Glassdoor Data Scientist Salary Prediction using Regularized Linear Regression

Sasha Nanda, Jack Ellis, Elisa Du

(Dated: 15/12/2020)

ABSTRACT

The goal of this research was to predict annual salaries for data science positions. Job listing data, as well as company information, was scraped from Glassdoor. Some of the variables scraped include company size, company location, job title, job description, annual salary, and salary bounds. Filters were created to categorize job positions into five levels and remove irrelevant jobs. Other general cleaning processes were also performed. Exploratory analysis was performed on the job position, company location, and job descriptions features. From this, it was found the best ways to deal with these features include mapping the filtered positions to integers, using a Bag Of Words model for the job descriptions, and using the latitude and longitude of the job location. Three different models were used during the training phase: a forward selection Linear Regression, a Lasso Regularized Linear Regression, and a Ridge Regularized Linear Regression. All models performed similarly with Mean Squared Error's falling between 0.13 and 0.18. The Ridge model outperformed the Lasso model in terms of Mean Squared Error, Mean Absolute Error, and R^2, indicating that a large portion of the features are important rather than a select few. The job position feature was found to be important, as it was in the top 5% of most important features. The Ridge model found that a promotion by one job position results in a salary increase of $5215. The most influential word features included the sequence (color,religion,sex), and the words (experience) and (learning). This suggests that jobs that prioritize prerequisites, self-growth, and non-discriminatory hiring practices result in higher salaries. Additionally, the Ridge model found that a US based job pays $14911.50 more than a UK based job on average. Also, a UK based job pays $654.86 more than a Canadian based job.

I. INTRODUCTION

As organizations have shifted towards a more data-driven approach to problem solving, the demand for data science positions has risen considerably. Qualifications and compensations have become highly varied and, at times, ambiguous. When applying for data science roles, understanding the current landscape of compensations can help applicants determine which roles are suitable for them and where to apply for. Learning about the key factors that affect salaries can also help applicants better filter out potential employers, as well as strive to fulfill certain qualifications and skills needed.

The objective of this project is to predict the annual salary for data science job positions, ranging from intern, junior, senior, lead, to general data scientists, using Glassdoor [1] data. The three research questions addressed are 1) how job position affects salary, 2) recurring words across job descriptions that affect salary, and 3) how location impacts salary. Section

II delineates the technical and statistical methods used. Section III describes and justifies the feature selection methods and models chosen. Section IV reports the metrics obtained from each model. Section V addresses the three research questions based on the results, as well as discusses other significant findings. It also discusses some limitations and shortcomings of the project. Section VI concludes the report and gives some future considerations.

All scripts and output are available at the Github repo [2].

II. METHODS

The general workflow is shown in Figure 1, and the following sections cover the workflow in more detail.

FIG. 1: General project workflow

A. Data Collection

Three types of data were scraped from Glassdoor: salary data, company data, and job listings data. The detailed fields scraped are listed in Table I.

Salary and company data were scraped using the Python `BeautifulSoup` library. Job listings data, which were dynamic sites with JavaScript redirects, were scraped using the Python `selenium` library with Chrome WebDriver.

Types of Scraped Data		
Salary data	Company data	Job listings data
• salary • cash and stock bonuses • lower and upper bound salary • base and bonus pay	• company name • number of employees • company sector, type • location of HQ • revenue • Glassdoor rating	• job title • job description • company info

TABLE I: Data scraped from Glassdoor

B. Data Cleaning

First, some standard data cleaning techniques were implemented. Numerical information was stripped of unit signs (such as a dollar sign on price) and converted to integers. Salaries quoted monthly or hourly were converted to yearly by assuming a forty-hour work week, or a twelve-month work year. Units were standardized (such as converting all money-based features to CAD). Furthermore, companies with `NaN` salaries were dropped.

Next, position names were cleaned. There were a variety of positions that were found to be irrelevant to our analysis, such as Data Analyst, Data Engineer, and Research Scientists. Such positions were dropped. Based on the remaining data, five overarching Data Scientist position titles were found to be appropriate: Data Science Intern, Junior Data Scientist, Data Scientist, Senior Data Scientist, and Lead Data Scientist.

It was discovered that all Data Science positions fit into one of the aforementioned roles. For instance, a role like "Principal Data Scientist" was converted to Lead Data Scientist. Note that the position of *Data Scientist* refers to any positions that originally had the same title, as well as job titles with ambiguous responsibilities that made it difficult to classify then under any of the Junior, Senior, or Lead job titles. At times certain job titles (e.g. manager of advanced analytics) were manually classified as one of the five titles by reviewing the responsibilities described in the job listings and using one's general domain knowledge. As such, the process of standarization may contains some biases. The position standarization workflow is detailed in the cleaning scripts.

Finally, upon cleaning, the data was merged. In particular, the company data was first merged (on company name) with the salary data. Finally, this data was merged (on company name and job position) with the job listings data. This produced a complete dataset where each data entry had salary data, company data, and job description data.

