Greenwich HR Project

Everything Starts with Data - MSiA 400

Jake Atlas | Srividya Ganapathi | Daniel Halteh | Adi Tyagi

"Bryce Canyon"

Introduction

The analysis conducted in this project serves as an important subsection of a collaborative effort amongst 4 groups of Northwestern University Master of Science in Analytics (MSiA) students. The overall goal in this multi-team MSiA investigation is to determine certain preliminary details regarding the viability of using job posting information for selecting stocks that will outperform the market. The analysis presented in this report pertains to 2 of the overarching project goals: (1) finding companies for which there is an association between job postings and stock outperformance and (2) building a machine learning model whose output can be used to make predictions of stock price and stock outperformance. Both of these objectives have been pursued with the goal of making estimations associated with stock price 90 days following job postings. This report will first investigate objective (1), hereafter referred to as "Association Analysis," and will then discuss objective (2), hereafter referred to as "Machine Learning Modeling."

Association Analysis

Summary

The given instructions to "find the correlation between job postings and stock outperformance 90 days following" immediately presented some fundamental issues, particularly pertaining to the use of the word "correlation." As statisticians, the team found it unusual to request correlation between a value that is binary – whether a job was posted or not – and stock outperformance – possibly a binary value, and possibly a vector. Regardless, a statistical correlation cannot be computed and analyzed effectively here. Consequently, two methods of measuring association were developed, Method A and Method B. Each of these methods will be explored in detail in the following Methodology section. An important note is that both methods seek to account for market movement so as to eliminate the impact of general market-wide factors that play a role in driving stock prices.

Methodology

Method A

This method of computing a numerical association between company job postings and that company's likelihood of stock outperformance after 90 days does not leverage the concept of statistical correlation. The following process was used to score companies:

Define for each job $j \in J$ posted by each company $c \in C$ four vectors:

1. $x_{j,c}^{*stock}$ of 90-day percent changes in stock price for each of the 90 days following the posting of job j. Then each x_{j,c_i}^{*stock} for $i_c \in I_C$ is the percent change in stock price on day $i+90$ as compared to day i. The first instance of a job posting by a company in 2018 or later corresponds to $i_c = 0$ and the most recent job posting by that company corresponds to (i_c). This is therefore a vector that provides information on how the stock performed each day following the job posting (for 90 days).
2. $x_{j,c}^{*market}$ is the same as $x_{j,c}^{*stock}$ except it contains entries corresponding to percent change in a market composite index. Here, each daily percent change is the average of the

percent change in the S&P 500 market index and the NASDAQ Composite market index. This is therefore a vector that provides information on how the market performed each day following the job posting (for 90 days).

3. $x^{0_{stock}}_{j,c}$ is the same as $x^{*_{stock}}_{j,c}$ except it contains entries that correspond to the 90 days leading up to the job posting. Each entry is still a percent change over 90 days. This is therefore a vector that provides information regarding the baseline for the stock performance, which can be compared to the performance following the job posting (detailed by $x^{*_{stock}}_{j,c}$).
4. $x^{0_{market}}_{j,c}$ is the same as $x^{0_{stock}}_{j,c}$ except it contains entries that correspond to the market index detailed in (2). This is therefore a vector that provides information regarding the market performance before the job posting.

Then for each company $c \in C$, define vector m_c that contains j elements m_{c_j} for jobs $j \in J$. Each element $m_{c_j} = (\overline{x^{*_{stock}}_{j,c}} - \overline{x^{0_{stock}}_{j,c}}) - (\overline{x^{*_{market}}_{j,c}} - \overline{x^{0_{market}}_{j,c}})$. Then the final metric calculated for each company $c \in C$ is the average of each element m_{c_j} in m_c, $\overline{m_c}$.

The companies are ranked according to the metric $\overline{m_c}$. The higher the magnitude of the metric, the stronger the association between job posting and stock performance. Highly positive metrics indicate that job postings yield stock outperformance after 90 days, while highly negative metrics indicate that job postings yield stock underperformance after 90 days, as compared to the market.

