HYBRID APPROACH TO TRADING VOLATILE DIGITAL ASSETS

AGENDA

1. Business Problem / Data Problem
2. Stakeholder Identification
3. Process Workflow
4. Dataset Characteristics
5. Exploratory Data Analysis
6. Classification Machine Learning Model & Evaluation
7. Time Series Forecasting Model & Evaluation
8. Proposed Approach to Production
9. Future Developments / Limitations

BUSINESS PROBLEM / DATA PROBLEM

Institute of Data

BUSINESS PROBLEM

- Digital assets are a highly volatile speculative asset class.
- The high volatility of cryptocurrencies not only increases the risks of crypto trading but also has the potential to make it more profitable than any other form of investment.
- Buy & hold strategy has shown to outperform daily trading; however, we propose that a portion of any investor's portfolio could be traded to reduce risk and increase alpha.
- For this case study we will analyze bitcoin prices.

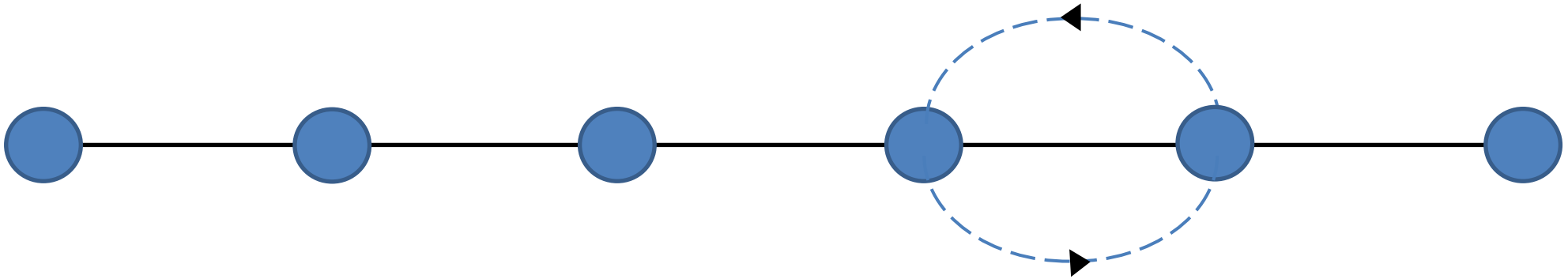
DATA PROBLEM

- Data containing open, high, low, close, volume data on digital assets
- On-chain data
- Economic indicators
- Stock market information

STAKEHOLDER IDENTIFICATION

Institute of Data

PROCESS WORKFLOW



Collection

- Data sourced from the following streams:
  - Binance API
  - Alphavantage API
  - Datascraping using JSON from blockchain.com containing onchain data

Data Cleaning

- Dataset was cleaned
  - Unnecessary columns dropped
  - Null rows removed
  - Ensured 0 duplicates
  - Data types formatted for analysis

Exploration

- Statical and visual exploratory data analysis was conducted to gain a better understanding of our dataset
- Identified unnecessary columns for the model

Classification

- Identified that predicting if future price is up/down as a classification problem
- Model's evaluated for accuracy, AUC/ROC
- Model built and metrics analyzed

TS Forecast

- Evaluated multiple models for time series forecasting and scored based on MAPE
- Model built and metrics analyzed

Reporting

- Reporting of findings in PowerPoint for non-technical stakeholders
- Jupyter notebook presented for technical stakeholders

DATASET CHARACTERISTICS (CLASSIFICATION)

Institute of Data

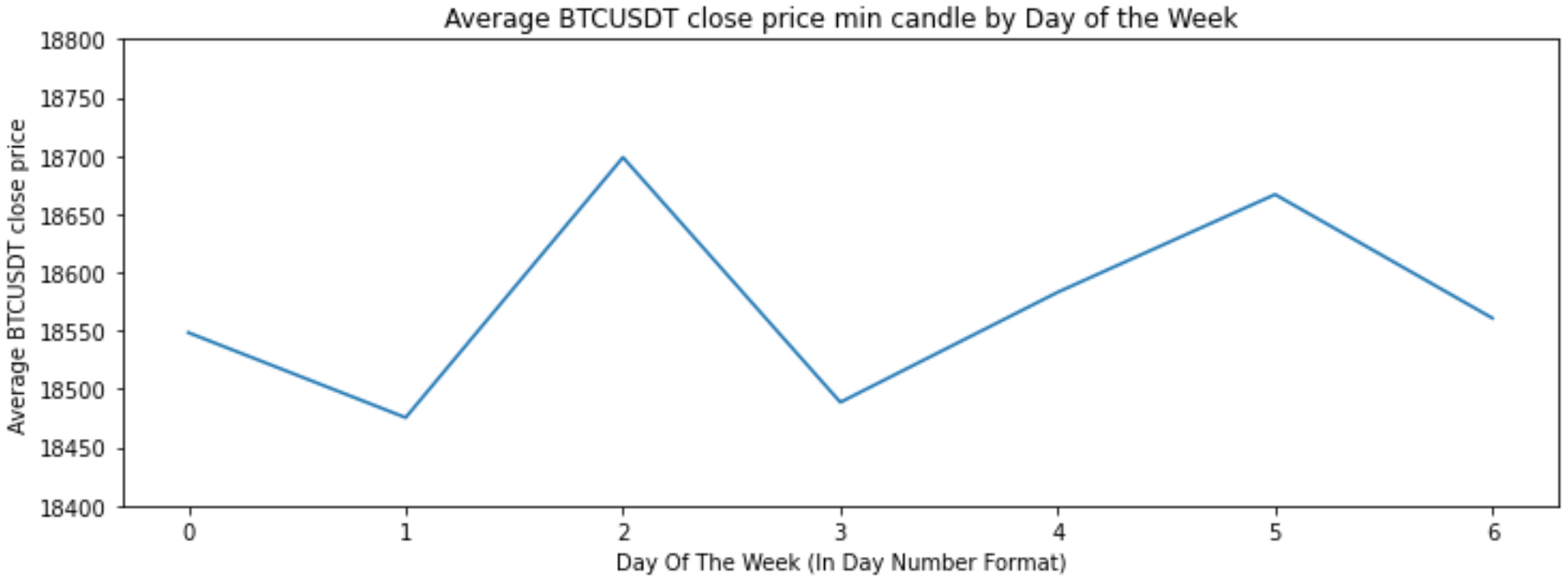
Model	# of Rows	# of Columns	# of Numeric Cols	# of Categorical cols	# of DateTime cols
TS Forecast	1190	27	27	0	1
Classificagtion 1190	1190	11	9	2	0