C. Exploratory Analysis

1. *Exploring the effect of job position on salary*

The first research area explored is the effect that the aforementioned five data science job positions have on their salary. Figure 8 below shows the distribution of these five Position levels, with "Normal" level denoting the *Data Scientist* position. There is an imbalanced distribution of the five position levels, with "Normal" Data Scientist taking up the largest fraction out of all cleaned job entries, and Junior Data Scientist taking up the least. The high frequency of "Normal" Data Scientist entries can be attributed to two factors: 1) many scraped job titles were simply "Data Scientist" and had no associated hierarchical semantics (e.g. junior, senior) and 2) the imprecise Glassdoor search functionality yielded more ambiguous job titles (e.g. Data Science analyst) than one would like, so they were then classified under the "Normal" Data Scientist level.

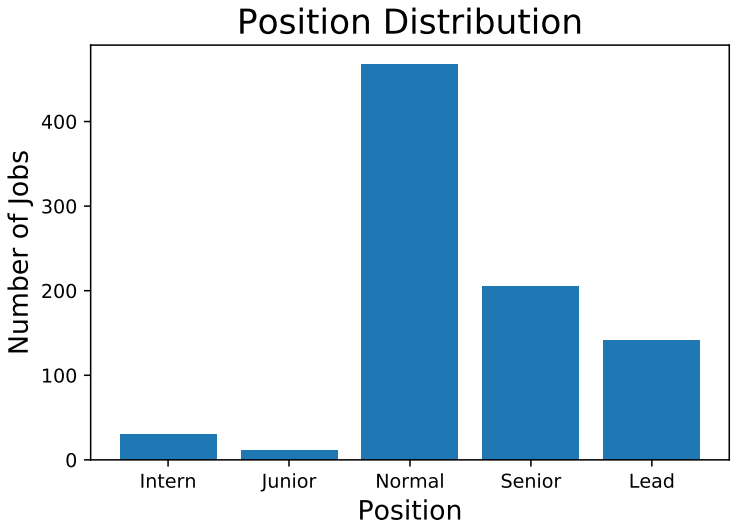


FIG. 2: Distribution of Position Labels from Clean Data

Figure 3 shows the mean salaries for each position level. There is a clear upward trend in the order of intern, Junior, Normal, Senior, and Lead. This is intuitive, as this would most likely be the case for any job field.

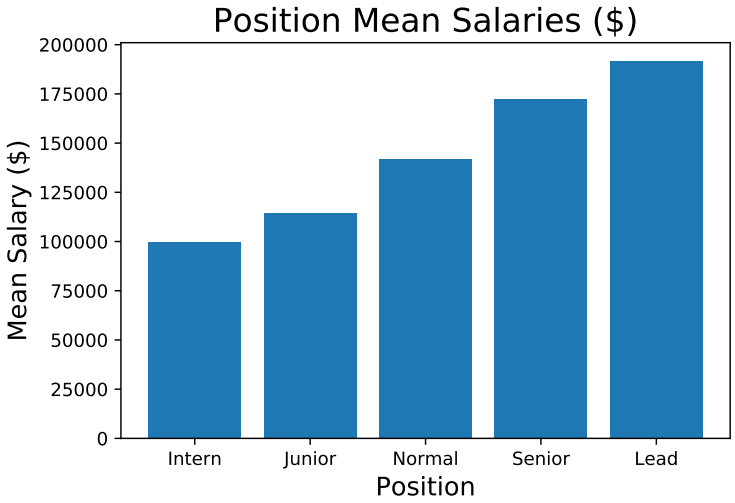


FIG. 3: Mean Salaries for Each Position Label

Figure 4 below shows the salaries by position label. Along the x-axis, a label of 1 represents Intern, 2 represents Junior, 3 represents Normal, 4 represents Senior, and 5 represents Lead. It is evident that there is a large variance across salary for each position level. For example, annual salary for the "Normal" Data Scientist level ranges from under \$50K to just under \$300K. The high variance in salaries indicates that the five position labels may not be as significant of a predictor of salary for this particular dataset.

FIG. 4: Salaries for Each Position Label

The high variance in salaries could also be due to the high variance in company size (e.g. difference in pay between small and large companies) and job location. In terms of job location, three countries' job listings data were collected: Canada, the US, and the UK. Figure 5 shows salaries by position label *and* by country. The same position-integer mapping from Figure 4 was used. It is clear that for each position label, the US has higher potential salaries, as well as considerably more data points, than Canada and the U.K.; however, the variance within US salaries is still very high on its own.

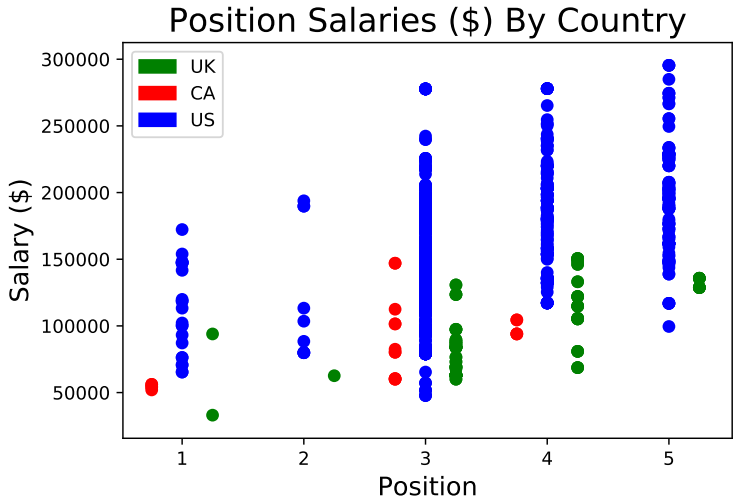


FIG. 5: Salaries for Each Position Label and Country

2. Exploring common words in job descriptions

The word cloud for all job descriptions is displayed in Figure 6. The frequency that a word occurs is proportional to its size. The most common words are: data scientist/science, machine learning, solution, year, customer, and product. The latter two words indicate that most listings scraped are based in industry rather than academic institutions, as expected.

