O PODER PREDITIVO DOS MODELOS *BOOSTING* DE *MACHINE LEARNING* NO MERCADO BRASILEIRO DE AÇÕES

THE PREDICTION ABILITY OF MACHINE LEARNING BOOSTING MODELS IN THE BRAZILIAN STOCK MARKET

José Erasmo Silva
Doutor em administração
MBA USP/Esalq
e-mail: jose.erasmo@natelcontact.com.br

Renato Máximo Sátiro
Doutor em administração
Universidade Federal de Goiás (UFG)
e-mail: renato_maximo_satiro@ufg.br

Resumo:
O mercado de ações está entre os pilares mais abrangentes e importantes da economia de qualquer país, pois desempenha um papel crucial em seu processo de crescimento, alimentando, ao longo do tempo, a atividade econômica nacional. A previsão do retorno de ações é uma área relevante de pesquisa e tem atraído muita atenção de pesquisadores e investidores em virtude dos potenciais benefícios monetários decorrentes dessa previsão. O objetivo deste estudo foi verificar o poder preditivo dos algoritmos *CatBoost*, *gradient boosting*, *AdaBoost*, *LightGBM* e *XGBoost* na previsão de retornos mensais de ações no mercado brasileiro. Os modelos foram treinados e avaliados no período de 2017 a 2021. Os resultados apontam, sob a perspectiva da maioria das métricas, que o *CatBoost* se destacou como mais eficiente, sendo um dos mais promissores e sofisticados entre os algoritmos apresentados, derivado das árvores de decisão.

Palavras-chave: mercado de ações; modelos de previsão; *CatBoost*.

CONTABILOMETRIA

Resumo do processo editorial:
a) Submissão em: 21/07/2022.
b) Envio para avaliação em: 21/09/2022.
c) Término da avaliação em 27/09/2022.
d) Correções solicitadas em: 27/09/2022.
e) Recebimento da versão ajustada em: 21/10/2022.
f) Aprovação final em: 25/10/2022.

Abstract:
The stock market is among the most comprehensive and important pillars of any country's economy for its crucial role in their growth process over time feeding the economic labor force. Forecasting stock returns is a relevant research area that has increasingly attracted researchers and investors' attention due to its potential monetary benefits. This study aims to verify the prediction ability of the CatBoost, gradient boosting, AdaBoost, LightGBM, and XGBoost algorithms in forecasting monthly stock returns in the Brazilian market. The models were trained and assessed between 2017 and 2021, highlighting CatBoost as the most efficient from the perspective of most metrics and a very promising and sophisticated algorithms derived from decision trees.

Keywords: stock market; forecasting models; CatBoost.

1 Introdução

O mercado de capitais é visto como uma peça de grande importância no desenvolvimento dos países, pois contribui para as indústrias de maneira geral, as quais, por sua vez, fomentam a atividade econômica de cada país (GHASHAMI; KAMYAR, 2021).

Nesse sentido, há que se observar que a previsão do retorno de ações é um campo fundamental de pesquisas, o qual tem atraído atenção considerável de pesquisadores e

investidores devido aos seus potenciais de benefícios monetários. O desenvolvimento de um modelo adequado e a seleção de características representativas do fenômeno em estudo são partes fundamentais de uma previsão acurada (HAQ *et al*., 2021).

Como um dos indicadores da importância da temática e do seu crescimento, observa-se que o número de investidores na bolsa de valores brasileira tem crescido de forma vigorosa, tendo atingido mais de 5 milhões de CPFs registrados, segundo dados da B3 (2022). Ao mesmo tempo, nota-se um incremento nas publicações de artigos na área de finanças utilizando abordagens de aprendizado de máquina e inteligência artificial (KOLANI, 2022). Esses tipos de abordagem permitem capturar comportamentos lineares e, especialmente, os não lineares, que são mais comuns no mercado acionário (AHMED *et al*., 2022).

Nessa linha, ressalta-se que, atualmente, o mercado de ações tem se tornado um dos principais campos de investimento. Ferramentas de aprendizado de máquina e inteligência artificial têm despertado o interesse tanto de estudiosos quanto de investidores. Nesse contexto, previsões acuradas para o movimento das ações no mercado podem reduzir riscos e gerar retornos abundantes (JI *et al*., 2022).

Ainda, é importante observar que muitos são os fatores que influenciam o preço das ações, podendo levar ao aumento ou à diminuição dos negociadores, incluindo oferta e demanda, tendências de mercado, economia local e global, resultado das empresas, preços históricos, notícias em geral (positivas ou negativas), informações financeiras confidenciais e popularidade da empresa(HO; DARMAN; MUSA, 2021; SRIVINAY *et al*., 2022).

Guerard, Xu e Wang. (2019) destacam a necessidade de uma avaliação constante dos modelos de investimento, pois ninguém pode testar somente um algoritmo ou método e investir dinheiro a partir deles. Dessa forma, é importante que se desenvolvam novos modelos e se reavaliem os antigos, tendo em vista que o comportamento do mercado está em contínua transformação, principalmente em virtude de novos investidores, empresas, governos e tecnologias.

Com base nesse cenário, este trabalho tem o objetivo de proporcionar a pequenos investidores *insights* sobre quais ferramentas de *machine learning* poderiam ser utilizadas na análise de ações. Consideram-se a necessidade de avaliação e de criação de modelos de investimento e o crescente interesse em investimento em ações, dada a possibilidade de maiores retornos desde que se assumam maiores riscos. Nesse sentido, este estudo tem como objetivo verificar o poder preditivo dos principais e mais atuais algoritmos de *machine learning* nos retornos mensais no mercado de ações brasileiro.

2 Fundamentação Teórica

A previsibilidade dos retornos no mercado de ações é um dos mais controversos assuntos em finanças. Apesar de economistas, estatísticos, investidores e pesquisadores se preocuparem em melhorar ou criar novos modelos de previsão, ainda não há consenso entre eles se isso é possível (ELMSILI; OUTTAJ, 2021).

Ainda nesse sentido, observa-se que a ideia da imprevisibilidade do mercado de ações é suportada pela hipótese dos mercados eficientes, que é um dos pilares da teoria clássica de finanças, proposta por Fama em 1970 (FAMA, 1998). Tal hipótese assume que os preços de uma ação incorporam todas as informações disponíveis e são baseados nas expectativas racionais dos investidores que buscam aumentar seus lucros. No entanto, no mundo real, investidores usam diferentes estratégias para tomadas de decisão e interpretam de forma heterogênea as informações disponíveis (ABDELAZIZ; MRAD, 2021).

Menciona-se ainda que as abordagens de previsão do retorno de ações são categorizadas em análise fundamentalista e análise técnica de acordo com o tipo de informação que cada uma delas utiliza. Na análise fundamentalista, investidores estimam o preço intrínseco de uma ação

examinando vendas, lucros, dívidas e dividendos da empresa. Por outro lado, a análise técnica avalia as ações observando tendências e estatísticas geradas pela atividade do mercado, como preços históricos e volume de negócios (HAQ *et al*., 2021; RAJKAR *et al*., 2021; XU *et al*., 2021).