EXPLORATORY DATA ANALYSIS

2021 Bitcoin/USD Price

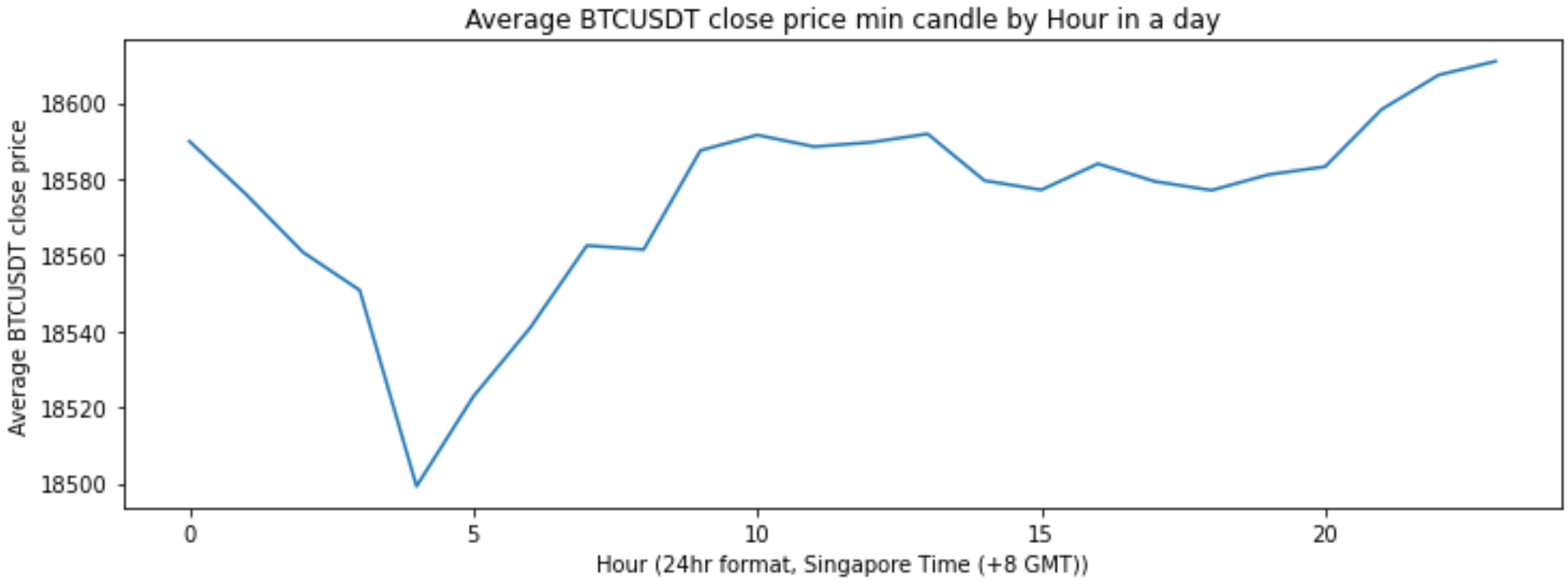
Source: Binance API (Jan 2021 to Sep 2021 Daily Close data)

EXPLORATORY DATA ANALYSIS



Source: Binance API (Jan 2019 to Sep 2021 minute on minute data)

