Kendall Smith

UW Data Science 410

Final Project Write-Up

**Summary of Findings:**

My initial hypothesis was that a change in tweet volume would correspond with a change in the price of a given cryptocurrency, although the latency/lag time between tweets and price action would be unknown. Therefore, the first linear regression attempted was between the count of tweets with Bitcoin related hashtags, and the closing price of the Bitcoin for each moving period (hourly in the dataset downloaded). This resulted in a low adjusted R^2 value of 0.057 and an AIC of -193.7. Obviously, the price of an asset can either go up or down, irrespective of the volume of trades or tweets. This lead me to believe perhaps I should compare the absolute price change to the volume of tweets. This however, resulted in an R^2 of -0.005 and AIC of -183.1.

The second linear regression attempt used multiple price features as the independent variables, and the volume of Bitcoin related tweets as the dependent variable (opposite of the first approach). Using a multitude of independent variables (with a high level of collinearity) resulted in worse model performance, adjusted R^2 of 0.032 and AIC of 2139. Lastly, I believed that if feature selection was used to fit the linear model on the top 2 or 3 variables that the model performance would get better. However, contrary to expectation using the chi2 test with the SelectKBest method from sklearn preprocessing library, the library performance got even slightly worse. This time the R^2 was 0.006 with an AIC of 2138. Due to such low performance, even with scaled data and selected features, my conclusion is that a linear model is not appropriate for solving this problem.

**Future Directions of Analysis:**

1. It became apparent early on that I needed larger dataset of tweets from my streamer script, both by length of time recording tweeps to align with historical pricing data, and in the breadth of scope in keywords used to catch tweets in the stream listener.
2. If I had more time and previously training on NLP techniques, I would have wanted to develop a sentiment analysis for Bull or Bear Market indication in each tweet using text analytics and using some form of Bayesian filtering to remove spam tweets generated by bots.
3. Within the limitations of linear regression model, I could have also applied Elastic Net Regression to employ Lasso and/or Ridge regression in weight regularization.
4. Perhaps I could have experimented with more complex feature selection process, such as forward variable selection, PCA, etc. or using Random Forest to weigh feature importance.
5. Another area of research I wanted to explore but did not have time was for back and forward propagation of time series data in Bitcoin prices to identify some kind of lead/lag time in the price action corresponding to twitter dialogue. In order to do this, I wanted to graph the pricing data on the hourly level for each day, but ran into difficulties using multiindex for the x-axis range. A visual check on lead/lag between the variables might have helped to improve training.
6. Lastly, linear regression did not seem appropriate for this type of data, so perhaps I could experiment using a different regression model, possibly logistic or Gradient Boosting Machine.