Abordagem para Predição de Séries Temporais baseada em Algoritmos de Aprendizado de Máquina em Ações da Bolsa de Valores Brasileira

Ana P. S. Gularte[1,2], Danusio G. G. Filho[1,2], Gabriel O. Torres[1], Thiago C. N. Silva[1]

[1]Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – SP – Brasil

[2]Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brasil

gularte@ita.br,danusio.gadelhafilho@gmail.com,

torres@ita.br,thg.cns@gmail.com

Resumo. *Muitos estudos utilizam técnicas de aprendizado de máquina (AM) para prever os valores de retorno de ações. A abordagem de seleção de atributos – que atinge melhor desempenho no tempo de treinamento, na precisão da previsão e na melhor capacidade de selecionar um subconjunto de atributos – tem atraído atenção no campo de AM em grandes conjuntos de dados. Esse artigo implementa ensemble de modelos e seleção de atributos para prever os retornos de dez ações 180 dias à frente. Os resultados sugerem que há potencial de predição em retorno de ações com base nas variáveis mais importantes selecionadas pelo algoritmo, e mostram que o modelo ensemble supera o ARIMA nas métricas de avaliação (R^2, Willmott, MAE e Curtose).*

1. Introdução

O mercado financeiro desempenha um papel fundamental nas organizações social e econômica da sociedade ao propiciar a captação de recursos, pois permite a alocação de capital em investimentos produtivos, gera emprego e renda e oportuniza o desenvolvimento econômico [Lin et al. 2012]. Na mesma linha, a democratização do acesso ao mercado financeiro, oriunda da aceleração digital, proporciona que muitos setores mantenham suas atividades de formas remota e produtiva, o que demonstra a importância desses mecanismos digitais como ferramentas para o relacionamento com os investidores. Dessa forma, a análise de movimentos que influenciam o setor tem sido amplamente estudada por acadêmicos e executivos. O desenvolvimento de métodos e algoritmos computacionais que se utilizam da inteligência artificial auxiliam os tomadores de decisão em seus mais diversos negócios [Leles et al. 2019], [Izbicki and dos Santos 2020], [Khaidem et al. 2016].

Em geral, o valor de retorno das ações é considerado séries estocásticas estacionárias, caracterizadas quando a média e a variância não oscilam sistematicamente ao longo do tempo [Bueno 2018], [Morettin 2017], [Gujarati and Porter 2011]. Para o propósito de predição essa característica é fundamental, caso contrário, uma série não estacionária possui pouco valor prático, tendo em vista que seu processo é puramente aleatório.

Com base em uma análise gráfica de [Cullen et al. 1999], é possível identificar a natureza provável da série temporal, cujos movimentos nos valores de retorno e variância dos ativos financeiros apresentam distribuição normal. No entanto, prever os preços dos ativos financeiros é uma difícil tarefa, pois geralmente são não lineares, dinâmicos e caóticos. Entre as técnicas mais recentes, o aprendizado de máquina vem ganhando cada vez mais espaço no mercado financeiro, entre as vantagens, está a capacidade de encontrar padrões complexos em grandes volumes de dados [Henrique et al. 2019], [Huck 2019].

Portanto, diante do exposto, surge o questionamento: Como os algoritmos de aprendizado de máquina podem ser utilizados para predição de retornos em ações da bolsa de valores de São Paulo?

Este trabalho objetiva prever a direção dos valores de retorno de 10 ações de diversos setores negociadas na B3 180 dias à frente. Utilizou-se dos seguintes algoritmos de aprendizado de máquina para seleção de variáveis: 1) *One Rule (OneR)*, 2) *Information Gain* e 3) *Chi-Squared*; e para predição: 1) *Gradient Boosting Machine (GBM)*, 2) *K-Nearest Neighbor (KNN)* e 3) *Bayesian Regularized Neural Networks (BRNN)*, além de *ensemble* dos modelos. Outrossim, faz-se engenharia de variáveis com intenção de aperfeiçoar o modelo preditivo, de contribuir no uso das técnicas de aprendizado de máquina e de proporcionar uma melhor interpretação dos padrões extraídos pelo modelo [Faceli et al. 2021].

2. Bibliografia Correlata

As técnicas de aprendizado de máquina para análise e previsão de séries temporais vêm se desenvolvendo e propondo aperfeiçoamento nas predições das tendências dos mercados financeiros. Resultados mostram que a precisão desses métodos de inteligência artificial é superior aos métodos estatísticos tradicionais, principalmente no que diz respeito a padrões não lineares. Esses métodos foram extensivamente estudados por [Huang and Wu 2008], [Yu et al. 2008], [Bahrammirzaee 2010], [Lin et al. 2012], [Khaidem et al. 2016], [Zhang et al. 2017] e mais recentemente, [Huck 2019], que apresenta algumas das técnicas de ponta baseadas em aprendizado de máquina em centenas de ações e preditores.

