Apoena Revista Eletrônica

APLICAÇÃO DATA SCIENCE: EM UM ESTUDO COMPARATIVO ENTRE REDES NEURAIS E REGRESSÃO LINEAR NA PREDIÇÃO DOS VALORES DE AÇÕES NA BOLSA BRASILEIRA

Arthur Hupsel[1], Pedro Cunha[1], Celso Barreto[1]

[1]Instituto de Informática – Universidade Jorge Amado (UNIJORGE)
CEP 41.730-101 – Salvador – BA – Brasil

2030000117@unijorge.com.br, 221000718@unijorge.com.br

Abstract. *This article proposes, through statistical prediction models based on the use of historical data from a time series, to verify the behavior and performance of two Machine Learning algorithms in predicting data from an investment from an economic point of view. The two computational models studied in question were Linear Regression and Artificial Neural Networks, both being applied to predict the results of the asset value of a Brazilian company on the stock exchange, with the objective of producing a comparative study between the methods and determining their differences in applicability in financial decisions..*

Keywords: *Machine Learning, Economics, Stocks, Linear Regression, Neural Networks.*

Resumo. *Este artigo propõe através de modelos de previsão estatísticos baseados na utilização de dados históricos a partir de uma série temporal, verificar o comportamento e a performance de dois algoritmos de Machine Learning na predição de dados de um investimento do ponto de vista econômico. Os dois modelos computacionais estudados em questão foram a Regressão Linear e as Redes Neurais Artificiais, sendo ambos aplicados na predição dos resultados do valor do ativo de uma empresa brasileira na bolsa de valores, com o objetivo de produzir um estudo comparativo entre os métodos e determinar suas diferenças na aplicabilidade nas decisões financeiras.*

Palavras Chave:*Machine Learning, Economia, Ações, Regressão Linear, Redes Neurais.*

1. Introdução

Modelos de Machine Learning consistem no reconhecimento de padrões em dados relevantes ao problema abordado, denominados *features*, que a partir de então podem ser treinados e testados estruturando assim um modelo capaz de fazer uma previsão quando apresentado a dados nunca antes vistos (MARONI NETO).

De acordo com KUMAR (2015), técnicas de mineração de dados associadas a esses modelos de aprendizagem de máquina estão desempenhando papéis cada vez mais importantes em diversas aplicações e setores, sobretudo aqueles com grande volume de dados, como a previsão do mercado de ações; considerada uma tarefa desafiadora, visto que seus preços reagem a fatores econômicos, financeiros, políticos e sociais. Entretanto, com o avanço da inteligência artificial e a capacidade de processamento de dados nas últimas décadas, tornou-se possível prever os movimentos dinâmicos do mercado financeiro, seja com modelos de regressão, ou até mesmo de classificação (VALENCIA et al., 2019).

Conforme Hoffman (2016), dentre os modelos mais utilizados, a análise de regressão é o método mais importante da econometria, e o seu estudo e a comparação com outros se faz muito importante para um melhor entendimento e aplicação desses algoritmos no auxílio de problemas econômico-financeiros, o que sugere o modelo de regressão linear como um dos mais adequados a essa questão. Por outro lado, entre todas essas técnicas, as redes neurais artificiais (RNAs) são amplamente populares em todo os campos principalmente devido à sua capacidade de analisar relações não lineares complexas entre variáveis de entrada e saída, diretamente aprendendo com os dados de treinamento (Baba and Suto,2020). Vários trabalhos utilizam redes neurais na previsão de séries temporais (BUSTOS), sendo uma das técnicas mais utilizadas para prever tendências do mercado financeiro justamente pelo seu bom desempenho (MARTINEZ).

Diante desse exposto e atrelado ao fato de as abordagens no país ainda serem escassas, permitindo assim, que uma vasta gama de problemas seja explorada (GOMBOSKI) pode-se verificar a importância do estudo desses modelos de predição de dados. Apropriando-se melhor do conhecimento, são diversos os ganhos econômicos que podemos ter com a utilização da IA na economia, novas tecnologias podem ser incorporadas para melhorar a tomada dessas decisões em empresas e órgãos públicos, considerando diversos aspectos que incluem os pontos de vista econômico, ético, político e até mesmo o operacional (SILVA, DE FRANÇA).

