Crypto Prediction with Alternative Data

Group 18: Rupinder Kaur & Janani Ravichandran

Data Summary

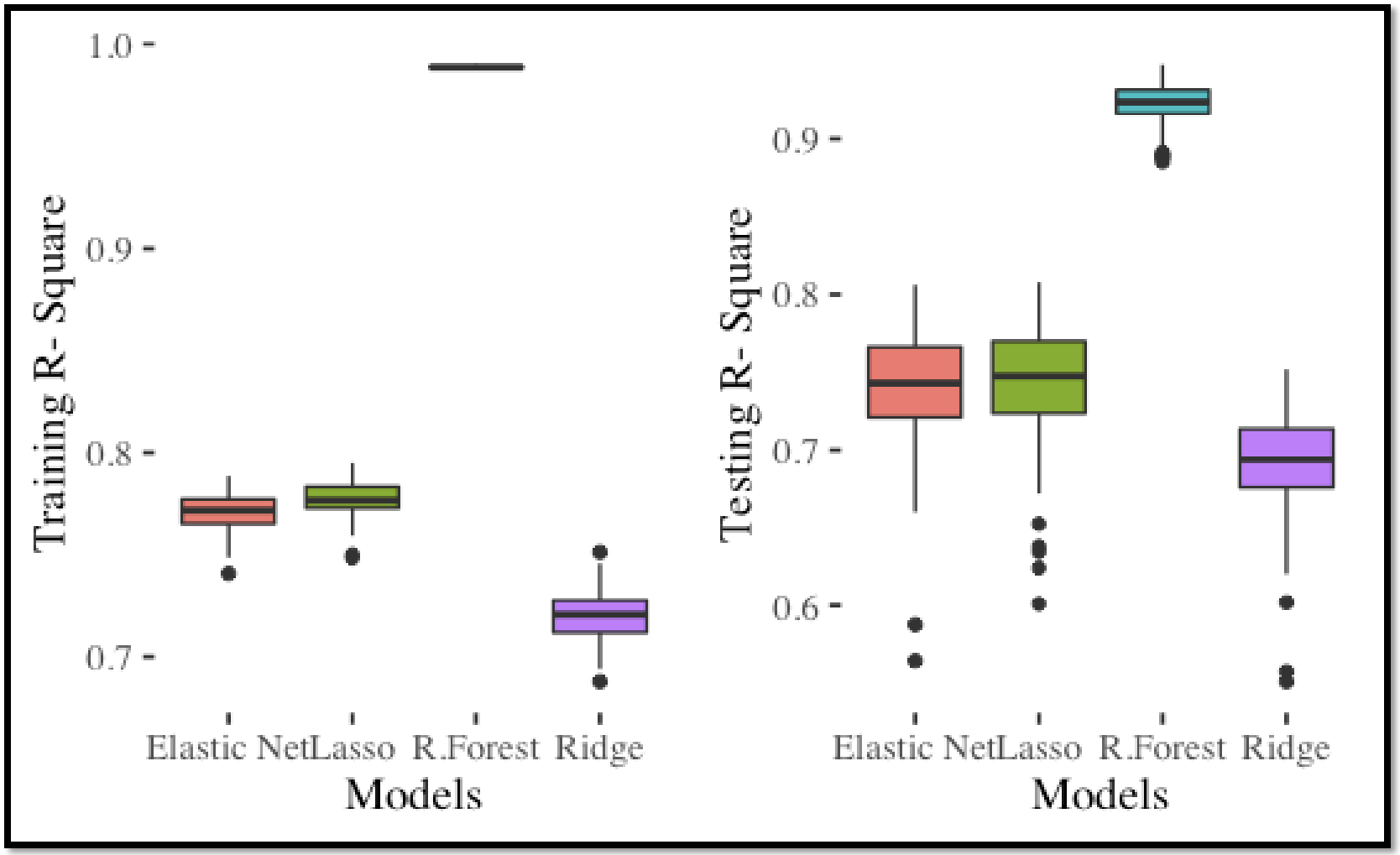
- Original Data: n = 8,617 and p = 48
- Cleaned Data: n = 2,391 and p = 40
- **Variables**:
  - All features are numerical