Nesse sentido, a previsão de retorno pode ainda ser vista sob outra perspectiva, em que são empregados métodos de *machine learning* no processo de análise e de tomada de decisão, os quais utilizam algoritmos, entre eles: *Support Vector Machines* (SVM), *Recurrent Neural Network* (RNN) e *Extreme Learning Machine* (EML), desenvolvidos por cientistas da computação e visando avaliar e conduzir a descoberta de informações em larga escala em um curto período (RAJKAR *et al*., 2021). Vale destacar que a maioria desses algoritmos utiliza como base as informações de preços dos ativos e a análise técnica (XU *et al*., 2021).

Com o rápido desenvolvimento do *machine learning*, tais algoritmos vêm sendo amplamente utilizados na solução de vários problemas práticos das indústrias. Mais recentemente, pesquisadores têm se dedicado a estudar e a aplicar esses mesmos algoritmos para previsões no mercado de ações, o que tem proporcionado melhores investimentos (LIU *et al*., 2021).

Apesar de não ser um assunto novo, os sistemas baseados em *machine learning* vêm despertando cada vez mais interesse de tomadores de decisão nos últimos anos devido à sua habilidade de entregar o estado da arte em uma variedade de domínios, por exemplo, visão de computador, entendimento de linguagem natural e reconhecimento de fala. Assim como em outras áreas, na literatura financeira, apesar da crença dos mercados eficientes, essas ferramentas também têm sido testadas (ELMSILI; OUTTAJ, 2021).

3 Procedimentos Metodológicos

Os dados para o estudo foram coletados na plataforma Economatica e processados para obtenção dos resultados utilizando a linguagem Python. Foram utilizadas como amostra as informações mensais ajustadas dos preços de fechamento de todas as empresas listadas na B3 referentes ao período de 2017 a 2021. Como critério de filtragem de ações de baixa liquidez, selecionaram-se somente as ações com negociação em todos os dias no período de treino e teste. Nesses moldes, 250 empresas foram testadas nos modelos, gerando 14.640 registros. Ressalta-se que nem todas as empresas estiveram em todos os períodos de treino e teste.

Depois de definida a amostra, é muito comum a discussão sobre eliminar ou não os *outliers*. No contexto deste trabalho, optou-se por eliminá-los quando estivessem fora do patamar comum do mercado, geralmente se caracterizando por erros. Desse modo, os retornos das ações foram classificados em percentis e foram eliminadas as ações cujos retornos se enquadrassem nos percentis 1 e 99.

Após a eliminação dos *outliers*, foi executado o processo de engenharia de características (*feature engineer*), o qual consiste em criar os indicadores que serão utilizados como variáveis explicativas no modelo a partir das informações coletadas das quatro cotações (de abertura, máxima, mínima e de fechamento) e do volume e de negócios. A acurácia nas previsões é função principalmente da seleção de indicadores representativos e do desenvolvimento de um modelo apropriado (HAQ *et al*., 2021). A Tabela 1 apresenta as referidas variáveis.

Após o tratamento dos dados e a criação dos indicadores técnicos, foram utilizadas as abordagens *boosting* em árvores de decisão para atingir o objetivo deste estudo, isto é, avaliar o uso dessas abordagens na previsão de retornos mensais positivos no mercado de ações brasileiro. Para representar a variável dependente, os retornos foram transformados em uma variável binária que recebeu 1 para retorno positivo e 0 para retorno negativo.

Tabela 1 - Variáveis explicativas

Variável	Descrição
Retorno	Retorno de 1, 2, 3, 6, 9 e 12 meses
Retornos defasados	Referentes ao 1º, 2º, 3º, 4º, 5º e 6º mês
Momentum	Referente a 2, 3, 6, 9 e 12 meses
Ano	Ano referente aos dados
Mês	Mês referente aos dados

Fonte: elaborada pelos autores (2022)

Árvore de decisão é o modelo de classificação mais utilizado em conjuntos de aprendizagem. A ideia básica dessa ferramenta é a seguinte: primeiramente, em cada nó, a árvore é baseada em múltiplas regras de recursos, e a seleção de variáveis e de valores de variáveis é realizada de acordo com o desempenho do efeito de classificação. Na sequência, com base nas variáveis selecionadas, após dividir a área dos dados, compara-se o efeito da classificação e a complexidade do modelo, determina-se o método de divisão ideal, divide-se o próximo nó da folha até que o nó não possa mais ser dividido e tomam-se decisões de classificação (CHEN *et al.*, 2020). Apesar de as árvores de decisão ainda serem utilizadas em alguns estudos, elas têm caído em desuso devido ao surgimento de modelos que combinam essas árvores para obter resultados mais consistentes, são os chamados *ensemble models*.

No contexto do *ensemble*, parte-se da premissa de que os resultados em conjunto são melhores que os resultados individuais. Apesar de envolver alguns perigos no sentido de se propagarem erros por meio de vários modelos, de maneira geral, os modelos *ensemble* apresentam melhores resultados que os modelos individuais (NABIPOUR *et al*., 2020).

Entre os *ensembles* mais famosos estão os *baggings* e os *boostings*. A principal diferença entre eles é que os *baggings* trabalham com um conjunto de árvores em paralelo e o resultado geral é dado pela média dos resultados das árvores. Os modelos *boosting* trabalham com árvores sequenciais nas quais o objetivo de cada uma delas é melhorar o resultado da árvore anterior. Tendo em vista que os modelos *boosting* têm uma amplitude maior de modelos, bem como têm apresentado melhores resultados e frequentes implementações, optou-se neste estudo por explorar somente tal grupo.

O método *boosting* é um dos mais poderosos conceitos de aprendizado de máquina introduzidas nos últimos 30 anos. O primeiro modelo de *boosting* foi desenvolvido por Robert Schapire e Yoav Freund por volta de 1990 (JANSEN, 2020). A ideia principal se baseia na possibilidade de se construir um modelo "forte" a partir da combinação de vários modelos "fracos" (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Ou seja, a motivação por trás do *boosting* foi encontrar um método que combina os resultados de muitos modelos fracos, em que "fraco" significa um desempenho apenas ligeiramente melhor que um palpite aleatório em uma previsão conjunta altamente precisa (SCHAPIRE; FREUND, 2013).

De maneira geral, o *boosting* aprende uma hipótese aditiva, H_M, de forma similar com uma regressão linear. No entanto, cada um dos m=1,...,M elementos de uma somatória é um aprendiz fraco, chamado de h_t, que por si só requer treinamento (JANSEN, 2020). A equação (1) sumariza essa abordagem.

$$H_M(x)=\sum_{m=1}^{M} h_t(x) \qquad (1)$$

em que: $h_t(x)$: são os aprendizes fracos.

A partir desse conceito central, outros algoritmos surgiram como extensões, sendo que o mais popular é chamado de *adaptive boosting*, ou AdaBoost, e foi criado por Freund e Schapire (1997). É um algoritmo iterativo que treina diferentes subclassificadores para o mesmo conjunto de dados, de modo específico, treina continuamente os dados classificados erroneamente e, então, serializa esses classificadores fracos para formar um classificador forte (CHEN *et al*., 2020; HASTIE *et al*., 2009). Um dos pontos a se destacar a respeito desse modelo é que ele faz ponderações nos erros e acertos de forma que os erros sejam mais trabalhados nas etapas posteriores do algoritmo.

