UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
DEPARTAMENTO DE FINANÇAS E CONTABILIDADE
CURSO DE CIÊNCIAS CONTÁBEIS

GLAUCO GRACO NÓBREGA PORDEUS

SELEÇÃO DE CARTEIRAS POR MEIO DE *MACHINE LEARNING* E DA INFLUÊNCIA DA INFORMAÇÃO ASSIMÉTRICA

JOÃO PESSOA
2018

GLAUCO GRACO NÓBREGA PORDEUS

SELEÇÃO DE CARTEIRAS POR *MEIO DE MACHINE* LEARNING E DA INFLUÊNCIA DA INFORMAÇÃO ASSIMÉTRICA

Monografia apresentada ao Curso de Ciências Contábeis, do Centro de Ciências Sociais Aplicadas, da Universidade Federal da Paraíba, como requisito parcial a obtenção do grau de Bacharel em Ciências Contábeis.

Orientador: Prof. Dr. Luiz Felipe de Araújo Pontes Girão

JOÃO PESSOA

2018

GLAUCO GRACO NÓBREGA PORDEUS

SELEÇÃO DE CARTEIRAS POR MEIO DE *MACHINE LEARNING* E DA INFLUÊNCIA DA INFORMAÇÃO ASSIMÉTRICA

Esta monografia foi julgada adequada para a obtenção do grau de Bacharel em Ciências Contábeis, e aprovada em sua forma final pela Banca Examinadora designada pelo Departamento de Finanças e Contabilidade da Universidade Federal da Paraíba.

BANCA EXAMINADORA

Presidente Profº Drº Luiz Felipe de Araujo Pontes Girão (Orientador/a)

Instituição: UFPB

Membro: Profª Dr(a) Thais Gaudencio do Rego

Instituição: UFPB

Membro: Profº Me Werton Jose Cabral Rodrigues Filho

Instituição: UFPB

João Pessoa, 05 de novembro de 2018.

Catalogação na publicação
Seção de Catalogação e Classificação

P835s Pordeus, Glauco Graco Nobrega.
SELEÇÃO DE CARTEIRAS POR MEIO DE MACHINE LEARNING E DA INFLUÊNCIA DA INFORMAÇÃO ASSIMÉTRICA / Glauco Graco Nobrega Pordeus. - João Pessoa, 2018.
43 f. : il.

Orientação: Luiz Felipe de Araujo Pontes Girão.
Monografia (Graduação) - UFPB/CCSA.

1. insider trading. 2. aprendizado de máquina. 3. seleção de ativos. I. Girão, Luiz Felipe de Araujo Pontes. II. Título.

UFPB/BC

AGRADECIMENTOS

Agradeço principalmente a minha mãe, Adriana, por toda paciência e motivação que vem me proporcionando durante toda minha vida.

Aos meus professores Dr. Sinézio Maia, quem me abriu os olhos ao mundo do mercado de capitais e proporcionou um primeiro contato com alunos e professores da área por meio da Sala de ações e ao Dr. Felipe Girão, que vem sendo meu orientador e amigo durante estes últimos anos de graduação e com toda certeza caminharemos para os próximos anos de mestrado.

Aos meus amigos Profª Geisa Paulino e o Profº Filipe Duarte, pelos debates instigantes e a tantas idéias.

Agradeço também as pessoas do CA de Economia, que são um alívio nas tensões universitárias. Ao grupo da Liga de Mercado Financeiro da UFPB, aos companheiros fundadores e aos atuais membros, pelos debates e crescimento crítico que proporcionam.

Ademais, agradeço a todos os que dividiram estes últimos quatro anos de desenvolvimento, aos professores dos cursos de Ciências Contábeis, Economia e Administração e seus alunos.

“São demônios, os que destroem o poder bravio da humanidade”

Chico Science

RESUMO

Este estudo teve como principal motivação a criação de uma metodologia de seleção de ativos, tendo como base informações de microestrutura do mercado brasileiro adquiridas por meio da ferramenta desenvolvida por Perlin e Ramos (2016) e aplicadas ao modelo de *proxy* de assimetria informacional *Probability Informed Trading* desenvolvida por Easley, Hvidkjaer e O'Hara (2002; 2010) adaptada por Lin e Ke (2011). De posse das informações deste modelo, mais especificamente em seus *outputs*, a PIN e o Delta, foi desenvolvido um modelo de seleção de ativos, estabelecendo estas duas variáveis como *input* de um algoritmo de *machinelearning* desenvolvido por Chen e Guestrin (2016), denominado de *eXtremeGradientBoosting – XGBoost*, com objetivo de identificação das características dos ativos que possuem retorno anormal, ou seja acima do custo de capital captado pelo *Capital AssetPricingModel – CAPM*. Os resultados demonstraram que existe a possibilidade de obtenção de retorno anormal utilizando apenas a PIN e o Delta como parâmetro de seleção de ativos, encontrado 32,92% e 16,46% para as carteiras de 10 e 5 ativos, respectivamente. Além disso, em 57,62% das ocasiões o investidor obteria retorno positivo, selecionando os 10 ativos com maiores predições de retorno anormal, e em ao menos 52,38% dos casos foi encontrado um retorno acima do precificado pelo CAPM. Estes resultados se mostraram inferiores quando analisados apenas os 5 primeiros ativos, o que pode ser explicado pela diversificação do portfólio proposta por Markowitz (1952).

Palavras-chave:*insider trading; machinelearning;* seleção de ativos.

ABSTRACT

This study had as main motivation the creation of an asset selection methodology, based on information from the Brazilian market microstructure acquired through the tool developed by Perlin and Ramos (2016) and applied to the informational asymmetry proxy model Probability Informed Trading developed by Easley, Hvidkjaer and O'Hara (2002; 2010) adapted by Lin and Ke (2011). Based on information from this model, specifically on its outputs, PIN and Delta, an asset selection model was developed, establishing these two variables as input of a machine learning algorithm developed by Chen and Guestrin (2016), called of eXtreme Gradient Boosting - XGBoost, with the objective of identifying the characteristics of the assets that have an abnormal return, that is, above the cost of capital captured by the Capital Asset Pricing Model (CAPM). The results showed that there is a possibility of obtaining an abnormal return using only the PIN and Delta as the asset selection parameter, found 32.92% and 16.46% for the 10 and 5 asset portfolios, respectively. In addition, at 57.62% of the occasions the investor would obtain a positive return, selecting the 10 assets with the highest predictions of abnormal return, and at least 52.38% a return was found above the CAPM price. These results were lower when only the first five assets were analyzed, which can be explained by the diversification of the portfolio proposed by Markowitz (1952).

Key words:insider trading; machinelearning;portfolio selection.

LISTA DE ABREVIATURAS

B3	Bolsa, Brasil, Balcão.
BM&FBovespa	Bolsa de Valores, Mercadorias e Futuros
CAPM	*Capital AssetPricingModel*(Modelo de Precificação de Ativos deCapital)
IBOVESPA	Índice Bovespa
PIN	*ProbabilityofInformed Trading* (Probabilidade de negociação com informações privilegiadas)
Ra	Retorno Anormal
Ro	Retorno Observado
VPIN	*Volume-synchronizedProbabilityofInformed Trading* (Probabilidade de negociação com informações privilegiadas)
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

A obtenção de retorno anormal é um dos principais objetivos daqueles que atuam no mercado financeiro. Os agentes buscam, por meio da racionalidade, encontrar investimentos que paguem mais do que o mínimo exigido pelo custo do capital. Porém, tal façanha é de extrema dificuldade, pois segundo a Hipótese do Mercado Eficiente de Fama (1970), os ativos estão precificados de forma eficiente, cenário em que osatuantes no mercado possuem todas as informações necessárias para o estabelecimento de um preço coerente com a realidade de cada ativo.

Os dados são a base para que os agentes citados anteriormente possam decidir onde irão alocar seu capital, além disso, principalmente, o processamento destes dados para elaboração de modelos de decisão,faz com que o acesso à informação e componentes tecnológicos, seja cada vez mais necessário em diversos âmbitos da economia.

O MIT Technology Review Insights (2018), por meio de uma pesquisa feita com mais de 2.300 executivos, obteve o *feedback* para os principais pontos deste cenário tecnológico,tendo em vista o seu desenvolvimento: 84% dos entrevistados afirmam que o foco é o tempo em que os dados levam de recebidos, serem analisados e interpretados, e finalmente utilizados no processo decisório; 81% acreditam que a inteligência artificial terá um impacto positivo na indústria do futuro; e para este avanço os entrevistados colocam como principais barreiras são: custo e despesas (53% dos entrevistados), infraestrutura de dados (43% dos entrevistados) e recursos e talentos (41% dos entrevistados).

