UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

JHONATAN KAZUHIRO TANABE

Uma comparação entre classificadores de aprendizado de máquina para predição de preços no mercado acionário

Campinas
2022

JHONATAN KAZUHIRO TANABE

Uma comparação entre classificadores de aprendizado de máquina para predição de preços no mercado acionário

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do Prof. João B. Florindo.

Sumário

Lista de Figuras

1 Introdução

O mercado financeiro é de fundamental importância para o desenvolvimento da sociedade, sendo que instituições financeiras como bancos e fundos de investimentos desempenham papel crucial no crescimento das empresas e no desenvolvimento econômico e social global ou de uma nação [1].

Um componente essencial desse universo são as ações negociadas em bolsas de valores ao redor do mundo. Ações de determinada empresa negociadas em bolsa representam a participação de seus investidores individuais ou institucionais nos lucros dessa empresa.

O presente trabalho visa utilizar classificadores de Aprendizado de Máquinas para prever sinais de posicionamento de compra ou venda de ações da empresa Nubank (NU) no mercado. As ações da *fintech* foram escolhidas para a análise, em detrimento de algum ativo com maior volume de negociações que componha índices como o Ibovespa, por ser um *case* brasileiro de destaque em que uma *fintech* brasileira abriu capital em uma bolsa internacional. Ademais, por ser uma *smallcap*, a análise pode trazer *insights* valiosos e específicos sobre a empresa e sobre o setor como um todo que podem não ser capturados em grandes índices.

O trabalho tem como principal fim fornecer ferramentas adicionais às ferramentas de análise tradicionais para tomadas de decisão em *trading*.

2 Base de dados

A base de dados utilizada para treinar e testar os modelos consiste na série histórica de preços, entre 26/04/2022 e 25/04/2023, da ação NU (Nu Holdings LTD) negociada em dólares na bolsa de valores de Nova Iorque (NYSE).

Originalmente, a base continha 251 observações com 7 colunas. A fim de facilitar a análise, uma manipulação foi performada na base original e a coluna "Date" com as datas das negociações foi desconsiderada, além dos valores faltantes terem sido removidos. Desse modo, o *dataset* definitivo é mostrado na Figura 1.

Data	Abertura	Max	Min	Fechamento	Preco Ajust	Volume
2022-04-26	6.81	6.81	6.22	6.31	6.31	8152300
2022-04-27	6.36	6.53	6.28	6.33	6.33	6803200
2022-04-28	6.33	6.40	5.90	6.16	6.16	14417100
2022-04-29	6.19	6.34	5.91	6.01	6.01	7716600
2022-05-02	6.02	6.28	5.90	5.96	5.96	15326100

Figura 1: Cinco primeiras observações do *dataset* escolhido para análise

As variáveis do *dataset* são explicadas por:

- Abertura: preço da ação na abertura do pregão
- Max: Preço máximo da ação na data indicada
- Min: Preço mínimo da ação na data indicada
- Preco Ajust: Preço no fechamento ajustado após desconto de dividendos
- Volume: Número de ações negociadas na data indicada

3 Metodologia

Para a aplicação dos classificadores de aprendizado de máquinas, dois atributos foram criados: o Movimento, que indica a movimentação dos preços no dia indicado, e Volatilidade, que indica a volatilidade dos preços por dia de pregão. Os valores foram dispostos em forma de *dataframe* utilizando as bibliotecas *pandas* e *NumPy*, o qual é mostrado na Figura 2, assim como a série temporal represantando a variação da medida, que pode ser observada na Figura 3.

Data	Movimento	Volatilidade
2022-04-26	-0.50	0.59
2022-04-27	-0.03	0.25
2022-04-28	-0.17	0.50
2022-04-29	-0.18	0.43
2022-05-02	-0.06	0.38

Figura 2: Cinco primeiras observações do *dataframe* criado com o movimento e volatilidade dos preços