Em 1999, Jerome Friedman criou o *gradient boosting*, modelo que tem produzido o estado da arte em termos de desempenho tanto em classificações quanto em regressões. Ele é provavelmente o mais popular em competições de *machine learning*, além de ser utilizado em aplicações do mercado, por exemplo, para prever taxas de cliques em anúncios *on-line* (JANSEN, 2020).

O *gradient boosting* aplica uma abordagem diferente daquela do AdaBoost. Nesse sentido, também se ajusta baseado nas previsões incorretas, porém, leva essa ideia a um passo adiante, ou seja, ajusta-se a cada nova árvore inteiramente fundamentada nos erros das previsões da árvore anterior. Em outras palavras, para cada árvore, ele analisa os erros e constrói uma nova árvore completamente em torno desses erros, a qual não leva em consideração as previsões que já estão corretas (WADE, 2020).

Nos últimos anos, muitas implementações de *gradient boosting* passaram a utilizar inovações que aceleram o treinamento dos modelos ou dos algoritmos, melhoram a eficiência dos recursos e permitem que o algoritmo seja dimensionado para conjuntos de dados muito grandes. Entre elas, estão os algoritmos *Extreme Gradient Boosting* (*XGBoost*), *Light Gradient Boosting Machine* (*LightGBM*) e *Categorical Boosting* (CatBoost) (JANSEN, 2020).

O *XGBoost* foi iniciado em 2014 por Tianqi Chen em seu processo de doutorado e ganhou muita popularidade: entre as 29 soluções vencedoras publicadas pelo *Kaggle* – plataforma em que se pode participar de competições de algoritmos e também trocar ideias, códigos e compartilhar dados –, em 2015, 17 usaram *XGBoost*, sendo que 8 delas o empregaram isoladamente, e o restante combinou com outros algoritmos (CHEN; GUESTRIN, 2016). Atualmente, ele tem sido utilizado em várias áreas, tal como energia, assistência médica e *score* de crédito (JABEUR; MEFTEH-WALI; VIVIANI, 2021; MO *et al*., 2019). A fórmula de saída é calculada com a equação. (2):

$$\text{yhat}_i^T = \sum_{k=1}^{T} f_k(x_i) = \text{yhat}_i^{T-1} + f_T(x_i) \qquad (2)$$

em que, yhat_i^{T-1}: é a árvore de decisão gerada; $f_T(x_i)$: é o modelo de árvore recém-criado; e T: é o total de árvores no modelo. De maneira geral, as principais vantagens do *XGBoost* em relação as implementações anteriores são: 1) Possibilidade de utilização de computação paralela, 2) Utilização de matrizes esparsas, 3) Alocação dos dados em blocos.

O *LightGBM* foi criado pela *Microsoft* e liberado para uso público em janeiro de 2017 (KE *et al*., 2017). Estudos mostraram que esse algoritmo é mais eficiente e possui maior acurácia, principalmente no sentido de consumo de memória e velocidade de treinamento e é um dos mais avançados algoritmos de elevação gradiente. (JABEUR; MEFTEH-WALI; VIVIANI, 2021). Ele usa os seguintes métodos: aprimoramento gradiente integrado para melhorar a robustez do classificador, árvore de decisão com profundidade limitada para evitar o ajuste excessivo, amostragem unilateral baseada em gradiente para representar a importância das instâncias de dados e algoritmo de histograma para encontrar a melhor divisão (YIN *et al*.,

2021). Segundo Jabeur, Mefteh-Wali e Viviani, (2021), a função estimada do *LightGBM* integra várias árvores de regressão T e é definida da seguinte forma (equação 3):

$$Y_t = \sum_{h=1}^{T} f_t(x) \quad (3)$$

em que, $f_t(x)$: denota a árvore de regressão. No *LightGBM*, o método de Newton foi usado para estimar a função objetivo. As principais inovações dessa implementação em relação aos anteriores são: 1) Usa um método chamado *gradiente-based one-side sampling* para excluir uma parte significativa da amostra que é pouco significativa para os resultados, 2) Usa um pacote de recursos exclusivos para combinar recursos que são mutuamente exclusivos, 3) Prioriza as quebras das árvores de maneira diferenciada.

Por fim, o *CatBoost* foi criado pela empresa *Yandex* e liberado para uso público em abril de 2017 (PROKHORENKOVA *et al*., 2018). Assim como os modelos anteriores, esse algoritmo também utiliza o *gradient boosting* em árvores de decisão. Uma das vantagens que o *CatBoost* e o *LightGBM* têm sobre os demais modelos é que não há necessidade de transformar as variáveis categóricas em *dummies*, pois ambos lidam com elas diretamente. Além disso, o *CatBoost* pode combinar diferentes categorias para criar novas categorias (JANSEN, 2020).

Ainda a respeito do *CatBoost*, ele resolveu um problema conhecido como *missing target*, que é encontrado nos algoritmos *XGBoost* e *LightGBM*. Tal problema faz com que as previsões feitas por esses modelos dependam de todos os *targets* utilizados no treinamento e, caso haja mudanças na amostra, os resultados das previsões ficam comprometidos (XU *et al*., 2021). Ademais, o *CatBoost* tem uma versão que utiliza G*raphics Processing Units* (GPU) e aumenta significativamente a velocidade de treinamento. De acordo com Jabeur, Mefteh-Wali e Viviani (2021), a função h da árvore de decisão pode ser escrita a partir da equação (4):

$$h^t \arg\min \frac{1}{N} \sum \left(-f^t(X_k, Y_k) - h(X_k)\right)^2 \quad (4)$$

em que, X_k: é um vetor aleatório de N variáveis de entrada, Y_k : é a saída e f: é uma aproximação de mínimos quadrados pelo método de Newton. A Tabela 2 apresenta os hiperparâmetros definidos no modelo LightGBM.

Tabela 2 - Hiperparâmetros utilizados com o modelo *LightGBM*

Parâmetro	**Descrição**
***Boosting_type*='gbdt'**	
***Objective*='*binary*'**	# Tarefa de aprendizagem
***Metric*='auc'**	
Num_*leaves*=31	# Máximo de folhas por árvore
Max_*depth*=-1	# Profundidade máxima da árvore, -1 significa sem limite
***Learning_rate*=0.1**	# Taxa de aprendizagem
N_*estimators*=100	# Número de árvores
***Subsample_for_bin*=200000**	# Número de amostras para construção de caixas
Class_weight=None	# dicionário, "*balanced*" ou "*none*"
Min_*split_gain*=0.0	# Redução de perda mínima para redução adicional
Min_*child_weight*=0.001	# Soma mínima do peso da instância
Min_*child_samples*=20	# Número mínimo de dados necessários em uma folha filha
***Subsample*=1.0**	# Proporção de subamostras de amostras de treinamento
***Subsample*_freq=0**	# Frequência de subamostragem, <=0: desabilitado

***Colsample_bytree*=1.0**	# Proporção de subamostragem de recursos
Reg_*alpha*=0.0	# Prazo de regularização L1 em pesos
Reg_*lambda*=0.0	# Prazo de regularização L2 em pesos
***Random_state*=42**	# Número aleatório semente; padrão: C++ semente
N_*jobs*=-1	# Número de "threads" paralelo. -1 significa sem limite
Silent=False	
***Importance_type*='gain'**	# padrão: *'split*' ou '*gain*'

Fonte: Jansen (2020)

Para Sun *et al*. (2020), a acurácia dos modelos depende fortemente dos hiperparâmetros selecionados para cada um deles. Portanto, antes de executar cada modelo, devem-se definir quais configurações de hiperparâmetros melhor contribuem para os resultados. Vale evidenciar que os hiperparâmetros são diferentes entre os modelos e que a documentação de cada um deles deve ser consultada para melhor adequação. Destaca-se ainda que há bibliotecas, por exemplo, a scikit-learn, que permitem definir e testar vários intervalos de hiperparâmetros automaticamente. Esse processo de testar vários hiperparâmetros para ver qual gera melhores resultados é chamado de grid search (MEHTAB; SEN, 2020).

