https://eventos.utfpr.edu.br//sicite/sicite2020

Aprendizado conjunto aplicado na predição do mercado de ações brasileiro

Ensemble learning applied in the prediction of the Brazilian stock market

Alvaro Pedroso Queiroz
alvaroq@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Giovani Volnei Meinerz
giovaniimeinerz@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

RESUMO

A previsão do mercado de ações é considerada uma tarefa complexa, envolvendo áreas interdisciplinares como estatística, economia e computação. Com o avanço das técnicas de *machine learning*, tal prática se tornou um recurso bastante utilizado para a maximização de lucros do mercado. É crescente o número de investidores e pesquisadores desse âmbito em cenário nacional, no entanto, a utilização de modelos que fazem combinação de técnicas de *machine learning* ainda não foi explorada. Este trabalho apresenta o *Stock Market Ensemble Predictor* (SMARTER), um modelo de aprendizado de máquina que combina diferentes técnicas de regressão, desenvolvido com o objetivo de realizar análises preditivas sobre dados históricos do mercado de ações brasileiro, visando aumentar a acurácia da precisão por meio da combinação dos resultados de múltiplas abordagens. Obteve-se com as técnicas *OLS* e *Bayesian Ridge,* aliado à técnica de *Voting,* o maior coeficiente de determinação R^2 médio dentre as combinações testadas. O valor alcançado foi de 0,914864, formando um modelo mais confiável, aumentando a acurácia da precisão do resultado.

PALAVRAS-CHAVE: Bolsa de valores. Previsão. Aprendizado do computador.

ABSTRACT

FORECASTING THE STOCK MARKET IS CONSIDERED A COMPLEX TASK, INVOLVING INTERDISCIPLINARY AREAS SUCH AS STATISTICS, ECONOMICS AND COMPUTING. WITH THE ADVANCEMENT OF MACHINE LEARNING TECHNIQUES, THIS PRACTICE HAS BECOME A WIDELY USED RESOURCE FOR MAXIMIZING MARKET PROFITS. THE NUMBER OF INVESTORS AND RESEARCHERS IN THIS FIELD IS GROWING IN THE NATIONAL SCENARIO, HOWEVER, THE USE OF MODELS THAT COMBINE MACHINE LEARNING TECHNIQUES HAS NOT YET BEEN EXPLORED. THIS PAPER PRESENTS THE STOCK MARKET ENSEMBLE PREDICTOR (SMARTER), A MACHINE LEARNING MODEL THAT COMBINES DIFFERENT REGRESSION TECHNIQUES, DEVELOPED WITH THE OBJECTIVE OF PERFORMING PREDICTIVE ANALYZES ON HISTORICAL DATA FROM THE BRAZILIAN STOCK MARKET, AIMING TO INCREASE THE ACCURACY OF PRECISION THROUGH COMBINING THE RESULTS OF MULTIPLE APPROACHES. WITH THE OLS AND BAYESIAN RIDGE TECHNIQUES, COMBINED WITH THE VOTING TECHNIQUE, THE HIGHEST AVERAGE COEFFICIENT OF DETERMINATION R^2 AMONG THE TESTED COMBINATIONS WAS OBTAINED. THE VALUE REACHED WAS 0.914864, FORMING A MORE RELIABLE MODEL, INCREASING THE ACCURACY OF THE PRECISION OF THE RESULT.

KEYWORDS: Stock exchanges. Forecasting. Machine learning.

INTRODUÇÃO

O mercado de ações é o ambiente no qual empresas de capital aberto e que estejam listadas na bolsa, negociam frações de seu patrimônio através de operações de compra e venda. Essas frações denominam-se ações, que correspondem à participação que o investidor tem em uma empresa, representando um direito sobre os ativos e lucros dessa companhia (NETO, 2018, p. 294-295). O setor acionário brasileiro está concentrado em uma bolsa de valores denominada Brasil, Bolsa, Balcão (B3).