Method B

Unlike Method A, this method of computing a numerical association between company job postings and that company's likelihood of stock outperformance after 90 days does leverage the concept of statistical correlation. The following process was used to score companies:

Define for each company $c \in C$ a vector L_{1_c} with $n_c \in N_C$ elements corresponding to the number of jobs posted each day. $n_c = 0$ corresponds to the first day in 2018 or later for which a job was posted and (n_c) corresponds to the most recent day for which we have job posting data for company c.

Define also for each company $c \in C$ a vector L_{2_c} with elements corresponding to those in L_{1_c}; each element in L_{2_c} is the market-adjusted percent change in the company's stock closing price 90 days from the day of the corresponding job posting in L_{1_c}. This is computed by:

$$\left(\frac{\text{\% change in the stock's closing price}}{\text{90 days after the job posting}}\right) - \left(\frac{\text{average \% change in the market indices' closing prices 90 days after the job posting}}{}\right)$$

Note that the average percent change in the market indices' closing prices is calculated in the manner of $x^{*_{market}}_{j,c}$ from Method A, by averaging the percent changes across the S&P 500 and the NASDAQ Composite. Then for each company $c \in C$, the correlation of vectors L_{1_c} and L_{2_c}, $cor(L_{1_c}, L_{2_c})$, can be computed and taken as a measure of the degree and direction of association between job postings and stock performance relative to the market, 90 days following the job posting.

The companies are then ranked by correlation value. Highly positive correlations indicate that more job posting yields stock outperformance, while highly negative correlations indicate that more job

postings yields stock underperformance. The p-value of the correlation for each company can be used to select only those companies with statistically significant correlations.

Analysis

It was found that Method A and Method B somewhat corroborate each other, but not entirely, as expected. There is less bias towards low-price, high-fluctuation stocks in Method B, whereas Method A tends to compute high magnitude metric values for penny stocks due to the fact that small changes in low-priced stocks represent very large percent changes. Across the first 100 values of each of the lists, there were 9 shared tickers. In consideration of the fact that Greenwich HR is seeking a small subset of around 10 companies that have strong association between job postings and stock outperformance, this list of 9 appears to successfully satisfy intuition and the project goals. However, given the clear variation across these two methods, it is suggested that Greenwich HR consult both of the lists created before investing. Extra caution is urged when considering investment in penny stocks. See Appendix A for visualizations from this segment of the analysis and further insights regarding the viability of using this approach for stock picking.

Conclusions and Next Steps

The team is confident that the request for "correlation" analysis has been satisfied through the association analysis that has been done. Using each of the output lists from Method A and Method B allows for validated selections of companies in which to invest; though it is important to consider many other factors when making investment decisions. This is especially true in this case, as it is unclear at this stage whether there is valuable information that would actually lead to consistent financial gains. Having completed this stage of the project successfully, the next step that the team recommends is to create simulated portfolios of stocks contained in the lists derived from Method A and Method B detailed above, which other MSiA teams are investigating. The simulated performances of these test portfolios can be used and compared to each other to develop a more concrete assessment of the best way to practically use the results of the analysis conducted. Finally, it is important to consider the fact that all analysis done here utilizes a 90-day window for assessing the performance of a stock. Other MSiA teams have conducted similar analyses with other time windows, and it is important to investigate whether there is a particular window that is optimal.