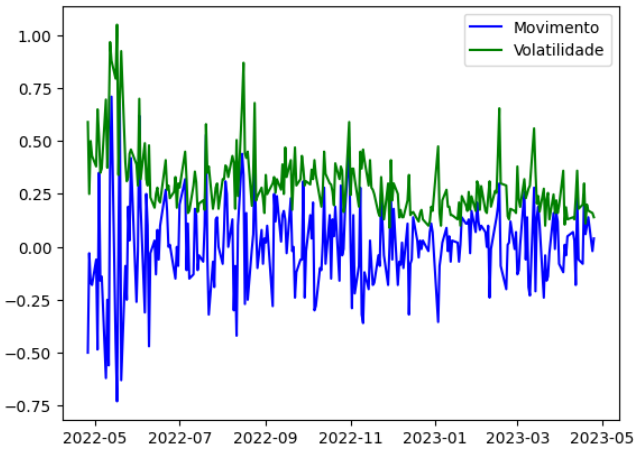

Figura 3: Série temporal dos atributos Movimento e Volatilidade

Em seguida, foi criado um *array* com valores de *trading signals* para indicar se o dia seguinte possui preço maior do que o dia anterior, ambos no fechamento, com 1 sendo um sinal para posicionamento com $Pf > Pi$ e 0 sendo um sinal para não-posicionamento com $Pf < Pi$, em que Pi é o preço da ação no fechamento na data analisada e Pf é o preço da ação no fechamento do dia seguinte. O *array* foi transformado em *dataframe* e os cinco primeiros valores podem ser observados na Figura 4.

	Data	Sinal
0	2022-04-26	1
1	2022-04-27	0
2	2022-04-28	0
3	2022-04-29	0
4	2022-05-02	0

Figura 4: Cinco primeiras observações do *dataframe* criado a partir do *array* com os sinais

Com base nos atributos criados, o *dataset* foi dividido em conjuntos de treino e teste, sendo 75% dos dados para treino e 25% para teste. A métrica utilizada para a avaliação da performance dos modelos será a acurácia, expressa por:

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN}, \tag{1}$$

em que:

VP = Verdadeiro Positivo

VN = Verdadeiro Negativo

FP = Falso Positivo

FN = Falso negativo

A acurácia foi escolhida como melhor métrica pois os dados do problema são balanceados, de modo que a mesma funciona de forma simples para compreender a eficácia do classificador testado.

Após a comparação entre os classificadores, foram calculadas médias móveis com variação entre janelas para analisar a tendência dos preços conforme as janelas escolhidas.

3.1 Modelos

Os classificadores testados serão comparados a fim de se obter a melhor acurácia na previsão de *trading signals* a partir da série temporal de preços da ação escolhida e os parâmetros serão dados pelo *default* da biblioteca Scikit-Learn.

Os modelos escolhidos para análise são: Máquinas de Vetores de Suporte, Perceptron, Naive Bayes, Regressão Logística, Floresta Aleatória e K-ésimo Vizinho mais Próximo.

3.1.1 Maquinas de Vetores Suporte (SVM)

Máquinas de Vetores Suporte são um modelo que classifica os dados separando-os por meio de um hiperplano. Os dados adjacentes à margem recebem o nome de vetores de suporte e devem estar o mais distante possível do hiperplano. O modelo pode ser esquematizado pela Figura 5.

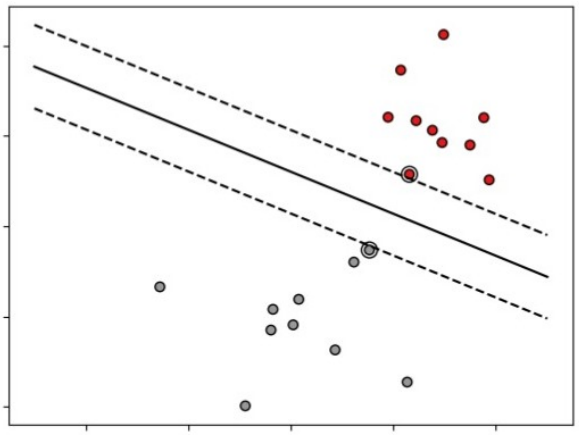

Figura 5: Gráfico que esquematiza o modelo de Maquinas de Vetores Suporte

O algoritmo resolve um problema de otimização convexa, ou seja, maximiza a margem garantindo que cada observação esteja do lado correto do hiperplano. Essa otimização pode ser formalizada por:

$$max_{w,b}||w||^{-2} \tag{2}$$

$$w^T \cdot x_i + b \geq 1 \qquad \forall x_i \in C_1 \tag{3}$$

$$w^T \cdot x_i + b \leq -1 \qquad \forall x_i \in C_2 \tag{4}$$

em que w é o vetor com os pesos, xi são as entradas e b é o termo que representa o viés do modelo.

3.1.2 Perceptron

O Perceptron é um classificador linear que simula um neurônio biológico e serve como substrato para a construção de redes neurais mais complexas, como as redes neurais de Aprendizado Profundo. O modelo recebe os dados, aplica pesos e devolve resultados a partir de uma função de ativação.

