BITCOIN ESTIMATOR

João Victor Magalhães Souza - 3483
Lucas Ranieri Oliveira Martins - 3479

SUMÁRIO

- Nosso problema;
- *Data acquisition;*
- *Data visualization;*
- *Data preparation;*
- *Feature engineering;*
- Hipóteses de modelos;
- Construção e otimização do *XGBoost;*
- Métricas, curvas e remodelagem dos dados;
- *Features* mais importantes;
- Gráfico de resíduos;
- Construção da tela;
- Arquitetura da aplicação.

NOSSO PROBLEMA

- Comportamento do BTC ao longo dos anos;
- Podemos prever o preço de fechamento do *Bitcoin* com um dia de antecedência ?

Fonte: Google - *Bitcoin.*

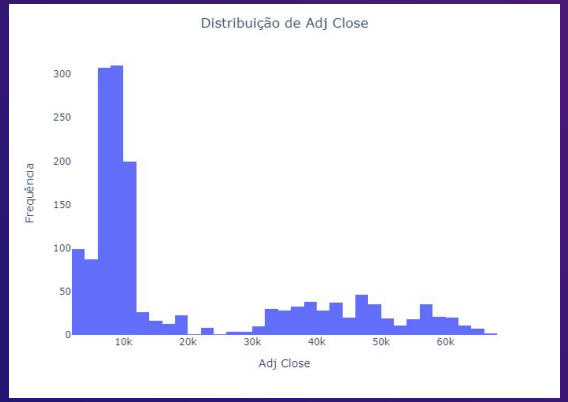
DATA ACQUISITION

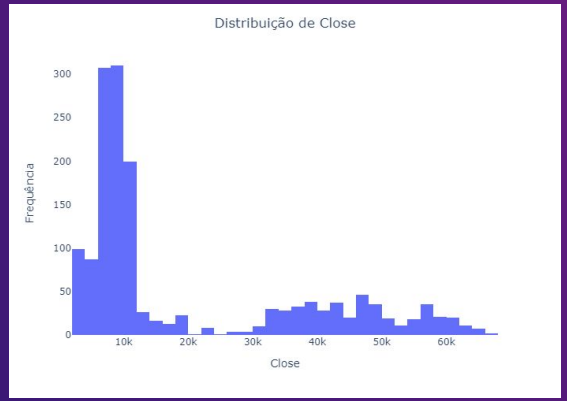
- API do *Yahoo Finance;*
- Dados desde 2018;
  - A etapa de *Data Visualization* foi preponderante para escolha do período !
- Atualização **diária** dos dados.

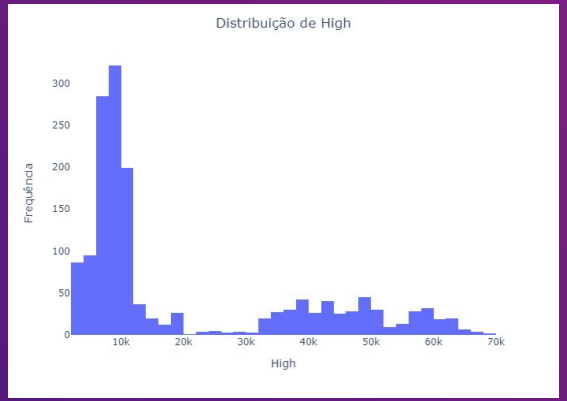
Date	High	Low	Open	Close	Volume	Adj Close
2018-01-01	14112.200195	13154.700195	14112.200195	13657.200195	10291200000	13657.200195
2018-01-02	15444.599609	13163.599609	13625.000000	14982.099609	16846600192	14982.099609
2018-01-03	15572.799805	14844.500000	14978.200195	15201.000000	16871900160	15201.000000
2018-01-04	15739.700195	14522.200195	15270.700195	15599.200195	21783199744	15599.200195
2018-01-05	17705.199219	15202.799805	15477.200195	17429.500000	23840899072	17429.500000
...	...	...	...	...	...	...
2022-03-23	42893.507812	41877.507812	42364.378906	42892.957031	25242943069	42892.957031
2022-03-24	44131.855469	42726.164062	42886.652344	43960.933594	31042992291	43960.933594
2022-03-25	44999.492188	43706.285156	43964.546875	44348.730469	30574413034	44348.730469
2022-03-26	44735.996094	44166.273438	44349.859375	44500.828125	16950455995	44500.828125
2022-03-27	44859.601562	44449.078125	44505.839844	44820.730469	19630307328	44820.730469

Fonte: Acervo próprio.

