Intermediate Report - Prediction of Bitcoin Prices

Team Members: Ambuj Arora, Vai Suliafu, Yicong Xiao

Progress towards the proposed goal

Our goal is to predict the Close Price for the next time window given historical trade data of BTC. For our analysis, we started with 5-minute rolled up data. As a starting step, we analyzed and explored the data and made some basic observations before training machine learning models on it. Next, we created around 70 features using the TA package in Python which includes technical indicators like Bollinger bands, RSI, MACD, etc. After dropping unwanted features(with most null values), we had 58 features in total. These features are one of volume, volatility, trend, and momentum type indicators to capture the price variation with time in all these four aspects. We have trained five different machine learning models to predict the close price for a given time window based on the data of the last 360 time windows.

Data Exploration

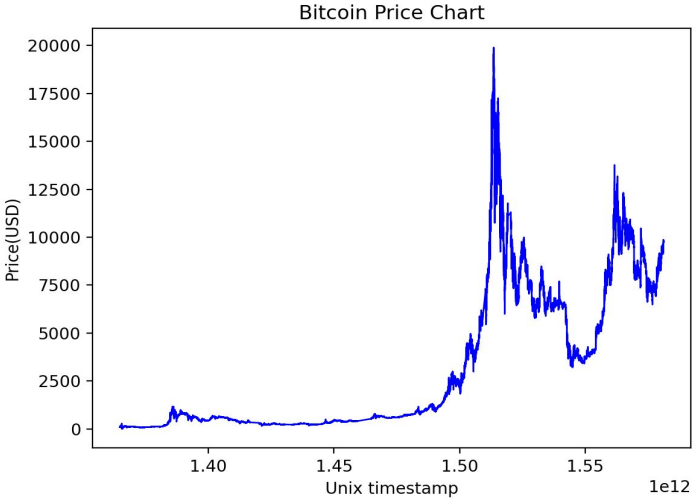
Data exploration is an important task when dealing with complex datasets. We use data visualization techniques to help explore and analyze raw data effectively. We made the following observations:

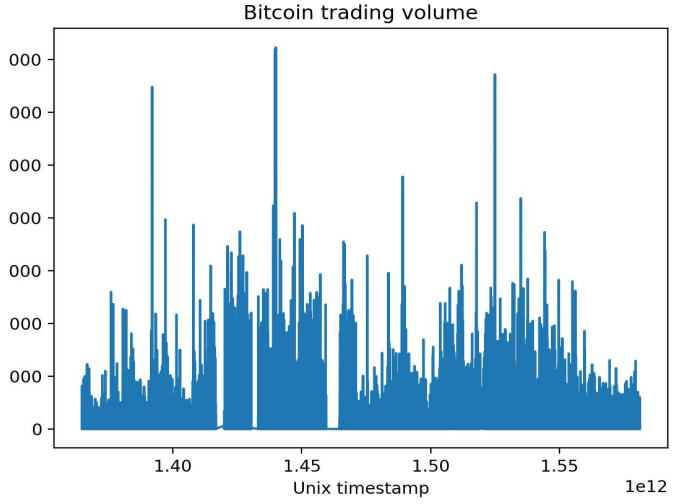
From Price Charts

- Until July 2017 (When the bitcoin became famous), the price of bitcoin was relatively stable.
- In December 2017, the price of bitcoin peaked at around $20,000.
- Starting in 2019, the price of bitcoin starts to fluctuate between $3,000 and $13,000.
- Also, we observe that generally, the price at a particular timestamp depends on the trend of the price in the last few hours.

From Volume Charts

- There have been three trading peaks, the first two appeared before bitcoin became famous and the third appeared right after the bitcoin price peaked.
- After December 2017, the trading volume of bitcoin generally showed a downward trend.
- Apart from a few large transactions, the majority of transactions are under 2,000.





Data Preprocessing

- Time Stamp Conversion: The UNIX timestamp was converted to DateTime format using Pandas' datetime function.
- Rolling Up: The raw data was rolled up from 1-minute candlestick to 5, 10, 15, 30 and 60-minute candlesticks using Pandas. Currently, we have done our analysis on 5-minute data.
- Feature Creation: 68 features like Bollinger bands, MACD, RSI, PSAR, etc. were created using the given attributes using the TA Python package.
- Null Value Treatment: Those features were dropped which had most of the values as null. Also, there were some features that had discontinuous null values. To preserve the continuity in the data points according to 5-minute window timestamps, we dropped these features too. Finally, all those rows were dropped which still had null values. These were the initial rows in the data that had null values for some features.

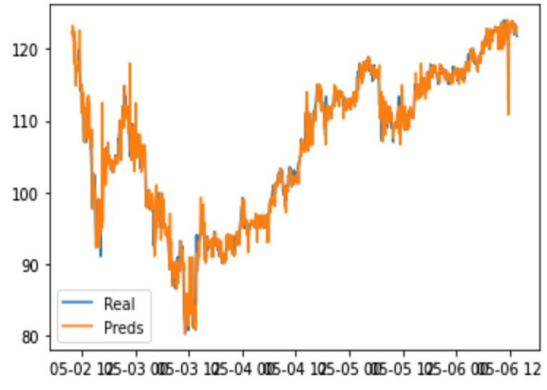
Intermediate Modeling

After cleaning and pre-processing the data, we trained five different models on it. These models are Decision Tree, Light GBM, XGBoost, Linear Regression, and SVM Regressor.

Based on our initial observations on the raw data we found that generally, the price at a particular timestamp depends on the trend of the price in the last few hours. Thus, for any given timestamp, we took the price trend in the last 30 hours as the training data. As the data was rolled up to 5-minute window timestamps, we took the last 360 data points for any given timestamp as its training data and we were able to predict the price at that timestamp. We hypothesized that because we have captured the volume, volatility, trend, and momentum of the price variation using the custom made features, the models should perform well.

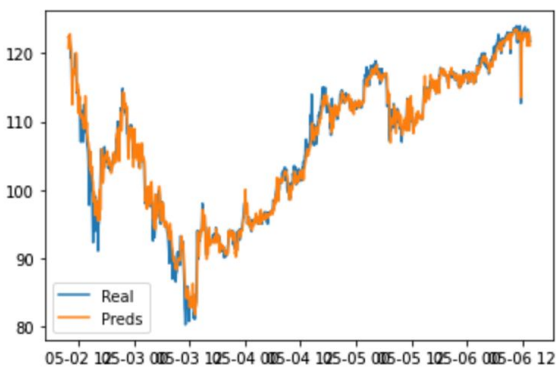
The following are the graph plots of the true and predicted prices from 2013-05-02 09:45:00 to 2013-05-06 13:40:00 with the Average Root Mean Square Error and the Standard Deviation between the predicted and actual close prices, for all the five models. The training time distributions have also been plotted.

RMSE mean:1.1082090833333333, std:1.3589983042244376



Decision Tree

RMSE mean:0.976544255629993, std:1.0551912545762463



Light GBM

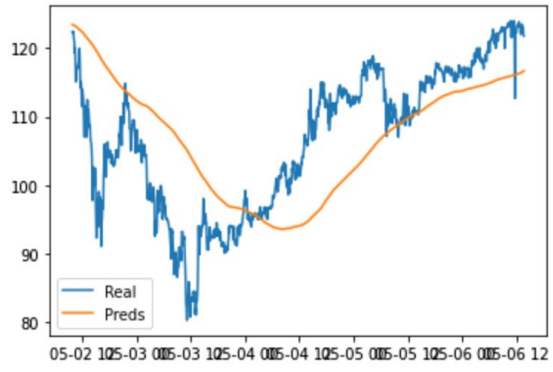
RMSE mean:0.9756259522739466, std:1.06480612790974

Linear Regression

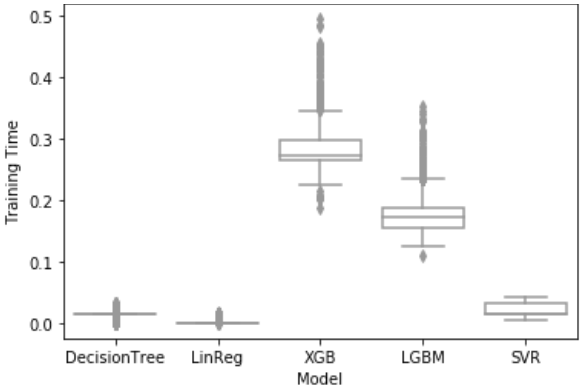
RMSE mean:1.007806593961111, std:1.073035377178316

XGBoost

RMSE mean:8.040499649785465, std:5.634449785913653



SVM Regressor



Training Time Distributions

Intermediate Results

Based on the RMSE and std values, we found Light GBM and Linear Regression to work the best. We also found that SVM Regressor just learns the general trend of the price in the given time range. It is not able to capture the price variation. The other four models perform well on the selected test sets and can be improved by using feature selection, hyperparameter tuning, etc. We also found that XGBoost and Light GBM took more time to train as compared to the other models.

Future Plan

For the next steps, we will continue to build and test more models. Specifically, we plan to further test the rolling window parameter and hypertune the existing models. We also plan to increase our test sizes for more robust error generalization. Lastly, we would also like to try a few more models if possible, or even possibly some custom ensembles, and then select the best model for our final results. It would be interesting to see our best model to predict the BTC price in real-time with minimum error. So, this would be one of the end goals that we would want to achieve.