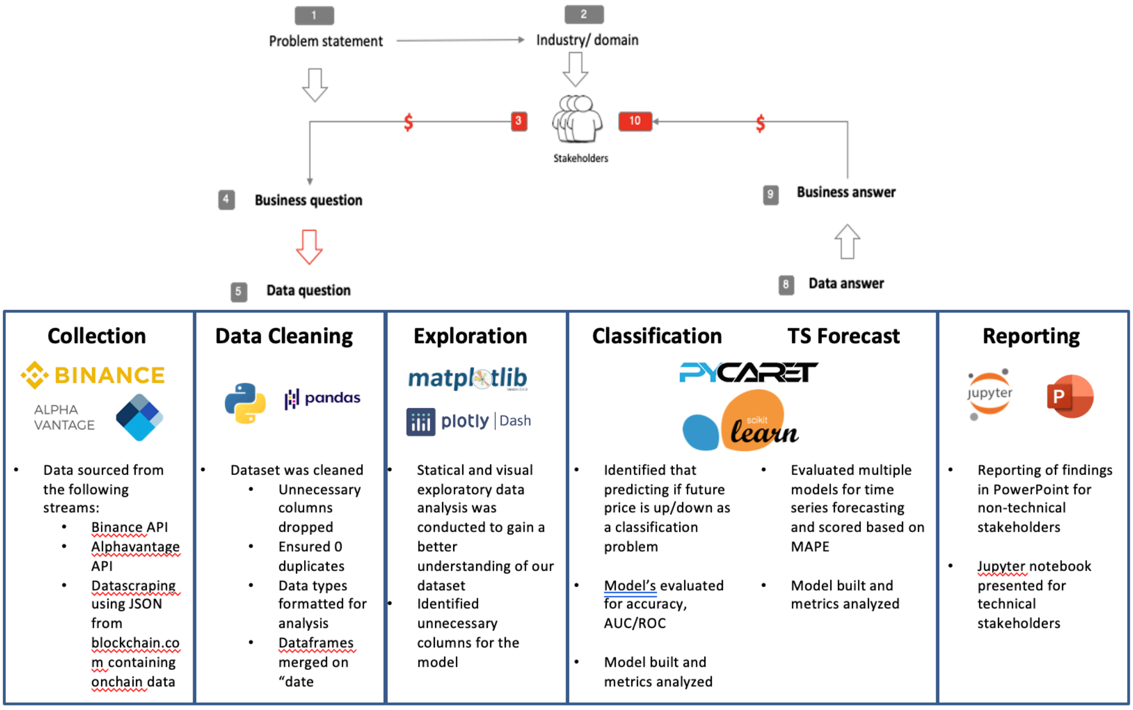
Cryptocurrency Capstone Project

Process overview

The following diagram shows the overall end-to-end process for defining, designing, and delivering the Capstone project.



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
    - Ensured 0 duplicates
    - Data types formatted for analysis
    - Dataframes merged on “date

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

Problem statement

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.

Predictive analysis of cryptocurrency provides the right methodology for explicit predictions of cryptocurrency markets helps the investors to make the right investment decisions to attain profits.

Using Machine Learning algorithms such as numerous supervised learning and neural networks we will be able to analyse the price fluctuations of the cryptocurrencies through historical data.

Industry/ domain

Cryptocurrency is one of the fast-growing digital money in the present world that is still in its infancy.

The industry provides investor’s with multiple means of exposure and the easiest entry into the industry is through directly trading the assets on cryptocurrency exchanges.

Other more sophisticated methods include trading futures/derivatives, staking of tokens to liquidity pools, and trading in non-fungible tokens (NFT’s)

Stakeholders

The stakeholder for this project would be a cryptocurrency investor who would like to invest utilizing a data driven strategy

Business question

How can an investor use data to identify entry and exit positions in trading of digital assets

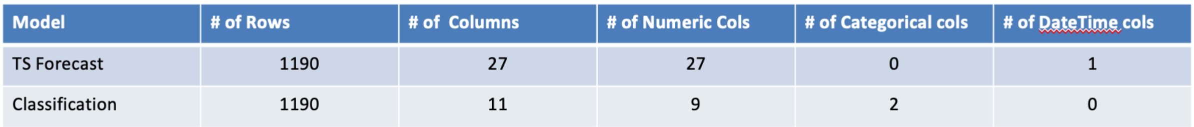
Data question

Historical data on digital assets need to be scraped
On chain blockchain data need to be scraped