DATA VISUALIZATION

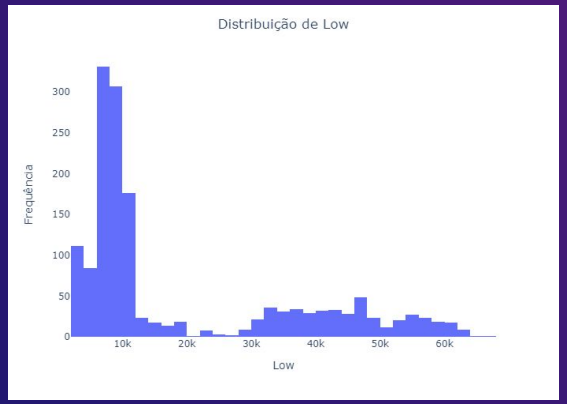


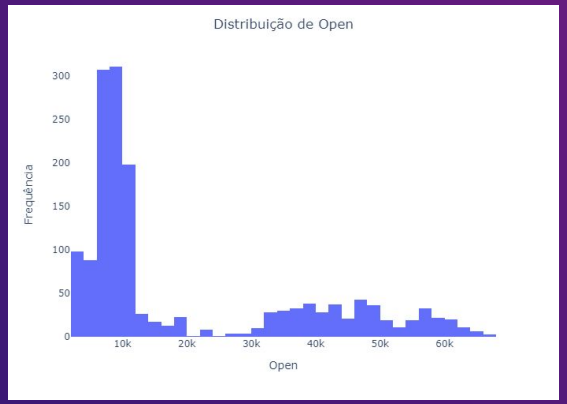


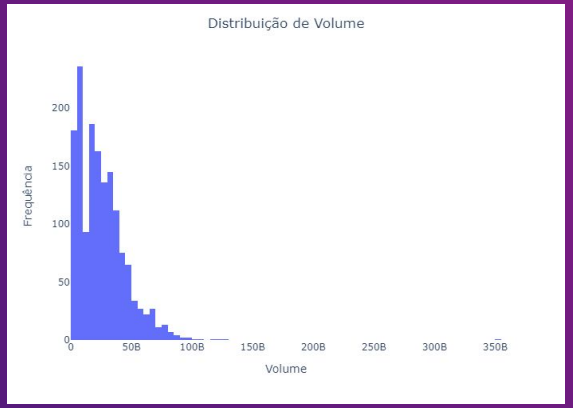


Fonte: Acervo próprio.

DATA VISUALIZATION







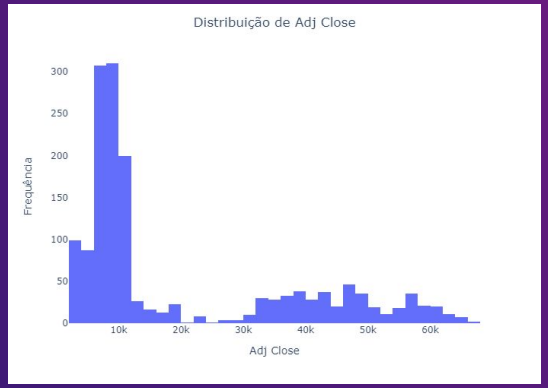
Fonte: Acervo próprio.

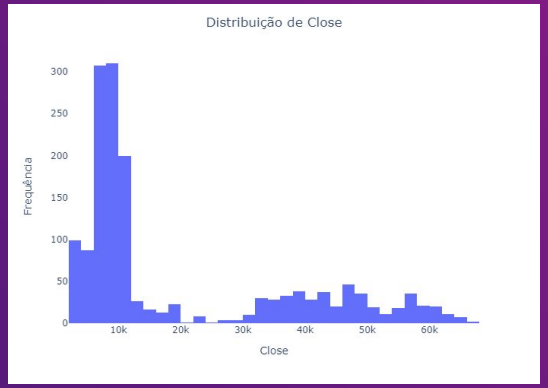
DATA PREPARATION

- Colunas redundantes;

Date	Close	Adj Close
2018-01-01	13657.200195	13657.200195
2018-01-02	14982.099609	14982.099609
2018-01-03	15201.000000	15201.000000
2018-01-04	15599.200195	15599.200195
2018-01-05	17429.500000	17429.500000
...	...	...
2022-03-23	42892.957031	42892.957031
2022-03-24	43960.933594	43960.933594
2022-03-25	44348.730469	44348.730469
2022-03-26	44500.828125	44500.828125
2022-03-27	44785.332031	44785.332031

Fonte: Acervo próprio.



