Stock Trading: Prediction of Auction Volumes

Emma Fannin, Anthony Ramirez, Bailey Schepke, Donovan Pinto
CSC 3220 – Dr. William Eberle

Abstract—With data analyzation methods comes the ability to make predictions about the future behavior of numerous types of data. This project seeks to analyze stock market closing auctions. Stock exchanges hold closing auctions at the end of a trading day to determine the closing price for each stock. We analyzed volume, the total value of stock exchanged, during the closing auctions from stock exchanges covering a set of 900 stocks over 800 days. In order to complete this project, we built a model to predict stock trade volumes utilizing data provided on challengedata.ens.fr. A benchmark was created by Capital Fund Management, the providers of the challenge, that utilizes a linear fit method for our model to be compared against. We have interpreted and cleaned the data in order to predict exchanged volume of a given stock during closing auctions.

Problem

At the end of a trading day, stock exchanges hold closing auctions to determine the closing price for each stock.

Predicting the trading volume of these auctions would be useful to optimize trades during the auctions as well as the trader's after-hours position.

Goals

The primary goal of this project is to reasonably estimate closing auction volume for a given stock and day.

Past data can be used to project and estimate future data, and to do this we construct a model.

Tools Used

R – used for data manipulation, visualization, and writing scripts.

Provided Input Data

- **id:** each datapoint has an id, unique for a stock on an individual day
- **pid:** the product ID of a stock
- **day:** day of the sample
- **abs_retn (n = 0 to 60):** absolute values of stock returns over the day separated into 61 periods
- **rel_voln (n = 0 to 60):** fraction of volume traded over the day separated into 61 periods (totals to 1 each day)
- **LS and NLV:** mysterious variables whose purpose is not disclosed in the challenge

Provided Output Data

- **id:** matches id in the input data
- **target:** natural log of the auction volume as a total volume in the 61 periods
  - If the auction volume is 10% of the day's volume, target = log(0.10) = -2.30

Approach

Before anything else, the data was cleaned as the challenge website suggested, replacing N/A values with zero.

In order to predict the target value, we need to determine which variables influence it.

We use gradient tree boosting to find the relative influence of each variable in the input data, as seen in **Figure 1.1.**

