Predição do Preço do Bitcoin Utilizando Aprendizado de Máquinas e Informações do Mercado Financeiro Mundial

Alexandre R. Lemos[1]

[1]Departamento de Informática – Universidade Federal do Espírito Santo (UFES)

Caixa Postal 01-9011 – 29075-910 – Vitória – ES – Brazil

alexandre.r.lemos@edu.ufes.br

Abstract. *The prediction of asset values in financial markets is a task widely studied in academia and business. Statistical models are widely used, and the advancement of Machine Learning algorithms has allowed new discoveries and results to be obtained. The proposal of this project is the development of a statistical model for the prediction of the price of the crypto-asset Bitcoin using historical data of the cryptocurrency itself and of different assets and indices of the world financial market. It was possible to develop an XGBoost model that obtained an R^2 value of approximately 94.97% and RMSE of approximately 1610.52 using data from approximately 3 years.*

Resumo. *A predição de valores de ativos em mercados financeiros é uma tarefa bastante estudada no meio acadêmico e empresarial. Modelos estatísticos são muito utilizados, e o avanço dos algoritmos de Aprendizado de Máquinas vem permitindo que novas descobertas e resultados sejam obtidos. A proposta desse projeto é o desenvolvimento de um modelo de estatístico para a predição do preço do criptoativo Bitcoin utilizando dados históricos da própria criptomoeda e de diferentes ativos e índices do mercado financeiro mundial. Foi possível desenvolver um modelo XGBoost que obteve um valor de R^2 de aproximadamente 94,97% e RMSE de aproximadamente 1610,52 utilizando dados de aproximadamente 3 anos.*

1. Introdução

Bitcoin é um ativo virtual da categoria de criptomoedas com estrutura em *blockchain*, desenvolvido para ser irrastreável, descentralizado e independente de bancos ou governos. Isso permite o comércio via Bitcoin ser ágil, anônimo e de baixo custo, mas também mais arriscado por não possuir nenhuma legislação ou proteção por parte de nenhum governo [Segendorf, 2014].

O uso de modelos estatísticos de Aprendizado de Máquinas é bastante utilizado para a predição de valores de ativos financeiros. O crescimento do volume de dados e o avanço tecnológico nas técnicas de Aprendizado de Máquinas vem proporcionando novas oportunidades de se alcançar resultados cada vez melhores nas análises e projetos realizados [Ndikum, 2020].

2. Trabalhos Relacionados

Existem diversas pesquisas que tentam predizer o preço do Bitcoin a partir de métodos de Aprendizado de Máquinas. Em [Phaladisailoed, Numnonda, 2018] foram utilizados diversos modelos para predizer o valor do Bitcoin utilizando dados de negociações a cada minuto, conseguindo um valor de R^2 máximo de aproximadamente 99,2% e RMSE de 0.00002 utilizando redes GRU (*Gated Recurrent Unit*).

[Chen, Li, Sun, 2020] utilizoaram conjuntos com alta de dimensionalidade para predizer os preços do bitcoin diários e em alta frequência. Foram utilizados dados do mercado financeiro para a predição diária do preço e dados oriundos de negociações de criptomoedas para a predição no intervalo de 5 minutos. Os modelos utilizados no estudo alcançaram uma acurácia máxima de 66% para as predições diárias e uma acurácia máxima de 67,2% para as predições em intervalos de 5 minutos.

3. Objetivos

O objetivo deste trabalho é utilizar aprendizado de máquinas para desenvolver um modelo estatístico capaz de realizar a predição do preço da criptomoeda Bitcoin utilizando informações históricas de outros criptoativos, de índices de bolsas mundiais e de preços de commodities.

4. Descrição do conjunto de dados

Para o desenvolvimento do projeto foram utilizados dados históricos de catorze conjuntos de dados, com dados a partir de 01/02/2018 até 31/12/2021. Treze dos catorze conjuntos são dados de cotações de ativos ou índices do mercado financeiro, trazendo informações de preço de fechamento e abertura, preço máximo e mínimo e volume movimentado no dia.