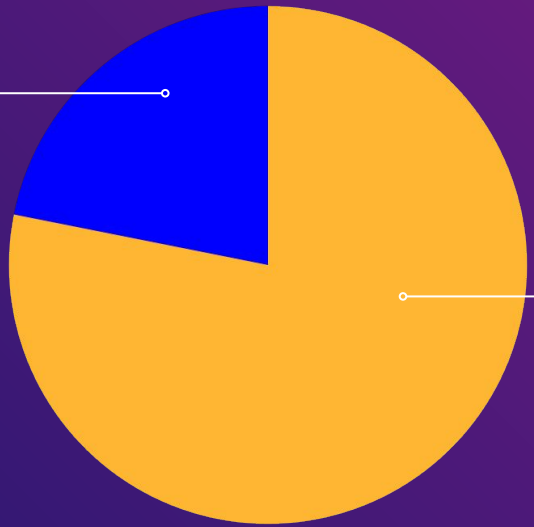


Fonte: Acervo próprio.

FEATURE ENGINEERING

20.0%

Olhar para os dados
(*data mining*)



80.0%

Olhar para o processo
(conhecimento de
causa)

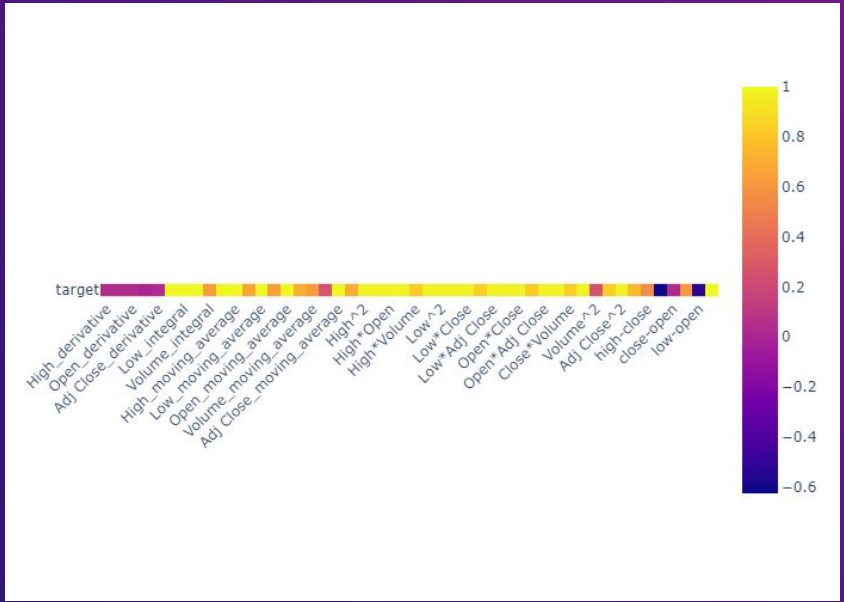
FEATURE ENGINEERING

- Baseadas em conhecimentos de causa:
  - Derivada;
  - Momentos estatísticos (média e desvio-padrão);
  - Variações:
    - Preço de pico - preço de vale;
    - Preço de pico - preço de fechamento;
    - Preço de vale - preço de fechamento;
    - Preço de fechamento - preço de abertura;
    - Preço de pico - preço de abertura;
    - Preço de vale - preço de abertura;

FEATURE ENGINEERING

- Baseadas em *data mining:*
  - Integral;
  - Combinações polinomiais;

FEATURE ENGINEERING



Fonte: Acervo próprio.