A previsão da oscilação do mercado de ações é considerada uma tarefa complexa para muitos analistas da área de finanças quantitativas, tendo características interdisciplinares, envolvendo áreas como estatística, economia e computação (PASUPULETY, 2019).

Com o avanço dos algoritmos de *machine learning* nos domínios financeiros, essa prática se tornou um dos principais recursos na análise técnica atualmente, podendo estipular modelos que possam maximizar os lucros do mercado (LIU, 2017, p. 228-229).

Em termos estatísticos, segundo Pahwa e Agarwal (2019), existem métodos de regressão linear que utilizam valores e atributos e estabelecem um relacionamento entre eles, sendo muito utilizado devido a sua simplicidade e eficácia na previsão. Enquadrando esses conceitos no aprendizado de máquina, é possível adaptar a mesma técnica utilizando recursos para treinamento de um classificador que prediz o valor do rótulo com determinada precisão.

Dessa forma, existe uma gama de técnicas de *machine learning* presentes em diferentes métodos, capazes de realizar previsões através de dados. Logo, surge a indagação de que tipo de técnica deve ser escolhida dentre as muitas existentes. Uma das alternativas é o emprego do *ensemble learning*, em que é possível estabelecer um modo de aprendizado capaz de realizar a união de várias técnicas de *machine learning* (ZHANG & MA, 2012, p. 1-2).

De acordo com Polikar (2009), um sistema baseado em *ensemble* é obtido por meio da combinação de vários modelos que, dentre suas aplicações, pode ser utilizado principalmente para melhorar o desempenho de um modelo ou reduzir a probabilidade de uma má seleção em um modelo ruim.

Assim, ponderado o impacto e a crescente quantidade de estudos publicados sobre o mercado acionário aliado a métodos de *machine learning*, este trabalho tem por objetivo desenvolver um modelo de análise preditiva baseado em métodos de aprendizado conjunto (*ensemble learning*), denominado *Stock Market Ensemble Predictor* (SMARTER), para aplicação sobre dados históricos do mercado de ações brasileiro, visando aumentar a acurácia da previsão por meio da combinação dos resultados de múltiplas abordagens.

MATERIAIS E MÉTODOS

Para a realização deste trabalho de pesquisa, foram utilizadas algumas tecnologias, tais como: MongoDB, sistema gerenciador de banco de dados (SGBD) responsável pelo armazenamento dos dados coletados; linguagem de programação Python; Jupyter-Notebook, como ambiente de desenvolvimento

integrado; biblioteca Scikit-learn, para análise preditiva dos dados; e as bibliotecas do projeto *open-source* Anaconda, para utilização de ferramentas importantes para a realização da ciência de dados, como a manipulação dos dados com o Pandas, a utilização de estruturas de dados e manipulação dessas com Numpy e Scipy, a criação de gráficos e visualizações de dados em geral com o Matplotlib e o Seaborn, bem com a utilização de funções matemáticas com o Math.

Dessa forma, para a criação dos modelos, definiu-se um *workflow*, em que se estipulou um conjunto de etapas que sistematicamente transforma e processa os dados com a finalidade de criar soluções preditivas, sendo organizado pelos seguintes processos: análise do problema, preparação dos dados, adequação ao problema, pré-processamento, treinamento, teste e avaliação do modelo.

Na etapa de **análise do problema**, focou-se no conhecimento do problema em questão, na realização de estudos e pesquisas em artigos que permitiram situar conclusões e observações já formalizadas a respeito desse âmbito e consequentemente proporcionar uma definição de resultado esperado.

Em seguida, iniciou-se a etapa de **preparação dos dados**, que foram coletados a partir das cotações históricas da B3 disponíveis em seu site[1]. Os dados foram armazenados no SGBD e em seguida realizou-se uma análise exploratória, com a finalidade de obter uma visão geral, como a identificação de atributos que possuíam valores constantes, os que possuíam valores ausentes, além de obter informações estatísticas destas, para ajudar em tomadas de decisões.