Nomercado de capitais,este assunto já é recorrente, segundo Aldridge (2010) o uso de *High Frequency Trading - HFT* (robôs de negociação) em *Wall Street* no ano de 2009 representava em torno de 60% do volume negociado, destes negociantes, segundo este levantamento feito pelo autor o Medallion Fund obteve um rendimento médio de 35%, entre 2000 e 2010. No Brasil, este setor vem tendo investimento por parte de agentes privados, como é o caso da iniciativa do BTG Pactual (2018) que está desenvolvendo a plataforma Stratsphera, atualmente em fase Beta, que tem como objetivo introduzir aos interessados no assunto ferramentas, dados e conteúdos preparatórios para novos programadores de estratégias de operação.

Desta maneira, o presente trabalho tem como objetivo principal o **desenvolvimento de uma metodologia para a seleção de ativos por meio de informações de microestrutura de mercado**, mais especificamente a*proxy*de assimetria*ProbabilityofInformedTrading*– PIN de Lin e Ke (2011), captada por meio das negociaçõesde compra e venda em alta frequência (todas negociações efetuadas) das ações listadas naBrasil, Bolsa e Balcão – B3, para o período

de outubro de 2015até dezembro de 2017.

Para atingir este objetivo, foi aplicado o modelo de *proxy* de assimetria informacional PIN desenvolvida por Easley, Hvidkjaer e O'Hara (2002; 2010) adaptada por Lin e Ke (2011).De posse das informações deste modelo, mais especificamente a PIN e o Delta, foi desenvolvido um modelo de seleção de ativos,estabelecendo estas duas variáveis como *input* de um algoritmo de *machinelearning* desenvolvido por Chen e Guestrin (2016), denominado de *eXtremeGradientBoosting – XGBoost,* com objetivo de identificação das características dos ativos que possuem retorno anormal, ou seja acima do custo de capital captado pelo *Capital AssetPricingModel – CAPM*.

Por fim, com base nas predições deste modelo, foram selecionados os grupos de5 e 10 ativos com maior projeção de retorno futuro. Já a análise dos resultados foi feita por meio da observação do desempenho financeiro das carteiras, comparando os desempenhos econômicos com o Ibovespa, o LTN com vencimento em 5 anos e indicadores de eficiência de risco e retorno, como os Índices de Sharpe e de Treynor.

2 REVISÃO DA LITERATURA

Este estudo é fundamentado basicamente na intercessão de três pontos amplamente abordados e em constate desenvolvimento na literatura financeira, quais sejam: Eficiência de Mercado, Assimetria Informacional e Inteligência Artificial. A seguir serão desenvolvidos os temas, elencado a convergência entre eles.

2.1 Eficiência de Mercado

A produção de conhecimento é farta no que diz respeito à eficiência de mercado. Em Callado (2009) são encontradas diversas referências no que diz respeito ao início da construção do pensamento de eficiência de mercado, onde são citados os trabalhos de Fama (1970); Leroy (1989); Campbell, Lo e MacKinlay (1997); e Ceretta (2001). Sendo exaltada a importância do estudo de Samuelson (1965), no qual foi introduzida a ideia de que os preços dos ativos têm movimentos aleatórios.

Na formulação da Hipótese do Mercado Eficiente – HME, Fama (1970) assume que a eficiência em sua forma forte, todas as informações, mesmo aquelas que ainda não são de domínio público, já se encontram precificadas; em sua forma semiforte, as informações são precificadas assim que são divulgadas ao público; por fim, em sua forma fraca, em que o autor coloca que as informações passadas ainda se mantêm absorvidas nos preços presentes e futuros. Segundo o autor, a função do mercado de capitais é disponibilizar um meio para que os recursos fluam entre os agentes superavitários e os agentes deficitários da economia. Seu funcionamento é baseado em três premissas chaves: (1) todos os agentes interessados em investir aceitam que os preços dos ativos sejam influenciados pelas *disclosures*; (2) as negociações são feitas sem custo de transação; e (3) todas as informações disponíveis são gratuitas para todos agentes do mercado.

Com base na ideia de eficiência de mercado e racionalidade do investidor, Treynor (1961), Sharpe (1964) e Lintner (1965) deram início ao pensamento de um modelo de precificação de ativos financeiros, do inglês *Capital AssetPricingModel*, ou CAPM.

O CAPM leva em conta três componentes básicos que determinam o custo de oportunidade para o agente investir em determinado ativo financeiro, sendo eles: (1) A taxa livre de risco representando o retorno mínimo garantido para o capital; (2) O coeficiente beta-mercado que representa a exposição do ativo financeiro ao risco sistemático do mercado; e (3) O retorno de mercado.

Algumas características do mercado brasileiro podem ser encontradas nos seguintes estudos: Machado (2009) analisou a existência de um prêmio pela liquidez (utilizando cinco

medidas diferentes) do mercado nacional, e se ela está precificada nos ativos nacionais. Os resultados deste estudo demonstraram a existência de prêmio de liquidez, independente da *proxy*utilizada, tendo uma variação de 0,04% a 0,77% ao mês; e Callado (2009) no qual investigou diferentes modelos e as formas de eficiência do mercado Brasileiro.

2.2 *Probability of Informed Trading* - PIN

A*proxy*de assimetria abordada neste estudo, a PIN, é um modelo de microestrutura de mercado desenvolvida por Easley*et al*. (2002), e reflete a assimetria a partir dos dados *intraday* das negociações dos ativos analisados. Sua mensuração é captada por meio do desequilíbrio entre as operações de compra e de venda dos agentes presentes no mercado. Este é um modelo que dispõe aos analistas uma medida sólida do grau de assimetria informacional para o período analisado (ABAD; RUBIA, 2005). No Brasil, Bosque (2016) desenvolveu uma abordagem bayesiana da PIN, com base no modelo aprimorado por Lin e Ke (2011), que, segundo Bosque (2016), esta abordagem possibilitaria introduzir opiniões dos analistas sobre os parâmetros da PIN.

Ainda no âmbito do uso da PIN no mercado brasileiro, Martins, Paulo e Girão (2014)encontraram que a PIN agrega informação na avaliação de empresas, ao analisarem o *valuerelevance* no modelo de Ohlson (1995) de assimetria de informação captada pelo o referido modelo de Easley*et al*. (2002). O estudo foi baseado em 198 ações da Bolsa de Valores, Mercadorias e Futuros de São Paulo (atual B3), durante o ano de 2011, e foi identificado que a PIN é mais expressiva na precificação de companhias que não se encontram em segmentos de governança corporativa reconhecidos.

Os mesmos autores (MARTINS; PAULO; GIRÃO, 2014) ao investigarem o possível caso de *insider* trading na OGX, encontraram que o preço da ação da companhia, entre os anos de 2008 e 2014, foi afetado positivamente pelas notícias no site da empresa e comunicados ao mercado, já quando relacionado aos fatos relevantes e formulários da ICVM 358, foi encontrada uma influência negativa. Com relação à PIN os autores ainda constataram que o mercado precificou o modelo de microestrutura até meados de 2011, o que para este período, entrou em consonância ao proposto por Demsetz (1986), em relação à possibilidade de os portadores de informações privilegiadas obterem retornos anormais.

Desta maneira, levando em conta o potencial e a relevância do modelo da PIN e da abordagem proposta por Lin e Ke (2011) – PIN LK, seu desenvolvimento teórico e operacional no Brasil, e sua relevância em estudos internacionais nos quais a PIN LK foi adotada como *benchmark* nos estudos mais recentes relacionadosà probabilidade de

negociação com informação privada (LIN; KE, 2011; YAN; ZHANG, 2012; YAN; ZHANG, 2014; GAN; WEI; JOHNSTONE, 2015; ERSAN; ALICI, 2016; BOSQUE, 2016) o que fundamenta a utilização da PIN LK (2011) como *proxy* para assimetria informacional neste trabalho.

2.3 *Machine Learning*

Já para a Inteligência Artificial, será empregado um modelo de Aprendizado de Máquina, ou*Machine Learning*, termointroduzido por Samuel (1959) quando, por meio do experimento de jogos de dama, verificou que um programa desenvolvido com apenas algumas regras que direcionaram seu aprendizado a programação capaz de evoluir a níveis melhores dos de quem oescreveu, isto em um período notavelmente curto, oito ou dez horas de aprendizagem. O autor ainda coloca que os princípios de aprendizado de máquina podemser reproduzidos e aplicados em diversas situações, que é o caso deste estudo.