  - Response Variable: Bitcoin stock price
    - Mean: $14,652 , Min = $5117, Max = $60,961, Std Dev = $10,785

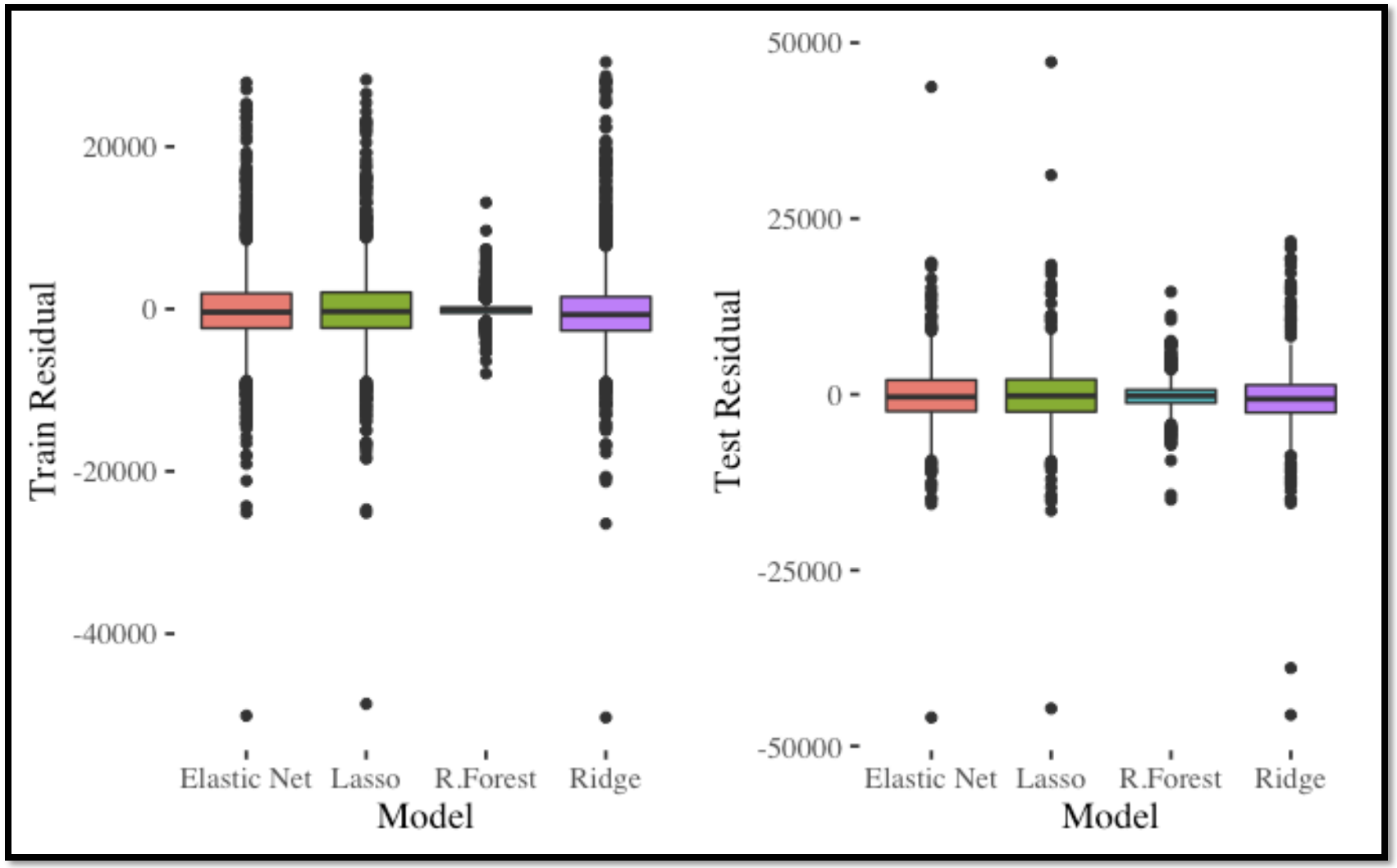
  - Input Variable: Various cryptocurrency data from social media such as Twitter, Reddit etc.
    - Some Examples of input variables:
      - Related volume of posts from Reddit
      - Related volume of comments from Reddit
      - Volume of tweets (from related keywords/topics)
      - Related retweets in Twitter
      - Bulling sentiments extracted from relevant twitter conversations
      - Bearish sentiments extracted from relevant twitter conversations
      - Rank based variables that compare the volume of relevant posts/comments/tweets to total volume(proxy for trending variables)