Para treinar e testar os modelos, criou-se um esquema de treino e teste (*cross validation*) da seguinte maneira: treinaram-se 12 meses e o modelo foi testado no mês subsequente. Posteriormente, em um mês, o período de treino foi incrementado e testou-se o mês subsequente, sempre mantendo 12 meses como treino e um mês como teste. A Figura 1 exemplifica o esquema dos períodos de treino e teste.

Figura 1 - Exemplificação dos períodos de treino e teste

Fonte: elaborada pelos autores (2022)

Os desempenhos desses modelos podem ser avaliados com base em várias métricas e é importante que elas sejam selecionadas de forma que melhor se adaptem à tarefa de previsão e à estrutura da variável resposta.

Em tarefas de regressão, diferentes métricas de desempenho podem ser empregadas, sendo que as mais comuns são *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE) e R^2. Para as tarefas de classificação utilizadas neste estudo, o método mais direto é comparar os resultados reais com os resultados previstos por meio de uma matriz de 2x2 chamada de matriz de confusão. A partir dela, muitas métricas de avaliação dos modelos podem ser geradas (CONSOLI; RECUPERO; SAISANA, 2021). A Tabela 3 mostra a estrutura dessa matriz.

Para os modelos de previsão que foram testados, a variável resposta é binária. Nesse sentido, não se desejava saber qual o retorno de uma ação para o próximo mês, mas sim se ele seria positivo ou negativo. Desse modo, as variáveis binárias foram criadas da seguinte forma: as ações com retornos maiores que zero receberam o valor 1 (um), senão o valor 0 (zero). Os retornos considerados nesse caso foram sempre referentes ao primeiro mês subsequente aos

dados utilizados. Exemplo: para os indicadores criados com informações até 12/2021, utilizou-se o retorno do mês 01/2022 para a criação da variável dependente.

Tabela 3 - Matriz de confusão

		Valor real	
		1	0
Valor previsto	1	TP	FP
	0	FN	TN

Fonte: elaborado pelos autores (2022)

Depois de rodar os modelos, as ações recebem classificações. Assim, as que foram classificadas como retornos positivos e de fato apresentaram valor positivo são chamadas de True Positive (TP). As ações que foram classificadas como retornos positivos, mas apresentaram retornos negativos são chamadas de False Positive (FP). As ações classificadas como retorno negativo e apresentaram retornos positivos são chamadas de False Negative (FN). Por fim, as ações classificadas como retornos negativos e que de fato obtiveram resultados negativos são chamadas de True Negative (TN). A partir dessas denominações, os seguintes indicadores puderam ser construídos por meio das equações (5) a (8):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = Sensitivity = \frac{TP}{TP+FN} \tag{7}$$

$$\text{F1 } Score = 2\frac{Precision \text{ x } Recall}{Precision \ + Recall} \tag{8}$$

A *accuracy* (acurácia) mostra o percentual de previsões acertadas em meio a todas as previsões, ou seja, tanto eventos quanto não eventos. Já a *precision* (precisão) apresenta quais as previsões de eventos acertadas entre as precisões feitas como evento. Essas duas medidas são muito interessantes para serem utilizadas no mercado acionário, sendo que a primeira é utilizada tanto para avaliar posições *long* (compradas) quanto posições *short* (vendidas a descoberto), já a segunda, somente para avaliação de posições long. O *recall* (sensitividade) é o indicador que mostra o percentual de acertos quando considerada a soma dos verdadeiros positivos com os falsos negativos. Por fim, F1 score é uma média harmônica criada a partir das métricas *precision* e recall.

Além das métricas citadas acima, que são geradas a partir da matriz de confusão, outras duas foram avaliadas: a função *log loss* e a *Area the Under the ROC Curve* (AUC). A função log loss tem o objetivo de avaliar se o modelo faz previsões boas ou ruins. Ela mensura o desempenho de um modelo de classificação cuja saída é uma probabilidade. Apesar de ainda não haver um consenso entre os pesquisadores, segundo Prado (2018), essa é uma das melhores métricas para se utilizar no mercado financeiro. Quanto mais divergentes forem os valores previstos dos valores atuais, maior será o valor da log loss, isto é, para essa métrica, quanto menor o valor, melhores são as previsões. De acordo com Ghatak (2019), ela pode ser calculada a partir da equação (9).

$$H(y,y\_hat)= \sum_i y_i \log \frac{1}{y\_hat_i} - \sum_i y_i \log y\_hat_i \quad (9)$$

Para entender o conceito de curva ROC e o de AUC, é necessário conhecer o conceito de especificidade. Segundo Fávero e Belfiore (2017), a especificidade se refere ao percentual de acerto, para um dado "cutoff", considerando-se apenas as observações que não são eventos. Utilizando-se a matriz de confusão, a especificidade (specificity) pode ser calculada utilizando a equação (10):

$$Specificity = \frac{TN}{TN + FP} \quad (10)$$

em que: TN significa true negative e FP significa false positive, da mesma forma que foram abordados nas equações (5) a (9).

A curva ROC mostra o comportamento propriamente dito do trade off entre a sensitividade e a especificidade, e ao trazer no eixo das abscissas os valores de (1-especificidade), apresenta formato convexo em relação ao ponto (0,1). Nesse sentido, um determinado modelo com maior área abaixo da curva ROC apresenta maior eficiência global de previsão (FÁVERO; BELFIORE, 2017). Na próxima seção, apresentam-se os resultados da pesquisa.

4 Resultados e Discussão

Na Tabela 4, é possível observar a análise descritiva das variáveis explicativas utilizadas nos modelos. Observa-se, por exemplo, que foram trabalhados cinco anos (de 2017 a 2021), correspondentes a 60 meses, e que as empresas estão distribuídas em 18 setores em conformidade com a classificação North American Industry Classification System (NAICS).

A Tabela 5 mostra como estão classificadas as informações na variável resposta. Conforme exposto na metodologia do trabalho, as variáveis binárias apresentadas nessa ilustração foram criadas a partir dos retornos das ações. Observa-se, por exemplo, que 53,3% dos retornos foram classificados como 0 e 46,7%, como 1, o que caracteriza uma base de dados relativamente balanceada.