Sendo assim, a quantidade total coletada foi de **2.245.476 registros** em um intervalo temporal compreendido entre 01/01/2015 à 31/05/2019. No entanto, como o presente trabalho visa o mercado acionário, só interessam os registros relacionados aos tipos de mercado **à vista** e **fracionário**. Do total coletado, cerca de **36,4%** correspondem a esses tipos de mercado, obtendo um conjunto de dados a ser trabalhado com **817.789 registros**.

O mercado de ações possui diferentes formas de influências no preço, o que pode gerar oscilações para determinadas ações. Dessa forma, na etapa de **adequação ao problema**, foi realizada uma divisão em diferentes cenários e intervalos temporais, a fim de verificar como os modelos se comportam em tais configurações.

Para a realização dos testes dos modelos, foram definidas ações de **5 empresas,** baseado em sua importância para o mercado acionário (as mais negociadas, com os maiores volumes). São elas: ABEV3, BBDC4, ITUB4, PETR4 e VALE3.

Ademais, foram estabelecidos **cenários** distintos, compostos por diferentes **períodos**, com a ideia de testar o comportamento dos modelos em **intervalos temporais** com diferentes quantidades de registros.

Na sequência, tais intervalos foram divididos em duas situações distintas quanto a volatilidade, obtendo cenários com **alta volatilidade** (períodos em que possui altas variações no preço da ação) e **volatilidade normalizada** (períodos com pouca variação no preço da ação).

[1] URL do site: http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/historico/mercado-a-vista/series-historicas/

Para que os modelos de *machine learning* tenham um bom desempenho, é necessário que esses recebam dados com qualidade e úteis para a predição, evitando possíveis problemas com dados ruidosos, redundantes, perdidos, dentre outros (KOTSIANTIS et al., 2007).

Sendo assim, na etapa de **pré-processamento** foram selecionados os seguintes atributos: **Código de negociação**, para a identificação das ações; **Data do pregão** (data em que a ação foi destinada a fechamento de negócios de compra e venda), para a formulação dos intervalos temporais; **Preço de abertura**, **Preço máximo**, **Preço mínimo** e **Volume total**, para serem os atributos preditores; **Preço de fechamento**, como o atributo a ser predito.

Como os atributos preditores apresentavam unidade de medida diferentes (os preços estão em reais e o volume total em quantidade), tal diferença poderia ocasionar de enviesar os algoritmos para as variáveis com maior ordem de grandeza (SKIENA, 2017, p. 103-104). Logo, a **padronização** desses dados se tornou necessária, utilizando a fórmula **z-score** representada pela Eq. (1).

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

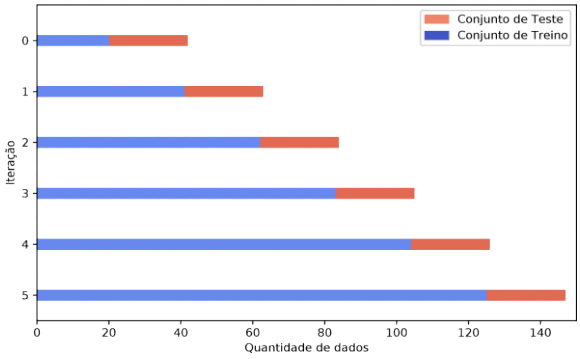
Onde a variável $\mathbf{x_i}$ representa o valor do atributo, $\boldsymbol{\mu}$ é equivalente à média da distribuição do atributo e $\boldsymbol{\sigma}$ o desvio padrão associado. Essa formula foi aplicada para cada valor presente em cada atributo, deixando todos em uma mesma escala.

Após toda a preparação realizada, os dados estavam prontos para serem utilizados para a criação dos modelos. Nessa etapa, o foco principal foi o **treinamento** destes, utilizando diversas técnicas e métodos, a fim de realizar combinações e aperfeiçoamentos, conquistando melhores resultados.