Data Source: https://www.kaggle.com/rafat97/cryptocurrency-price-prediction-by-ieee-ensi-sb

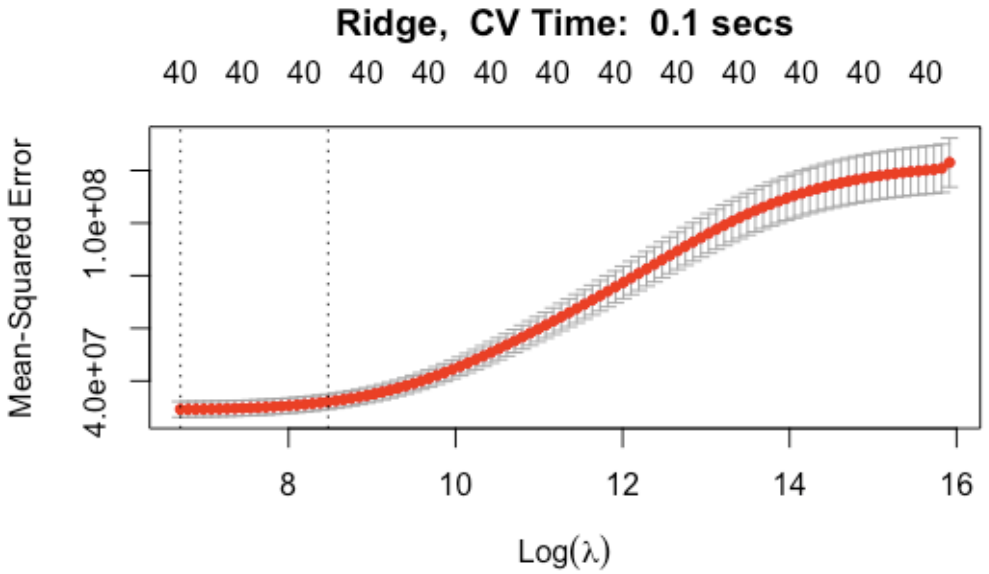
Comparing R-squared Values Across Models