The word clouds for Junior, Senior, and Lead data scientists are shown in Figure 7. Comparing them, all three job position listings have in common the word "machine learning", indicating the heavy overlap between predictive modeling and data science-related tasks. The words "deep learning" and "artificial intelligence" appear in the word clouds of both lead and senior data science job descriptions, but neither appear in the word cloud for junior data scientists, for which "statistics" and "people" are the most common words. This suggests that junior roles are more focused on leveraging statistical knowledge and communicating insights to others, rather than on advanced predictive modeling where more complex and less interpretable models are applied.

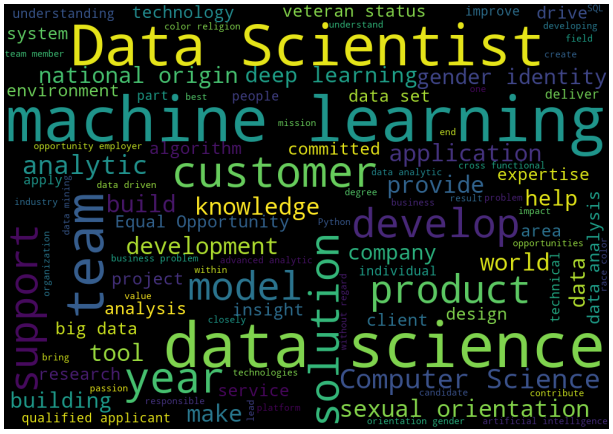
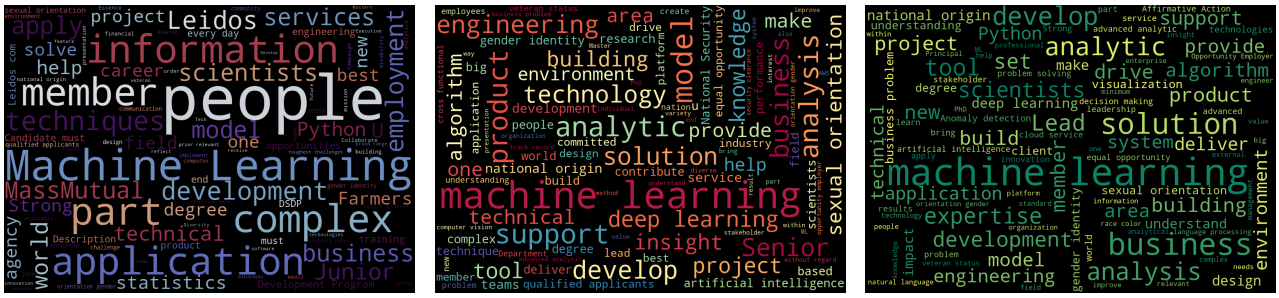


FIG. 6: Word cloud for all job descriptions.



(a) Junior Data Scientist (b) Senior Data Scientist (c) Lead Data Scientist

FIG. 7: Word clouds by position.

3. Exploring the effect of company location on salary

The final research question was whether job location affects annual salary. The histogram shown in figure 8 shows that the data is imbalanced, with a vast proportion of jobs being located in the US. This suggests that the models fitted in the future would more accurately predict salaries for the US than for the other two countries.

Furthermore, figure 9 shows that, on average, jobs in the US pay vastly more than those in the UK, which pay more than those in Canada, suggesting that job location does impact salary. Remote jobs have the highest average salary (even higher than the US), although due to data imbalance (i.e.: we have significantly more US data compared to remote data), it is not possible to do a concrete analysis of remote salaries.

In order to gain a more fine-grained understanding of the impact of location, the company headquarters data were converted from strings to latitudes and longitudes and explored. Figure 10 is a scatterplot of headquarter latitude and longitude versus salary. There doesn't seem to be much of an impact of latitude or longitude on company salary. Instead, we see

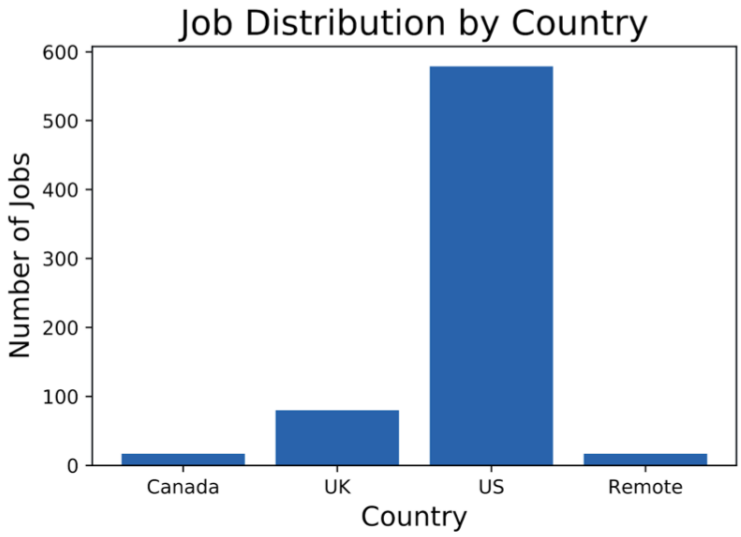


FIG. 8: Histogram of Jobs by Country.

