Data Science Project Report

Salary Prediction using Regression Analysis

May 7, 2019

Professor	Dr Atif Tahir
Project Member 1	Hamza Mustafa Khan (K15-2832)
Project Member 2	Sarim Balkhi (K15-2828)
Submission Date	May 7, 2019

Task	Performed By
Setting Research Goal	Member 2
Retrieving Data	Member 1
Data Preparation	Both
Data Exploration	Both
Data Modeling	Both
Data Presentation	both

Contents

1 Setting Research Goal

Nowadays, everyone is searching for jobs using several online job sites. These sites include Indeed, Rozee, etc. Several job postings on these sites have no salary associated with them as they are undecided and the job seekers hesitate to apply for them [1]. Therefore, this project aims to predict the salaries in this scenario using Regression Analysis. Salary prediction is challenging; however, on

2 Retrieving Data

The data is retrieved through web scraping from an online job portal namely, Indeed.com. A package called Beautiful Soup, an HTML parser is used to gather the data. The data comprises of job postings with attributes namely, city, job title, company name, location, summary and salary. This data will help in predicting the salary of the job postings.

3 Data Preparation

Data is loaded into a Pandas dataframe and is cleaned. The cleaning process includes certain processes like if there is any integer in the name of location, its replaced with an empty string, postings with salary = 0 have been dropped, and duplicates have also been dropped. One hot encoding and label encoding is applied so as to convert the data from categorical type to numerical type in order to perform analysis efficiently. By using the existing attributes, another attribute ”Higher Position” is created which indicates if the position belongs ”Senior”, ”Head” or ”Lead” category.

4 Data Exploration

This data is explored using descriptive analysis like head, shape, etc. Skewness and Kurtosis is determined. When Salary was plotted, it was discovered that the it is not normally distributed. Several kinds of plots of salary have been plotted like distplot, probplot, scatter plot etc. A heatmap of the data is also generated.

Moreover, top words in job title involve data scientist, analyst, engineer etc. Also, the probability curve showed that regression can be applied. Besides, the box plot