Neste estudoo algoritmo*eXtremeGradientBoosting* (XGBoost) foi a ferramenta utilizada paraseleção das operações de compra. Este algoritmo foi desenvolvido por Chen eGuestrin (2016) e é, segundo os autores, amplamente usado por *data scientists* para obtenção de seus resultados em competições de aprendizado de máquina. O modelo foi vencedor do evento HiggsMachine Learning Challenge, proposto pelaEuropeanOrganization for Nuclear Research – CERN em 2014na plataforma Kaggle, no qual o objetivo era a identificação da partícula Bóson de Higgs por meio de algoritmos de *machinelearning*; sendo, este algoritmo, também premiado em 2016 com John ChambersAwardna categoria de estatística computacional, designado aos melhores softwares desenvolvidos no ano.

O XGBoost é um método de classificação com base no conceito de árvores de decisão, porém com diferenças significantes. Este modelo basicamente treina as novas árvores de decisão com base em suas estruturas fracas anteriores. Inicialmente são modeladas regressões para explicar os dados, de forma randômica, obtendo seus erros, e em seguida, com base nestes erros, são desenvolvidas novas regressões das quais éextraído um novo modelo com melhor desempenho (i.e., melhor capacidade preditiva). Desta forma, este processo segue os seguintes passos: (1) aprendizagem do modelo preditivo; (2) computação dos resíduos; e (3) aprendizagem para predição dos resíduos (DEY; *et al.*, 2016).

Este modelo foi utilizado em aplicações no mercado financeiro para previsão das tendências das ações, no qual teve um acerto de 87% para os períodos de 60 e 90 dias (DEY; *et al*., 2016). Yu(2017) utilizou esta ferramenta para prever o *Chicago BoardOptions Exchange (CBOE) Voltatility Index – VIX*, do primeiro ao sexto pregão subsequente da

análise, obtendo uma taxa de acerto entre 55% e 65%, dependendo da janela analisada, sendo a melhor acurácia obtida aplicando o modelo para cinco e seis dias.

Sob outra perspectiva, Carmona *et al.* (2018) utilizou o algoritmo XGBoost com o objetivo de prever a falência de bancos no mercado dos Estados Unidos entre os anos de 2001 e 2015, com uma serie de 30 indicadores financeiros de 156 bancos comerciais, obtendo um acurácia de 94,74%, ou seja, o melhor desempenho quando comparado aos demais modelos testados em seu trabalho, a Regressão Logística e *Random Forest* que obtiveram acertos de 84,21% e 92,11% respectivamente.

3 MÉTODO

Os dados, referente às negociações, utilizados para o cálculo do modelo da PIN e dos retornos foram extraídos por meio do programa R, utilizando principalmente pacotes GetHFData (PERLIN; RAMOS, 2016) para download e tabulação destes dados de negociações diárias do período analisado e xgboost para cálculo do modelo de predição. Para o cálculo do modelo da PIN foi utilizado o script disponibilizado por Bosque (2016) em seu apêndice C – PROGRAMAÇÕES UTILIZADAS.

Já para a classificação das operações de compra e construção das carteiras, foi utilizada a PIN e seu parâmetro Deltacomo *input*do algoritmo XGBoost, visto que segundo a fundamentação de EHO (2002) a PIN representa a probabilidade de negociação informada naquele determinado dia, já o delta indica o sentido da negociação, se negativa ou positiva. Este modelo foi constituído com objetivo em obtenção de retorno anormal positivo. Para a sua construção, foi utilizado o pacote do R denominado de xgboost, que é atualizado e mantido pelos próprios autores do algoritmo. Também foi aplicada a metodologia proposta por Lemagnen (2017) para ajuste fino dos parâmetros do modelo (*parametertunning*) por meio de validação cruzada em *k-folds*, tal metodologia será abordada mais a frente no item “3.2 Fluxo de Aprendizado” deste estudo.

3.1. Classificação das Operações

*FeatureEngineering*é um método utilizado na ciência de dados com objetivo de extrair mais informações das variáveis, do que as mesmas demonstram à primeira vista.

Yu (2017) coloca que a maior parte do sucesso de uma abordagem com *machinelearning* se dá por meio da aplicação do procedimento de *featureengineering*.O autor coloca que este procedimento é feito por meio da transformação dos valores dos *inputs*em outros valores que sejam mais fáceis de serem relacionados com o objetivo do modelo. Desta maneira este estudo se baseia, no número de agentes compradores e vendedores para identificação de ativos para compra, tendo em vista que estas variáveis são os *inputs* necessários para a computação do algoritmo da PIN, que após o cálculo da probabilidade de negociação informada, seus *outputs* PIN e Delta serão utilizados para predição dos ativos e seleção das carteiras de ações.

A mensuração da *proxy* de assimetria de informação nas negociações realizadas nas ações das companhias listadas B3 foi captada por meio da probabilidade de negociação informada (PIN) de Easley, Hvidkjaer e O’Hara (2002; 2010) adaptada por Lin e Ke (2011). Esta é obtida a partir do número e sentido dos negócios das ações (ordens de compra e venda)

em determinado período. O modelo da PIN desenvolvida por EHO identifica uma quantidade recorrente de envio de ordens de compra (ε_b) e venda (ε_s) das ações, e toma como base para sinalizar as negociações dos negociantes desinformadas, então um nível anormal de compras e vendas é utilizado para classificar negociações dos agentes informados (μ). A probabilidade de uma negociação de agentes informados (α) é obtida por meio do número de dias em que o nível de negociação foi anormal. A partir disso, a PIN é estimada por meio da união desses fatores simultaneamente, como demonstra a Equação (1).

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} \tag{1}$$

Em que:

α é a probabilidade de ocorrer um evento informacional;

μ é taxa de negociações de agentes informados;

ε_bé a taxa de compra de agentes desinformados; e

ε_sé a taxa de venda de agentes desinformados.

Bosque (2016) coloca que os autores Lin e Ke (2011) propuseram um arranjo para o melhoramento do modelo citado anteriormente, essa melhoria foi instigada pela presença de imprecisões no processo computacional da PIN de EHO (2002; 2010). Além disso, outros trabalhos corroboram e abordaram o mesmo ponto no que diz respeito à falha do mesmo modelo, este trabalho são: Yan e Zhang (2012), GanWei e Johnston (2015) e Ersan e Alice (2016). Neste sentido foi escolhido para este estudo o modelo adaptado da PIN por LK (2011), pelo fato dele atualmente ser, um benchmark para novos estudos (BOSQUE, 2016).

Assim, segue o método abordado para o modelo LK em que, pós-classificação e apuração das transações de compra e venda, são estimados os parâmetros do modelo apresentados na Equação (1), por meio de um modelo de negociação sequencial. Tal estimativa será desenvolvida pela maximização de uma função máxima verossimilhança,como demonstra a Equação (2).

$$L(\theta|B,S) = \sum_{i=1}^{n} \left\{ log \begin{bmatrix} \alpha\delta\, exp(e_{1i} - e_{maxi}) - \\ \alpha(1 - \delta) exp(e_{2i} - e_{maxi}) + \\ (1 - \delta) exp(e_{3i} - e_{maxi}) \end{bmatrix} + B_i log\,(\varepsilon_b + \mu) \right.$$

$$\left. + S_i\, log(\varepsilon_S + \mu) - (\varepsilon_b - \varepsilon_s) + \varepsilon_{maxi} - log(B_i!\, S_i!) \right\}, \tag{2}$$

Sendo:

$$e_{1i} = -\mu - B_i \, log\left(1 + \frac{\mu}{\varepsilon_b}\right)$$

$$e_{2i} = -\mu - S_i \, log\left(1 + \frac{\mu}{\varepsilon_S}\right)$$

$$e_{3i} = -\mu - B_i \, log\left(1 + \frac{\mu}{\varepsilon_b}\right) - \mu - S_i \, log\left(1 + \frac{\mu}{\varepsilon_S}\right)$$

$$e_{maxi} = max(e_{1i}, e_{1i}, e_{1i}), \qquad i = 1, 2, \dots, n$$

Em que:

B e S são os volumes de compras e vendas;

B_ieS_i são as negociações no dia i;

θ é o vetor de parâmetros (α, μ, δ, ε_b e ε_s);

(1-δ) é a probabilidade de ser uma “boa notícia”; e

δ é a probabilidade de ser uma “má notícia”.