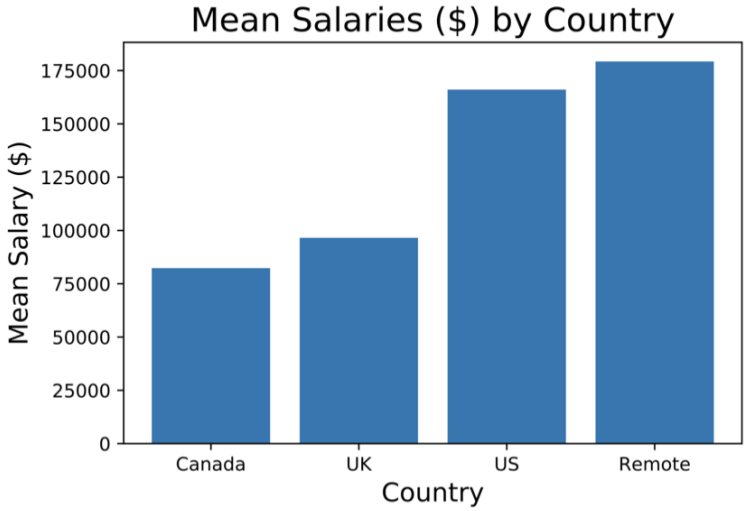


FIG. 9: Histogram of Mean Salary by Country.

some clusters forming in the Silicon Valley, Philadelphia, and New York, which suggests that almost all company headquarter locations scraped are in these three cities.

4. Other

The continuous features were explored by plotting a heatmap of the feature correlation matrix (shown in figure 11) and analysing the correlations between each feature and the target variable of salary. Features such as the lower and upper bound on salaries, as well as stock bonus, exhibit extremely high correlations with salary. Interestingly, user-generated features such as company rating exhibit low correlation with salary. This suggests that the reviews on Glassdoor are not particularly accurate, since it would logically stand to

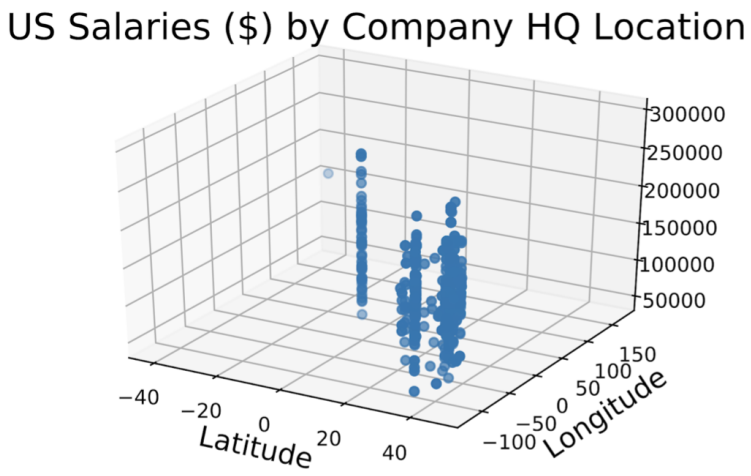


FIG. 10: Scatterplot of US Headquarter Locations versus Salaries.

reason that a highly rated company would pay their employees more. Finally, we note the multicollinearity between features; for instance, lower bound salary and bonus pay have a correlation of 0.66, which is quite high. The lack of independence of the features suggests that there might be some bleeding of information between features in the model, which might obfuscate the interpretation of models trained on the data. This also calls for feature selection techniques to be used.

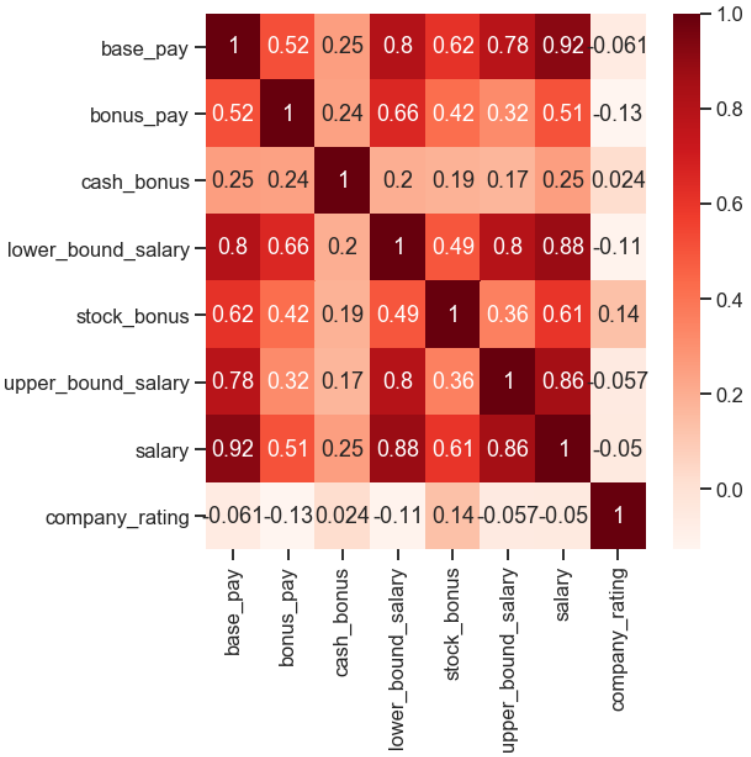


FIG. 11: Correlation matrix of salary features, company features, and salaries.