Machine Learning Modeling

Dataset Overview

Each observation in our dataset represented a 'single job posting occurrence'. The following variables associated with a job posting were recorded:
Covariates

- stock ticker of the company that posted the job (ticker)
- the job post date (post_date)
- the number of other jobs posted by the company that day (njobs_on_date)
- the total number of jobs posted by the company during the time horizon of the entire dataset (njobs)
- the salary associated with the job posted (salary)
- the stock market closing price for that day (close)
- the look back 90 day percent change in the company' stock price (i.e. how much has stock price changed since 90 days ago?) (prev_90_day_pct_change)
- the look back 90 day percent change in the market as defined by the S&P500 index value 90 days ago (pct_change_90_market)

Dependent Variable:

- the change in the company's stock price in the next 90 days (pct_change_90_stock)

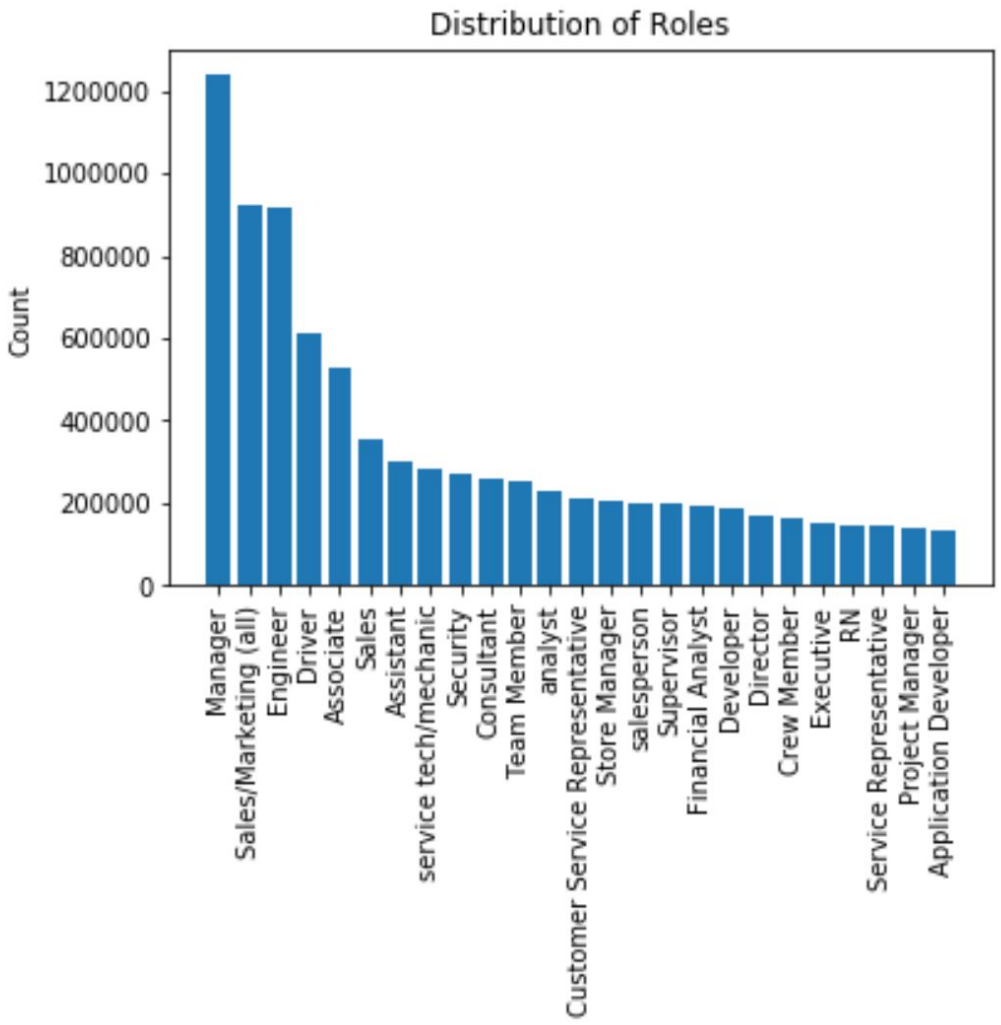
Approach High Level Overview

We divided our approach into two phases:

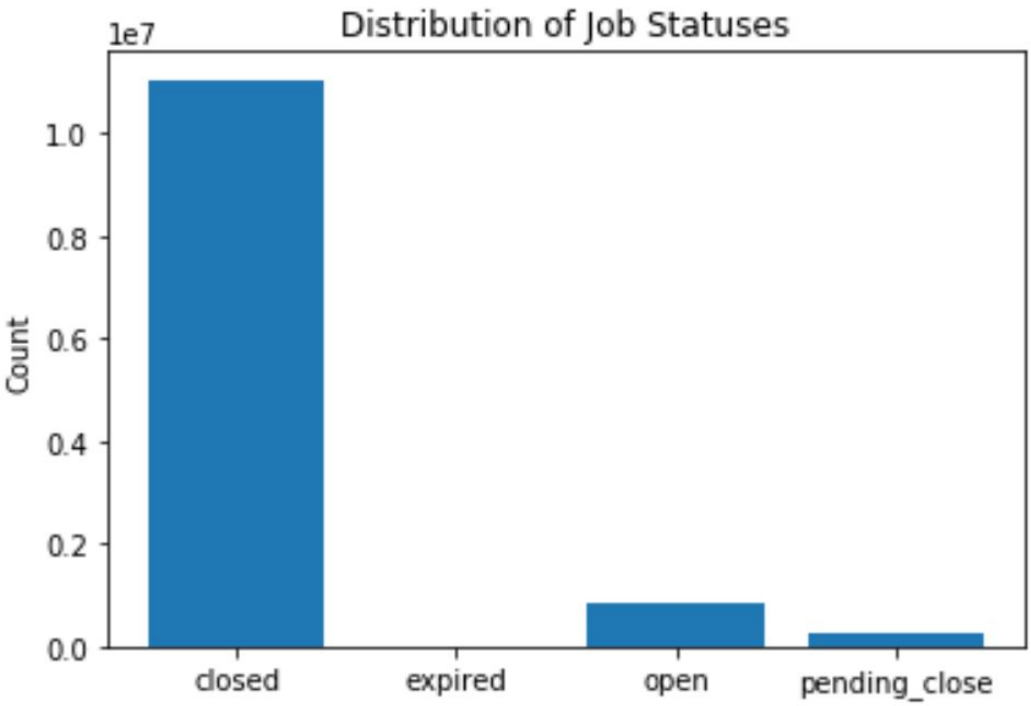
I. Exploratory Data Analysis
II. Data Preparation
III. Analytical Modelling

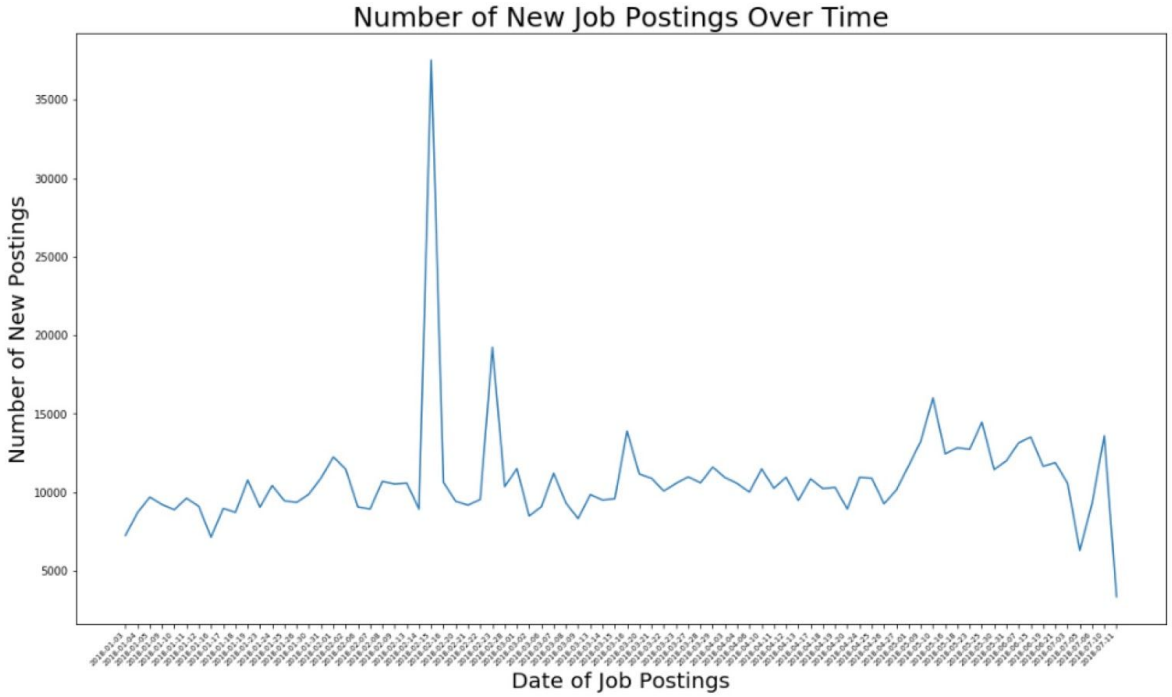
Phase I: Exploratory Data Analysis

During the EDA phase, our goal was to explore general trends and patterns in the data to get an idea of what modelling approaches to use.

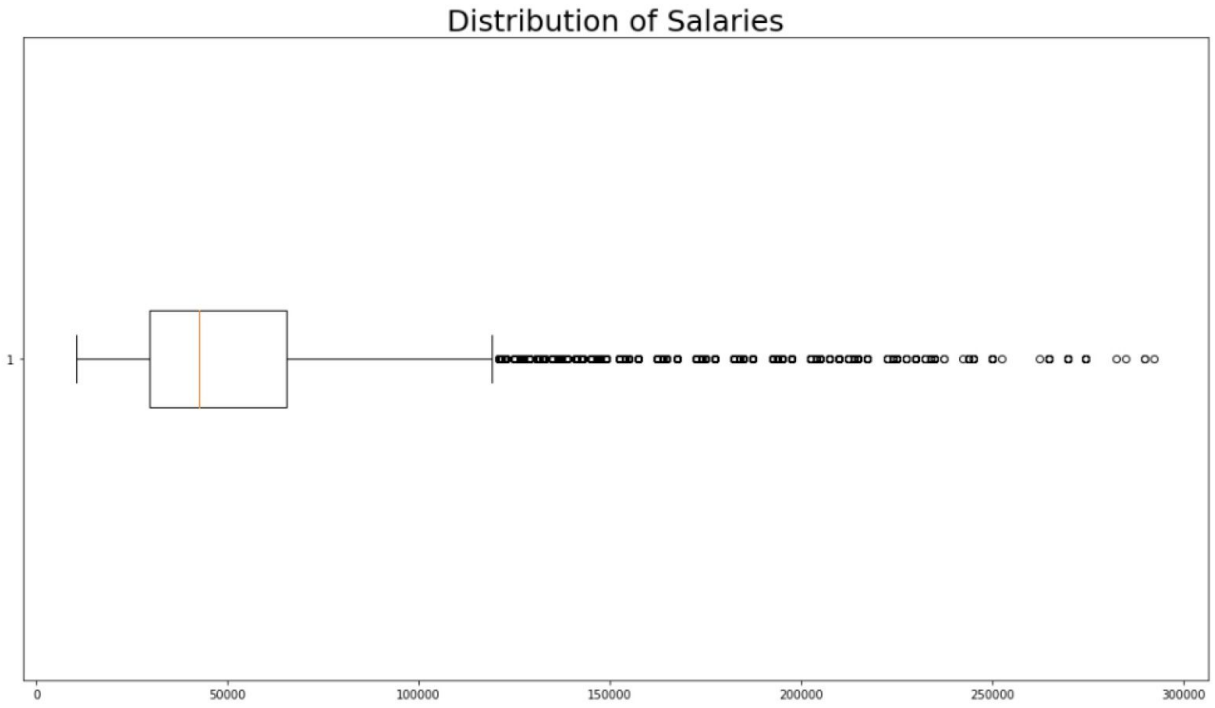


Distribution of Roles for Job Postings





We notice an interesting spike in the number of job postings



As expected, salary has a right tailed distribution.