O procedimento de *Machine Learning* é dado em função de um determinado objetivo. Em Samuel (1959) o objetivo era a melhoria do resultado do jogo de damas; Carmona *et al.*(2018) projetou seu algoritmo para identificar a falência de bancos, Yu (2017) utilizou a mesma técnica para previsão da volatilidade de *Chicago BoardOptions ExchangeVoltatility Index –VIX*, como citado anteriormente, já o presente trabalho se desenvolve na busca de ativos que sejam possível a obtenção de retorno anormal pelos seus compradores, desta maneira foi aplicado o modelo XGBoost com objetivo em uma regressão linear para predição dos retornos anormais futuros, parâmetroencontrado como *reg:linear* na bibliografia do algoritmo. Assim para o cálculo do retorno anormal, se adotou o CAPM como estimativa do custo de capital seguindo a seguinte formula (3):

$$Ke_i = R_f + B_i(R_m - R_f) \tag{3}$$

Em que:

Ke_ié o custo do capital próprio da ação *i;*

R_fé a taxa livre de risco;

B_ié o coeficiente beta-mercado da ação *i*; e

R_mé o retorno da carteira de mercado.

Para realização deste cálculo, este trabalho teve como base as seguintes variáveis: Título do tesouro nacional LTN com vencimento em 2023, para a taxa livre de risco (R_f); o retorno mensal do Índice Ibovespa, para estimação do beta-mercado para as ações (B_i); e o mesmo índice Ibovespa desde 1994, para o cálculo do retorno esperado de mercado (R_m).

Por fim, com o custo de capital em mãos é verificado o retorno anormal de cada ativo, e finalmente os *inputs* do modelo XGBoost estão completos. Para obtenção do Retorno

Anormal – Ra é efetuada a seguinte equação (4):

$$Ra_i = Ro_i - Ke$$

$Sendo$: (4)

$$Ro_i = \frac{P_t}{P_{t+20}}$$

Em que:

Ro_ié o retorno observado para a ação *i*;

P_té o preço da ação na data *t*;

P_{t+20}é o preço da ação no 20º pregão subsequente a data *t*; e

Ro_ié o retorno observado para a ação *i*;

O algoritmo de *machinelearning*utilizado para classificação das ações com maior potencial de retorno anormal futuro foiXGBoost. Este modelo tem em sua bibliografia informações referentes a todos os parâmetros passiveis de alteração pelos usuários, estas alterações são feitas com intuito de adequar da melhor forma possível o sistema à amostra estudada.

O ajuste fino do algoritmo para este trabalho foi feito com base na literatura disponibiliza pelo CambridgeSpark, elaborada por Lemagnen (2017), o autor propõe a busca dos valores ideais para os parâmetros descritos no Quadro (1) abaixo, com suas respectivas características.

Quadro 1 – Descrição dos Parâmetros Aprimorados.

Parâmetro	Característica
max_depth	É o número máximo de nós permitidos da raiz até a folha mais distante da arvore de decisão. Quanto mais profunda é a árvore, mais complexo será o modelo, porém é preciso ter cuidado para não chegar ao ponto de que a partição da amostra seja menos relevante, o que pode causar *overfit* ao modelo.
min_child_weight	Permite que o modelo crie árvores menores com menor número de amostra, o que torna o modelo mais complexo em contrapartida propenso a *overfit*.
subsample	Corresponde a fração de observações (na linha) usadas em cada *round.* O valor 1 significa que serão utilizadas todas as linhas.
colsample_bytree	Corresponde a fração de observações (na coluna) usadas em cada *round.* O valor 1 significa que serão utilizadas todas as colunas.
eta	Este parâmetro controla a taxa de aprendizagem do modelo. Isto corresponde a diminuição dos pesos associados a cada variável após cada *round*, ou seja, o tamanho da correção que é feita. Na prática, quanto menor é o valor deste parâmetro, mais robusto será o modelo contra *overfitting.* Porém quanto menor o valor do eta, será necessário rodar mais interações, o que leva mais tempo.

Adaptado de Lemagnen (2017)

Nó próximo tópico será demonstrado o método de *ParameterTunning* para os possíveis valores dos parâmetros informados acima.

3.2. Consolidação da Amostra

Inicialmente, a amostra do deste estudo foi composta por 552 ações(*tickers*) encontrados ao se consolidar as negociações diárias do período de outubro de 2015 até dezembro de 2017. Esta amostra foi utilizada para o cálculo da PIN de todos os *tickers* para todos os dias de pregão para o período descrito acima.A amostra ainda passou por dois filtrospara compor o banco de dados de *input* do XGBoost.

Primeiro filtro: de acordo com a teoria da PIN, para que o ativo esteja apto ao cálculo da PIN, o mesmo deve ter sido negociado nos últimos 60 pregões antecedentes à data de cálculo do modelo, tal condicional é aplicada desta maneira no programa deste trabalho: *if (numero_de_pregoes<= 59){nextdate}*, ou seja, caso o ativo não esteja em acordo com o pré-requisito da PIN na data X, o algoritmo seguirá para a próxima data da amostra.Ao término do processo, a amostra se reduziu para 232 ações.

Segundo filtro: após a obtenção de todos os parâmetros da PIN para as ações que se adequaram ao modelo, é efetuado o cálculo do custo do capital, convertendo-o para a taxa de equivalência de 20 pregões, período que será formulada o rearranjo das carteiras. Neste segundo filtro, foram selecionadas apenas as empresas que tiveram o Beta positivo, sendo

excluídas as que possuam valores negativos ou nulos, pois tais empresas carecem de outras informações para o cálculo eficiente do seu custo de capital. Ao término deste procedimento, a amostra contém informações de 171 companhias (amostra final).

De posse do Ke, o próximo passo é estruturar a tabela com os valores da PIN, Delta e do Retorno Anormal, e finalmente se inicia o processo de aprendizado de máquina, de acordo com o procedimento que segue nos próximos parágrafos.

3.3. Aprendizagem e Predição

O método inicia com a segmentação da amostra para aprendizado, estabelecida neste estudo por 60 pregões de cálculo da PIN, o que corresponde ao todo em 120 pregões analisados, visto que serão 60 com informações da PIN, e outros 60 que são necessários para o cálculo desta*proxy*, seguindo sua teoria.

Segundo passo é feito por meio do aprimoramento dos parâmetros do XGBoost por meio de *cross-validation.* O método utilizado para efetuar este procedimento foi adaptado da linguagem *Python* para *R*e foi proposto por Lemagnen (2017),sendo efetuadas modificações em cinco parâmetros do algoritmo XGBoost. Desta maneira, de acordo com os parâmetros citados anteriormente, seguem os valores testados para cada um deles: *max_depth(9;10;11;12);min_child_weight(5;6;7;8);*
subsample(0,7;0,8;0,9;1);colsample_bytree(0,7;0,8;0,9;1); e eta(0,3; 0,2; 0,1; 0,05; 0,01; 0,005).

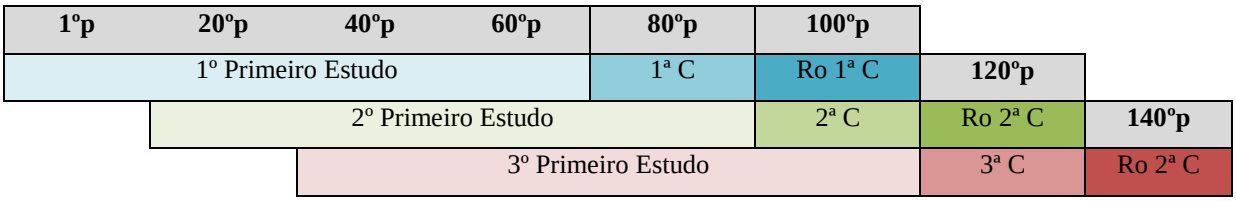
Ainda segundo o autor, para controlar as distorções que o algoritmo sofre ao se alterar cada um destes parâmetros, o procedimento foi dividido em três partes: primeiro foram calibrados o *max_depth*e o *min_child_weight*,sendo 16 diferentes modelos, incluindo o melhor resultado deles para a próxima etapa, calibrando os parâmetros*subsample* e*colsample_bytree* sendo extraídos mais 16 modelos, da mesma maneira selecionando o melhor resultado e finalmente, este resultado é testado com os valores propostos para o *eta*, adicionando 5 simulações, o que totalizou 37 modelos testados para seleção do que obtiver menor erro de predição.

Ao término do procedimento citado acima, os parâmetros ideais para amostra são aplicados ao modelo de treinamento, com a amostra citada no início da descrição deste procedimento, em seguida é feita a predição para os valores do 20º subsequente ao dia de aprendizado e selecionados os 10 ativos que obtiverem maior expectativa de retorno futuro.