Tabela 4 - Análise descritiva das variáveis explicativas

Variável	Quant.	Média	Desvio Padrão	Min.	25%	50%	75%	Max.
***Return*_1m**	14640	0,0177	0,1455	-0,3792	-0,0557	0,0000	0,0724	0,6157
***Return*_2m**	14640	0,0135	0,1066	-0,2855	-0,0404	0,0023	0,0612	0,4142
***Return*_3m**	14640	0,0118	0,0880	-0,2362	-0,0330	0,0050	0,0526	0,3228
***Return*_6m**	14640	0,0115	0,0597	-0,1623	-0,0203	0,0075	0,0407	0,2105
***Return*_9m**	14640	0,0127	0,0486	-0,1264	-0,0135	0,0094	0,0370	0,1709
***Return*_12m**	14640	0,0132	0,0419	-0,1042	-0,0100	0,0106	0,0353	0,1482
***Momentum*_2**	14640	-0,0042	0,1060	-0,8696	-0,0467	0,0000	0,0455	0,7934
***Momentum*_3**	14640	-0,0059	0,1233	-0,8519	-0,0578	0,0000	0,0560	0,7020
***Momentum*_6**	14640	-0,0062	0,1360	-0,7780	-0,0617	0,0024	0,0644	0,5897
***Momentum*_9**	14640	-0,0050	0,1391	-0,7422	-0,0609	0,0045	0,0654	0,5501
***Momentum*_12**	14640	-0,0045	0,1396	-0,7199	-0,0585	0,0052	0,0651	0,5274
***Momentum*_3_12**	14640	0,0014	0,0772	-0,4191	-0,0372	0,0019	0,0408	0,3844
***Return*_1m_t-1**	14616	0,0173	0,1453	-0,3792	-0,0562	0,0000	0,0714	0,6157

***Return*_1m_t-2**	14592	0,0169	0,1455	-0,3792	-0,0567	0,0000	0,0711	0,6157
***Return*_1m_t-3**	14568	0,0190	0,1454	-0,3792	-0,0543	0,0000	0,0729	0,6157
***Return*_1m_t-4**	14543	0,0199	0,1452	-0,3792	-0,0526	0,0000	0,0732	0,6157
***Return*_1m_t-5**	14516	0,0208	0,1455	-0,3792	-0,0517	0,0000	0,0738	0,6157
***Return*_1m_t-6**	14488	0,0242	0,1464	-0,3792	-0,0489	0,0000	0,0769	0,6157

Fonte: resultados da pesquisa (2022)

Tabela 5 - Distribuição das frequências da variável dependente

Variável Y	**Quant.**	**Perc.**
Retorno negativo ou zero (0)	7805	0.5331
Retorno positivo (1)	6835	0.4669
Total	14640	

Fonte: resultados da pesquisa (2022)

Por fim, finalizando essa parte introdutória dos dados, observa-se, por intermédio da Tabela 6, em quais setores estão enquadradas as empresas utilizadas na amostra. Destaca-se que a classificação NAICS é bastante ampla e que só uma pequena parte dos setores aparece na pesquisa.

Tabela 6 - Quantidade de empresas utilizadas na pesquisa de acordo com cada setor

Setor NAICS	**Quant.**
Administração de empresas e empreendimentos	9
Agricultura, pecuária, silvicultura, pesca e caça	3
Artes, entretenimento e recreação	1
Assistência médica e social	7
Comércio atacadista	4
Comércio varejista	11
Construção	18
Educação	5
Empresa de eletricidade, gás e água	38
Hotel e restaurante	1
Imobiliária e locadora de outros bens	11
Indústria manufatureira	82
Informação	10
Mineração, exploração de pedreiras e extração de petróleo e gás	6
Serviços de apoio a empresas e gerenciamento de resíduos e remediação	3
Serviços financeiros e seguros	27
Serviços profissionais, científicos e técnicos	3
Transporte e armazenamento	11

Fonte: resultados da pesquisa (2022)

Os resultados da pesquisa são mostrados de maneira resumida na Tabela 7, contendo somente o período de testes de cada métrica. O modelo criado a partir do algoritmo CatBoost apresentou melhores indicadores em praticamente todas as métricas. Conforme será comentado a seguir, notou-se certa similaridade entre as precisões no mercado brasileiro e estudos realizados em outros países. Apesar de a vantagem do CatBoost ser aparentemente pequena em relação aos outros, no mercado financeiro isso pode representar grandes valores.

Tabela 7 - Resumo dos resultados dos modelos no período de teste

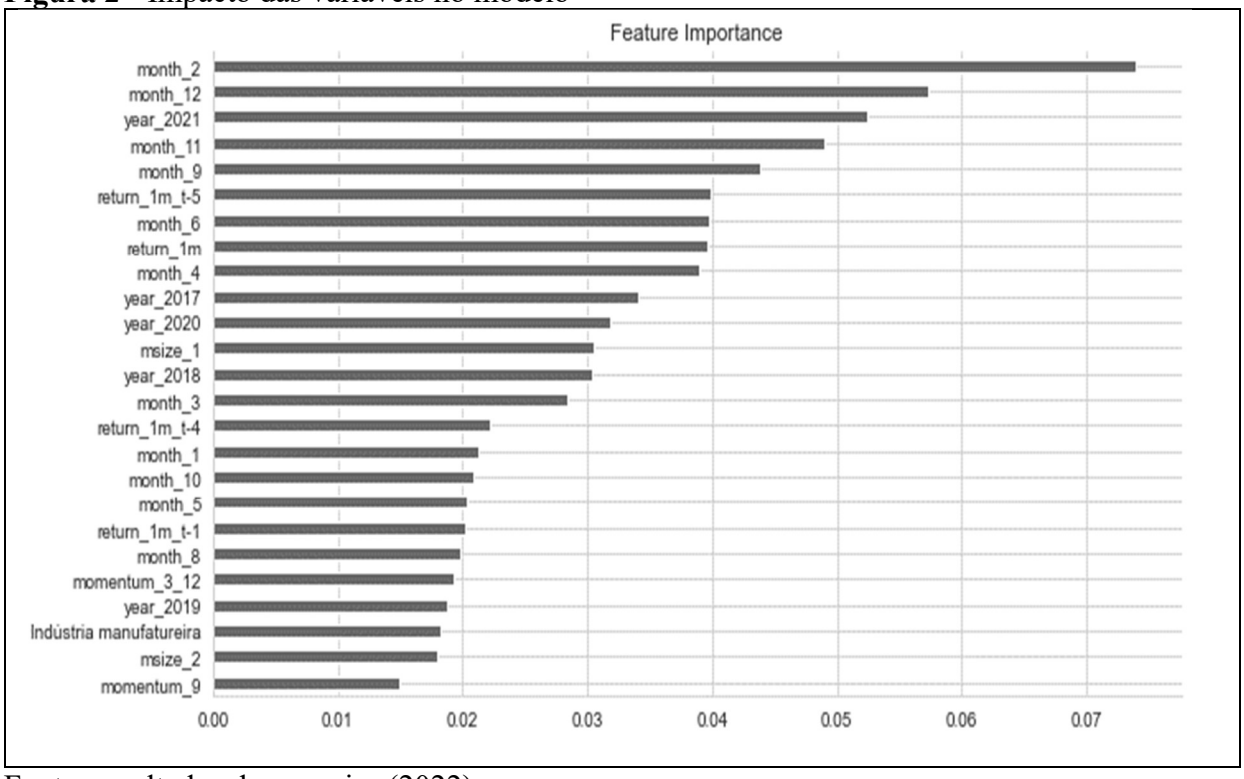
Métrica	AUC	*Accuracy*	F1	*Log loss*	*Precision*	*Recall*
CatBoost	0.565053	0.559442	0.548607	-0.69806	0.635065	0.535835
Gradient boosting	0.555262	0.512103	0.469639	-0.78902	0.587741	0.519400
AdaBoost	0.552201	0.509216	0.479508	-0.69271	0.627196	0.522487
LightGBM	0.551736	0.531373	0.467815	-0.74044	0.649184	0.497354
XGBoost	0.547973	0.533237	0.481243	-0.71442	0.618827	0.511785