Data

- Where was the data sourced?
  - Binance API
  - AlphaVantage API
  - Blockchain.com web scraping

- What is the volume and attributes of the data?
  End of day historical prices were used for the Time Series forecast and classification problem



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

- How reliable is the data?
    - As the data is primary data taken directly from the data sources it is accurate and reliable
- What is the quality of the raw data?
    - Quality of the raw data was verified to be true when compared to historical data on other websites
- How was this data generated?
    - Data from cryptocurrency exchange, Binance, is recorded after each trade and we collected minute on minute and end of day data as a comparison
    - Data from AlphaVantage is scraped from multiple providers that publish economic indicators, this data is stored into a database and AlphaVantage provides a premium service as well
    - Data scraped from blockchain.com was recorded from historical transactions on the bitcoin blockchain
- Is this data available on an ongoing basis?
    - The data is available for access on an ongoing basis

Data science process

Data analysis

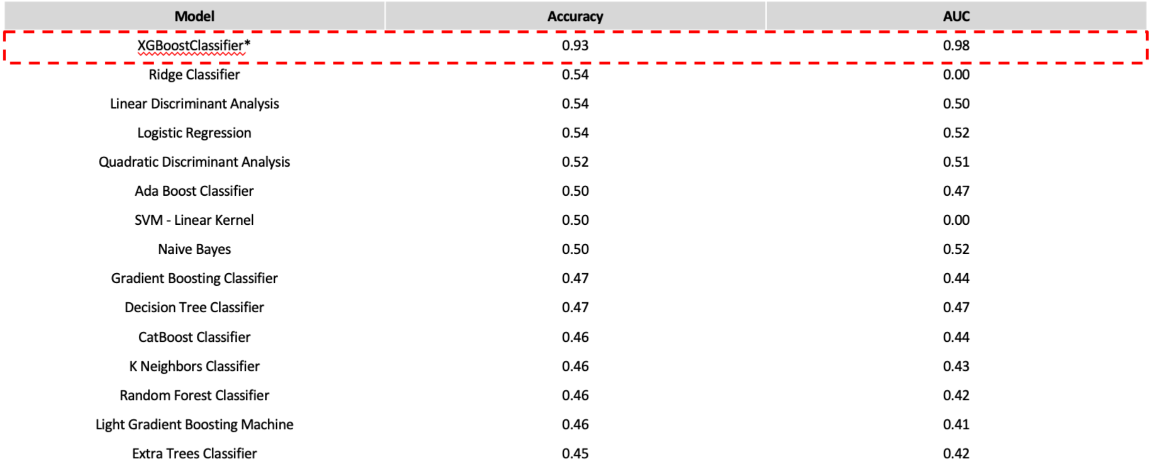
- What data pipeline was to wrangle the raw data?
    - A custom function was created to use API from the 3 different streams and gather the data required, within the function each column was cleaned to the required readable text
    - A forwardfill was used for the economic data, which is either released on a monthly, quarterly basis
    - Each column is put into its own DataFrame, and the resulting DataFrames are merged
    -
- What are the highlights of the Exploratory Data Analysis (EDA)?
    - Onchain data, and the assets actual price data (open, high, low, close, volume) were more closely correlated to bitcoin price movements compared to stock market indices and economic data
- Is the pipeline reusable?
    - The pipeline is reusable, and easily modifiable for different or multiple digital assets, and other stock market data not analyzed in this project

Modelling

(i) <u>Classification – XGBoostClassifier</u>

We decided to design a model that would predict if next day returns were either a percentage gain/lost and use a classification model to solve this problem.