Este procedimento é efetuado novamente tendo como dias de aprendizado, acrescentando os pregões entre o último pregão utilizadopara o aprendizado anterior e o

utilizado para a predição da carteira, sendo excluídos os 20 primeiros pregões utilizados na primeira predição. Tal procedimento é repetido para a constituição das 21 carteiras.O Gráfico (1) abaixo descreve a metodologia descrita:

Gráfico 1 – Descrição do Fluxo de Aprendizado Empregado.



Nota: *p* representa o pregão referido; 1ª C representa a formulação da primeira carteira; e Ro 1ª C o retorno observado da primeira carteira. Elaborado pelo autor (2018)

Já na Tabela (1),a seguir, está demonstrado o resumo do modelo aplicado neste estudo:

Tabela 1 – Descrição do Modelo Utilizado.

Item	Descrição
Modelo	*eXtremeGradientBoosting.*
Objetivo	*reg:linear*
Métrica de avaliação	*Root Mean Square Error - RMSE*
Parameterstuning	*Cross-validation k-folds.*
Variáveis explicativas	Número de compra e venda diária.
FeatureEngineering	PIN - LK (2011).
Variável-objetivo	Retorno anormal em t+20
Inputs	Pin e Delta.
Outputs	Projeção do retorno anormal em t+20

Elaborado pelo autor (2018)

A análise do desempenho financeiro das carteiras foi feita por meio da Acurácia,Retorno Acumulado, do Retorno Anormal (Fórmula 4), doÍndice de Sharpe (1964) e do Índice de Treynor(1961). O Quadro (2) a seguir demonstra as fórmulas e os parâmetros dos indicadores que não foram citados até o momento neste estudo.

Quadro 2 – Indicadores Para Medida de Desempenho do Modelo, Formula e Descrição.

Indicador	Formula	Descrição
Acurácia	$Ac = \frac{PrediçõescomRopositivo}{NúmerodePredições}$	Ou seja, representa o quanto que o modelo acertou no que diz respeito à evolução do ativo analisado. Quanto das sugestões obteve retorno positivo.
Retorno Acumulado	$C_{pt} = C_{pt-1} * (1 + Ro_t)$	Ou seja, o capital "*C*" do portfólio "*p*" na data "*t*" é obtido pelo capital do portfólio na data "*t-1*", mais o retorno observado R_o na data "*t*".
Índice de Sharpe	$IS = \frac{[E(Ro) - Rf]}{DP_{Ro}}$	Ou seja, o Índice de Sharpe mede a relação entre a esperança do prêmio pelo risco do ativo e o risco do mesmo, representado pelo Desvio-padrão do retorno observado.
Índice de Treynor	$IT = \frac{[E(Ro) - Rf]}{B_P}$	Ou seja, o Índice de Treynor mede a relação entre a esperança do prêmio pelo risco do ativo e risco sistemático do ativo, representado pelo Beta-mercado.

Elaborado pelo autor (2018)

4 RESULTADOS

Neste trabalho foram selecionadas as 10 ações com maior predição de retorno anormal em 20 pregões, sendo reformuladas após a observação do retorno projetado, como demonstrou o Gráfico (1).Esta reformulação se deu por meio do algoritmo XGBoost, entre os pregões 02/05/2016 e 07/12/2017, totalizando 21 carteiras observadas. A Tabela (2) demonstra todos os ativos selecionados em cada carteira, assim como sua classificação de acordo com o valor do seu retorno esperado (com base na predição do modelo) e o retorno observado para os 10 e 5 ativos com maiores valores preditos.

Tabela 2 – Relação dos Ativos Selecionados Pelo Modelo, Entre 02/05/2016 e 07/12/2017.

Carteiras de 02/05/2016 a 20/10/2016							
Posições	**02/05/2016**	**31/05/2016**	**28/06/2016**	**26/07/2016**	**23/08/2016**	**21/09/2016**	**20/10/2016**
1ª	CPLE3	ESTC3	JHSF3	CSNA3	UNIP6	TIMP3	BRPR3
2ª	FESA4	LOGN3	CSMG3	SANB4	USIM5	JHSF3	AGRO3
3ª	PRML3	BTOW3	BRSR6	IGTA3	PDGR3	CARD3	HBOR3
4ª	SULA11	LLIS3	OIBR3	RENT3	PFRM3	ROMI3	LUPA3
5ª	RSID3	SGPS3	SLED4	LAME3	CMIG3	CPLE3	ALPA4
6ª	ETER3	TCSA3	AGRO3	LREN3	TUPY3	VALE3	RAPT4
7ª	DIRR3	LREN3	CARD3	POMO3	CGAS5	HYPE3	COCE5
8ª	ABEV3	BVMF3	POMO3	USIM3	BEEF3	KLBN4	PFRM3
9ª	MDIA3	OGXP3	HBOR3	RAPT4	BRIN3	LUPA3	ELET3
10ª	HGTX3	PETR4	POMO4	KLBN4	JHSF3	ECOR3	JHSF3
Ro 10 Atv	*-6,99%*	*-0,55%*	*36,06%*	*0,48%*	*0,16%*	*6,41%*	*-10,73%*
Ro 5 Atv	*-11,84%*	*0,87%*	*42,39%*	*-0,17%*	*-3,05%*	*9,43%*	*-13,76%*
Carteiras de 21/11/2016 a 17/05/2017							
Posições	**21/11/2016**	**19/12/2016**	**17/01/2017**	**15/02/2017**	**17/03/2017**	**17/04/2017**	**17/05/2017**
1ª	MILS3	PSSA3	KROT3	BRIN3	FHER3	CMIG3	TRPL4
2ª	BBRK3	MYPK3	ODPV3	AMAR3	PFRM3	GRND3	TCSA3
3ª	CSMG3	GGBR4	BBDC3	TESA3	ITSA3	BBRK3	ENGI11
4ª	BBAS3	PCAR4	VIVT4	PTBL3	LOGN3	FHER3	RENT3
5ª	CARD3	DIRR3	BBDC4	SBSP3	DIRR3	DIRR3	TGMA3
6ª	CCRO3	CIEL3	GRND3	MYPK3	CCRO3	RCSL4	POMO4
7ª	ESTR4	PFRM3	GGBR4	LIGT3	ABCB4	TCSA3	BRPR3
8ª	AGRO3	ESTC3	BRFS3	EVEN3	BPAN4	GUAR3	SAPR4
9ª	GOAU4	ITUB4	ENBR3	SLED4	GRND3	LPSB3	MAGG3
10ª	SHUL4	MRVE3	ALPA4	ALSC3	VIVT3	SCAR3	TIET4
Ro 10 Atv	*-1,17%*	*12,57%*	*2,89%*	*0,20%*	*-0,21%*	*3,81%*	*-3,08%*
Ro 5 Atv	*3,87%*	*13,69%*	*0,02%*	*0,41%*	*1,30%*	*2,23%*	*-6,28%*
Carteiras de14/06/2017 a 07/12/2017							
Posições	**14/06/2017**	**13/07/2017**	**10/08/2017**	**08/09/2017**	**06/10/2017**	**07/11/2017**	**07/12/2017**
1ª	VIVT4	PCAR4	OIBR4	AGRO3	TRPN3	TIET3	CRPG5
2ª	PTBL3	FIBR3	PRML3	GFSA3	CGAS5	VVAR3	LPSB3
3ª	ECOR3	EQTL3	MAGG3	ENBR3	TCSA3	HBOR3	OIBR4
4ª	BVMF3	ITUB3	JHSF3	BTOW3	DTEX3	FJTA4	ELET6
5ª	CCRO3	KLBN4	LOGN3	FIBR3	ABEV3	BRFS3	MILS3
6ª	CYRE3	GOLL4	ITSA3	WEGE3	FHER3	ALPA4	RSID3
7ª	GGBR3	TIET4	JSLG3	JHSF3	BBRK3	CGAS5	ABCB4
8ª	SANB11	IGTA3	ABCB4	GOAU3	FJTA4	OIBR4	VVAR3
9ª	SLCE3	RSID3	CIEL3	ITUB3	PRML3	MAGG3	BRML3
10ª	LREN3	BOBR4	GOAU4	LAME3	TIET3	KROT3	CESP6
Ro 10 Atv	*7,90%*	*6,41%*	*9,20%*	*7,77%*	*-5,26%*	*1,88%*	*2,58%*
Ro 5 Atv	*6,23%*	*4,89%*	*9,19%*	*12,00%*	*-4,80%*	*-3,28%*	*-1,79%*

Nota: Ro 5 Atv e Ro 10 Atv referem-se aos retornos observados para a carteira de 10 e 5 ativos, respectivamente. Elaborado pelo autor (2018)

No total foram selecionadas 210 predições para compor a carteira de 10 ativos e 105 para compor a carteira dos 5 ativos com maior retorno predito. Da careira de 10 ativos, 121 (57,62%) obtiveram Retorno Observado positivo, e 110 (52,38%) obtiveram Retorno Anormal positivo, o que indica que o modelo teve êxito na sugestão de mais da metade dos ativos da carteira.