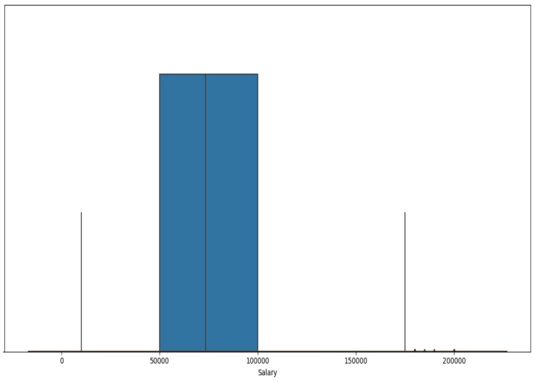

Figure 1: Box plot of Salary

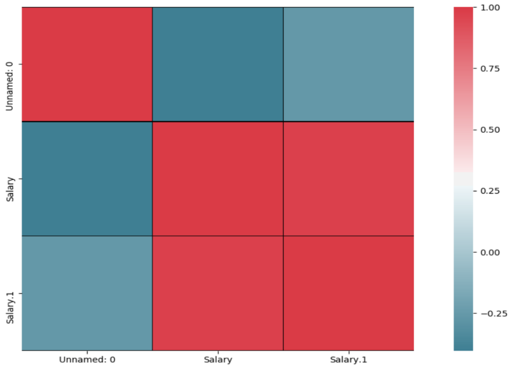

Figure 2: Heat map

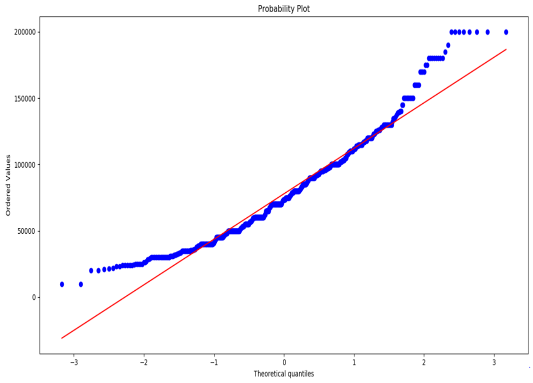

Figure 3: Probability plot

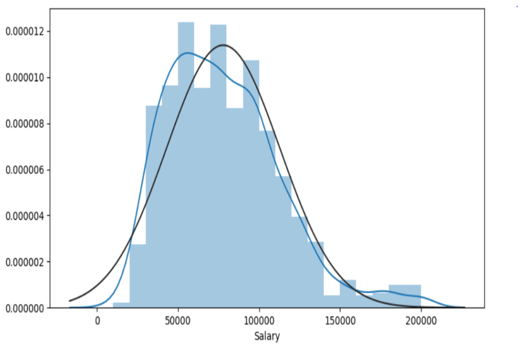

Figure 4: Salary Distribution

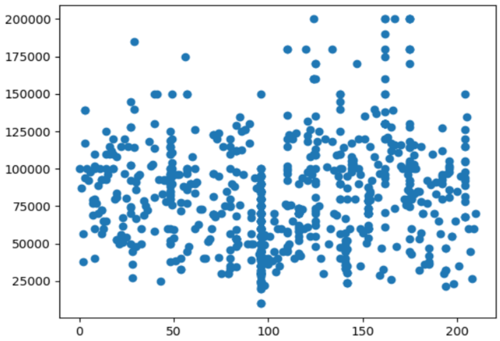

Figure 5: Scatter plot

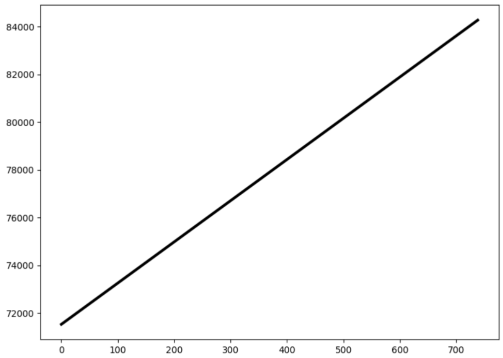

Figure 6: Regression Line

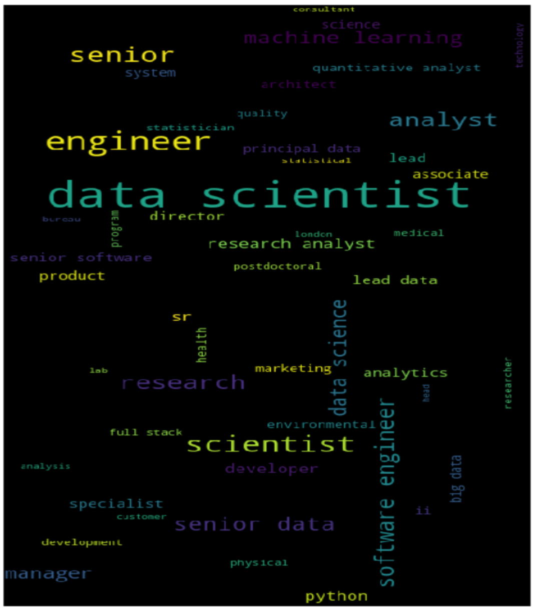

Figure 7: This cloud indicates the most frequent job titles.

5 Data Modeling

Linear Regression has been applied because our target data is continuous and the graph of probability plot indicated that the regression is linear. Hence, the classifier works best in the scenario.

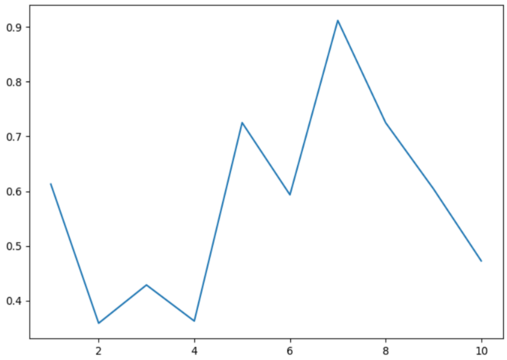

Figure 8: Cross Validation

6 Presentation and Automation

This data is composed of job postings and is presented with their salaries. The WordCloud generated shows that ”data scientist” is the most frequently occurring job title in the extracted postings.

7 References

[1] McCulley, W. L., Downey, R. G. (1993). Salary compression in faculty salaries: Identification of a suppressor effect. Educational and Psychological Measurement, 53(1), 79-86.