O método não paramétrico dos K-vizinhos mais próximos (K-NN) é comparado a estudos de redes neurais que apresentaram desempenho ligeiramente melhor em algumas redes [Bahrammirzaee 2010], no entanto, ambos os métodos superam os clássicos estatísticos de regressão [Bontempi et al. 2012]. Pesquisas revelam que o K-NN provou ser um método potencial e amplamente aplicado em várias previsões de séries temporais [Zhang et al. 2017], [Bontempi et al. 2012].

Outrossim, modelos de previsão que consideram a importância da otimização dos parâmetros do modelo, e o uso de dados recentes como a abordagem *BRNN* se destacam na pesquisa de [Yan et al. 2017], que propõe um incremento no algoritmo para lidar com problemas de capacidade de generalização e sobreajuste na previsão.

A técnica de *ensemble learning* contribui no aprendizado de máquina pela vantagem estatística, computacional e representacional ao permitir a combinação de várias hipóteses no espaço de solução da função, pois gera um modelo com maior acurácia [Flach 2012], [Krauss et al. 2017]. Uma das abordagens de *ensemble learning* é a *bo-*

osting explorada neste trabalho: o método *gradient boosting*. [Hastie et al. 2009] mencionam em seu livro que "*boosting* é uma das ideias de aprendizagem mais poderosas introduzidas nos últimos vinte anos. [...] p. 337.", além de ser computacionalmente eficiente se comparado a métodos como redes neurais [Krauss et al. 2017].

A abordagem de seleção dos atributos baseada em filtro permite melhoria no tempo de treinamento, na precisão da previsão e na capacidade de selecionar um subconjunto de atributos melhor. Essa abordagem é discutida na pesquisa de [Huang and Tsai 2009], que faz uso desse método para reduzir a complexidade dos dados e melhorar a acurácia da previsão.

Todos os itens acima contribuem para o desenvolvimento de algoritmos computacionais que preenchem a lacuna no uso de técnicas com maior precisão para previsão de séries temporais na área de finanças acadêmica e profissional, o que torna este estudo relevante para ambas as partes.

3. Materiais e Métodos

São apresentados o processo de obtenção da base de dados utilizada, os ativos financeiros selecionados, o período avaliado, bem como as técnicas e algoritmos adotados.

3.1. Conjunto de Dados

Compôs a amostra preços de fechamento diário de dez ações de empresas de capital aberto de oito setores econômicos distintos, ordenados cronologicamente. Essas empresas fazem parte do índice Bovespa, que mede o desempenho dos ativos de maior representatividade do mercado brasileiro e que se destacam pela eficiência de mercado (alcançada quando a alocação dos recursos maximiza o excedente total) e liquidez. O período de análise é de 1º/Jan/2016 a 30/Dez/2020 consistindo em 1.254 dias de negociações.

A relação de uma pequena amostra da base de dados considerando as variáveis coletadas para análises posteriores está representada na Tabela 2 no Apêndice A.

3.2. Técnicas

A Figura 1 ilustra as principais técnicas utilizadas, as quais são detalhadas na sequência.

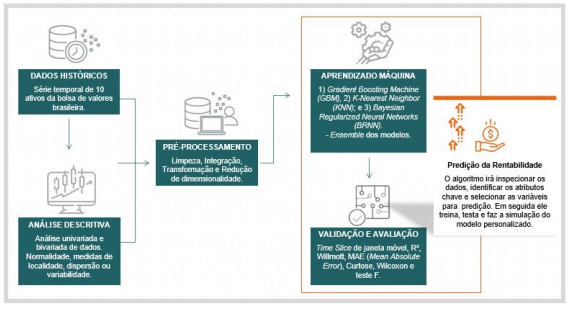


Figura 1. Pipeline do Aprendizado de Máquina. Fonte: Autoria própria, 2021.

3.2.1. Seleção de variáveis

Utilizamos abordagens dos tipos embutida e baseada em filtros para que cada atributo fosse avaliado por sua característica, inerente ao ganho de informação. A combinação da

magnitude das importâncias traz uma visão menos enviesada da relevância dos preditores.

Por conta do longo período de experimentação e pelo número de ações utilizadas, o tempo demandado na seleção de atributos torna-se fundamental. O *One Rule* é um algoritmo que fornece os resultados das tarefas de classificação de forma rápida e precisa quando aplicado em seleção de atributos; tem-se os preditores que, individualmente, entregam o melhor resultado de predição [Holmes and Nevill-Manning 1995].

O Ganho de Informação também possui rápida execução e permite escolher os preditores de forma mais genérica independente do algoritmo, diferente do *One Rule*, que implícita uma regressão ou classificação [Azhagusundari et al. 2013]. Para defini-lo, utiliza-se uma medida de aleatoriedade do atributo alvo, chamada entropia – medida de incerteza de uma variável aleatória, e que especifica o número mínimo de bits de informação para codificar a classificação de um membro arbitrário de H(A) [Mitchell et al. 1997]. Em geral, a entropia de uma variável aleatória A com valores a_v e probabilidades p_v é definida como $H(A) = -\sum_i p_i * log_2(p_i)$ [Faceli et al. 2021].