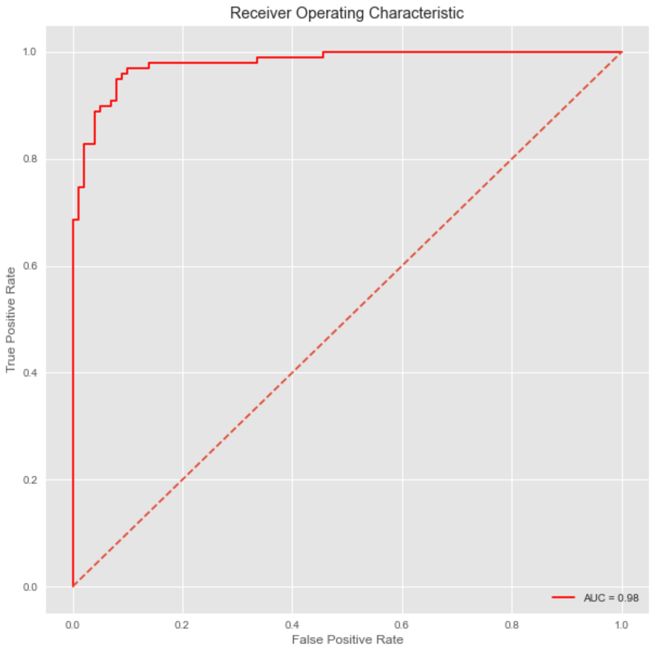
Multiple classification models were screened using Pycaret as a base and these were our results



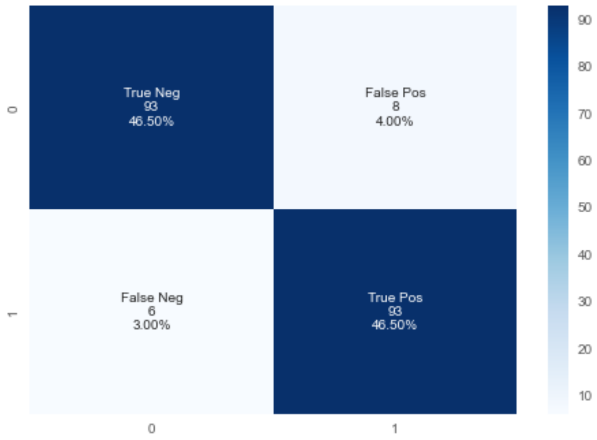
Model	Accuracy	AUC
XGBoostClassifier*	0.93	0.98
Ridge Classifier	0.54	0.00
Linear Discriminant Analysis	0.54	0.50
Logistic Regression	0.54	0.52
Quadratic Discriminant Analysis	0.52	0.51
Ada Boost Classifier	0.50	0.47
SVM - Linear Kernel	0.50	0.00
Naive Bayes	0.50	0.52
Gradient Boosting Classifier	0.47	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

On further analysis using XGBoostClassifier, we used an AUC-ROC Curve as a performance metric and looked at the confusion matrix to see if the model could accurately predict next day returns

