Projeto de Previsão de Preços do Bitcoin Utilizando Machine Learning

**Leonardo Fernandes Contrera**

**Matheus Rossi**

**Vinicius Kenzo Hiraoka**

***Resumo.*** *Este estudo tem como objetivo apresentar um modelo de aprendizado de máquina para previsão dos preços mínimo e máximo do Bitcoin no dia seguinte, utilizando dados reais do mercado entre 25/08/2024 e 11/06/2025. Foram extraídas e engenheiradas variáveis relevantes, como volume financeiro, médias móveis, indicadores relativos e booleanos de rompimento de topo histórico. Três algoritmos de regressão foram comparados: MLP, XGBoost e Random Forest. A métrica R² e o erro percentual foram utilizadas para avaliação. Os resultados indicam que a combinação de variáveis históricas com volume e variação recente é eficaz na modelagem do comportamento do Bitcoin.*

***Abstract.*** *This study aims to present a machine learning model for predicting Bitcoin's minimum and maximum prices on the next day, using real market data between 08/25/2024 and 06/11/2025. Relevant variables, such as financial volume, moving averages, relative indicators, and historical top breakout Boolean indicators, were extracted and engineered. Three regression algorithms were compared: MLP, XGBoost, and Random Forest. The R² metric and percentage error were used for evaluation. The results indicate that the combination of historical variables with volume and recent variation is effective in modeling Bitcoin's behavior.*

**2. Introdução**

O Bitcoin, principal criptomoeda do mercado, é caracterizado por elevada volatilidade, o que torna a previsão de seus preços uma tarefa desafiadora. Este trabalho tem como objetivo desenvolver e comparar modelos de aprendizado de máquina para estimar os valores mínimo e máximo do Bitcoin no dia seguinte, com base em dados históricos e indicadores técnicos.

O conjunto de dados analisado abrange o período de 25/08/2024 a 11/06/2025, totalizando mais de 290 registros diários. Além das variáveis originais, foram criadas novas features derivadas, como médias móveis, volume financeiro relativo e indicadores de rompimento de topo histórico, com o intuito de aprimorar o desempenho preditivo dos modelos.

**3. Metodologia**

O pré-processamento dos dados incluiu a conversão de colunas financeiras para o formato numérico adequado, abrangendo valores monetários, percentuais e volumes. Em seguida, foram criadas variáveis derivadas com o objetivo de enriquecer o conjunto de dados e potencializar a capacidade preditiva dos modelos. Entre essas variáveis destacam-se:

* Volume em USD, obtido pela multiplicação entre o volume negociado e o preço de fechamento do dia;
* Médias móveis, como a média da máxima dos últimos três dias;
* Variação percentual média, considerando os três dias anteriores;
* Volume relativo diário e semanal, que representam a relação do volume atual com os valores anteriores e com a média semanal, respectivamente;
* Indicador de rompimento de topo histórico, que sinaliza se o preço do ativo superou o valor máximo dos últimos 90 dias.
* Após a criação dessas variáveis, foi realizada a padronização dos nomes das colunas e o tratamento dos valores ausentes, garantindo a consistência do dataset.

Para entender melhor a distribuição e as correlações dos dados, foram geradas diversas visualizações exploratórias:

* Histogramas das variáveis numéricas;
* Boxplot do volume financeiro (USD), útil para identificar outliers e assimetrias;
* Gráfico de contagem (countplot) para a variável booleana rompeu\_topo\_historico;
* Matriz de correlação entre as variáveis numéricas, acompanhada de uma legenda explicativa das intensidades de correlação;
* Gráficos de dispersão (scatterplots) entre as principais variáveis explicativas e os valores-alvo (minima\_amanha e maxima\_amanha).