Como base em teste estatístico, escolhemos o algoritmo Qui-Quadrado, o qual é uma metodologia bastante utilizada na escolha de atributos [Hastie et al. 2009], pois testa a independência ou associação entre as variáveis categóricas, e se o modelo estatístico ajusta-se aos dados de forma adequada.

3.2.2. Modelo preditivo

Assim como na seleção de atributos, optamos por combinar modelos de forma a explicar uma maior amplitude da variável alvo. Utilizamos a técnica *ensemble learning* e escolhemos o método *gradiente boosting* para a combinação de modelos fracos a fim de gerar um modelo com maior acurácia [Flach 2012], [Krauss et al. 2017]. O objetivo do *gradiente boosting* é trabalhar com preditores em sequência, de forma que cada preditor corrija os erros do anterior. Ao se basear nos erros residuais dos preditores, o algoritmo gera um novo preditor para melhorar a eficiência do preditor em geral [Cavalcante et al. 2018]. Já a escolha do *Bayesian Regularization Neural Networks (BRNN)* foi devido ao algoritmo apresentar precisão de convergência rápida e por contribuir na redução dos efeitos de *overfiting* [Lahmiri and Bekiros 2020].

Implementar predições em séries financeiras baseadas em padrões é uma metodologia amplamente utilizada, como já citado anteriormente na bibliografia correlata. A escolha do (KNN) além de apresentar resultados satisfatórios armazena o histórico que compõe os preditores, já que o resultado se dá por padrões semelhantes do passado. A Tabela 3 constante no Apêndice A especifica os parâmetros de todas as técnicas utilizadas.

4. Experimentos e Resultados

A predição dos valores de retornos dos preços dos ativos em $D + s$ é feita a partir do valor dos preditores em $D + 0$, onde s é igual a 180 dias corridos. O desempenho do algoritmo é medido por meio do coeficiente de determinação ($R^2 = 1 - \sum_{i=1}^{n} u_i^2 / \sum_{i=1}^{n} (y_i - \bar{y})$), que mede a proporção da variação total da variável-alvo explicada pelo modelo de regressão [Gujarati and Porter 2011], em que $\sum_{i=1}^{n} u_i^2$ é a soma dos quadrados do resíduos do modelo e $\sum_{i=1}^{n} (y_i - \bar{y})$, a soma dos quadrados totais; do erro absoluto médio

($MAE = \frac{1}{n}\sum_i^n |y_i - f(x_i)|$), que mensura a distância entre os valores preditos e reais, em que y_i é o valor real do alvo i e $f(x_i)$, o valor predito do mesmo; e do Índice de willmott ($d = 1 - \sum_{i=1}^{n}[(y_i - \bar{x}) - (x_i - \bar{x})]^2 / \sum_{i=1}^{n}[|y_i - \bar{x}| + |x_i - \bar{x}|]^2$), que é uma medida que reflete a precisão entre os valores preditos e os valores observados. Ainda, utilizou-se o cálculo da curtose a fim de medir a dispersão dos erros entre as predições e os valores observados. Por fim, dada a aplicação dos dois modelos (*ensemble* e ARIMA), foram realizados o teste Wilcoxon, para comparar a mediana dos desvios absolutos das predições entre os modelos; e o teste F, para comparar as diferenças nas variâncias das predições nos dois modelos. As equações das métricas utilizadas encontram-se na Tabela 5, no Apêndice A.

4.1. Pré-processamento dos Dados

Esta etapa consistiu em proporcionar robustez ao algoritmo e a minimizar a influência de eventuais problemas presentes na base de dados, tais como: ruídos e valores ausentes, entre outros. Fizeram parte dessas operações: 1) limpeza dos dados; 2) integração de dados; 3) transformações dos dados e 4) redução de dimensionalidade. Cada etapa realizada é descrita abaixo:

- 1) Na limpeza dos dados, os valores faltantes foram preenchidos pelo valor imediatamente anterior da série, a fim de preservar a variabilidade da série temporal contínua [Rihbane 2014].
- 2) Realizou-se o cálculo da rentabilidade logarítmica dos valores de fechamento dos ativos, conforme : $rcc = ln\left(\frac{P[D+s]}{P[D+0]}\right)$, onde s é o *lag* (janela de previsão) e $P[x]$ é o preço do ativo no dia x. Houve um total de 18 variáveis concebidas após aplicação de *feature engineering*, divididas da seguinte forma: 4 técnicas, 2 mercadológicas, 2 da ação, 3 artificiais de preço e 6 artificiais de ROC, além da variável alvo, descritas parcialmente na Tabela 4, Apêndice A.
- 3) Padronizou-se a variável rentabilidade por meio da expressão $\tilde{X} = \frac{x - \bar{x}}{s_x}$, em que $\bar{x}$ é a média de x, s_x é o desvio-padrão de x e $\tilde{X}$ é o vetor padronizado, para que as variáveis com grande escala não prevaleçam no processo de predição.
- 4) Utilizamos seleção de variáveis, que atua como filtros que descartam as variáveis redundantes ou irrelevantes e mantêm uma parte das variáveis originais na base de dados. Assim, fez-se o rankeamento das variáveis com base no resultado de três algoritmos de seleção: *One Rule*; *Information Gain*; e Qui-Quadrado.