HIPÓTESES DE MODELOS

dmlc XGBoost

Fonte: *XGBoost - Documentation.*

LSTM

LONG SHORT-TERM MEMORY

CatBoost

Fonte: *Catboost - Documentation.*

CONSTRUÇÃO E OTIMIZAÇÃO DO MODELO

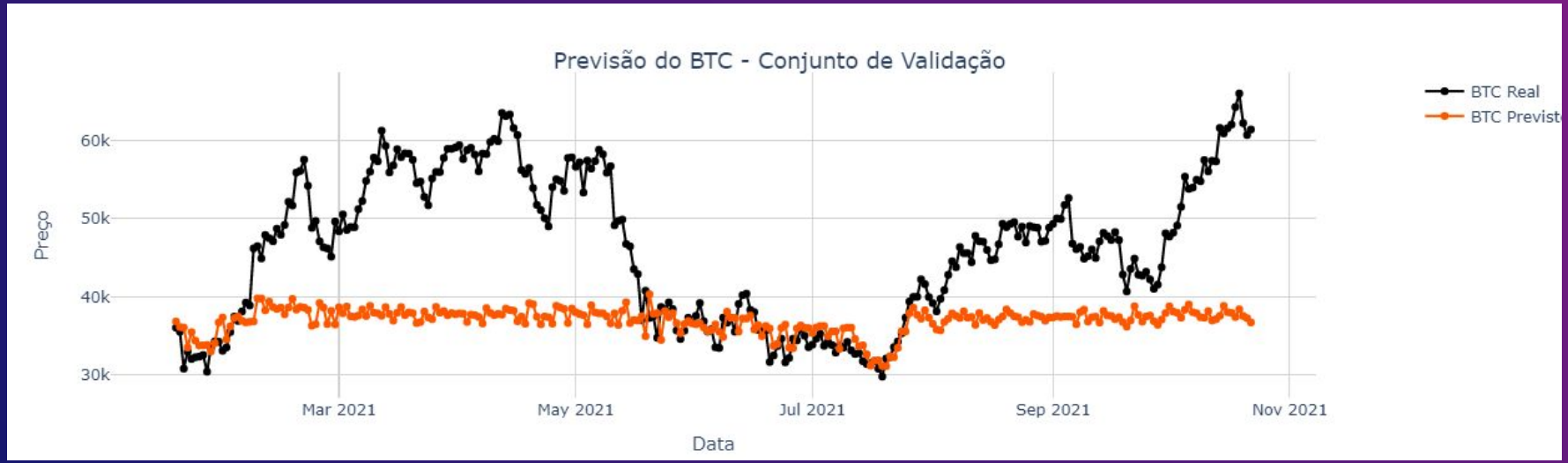
- *XGBoost Regressor;*
- Hiperparâmetros:
  - 1500 árvores;
  - *learning rate* de 0.05;
  - profundidade máxima igual a 12;
  - *MAE* como métrica de avaliação;
  - regularizações L1 e L2 para combate ao *overfiting*;
  - amostra de 80% para construção da árvore inicial.

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- *Split* dos dados:
  - 80% para treino;
  - 10% para testes;
  - 10% para validação;

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- Métricas no conjunto de validação:
  - *MAE:* ***$ 9838.87***
  - Erro Percentual: **+- 72.75%**



Fonte: Acervo próprio.