D. Data Preprocessing

1. Bag of Words

A BOW (bag-of-words) model was used to encapsulate the job descriptions. A BOW model creates indicator functions for each word or sequence of words from the job description corpus. For simplicity, the top sixty most common word features, with twenty in each of the sequences of length 1, 2, and 3, were used. BOW was chosen for its effortless implementation and interpretability. The model uses frequency counts of the sequences as a measure of relative importance, and makes it easy to detect the presence of keywords common across all job descriptions.

2. Categorical Variables

The job position was encoded as an ordinal feature and mapped to an integer from 1-5, where: Data Science Intern $\rightarrow$ 1, Junior Data Scientist $\rightarrow$ 2, Data Scientist $\rightarrow$ 3, Senior Data Scientist $\rightarrow$ 4, and Lead Data Scientist $\rightarrow$ 5. This ordinal encoding makes sense due to the inherent rank ordering amongst the position labels - a lower integer value implies a less experienced role (which likely pays less) whereas a higher integer value implies more experience and responsibility and thus higher compensation.

Since other categorical variables such as company or industry type do not share the inherent rank order (i.e.: it wouldn't make sense to make a statement like E-Commerce $>$ Technology), these categorical variables were encoded as one-hot vectors. Similarly, since exploratory analyses showed that country location does affect salary, the job country was one-hot encoded.

III. MODELS

An 80/20 split on the cleaned data was used to generate the train and test datasets. For each model, training was performed on an unnormalized version of the train set in order to generate model weights (i.e. coefficient estimates) that were easily interpretable, and training was also performed on a normalized version of the train set in order to generate feature importance plots. The following three models were fit to the data:

1. Forward selection, with a significance threshold of 0.025, was used to choose the most important features. Then, linear regression was fitted to the data.

2. The full feature set was used, and Lasso Regularization was used to reduce over-fitting and perform automatic feature selection. Since Lasso enforces a linear penalty (by penalizing the magnitude of coefficients) on the loss function, this drives many of the feature weights down to zero, resulting in a sparse model and implicitly feature selection

3. The full feature set was used, and Ridge Regularization with the default `sklearn`penalization parameter of 1.0 was used to reduce over-fitting. Unlike Lasso regression, Ridge regression enforces a quadratic penalty on the loss function, which drives many of the feature weights down to small but non-zero values, resulting in a dense model where all the features are deemed relevant enough to include and the multicollinearity amongst them is reduced.

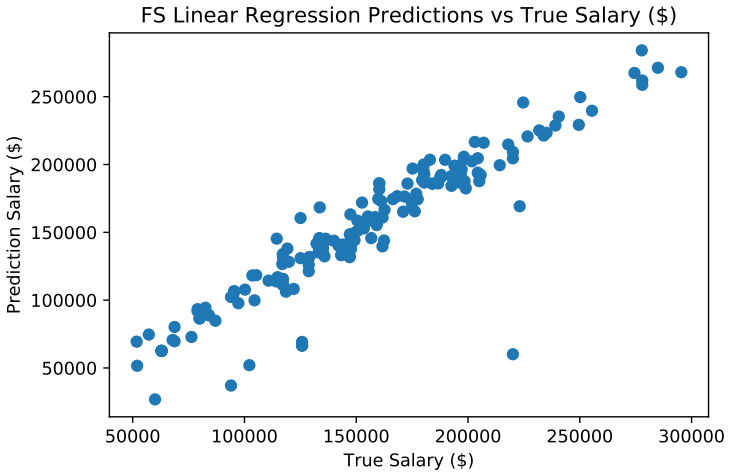
IV. RESULTS

Table II below shows the *test* metrics for each model. Each model performed similarly, despite the variation in the features selected by each model (discussed in section [V D 1]). All models had a Mean Squared Error (computed for the normalized data) between 0.13 and 0.18, an average salary error (i.e. mean absolute error computed for the unnormalized data) between \$11,000 and \$12,000, and an R^2 term between 0.85 and 0.86. The relatively high R^2 indicates that the features selected explain the variance in target salary relatively well. However, the R^2 may also be inflated due to including more features in the forward selection and ridge-regularized model, which results in them having higher R^2 values than the Lasso model that eliminates many features by shrinking them to zero. This is why it is important to report prediction error metrics as well. On average, the predicted salary is off by an amount between \$11,000 and \$12,000 amongst the three models. The forward selection had the best prediction performance in terms of the three metrics, where the Ridge Regression Model was the second best-performing.

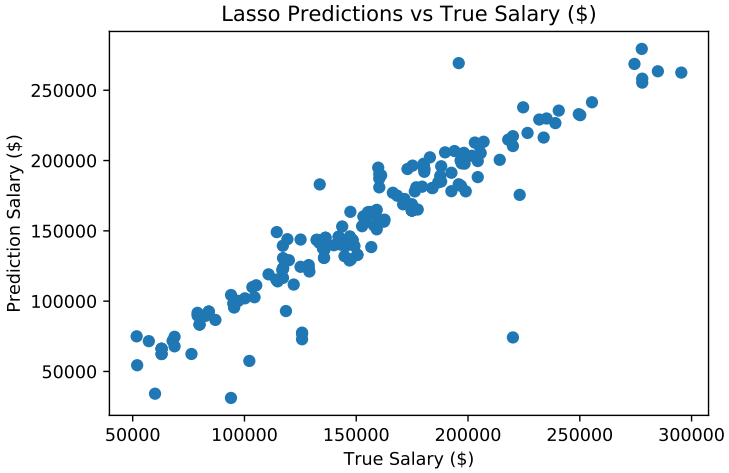
Model	MSE	Average Salary Error	R^2
Forward Selection Linear Regression	0.1397	11130	0.8585
Lasso Regularizer Linear Regression	0.1782	11861	0.8543
Ridge Regularizer Linear Regression	0.1447	11681	0.8575

TABLE II: Test Metrics for Each Model.

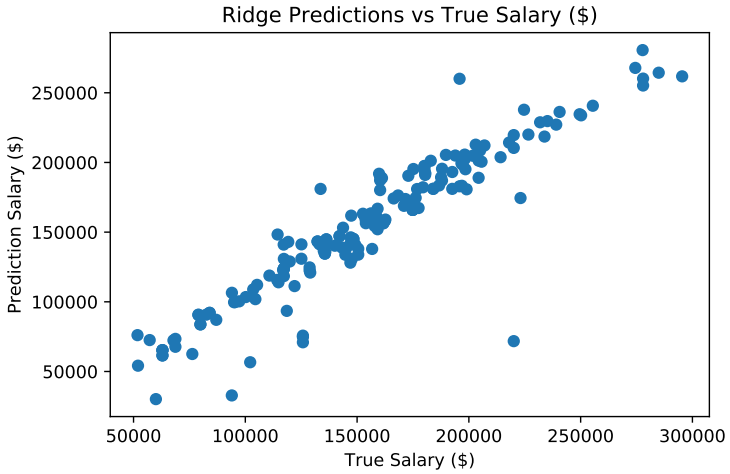
Predicted Salaries versus True Salaries are plotted in Figure 12 for each of the three models fitted. By comparing these plots, it is clear why the three models performed similarly; each model predicted the same outliers in similar manners. From first glance, it is almost impossible to tell the difference between Figure 12b and Figure 12c, as the difference in predictions is extremely small.



(a) Forward Selection Linear Regression Predictions with True Salaries



(b) Lasso Linear Regression Predictions with True Salaries



(c) Ridge Linear Regression Predictions with True Salaries

FIG. 12: Model Predictions

V. DISCUSSION

The next sections discuss the three research questions in light of the modeling results: 1) how the job position feature affects salary, 2) word features that affect salary, and 3) how location features impact salary. The ridge-regularized linear regression model was selected to answer these questions. Although this model did not have the lowest MSE (its MSE was slightly higher than that of forward selection with linear regression, which had the lowest MSE), it not only aims to select the coefficients that reduce the residual sum of squares, but also applies a shrinkage penalty that reduces overfitting and multicollinearity. It is also more computationally efficient than subset selection methods such as forward selection.

A. Research Question 1

The position feature was the 7th most important feature out of the 143, putting it in the top 5%. Its coefficient estimate from the Ridge Regression model was 5215, implying that going up one job level increases a salary by $5215, while holding all other features constant. According to indeed [3], the average promotion involves a 3% pay raise. Since data science is a highly-skilled job field, this value is not unreasonable. However, for large companies like Facebook, it is hard to imagine that the difference in salary from a promotion is about $5000. For example, glassdoor has a job salary for a Facebook data scientist intern at about $75K [4], whereas the average data scientist salary at Facebook is about $150K [5]. This may be due to the large variance in salary across each position, as well as a weak representation of company size (in terms of pay to data scientists) in the dataset. It is also important to note that the position feature was found most important by the Forward Selection Linear Regression, implying that position is a strong feature for salary prediction.

B. Research Question 2

The word features in the top 15% of the most important features are the trigram "('color', 'religion', 'sex')" and the unigrams "experience" and "learning". Their (unnormalized) coefficient estimates are 9019.86, 7006.60, and 6432.90, respectively. The two aforementioned unigrams suggest that positions with more pre-requisites (i.e. experience) and an expectation to pick up new skills (i.e. learning) are associated with a higher salary. The trigram "('color', 'religion', 'sex')" comes from the diversity and equity hiring statement. It ranks higher in importance and has a larger coefficient than both unigrams, implying that companies that incorporates non-discriminatory hiring statements are associated with higher pay.