Dos catorze conjuntos de dados, seis são relacionados ao mercado de criptomoedas, sendo cinco deles cotações das criptomoedas Bitcoin, Ethereum, Cardano, Binance Coin e Tether, e o sexto sendo o histórico *Fear and Greed Index*, índice que tenta calcular qual o sentimento, em geral, do mercado em relação ao Bitcoin [Gurdgiev, O'Loughlin, 2020]. Esse índice varia de 0 a 100, onde os valores mais baixos simbolizam que o mercado está receoso (com medo) e valores mais altos simbolizam que o mercado está otimista (ganancioso) em relação ao preço da criptomoeda.

Foram utilizados dados historicos de seis índices de bolsas mundiais, sendo eles o Ibovespa, S&P 500, Nasdaq, Nikkei, Shanghai e Euro STX. Também foram utilizados dados históricos do Ouro e Petróleo (Commodities). Ao todo, o conjunto de dados utilizado contém inicialmente 1430 amostras e 66 características incluindo a variável alvo (preço de fechamento do bitcoin).

5. Metodologia

O desenvolvimento do projeto se divide nas seguintes etapas: Aquisição dos dados, Engenharia de Variáveis e Desenvolvimento do Modelo.

5.1. Aquisição dos dados

Todos os conjuntos de dados utilizados para o desenvolvimento do projeto foram inteiramente obtidos do site Investing.com [Investing, 2020], que contém informações de diferentes ativos, nacionais e internacionais. Nele é possível escolher a faixa histórica desejada dos dados e baixá-los em formato csv. Para esse projeto, utilizou-se a faixa de dados iniciando no dia 01/02/2018 e finalizando no dia 31/12/2021, totalizando aproximadamente três anos de informação.

Utilizou-se como conjunto de treino os dados no período de 16/02/2018 até 31/07/2021 e como conjunto de teste os dados de 01/08/2021 até 31/12/2021.

5.2. Engenharia de Variáveis

Engenharia de Variáveis consiste na técnica de se utilizar informações presentes no conjunto de dados para gerar novas informações que possam ser valiosas na etapa de desenvolvimento do modelo [Zheng, Casari, 2018].

Nesse projeto, para cada ativo do conjunto inicial, calcula-se primeiramente a variação máxima diária (diferença entre a máxima e minima no dia) e a variação diária (diferença entre o preço de fechamento e preço de abertura).

Em seguida, para cada variável do conjunto de dados, incluindo as de variação recém criadas, calcula-se a média, desvio padrão e acumulado nos últimos três, sete e quinze dias. O propósito na criação dessas variáveis é conseguir capturar as mudanças no preço do ativo ao longo do tempo, para conseguir identificar momentos de subida/queda e preço médio ao longo do tempo.

O conjunto de dados após todas as etapas consiste de 1430 amostras e 829 características, incluindo a variável alvo.

5.3. Desenvolvimento do Modelo

A etapa de Desenvolvimento do Modelo se divide em quatro sessões: Criação do Modelo *baseline*, Seleção de Características, Otimização de Hiperparâmetros, Seleção do Modelo.

A primeira etapa consiste na utilização de todos os dados resultantes da etapa de Engenharia de Variáveis para treinamento e teste de cinco modelos diferentes, sendo eles: *Gradient Boost, Random Forest,* XGBoost*, Multi-Layer Perceptron* (MLP) e *K-Nearest Neighbors* (K-NN). Os modelos *baseline* servirão de comparativo nas as etapas seguintes, com o intuito de identificar se os procedimentos realizados estão melhorando o resultado obtido incialmente. Como métricas de performance dos modelos foram utilizados o coeficiente de determinação (R^2) e a raiz quadrada do erro médio (RMSE).

Os modelos *Gradient Boost, Random Forest*, MLP e K-NN utilizados são da biblioteca *scikit-learn* e foram utilizados com seus parâmetros *default*. O modelo XGBoost utilizado é proveniente da biblioteca xgboost e também foi utilizado com seus parâmetros *default*. Ao final dessa etapa, apenas os algoritmos *Gradient Boost*, *Random Forest* e *XGBoost* foram selecionados para serem testados nos próximos passos da metodologia, enquanto os algoritmos MLP e K-NN foram descartados por não terem alcançado uma performance inicial satisfatória.

Em seguida, tem-se a etapa de Seleção de Características, cujo objetivo é diminuir a dimensionalidade dos dados visando remover variáveis que possam estar adicionando complexidade desnecessária ao modelo.