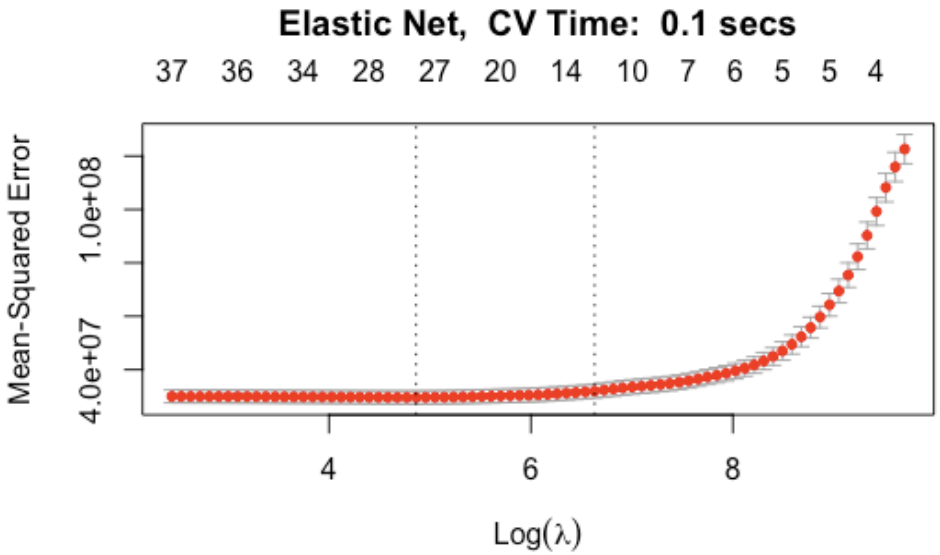


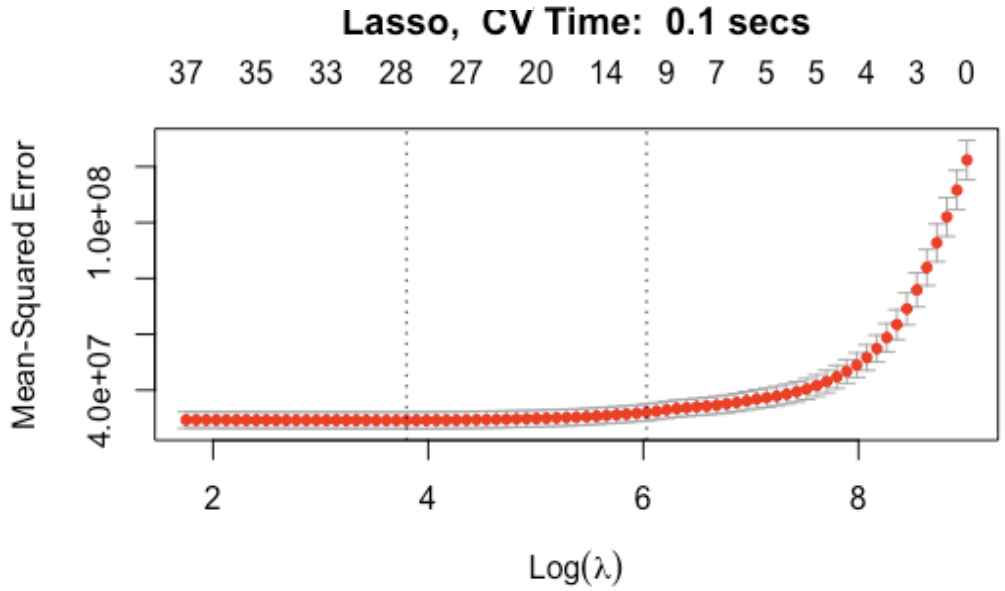
Comparing Residuals Across Models



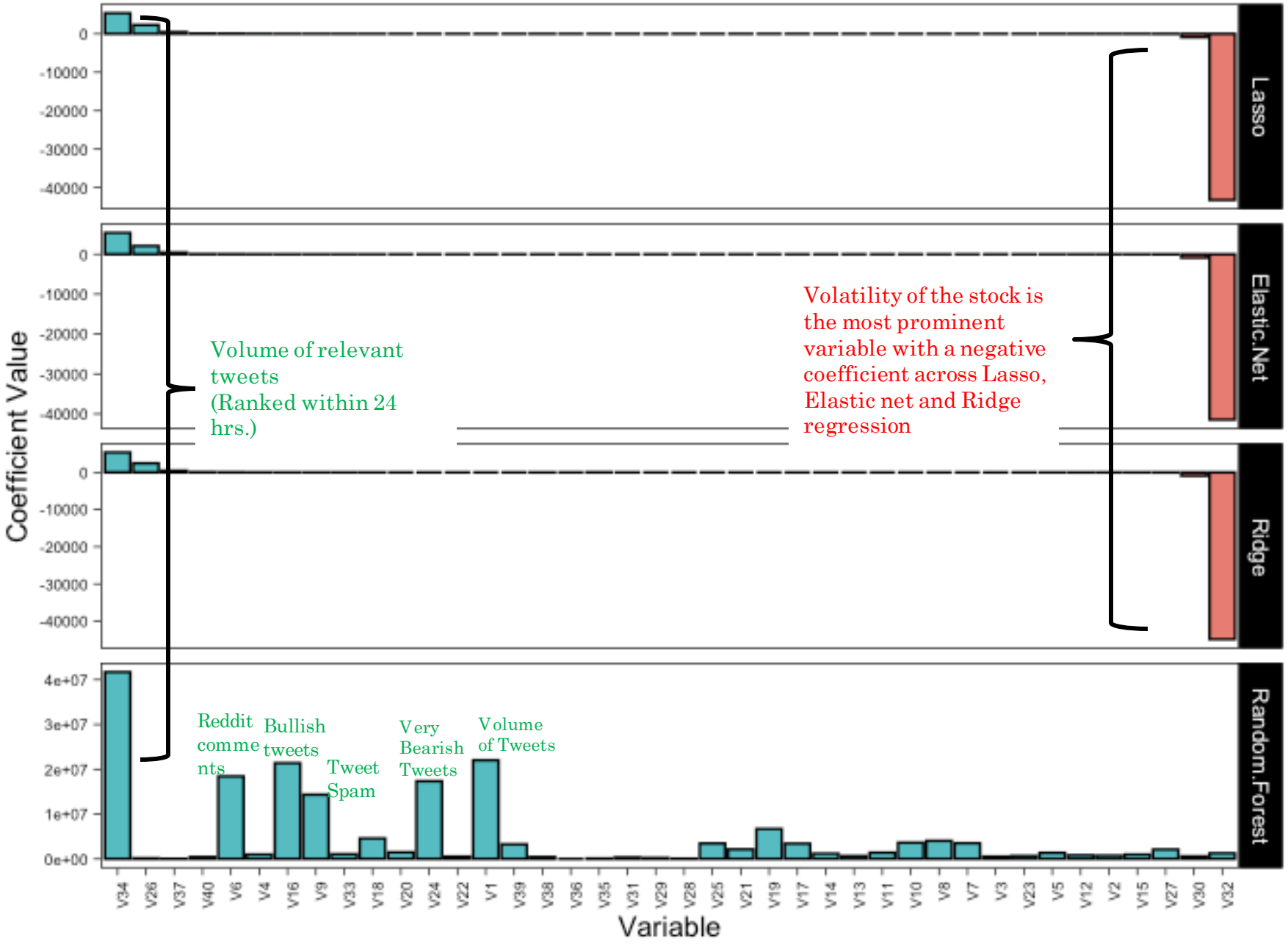
Cross Validation Curves





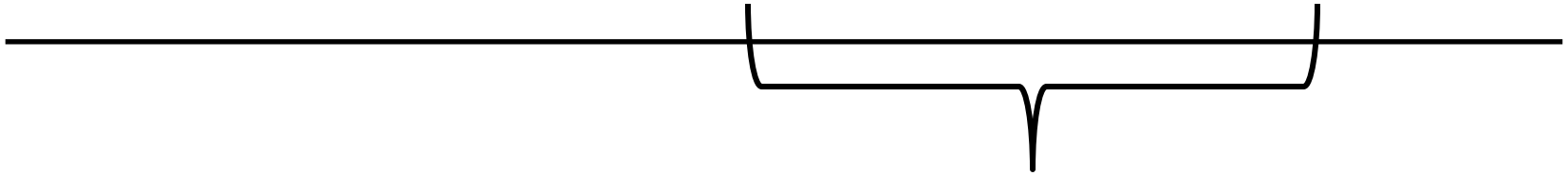


Variable Importance



90% R-squared Interval and Running Time

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



Is the accuracy improvement worth the increased time?

Conclusion

Random Forest accuracy outperforms the rest of the models. However, there is a significant time trade-off. Increase in the data size in the future might further highlight this trade-off.

The other three models have comparable time. Lasso has the highest accuracy interval.

This model tends to agree with the hypothesis of social media's impact on crypto. Across all the models, the ranked Twitter volume is one of the most important predictors.

The current dataset is also constrained by the listed variables. The accuracy of the model can be further tested with the additional media variables.