Essas etapas permitiram uma compreensão mais profunda dos dados e fundamentaram as decisões de modelagem e seleção de features, essas imagens estão disponíveis em:

<https://docs.google.com/document/d/1yn2pbSyTiee87ggKX7qUEg3mV-BBRwLW/edit?usp=sharing&ouid=102805005150896677393&rtpof=true&sd=true>

**3.2 Modelagem Preditiva**

Para a etapa de modelagem, os dados foram divididos em 70% para treinamento e 30% para teste, respeitando a ordem temporal das observações para evitar vazamento de informações futuras. Foram avaliados três algoritmos de regressão: MLP Regressor (rede neural com três camadas ocultas), XGBoost Regressor (baseado em boosting de árvores) e Random Forest Regressor (ensemble de árvores de decisão).

A busca pelos melhores hiperparâmetros foi realizada por meio do RandomizedSearchCV, utilizando validação cruzada com 5 folds no conjunto de treinamento. A comparação de desempenho dos modelos considerou as seguintes métricas: erro absoluto médio (MAE), coeficiente de determinação (R²), erro relativo percentual, além dos percentuais de acerto dentro das margens de erro de até 2% e 5% entre os valores reais e previstos.

Essas etapas permitiram uma compreensão mais profunda dos dados e fundamentaram as decisões de modelagem e seleção de variáveis. Para facilitar a reprodutibilidade e o acompanhamento técnico do projeto, os códigos-fonte completos estão disponíveis em notebooks públicos no Google Colab, conforme os modelos utilizados:

MLP Regressor: disponível em

<https://colab.research.google.com/drive/1xEzH7eDDICupHp0-Qbt-rAwSvBBb8KXF?usp=sharing>

XGBoost Regressor: disponível em:

<https://colab.research.google.com/drive/1I9tpl05ICXxCHav7-qc1NF1mIo0mnnWZ?usp=sharing>

Random Forest Regressor: disponível em:

<https://colab.research.google.com/drive/1Be6HB2935wKLz486XeJzRDeVGkkPlEX5?usp=sharing>

**4. Resultados**

Tabela 1 – Métricas de desempenho dos modelos preditivos para o preço mínimo do Bitcoin

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | MAE Mín | Erro Rel. Mín (%) | R² Mín (%) | ≤2% Mín | ≤5% Mín |
| MLP Regressor | 10.094,70 | 10,83% | -30,98 | 8,14% | 26,74% |
| XGBoost Regressor | 2.016,01 | 2,16% | 92,86 | 61,63% | 88,37% |
| Random Forest | 1.785,31 | 1,91% | 94,57 | 63,95% | 93,02% |

Fonte: Elaborado pelo autor com base nos resultados dos modelos treinados em Python com Random Forest, XGBoost e MLP (dados de teste Período: 25/08/2024 a 11/06/2025)

Tabela 2 – Métricas de desempenho dos modelos preditivos para o preço máximo do Bitcoin

