**Data Scientist Salaries**

Carson Brower, [cbrower@bellarmine.edu](mailto:cbrower@bellarmine.edu)

Garrett Ward, gward2@bellarmine.edu

**ABSTRACT**

For this Final Project, our group used a Data Scientist Salary dataset from Kaggle. We used this dataset because of the question we had, can we predict a salary of a Data Scientist based on variables that are in connection and are closely related to what affects salaries. Within the document using our data will be shown with graphs, plots, and a correlation map of our variables. As well as the use of linear and logistic regression will allow for the most accurate results to allow us to produce a conclusion of our findings.

1. **INTRODUCTION**

To start off, with the use of Kaggle, Garrett and I found a dataset that was interesting to us and consisted of a topic we thought would be good to present to the class since it is related to the class. The data set we chose was data relating to data scientist salaries. Within the data set there are variables such as job title, salary estimate, rating, location, size of company, company age, different software is, and a few more. After looking at the data set in its entirety Garrett and I thought this was a good data set to do our final project on. From the dataset, the classification target, or the question we had, can we predict what salary a data scientist is going to have? Using the variables that closely relate and using the data for graphs, plots, matrix, and testing will allow us to produce the most accurate results and allow us to find our conclusion and solutions to the classification target of the data.

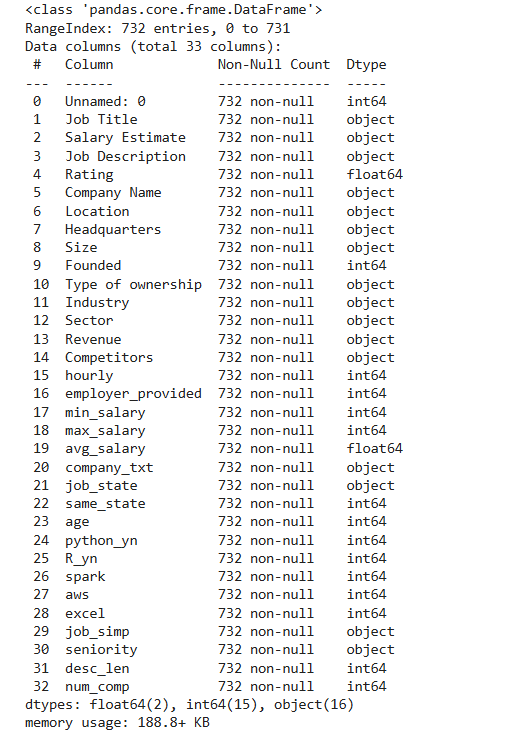
1. **BACKGROUND**

Data scientists are analytical people, they use data to understand and explain the things that happen around them and help organizations make better decisions. Careers in data are growing rapidly and is it becoming a more popular and needed career choice. Within their job they often develop predictive models for theorizing and forecasting. Data scientist find patterns and trends to uncover insights as well as create algorithms and data models to forecast outcomes. The use of machine learning helps improve the quality of the data and allows for communication of recommendations to others. For the data set we used itself, it contains job postings from Glassdoor from 2017. It can be used to analyze current trends based on job positions, company size, etc. This data set can be used to identify which factors most affect data science salaries, determine which states and cities offer higher pay, and predict what data science job will pay based on job description.

1. **EXPLORATORY ANALYSIS**

This data set has 732 entries with thirty-two columns. From our data there were no missing values or missing data, however there were a few miscues within one of our variables that we had to fix and fill in ourselves. We had a wide range of variable types in our data from int, float, and even object. When we first started working with our data, we noticed an age column and thought it meant the person’s age, but it meant the age of the company and that required us to use that in a unique way than originally. Something that stood out to us was the skew to the right on salary distribution where majority of the data scientist made anywhere between $50,000 and $100,000. Something that was surprising to us was the fact that company age really did not have an influence on salary. Our EDA graphs, plots, and more are shown below.