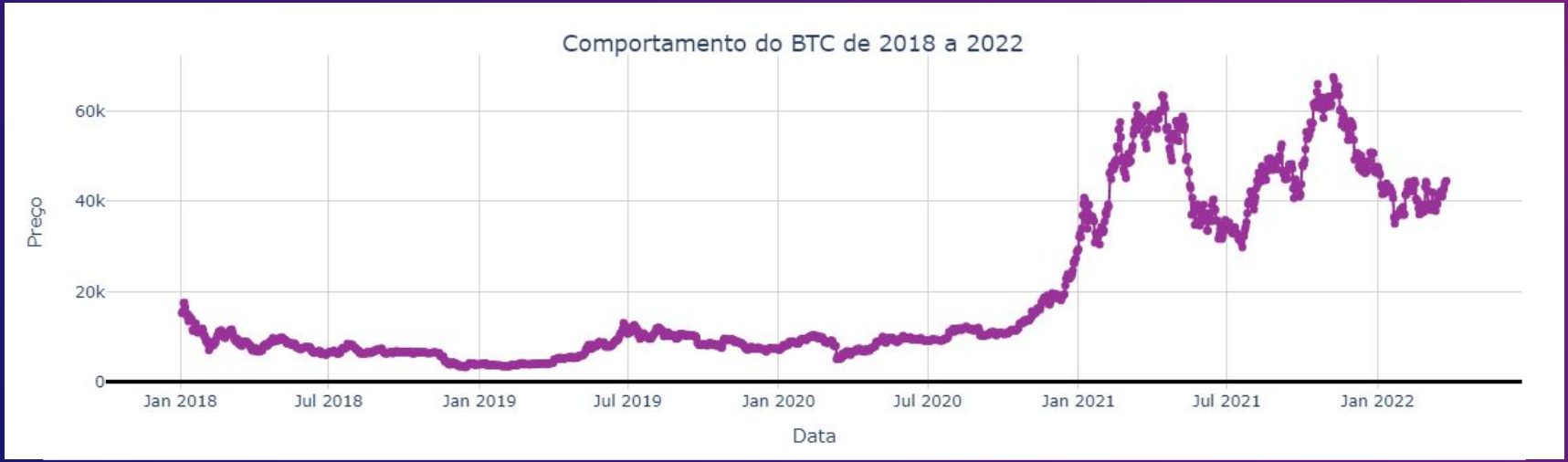
MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- Métricas no conjunto de testes:
  - *MAE:* ***$ 10328.31***
  - Erro Percentual: **+- 42.84 %**

Fonte: Acervo próprio.

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- Qual o problema ?
  - Vamos observar o comportamento dos dados …



Fonte: Acervo próprio.

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

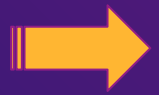
- Como você pode ter uma predição aceitável de um padrão com que não foi treinado, isto é, que não viu nada minimamente parecido ?

Fonte: Acervo próprio.

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- Aliviador: aplicação da técnica de *shuffle* nos dados;
  - *Gap* comportamental exorbitante nos conjuntos de treino, validação e testes;
  - Cobrir mais hipóteses de treinamento;
  - Misturar os períodos de treino e validação.

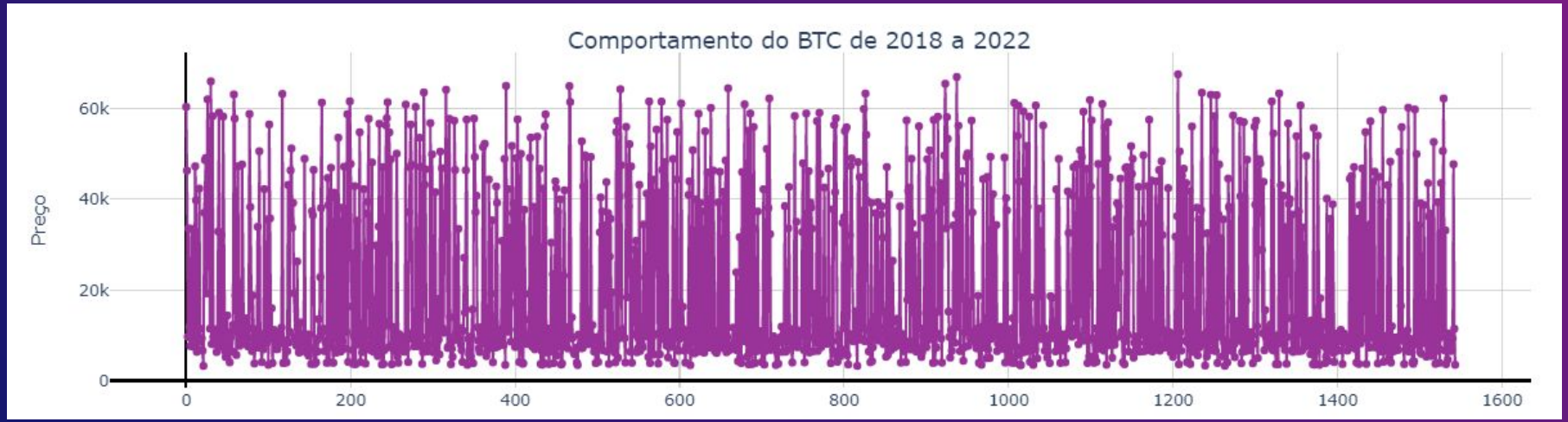
Date	High	Low	Open	Close	Volume
2018-01-03	-0.246028	-0.240612	-0.257297	15201.000000	-0.431134
2018-01-04	-0.237013	-0.259057	-0.241074	15599.200195	-0.193841
2018-01-05	-0.130847	-0.220106	-0.229620	17429.500000	-0.094422
2018-01-06	-0.130458	-0.130725	-0.119527	17527.000000	-0.361429
2018-01-07	-0.137631	-0.169464	-0.115910	16477.599609	-0.479735
2018-01-08	-0.193898	-0.277027	-0.174210	15170.099609	-0.356631
2018-01-09	-0.250095	-0.264677	-0.249227	14595.400391	-0.441372