AUC-ROC Curve



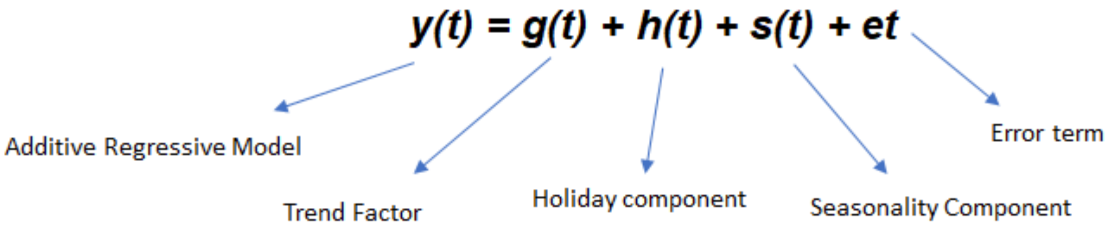
Confusion Matrix



(ii) Time Series - AR – Facebook Prophet

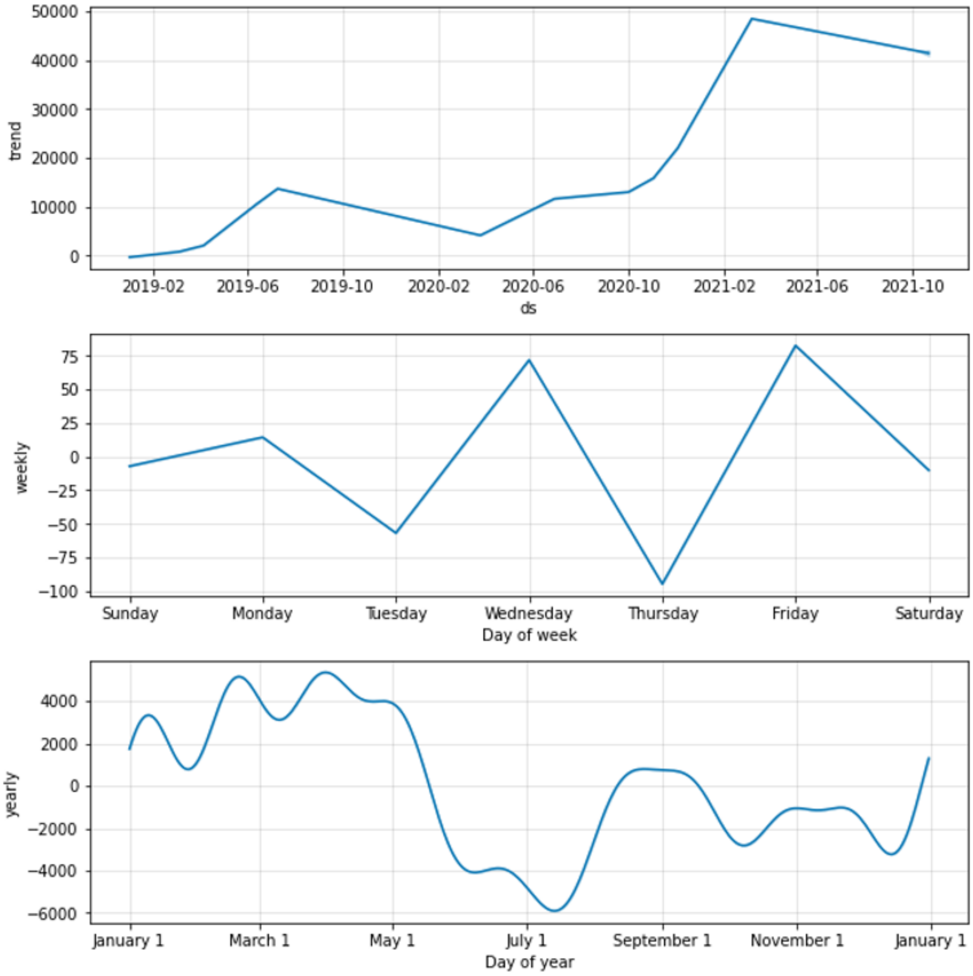
The model that we decided worked best was the Additive Regressive model, using Facebook Prophet which takes in only 1 feature column, in this case historical price

The model then looks for trend factors, holiday components, a seasonality component and the error term with the following mathematical formula