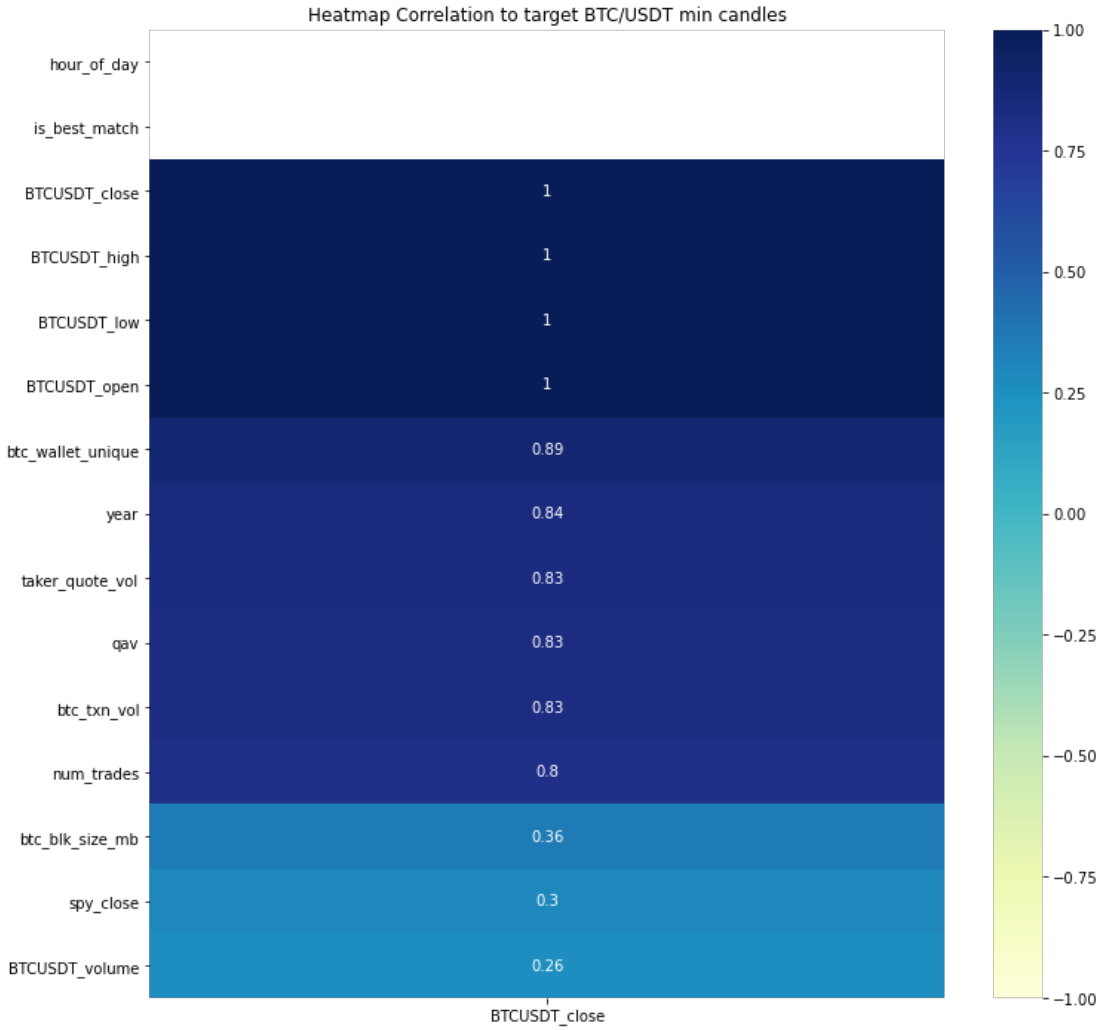
EXPLORATORY DATA ANALYSIS



Source: Binance API (Jan 2019 to Sep 2021 minute on minute data)

EXPLORATORY DATA ANALYSIS

- Onchain data and OHLCV data are better correlated to the target.
- We should possibly feature engineer to find technical indicators as they could be really good predictors to the target.
- Stock market indices and economic data were poorly correlated to the target.