C. Research Question 3

The coefficient estimates (i.e. weights) for the one-hot encoded country features and the latitude and longitude are shown in Table III. As seen in the table, a US job on average pays $10159.29 - (-4752.21) = \$14911.50$ more than a UK job, and a UK job pays $-4752.21 -$

$(-5407.07) = \$654.86$ more than a job in Canada. Clearly, the model picked up on the fact that US salaries are higher than UK salaries, which are higher than Canadian salaries (as discussed in the exploratory analysis). However, based on the results displayed in figure 9, one might expect the US, UK, and Canada weights (especially US) to be much higher. Due to the multicollinearity of the data (for instance, the fact that the US feature was correlated with lower bound salary), it could be possible that other important features such as lower bound salary accounted for a portion of the "true" magnitude of the US feature. Since the features are not uncorrelated, the analysis would not be as literal as one would hope to be.

Finally, latitude and longitude features are not particularly important. This supports the exploratory analysis, since there was a clustering of locations where most company headquarters tend to reside, and hence there wasn't enough variance in the latitude or longitude features to make a notable contribution to the model. Perhaps if Glassdoor offered the exact job locations rather than the company headquarters then a fine-grained analysis of location would have boosted the importance of the fine-grained company location feature.

Feature	Weight
US	10159.29
UK	-4752.21
CAN	-5407.07
lat	-27.07
long	-46.22

TABLE III: Weights for Location-Related Feature

D. Other

1. Other important features selected by ridge regularization

The top 10 most important features selected by the ridge regularized model are plotted in Figure 13. The feature importance metric used in the y-axis is the absolute value of coefficient estimates from the normalized model. The higher this metric is, the more important the model interprets that feature to be. In the plot, the top two features are lower bound salary and upper bound salary. Their corresponding coefficient estimates in the model fitted on the non-normalized data are 0.39 and 0.29, respectively, which are quite small. This is expected due to both features being on the same scale with the target salary, so a dollar increase in either lower or upper bound salary, while holding all other predictors constant, would only result in a slight increase in the predicted salary. The third, fourth, sixth, and eighth most important features are word trigrams that appear in the equity hiring statement, i.e. "*all qualified applicants* will *receive consideration for employment without* discriminating against *sexual orientation*, race, color, ...". These trigrams are seen as important factors that affect salary since it may be that companies that are more conscious of their equity and

diversity hiring practices are of larger size which tends to pay better, or these companies put in more effort to compensate their employees better. Also, the 5th and 9th most important features may be unsurprising since jobs in the "Computer Hardware & Software" (i.e. tech) and "Internet" industry generally pay considerably better than other sectors, and thus can more heavily impact the predicted salary.

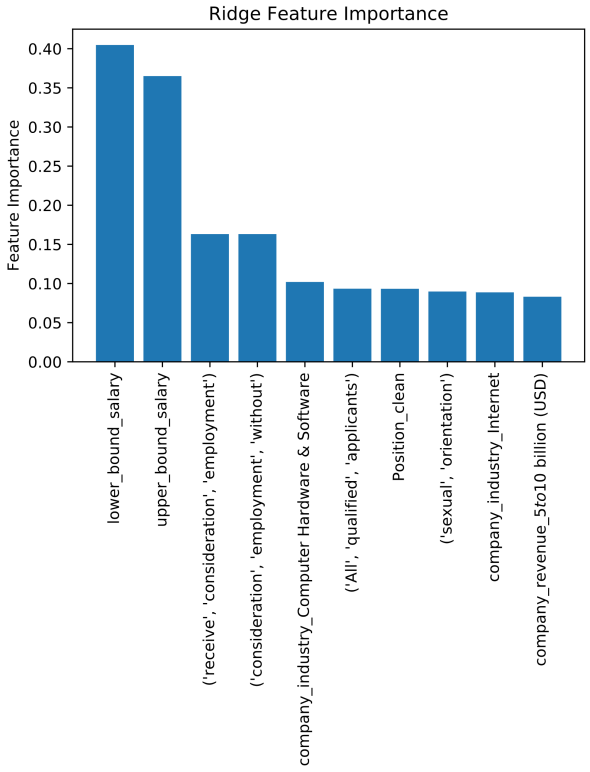


FIG. 13: Top 10 most important features selected by ridge regularization.

2. Features selected by other models

The top features selected by forward selection are shown in Figure 14. In total twenty features were selected by forward selection. The top two most important features are position and bonus pay. The other 8 features all belong to company sector (e.g. aerospace, banking, education).

The top features selected by lasso regularization are shown in Figure 15. In this case, only six features are selected (i.e. have non-zero coefficient estimates). Lower and upper bound salary are ranked as the the most important features in determining salary, followed by stock bonus and it being a job in the US. It is interesting how positions in "Colleges and Universities" also are interpreted as important in predicting salary. The corresponding (unnormalized) coefficient estimate is 25339, which is higher than that for "Hardware & Software" jobs which is 14571. This may be because the cleaned dataset is non-representative of the salary landscape (discussed in next section). Namely, academic job listings scraped may be from