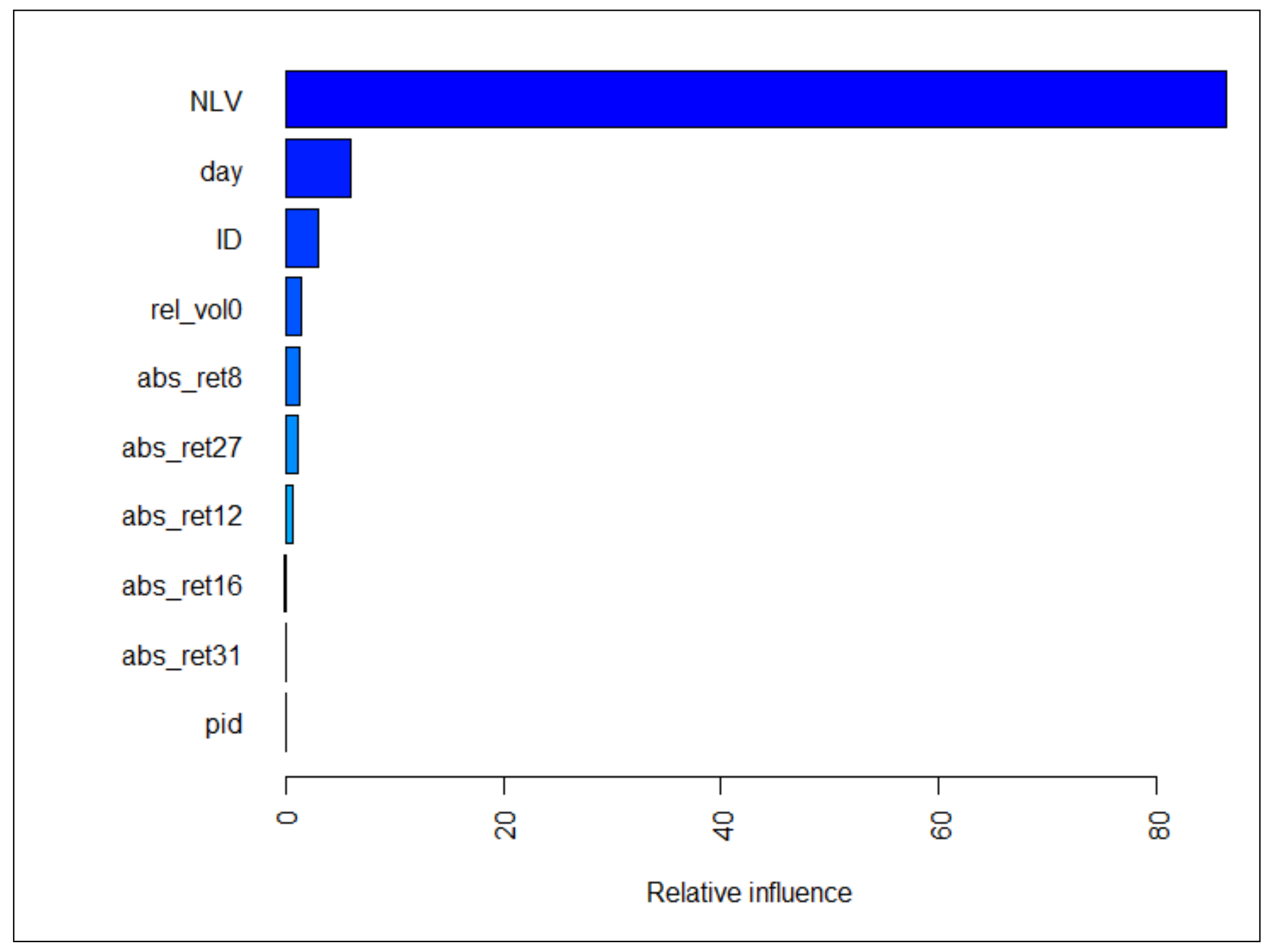

Figure 1.1 - Relative Influence

Model

The gradient tree boosting result shows us that the target is most subject to the mysterious NLV variable.

We create a simple linear regression model based on NLV.

A model was generated for each of the 900 pid's.

Shown below in **Figure 1.2,** the data represented in a scatterplot and our model for pid 69 represented as a red line.