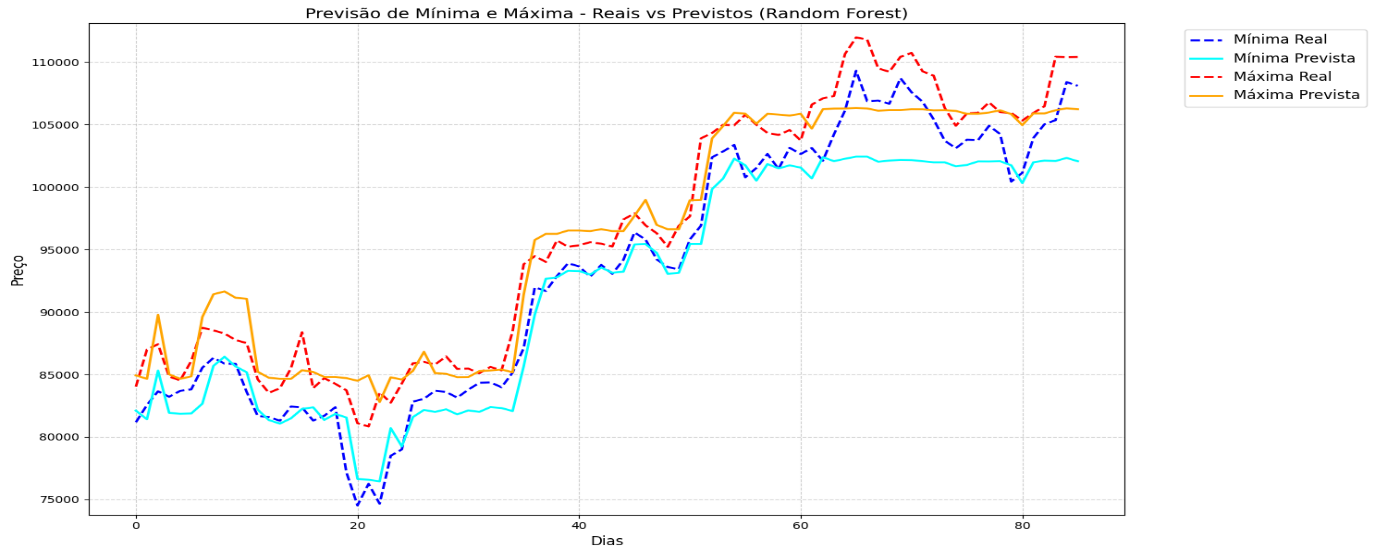
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | MAE Máx | Erro Rel. Máx (%) | R² Máx (%) | ≤2% Máx | ≤5% Máx |
| MLP Regressor | 10.457,69 | 10,87% | -48,45 | 5,81% | 27,91% |
| XGBoost Regressor | 1.764,19 | 1,83% | 93,99 | 70,93% | 95,35% |
| Random Forest | 1.608,08 | 1,67% | 95,31 | 67,44% | 97,67% |

Fonte: Elaborado pelo autor com base nos resultados dos modelos treinados em Python com Random Forest, XGBoost e MLP. (dados de teste Período: 25/08/2024 a 11/06/2025)

O modelo Random Forest apresentou o melhor desempenho entre os algoritmos avaliados, com destaque especial para a previsão do valor máximo diário do Bitcoin. O modelo obteve um coeficiente de determinação (R²) de 95,31%, indicando que foi capaz de explicar com elevada precisão a variabilidade dos dados observados. Além disso, atingiu uma acurácia de 97,67% dentro da margem de erro de até 5%, demonstrando alto grau de confiabilidade mesmo diante da notória volatilidade do ativo.

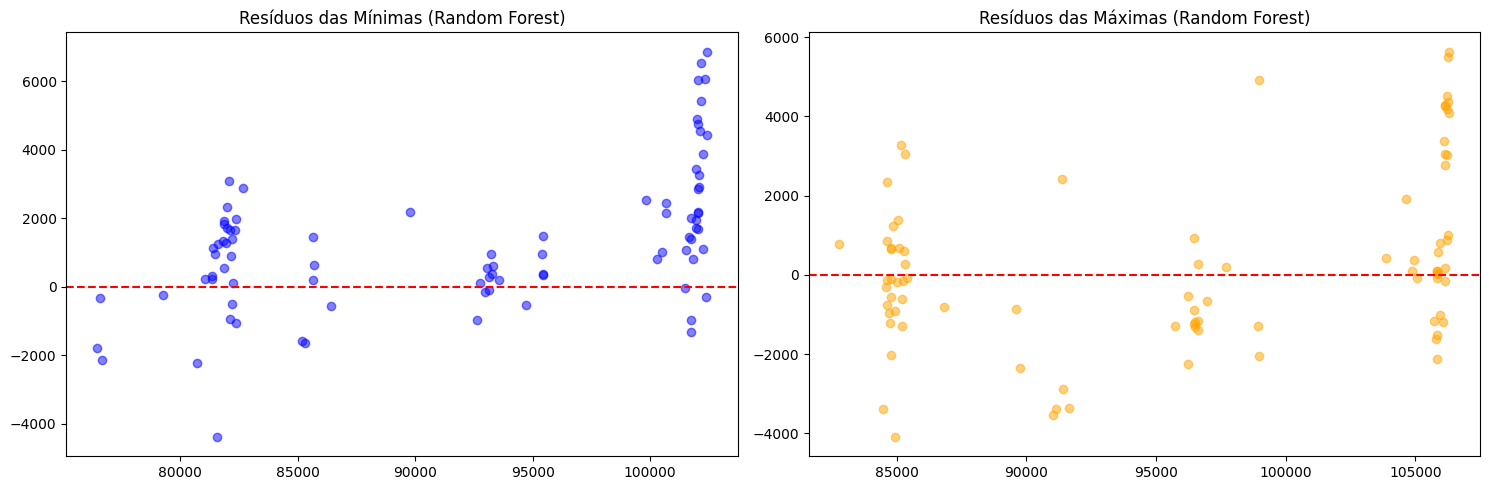
Esse desempenho evidencia a robustez do Random Forest, especialmente por sua capacidade de lidar com relações não lineares e variáveis com diferentes escalas, características comuns em séries financeiras. Para ilustrar visualmente a performance do modelo, foi elaborado um gráfico com os resultados do conjunto de teste. No gráfico, as linhas tracejadas representam os valores reais de mínima e máxima diários, enquanto as linhas contínuas correspondem às previsões feitas pelo modelo. Essa visualização permite observar a aderência entre os valores previstos e os observados, reforçando a eficácia do modelo na tarefa de regressão.