Fonte: resultados da pesquisa (2022)

Além dos resultados das métricas apresentados na Tabela 7, há ainda a necessidade de se avaliar quais são as outras vantagens dos modelos, pois diferenças de performance podem comprometer o resultado de trabalhos que demandem retornos rápidos. No que se refere ao *LightGBM*, por exemplo, o modelo tem um nível de resultados ligeiramente inferior quando comparado aos outros modelos aqui empreendidos. No entanto, é muito mais rápido para convergir que os demais, uma vantagem significativa ao se trabalhar com um grande volume de dados, em especial quando se deseja avaliar intervalos de hiperparâmetros. De acordo com testes realizados em vários *datasets* públicos, ele se mostrou até 20 vezes mais rápido no treinamento que os modelos anteriores (KE *et al*., 2017).

Yin *et al*. (2021) utilizaram o modelo *LightGBM* para prever a tendência de quatro empresas listadas no mercado americano. Com informações diárias referentes ao período de janeiro de 2015 a outubro de 2020, os pesquisadores alcançaram um AUC de 0,87. O resultado encontrado pelos autores é bastante superior ao encontrado neste estudo, o que talvez se justifique pelo fato de os autores utilizarem empresas, escala de tempo e período diferentes. Tais divergências poderiam ser exploradas em estudos futuros.

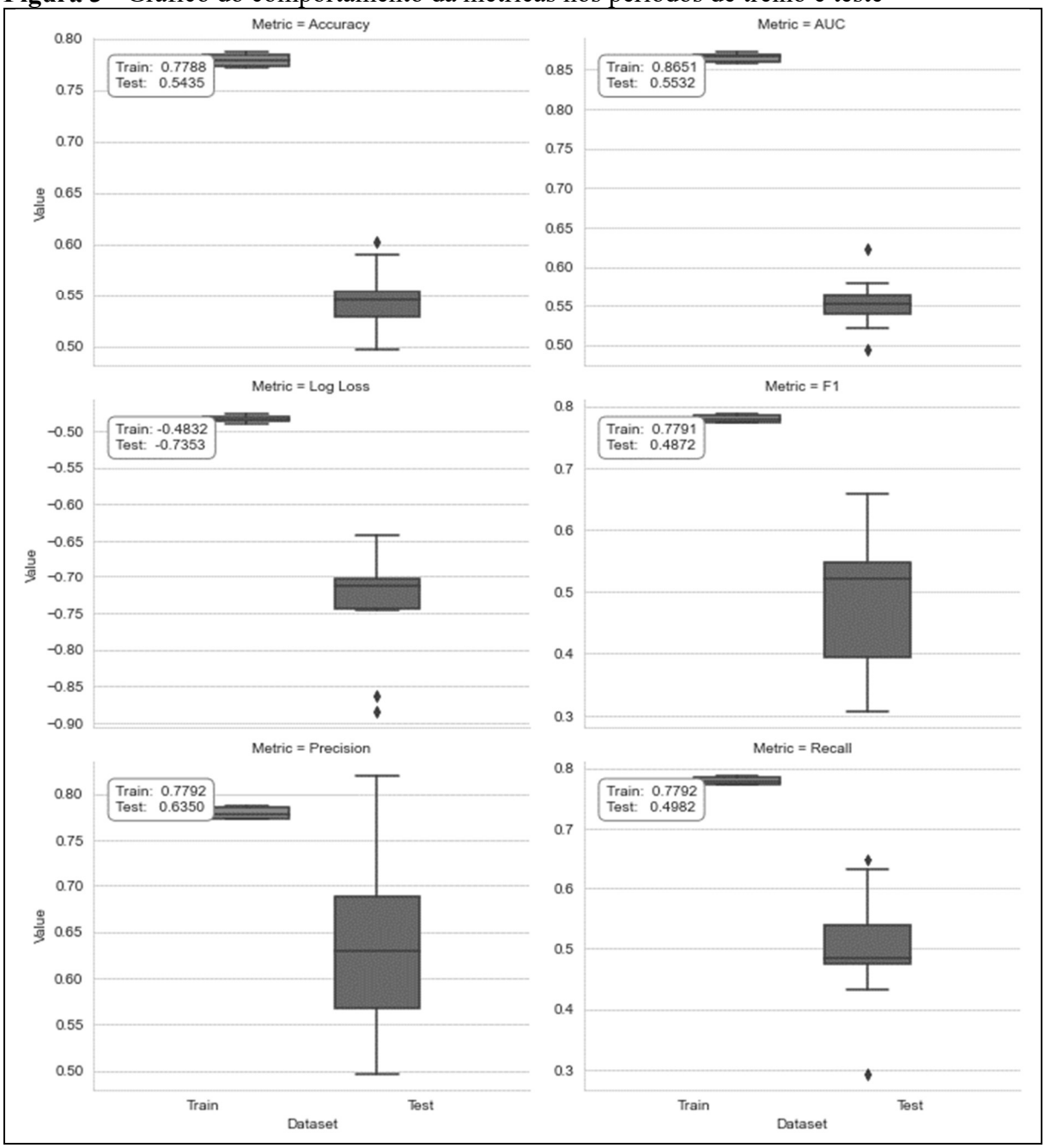
Figura 2 - Impacto das variáveis no modelo



Fonte: resultados da pesquisa (2022)

Xu *et al*. (2021) empregaram os modelos *XGBoost*, *LightGBM* e *CatBoost* para prever o comportamento dos seis maiores índices de ações do mercado chinês. Os autores fizeram previsões diárias do fechamento dos índices no período de 2009 a 2020, sendo 70% desse intervalo de tempo como treino e o restante como teste. As características que serviram de variáveis explicativas para os modelos foram criadas a partir das cotações de diferentes períodos, sendo a mínima de um minuto e a máxima de 60 dias. Dessa forma, os autores alcançaram a acurácia de 0,53 para o *XGBoost*, 0,52 para o Light GBM e 0,55 para o *CatBoost*. Apesar de não haver muitas similaridades entre o mercado chinês e o brasileiro, o resultado das previsões deste estudo foi muito parecido com o encontrado por Xu *et al*. (2021).

Figura 3 - Gráfico do comportamento da métricas nos períodos de treino e teste



Fonte: resultados da pesquisa (2022)

Jabeur, Mefteh-Wali e Viviani, (2021) testaram esses modelos no modo de regressão de maneira diferente da realizada neste trabalho, no qual se utilizou o modo classificação. O objetivo dos autores foi prever o retorno do ouro com base em indicadores de outras *commodities*. Para isso, eles coletaram informações mensais para o período de janeiro de 1986 a dezembro de 2019. Diferentemente dos achados deste estudo, o melhor modelo preditivo para os autores foi o *XGBoost*, seguido pelo *CatBoost*, *Random Forest* e *LightGBM*.