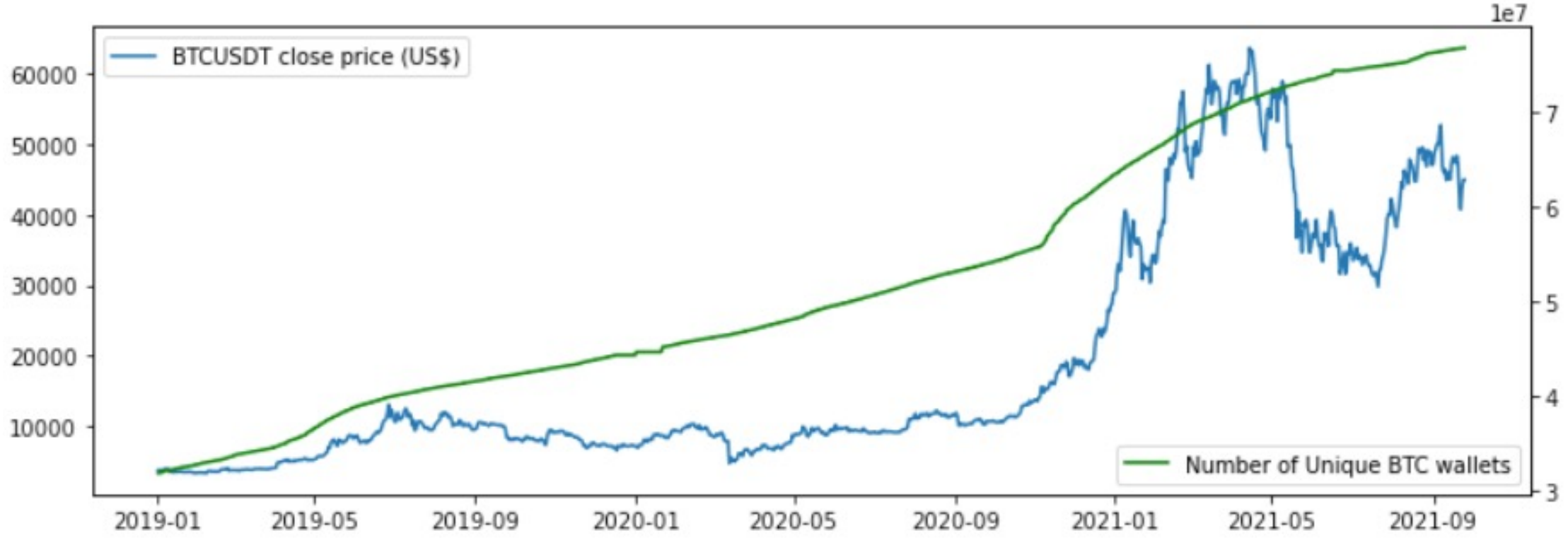


Source: Binance API, blockchain.com, Alphavantage API

EXPLORATORY DATA ANALYSIS

Institute of Data

2021 Bitcoin USD Price VS # of Unique BTC wallets

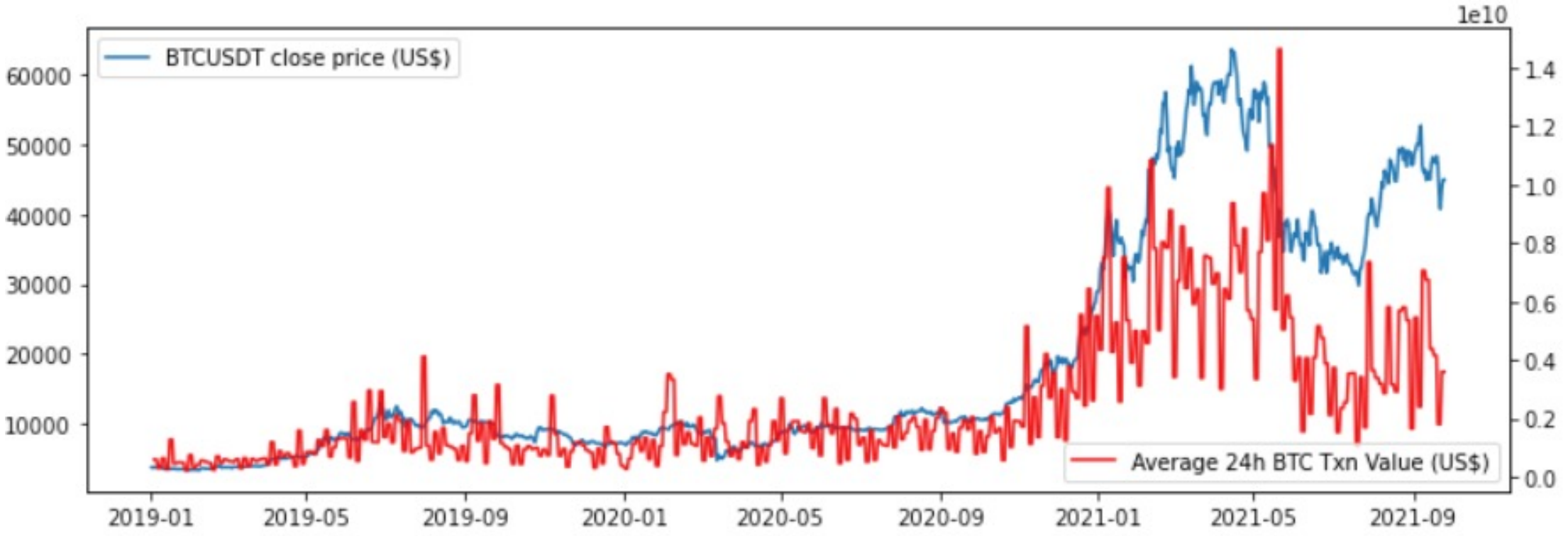


Source: Binance API, blockchain.com

EXPLORATORY DATA ANALYSIS

Institute of Data

2021 Bitcoin USD Price VS # of Unique BTC wallets



Source: Binance API, blockchain.com

CLASSIFICATION MODEL EVALUATION (DETERMINE PRICE DIRECTION)