Sua estrutura com as entradas $(x_1, x_2 ... x_p)$ e os pesos $(w_1, w_2 ... w_p)$ para aplicação na função de ativação $F(z)$ e predição da variável resposta $\hat{y}$ pode ser observada na Figura 6

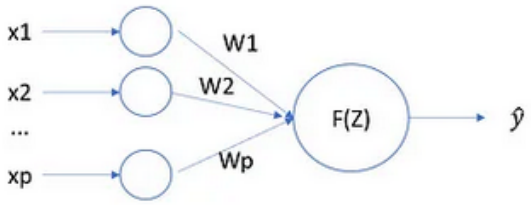

Figura 6: Esboço esquemático do Perceptron.
Fonte:
https://medium.com/analytics-vidhya/understanding-the-perceptron-algorithm-4a368f493109

O modelo recebe os dados, aplica pesos e devolve resultados binários a partir de uma função de ativação dada por:

$$F(z) = \sum_{j=1}^{p} w_j x_j \begin{cases} 1 & se \quad z \geq \theta \\ 0 & se \quad z \leq \theta \end{cases} \tag{5}$$

em que w_j é o vetor com os pesos, x_j é o vetor com as entradas e p é a quantidade de entradas e pesos.

A função de ativação mais adequada irá depender do problema e pode ser uma função sigmoide, tangente hiperbólica, etc.

3.1.3 Regressão Logística

A Regressão Logística é um algorítmo que busca classificar os dados com a variável resposta binária. O modelo estima a probabilidade da variável resposta, 0 ou 1, a partir da função sigmoide:

$$p(y = 1|x) = f(x) = \frac{1}{1 + e^{-z}}, \tag{6}$$

em que y é a variável resposta, x é a entrada dos dados e z a combinação linear entre os valores da entrada e seus coeficientes.

3.1.4 Floresta Aleatória

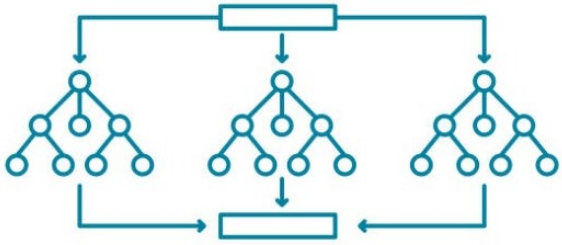

Figura 7: Esquematização da Floresta Aleatória que indica a entrada, suas árvores de decisão e a saída.
Fonte: https://stock.adobe.com/br/images/random-forest-line-icon-decision-trees-symbol-machine-learning-technique-that-s-used-to-solve-regression-and-classification-problems-complex-problems-solution-vector-illustration-flat-clip-art/474661732

Floresta Aleatória é um modelo de classificação que utiliza métodos de *ensemble*, isto é, combina vários modelos em um só para obter o melhor resultado na predição da variável resposta.

No caso da Floresta Aleatória, o classificador combina diversas árvores de decisão (modelos mais simples) para obter uma classificação.

3.1.5 K-ésimo Vizinho mais Próximo

O K-ésimo Vizinho mais Próximo é um classificador que busca semelhança entre as variáveis da entrada, para classificá-las.

O modelo não possui fases de aprendizagem e realiza sua classificação com base na distância entre os dados em vizinhanças especificadas.

3.2 Média Móvel

A média móvel simples é denotada por :

$$MMS = \frac{A_1 + A_2 + ... + A_n}{n} \tag{7}$$

Onde: MMS = Média Móvel Simples

A = Valor no período indicado

n = Número de períodos

No contexto de análise de séries temporais financeiras, a média móvel funciona como um inibidor de ruídos por computar a média da variação de preços sem considerar as pequenas flutuações que não indicam tendência de queda ou crescimento.

4 Resultados

Após treinar e testar os modelos selecionados, verificou-se que, ao utilizar a métrica escolhida, o melhor desempenho se deu pelo modelo de Máquinas de Vetores Suporte (SVM), com uma acurácia de 0.62. Este resultado pode ser explicado pela performance do modelo frente aos possíveis *outliers* presentes nos dados e bom ajuste aos dados não-lineares, como numa série temporal financeira.

O gráfico da Figura 8 dispões todos os resultados de acurácia para cada modelo.