Sun *et al*. (2020) aplicaram os modelos *LightGBM*, *Random Forest* e SVM para a previsão do retorno de criptomoedas. Apesar de utilizarem uma amostra consideravelmente pequena (01/2018 a 06/2018), a acurácia foi promissora, a qual mostrou maior taxa de acerto para o *LightGBM*.

A Figura 2 demonstra que a característica que mais influenciou os retornos foi a variável categórica month_2 (referente ao mês de fevereiro), seguida de month_12 e year_2021 (representando o ano de 2021). A Figura 3 apresenta os resultados deste estudo sob uma perspectiva mais ampla. Nessa ilustração, é possível observar que os valores variam consideravelmente em torno das médias.

A Figura 3 mostra que a variação no período de treino é pequena, fator que já era esperado. Observa-se, ainda, que há grande variação nos indicadores nos períodos de teste. Compreende-se que três questões poderiam ser abordadas futuramente para minimizar essa variação e tornar os modelos mais estáveis e precisos. São elas: aumento do período de treino, um aumento do número de variáveis explicativas (ou mesmo um melhor refinamento na escolha de tais variáveis para a composição do modelo), bem como uma modificação da frequência de negócios para a escala diária.

Apesar de os resultados terem se apresentado muito próximos de um palpite aleatório, alguns autores têm defendido que quando se faz um grande volume de negócios utilizando um sistema de *High Frequency Trading* (HFT), uma acurácia entre 51% e 52% é suficiente para se obter lucros. Não é o caso deste trabalho, em que se testaram retornos na escala mensal.

Muitas companhias comerciais e instituições de pesquisa têm desenvolvido preditores de ações baseados em modelos estatísticos esperando obter informações mais regulares na análise dos dados históricos ou direcionamento no processo decisório (YIN *et al*., 2021).

Este estudo utilizou como amostra as ações listadas na B3. Dadas as características simples dos dados utilizados (preço de fechamento), esses algoritmos podem ser aplicados também em outros ativos, por exemplo, *Exchange Traded Fund* (ETF), títulos públicos, criptomoedas, dentre outros.

5 Considerações Finais

Considera-se que este estudo alcançou seu objetivo e mostrou os resultados da aplicação de diferentes modelos de *machine learning* no mercado de ações brasileiro. O modelo gerado a partir do algoritmo *CatBoost* foi o que apresentou melhores resultados em praticamente todas as métricas, especialmente no que diz respeito à acurácia e à AUC. Destaca-se que esse algoritmo é o mais novo entre os testados e representa o estado da arte em termos de algoritmos baseados em árvores de decisão.

A principal dificuldade encontrada no melhor modelo foi a necessidade de se utilizar (GPU) para a obtenção de um bom desempenho, dispositivo que ainda não é comum em computadores pessoais. Estima-se que em poucos anos isso não deverá ser um problema, haja vista a velocidade de desenvolvimento dos dispositivos de *hardware*. Nesse sentido, um processo que exige agilidade nas respostas demandará um alto poder de processamento computacional. Uma das limitações reside no fato de haver pouca literatura científica relativa ao tema, principalmente considerando o contexto brasileiro, poucos são os estudos que se utilizam do modelo *Catboost*, o que dificultou a comparabilidade com outros estudos nacionais.

Sugere-se que trabalhos futuros explorem outros indicadores técnicos como variáveis explicativas. Eles podem ser gerados a partir de diferentes pacotes tanto para a linguagem Python quanto para a R. Conforme destacado por Xu *et al*. (2021), o desempenho das previsões no mercado financeiro depende não somente dos algoritmos, mas também das características do mercado e das ações que foram inseridas no referido modelo. Sugerem-se, em especial, o cálculo e a inclusão dos cinco fatores de *Fama-French*, que são largamente estudados e melhoram o poder preditivo de alguns cenários.

Referências

ABDELAZIZ, Fouad Ben; MRAD, Fatma. Multiagent systems for modeling the information game in a financial market. **International Transactions in Operational Research**, Oxford, v. 0, p. 1 – 14, 2021. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/itor.12944. Acesso em: 12 jan.2022.

AHMED, Shamima; ALSHATER, Muneer M.; AMMARI, Anis El; HAMMAMI, Helmi. Artificial intelligence and machine learning in finance: a bibliometric review. **Research in International Business and Finance**, [*s. l.*], v. 61, p. 1 – 34, 2022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0275531922000344. Acesso em: 1 maio 2022.

B3. **Investidor pessoa física**: total de investidor pessoa física cresce 43% no primeiro semestre, mostra estudo da B3. São Paulo, 2021. Disponível em: https://www.b3.com.br/pt_br/noticias/porcentagem-de-investidores-pessoa-fisica-cresce-na-b3.htm. Acesso em: 1 abr. 2022.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost. *In*: KDD '16: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 2016, San Francisco, CA. **Anais** [...]. San Francisco, CA: ACM, 2016. p. 785 – 794. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939785. Acesso em: 10 jan. 2022.

CHEN, Yanyu; LI, Xuechen; SUN, Wei. Research on stock selection strategy based on adaboost algorithm. *In*: CSAE 2020: PROCEEDINGS OF THE 4TH INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND APPLICATION ENGINEERING 2020, Sanya. **Anais** [...]. Sanya: ACM, 2020. p. 1 – 5. Disponível em: https://dl.acm.org/doi/10.1145/3424978.3425084. Acesso em: 16 fev. 2022.

CONSOLI, Sergio; RECUPERO, Diego Reforgiato; SAISANA, Michaela. **Data science for economics and finance**. Cham: Springer International Publishing, 2021. *E-book* (357 p.). Disponível em: https://link.springer.com/10.1007/978-3-030-66891-4. Acesso em: 15 fev. 2022.

ELMSILI, Bilal; OUTTAJ, Benaceur. Predicting stock market movements using machine learning techniques. **International Journal of Accounting, Finance, Auditing, Management and Economics**, [*s. l.*], v. 2, n. 3, p. 390 – 405, 2021. Disponível em: http://ijafame.org/index.php/ijafame/article/view/171. Acesso em: 18 fev. 2022.

FAMA, Eugene F. Market efficiency, long-term returns, and behavioral finance. **Journal of Financial Economics**, [*s. l.*], v. 49, n. 3, p. 283 – 306, 1998. Disponível em:

https://www.sciencedirect.com/science/article/pii/S0304405X98000269. Acesso em: 16 fev. 2022.

FREUND, Yoav; SCHAPIRE, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, [*s. l.*], v. 55, p. 119 – 139, 1997. Disponível em: https://doi.org/10.1006/jcss.1997.1504. Acesso em: 15 jan. 2022.

GHASHAMI, Farnaz; KAMYAR, Kamyar. Performance evaluation of ANFIS and GA-ANFIS for predicting stock market indices. **International Journal of Economics and Finance**, [*s. l.*], v. 13, n. 7, p. 1 – 6, 2021. Disponível em: https://doi.org/10.5539/ijef.v13n7p1. Acesso em: 15 jan. 2022.