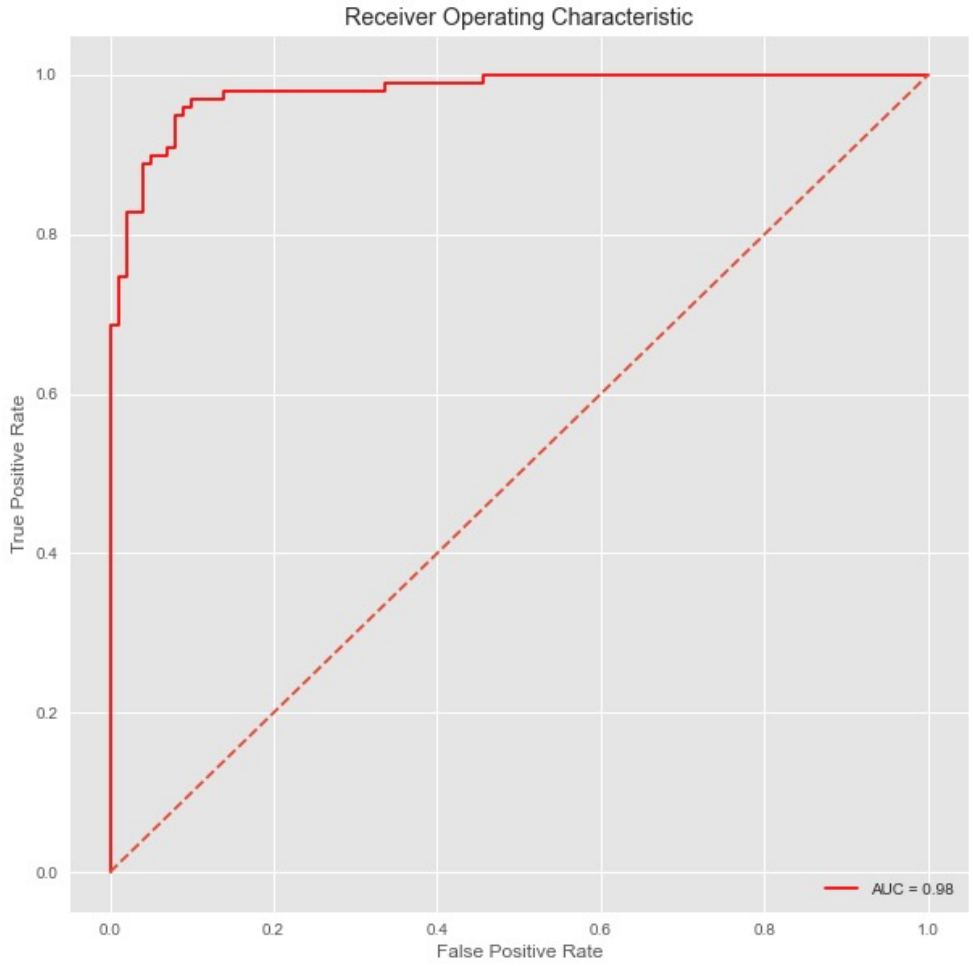
Model	Accuracy	AUC
XGBoostClassifier*	0.93	0.98
Ridge Classifier	0.54	0.00
Linear Discriminant Analysis	0.54	0.50
Logistic Regression	0.54	0.52
Quadratic Discriminant Analysis	0.52	0.51
Ada Boost Classifier	0.50	0.47
SVM - Linear Kernel	0.50	0.00
Naive Bayes	0.50	0.52
Gradient Boosting Classifier	0.47	0.44
Decision Tree Classifier	0.47	0.47
CatBoost Classifier	0.46	0.44
K Neighbors Classifier	0.46	0.43
Random Forest Classifier	0.46	0.42
Light Gradient Boosting Machine	0.46	0.41
Extra Trees Classifier	0.45	0.42

Source: Binance API

CLASSIFICATION MODEL EVALUATION (AUC-ROC CURVE)

Institute of Data

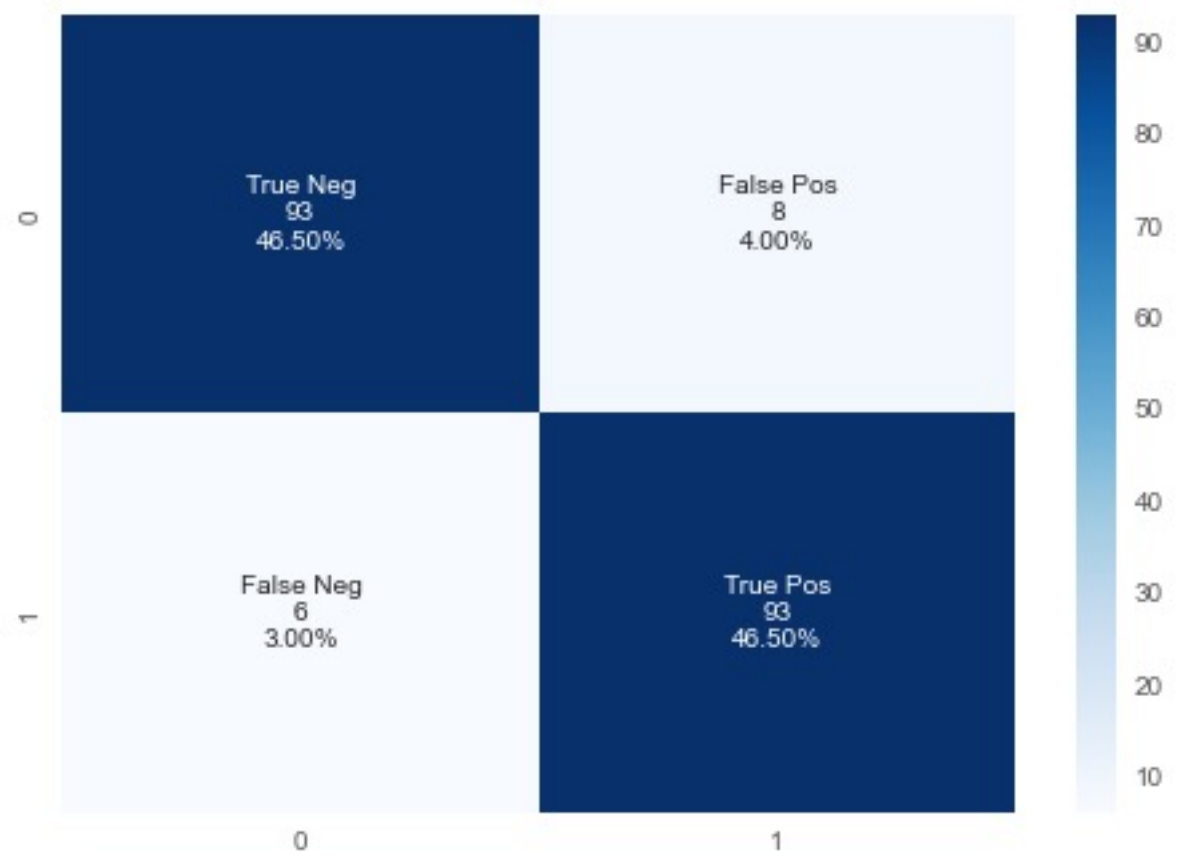
- AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.