Date	High	Low	Open	Close	Volume
2018-01-06	-0.130458	-0.130725	-0.119527	17527.000000	-0.361429
2018-01-12	-0.318564	-0.337124	-0.341843	13980.599609	-0.663349
2018-01-04	-0.237013	-0.259057	-0.241074	15599.200195	-0.193841
2018-01-08	-0.193898	-0.277027	-0.174210	15170.099609	-0.356631
2018-01-10	-0.278410	-0.306615	-0.278912	14973.299805	-0.352433
2018-01-05	-0.130847	-0.220106	-0.229620	17429.500000	-0.094422
2018-01-11	-0.275952	-0.340111	-0.257852	13405.799805	-0.447455

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

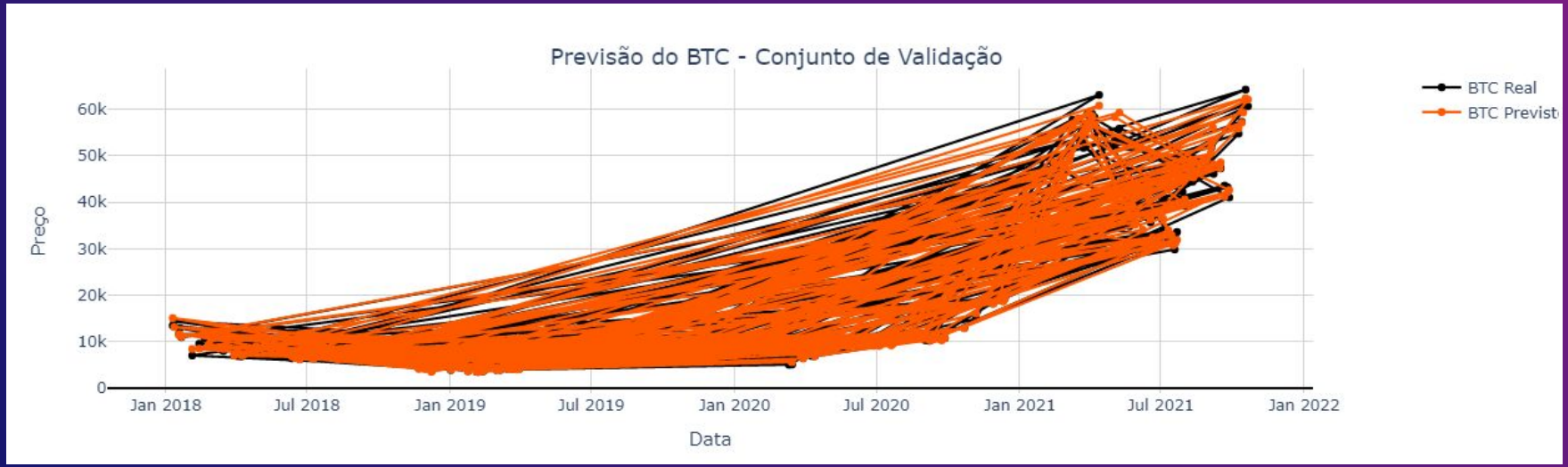
- Aliviador: aplicação da técnica de *shuffle* nos dados;
  - *Gap* comportamental exorbitante nos conjuntos de treino, validação e testes;
  - Cobrir mais hipóteses de treinamento;
  - Misturar os períodos de treino e validação.



Fonte: Acervo próprio.