Para reduzir a dimensionalidade do conjunto de dados, utilizou-se dois métodos baseados em filtro. O primeiro método consiste em calcular a correlação de cada variável do conjunto de dados com as demais. Em seguida, para cada par de variável cuja correlação seja maior ou igual que 90%, remove-se uma das variáveis. O intuito desse filtro é remover variáveis que são muito correlacionadas. Foram removidas ao todo 613 características.

O segundo método consiste na exclusão de variáveis baseado na Informação Mútua, onde foram selecionadas as cem variáveis que obtiveram os valores de Informação Mútua mais elevados. Ao todo, o processo de Seleção de Características reduziu consideravalmente a dimensionalidade do conjunto de dados de 828 características para 100.

A próxima etapa consiste na otimização dos hiperparâmetros com o intuito de melhorar a performance dos modelos selecionados. Nessa sessão utilizou-se para o *framework* Optuna para a otimização dos hiperparâmetros dos três modelos [Akiba, Sano, Yanase, Ohta, Koyama, 2019].

Para o *Gradient Boost* variou-se os valores da taxa de aprendizado, do número de árvores, da profundidade das árvores e do parâmetro alfa. Para o *Random Forest* variou-se os valores do mínimo de amostras para divisão de um nó, a profundidade das árvores e o número de árvores. Para o XGBoost variou-se os valores da taxa de aprendizagem, da profundidade das árvores e dos coeficientes gamma, lambda e alfa.

6. Resultados

Os resultados obtidos para cada modelo testado no desenvolvimento do projeto estão apresentados na Tabela 1.

Tabela 1. Resultados obtidos pelos modelos testados

Versão	Modelo	R^2 (%)	RMSE
Baseline	Random Forest	74,73	3608,68
	Gradient Boost	94,04	1752,46
	XGBoost	83,44	2921,65
	MLP	0	9251,02
	K-NN	0	13609,16
Otimizada	Random Forest	80,56	3165,33
	Gradient Boost	94,28	1716,85
	XGBoost	94,97	1610,52

O modelo que obteve a melhor performance os melhores valores de R^2 e RMSE foi o XGBoost, sendo portando o modelo escolhido. A Figura 1 mostra a comparação entre os valores calculados pelo modelo e os valores reais para o preço do Bitcoin.

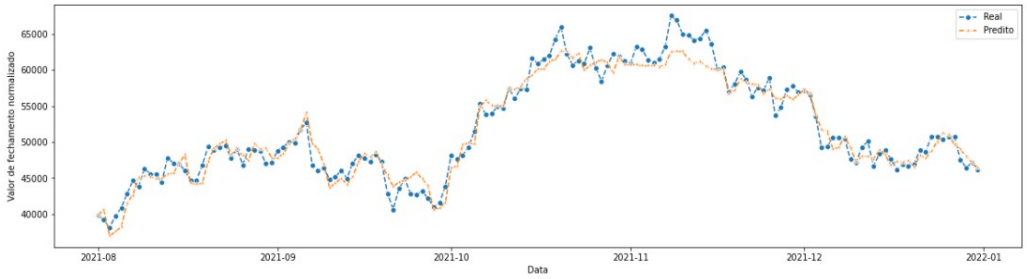

Figura 1. Comparação entre valor real (azul) e o valor calculado pelo modelo (laranja) para o preço do Bitcoin.

A Figura 2 apresenta informações a respeito dos erros do modelo. A Figura 2(a) mostra a distribuição normal da diferença dos valores calculados pelo modelo e os valores reais e a Figura 2(b) mostra a diferença dos valores calculados pelo modelo e os valores reais a cada amostra (observação) do conjunto de teste.