Esse trabalho propõe-se a uma revisão de artigos anteriores sobre o tema em questão, buscando assim compreender os resultados obtidos nos mesmos, bem como adquirir um embasamento teórico e planejar a metodologia utilizada neste documento. Posteriormente foram selecionados os modelos de Regressão linear e Redes Neurais, que são os mais adequados para esse tipo de previsão de acordo com estudos de autores anteriores acerca do mesmo tópico.

2. Metodologia

Essa pesquisa partiu da seguinte questão norteadora: como uma inteligência artificial (IA) poderá prever ganhos econômicos juntamente com Sistemas Distribuídos e Machine Learning, na criação de um Sistema Especialista em Econometria?, trata-se de uma revisão integrativa, processo pelo qual permite a busca e o embasamento teórico através de fontes seguras de grande relevância para o conhecimento e apropriação dos resultados (SOARES et al., 2014), realizado no período de junho a outubro de 2022, utilizando-se de dados históricos com as informações sobre o valor na bolsa de São Paulo da empresa Petrobrás como objeto de estudo, sobre as estratégias de PICO (Problema, Intervenção, Controle, Outcome).

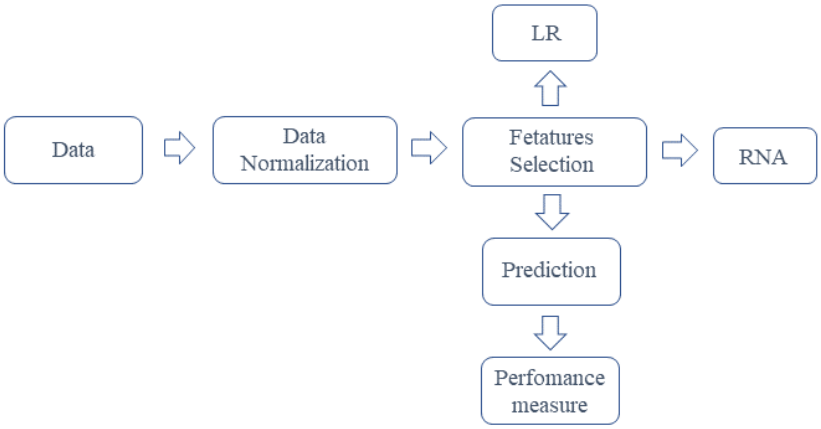


Figura 1: Diagrama da Metodologia.

Comentado [CB1]: Colocar o titulo da figura acima da mesma e abaixo a fonte, ou seja o local de onde foi retirada a imagem.

Foram utilizadas todas as principais etapas do aprendizado de máquina, iniciando-se pela abordagem dos dados que trazem informações da data do pregão, nome da ação, preço de abertura, preço máximo, preço mínimo, preço de fechamento e quantidade de negócios fechados no dia, entre o período do ano de 2020 e 2022 para a construção do Data Frame.

Após a obtenção dos dados, a rotulação consistiu-se basicamente na data de fechamento e valor de fechamento, tendo como classe principal o valor da ação. O balanceamento dos dados não se fez necessário uma vez que a base se mostrou muito concisa e coerente com suas informações, as principais *features* extraídas para o modelo foram a data do pregão e o preço de fechamento.

O treinamento deu-se da seguinte forma: 75% dos dados utilizados para treino, 24% para testes e 1% para validação e verificação da acurácia do modelo gerado. Após a extração dos dados preditos, os mesmos foram submetidos a mais uma etapa de avaliação fazendo uso das principais técnicas de cálculo de erro e métricas de avaliação de séries temporais

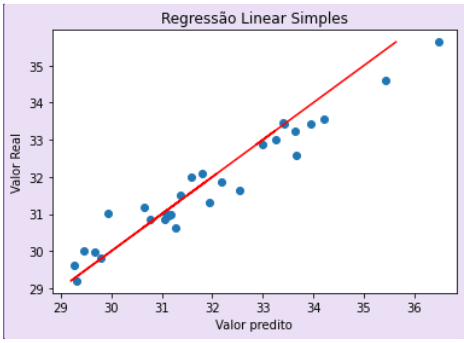
Os algoritmos citados foram treinados para produzir previsões de 30 dias. Portanto foi observado os resultados obtidos dos modelos e feito um comparativo estatístico de seus desempenhos através do valor observado e o valor previsto no dia de fechamento do pregão, sendo o desvio entres estes dois valores o erro previsto.

3. Regressão Linear

Segundo Pasim (2004), a regressão é uma técnica estatística multivariada que possibilita prever os valores de uma variável, com base nos valores de uma variável independente (regressão linear simples), ou de diversas variáveis independentes (regressão linear múltipla).