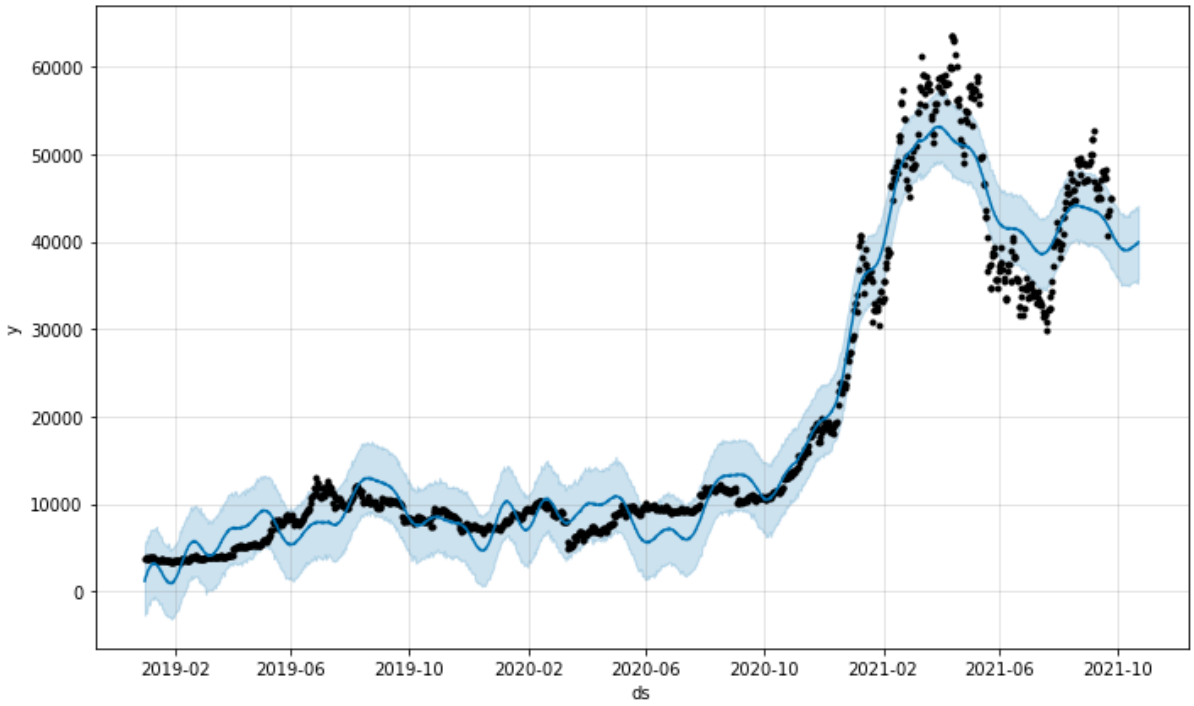


Prophet Forecasting Model Equation

A chart of these components can be seen as follows:



The model was able to make predictions with a mean absolute error percentage of 10% which is considered “excellent” and made the following predictions as below



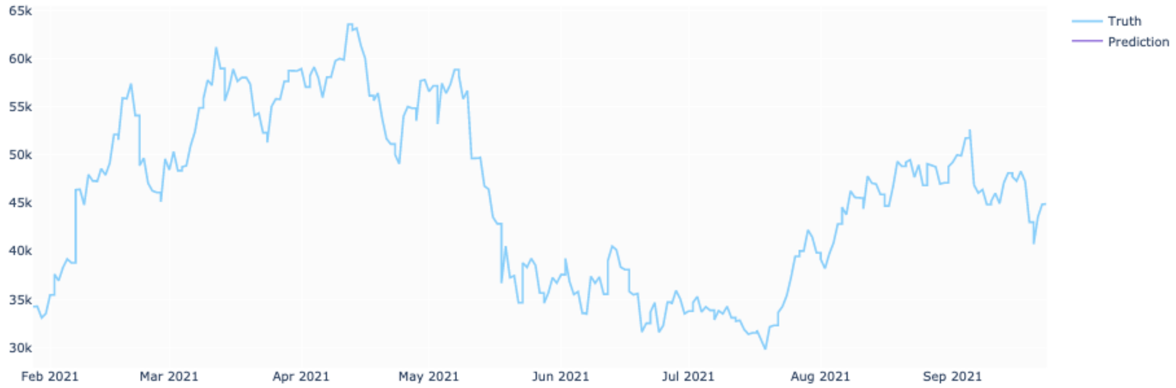
(iii) Time Series - XGBoostRegressor

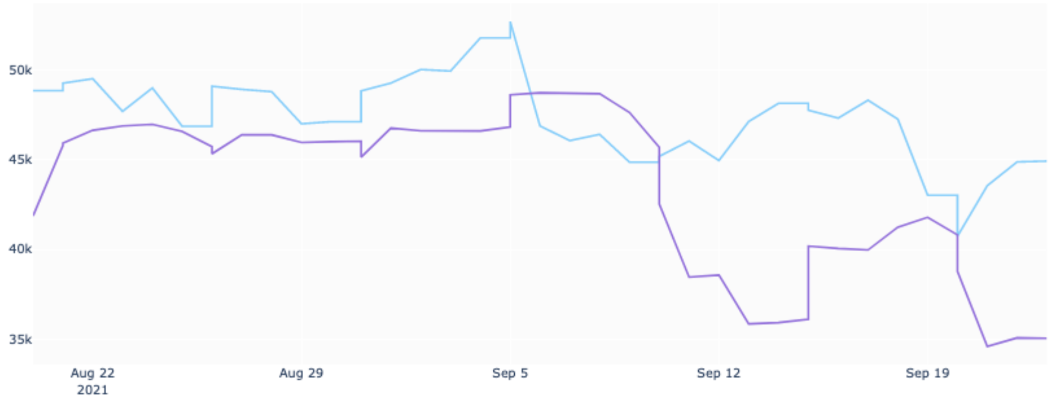
The second contender we had for a model was XGBoost. Besides the open high low close volume candles, we also analyzed some technical indicators; simple moving averages, exponential moving averages, relative strength indicator and the MACD

