Executive Summary

The aim of this report was to gauge salary amounts for data scientists by investigating the industry factors that influence the pay scale for these professionals. The data was collected by scraping indeed.com, an aggregator that regularly pools job postings from various resources. The results suggest that the keywords used in the job titles are the most significant predictors for salary levels. It is recommended that a position should be understood and described accordingly in the job title, especially when highlighted keywords are used.

Introduction

Our company is planning to hire more data scientists and wants to be competitive in the hiring market. To do this, we want to be able to predict whether the salary for a data scientist is higher or lower than national median. We also want to understand which industry factors have impact on the pay scale.

Methods

The research was conducted by collecting job posting data from indeed.com. Specifically, queries were made with the keywords “data scientist” combined with 90 different cities within 100 miles radius. The company name, job title, job description snippet, the location, and the annual salary were extracted from each of the results. Natural Language Processing technique was used to extract text features from the job titles and job description snippets, which were then analyzed by a few classification models, including Random Forest, Logistic Regression, K-Nearest Neighbors, and Support Vector Machine, for statistical inference and prediction. The performance (in terms of accuracy score) of these models were evaluated with cross-validation.

Results

3,215 job listings with annual salary information was collected. These listings covered 702 cities in 48 states in the US. The salary median was $74,669. The best accuracy score was 73% ± 6%, obtained from Support Vector Machine technique. The job title was identified as the most impactful factor in predicting salary levels. The most important keywords in job titles and the direction of their impact are listed in the table on the next page.

Positive Keywords	Negative Keywords
data	analyst
engineer	specialist
senior	research
scientist	technician
developer	associate
sr	ii
software	assistant

Interpretation of Results

The importance of the keywords was determined by Random Forest, while the direction of impact was inferred from Logistic Regression. It is worth noting that both techniques identify similar keywords of significance, which strongly suggest the validity of the findings. However, the models are limited by their accuracy performances, which must be taken into consideration when using the models to make predictions.

Conclusion

The job title was identified as the most impactful factor in predicting salary levels. Certain words describing job function, experience, or skills are particularly important for prediction.

Recommendations

It is recommended that when a job listing is composed, the skills, experience, and responsibilities needed for the position should be understood and described accordingly in the job title, especially when the keywords highlighted above are used.

To improve model performance, we should continue to accumulate more data over a longer period of time; we may also consider collecting data using other methods such as surveys. To gain more insight on industry factors that influence the pay scale, we should collect the full job descriptions (instead of just the snippets) that contains a fuller picture of the skills, experience, and responsibilities required of the data scientist positions.