Elsevier, v. 25, n. 2, p. 218–238, 2009.

PERLIN, M.; RAMOS, H. Gethfdata: Ar package for downloading and aggregating high frequency trading data from bovespa. 2016.

PRADO, M. L. de. **Advances in financial machine learning**. [S.l.]: John Wiley & Sons, 2018.

QUINLAN, J. R. C4. 5: Programming for machine learning. **Morgan Kauffmann**, v. 38, p. 48, 1993.

QUINLAN, J. R. et al. Learning with continuous classes. In: SINGAPORE. **5th Australian joint conference on artificial intelligence**. [S.l.], 1992. v. 92, p. 343–348.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>.

RAFTERY, A. E.; KÁRNỲ, M.; ETTLER, P. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. **Technometrics**, Taylor & Francis, v. 52, n. 1, p. 52–66, 2010.

ROSS, S. M. **Applied probability models with optimization applications**. [S.l.]: Courier Corporation, 2013.

SCHAPIRE, R. E. **The design and analysis of efficient learning algorithms**. [S.l.], 1991.

SCHAPIRE, R. E.; FREUND, Y. **Boosting: Foundations and Algorithms**. [S.l.]: The MIT Press, 2012. ISBN 0262017180, 9780262017183.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 267–288, 1996.

TSANTEKIDIS, A. et al. Forecasting stock prices from the limit order book using convolutional neural networks. In: IEEE. **Business Informatics (CBI), 2017 IEEE 19th Conference on**. [S.l.], 2017. v. 1, p. 7–12.

TSAY, R. S. **Analysis of financial time series**. [S.l.]: John Wiley & Sons, 2005. v. 543.

UCCI, D.; ANIELLO, L.; BALDONI, R. Survey of machine learning techniques for malware analysis. **Computers & Security**, Elsevier, 2018.

VIOLANTE, F.; LAURENT, S. Volatility forecasts evaluation and comparison. In: **Handbook of Volatility Models and Their Applications**. [S.l.]: John Wiley and Sons, 2012. p. 465–486.

WANG, Y.; WITTEN, I. H. Inducing model trees for continuous classes. In: **Proceedings of the Ninth European Conference on Machine Learning**. [S.l.: s.n.], 1997. p. 128–137.

WASSERSTEIN, R. L.; LAZAR, N. A. et al. The asa's statement on p-values: context, process, and purpose. **The American Statistician**, v. 70, n. 2, p. 129–133, 2016.

WITTEN, I. H. et al. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.

ZOU, H. The adaptive lasso and its oracle properties. **Journal of the American statistical association**, Taylor & Francis, v. 101, n. 476, p. 1418–1429, 2006.