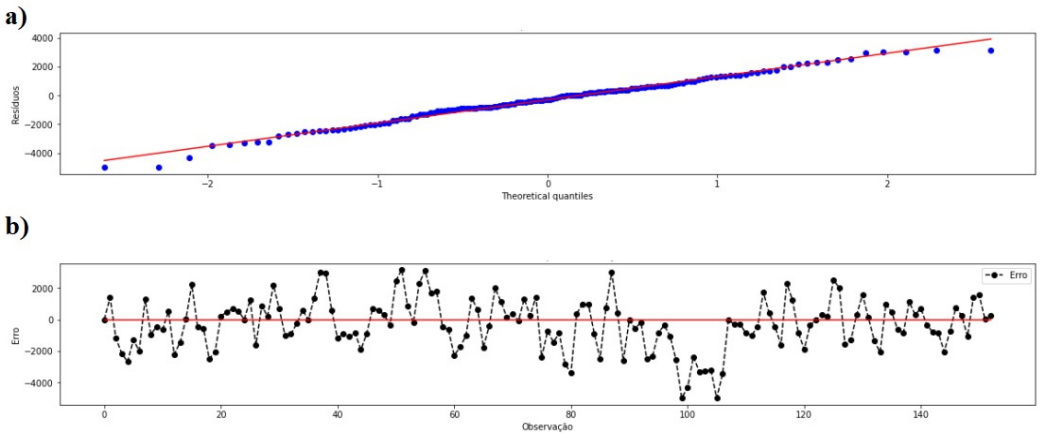

Figura 2. Análise dos erros do modelo. Distribuição normal da diferença (a), diferença entre valor calculado e real para cada amostra (b)

A Figura 3 mostra a análise dos valores SHAP do modelo, para determinar quais variáveis mais influenciam no cálculo do valor predito.

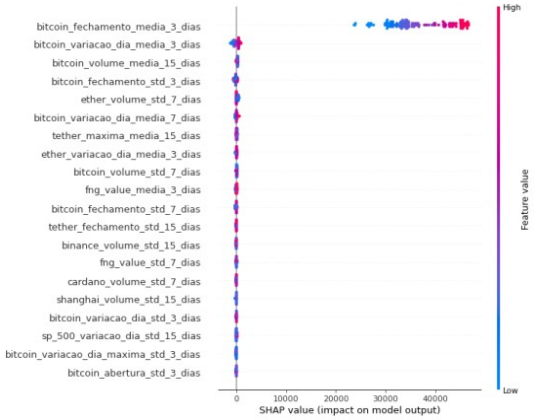

Figure 3. Análise SHAP do modelo

7. Conclusão

Nesse projeto, foi possível a obtenção de um modelo cuja performance alcançada foi considerada bastante satisfatória, onde o modelo XGBoost desenvolvido conseguiu obter um score R2 de aproximadamente 94,97% e um valor de RMSE de 1610,52 a partir dos dados históricos de feveveiro de 2018 até dezembro de 2021.

Foram realizadas diversas etapas de um projeto típico de Ciência de Dados, desde a aquisição e tratamento dos dados iniciais, criação de características a partir das características iniciais (Engenharia de Variáveis), criação de um modelo de referência (baseline), seleção de características através de diferentes algoritmos (no caso desse projeto foram utilizados algoritmos baseados em filtro), otimização dos hiperparâmetros dos modelos e, por fim, a seleção do melhor modelo utilizando métricas de avaliação de performance adequadas.

Para projetos futuros, é possível que variáveis obtidas a partir de análise de redes sociais como sejam características importantes, visto que esse esse tipo de ativo financeiro (criptomoedas), por ainda ser algo bastante recente e novo no mercado, ainda possui um grande fator especulativo cujo valor sofre grande influência de ações/declarações feitos por grandes *players* (pessoas e empresas).

Como principal ponto de melhoria do modelo, destacaria a melhora da capacidade do modelo acompanhar variações repentinas nos valores. É possível observar que as observações que obtiveram o maior erro foram aquelas onde o preço variou a direção de subida/descida bruscamente.

Referências

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395.

Gurdgiev, C., & O'Loughlin, D. (2020). Herding and anchoring in cryptocurrency markets: Investor reaction to fear and uncertainty. *Journal of Behavioral and Experimental Finance*, *25*, 100271.

Investing, I. (2020). Investing. com. Acesso em: 04 de Fevereiro de 2022.

Ndikum, P. (2020). Machine learning algorithms for financial asset price forecasting. *arXiv preprint arXiv:2004.01504*.

Phaladisailoed, T., & Numnonda, T. (2018, July). Machine learning models comparison for bitcoin price prediction. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 506-511). IEEE.

Segendorf, B. (2014). What is bitcoin. *Sveri gesRiksbankEconomicReview*, *2014*, 2-71.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. "O'Reilly Media, Inc.".