**Table 1: Data Description**



**Table 2: Data Sample**

**A screenshot of a computer

Description automatically generated**

**A graph of a number of employees

Description automatically generated**

**Figure 1: Size of Company Distribution**

**A graph of a distribution of salary

Description automatically generated**

**Figure 2: Distribution of Salaries**

**A diagram of a box diagram

Description automatically generated**

**Figure 3: Box Plot of Salaries**

**A graph of a number of people

Description automatically generated**

**Figure 4: Top Job Locations**

**A graph of a number of blue bars

Description automatically generated**

**Figure 5: Top Job Titles**

**A graph of blue bars

Description automatically generated with medium confidence**

**Figure 6: Job Titles and Average Salary**

**A graph of a company age

Description automatically generated**

**Figure 7: Scatterplot of Company Age and Average Salary**

**A graph of a number of numbers

Description automatically generated with medium confidence**

**Figure 8: Correlation Heatmap of all Variables**

**A graph of a number of blue and orange squares

Description automatically generated**

**Figure 9: Python Knowledge**

1. **METHODS**
   1. *Data Preparation*

To start off we got a sample of our data and realized no missing values were present, so we did not have to drop anything. We did have to fix a variable, but it was a simple task. After that we did EDA and gathered graphs, plots, and heatmap to give is and insight of the variables and our data that was present. The correlation heatmap gave us the greatest insight to see what variables directly correlated with each other. After we finished with our EDA, we then went into our experiment with logistic and linear regression which will be explained and shown below.

* 1. *Experimental Design*

A computer screen shot of a program code

Description automatically generated

Listed above is our experimental design on Python for our train and test split for the Linear Regression prediction model.

A screenshot of a computer program

Description automatically generated

Listed above is the experimental design we used using Python for our train and test split for the Logistic Regression prediction model.

* 1. *Tools Used*

The following tools were used for this analysis: Python v3.11.11 running the Anaconda 22.9.0 environment for HP ENVY Laptop computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 2.2.3, NumPy 2.1.0, Matplotlib 3.9.0, Seaborn 0.13.0, SKLearn 1.3.0, and Patsy 1.0.1. These were the tools we used because they are what we had and have been using for the whole semester. We also used these because it is what we are most familiar with and gave us the most accurate results and allowed us to find our solutions.

1. **RESULTS**
   1. *Classification Measures/ Accuracy measure*

A black text with black numbers

Description automatically generated with medium confidence

**Results from Linear Regression Model**

A screenshot of a computer

Description automatically generated

**Results from Logistic Regression**



**Accuracy Score**

A graph of a number of colored squares

Description automatically generated with medium confidence

**Confusion Matrix Model**

* 1. *Discussion of Results*

After going through our results, we got clear information regarding the linear and logistic regression tests we ran. When looking at the linear regression, a mean square error of 8.32 is low which emphasizes that the model’s results are similar or close to the actual results. With an R Square of 0.995 it is high, which is a good sign which means the model is fitting the data in a correct way. When looking at Logistic, we got an accurate score of approximately 73% which suggests that majority of the time the model correctly predicted the class of 0 or 1. An F1-Score of 0.70 and .76 indicates a decent balance oof the precision and recall of 0 and 1.

* 1. *Problems Encountered*

When looking at problems we encountered the first thing was we had invalid entries in the company size column and had to fix the entries so that our data was not messed up and allowed us to have accurate results. Overall, this was a simple task to fix but was a setback. Another problem we had was a problematic entry for a company which was like the first problem we had. The last problem we encountered was output errors when running the linear regression and we had to redo the code multiple times.

* 1. *Limitations of Implementation*

The main limitation to the dataset and the model is how the model predicts a lot of false positives. Our dataset is also limited due to the sample size and the number of variables available for testing. If the dataset had more variety and more data would allow for the model to have better and more accurate results. Between the logistic and linear regression models the linear had better and more accurate results and gave better insight than the logistic regression.

* 1. *Improvements/Future Work*

Areas to improve and to allow for better use for future work are things like resampling techniques for over sampling for more results. As well as adding more variables and collecting more data for more validation from both models. As well as managing outliers and evaluating more data, all these things will allow for better results for future work and improvements for the models.

1. **CONCLUSION**

After going through this whole project, it was a good project, and it allowed Garrett and I to learn a lot and use our skills. We were able to gain insight into EDA and logistic and linear regression and how the models work. After getting a data set, we were unfamiliar with and developing some EDA and training and evaluating our data we were able to develop a conclusion. Overall, we can conclude that for the most part and majority of the time we are able to predict a Data Scientist Salary. Using the variables provided from the data as well as using the models we can also see that there were many variables that are related to the salary. The model was good, but it most defiantly can still use some work and improvements. Lastly, this was a great project that allowed us to learn more, and it allowed us to look at variables and use models to answer our question. Which we were able to with a good model that yes, we can predict a data scientist salary.

**REFERENCES**

*https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor*