Cryptocurrency Price Prediction

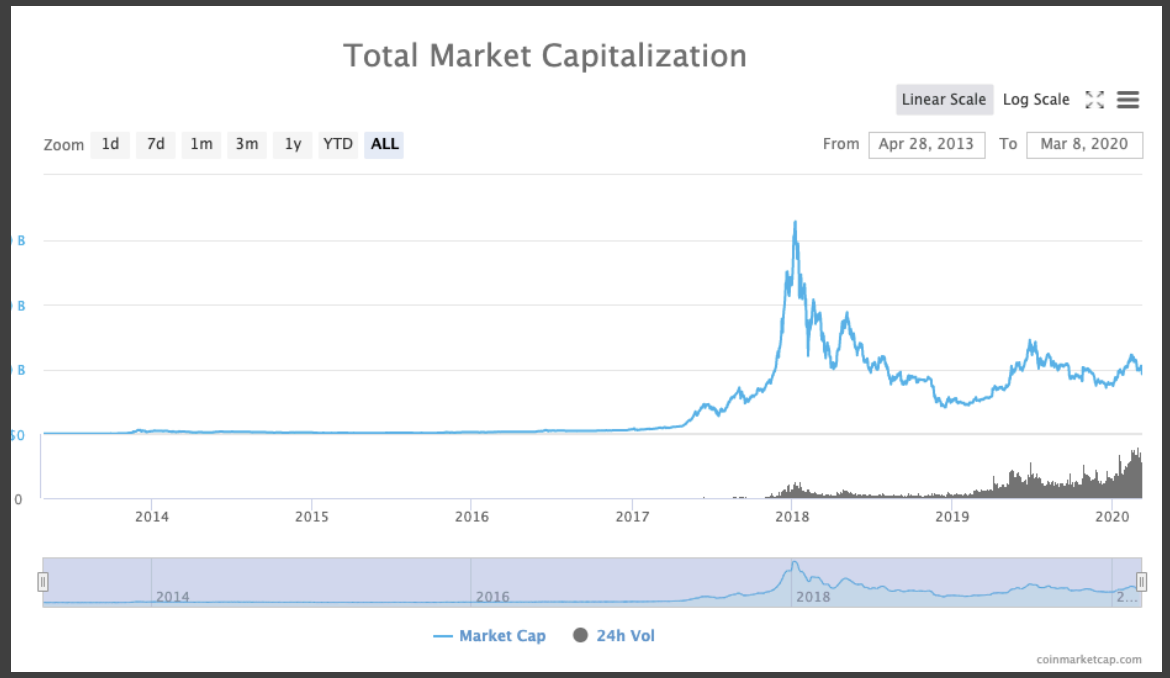
Ruinan Lu

Outline

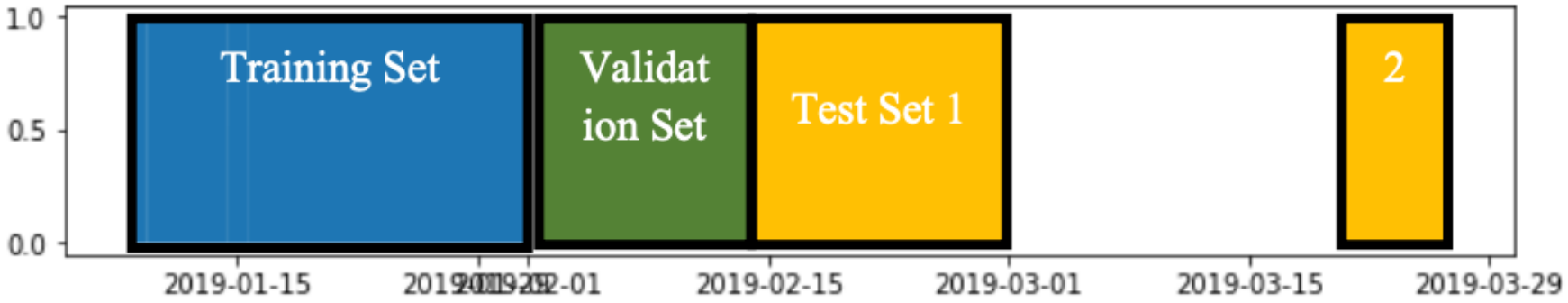
- Background and Introduction
- Data Description and Exploration
- Machine Learning Models
- Results and Discussion

Business Objective

- The cryptocurrency market has experienced rapid growth in the past decade. On an almost daily basis, new cryptocurrencies are created, and the public is paying increasing attention to the new asset class. This market provides chances for companies to raise money without involving venture capitalists and to trade cryptos without being listed on stock exchanges. The set of coins in the crypto market ranges from the best-known cryptocurrency of our time, the Bitcoin, the very popular coins like Ripple, and Ethereum to much more other obscure coins. There are over 1900 cryptos issued up to 2019, which resulted in a market of more than $850 billion. Many investment firms have been investing in and maintaining a portfolio of cryptos. Some even are specialized in crypto trading. More than 1,500 cryptocurrencies are being actively traded by individual and institutional investors worldwide across different exchanges. Over 170 hedge funds, specialized in cryptocurrencies, have emerged since 2017 and in response to institutional demand for trading and hedging, Bitcoin futures have also been launched. Although many people are struggling to understand what cryptocurrencies are, and some consider perhaps most coins are the representative of bubbles, more and more people take a position in the crypto market. So predicting the cryptocurrency price would be beneficial to both crypto investors as well as hedge funds who want to make profit from.



Data Description

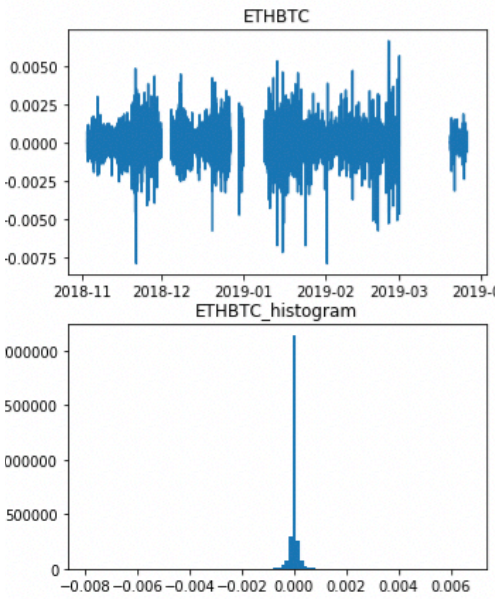


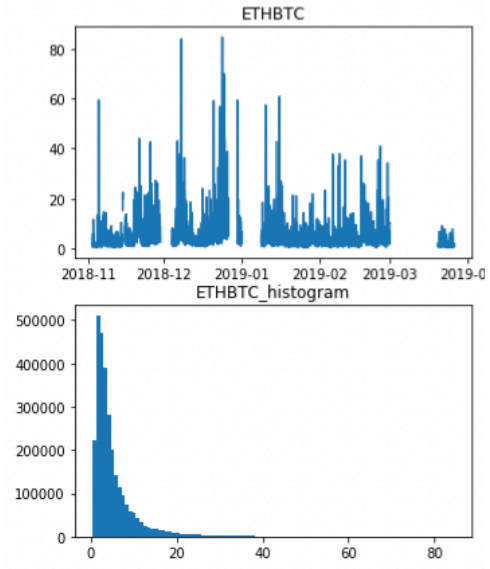
- High frequency data for 3 currency pairs ETH/BTC, ETH/USDT, BTC/USDT from Binance: Binance is a global cryptocurrency exchange that provides a platform for trading more than 100 cryptocurrencies. Since early 2018, Binance is considered as the biggest cryptocurrency exchange in the world in terms of trading volume ETH/BTC, ETH/USDT, BTC/USDT

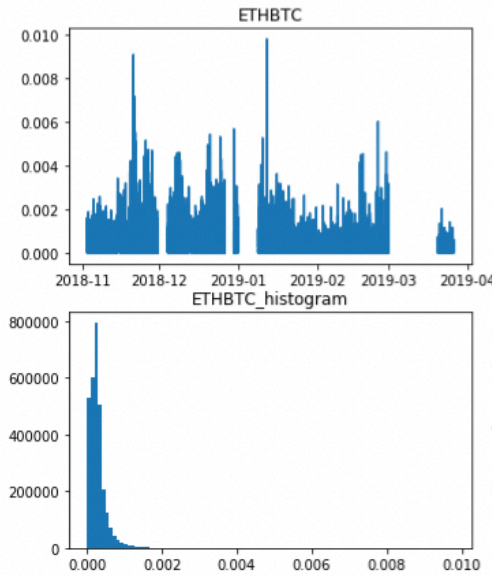
- BTC (Bitcoin) – top 1 in terms of trade volume, market cap and popularity
- ETH – second large, more flexible
- USDT – stable coin, equivalent to US dollar in crypto market
- All data in milliseconds

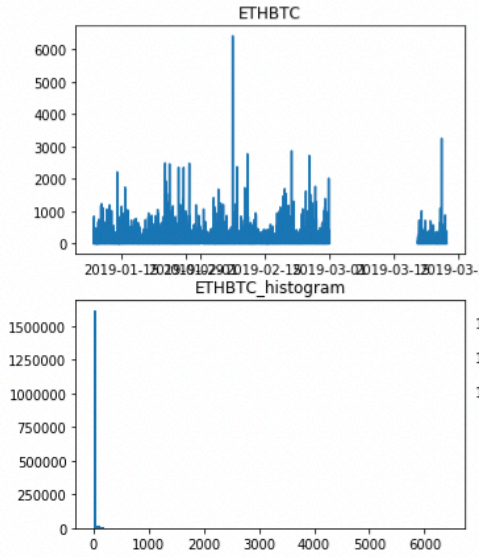
- Timeframe: Jan 8th – Mar 26th, 2019
- Train/Test/Validate Split
- 47 Market Microstructure factors

Data Exploration – ETHBTC









Return

$$r(t_j) = x(t_j) - x(t_j - \Delta t)$$

Volatility

$$v(t_i) = \sqrt{\frac{1}{n-1}\sum[r(t_j) - \frac{1}{n}\sum r(t_j)]^2}$$

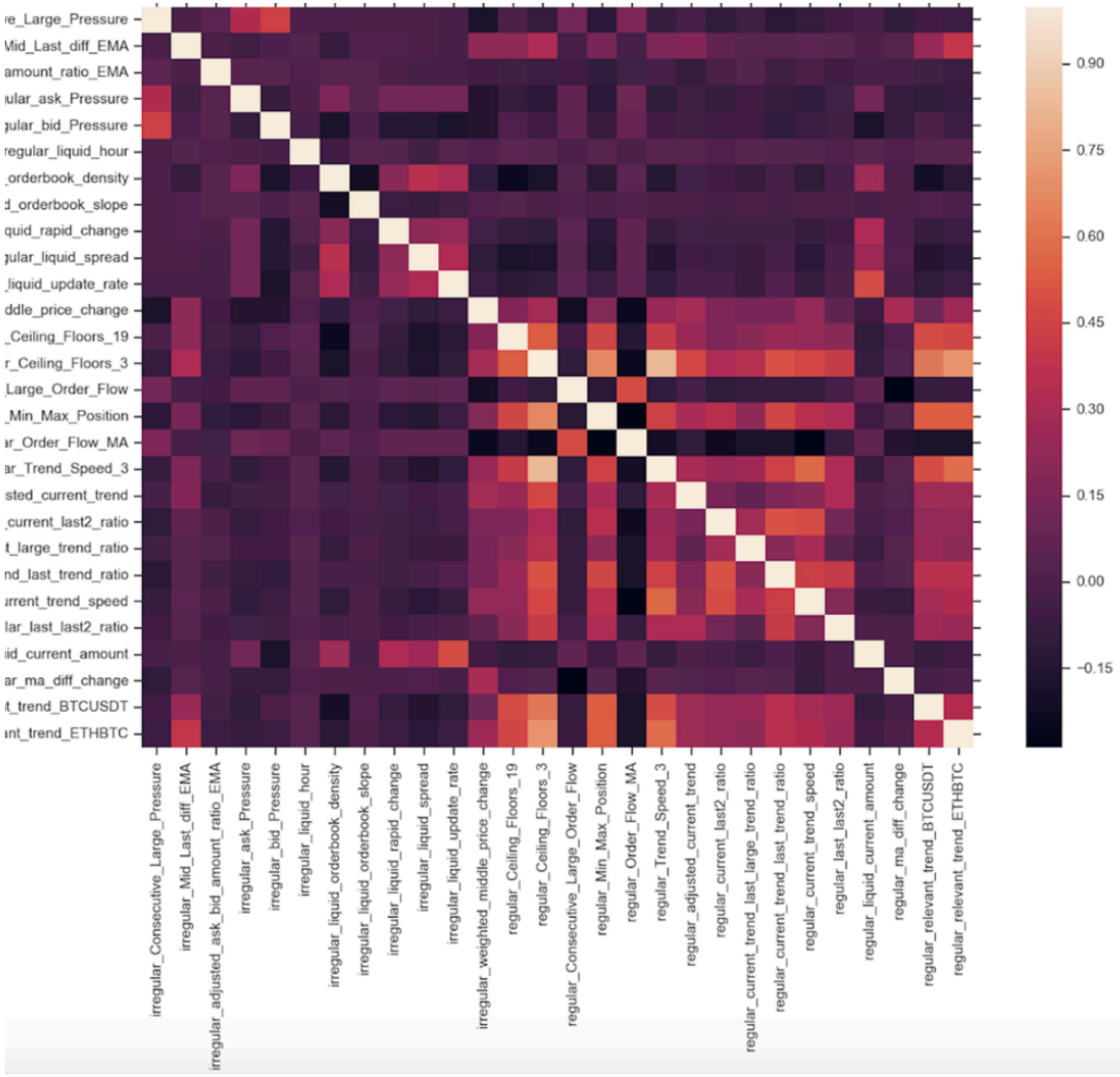
Spread

$$s(t_j) = \log p_{ask}(t_j) - \log p_{bid}(t_j)$$

Amount

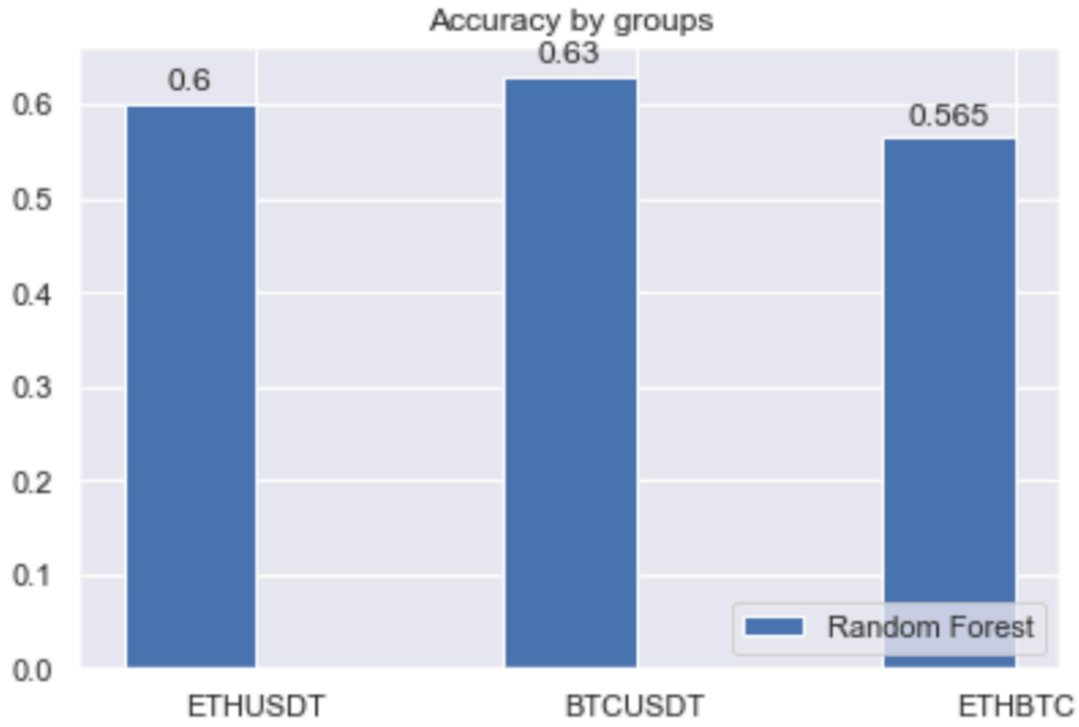
Forecast models Selection - Machine Learning Algorithms for Optimal Price Prediction

- Feature Selection:
- Heatmap to visualize correlations among features
- Tree-based models and Neural Network can handle feature correlations



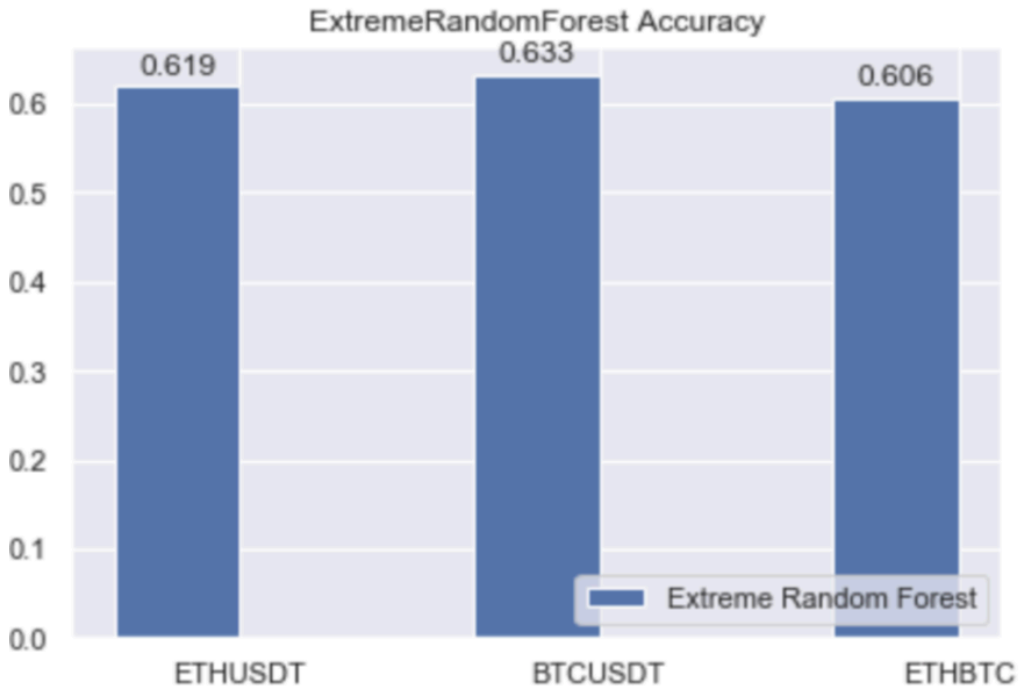
Naïve Approach and Random Forest

- A naïve approach is just making a guessing either the price will go up or down, like flip a coin. Also we can guess based on the previous day's moving direction. Since price is time series data, guessing from yesterday makes sense. Therefore the accuracy is ~50%.
- **Random Forest**
- An ensemble model, bagging method of multiple decision trees
- **Pros:** easy and quick to train, can handle large datasets with high dimensionality, feature selection
- **Cons:** does not give continuous accurate prediction for regression problems, may overfit the data when it is noisy; Little interpretability – "black box" approach
- I will use Random Forest as benchmark model for price moving direction prediction and output feature significance



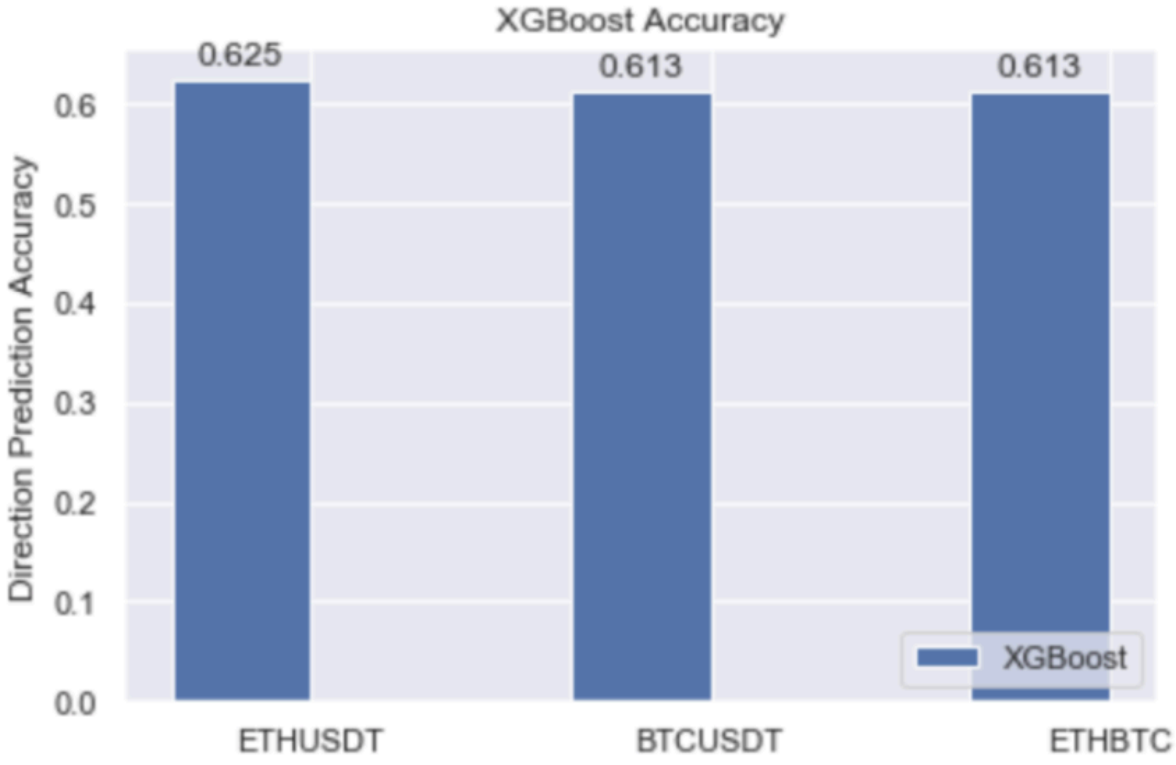
Extreme Random Forest

- **Random forest:** splits by computing locally optimal features
- **Extreme randomized trees:** more randomness
  - Draw candidate features randomly to set threshold
  - Pick the best to serve as the splitting rule
- Reduce variance but increase bias slightly
- More diversified trees and less splitter: Train faster, reduce time
- **Result:** Overall directional prediction accuracy is ~61%
- High performance in presence of noisy features
  - When all the variables are relevant, both methods seem to achieve the same performance



XGBoost (Extreme Gradient Boosting)

- **XGBoost** is based on the implementation of the Gradient Boosting method (reduce both variance and bias), but with more accurate approximations to find the best tree model.
- **Results:**Two main reasons I choose XGBoost:
  - Execution Speed: XGBoost training is very fast due to its block structure for parallel learning.
  - Model Performance: XGBoost could penalize complex models through both L1 and L2 regularization which improves model generalization to prevent overfitting.



The directional prediction accuracy for XGBoost is about 61%.

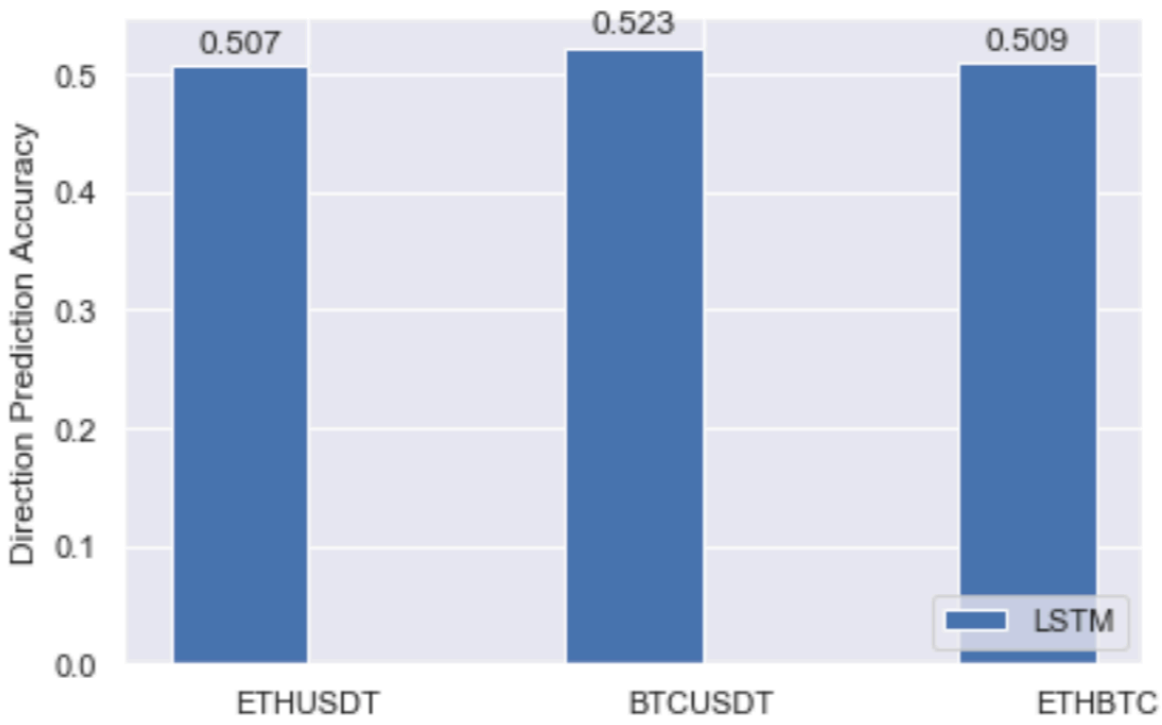
Recurrent Neural Network - LSTM

- LSTM is an ideal algorithm of RNN to deal with sequence data.
- It bridges long-time intervals without losing short-time lag capabilities by replacing the simple neurons with complex LSTM cells.
- LSTM is considered an ideal algorithm for modeling my huge time series dataset due to its efficiency and capability. But compared to the previous tree-based models, LSTM underperformed in my case.
- Potential reasons:

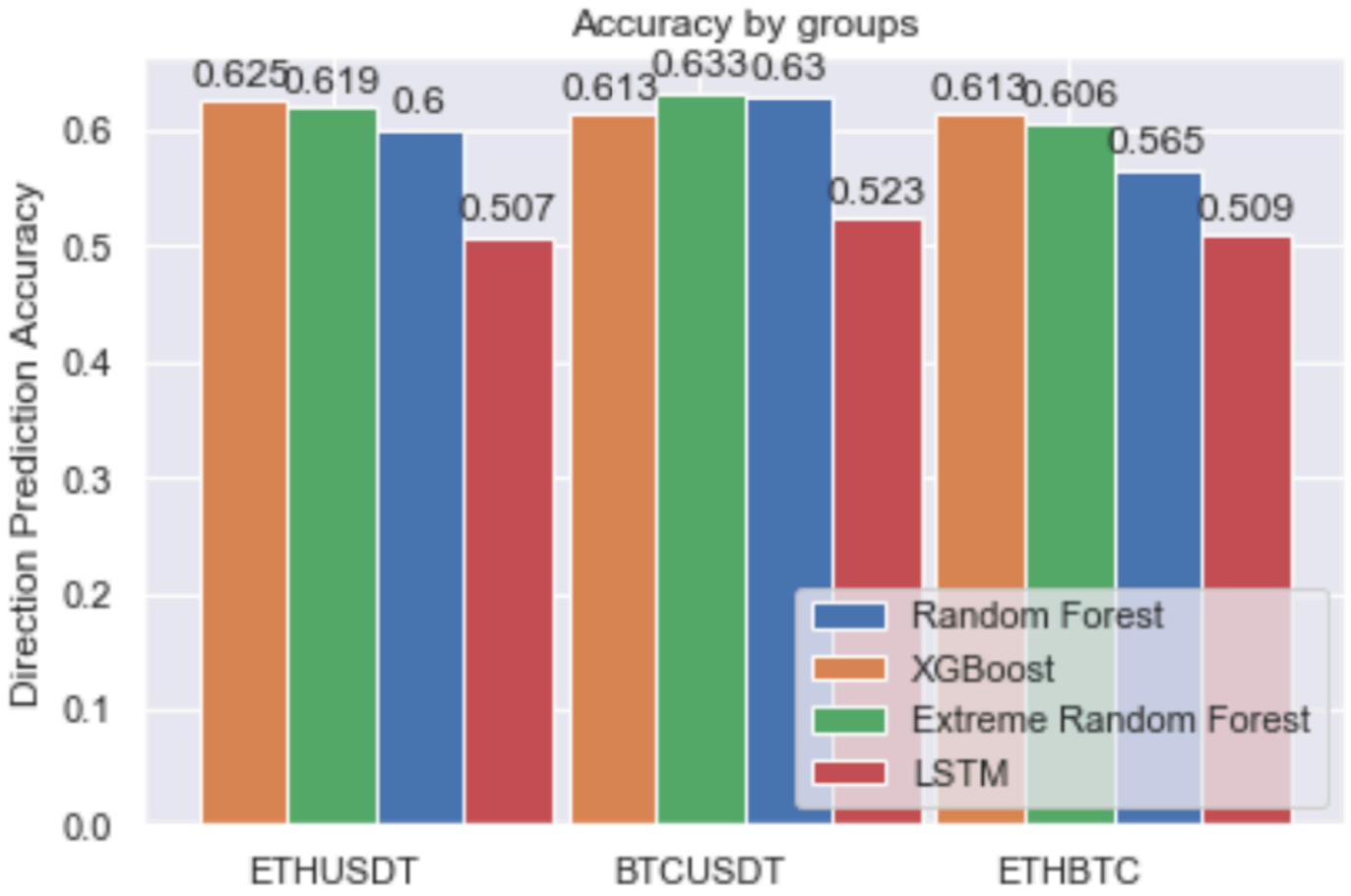
  1. LSTM is **computationally heavy**, so I didn't use the full data for training the model because of limited computer powers, Neural Network usually benefits from using a huge amount of data.

  2. The model is **relatively simple** in the sense that we cannot perform exhaustive hyperparameter tuning or add more layers to the model.

  3. My data is very **noisy** – it is hard to predict in nature, plus we only use a small sample of it to train the model.



Model Comparison



- Combining all the models I found the 3 tree models outperformed LSTM and with an almost even accuracy of ~61% and XGBoost and Extreme Random Forest are slightly better.

Conclusion

- From the analysis of the three cryptocurrency pairs, I found that there is still predictability in the cryptocurrency market. The project has demonstrated that high frequency methods can be used to achieve successful direction prediction of future out-of-sample price trend movement at a ~60% probability. Especially in the optimized model of XGBoost for the BTC/USDT pair, a direction prediction accuracy of 63.3% is achieved. Compared to a 50% success rate resulting from randomly guessing the direction, a ~60% success rate is definitely significant enough to give us an edge.
- However, whether it is possible to make profit through trading using this direction predictability I have obtained is another question that requires further extensive research. The model is capable for predicting the general future trend of price movement but does not accurately indicate how much the trend will move in the direction. At the same time, high frequency trading is usually done in large volumes, given a ~60% accuracy, the upside profit of a trading strategy might be limited when the prediction is correct, while the downside loss might be substantial when the prediction is wrong. In addition, the transaction cost of trading needs also to be considered when seeking profit.

Conclusion

- Besides seeking direct profit from trading on this predictability, the model results can also be used as an assistive tool for market makers to understand the current market situation. By having a little edge in where the market is moving in the next few minutes or second, market makers can prepare and time their trades better and make necessary adjustments in their books before extremely abrupt changes. A direction predictability of ~60% is definitely superior than predictability in the traditional equity market. The results can also be used as a marketing point for including cryptocurrency pairs in investment portfolios.
- If more high frequency cryptocurrency data is available, the model can be expanded to a more granular scale that involve more cryptocurrencies and exchanges with different time periods.