Já no que diz respeito à carteira de 5 ativos, o acerto relacionado ao Retorno Observado foi proporcionalmente similar à carteira anterior, 59 (56,19%), porém quando analisada a acurácia para obtenção de retorno anormal positivo, este indicador é reduzido para 48,57%, ou 51 ativos com Retorno Anormal positivo. Como demonstra a Tabela (3) abaixo:

Tabela 3 – Relação Entre Predições e sua Acurácia.

	Seleções	**Ro> 0**	**Ra > 0**	**AcRo**	**Ac Ra**
Carteira 10	210	121	110	57,62%	52,38%
Carteira 5	105	59	51	56,19%	48,57%

Elaborado pelo autor (2018)

Estes resultados se equiparam aos resultados encontrados por Yu (2017), que obteve êxito de entre 55% e 65%, porém ficam muito aquém da eficiência dos resultados obtidos nas pesquisas de Dey*et al* (2016), 87% de acerto e os 94,74% de acurácia obtida por Carmona *et al*. (2018). Porém, vale ressaltar, que são estudos com objetivos e amostras diferentes, portanto estão sendo levados em conta apenas para comparação da acurácia do algoritmo em si, visto que não foi encontrada na literatura metodologia similar.

Já com relação ao desempenho financeiro, o resultado cumpriu o objetivo proposto inicialmente, a obtenção de retorno anormal por meio da seleção de portfólio com base na assimetria de informação captada pelo algoritmo de microestrutura de mercado (PIN). A Tabela (4) abaixo demonstra os resultados financeiros obtidos pelas carteiras e os compara com a *proxy* de mercado e a taxa livre de risco adotada neste estudo.

Tabela 4 – Desempenho Financeiro das Carteiras Elaboradas Entre 02/05/2016 e 28/12/2017.

	Retorno Anormal	**Retorno Acumulado**	**Retorno Médio[1]**	**Desvio-padrão**	**Beta-mercado**	**Índice de Sharpe**	**Índice de Treynor**
Carteira 10	32,94%	85,40%	3,35%	9,35%	1,07	0,2695	2,35%
Carteira 5	16,46%	64,02%	2,93%	11,45%	1,23	0,1834	1,71%
Ibovespa		47,24%	2,07%	6,70%	1,00	0,1849	1,24%
LTN 5anos		18,97%	0,83%	0,08%	0,00		

[1] Retorno Médio das carteiras. Elaborado pelo autor (2018)

O retorno anormal obtido para o período do estudo foi positivo tanto na carteira de 10 ativos como na de 5 em 32,94% e 16,46%, respectivamente. Já em Retorno Acumulado, o desempenho foi das carteiras de 10 (85,40%) e de 5 (64,02%) foram superiores ao observado

pelo Ibovespa, que obteve um retorno acumulado de 47,24%, o qual teve um prêmio pelo risco de 28,27% no mesmo período. Estes resultados entram em consonância às evidências encontradas por Martins, Paulo e Girão (2014), com relação à assimetria captada pela PIN ter *value relevance* novalor de mercado das empresas, e mais recentemente, Siqueira, Amaral e Correia (2017) que encontraram indícios de que o risco informacional é precificado no mercado acionário brasileiro, quando utilizado o *volume-synchronizedprobabilityofinformed trading* (vPIN) como base, outro modelo aprimorado da PIN de EHO.

Quando observada a relação entre risco e retorno das carteiras, a Carteira de 10 ativos obteve a melhor relação, tanto no Índice de Sharpe, quanto no Índice de Treynor. Este resultado é fundamentado pela teoria de Markwitz (1952), quando o autor propões que a diversificação dos investimentos, efetuada de forma correta, dissolve o risco intrínseco à companhia restando apenas o risco sistemático do mercado.

5 CONCLUSÃO

Este estudo teve como principal motivação a criação de uma metodologia de seleção de ativos presentes na B3, tendo como base informações de microestrutura de mercado, adquiridas por meio da ferramenta desenvolvida por Perlin e Ramos (2016) e aplicadas ao modelo da PIN-LK (2011), só então, com os valores referentes a probabilidade de negociação informada dos ativos, e o sentido desta negociação foramintroduzidos no algoritmo *eXtreme Gradiente Boosting*.

O modelo foi desenvolvido com objetivo de obtenção de retorno anormal, os resultados demonstraram que existe a possibilidade de obtenção de retorno anormal utilizando apenas a PIN e o Delta como parâmetro de seleção de ativos, encontrando 32,92% e 16,46% de retorno acima do custo de capital para as carteiras de 10 e 5 ativos, respectivamente. Além disso,caso um investidor estivesse utilizado esta metodologia no período abordado, em 57,62% das ocasiões o teria obtido retorno positivo se houvesse selecionando os 10 ativos com maiores predições de retorno anormal, e em ao menos 52,38% dosobtiveram um retorno acima do precificado pelo CAPM. Estes resultados se mostraram inferiores quando analisados apenas os 5 primeiros ativos, o que pode ser explicado pela diversificação do portfólio proposta por Markowitz (1952).

Vale ressaltar que estes resultados são referentes ao modelo de classificação utilizado, com a *proxy*de assimetria utilizada, agregada ao custo de capital empregado e o período observado. Desta maneira, como proposta para futuros estudos sugere-se a inclusão de outras variáveis, tanto de assimetria, quanto financeiras, com intuito de verificar se estas*features*têm a capacidade de aumentar o poder preditivo do modelo. Também se sugere a ampliação do período de análise, tendo em vista que só foi possível coletar dados referentes às transações a partir de outubro de 2015. Outro aperfeiçoamento pode ser aplicado com o aprofundamento do teste de validação cruzada com o algoritmo empregado, visto que é necessário alto poder de processamento computacional para extração de todo potencial do modelo.

Importante informar que estes dados não estão mais em domínio público por conta da janela temporal de dois anos disposta para a base no repositório *ftp* da B3. Ademais, o estudo demonstra que o poder dos dados e o processo de obtenção de informações referente a estes dados sãodinâmicos e de evolução contínua, dominar ferramentas tecnológicas de ponta pode ser diferencial competitivo, tanto para empresas privadas, quanto para o governo quando disposto a aprimorar a eficiência e segurança do mercado de capitais.

REFERENCIAS

ABAD, D.; RUBIA, A. Modelos de estimación de laprobabilidad de negociación informada: uma comparación metodológica enel mercado Español. **Revista de EconomíaFinanceira**, n. 7, p. 1- 37, 2005.

BARANIAK, K. **"ISMIS 2017 Data Mining Competition: Trading Based on Recommendations-XGBoost Approach with Feature Engineering."**Intelligent Methods and Big Data in Industrial Applications.Springer, Cham, 2019.145-154.

BOSQUE, L. M. **Estimação da probabilidade de negociação privilegiada por meio de inferência bayesiana.** Dissertação (Mestrado). Programa de Pós-Graduação em Administração, Universidade de Brasília, Brasília, 2016.

BTG Pactual. **Conheça o AlgoTrading, técnica que automatiza a tomada de decisão no mercado,** 2018.Disponível em <https://www.btgpactualdigital.com/blog/investimentos/renda-variavel/conheca-o-algotrading-tecnica-que-automatiza-a-tomada-de-decisao-no-mercado>. Acesso em: 25/08/2018

CALLADO, A. A. C. **Eficiência do mercado acionário brasileiro: retorno das ações negociadas na Bovespa, variáveis macroeconômicas, causalidade e fatores condicionantes**. Tese (Doutorado). Programa de Pós-Graduação em Administração, Universidade Federal de Pernambuco, Recife, 2009.

CARMONA, P., CLIMENT, F., & MOMPALER, A. Predicting bank failure in the US banking sector: An extreme gradient boosting approach.**International Review of Economics & Finance**.2018.

CERN.**Machine Learning Wins the Higgs Challenge**, 2014. Disponível em:<https://atlas.cern/updates/atlas-news/machine-learning-wins-higgs-challenge> Acesso em: 15/01/2018.

Chen, Tianqi, Tong He, and Michael Benesty. Xgboost: extreme gradient boosting. **R package version** 0.4-2 2015: 1-4.

Demsetz, Harold. **Corporate control, insider trading, and rates of return**. The American Economic Review 76.2 (1986): 313-316.

DEY, S.; et al. **Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting**. Working paper, 2016. DOI: 10.13140/RG. 2.2. 15294.48968.

