Crypto Prediction with Alternative Data

Data

For this project, we used crypto data from Kaggle. The data set has 48 features and 8,617 rows. However after removing the rows with missing values, we only have 2,391 rows left. We also removed columns that contained id's and lagged values. Therefore, we have 40 features left. All of the features are numerical. Some examples of the features are:

- Related volume of posts from Reddit
- Related volume of comments from Reddit
- Volume of tweets
- Related retweets
- Bulling sentiments extracted from relevant twitter conversations
- Bearish sentiments extracted from relevant twitter conversations

The response variable:

- Bitcoin stock price

Model fitting and Analysis

Part I: R-Square, Residuals and Cross-Validation Curves

We used 10-fold cross-validation to tune the lambda values and used the minimum lambda value to fit lasso, ridge, and elastic net on the training data. We also fitted random forest. This was repeated on 100 samples and training & test R^2, training & test residuals were computed for each sample and stored in a data frame. The boxplots of R^2 and residuals are given below.

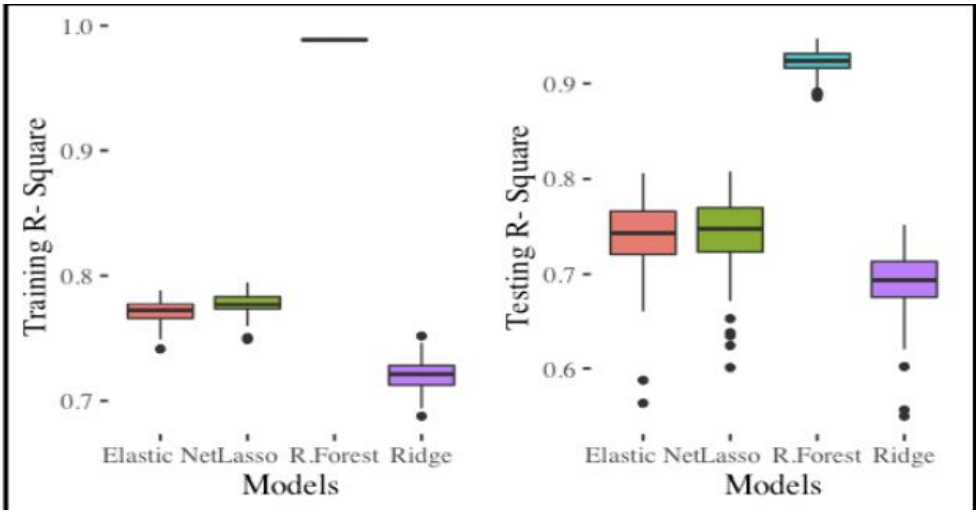

Figure 1: All four models performed similarly on the training and test data. But test R^2 values vary more compared to the training R^2 values. Performance of elastic net and lasso is very similar to one another. Ridge regression performed slightly poorer than both, but above 70% for R^2 is still decent. Random forest performs the best as it can account of more than 90% of the variation in the data.

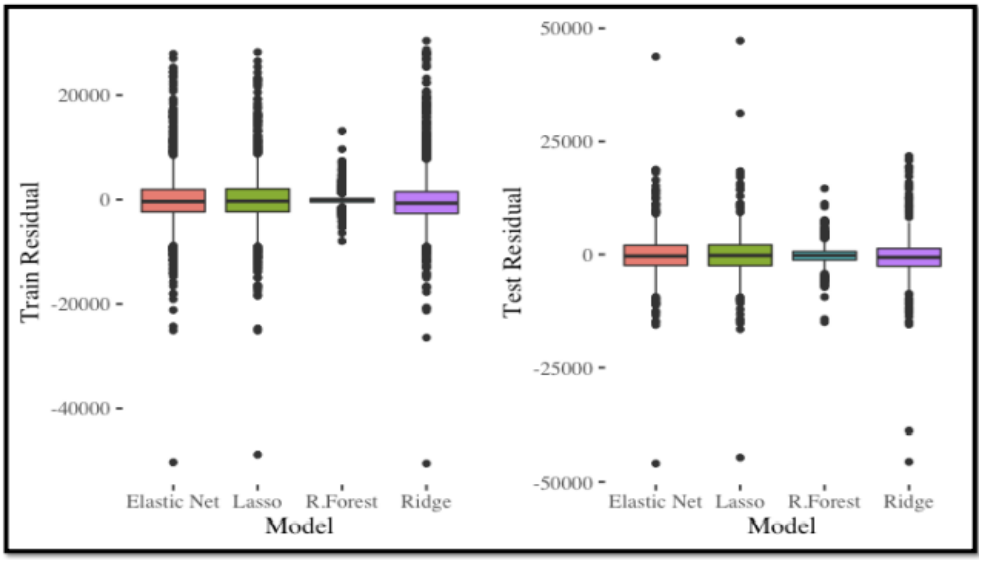

Figure 2: Both the train and test residuals are very similar for all four models. We also notice some outliers.

Then we plotted the cross-validation curves of elastic net, lasso, and ridge regression for the 100th sample. They are given below.