4.1.1. Divisão dos dados - Janela Deslizante

Para predição dos valores futuros da série temporal, utilizamos a técnica *time slice*, que se inicia com a montagem do conjunto de treinamento a partir da janela de tempo de 180 dias considerando os valores passados das variáveis preditoras e da própria variável que se deseja prever, bem como do horizonte de previsão definido de 5 anos.

O padrão da entrada é formado pelos dados históricos das variáveis preditoras, que pode incluir os valores passados da própria série que se deseja prever, e a saída desejada é o valor da série temporal no horizonte de previsão. O algoritmo constrói um dataset de treinamento movendo as janelas de entrada e saída ao longo de toda a série temporal até esgotarem-se as instâncias de teste, conforme procedimento demonstrado a seguir:

{Início da iteração time slice}

- Passo 1. Definir a instância de início do teste $\{i\}$;
- Passo 2. Calcular o número médio de pregões em 180 dias $\{n\}$;
- Passo 3. Definir a variável-alvo: preços de fechamento da instância $(n + 1)$ à instância $(i - 1)\{r1\}$;
- Passo 4. Definir uma variável preditora com os retornos da instância (1) à instância $(i - 1 - n)\{r0\}$;
- Passo 5. Calcular as demais variáveis preditoras com referência às instâncias entre (1) e $(i - 1 - n)$;
- Passo 6. Treinar o modelo no dataset gerado: alvo com referência às instâncias entre $(n+1)$ e $(i-1)$ e preditores com referência às instâncias entre (1) e $(i-1-n)$;
- Passo 7. Estimar o valor do alvo na instância (i) usando os atributos da instância $(i - n)$;
- Passo 8. Incrementar o valor de $\{i\}$ em uma unidade;
- Passo 9. Repetir os passos de 2 a 9.

Critério de parada: esgotarem-se as instâncias de teste.

{Fim da iteração time slice}

4.2. Resultados e discussões

Após alternar parâmetros dos algoritmos, optou-se pelos valores presentes na Tabela 3, Apêndice A, em que a quarta coluna apresenta a configuração final utilizada nos experimentos, pois aliam bom desempenho (dentro dos critérios usados neste trabalho) e economia de recursos computacionais. Foram realizadas 100 iterações por ativo, 1.000 iterações e 29 horas de execução no total. Nove preditores foram selecionados para o modelo.

A Figura 2 apresenta o tempo de execução por iteração para o modelo proposto, que resultou em uma média de 104 segundos por iteração.

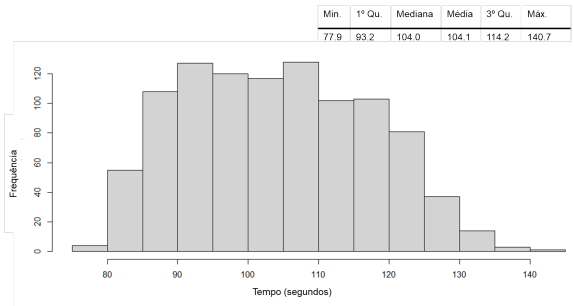


Figura 2. Tempo de execução (modelo AM). Fonte: Autoria própria, 2021.

4.2.1. Seleção de variáveis

Com os algoritmos *OneR*, Ganho de Informação e Qui-Quadrado, extraímos, por meio de rankeamento da força preditiva relativa de cada recurso, a importância das variáveis. Na Figura 3, à esquerda, apresentamos a ocorrência das variáveis mais frequentes, em cada uma das 9 posições entre os 18 preditores, concebidos na etapa de pré-processamento. O gráfico "Importância 1", por exemplo, descreve a variável "mml" (Média Móvel Lenta do preço da ação) como a mais frequentemente selecionada como variável mais importante.

Para agregar o resultado, foi usado o somatório ponderado das frequências de ocorrência das variáveis em cada uma das 9 posições. Foi atribuído peso 9 para a ocorrência na posição 1, 8 para a posição dois até 1 para a posição 9. Se a variável *X1* apareceu 40% das vezes na posição 3 (terceira variável mais importante) e 10% das vezes na posição 6, não ocorrendo em outras posições, sua importância total é: $7 \times 0,4 + 4 \times 0,1 = 3,2$.

Na Figura 3, à direita, são apresentadas as 14 variáveis selecionadas em todas as iterações do algoritmo, com a média da importância das variáveis mais frequentes em ordem decrescente, no eixo das abscissas. Destacamos as sete variáveis mais importantes e frequentes: “mml”, “mmr”, mmr_rcc”, “volat_acao”, “volat_ibov”, “mml_rcc” e “signal”, representadas na Tabela 4; nas demais variáveis, percebe-se uma queda de importância relativa a partir da variável “roc_acao”. Os resultados sugerem que pode haver um potencial significativo para estratégias de predição em retorno de ações com base nessas variáveis.