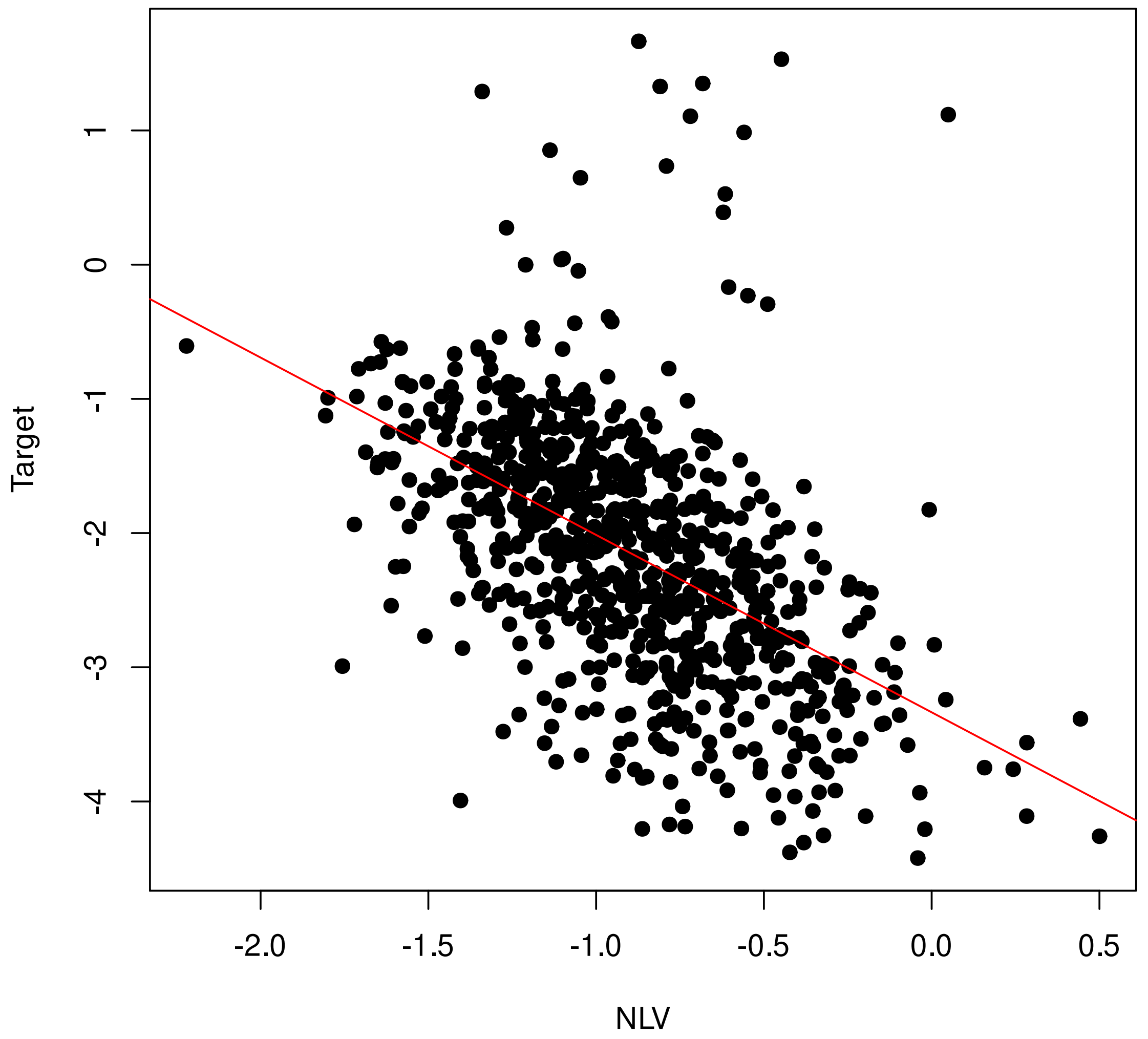

Figure 1.2 - Model for PID 69

Results

The model summary lists NLV as a significant coefficient with a standard error is 3.263e-3, t-value of -78.539, and p-value of <2e-16. The low p-value is a good sign the NLV is significant, though the t-value being so far from zero could indicate otherwise.

Further summary data includes:

- Residual Standard Error: 1.334 on 684479 degrees of freedom
- Multiple R-squared: 0.6182
- Adjusted R-squared: 0.6182
- F-statistic: 3.694e+5 on 3 and 684479 degrees of freedom, p-value < 2.2e-16

When submitted to the Challenge Data website, our output data was given a score of 1.5130.

With lower scores being better, the benchmark model was given a score of 0.4742.

Conclusions

Separating the data by pid worked well, as most stocks behaved differently.

We used gradient tree boosting like the benchmark did, but rather than simply using it for classifying variables, the benchmark's model is also based on gradient tree boosting using all columns.

It was unclear to us as to why rel_vol seemed to have little effect on target when volume was what we were trying to predict.

Our data so far seems to support the notion that NLV is the factor that influences stock value the most and that using it would be the best way for someone in the market to estimate the volume of stock traded.

Future Work Recommendations

For future work regarding this challenge, we would like to see this data be reexamined using analysis methods such as time series in order to gain better views on the data as well as perhaps using it to gain better understanding of why NLV is the best predictor.

We would also like to gain better understanding of the absolute values of stock returns (abs_retn) and the traded stock volume as a fraction (rel_voln). These were significant parts of the data, and we feel like their relevancy could be much clearer with further research.

Our challenge score was not as close to the benchmark as we would like, and with further research as described above, we believe the score could be improved.

References

1. B. Boehmke and B. M. Greenwell, "Chapter 12 - Gradient Boosting," in *Hands-On Machine Learning with R*, CRC Press, 2020.
2. C. Shah, *A Hands-On Introduction to Data Science*. New York, New York: Cambridge University Press, 2020.
3. CFM, "Stock Trading: Prediction of Auction Volumes," *Challenge Data*, 04-Jan-2021. [Online]. Available: https://challengedata.ens.fr/challenges/60. [Accessed: 10-Mar-2021].
4. R. Nau, "What to Look for in Regression Model Output," *Statistical Forecasting: Notes on Regression and Time Series Analysis*. [Online]. Available: https://people.duke.edu/~rnau/411regou.htm. [Accessed: 12-April-2021].
5. Suicasmo. (2017) New York Stock Exchange [JPEG]. https://commons.wikimedia.org/wiki/File:New_York_Stock_Exchange_20170311.jpg

We would like to extend acknowledgements to CFM for providing the challenge as well as our advisor, Dr. William Eberle.