Phase II: Data Cleaning

During the data preparation phase, we spent a lot of time getting the dataset into a modelling-ready form.

During the data cleaning part, we encountered the following issues:

- We grouped jobs by company, and deleted all those company's jobs who posted less than 500 jobs during the time horizon of the dataset. This preserved about ~95% of the dataset.
- We converted datetime values from a single string to separate columns for month, day, year.
    - Caveat: Year was dropped from the analysis as all jobs were from the year 2018 (hence no predictive power for year)
- \n values masquerading as N.A (Null values), that needed to be replaced.
    - Resolution: deleted job postings with null values
        - Exception: Salary, which was imputed using Median of all the jobs in the dataset.
- Categorical variables that were not encoded as dummy variables
    - Resolution: dummy encoded the State, Ticker variables. 'City' was dropped from the analysis as it had 11,900 unique values, thus Python memory errors were frequently encountered while dummy encoding it.
- Due to the sheer size of the dataset, carrying out simple tasks such as one-hot-encoding categorical variables took inordinately large amounts of time (~30 minutes)
- We finally carried out a 70-30 training/test split.
    - Caveat: Since we wanted to ensure at a similar distribution of companies posting jobs between training and test set, we carried out a 'stratified sampling' on ticker value.

Phase III: Analytical Modelling

For the modelling phase, rather than explore a breadth of models, we decided to pick a few linear models and spend time optimizing them. The overall goal was to use them as a benchmark for Method A/Method B outlined above. We chose to explore the following models:

- Multiple Linear Regression
- LASSO Regression
- Ridge Regression

A major advantage of choosing linear models is that it allowed us to characterize how a job posted by each company had different effects on the predicted stock price. We hoped to do this by comparing the different coefficients fitted to each stock ticker. Since high interpretability was a major concern from the outset (our client - Greenwich HR provides insights to stock managers, investment funds who are used to single measures of stock performance (alpha, gamma, beta, etc.)), the linear regression coefficients provide a great discriminant between various types of stocks and their job posting activity, and are easy to understand- hence they were used.

The key results from the modelling phase are outlined below.

Key Results

When deciding between our models, our team chose to focus primarily on linear implementations due to their ease of interpretability. More specifically, the team sought to extract the coefficients from the generated fitted models that would in turn allow for characterization of the manner in which a job posted by a certain company had different effects on the predicted stock price.

The first model built was multiple linear regression, while using the "pct_change_90_stock" (or percent change in stock closing price over the last 90 days) as our dependent variable . This model resulted in an R-squared value of approximately 0.58292.

The next two models implement regularization in the form of LASSO and Ridge regressions.

When first running LASSO regression, the team noticed a significant drop in the estimated R-squared value generated by the sklearn model. As this was most likely due to the invalid usage of the default alpha parameter, the team then chose to proceed with using k-Fold cross validation to help select an appropriate alpha. This methodology was also used to help select the appropriate tuning parameter alpha for our Ridge Regression model development.

For LASSO regression, the optimal tuning parameter (alpha) value was chosen to be 0.001, while in Ridge Regression, it was 0.01. The team utilized these two alpha values in the formulation of the final predictive models.

The final R-squared value for the LASSO regression model was approximately 0.52668, while that for the Ridge Regression model was 0.58292.

The next step in comparing the three models' performances was to calculate the Mean Squared Error for each on the test dataset. To calculate this performance metric, the team chose to use the mean_squared_error function from the sklearn module.