Na realização do treinamento, é muito importante que o **conjunto de treino** seja diferente do **conjunto de teste**, evitando a ocorrência de *overfitting* (LIU, 2017, p. 17). Assim, uma prática interessante é a **validação cruzada**, a qual consiste do treinamento e teste em diferentes subconjuntos de dados, a fim de entender como o modelo é generalizado (BROWNE,2000).

Como os dados possuem características temporais, a divisão dos dados foi estabelecida conforme a técnica **Time Series Split** que, a cada iteração utiliza a quantidade de dados que possui datas de pregões correspondentes aos períodos definidos para a realização do treino do modelo e o próximo período utilizado para teste, respeitando os padrões de tempo, ou seja, evitando que dados futuros sejam utilizados para previsão em tempos passados, como explicitado na Figura 1.

Figura 1 – Divisão de dados para treinamento e teste



Fonte: Próprio Autor (2020).

O **conjunto de treino** é variado, sendo composto pela quantidade de **21 datas de pregão**, correspondentes a um intervalo temporal de **1 mês**. A cada **iteração**, os dados do próximo mês são agrupados ao conjunto de treinamento até que se obtenha um conjunto de treino equivalente a **6 meses**, ou seja, **126 datas de pregão**.

Já o **conjunto de teste** é composto pelos dados que possuem as **próximas 21 datas de pregão** do conjunto de treinamento, ou seja, sempre é testado com os dados do **mês posterior**.

Por fim, na etapa de **avaliação**, estipulou-se uma métrica de avaliação que emite valores úteis para mensurar a eficácia dos modelos testados, a fim de realizar comparações e utilizar como parâmetro para melhorias. Logo, utilizou-se o **coeficiente de determinação R^2** para a verificação de como cada modelo se ajustou aos dados, ou seja, o quanto o modelo é capaz de generalizar e ser efetivo com dados desconhecidos e, assim, realizar comparações de comportamento entre as diferentes técnicas utilizadas.

Tal métrica possui como valor de coeficiente um limite superior igual a 1, ou seja, quanto mais próximo o **coeficiente** estiver do **valor 1**, significa, a princípio, que os dados estão próximos à reta de regressão (MARTINS, 2018).

RESULTADOS E DISCUSSÃO

Esta seção discutirá os resultados alcançados pelos modelos desenvolvidos após a aplicação dos procedimentos descritos.

Na Tabela 1, encontra-se a média dos coeficientes de determinação R^2, obtidos pelos modelos com base em diferentes técnicas, em cenários com alta volatilidade e volatilidade normal.

Tabela 1 – Coeficiente de determinação R^2 médio dos modelos em diferentes cenários de volatilidade

Técnica	Alta volatilidade	Volatilidade Normal
Voting (Ordinary Least Squares + Bayesian Ridge)	0,902867	**0,926860**
Ordinary Least Squares	0,902983	**0,926633**
Bayesian Ridge	0,901908	**0,927001**
Bagging (Bayesian Ridge)	0,900691	**0,927055**
Bagging (Ordinary Least Squares)	0,903929	**0,922914**
Boosting (Ordinary Least Squares)	0,898316	**0,923812**
Stacking (Ordinary Least Squares + Bayesian Ridge)	0,890688	**0,928051**
Boosting (Bayesian Ridge)	0,892249	**0,922674**

Fonte: Próprio Autor (2020).

Percebe-se que dentre os modelos citados, todos possuem maiores valores de coeficiente de determinação R^2 nos cenários com volatilidade considerada normal.

Para tanto, os resultados alcançados para intervalos com alta volatilidade, ainda assim são bem expressivos e com bastante relevância, indicando que mesmo

com a alta variação de preço da ação, os modelos conseguem ter eficácia na previsão.

Portanto, é importante que um modelo de *machine learning* tenha seu desempenho satisfatório em ambos os cenários considerados (WENG et al.,2018). Sendo assim, a Tabela 2 apresenta a média dos melhores coeficientes de determinação R^2 para os modelos desenvolvidos pelas diferentes técnicas, considerando os dois cenários de volatilidade.