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- Métricas no conjunto de validação:
  - *MAE:* ***$ ~~9838.87~~ 616.80***
  - Erro Percentual: **+- ~~72.75%~~ 9.16%**



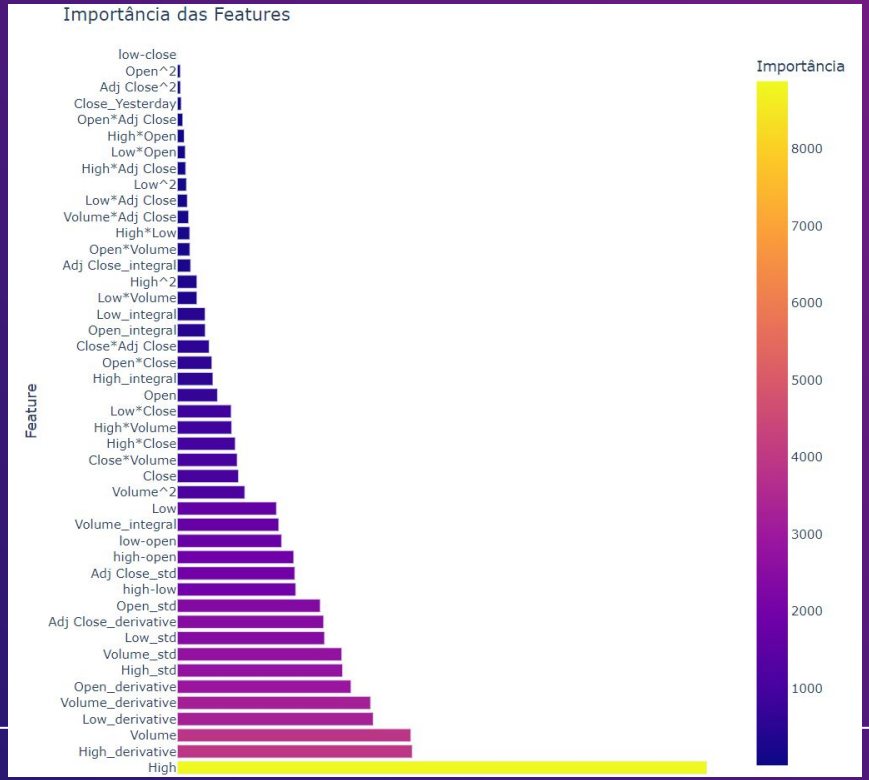
Fonte: Acervo próprio.

MÉTRICAS, CURVAS E REMODELAGEM DOS DADOS

- Métricas no conjunto de validação:
  - *MAE:* ***$ ~~10328.31~~ 1622.82***
  - Erro Percentual: **+- ~~42.84%~~ 5.44%**

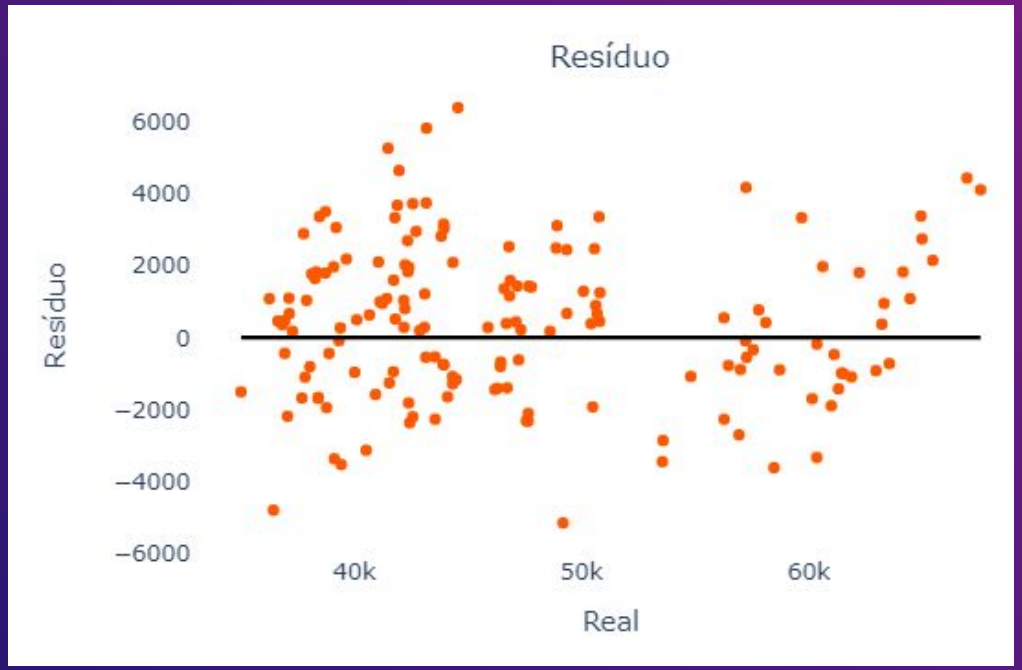
Fonte: Acervo próprio.