- What feature engineering techniques are used?
    - We did some feature engineering to calculate the technical indicators used

- How did we improve the model?
    - We applied a gridsearchCV to find the tune the hyperparametors to increase accuracy of the model

The model was able to make predictions with a mean absolute error percentage of 9.3% and made the following predictions as below



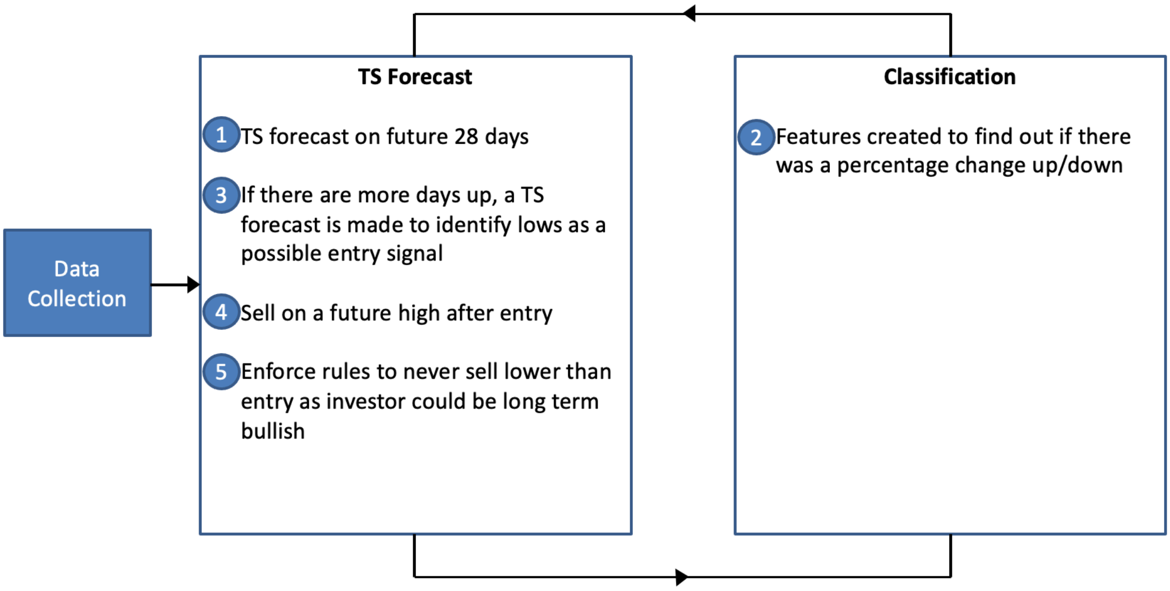


Outcomes

- The main outcomes show that the models picked were able to accurately predict bitcoin prices with what data science defines as excellent, based on the accuracy metrics chosen

Implementation/End to End Solution

- We will consider moving the model over to production with the following process



Data Collection

TS Forecast

1 TS forecast on future 28 days

3 If there are more days up, a TS forecast is made to identify lows as a possible entry signal

4 Sell on a future high after entry

5 Enforce rules to never sell lower than entry as investor could be long term bullish

Classification

2 Features created to find out if there was a percentage change up/down

Data answer

We feel that the data science question was not answered satisfactorily. As the models were able to predict to what are defined as satisfactory error levels for the problem