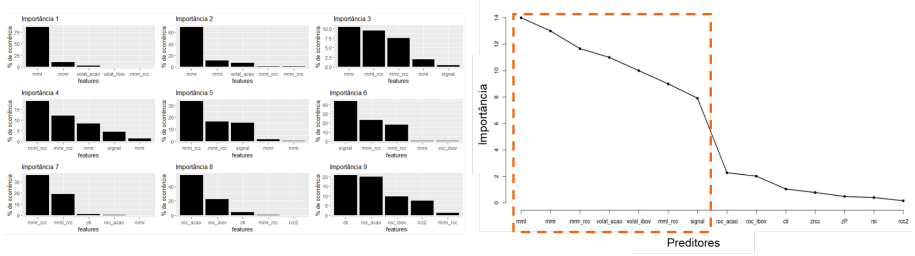


Figura 3. Esq.: 5 variáveis mais frequentes em cada uma das 9 posições dos preditores; Dir.: importância dos preditores. Fonte: Autoria própria, 2021.

4.2.2. Modelos Preditivos

Quatro métricas de avaliação (R^2, Willmott, MAE e Curtose) foram usadas para medir a estabilidade dos modelos desenvolvidos durante a fase de teste para predição dos valores de retorno, conforme Tabela 1. Os valores médios das métricas de desempenho são apresentadas nesta. Todas foram calculadas *out-of-sample*, ou seja, com os dados fora da amostra de treinamento, considerando instâncias não vistas pelo modelo treinado, na intenção de deixar a simulação mais próxima da vida real.

De fato, o modelo de aprendizado de máquina supera o modelo tradicional ARIMA em todas as métricas. Os erros de predição medidos pelo MAE apontam uma média de 3% para o modelo de AM, enquanto o ARIMA apresentou 32%; o índice de concordância ou de ajuste de Willmott indicou melhor ajuste entre os valores preditos e realizados no modelo de AM, com valor médio próximo de 1.

Na Figura 4 comparamos os valores de retorno predito com os valores reais considerando as 6 ações que melhor performaram no modelo de AM. É notável que os desvios da predição do modelo proposto apresentam valores menos dispersos, medidos pela curtose mais concentrada. Além de valores de R^2 mais altos do que o ARIMA, o que indica que a predição do modelo proposto é adequada e que explicam bem os dados reais.

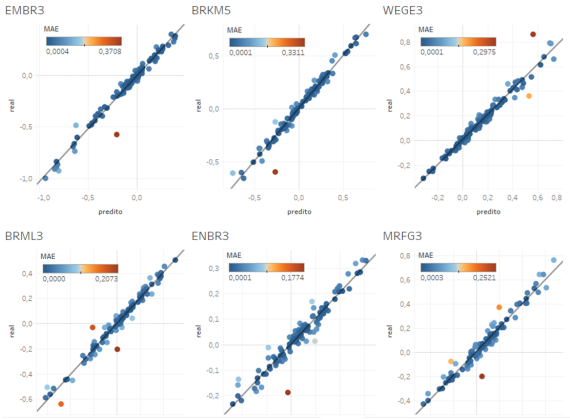
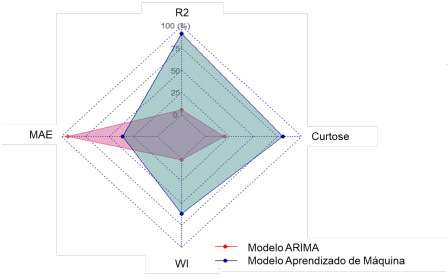


Figura 4. Comparação entre valores de retorno predito e realizado para os seis melhores modelos do *ranking*. Fonte: Autoria própria, 2021.

4.2.3. Teste de hipóteses

Na Figura 5, à esquerda, é apresentada a comparação entre as métricas dos modelos utilizados. Em todas as medidas, o ARIMA mostrou-se inferior ao modelo de AM. A fim de comprovar a diferença de performance entre os modelos, realizamos a análise estatística de significância por meio do teste não paramétrico de Wilcoxon e do teste F. Ambos apontaram para a aceitação da hipótese alternativa de que os desvios de predição do modelo de AM têm mediana e variância menores que os desvios gerados pelo ARIMA.

Analogamente, à direita da Figura 5, realizamos o ranqueamento dos ativos selecionados de acordo com cada métrica de avaliação e suas médias. Destacamos as seis ações com melhor desempenho: EMBR3, seguida por BRKM5, WEGE3, BRML3, ENBR3 e MRFG3. Essas ações foram submetidas aos testes estatísticos, e os resultados do teste de Wilcoxon apontaram que somente ficou abaixo do valor de 5% de significância as medianas das ações EMBR3 e WEGE3 no modelo de AM, e BRKM5, ENBR3 e WEGE3 no modelo ARIMA, ficando muito distantes da mediana os valores preditos e reais. Já os resultados do teste f apontaram que em todas as ações do modelo de aprendizado de máquina as predições apresentaram variância similar aos valores reais, já no modelo ARIMA as ações BRKM5, ENBR3, MRFG3 e WEGE3 ficaram fora do nível de significância, de acordo com os valores na Tabela 1.