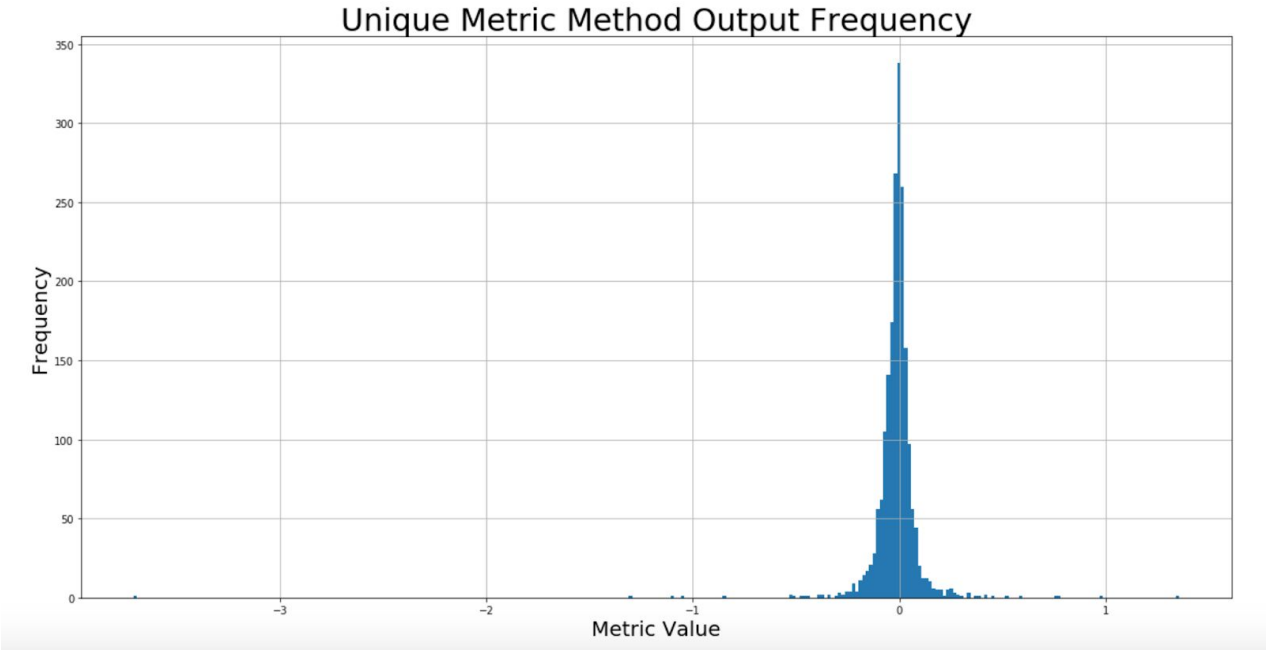
For LASSO regression, the final validated model resulted in a Mean Squared Error (MSE) of approximately 0.01481.

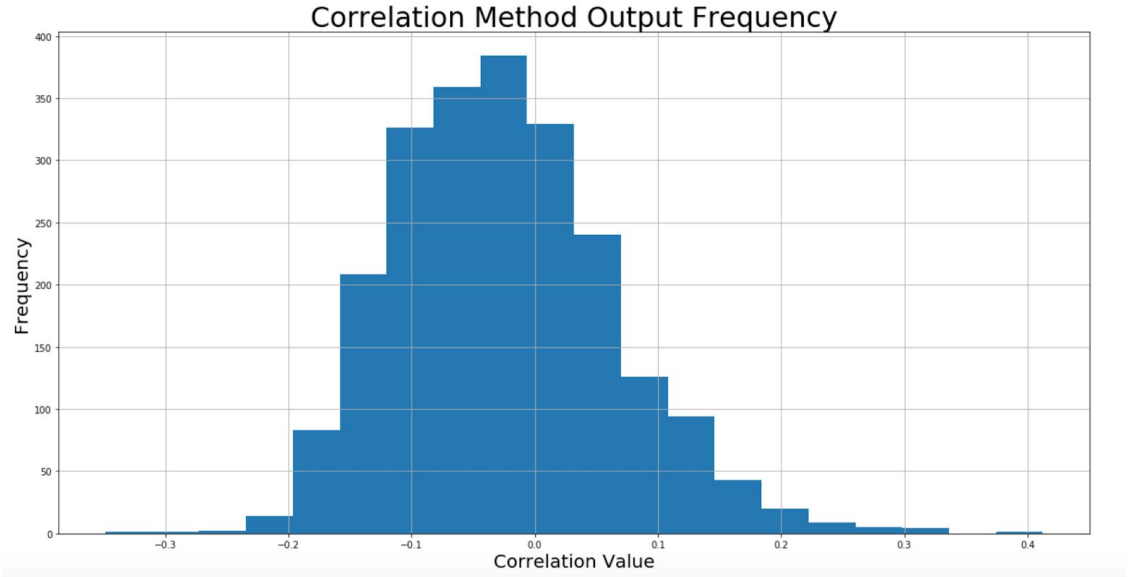
For Ridge regression, the final validated model resulted in an MSE of approximately 0.013029.