De acordo com Andressa (2021), um modelo de regressão simples inclui somente duas variáveis: uma independente e uma dependente. A variável dependente é aquela que está sendo explicada, enquanto a variável independente é aquela que é utilizada para explicar a variação na variável dependente. Um modelo de regressão linear é uma equação matemática que fornece uma relação linear, ou seja, de linha reta entre duas variáveis, comumente chamada de X e Y.

Quanto aos tipos de regressão linear existem duas categorias: a regressão linear simples e a regressão linear múltipla. O modelo de regressão linear simples define-se como a relação linear entre a variável dependente e uma variável independente. Enquanto que na regressão linear múltipla assume-se que exista uma relação linear entre uma variável dependente e várias variáveis independentes (RODRIGUES; NUNES, 2012).



Comentado [CB2]: idem

Figura 2: Exemplo de uma Reta da Regressão Linear do a partir do modelo criado.

4. Redes Neurais

As redes neurais são compostas por uma determinada quantidade de entradas e unidades de processamento, as quais são ligadas através de pesos sinápticos. As entradas são propagadas através da topologia da RNA, sendo transformadas pelos pesos sinápticos e pela função de ativação (AF) dos neurônios (MACHADO E FONSECA JÚNIOR, 2013). As RNAs são comumente utilizadas na resolução de problemas complexos, onde o comportamento das variáveis não é rigorosamente conhecido. Uma de suas principais características é a capacidade de aprender por meio de exemplos e de generalizar a informação aprendida, gerando um modelo não-linear, tornando sua aplicação na análise espacial bastante eficiente (SPÖRL et al., 2011).

O algoritmo de aprendizado mais utilizado para treinamento das redes MLP é o algoritmo de retropropagação, que utiliza pares (entrada e saída desejada) para ajustar os pesos por meio de um mecanismo de correção de erros. O treinamento ocorre em duas fases, em que cada fase percorre em um sentido. Estas duas fases são chamadas de fase forward e fase backward. A fase forward é utilizada para definir a saída da rede para um dado padrão de entrada. A fase backward utiliza a saída desejada e a saída fornecida pela rede para atualizar os pesos de suas conexões (BRAGA; LUDERMIR; CARVALHO, 2000, p.59)

Comentado [CB3]: palavras de origem estrangeira são escritas em italico

5. Análise dos Dados

Com o objetivo de medir e analisar a qualidade dos modelos recorreu-se a três técnicas de cálculos de métricas diferentes comumente usadas: a média das diferenças dos valores absolutos entre a previsão e a realização (Erro Médio Absoluto - MAE); a diferença entre o valor predito com o real elevado ao quadrado (Erro Quadrático Médio - MSE) e a raiz quadrada da média das diferenças ao quadrado dos valores previstos e realizados (Raiz Quadrada do Erro Médio - RMSE).

5.1 Erro Médio Absoluto(MAE)

Segundo WILLMOTT(2005),o cálculo para o Erro Médio Absoluto se resume na soma das magnitudes, obtendo assim o "erro total", que posteriormente é dividido pelo número de magnitudes envolvidas. Com isso, é encontrado uma média de erro entre os valores previstos pelo modelo e os valores reais que foram utilizados nos processos de teste.

$$\mathbf{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} | \boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x} |$$

5.2 Erro Quadrático Médio (MSE)

Já o paramêtro de Erro Quadrático Médio(MSE) consiste num cálculo similar, porém o mesmo consiste na elevação do resultado ao quadrado antes de realizar a divisão. A razão para esse artifício seria para que seja trivial analisar possíveis anomalias no modelo, já que um “desvio” maior será evidenciado nos cálculos.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

5.3 Raiz Quadrática do Erro Médio (RMSE)

Segundo WILLMOTT(2005), o cálculo do RMSE é realizado em 3 etapas. O 'erro quadrático total' é obtido primeiro como a soma dos erros quadráticos individuais; ou seja, cada erro influencia o total na proporção de seu quadrado, ao invés de sua magnitude. Grandes erros, como resultado, têm uma influência relativamente maior no erro quadrado total do que os erros menores. Isso significa que o erro quadrado total crescerá à medida que o erro total for concentrado em um número decrescente de erros individuais cada vez maiores. O erro quadrático total é então dividido por n, o que resulta no erro quadrático médio (MSE). Após isso é realizada a raiz quadrada do erro quadrático médio. Como a divisão pelo número de magnitudes e a raiz quadrada escalam o erro quadrático total, segue-se que o MSE e consequentemente o RMSE também irão aumentar a medida que a variância associada e distribuição de frequências das magnitudes de erro aumenta. Com isso, podemos concluir que esse é um método eficaz na detecção de anomalias nos dados previstos pelo modelo em questão.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Figura 3