Ativo	Modelo	R2	Willmott	MAE	Curtose dos desvios da predição	Média Ranking
EMBR3	Aprendizado de Máquina	1	3	5	1	2,50
BRKM5		2	1	6	2	2,75
WEGE3		4	4	3	3	3,50
BRML3		3	2	4	6	3,75
ENBR3		8	5	1	5	4,75
MRFG3		7	7	8	7	7,25
VALE3		5	6	9	9	7,25
CSAN3		9	10	7	4	7,50
VIVT3		10	8	2	10	7,50
GGBR4		6	9	10	8	8,25

Figura 5. Esq.: comparação de desempenho dos modelos de AM e ARIMA; Dir.: ranking do modelo de AM por métrica). Fonte: Autoria própria, 2021.

Tabela 1. Métricas dos modelos aplicados.

Ativo	Modelo	R2	Willmott	MAE	Curtose dos desvios da predição	Wilcoxon P-valor	F P-valor
Característica		Maior melhor (máximo = 1)		Menor melhor	referência distrib. normal = 3	-	-
BRKM5		0.970034	0.991709	0.032958	18,216822	0.240320	0.918488
BRML3		0.964421	0.991294	0.029572	9,109334	0.595266	0.758876
CSAN3		0.927291	0.986712	0.032980	15,010215	0.179380	0.663911
EMBR3	**Apredizado**	0.972139	0.991109	0.031039	25,079928	0.033453	0.667514
ENBR3	**de**	0.933676	0.990176	0.022744	10,569901	0.406338	0.601283
GGBR4	**Máquina**	0.952141	0.986788	0.048842	6,332533	0.952020	0.792186
MRFG3		0.949316	0.987818	0.037041	8,241648	0.631479	0.883291
VALE3		0.957942	0.989490	0.038638	5,237920	0.087798	0.838927
VIVT3		0.888672	0.987060	0.025419	4,403632	0.607231	0.744828
WEGE3		0.959745	0.990208	0.029251	18,026506	0.033740	0.679524
Médias		**0,947538**	**0,989236**	**0,032848**	**12,022844**	**0,376703**	**0,754883**
Ativo	**Modelo**	**R2**	**Willmott**	**MAE**	**Curtose dos desvios da predição**	**Wilcoxon P-valor**	**F P-valor**
Característica		Maior melhor (máximo = 1)		Menor melhor	referência distrib. normal = 3	-	-
BRKM5		0.096021	0.449834	0.275175	2,752926	0.000796	0.000000
BRML3		0.018008	0.357350	0.331181	3,318790	0.109479	0.091276
CSAN3		0.001351	0.399164	0.351276	2,824857	0.334820	0.191151
EMBR3		0.075595	0.084169	0.600571	2,187870	0.686210	0.987901
ENBR3	**ARIMA**	0.037713	0.466410	0.238525	3,521933	0.000958	0.000000
GGBR4		0.198983	0.557421	0.205132	2,458412	0.444255	0.825316
MRFG3		0.330618	0.246894	0.179208	2,395699	0.829849	0.000000
VALE3		0.036629	0.028636	0.640999	3,614053	0.784587	0.898107
VIVT3		0.003609	0.534696	0.277023	3,631085	0.322899	0.509939
WEGE3		0.110699	0.702129	0.088847	2,497368	0.005048	0.000000
Médias		**0,090923**	**0,382670**	**0,318794**	**2,920299**	**0,351890**	**0,350369**

5. Conclusão

O trabalho proposto abordou o tema de predição de ações da bolsa de valores brasileira com aplicação de técnicas de aprendizado de máquina (AM). Foi realizada uma análise de regressão das séries temporais das ações com uso de *ensemble* de modelos e utilização do método ARIMA, para posterior comparação com o aprendizado. Os resultados sugerem que pode haver um potencial significativo para estratégias de predição em retorno de ações com base nas 7 variáveis mais importantes sugeridas pelos algoritmos. As métricas de desempenho do modelo de AM – R^2, Willmott, MAE e Curtose – apontaram que o método utilizado na pesquisa provê resultados animadores com respeito a predição das ações. Ainda, notou-se que o modelo de *ensemble* resultou em um desempenho melhor que o modelo tradicional ARIMA, dados os valores das métricas utilizada.

Como sugestões para trabalhos futuros, estão o emprego de métodos heurísticos para seleção de ativos mais correlacionados, a fim de construir uma amostragem probabilística por agrupamento; a incorporação de novas variáveis independentes, como as sugeridas pelo pacote *tsfresh*, o qual se utiliza de métodos para avaliar o poder explicativo e a importância das variáveis; a aplicação de outros testes de hipóteses não paramétricos, como os de Friedman e de Nemenyi; e, com o intuito de reduzir o tempo de processamento, introduzir técnicas de paralelização no algoritmo.

Referências

Azhagusundari, B., Thanamani, A. S., et al. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(2):18–21.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications*, 19(8):1165–1195.