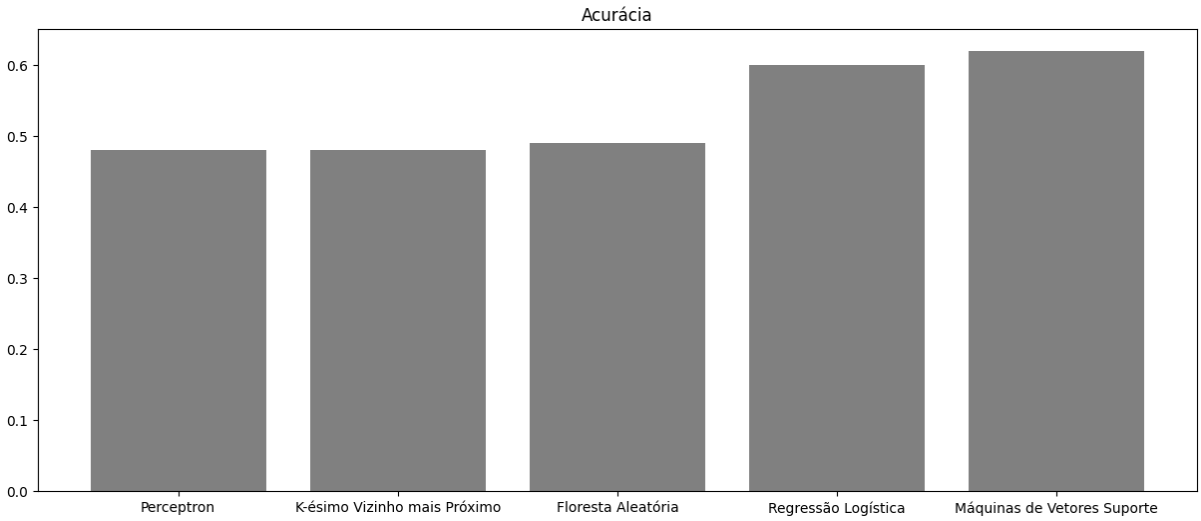

Figura 8: Gráfico de barras que compara a acurácia de cada modelo testado

Após a avaliação do modelo escolhido, médias móveis simples foram calculadas no conjunto de dados de treino e teste, afim de eliminar ruídos e reavaliar a performance do SVM.

Os intervalos para a média móvel simples foram 5, 10 e 30, inferiores aos intervalos comumente adotados em análises técnicas do mercado financeiro pelo conjunto de dados ser pequeno.

O gráfico de barras com a acurácia do SVM para cada janela utilizada na média móvel simples pode ser consultado na 9