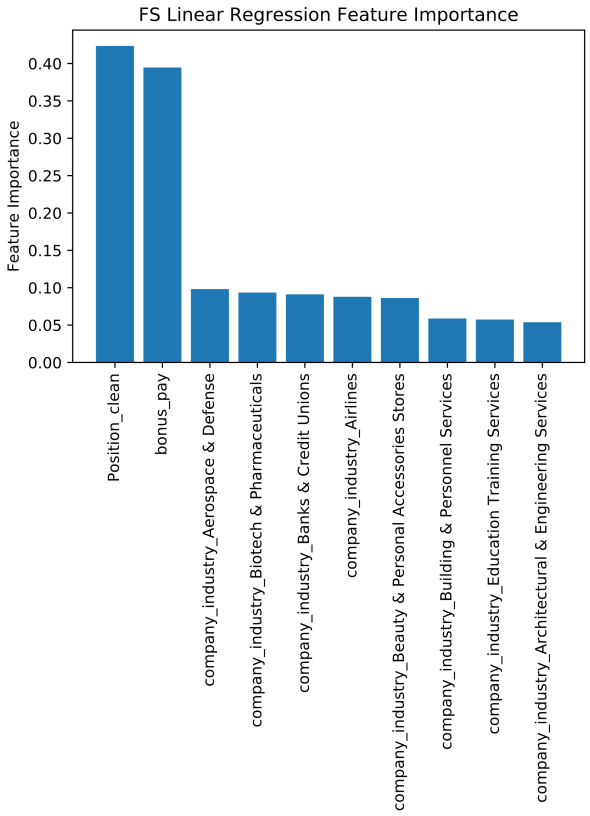


FIG. 14: Top 10 most important features selected by forward selection.

private research or academic institutions that have better compensations than their publicly-funded counterparts.

E. Limitations

Throughout the statistical process, there were trade-offs made. First, the final dataset was fairly small in size - only 854 job descriptions. This size would likely imply the dataset is not a fair representation of all job listings and the data science salary landscape. With more data, feature importance and model metrics may be different. This was heavily influenced by another limitation: Glassdoor's capabilities. Glassdoor was not made for data scientists to scrape information, as they have page limits for job searching. Due to these page limits, scraping all job descriptions on the site was incredibly hard to do and was not feasible given the project timeline. In addition, Glassdoor does not have extensive salary information for all companies. For example, for a particular company, salaries for "Senior Data Science" would be available but not for "Data Scientist Intern". Finally, a trade-off was made by using the Bag-of-Words model. Although the model is easy to implement and interpretable, the effectiveness of BOW does not compare to preprocesssing techniques such as pre-trained word embeddings or Word2Vec. Higher prediction accuracies may have resulted by employing ensemble models, but for sake of interpretability, linear regression was used.

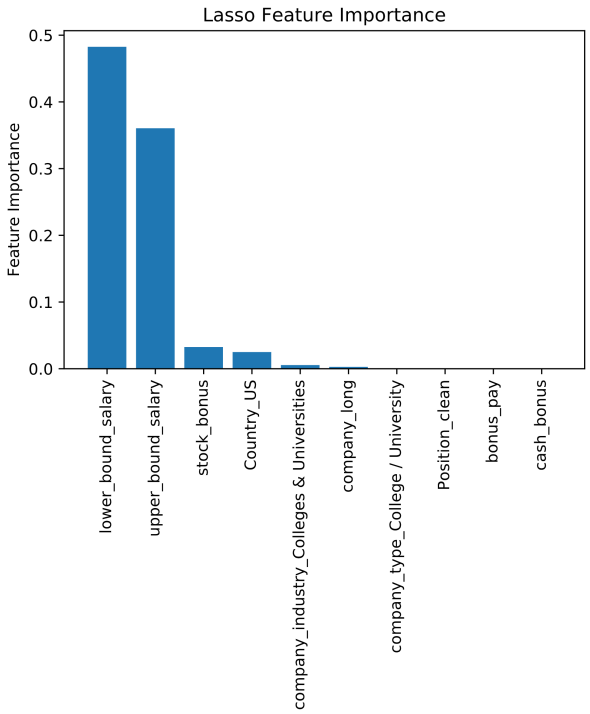


FIG. 15: Top 10 most important features selected by lasso regularization.

VI. CONCLUSION

To predict annual salary using scraped Glassdoor data, 143 total features were used and three model were used for feature selection and fitting to generate salary predictions. Their MSE, MAE, and R^2 values were reported and compared. The forward selection-linear regression model had the best prediction performance, and the ridge-regularized linear regression was chosen to address the three research questions. It was found that the position feature had a positive association with salary, and word features that connote equity hiring practices and prior job experiences are also associated with higher salaries. It was also found that the country feature was important, and that US salaries were the highest amongst the three countries. However, the more fine-grained company headquarter location feature was not a significant determinant of salary.

A. Future Considerations

Due to the time and resource limitations mentioned above, there are ample extensions that can be made to our work. First, more advanced language preprocessing and embedding techniques could have been implemented such as word lemmatization and using pretrained word embeddings to capture the semantic meaning of words in the job descriptions. Additionally, PCA could have been used as another potential feature selection method, or even

as a way to learn a latent distribution over the features.

---

[1] https://glassdoor.com.

[2] https://github.com/sashananda/sta_project_1.

[3] https://www.indeed.com/career-advice/pay-salary/average-promotion-raise#:~:text=According%20to%20the%20Bureau%20of,on%20average)%20a%20%241%2C200%20raise.

[4] https://www.glassdoor.ca/Intern-Salary/Facebook-Internship-Salary-E40772.htm?filter.jobTitleFTS=Data+Scientist.

[5] https://www.glassdoor.ca/Salary/Facebook-Data-Scientist-US-Salaries-EJI_IE40772.0,8_KO9,23_IL.24,26_IN1.htm.