Bontempi, G., Taieb, S. B., and Le Borgne, Y.-A. (2012). Machine learning strategies for time series forecasting. In *European business intelligence summer school*, pages 62–77. Springer.

Bueno, R. L. S. (2018). *Econometria de Séries Temporais*. Cengage Learning.

Cavalcante, H., H., Campos, H., M., and Sakurai, G., T. (2018). Trading de ações com base em análise de mídias sociais. monografia (tcc).

Cullen, A. C., Frey, H. C., and Frey, C. H. (1999). *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer Science & Business Media.

Faceli, K., Lorena, A. C., Gama, J., Carvalho, A., et al. (2021). Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2:304.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

Gujarati, D. N. and Porter, D. C. (2011). *Econometria básica-5*. Amgh Editora.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Henrique, B. M., Sobreiro, V. A., and Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124:226–251.

Holmes, G. and Nevill-Manning, C. G. (1995). Feature selection via the discovery of simple classification rules.

Huang, C.-L. and Tsai, C.-Y. (2009). A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. *Expert Systems with applications*, 36(2):1529–1539.

Huang, S.-C. and Wu, T.-K. (2008). Integrating ga-based time-scale feature extractions with svms for stock index forecasting. *Expert Systems with Applications*, 35(4):2080–2088.

Huck, N. (2019). Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research*, 278(1):330–342.

Izbicki, R. and dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

Khaidem, L., Saha, S., and Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

Krauss, C., Do, X. A., and Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. *European Journal of Operational Research*, 259(2):689–702.

Lahmiri, S. and Bekiros, S. (2020). Intelligent forecasting with machine learning trading systems in chaotic intraday bitcoin market. *Chaos, Solitons & Fractals*, 133:109641.

Leles, M. C., Sbruzzi, E. F., de Oliveira, J. M., and Nascimento, C. L. (2019). A matlab™ computational framework for multiagent system simulation of financial markets. In *2019 IEEE International Systems Conference (SysCon)*, pages 1–8. IEEE.

Lin, C.-S., Chiu, S.-H., and Lin, T.-Y. (2012). Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Economic Modelling*, 29(6):2583–2590.

Mitchell, T. M. et al. (1997). Machine learning.

Morettin, P. A. (2017). *Econometria financeira: um curso em séries temporais financeiras*. Editora Blucher.

Rihbane, F. (2014). *Preenchimento de Falhas Aleatórias de Séries Temporais Micrometeorológicas pela Técnica de Monte Carlo*. PhD thesis, Dissertação de Mestrado, IF, UFMT, Cuiabá.

Yan, D., Zhou, Q., Wang, J., and Zhang, N. (2017). Bayesian regularisation neural network based on artificial intelligence optimisation. *International Journal of Production Research*, 55(8):2266–2287.

Yu, L., Chen, H., Wang, S., and Lai, K. K. (2008). Evolving least squares support vector machines for stock market trend mining. *IEEE transactions on evolutionary computation*, 13(1):87–102.

Zhang, N., Lin, A., and Shang, P. (2017). Multidimensional k-nearest neighbor model based on eemd for financial time series forecasting. *Physica A: Statistical Mechanics and its Applications*, 477:161–173.

A. Apêndice

Tabela 2. Variáveis coletadas para análise (parcial, apenas para referência) com tipo e escala dos atributos.

Qualitativo Nominal	**Qualitativo Nominal**	**Quantitativo Intervalar**	**Quantitativo Racional**
Ativo	**Setor***	**Data**	**Valor de Fechamento (R$)**
BRKM5	Petroquímico	05/01/2016	25,93
BRML3	Imobiliário e Construção	05/01/2016	7,26
CSAN3	Petróleo e Gás	05/01/2016	24,12
EMBR3	Industrial	05/01/2016	29,32
ENBR3	Energia e Saneamento	05/01/2016	11,69
GGBR4	Siderúrgico e Mineração	05/01/2016	4,33
MRFG3	Consumo e Varejo	05/01/2016	5,95
VALE3	Siderúrgico e Mineração	05/01/2016	12,52
VIVT3	Telecomunicações	05/01/2016	30,40
WEGE3	Industrial	05/01/2016	11,81

*Fonte: https://www.infomoney.com.br

Tabela 5. Métricas de desempenho utilizadas no trabalho.

Abreviação	Métrica	Equação	Característica
R^2	Coeficiente de Determinação	$1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})}$	Maior é melhor (0 a 1)
MAE	Erro Absoluto Médio	$\frac{1}{n}\sum_{i}^{n} \lvert y_i - f(x_i)\rvert$	Menor é melhor
Wilcoxon	Teste de Hipótese de Wilcoxon	–	–
WI	Willmott Index	$1 - \frac{\sum_{i=1}^{n}[(f(x_i)-\bar{y})-(y_i-\bar{y})]^2}{\sum_{i=1}^{n}[\lvert f(x_i)-\bar{y}\rvert+\lvert y_i-\bar{y}\rvert]^2}$	Maior é melhor (0 a 1)
Curtose	Curtose	$\frac{1}{n}\sum_{i}^{n}\left[\frac{f(x_i)-\bar{f(x)}}{s_{f(x_i)}}\right]$	–
F	Teste F	–	–

$y_i = obervado; f(x_i) = predito; u_i = erro; n = instâncias$

Tabela 3. Parâmetros dos modelos de Aprendizado de Máquina.