GHATAK, Abhijit. **Deep learning with r**. Singapore: Springer Singapore, 2019. *E-book* (245 p.). Disponível em: http://link.springer.com/10.1007/978-981-13-5850-0. Acesso em: 16 fev. 2022.

GUERARD, John B.; XU, Ganlin; WANG, Ziwei. **Portfolio and investment analysis with SAS: Financial Modeling Techniques for Optimization**. Cary: Inc., SAS Institute, 2019.

HAQ, Anwar Ul; ZEB, Adnan; LEI, Zhenfeng; ZHANG, Defu. Forecasting daily stock trend using multi-filter feature selection and deep learning. **Expert Systems with Applications**, [*s. l.*], v. 168, p. 1 – 9, 2021. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S095741742031099X. Acesso em: 17 fev. 2022.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning**. New York: Springer New York, 2009. *E-book* (745 p.). Disponível em: https://doi.org/10.1007/978-0-387-84858-7. Acesso em: 10 jan. 2022.

HO, M. K.; DARMAN, Hazlina; MUSA, Sarah. Stock price prediction using ARIMA, neural network and LSTM models. *In*: (Journal of Physics: Conference Series, Org.)SIMPOSIUM KEBANGSAAN SAINS MATEMATIK KE-28 (SKSM28) 2021, Kuantan. **Anais** [...]. Kuantan: IOP Publishing Ltd, 2021. p. 1 – 11. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1988/1/012041. Acesso em: 15 fev. 2022.

JABEUR, Sami Ben; MEFTEH-WALI, Salma; VIVIANI, Jean-Laurent. Forecasting gold price with the xgboost algorithm and SHAP interaction values. **Annals of Operations Research**, [*s. l.*], p. 1 – 21, 2021. Disponível em: https://doi.org/10.1007/s10479-021-04187-w. Acesso em: 5 mar. 2022.

JANSEN, S. **Machine Learning for Algorithmic Trading**: Predictive models to extract signals from market and alternative data for systematic trading strategies with python. 2. ed. Birmingham: Packt Publishing, 2020. 821 p. ISBN 978-1-83921-771-5.

JI, Gang; YU, Jingmin; HU, Kai; XIE, Jie; JI, Xunsheng. An adaptive feature selection schema using improved technical indicators for predicting stock price movements. **Expert Systems with Applications**, [*s. l.*], v. 200, p. 1 – 12, 2022. Disponível em: https://doi.org/10.1016/j.eswa.2022.116941%20. Acesso em: 16 jan. 2022.

KE, Guolin; MENG, Qi; FINLEY, Thomas; WANG, Taifeng; CHEN, Wei; MA, Weidong; YE, Qiwei; LIU, Tie Yan. LightGBM: A highly efficient gradient boosting decision tree. *In*: 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2017) 2017, Long Beach, CA. **Anais** [...]. Long Beach, CA: NIPS, 2017. p. 3147 – 3155. Disponível em: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. Acesso em: 19 fev. 2022.

KOLANI, Daname. Portfolio selection using random forest algorithm. **International Journal of Computer Engineering and Data Science**, [*s. l.*], v. 2, n. 1, p. 28 – 36, 2022. Disponível em: http://www.ijceds.com/ijceds/article/view/32. Acesso em: 17 jan. 2022.

MEHTAB, Sidra; SEN, Jaydip. Stock price prediction using CNN and LSTM-based deep learning models. *In*: 2020 INTERNATIONAL CONFERENCE ON DECISION AID SCIENCES AND APPLICATION (DASA) 2020, Sakheer. **Anais** [...]. Sakheer: IEEE, 2020. p. 447 – 453. Disponível em: https://ieeexplore.ieee.org/document/9317207/. Acesso em: 18 fev. 2022.

MO, Hao; SUN, Hejiang; LIU, Junjie; WEI, Shen. Developing window behavior models for residential buildings using xgboost algorithm. **Energy and Buildings**, [*s. l.*], v. 205, p. 1 – 8, 2019. Disponível em: https://doi.org/10.1016/j.enbuild.2019.109564%20. Acesso em: 17 jan. 2022.

NABIPOUR, M.; NAYYERI, P.; JABANI, H.; MOSAVI, A.; SALWANA, E.; S., Shahab. Deep learning for stock market prediction. **Entropy**, [*s. l.*], v. 22, n. 840, p. 1 – 23, 2020. Disponível em: https://www.mdpi.com/1099-4300/22/8/840. Acesso em: 18 fev. 2022.

PRADO, M. L. de. **Advances in financial machine learning**. Hoboken: John Wiley & Sons, Inc., 2018. 393 p. ISBN 9781119482086.

PROKHORENKOVA, Liudmila; GUSEV, Gleb; VOROBEV, Aleksandr; DOROGUSH, Anna Veronika; GULIN, Andrey. Catboost: unbiased boosting with categorical features. **arxiv**, [*s. l.*], p. 1 – 23, 2017. Disponível em: http://arxiv.org/abs/1706.09516. Acesso em: 15 fev. 2022.

RAJKAR, Ajinkya; KUMARIA, Aayush; RAUT, Aniket; KULKARNI, Nilima. Stock market price prediction and analysis. **International Journal of Engineering Research & Technology**, [*s. l.*], v. 10, n. 6, p. 115 – 119, 2021. Disponível em: https://www.academia.edu/49509217/IJERT_Stock_Market_Price_Prediction_and_Analysis?from=cover_page. Acesso em: 20 jan. 2022.

SCHAPIRE, R. E.; FREUND, Y. **Boosting**: foundations and algorithms. Cambridge: The MIT Press, 2012. 544 p. ISBN 9780262017183.

SRIVINAY; MANUJAKSHI, B.; KABADI, Mohan; NAIK, Nagaraj. A Hybrid stock price prediction model based on PRE and deep neural network. **Data**, [*s. l.*], v. 7, n. 5, p. 51, 2022. DOI: 10.3390/data7050051. Disponível em: https://www.mdpi.com/2306-5729/7/5/51. Acesso em: 18 jan. 2022.

SUN, Xiaolei; LIU, Mingxi; SIMA, Zeqian. A novel cryptocurrency price trend forecasting model based on lightgbm. **Finance Research Letters**, [*s. l.*], v. 32, p. 1 – 6, 2020. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1544612318307918. Acesso em: 18 jan. 2022.

WADE, C. **Hands-on gradient boosting with xgboost and scikit-learn**: perform accessible machine learning and extreme gradient boosting with python. Birmingham: Packt Publishing Ltd., 2020. 311 p. ISBN 978-1-83921-835-4.

XU, Renzhe; CHEN, Yudong; XIAO, Tenglong; WANG, Jingli; WANG, Xiong. Predicting the trend of stock index based on feature engineering and catboost model. **International Journal of Financial Engineering**, [*s. l.*], v. 08, n. 2, p. 1 – 17, 2021. Disponível em: https://www.worldscientific.com/doi/epdf/10.1142/S2424786321500274. Acesso em: 21 jan. 2022.

YIN, Lili; LI, Benling; LI, Peng; ZHANG, Rubo. Research on stock trend prediction method based on optimized random forest. **CAAI Transactions on Intelligence Technology**, [*s. l.*], p. 1 – 11, 2021. Disponível em: https://doi.org/10.1049/cit2.12067. Acesso em: 1 maio 2022.