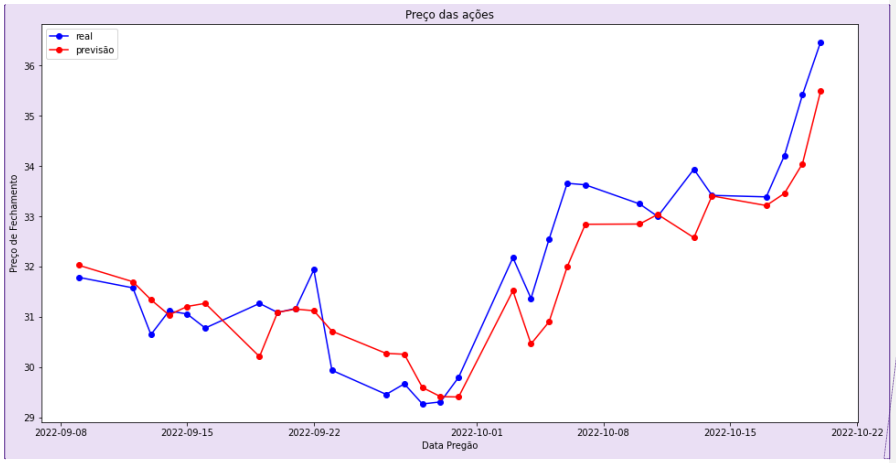
Algoritmo	Acurácia	MAE	MSE	RMSE
Linear Regression	76.19	0.866	1.306	1.143
RNA	78.53	0.800	1.178	1.085

Comentado [CB4]: qual o titulo da tabela? Quem a elaborou? Colocar a fonte

Após apurados os resultados obtidos, e como exemplificado na Figura 3, pode-se considerar que o algoritmo de redes neurais apresentou um melhor índice de precisão com um coeficiente de determinação cerca de 4,92% melhor que o de regressão linear, com uma porcentagem de erro absolutos inferior 0,30%.

O modelo linear utilizado para o cálculo regressão linear foi o média-móvel, e mostrou uma correlação adequada com os valores das séries apresentadas no período, como pode ser evidenciado na Figura 2 através da inclinação positiva de sua reta, indicando uma linearidade satisfatória.

Na figura 4 , percebe-se no gráfico de sazonalidade uma semelhança entre a tendência azul (valores reais) e a tendência vermelha (valores preditos), existindo momentos em que os pontos se encontram, sendo os mesmos valores. Indicando assim um bom desempenho do algoritmo nos resultados obtidos pela Regressão Linear.



Comentado [CB5]: idem

Figura 4: Gráfico de previsões do valor da ação utilizando Regressão Linear.

Sobre o modelo de Redes Neurais, foi utilizado no seu treinamento o parâmetro de épocas igual a 4000 e até 4 camadas escondidas; podendo-se perceber que o desempenho melhorou à medida que aumentamos o números de iterações (épocas) com a RNA, comportamento semelhante nota-se também com o número de camadas escondidas. Entretanto, em determinado momento o aumento dessa variável não proporciona mais resultados satisfatórios no coeficiente de determinação.

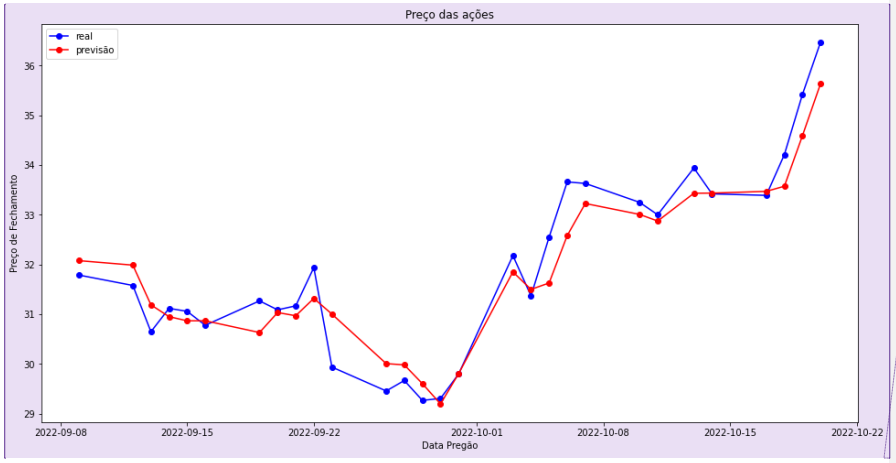


Figura 5: Gráfico de previsões do valor da ação utilizando RNA.