Função	Pacotes	Parâmetros	Valor dos parâmetros	Descrição
oneR	Fselector	Nenhum	Nenhum	Fornece a precisão preditiva dos atributos em ordem decrescente.
information.gain	Fselector	Nenhum	Nenhum	Fornece o peso dos atributos discretos com base em sua correlação com o atributo de classe continua.
chi.squared	Fselector	Nenhum	Nenhum	Determina se duas variáveis possuem uma correlação significativa.
preProcess	caret	Nenhum	Nenhum	Estima os parâmetros necessários para cada operação e os prediz.
predict	caret	Nenhum	Nenhum	Estima valores baseado nos dados de entrada.
gbm	gbm	shrinkage	0.1	Taxa de aprendizagem, comumente escolhido um valor menor para muitas iterações.
gbm	gbm	interaction.depth	4	Número de divisões que deve ser executado em uma árvore partindo de um único nó.
gbm	gbm	n.minobsinnode	10	Número mínimo de observações nos nós folhas das árvores.
gbm	gbm	n.trees	200	Número de iterações do algoritmo *gradient boosting machine (gbm).*
knn	kknn	kmax	5	Número máximo de k utilizado quando k não é especificado.
knn	kknn	distance	1	Parâmetro da distância de Minkowski, no caso, distancia euclidiana (distance =1).
knn	kknn	kernel	optimal	Tipo de núcleo a ser utilizado.
brnn	brnn	neurons	9	Número de neurônios por variável.
cubist	Cubist	committees	Automático	Número de interações que serão aplicadas.
cubist	Cubits	neighbors	Automático	Número de vizinhos próximos.

Tabela 4. Descrição das Variáveis para o modelo de Aprendizado de Máquina.

Tipo 1	Tipo 2	Nome(Variável)	Descrição
-	Variável Alvo	Retorno D+lag	Rendimentos (variações do preço de fechamento) dos ativos escolhidos, em D+lag.
-	Variável Preditora	Retorno D+0	Rendimentos dos ativos escolhidos em D+0.
Variável Técnica	Variável Preditora	MACD (*signal*)	*Moving Average Convergence-Divergence*: indicador de tendência construído pela relação entre duas médias móveis.
Variável Técnica	Variável Preditora	RSI (*rsi*)	*Relative Strength Index*: indicador de momentum que mede a variação dos preços recentes.
Variável Técnica	Variável Preditora	Médias Móveis (*mmr, mml*)	Médias móveis de 1 lag e 1/2 lag que visam a identificar tendências de diferentes origens temporais.
Variável Técnica	Variável Preditora	CTI (*cti*)	*Correlation Trend Indicator*: indicador de tendência baseado na medição da correlação entre o histórico de preços de cada ação e a tendência ideal.
Variável Mercadológica	Variável Preditora	ROC Ibov (*roc_ibov*)	Retornos do IBov em D+0.
Variável Mercadológica	Variável Preditora	Volat IBov (*volat_ibov*)	Desvio-padrão dos últimos 2*lag retornos do IBov.
Variável da ação	Variável Preditora	ROC Ação (*roc_ação*)	Retornos do ativo em D+0.
Variável da ação	Variável Preditora	Volat Ação (*volat_ação*)	Desvio-padrão dos últimos 2*lag retornos do ativo.
Variável artificial de preço	Variável Preditora	Derivada 1ª do preço (*dP*)	Derivada discreta de 1ª ordem aplicada à Média Móvel Exponencial de 10 períodos do preço do ativo.
Variável artificial de preço	Variável Preditora	Derivada 2ª do preço (*d2P*)	Derivada discreta de 2ª ordem aplicada à Média Móvel Exponencial de 10 períodos do preço do ativo.
Variável artificial de preço	Variável Preditora	Preço normalizado (*zP*)	Preço menos a média dos últimos lag preços, e o resultado dividido pelo desvio-padrão dos últimos lag preços.
Variável artificial de ROC	Variável Preditora	Derivada 1ª da ROC (*drcc*)	Derivada discreta de 1ª ordem aplicada aos rendimentos do ativo.
Variável artificial de ROC	Variável Preditora	Derivada 2ª da ROC (*d2rcc*)	Derivada discreta de 2ª ordem aplicada aos rendimentos do ativo.
Variável artificial de ROC	Variável Preditora	ROC² (*rcc2*)	Rendimentos do ativo ao quadrado.
Variável artificial de ROC	Variável Preditora	Médias Móveis da ROC (*mmr_rcc, mml_rcc*)	Médias móveis de 1 lag e 1/2 lag aplicadas aos rendimentos do ativo.
Variável artificial de ROC	Variável Preditora	ROC normalizada (*zrcc*)	Retorno menos a média dos últimos lag retornos, e o resultado dividido pelo desvio-padrão dos últimos lag retornos.