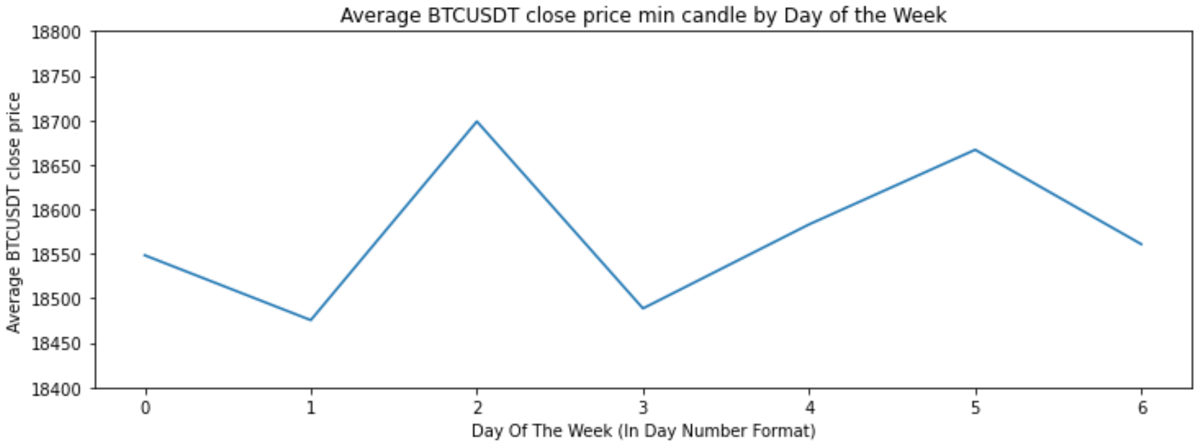
Business answer

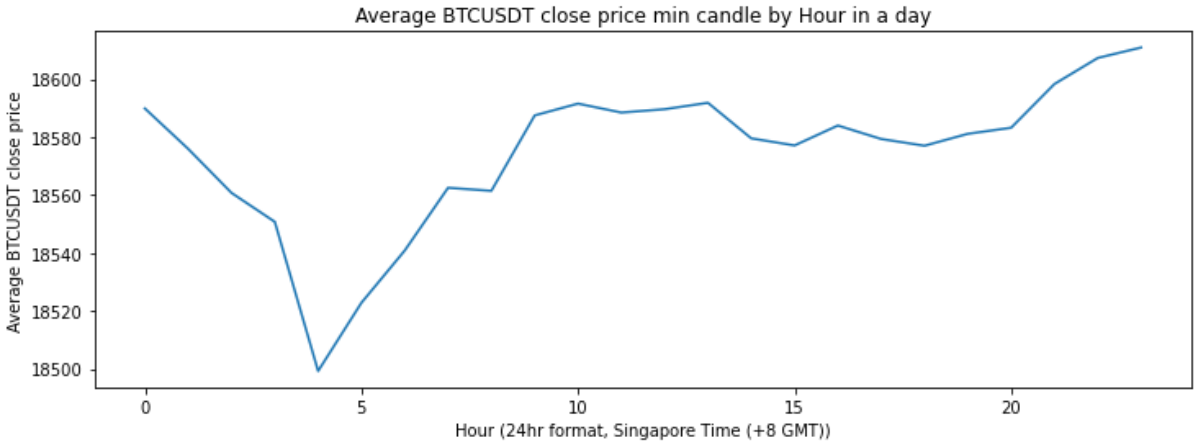
However the business question was not answered, as the model needs to be ported over to production and the trading strategy back tested and benchmarked

- We recommend to the stakeholder to take the following weekly buying opportunity's, and buying on the weekly/daily dips in the below charts while the model is ported over to production

Response to stakeholders

- We recommend to the stakeholder to take the following weekly buying opportunity's, and buying on the weekly/daily dips in the below charts while the model is ported over to production





References

Libraries used in the project

- Binance
- Pandas
- Numpy
- Alpha_vantage
- Requests
- Json
- Datetime
- Functools
- Sklearn
- Statsmodels
- Pylab
- Matplotlib

- Pycaret
- Sktime
- Plotly
- Xgboost
- Keras
- Fbprophet