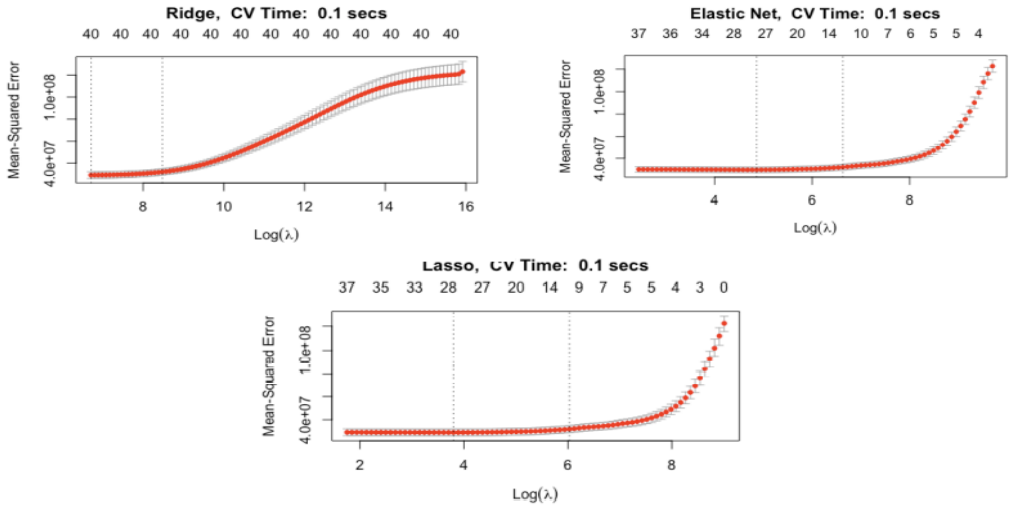

Figure 3: Ridge gives us the lowest MSE when log(λ) is less than 8. As for elastic net, it gives the lowest MSE for log(λ) values between about 5 and 6.5 (13-27 features). And lasso gives the lowest MSE when log(λ) is between 3.8 and 6 (11-28 features).

Part II: Variable Importance

We still fitted the same four models but this time we didn't divide the data into test and training sets and there was only one sample. In this section we investigate the variable importance.

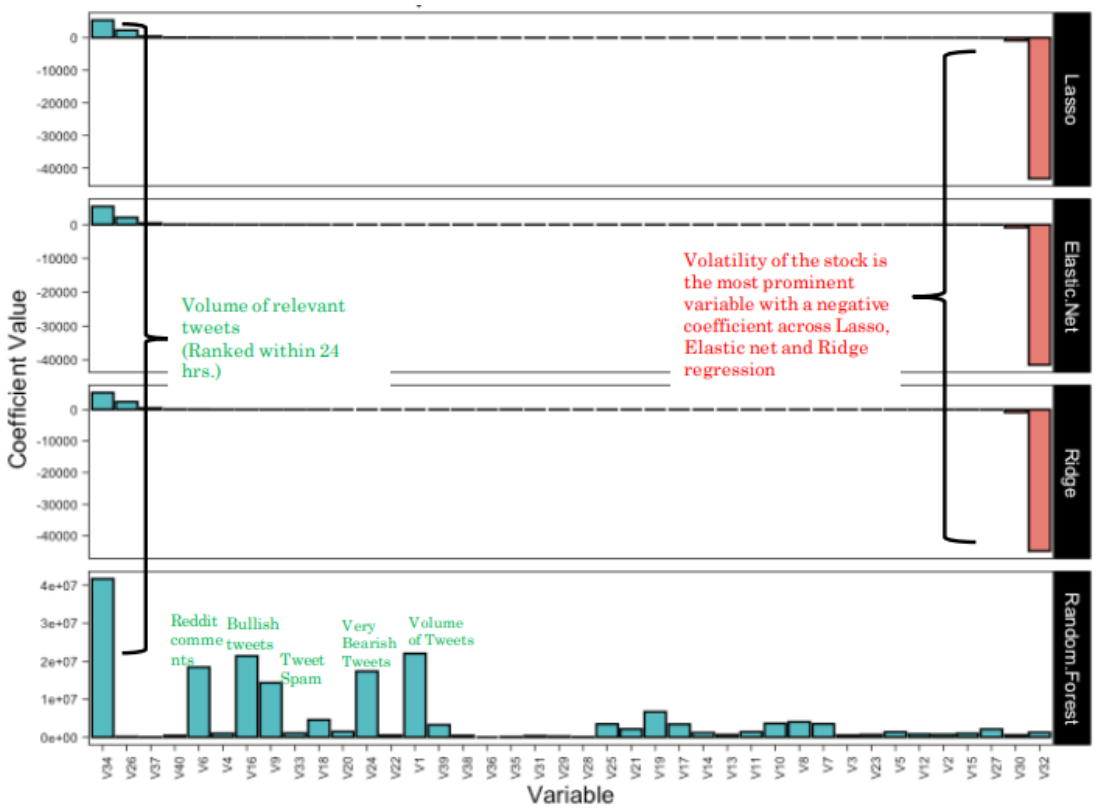

Figure 4: For all four models, "Volume of relevant tweets" is important. For lasso, ridge, and elastic net, "Volatility of the stock" is most prominent variable with negative coefficient. Random forest has several important variables including "Reddit comments", "Very bearish tweets", etc.

Part III: Accuracy Vs. Time

We compute 90% confidence interval of test R^2 using test r-squares from part I for 100 samples. We also use the time it took to cross-validate and/or fit the models once, from part II. The table containing test R^2confidence intervals and the time is given below.

Model	Interval	Time
Ridge	0.635-0.732	0.2 secs
Lasso	0.671 - 0.8	0.2 secs
Elastic Net	0.665 - 0.794	0.2 secs
RF	0.895 - 0.94	11.8 secs

Figure 5: It takes 0.2 seconds to cross-validate and fit lasso, ridge, and elastic net. But it takes 11.8 seconds to fit random forest, which performs much better than the other three.

Conclusion

In this project, we predict “Bitcoin stock prices” using alternative data. We fitted models such as random forest, elastic net, lasso, and ridge regression. Random forest outperforms the other three but there is a time trade-off.