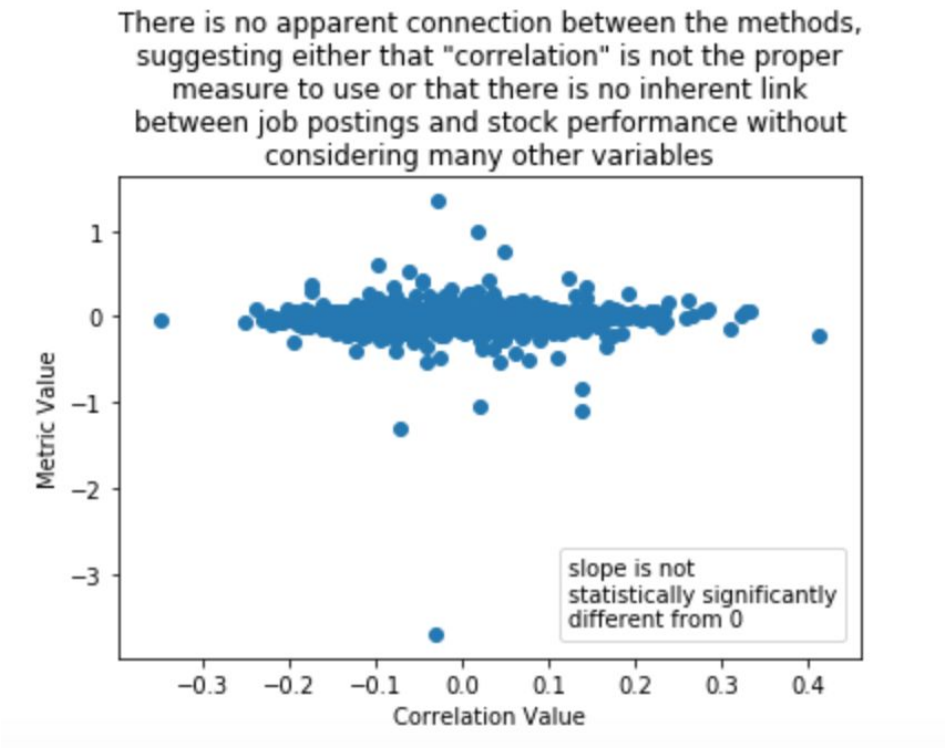
Lastly, for the final validated Multiple Regression Model, the test set generated an MSE of approximately 0.013029.

Even these consistent values for MSE and R-squared, our team nevertheless concludes that there is not a strong relationship between the number of job postings for a given company and it's expected change stock market closing price.

Appendix A


Unique Metric Method Output Frequency
Frequency
Metric Value
0
50
100
150
200
250
300
350
−3
−2
−1
0
1


Correlation Method Output Frequency
Frequency
Correlation Value
0
50
100
150
200
250
300
350
400
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4


There is no apparent connection between the methods, suggesting either that "correlation" is not the proper measure to use or that there is no inherent link between job postings and stock performance without considering many other variables
Metric Value
1
0
−1
−2
−3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
Correlation Value
slope is not statistically significantly different from 0