Source: Binance API

CLASSIFICATION MODEL EVALUATION (CONFUSION MATRIX)

- TPR = 99
- Specificity = 92
- FPR = 0.04



Source: Binance API

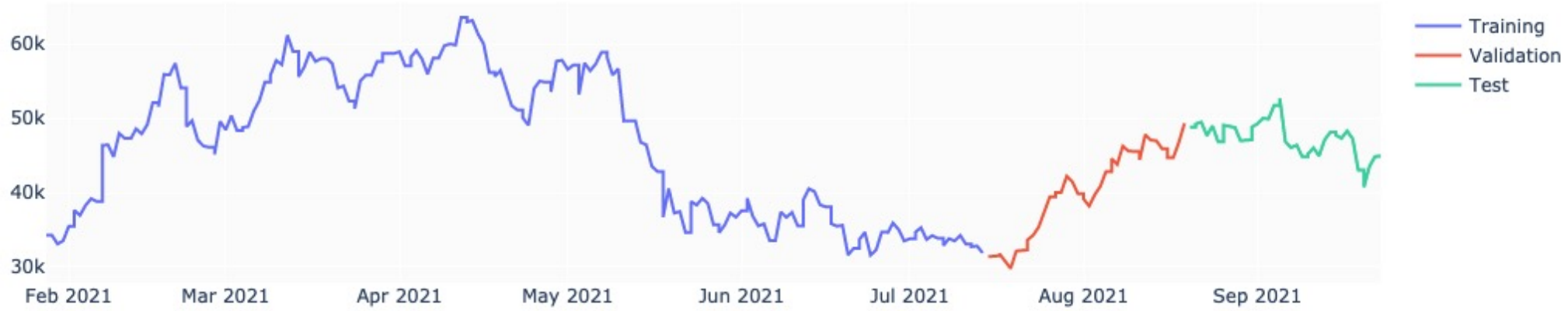
TIME SERIES FORECAST MODEL EVALUATION

Model	RMSE	MAPE
XGBoost	30203042	9%
RNN-LSTM	*	14%
Auto ARIMA	4071315	107%
ARIMA	4078818	190%
Decision Tree w/ Cond. Deseasonalize & Detrending	8088001	397%
Gradient Boosting w/ Cond. Deseasonalize & Detrending	5388718	411%

Source: Binance API

TIME SERIES XGBOOST FORECAST MODEL (TRAIN TEST SPLIT)

- DataFrame into three subsets: training (70%), validation (15%) and test (15%) sets. I calculated split indices and create three separate frames (train_df, valid_df, test_df). All three frames have been plotted in the chart below.