Figura 1 – Comparação entre valores reais e previstos para mínima e máxima com o modelo Random Forest



**Fonte:** Elaboração própria.

Figura 2 – Resíduos das previsões de mínima e máxima pelo modelo Random Forest.



**Fonte:** Elaboração própria

Além dos resultados do modelo Random Forest apresentados acima, os gráficos de comparação entre valores reais e previstos e de resíduos, para os modelos MLP Regressor e XGBoost Regressor também foram gerados e estão disponíveis para consulta no seguinte link:

<https://docs.google.com/document/d/1yn2pbSyTiee87ggKX7qUEg3mV-BBRwLW/edit?usp=sharing&ouid=102805005150896677393&rtpof=true&sd=true>

**5. Discussão**

O modelo Random Forest se destacou por sua robustez diante de dados com distribuição não linear e presença de outliers, apresentando o melhor desempenho geral, especialmente para previsão do valor máximo do Bitcoin. Já o modelo XGBoost teve desempenho sólido, mas ligeiramente inferior ao Random Forest, possivelmente por ser mais sensível a ruídos quando o volume de dados é limitado.

O modelo MLP (Multi-Layer Perceptron), por sua vez, inicialmente apresentou underfitting severo, com baixo R² e alta margem de erro, o que foi atribuído ao tamanho limitado do conjunto de dados original. No entanto, ao ser testado em um dataset ampliado com mais de 500 dias, o desempenho do MLP melhorou significativamente, atingindo:

* MAE Mínima: 1852.29 | R² Mínima: 98.51%
* MAE Máxima: 2677.11 | R² Máxima: 97.23%
* % Acerto ≤5% (mínima): 89.08% | % Acerto ≤2%: 46.89%
* % Acerto ≤5% (máxima): 76.97% | % Acerto ≤2%: 33.61%

Esses resultados indicam que modelos baseados em redes neurais requerem volumes de dados maiores para expressar seu potencial, e que o tamanho do dataset foi um fator crítico no desempenho inicial inferior do MLP.

Outro ponto importante foi o impacto da variável rompeu\_topo\_historico, introduzida para identificar quebras de padrão associadas ao rompimento do topo dos últimos 90 dias. Essa feature mostrou-se relevante, especialmente em períodos de forte valorização do Bitcoin, onde os modelos tradicionais perderam parte da capacidade preditiva.