Tabela 2 – Coeficiente de determinação R^2 médio dos modelos em ambos os cenários de volatilidade

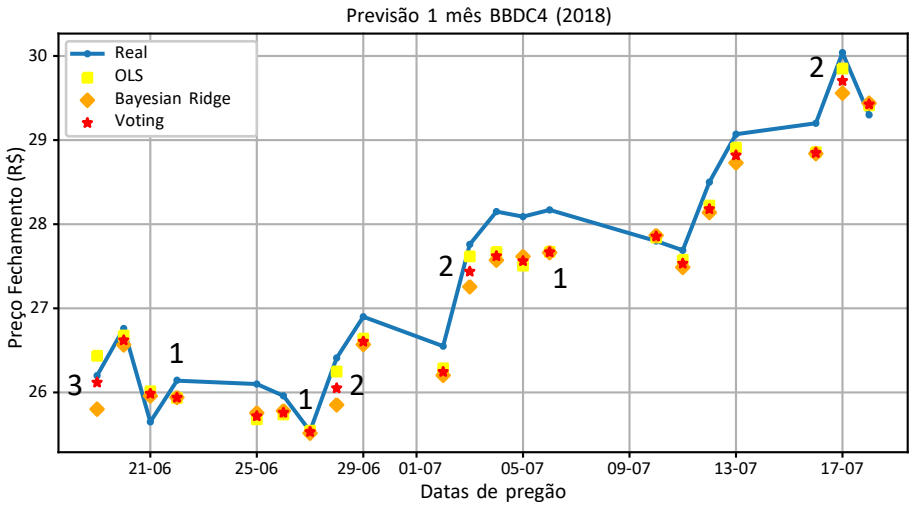
Técnica	Coeficiente de determinação R^2
Voting (Ordinary Least Squares + Bayesian Ridge)	**0,914864**
Ordinary Least Squares	0,914808
Bayesian Ridge	0,914455
Bagging (Bayesian Ridge)	0,913873
Bagging (Ordinary Least Squares)	0,913422
Boosting (Ordinary Least Squares)	0,911064
Stacking (Ordinary Least Squares + Bayesian Ridge)	0,909370
Boosting (Bayesian Ridge)	0,907462

Fonte: Próprio Autor (2020).

Nota-se que a combinação das técnicas *Ordinary Least Squares (OLS)* e *Bayesian Ridge,* por meio de votação, alcança um resultado médio com valores superiores aos obtidos de forma individual.

Dessa forma, aplicou-se um estudo sobre o seu comportamento de uma forma mais aprofundada, buscando identificar os principais motivos que possibilitaram um desempenho robusto ao realizar combinações de técnicas de regressão distintas, sendo destacado na Figura 2.

Figura 2 – Previsão de preço de fechamento pelos modelos para a ação BBDC4 utilizando 21 datas de pregão para treinamento e 21 datas para teste.



Fonte: Próprio Autor (2020)

Conforme ilustrado, é possível estabelecer três situações distintas que ocorrem nas predições dos modelos gerados pelas técnicas citadas. A **primeira situação** é quando as previsões são bem **semelhantes**, obtendo os mesmos resultados ao serem combinados.

Já a **segunda situação**, ocorre quando um determinado modelo realiza uma previsão que mais se **aproxima do valor real** em comparação a previsão do outro modelo. Assim, a combinação ao realizar a intermediação, diminui a taxa de erro que o modelo com menos precisão teve, proporcionando o **equilíbrio** de suas fraquezas individuais, sendo útil para os diferentes tipos de comportamentos desenvolvidos em ações distintas.

Por fim, pode-se observar uma **terceira situação**, em que se possui uma incongruência de previsão entre os dois modelos utilizados (um modelo realiza previsão **acima do valor real** e o outro realiza previsão **abaixo do valor real**) e a combinação atua como um **intermediador**, obtendo valores mais próximos do real.