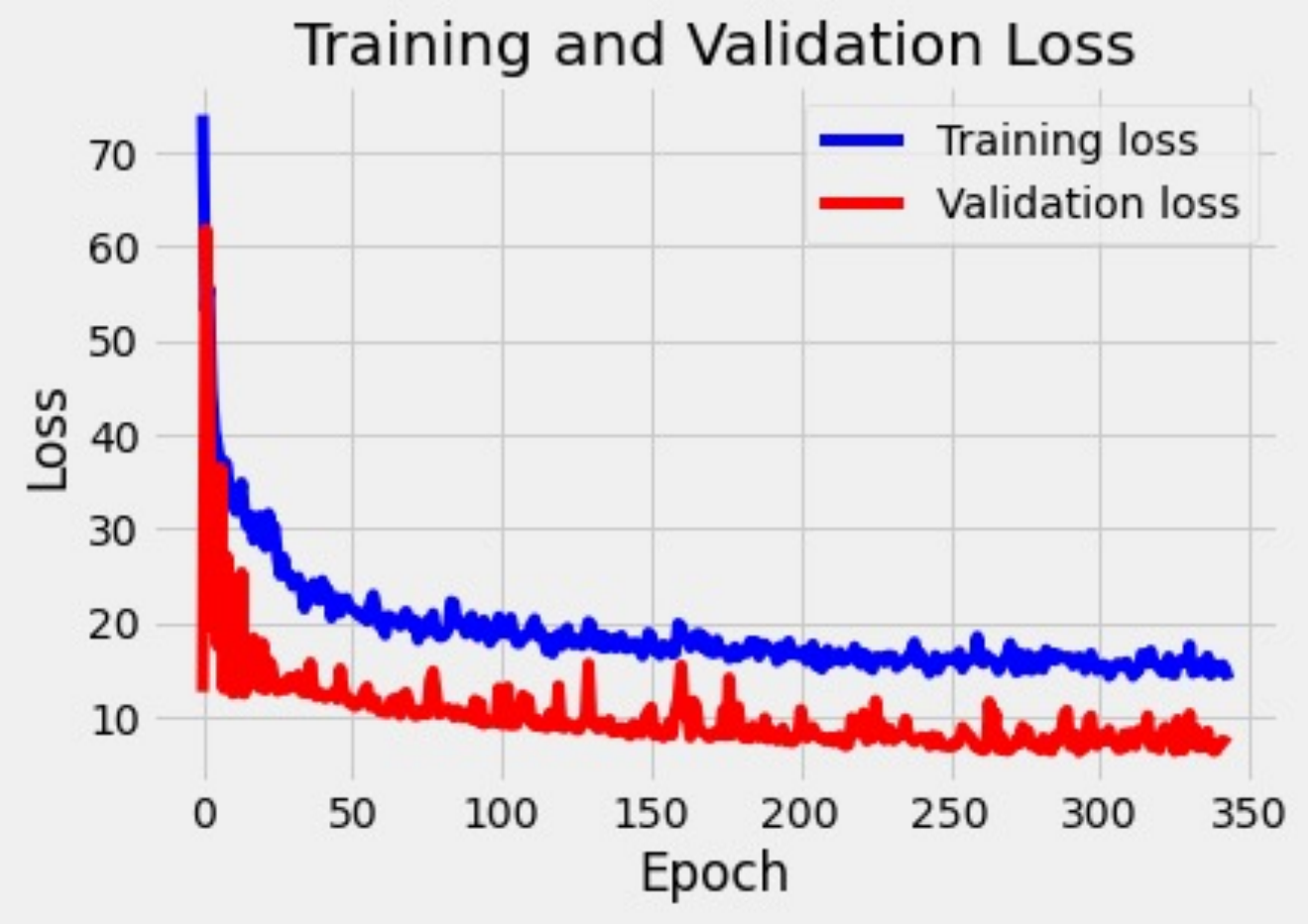
Source: Binance API

TIME SERIES XGBOOST FORECAST (PREDICTIONS)

Source: Binance API

TIME SERIES RNN-LSTM (TRAINING-VALIDATION LOSS)