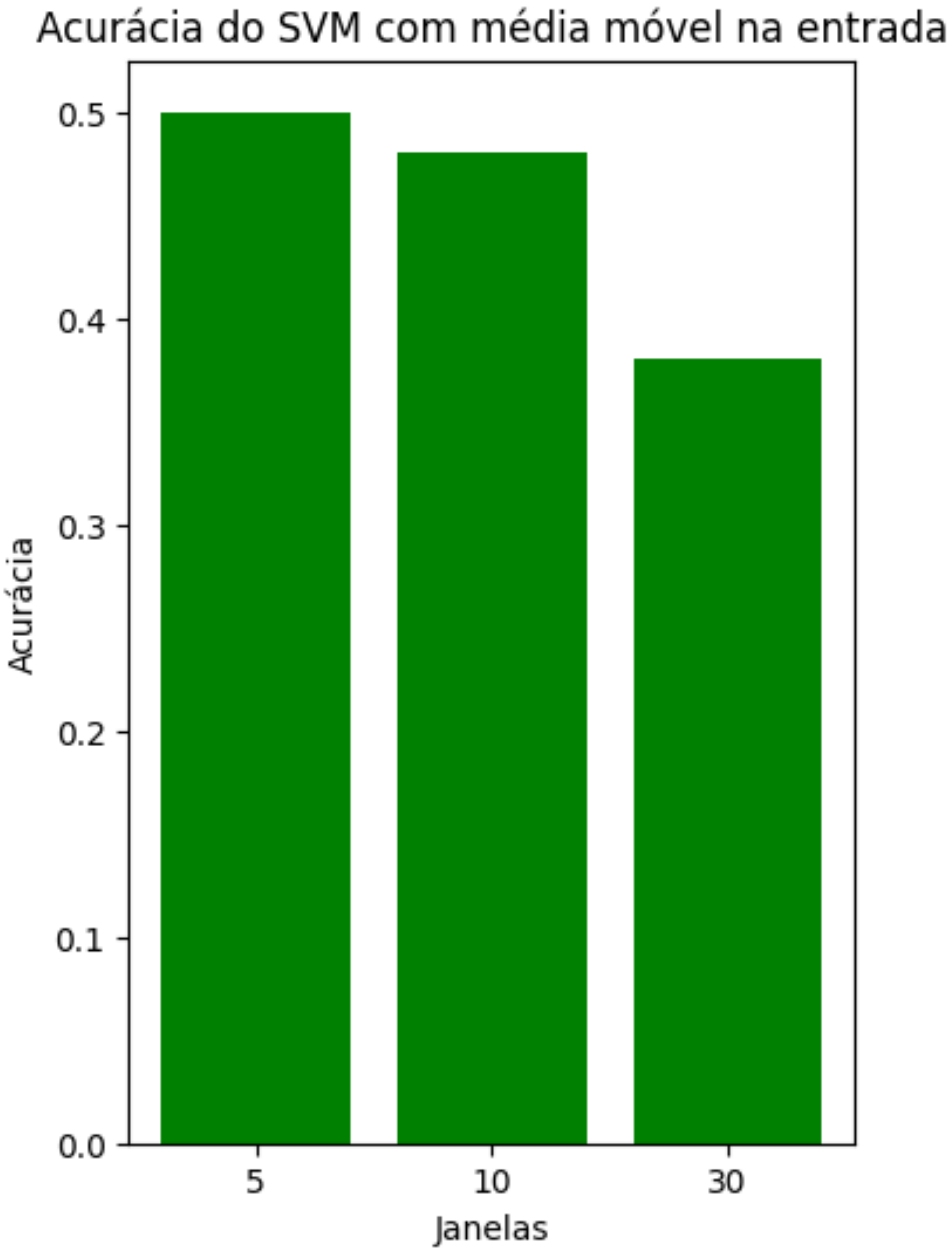

Figura 9: Gráfico de barra que compara as acurácias para as duas janelas, escolhidas na média móvel, para o modelo escolhido (SVM)

A partir da reavaliação das acurácias do modelo após aplicação das médias móveis, pôde-se observar que o modelo perdeu precisão, chegando a uma acurácia menor do que 0.40, equivalente a um chute aleatório simples para a direção do preço da ação.

Uma hipótese a ser levantada a respeito da redução da acurácia após a aplicação de médias móveis simples aos dados que alimentam o modelo é que ao utilizar a técnica estatística para eliminar ruídos, informações relevantes a partir das pequenas flutuações de preço

são perdidas, principalmente ao analisar o curto prazo.

5 Conclusão

Pode-se concluir que é possível aplicar algorítmos de classificação em Aprendizado de Máquinas para estimar a direção do preço de determinada ação no mercado, sobretudo o modelo de Máquinas de Vetores Suporte [2].

Dessa forma, é possível utilizar os modelos para auxiliar, como suporte à análise técnica e estratégias de *trading* na tomada de decisão para posicionamento.

No entanto, a depender de como se comportar os dados de entrada do modelo, a acurácia pode diminuir e a efetividade preditiva pode ser comparada a um simples *coin flip*.

A partir deste trabalho, é possível iniciar, posteriormente, uma investigação a respeito dos aspectos teóricos dos algorítmos e, com um eventual conhecimento de domínio mais amplo em Finanças, tirar melhores conclusões preditivas acerca dos preços no mercado acionário.

Referências

1. Beck, T. *The role of finance in economic development: benefits, risks*; rel. técn.; and politics. Discussion Paper, 2011.
2. Madge, S. *Predicting Stock Price Direction using Support Vector Machines. Independent Work Report Spring.(2015); Applied Mathematics Journal of Hindawi www. hindawi. com*, *2018*.
3. Autor, J. F. Investopedia, `https://www.investopedia.com/terms/m/movingaverage.asp`, Acessado em 23 de junho de 2023.
4. Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*; "O'Reilly Media, Inc.": 2022.
5. Autor, V. A. Medium, `https://medium.com/analytics-vidhya/understanding-the-perceptron-algorithm-4a368f493109`, Acessado em 23 de junho de 2023.
6. Autor, C. Medium, `https://medium.com/mlearning-ai/data-science-machine-learning-models-metrics-77f9f77c2ff4`, Acessado em 25 de junho de 2023.
7. Autor, R. D. Medium, `https://medium.com/mlearning-ai/a-little-about-perceptrons-and-activation-functions-aed19d672656`, Acessado em 26 de junho de 2023.