Vale destacar que a forma de combinação que a biblioteca utilizada implementa a técnica de *Voting*, é através do cálculo da **média** das previsões individuais dos vários regressores de base utilizados, formando a previsão final. Posto isso, foi de extrema importância a combinação de técnicas que geraram modelos com altos desempenhos, influenciando diretamente na melhora provocada com a utilização dessa combinação.

Assim sendo, devido aos resultados promissores alcançados na utilização da combinação das técnicas **OLS** e ***Bayesian Ridge***, empregando a técnica de ***Voting***, tal combinação foi escolhida para ser a base do modelo **SMARTER**.

CONCLUSÃO

Este trabalho desenvolveu o SMARTER, um modelo de análise preditiva baseado em métodos de aprendizado conjunto (*ensemble learning*), aplicados sobre dados históricos do mercado de ações brasileiro, com o objetivo de aumentar a acurácia da previsão da oscilação dos preços das ações.

Após a realização da análise e avaliação dos resultados obtidos, constatou-se que este trabalho de pesquisa conseguiu alcançar o objetivo proposto, ao realizar a combinação das técnicas de regressão mais adequadas para a formação de um novo modelo de aprendizado de máquina, sendo os resultados obtidos por este, apresentar melhores resultados médios estabelecidos através do coeficiente de determinação R^2 comparados aos demais modelos desenvolvidos.

Portanto, o modelo desenvolvido proporciona benefícios de aumentar a qualidade de generalização, obtendo uma melhor definição do comportamento das ações e consequentemente atingindo previsões mais confiáveis.

Futuramente, tal modelo pode ser incorporado a uma plataforma que utiliza dados extraídos em tempo real, tendo potencial de ser uma ferramenta de grande utilidade tanto para investidores e também pesquisadores da área.

REFERÊNCIAS

BROWNE, M. W. **Cross-Validation Methods**. Journal of Mathematical Psychology,v. 44, n. 1, p. 108–132, 2000. Disponível em:

http://www.sciencedirect.com/science/article/pii/S0022249699912798. Acesso em: 01 jun. 2020.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. **Data Preprocessing for Supervised Leaning**. World Academy of Science, Engineering and Technology,International Journal of Computer, Electrical, Automation, Control and InformationEngineering, v. 1, p. 4104–4109, 2007. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.8413&rep=rep1&type=pdf. Acesso em: 10 jun. 2020.

LIU, Y. **Python Machine Learning By Example**. Packt Publishing, 2017.

MARTINS, E. **Coeficiente de determinação**. Rev. Ciência Elem, v. 6, n. 1, p. 24. Disponível em: https://rce.casadasciencias.org/rceapp/art/2018/024/. Acesso em: 10 jul. 2020.

NETO, A. A. **Mercado Financeiro**. 14. ed. São Paulo: Atlas, 2018.

PAHWA, K.; AGARWAL, N. **Stock Market Analysis using Supervised Machine Learning**. IEEE, 2019. Disponível em: https://ieeexplore.ieee.org/document/8862225. Acesso em: 20 mar. 2020.

PASUPULETY, U. et al. **Predicting Stock Prices using Ensemble Learning and Sentiment Analysis**. IEEE, 2019. Disponível em: https://ieeexplore.ieee.org/document/8791689. Acesso em: 20 mar. 2020.

POLIKAR, R. **Ensemble learning**. Scholarpedia, v. 4, n. 1, p. 2776, 2009. Disponível em: http://www.scholarpedia.org/article/Ensemble_learning. Acesso em: 27 abr. 2020.

SKIENA, S. S. **The Data Science Design Manual**. 1st. Springer Publishing Company, Incorporated, 2017.

WENG, B. et al. **Predicting short-term stock prices using ensemble methods and online data sources**. Expert Systems with Applications, v. 112, p. 258–273, 2018. Disponível em: http://www.sciencedirect.com/science/article/pii/S0957417418303622. Acesso em: 17 mar. 2020

ZHANG, C.; MA, Y. **Ensemble Machine Learning: Methods and Applications**. Springer Publishing Company, Incorporated, 2012.