EASLEY, D.; et al. Liquidity, information, and infrequently traded stocks. **The Journal of Finance** 51.4 1996: 1405-1436.

EASLEY, D.; HVIDKJAER, S.; O'HARA, M. Factoring information into returns.**Journal of Financial and Quantitative Analysis** 45.2: 293-309, 2010.

EASLEY, D.; HVIDKJAER, S.; O'HARA, M.Is information risk a determinant of asset returns?.**The journal of finance** 57.5: 2185-2221, 2002.

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

FAMA, E. F. **Random walks in stock market prices.**Financial analystsjournal 51.1: 75-80, 1995.

GIRÃO, L. F. A. P.; MARTINS, O. S.; PAULO, E. **O Lado B do Insider Trading: Relevância, Tempestividade e Influência do Cargo**. In: Congresso USP de Controladoria e Contabilidade, São Paulo, 2014.

GIRAO, L. F. de A. P.; MARTINS, O. S.; PAULO, E.. Avaliação de empresas e probabilidade de negociação com informação privilegiada no mercado brasileiro de capitais. **Rev. Adm. (São Paulo)**, São Paulo , v. 49, n. 3, p. 462-475, Sept. 2014.

JENSEN, M. C.; MECKLING, W. H. Theory of the firm: managerial behavior, agency costs and ownership structure.**Journal of Financial Economics**, v. 3, n. 4, p. 305-360, 1976.
LEMAGNEN, K. **Hyperparameter tuning in XGBoost**, 2017. Disponível em <https://blog.cambridgespark.com/hyperparameter-tuning-in-xgboost-4ff9100a3b2f>. Acessoem: 02/02/2018.

LIN, H. W., KE, W. A computing bias in estimating the probability of informed trading.**Journal of Financial Markets** 14.4: 625-640, 2011.

LINTNER, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets.**Review of Economics and Statistics**, p. 13-37, 1965.
MARKOWITZ, H. H. Portfolio selection.**Journal of Finance**, v.7 n.77, p.91, 1952.

MARTINS, O. S.; PAULO, E.; GIRÃO, L. F. A. P.. Preço da Ação, Disclosure e Assimetria de Informação: o Caso OGX. **RevistaUniversoContábil**, v. 12, p. 6-24, 2016.

PERLIN, M., RAMOS, H. GetHFData: A R Package for Downloading and Aggregating High Frequency Trading Data from Bovespa. **Brazilian Review of Finance**, V. 14, N, 2016.
SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development** 3.3: 210-229, 1959.

SHARPE, W. F. Capital asset prices: a theory of market equilibrium under conditions of risk.**Journalof Financial**, v. 19, p. 425-442, 1964.

SIQUEIRA, L. S.; AMARAL, H. F.; LAÍSE, LAÍSE, F. C. O efeito do risco de informação assimétrica sobre o retorno de ações negociadas na BM&FBOVESPA**.RevistaContabilidade&Finanças** 28.75: 425-444, 2017.

SONG, Y. **Stock Trend Prediction: Based on Machine Learning Methods**. Tese (Mestrado), Departamento de Estatística da Universidade da California, 2018.

TREYNOR, J. **Toward a theory of the market value of risky assets**. Artigo não publicado, 1961.

VERAS, M. M. A.; MEDEIROS, O. R. Modelos de precificação de ativos e o efeito liquidez: evidências empíricas no mercado acionário brasileiro.**RevistaBrasileira de Finanças** 9.3, 2011.

YU, M. Y. **Predicting the Volatility Index Returns Using Machine Learning**. Tese (Mestrado), Deparatamento de Matemática da Universidade de Toronto, 2017.

ANEXO – Computação da PIN

Script da PIN-LK (2011) disponibilizado por Bosque (2016) em seu apêndice C – PROGRAMAÇÕES UTILIZADAS.

```
f.YZLKPIN <- function(buySell) {
  Bmean <- mean(buySell[,1])
  Smean <- mean(buySell[,2])
  solutions <- matrix(0, 125, 7)
  counter <- 1
  if (Bmean <= Smean) {
    for (alpha1 in seq(0.1, 0.9, 0.2)) {
      for (delta1 in seq(0.1, 0.9, 0.2)) {
        for (gamma in seq(0.1, 0.9, 0.2)) {
alpha <- alpha1
          delta <- delta1
          Eb <- gamma * Bmean
mu <- (Bmean - Eb) / (alpha * (1-delta))
          Es <- Smean - alpha*delta*mu
          if(Eb < 0 | Es < 0 | mu < 0 ) {
            next
          }
          xx <- f.LKPIN(buySell, c(alpha, delta, Eb, Es, mu))
          solutions[counter, 1] <- xx$PIN
          solutions[counter, 2] <- xx$neglogl
          solutions[counter, 3] <- xx$param[1]
          solutions[counter, 4] <- xx$param[2]
          solutions[counter, 5] <- xx$param[3]
          solutions[counter, 6] <- xx$param[4]
```

```
        solutions[counter, 7] <- xx$param[5]
        counter <- counter + 1
       }
     }
   }
  } else {
   for (alpha1 in seq(0.1, 0.9, 0.2)) {
     for (delta1 in seq(0.1, 0.9, 0.2)) {
       for (gamma in seq(0.1, 0.9, 0.2)) {
alpha <- alpha1
        delta <- delta1
        Es <- gamma * Smean
mu <- (Smean - Es) / (alpha * delta)
        Eb <- Bmean - alpha*(1-delta)*mu
        if(Eb < 0 | Es < 0 | mu < 0 ) {
          next
        }
        xx <- f.LKPIN(buySell, c(alpha, delta, Eb, Es, mu))
        solutions[counter, 1] <- xx$PIN
        solutions[counter, 2] <- xx$neglogl
        solutions[counter, 3] <- xx$param[1]
        solutions[counter, 4] <- xx$param[2]
        solutions[counter, 5] <- xx$param[3]
        solutions[counter, 6] <- xx$param[4]
        solutions[counter, 7] <- xx$param[5]
        counter <- counter + 1
       }
```

```
    }
   }
  }
  solutionSet <- solutions[solutions[,1] != 0,]
  PIN <- solutionSet[solutionSet[,2] == min(solutionSet[,2]), 1]
  if (length(PIN) > 1) {PIN <- PIN[1]}
  param <- solutionSet[solutionSet[,2] == min(solutionSet[,2]), 3:7]
  if(length(param) > 5) {param <- param[1,] }

  if(length(PIN) == 0) { PIN <- 0}
  if(length(param) == 0) { param <- c(0,0,0,0,0)}
  return(list(PIN= PIN, param = param, solutions = solutions))
}
f.LKML <- function(param, buySell) {
alpha <- param[1]
  delta <- param[2]
  Eb <- param[3]
Es <- param[4]
  mu <- param[5] # build LK (2011) dynamic factorization components
e1 <- -mu - B*log(1 + mu / Eb)
  e2 <- -mu - S*log(1 + mu / Es)
  e3 <- -B*log(1+ mu / Eb) - S*log(1+ mu / Es)
  emax <- apply(as.matrix(cbind(e1,e2,e3)), 1, max)

  logl <- log(alpha*delta*exp(e1 -emax) + alpha*(1-delta)*exp(e2 - emax) +
              (1-alpha)*exp(e3 - emax)) + B*log(Eb + mu) + S*log(Es + mu) - (Eb + Es) + emax
logl[is.infinite(logl)] <- 0
```

```
  logl[is.nan(logl)] <- 0
  return(-sum(logl)) # return negative log likelihood
}
f.LKPIN <- function(buySell, initial) {
  # param: pnews, pbad, mu, epsB, epsS
  # initial <- c(0.5, 0.5, 1, 1, 1)
  optim(initial, f.LKML, buySell = buySell, method = "L-BFGS-B",
        lower = c(0,0,0.01,0.01, 0.01), upper = c(1,1,Inf, Inf, Inf))-> xx
param <- xx$par
  neglogl <- xx$value
  PIN <- (param[1] * param[5])/(param[3] + param[4] + param[1]*param[5]) # PIN
return(list(PIN = PIN, param = param, neglogl = neglogl))
}
```