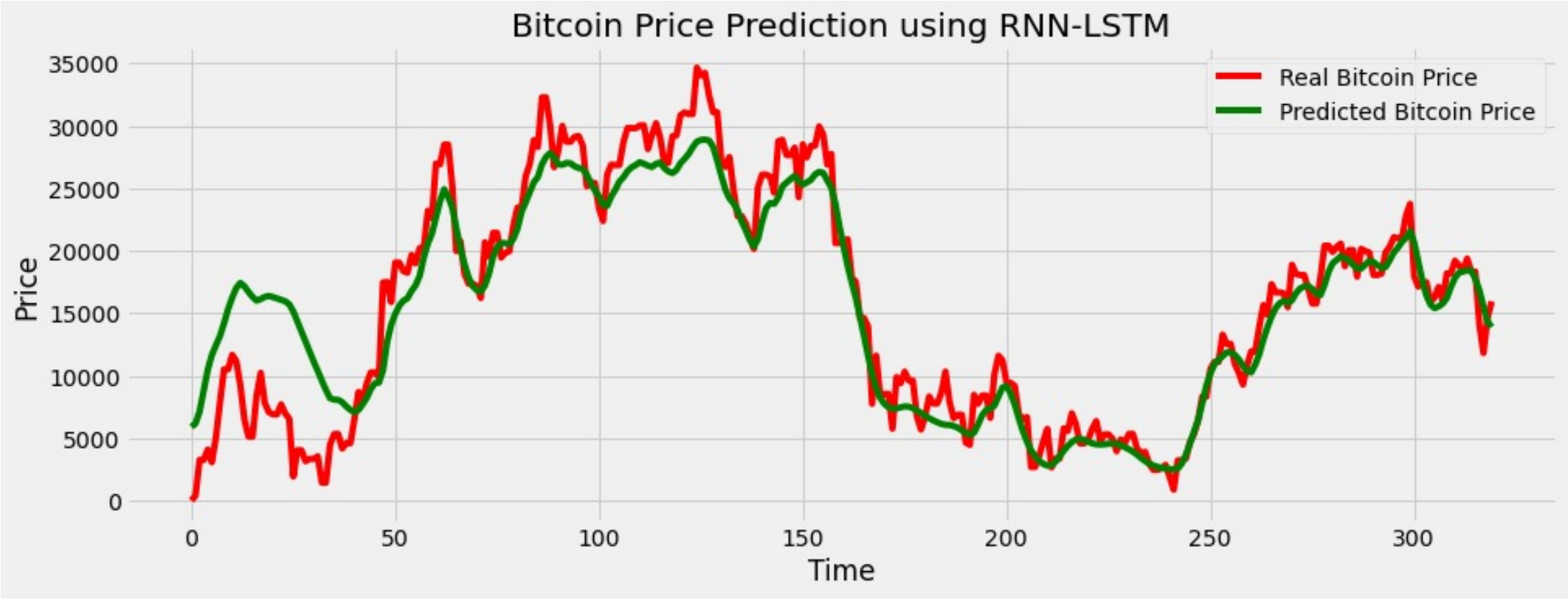
Institute of Data



Source: Binance API

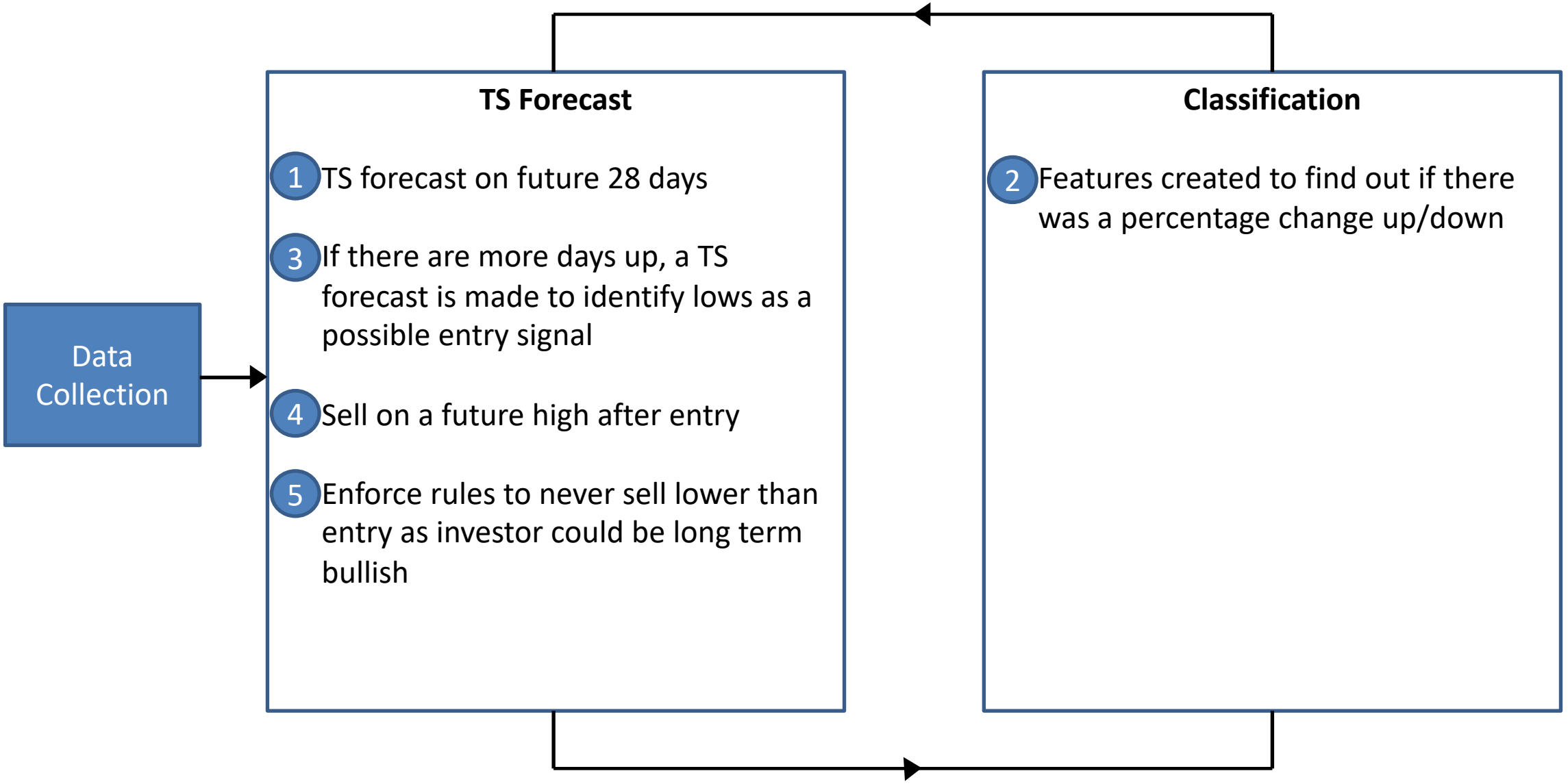
TIME SERIES RNN-LSTM (PREDICTIONS)

Institute of Data



Source: Binance API

PROPOSED APPROACH FOR PRODUCTION



FUTURE DEVELOPMENTS

- Porting over models to production grade using Gemini/FTX API due to regulatory concerns
- Testing of other models
- Backtesting of approach and models in testnet environment VS control of buy/hold strategy

Thank you