FEATURES MAIS IMPORTANTES



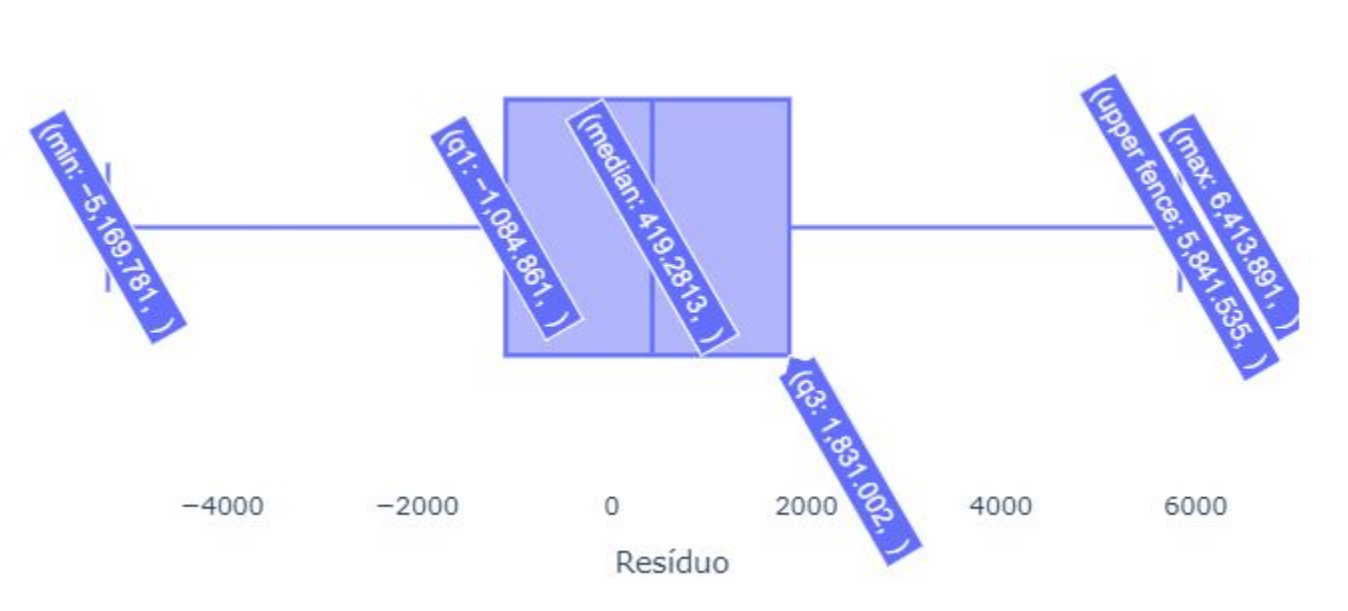
Fonte: Acervo próprio.

GRÁFICO DE RESÍDUOS



Fonte: Acervo próprio.

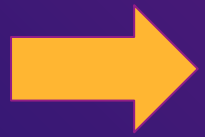
GRÁFICO DE RESÍDUOS

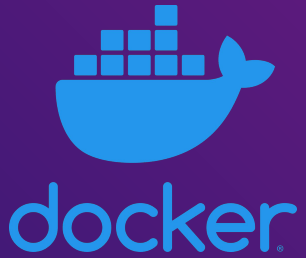


Fonte: Acervo próprio.

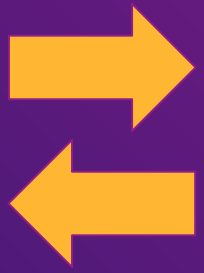
ARQUITETURA DA APLICAÇÃO

Flask Application





Google Cloud Run



Google Cloud SQL Server

VISUALIZAÇÃO DA APLICAÇÃO

https://projetodados-ju3qcqp27a-rj.a.run.app