Durante e após o rompimento do topo histórico (cerca de um mês antes do final do dataset), observou-se uma queda drástica no desempenho dos modelos Random Forest e XGBoost, com previsões que subestimavam significativamente as máximas, chegando a prever valores máximos abaixo das mínimas reais. Esse comportamento sugere que, nesses contextos de "preço descoberto", os modelos perdem referência histórica e tornam-se menos eficazes, evidenciando uma alta sensibilidade ao comportamento recente do mercado.

Esse cenário reforça a importância das features temporais e de volume, e aponta para a necessidade de mecanismos adaptativos ou de reentrenamento em cenários de mercado extremamente voláteis. Também indica que o uso de variáveis de tendência e volatilidade de prazo mais longo pode ser benéfico para lidar com mudanças abruptas de regime de preços.

**6. Conclusão**

Este projeto demonstrou que é possível prever, com boa precisão, os valores mínimo e máximo do Bitcoin utilizando algoritmos de regressão aplicados a dados históricos e variáveis técnicas derivadas. A engenharia de features desempenhou papel fundamental para capturar padrões relevantes do mercado.

Dentre os modelos testados, o Random Forest se destacou como o mais eficaz no conjunto de dados analisado, superando o XGBoost e o MLP em todas as métricas avaliadas para a maioria dos casos. No entanto, observou-se que o MLP apresentou desempenho significativamente melhor em datasets maiores, indicando seu potencial para cenários com mais amostras e padrões complexos.

Além disso, a análise revelou limitações importantes dos modelos tradicionais durante momentos de alta volatilidade e rompimento de topos históricos, quando os padrões do passado se tornam menos informativos. Isso reforça a necessidade de estratégias adaptativas ou modelos que incorporem dinâmicas temporais mais profundas.

Como aprimoramentos futuros, recomenda-se:

* Ampliação do dataset, incluindo janelas maiores e novos ciclos de mercado.
* Testes com modelos temporais avançados, como LSTM e transformers com mecanismos de atenção, para capturar melhor a sequência e a volatilidade.
* Integração de variáveis externas, como:
* Sentimento de mercado (notícias, redes sociais);
* Indicadores macroeconômicos;
* Dominância do BTC e comportamento de outras criptomoedas;
* Indicadores técnicos clássicos (RSI, MACD, etc.).

Esses avanços podem aumentar a robustez dos modelos e sua aplicabilidade em ambientes reais, especialmente em períodos de alta imprevisibilidade.

**7. Referências**

**GOOGLE.** **Google Colaboratory (Colab).** 2017. Disponível em: [https://colab.research.google.com](https://colab.research.google.com/). Acesso em: 10 jun. 2025.

**INVESTING.COM.** **Bitcoin – Dados Históricos (BTC).** Disponível em: <https://br.investing.com/crypto/bitcoin/historical-data>. Acesso em: 10 jun. 2025.

**PEDREGOSA, F.** et al. **Scikit-learn: Machine Learning in Python.** *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Disponível em: <https://scikit-learn.org/>. Acesso em: 10 jun. 2025.

**CHEN, T.; GUESTRIN, C.** **XGBoost: A Scalable Tree Boosting System.** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, San Francisco. p. 785–794. Disponível em: <https://xgboost.ai/>. Acesso em: 10 jun. 2025.

**THE PANDAS DEVELOPMENT TEAM.** **pandas: Powerful Python data analysis toolkit**, v.2.0.3, 2024. Disponível em: <https://pandas.pydata.org/>. Acesso em: 10 jun. 2025.

**WASKOM, M.** et al. **Matplotlib: Visualization with Python**, v.3.8.0, 2023. Disponível em: <https://matplotlib.org/>. Acesso em: 10 jun. 2025.

**BISHOP, C. M.** **Pattern Recognition and Machine Learning.** 2. ed. New York: Springer, 2016. ISBN 978-0387-31073-2.

**GÉRON, A.** **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.** 2. ed. Sebastopol: O’Reilly Media, 2019. ISBN 978-1492032649.