APÊNDICE – Treinamento e Predição

Programação desenvolvida pelo autor para construção das carteiras.

```
library(data.table)
library(GetHFData)
library(xgboost)
library(caret)
library(plyr)

{
  # Params base
    params <- list(objective = "reg:linear", eval_metric = "rmse")
  # Part 1
    max.depths = c(9, 10, 11, 12) # 4 valores (4 int)
    min.child.weight = c(5, 6, 7, 8) # 4 valores (12 int)
  # Part 2
    subsampless = c(0.7, 0.8, 0.9, 1) # 4 valores (16 int)
colsample.bytree = c(0.7, 0.8, 0.9, 1) # 4 valores (28 int)
  # Part 5
    etas = c(0.3, 0.2, 0.1, 0.05, 0.01, 0.005) # 6 valores (34 int)
# Results table
    cv.results <- matrix(nrow = 1, ncol = 9)
} # Parameters tuning

{
  all.days <- ddply(dataset,.(date),summarize,PIN=mean(pin))

  rank.f <- matrix(nrow = 1, ncol = 180)
  colnames(rank.f) <- c('ticker','date','pin','alpha','delta','eb','es','mu','ra','pred','pred1')
```

```
  # Tickers to dummyvars
  dummy.vars <- dummyVars(~ ., data = dataset)
  train.dummy <- predict(dummy.vars, dataset)
  dummys <- train.dummy[,1:171]
  dataset <- cbind(dataset1, dummys)

a <- 0 # Seleção do periodo de estudo
  b <- 0 # Seleção da data de predição
  c <- 20 # Intervalo da observação do retorno
  d <- as.integer((length(all.days[,1])/c)-1) # Numero de predições
} # Tratamento dos dados para o XGboost ###### SELEÇÃO DE CARTEIRAS

# ANÁLISE EFETUDA COM BASE NAS SEGUINTES VARIÁVEIS
# PIN
    # DELTA

for(i in 1:d){
{
  a <- c*i+40
  b <- a+c+1
  e <- a-c-40

  train <- all.days[e:a,1]
set.train <- subset(dataset, date %in% train)
set.train <- set.train[,-c(1,2,4,6,7,8)] # retirar 4,6,7,8 quando for PIN e DELTA

test <- all.days[b:b,1]
  test <- subset(dataset, date %in% test)
```

```
  # Results table
  cv.results <- matrix(nrow = 1, ncol = 9)
} # Inicio do processamento | Alterar var objetivo


{
  TimeI <- Sys.time()
for(d in 1:40){


  train.1 <- train[d:d,1]
  set.train.1 <- subset(set.train, date %in% train.1)


label <- set.train$ra #ifelse(set.train$ra > 0, 1, 0) # ALTERAR DEPENDENDO DO
OBJETIVO set.train$ra #
label <- t(label)


  cv.train <- as.matrix(set.train.1[,-3])
  cv.train <- xgb.DMatrix(cv.train, label = label)


  for(max_depth in max.depths){
    for(min_child_weight in min.child.weight){
      mcv.1 = xgboost(params = params,
                      data = cv.train,
                      nrounds = 1000,
                      early_stopping_rounds = 10,
                      nfold = 10,
                      verbose = 0,
                      max_depth = max_depth,
                      min_child_weight = min_child_weight)
```

```
    bint <- which.min(mcv.1$evaluation_log$test_rmse_mean) # modificar por tipo do
modelo

    metric <- min(mcv.1$evaluation_log$test_rmse_mean) # modificar (max ou min) por tipo
do modelo


    cv.1 <- matrix(nrow = 1, ncol = 9)

    cv.1[,1] = bint

    cv.1[,2] = max_depth

    cv.1[,3] = min_child_weight

cv.1[,4] = 0

    cv.1[,5] = 1

    cv.1[,6] = 1

    cv.1[,7] = 0

    cv.1[,8] = 0.3

    cv.1[,9] = metric

cv.results <- rbind(cv.results,cv.1)

    print(cv.results)

    print(paste0('Linha: ', which.min(cv.results[, 9]),' Resultado: ',
cv.results[which.min(cv.results[, 9]), 9]))

   }

  } # GRID: max_depth / min_child_weight

{

   metric <- which.min(cv.results[,9]) # alterar dependo do objetivo

max.depth.f <- cv.results[metric,2]

   min.child.weight.f <- cv.results[metric,3]

} # BACKUP GRID max_depth / min_child_weight

  print(paste0('Parcial Time: ',Sys.time()))


  for(subsamples in subsampless){

   for(colsample_bytree in colsample.bytree){
```

```
    mcv.3 = xgb.cv(params = params,
                   data = cv.train,
                   nrounds = 1000,
                   early_stopping_rounds = 10,
                   nfold = 10,
                   verbose = 0,
                   max_depth = max.depth.f,
                   min_child_weight = min.child.weight.f,
                   subsamples = subsamples,
                   colsample_bytree = colsample_bytree)


    bint <- which.min(mcv.3$evaluation_log$test_rmse_mean) # modificar por tipo do
modelo
    metric <- min(mcv.3$evaluation_log$test_rmse_mean) # modificar (max ou min) por tipo
do modelo


    cv.3 <- matrix(nrow = 1, ncol = 9)
    cv.3[,1] = bint
    cv.3[,2] = max.depth.f
    cv.3[,3] = min.child.weight.f
    cv.3[,4] = 0
    cv.3[,5] = subsamples
    cv.3[,6] = colsample_bytree
    cv.3[,7] = 0
    cv.3[,8] = 0.3
    cv.3[,9] = metric
    cv.results <- rbind(cv.results,cv.3)
    print(cv.results)
    print(paste0('Linha: ', which.min(cv.results[, 9]), ' Resultado: ',
cv.results[which.min(cv.results[, 9]), 9]))
```

```
    }
} # GRID: subsamples / colsample_bytree
  {
    metric <- which.min(cv.results[,9])
    max.depth.f <- cv.results[metric,2]
    min.child.weight.f <- cv.results[metric,3]
    gamma.f <- cv.results[metric,4]
    subsamples.f <- cv.results[metric,5]
    colsample.bytree.f <- cv.results[metric,6]
} # BACKUP GRID subsamples / colsample_bytree
  print(paste0('Parcial Time: ',Sys.time()))


  for(eta in etas){
    mcv.5 = xgb.cv(params = params,
                   data = cv.train,
                   nrounds = 4000,
                   early_stopping_rounds = 10,
                   nfold = 10,
                   verbose = 0,
                   max_depth = max.depth.f,
                   min_child_weight = min.child.weight.f,
subsamples = subsamples.f,
                   colsample_bytree = colsample.bytree.f,
                   eta = eta)


bint <- which.min(mcv.5$evaluation_log$test_rmse_mean) # modificar por tipo do modelo
    metric <- min(mcv.5$evaluation_log$test_rmse_mean) # modificar (max ou min) por tipo
do modelo


    cv.5 <- matrix(nrow = 1, ncol = 9)
```

```
    cv.5[,1] = bint
    cv.5[,2] = max.depth.f
    cv.5[,3] = min.child.weight.f
    cv.5[,4] = 0
    cv.5[,5] = subsamples.f
    cv.5[,6] = colsample.bytree.f
    cv.5[,7] = 0
    cv.5[,8] = eta
    cv.5[,9] = metric
    cv.results <- rbind(cv.results,cv.5)
    print(cv.results)
    print(paste0('Linha: ', which.min(cv.results[, 9]), ' Resultado: ',
cv.results[which.min(cv.results[, 9]), 9]))
  } # GRID: etas
  {
    metric <- which.min(cv.results[,9])
    max.depth.f <- cv.results[metric,2]
    min.child.weight.f <- cv.results[metric,3]
    subsamples.f <- cv.results[metric,5]
    colsample.bytree.f <- cv.results[metric,6]
    eta.f <- cv.results[metric,8]
    nruns <- cv.results[metric,1]
} # BACKUP GRID FINAL
  print(paste0('Parcial Time: ',Sys.time()))
}


  TimeF <- Sys.time()
  print(paste0('Initial Time: ',TimeI))
  print(paste0('Final Time: ',TimeF))
} # Cross-validation XGboost
```

```
{
  class = xgboost(params = params,
                  data = cv.train,
                  nrounds = nruns,
                  max_depth = max.depth.f,
                  min_child_weight = min.child.weight.f,
subsamples = subsamples.f,
                  colsample_bytree = colsample.bytree.f,
                  eta = eta.f,
                  verbose = 0)

pred = predict(class, newdata = as.matrix(test[,-c(1,2,4,6,7,8,9)]))
  rank <- cbind(pred, test)
  rank.f <- rbind(rank,rank.f)

} # XGBoost predict opperations

{
  setwd ("C:/Users/glauc/Desktop/PIBIC final")
  write.table(cv.results, paste0('cv.',i,'.txt'), sep=';')
  write.table(rank, paste0('rank.test.',i,'.txt'), sep=';')
  write.table(rank.f, 'rank.test.0.txt', sep=';')
} # Save results
} # Modelo completo para testes
```