Comentado [CB6]: idem

6. Resultados

Analisando o resultado da pesquisa, observa-se que a tendência linear foi bastante considerável para a série, apresentando um bom desempenho. No entanto o modelo de Redes Neurais demonstrou um aproveitamento melhor, provando ser um algoritmo complexo e que atende bem a qualquer cenário, induzindo a mais pesquisas para um melhor entendimento de seu algoritmo.

As redes neurais possuem camadas ocultas que interagem com a camada de entrada provocando diferentes resultados na camada de saída, esse número pode variar de acordo com o melhor resultado pretendido no coeficiente de determinação. O número de épocas que interagem com esses neurônios também segue a mesma linha de comportamento.

A previsão do comportamento dos preços das ações é uma tarefa complexa, no entanto as novas tecnologias podem ajudar a análise e tomada de decisões que necessitam da previsão de dados futuros. Ambos os modelos podem ser usados para auxiliar na tomada de decisões financeiras e mostram-se bastantes eficientes no que se propõe.

Referências

Comentado [CB7]: por que as referências estão com fonte tamanho 10?

BUSTOS, Oscar; POMARES-QUIMBAYA, Alexandra. Stock market movement forecast: A systematic review. **Expert Systems with Applications**, v. 156, p. 113464, 2020.

DA SILVA, Everton Josué. Modelagem e aplicação de técnicas de aprendizado de máquina para negociação em alta frequência em bolsa de valores. 2015.

FIGUEIREDO FILHO, Dalson et al. O que fazer e o que não fazer com a regressão: pressupostos e aplicações do modelo linear de Mínimos Quadrados Ordinários (MQO). **Revista Política Hoje**, v. 20, n. 1, 2011.

GOMBOSKI, Matheus. A utilização de algoritmos de machine learning na análise econômica. 2019.

HOFFMANN, Rodolfo. Análise de regressão: uma introdução à econometria. 2016.

KUMAR, R. R. Visualizing Big Data Mining: Challenges, Problems and Opportunities. v. 6, n. 4, p. 3933–3937, 2015.

MACHADO, W, C.; FONSECA JÚNIOR, E. S. Redes Neurais Artificiais aplicadas na previsão do VTEC no Brasil. Boletim de Ciências Geodesicas, v.19, n.2, p. 227-246, 2013.

MARTINEZ, Leonardo C. et al. From an artificial neural network to a stock market day-trading system: A case study on the bm&f bovespa. In: **2009 International Joint Conference on Neural Networks**. IEEE, 2009. p. 2006-2013.

MARONI NETO, Ricardo et al. Machine Learning aplicada a finanças: Previsão por meio de indicadores econômicos-financeiros. 2021.

NIED, A. Treinamento de redes neurais artificiais baseado em sistemas de estrutura variável com taxa de aprendizado adaptativa. 2007. Belo Horizonte, MG. Tese (Doutorado) – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais.

PATRIKAINEN, Tuomas. Using machine learning to forecast long-term equity price movement. **Journal of Applied Finance & Banking**, v. 7, n. 1, p. 1-40.

SILVA, Vitor Hugo Miro Couto; DE FRANÇA, João Mário Santos. MODELOS DE MACHINE LEARNING NA CLASSIFICAÇÃO DE POBREZA: UMA APLICAÇÃO PARA O ESTADO DO CEARÁ.

STEFFEN, Andressa Antunes et al. Predição de séries temporais aplicada ao mercado de ações utilizando regressão linear. 202

XIANGYAN, Pan. Prediction Algorithm of Digital Economy Development Trend Based on Big Data. **Mathematical Problems in Engineering**, v. 2022, 2022.]

VALENCIA, F.; GÓMEZ-ESPINOSA, A.; VALDÉS-AGUIRRE, B. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy, v. 21, n. 6, 2019.

WILLMOTT, Cort J.; MATSUURA, Kenji. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate research**, v. 30, n. 1, p. 79-82, 2005.