**PROJECT 4 REPORT**

**Glassdoor Jobs (2017-18)**

**Salary Prediction Models**

**Introduction**

Data from Glassdoor.com for 2 years (2017-18) has ~750 job postings in the US. Using the salaries listed for jobs, various models are tested and implemented to predict salaries. From data exploration, it becomes apparent that higher salaries are in tech sectors, especially for ‘data’-related jobs.

Therefore, such jobs and corresponding salaries form the basis of modeling for salary-projection in this study.

**PART I**

**Visual Understanding of Data**

Glassdoor data was obtained from Kaggle.com:

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

File name: eda\_data.csv

In Jupyter Notebook, a Python file ‘glassdoor\_Data’ was created to ‘clean’ the dataset.

The cleaned data has 742 records (job postings), and 16 features, *viz.*,

1. 'job\_state',
2. 'job\_region',
3. 'Sector',
4. 'Avg\_Company\_Revenue',
5. 'Company\_Age',
6. 'Avg\_Company\_Size',
7. 'Job\_Title',
8. 'Avg\_Salary',
9. 'Max\_Salary',
10. 'Min\_Salary'
11. 'Usr\_Rating',
12. 'Python\_y\_n',
13. 'R\_y\_n',
14. 'Spark\_y\_n',
15. 'AWS\_y\_n',
16. 'Excel\_y\_n'

An SQLite database, ‘glassdoor\_jobs.db’ is created to include a table named ‘glassdoor’, containing 16 features and all the data in the clean dataset. The SQLite file is accessed in Jupyter notebook for analysis using Python.

**An overview of job prospects around the country (USA) in graphs:**

Figure 1a shows a map of 5 geographical regions (courtesy National Geographic).

Figure 1a

A map of the united states

Description automatically generated

The following is a list of regions and corresponding state codes of states that belong to each region:

* 'West': WA, MT, OR, ID, WY, CA, NV, UT, CO, AK, HI,
* 'Southwest': AZ, NM, TX, OK,
* 'Midwest': ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH,
* 'Southeast': KY, WV, VA, AR, TN, NC, SC, GA, AL, MS, LA, FL,
* 'Northeast': PA, NY, VT, ME, NH, MA, RI, CT, NJ, DE, MD, DC

Figure 1b is a pie chart with average salary for each region listed in the legend. Average of salaries in West is ~25% of averages of salaries of all regions combined.

The chart shows that the earning capacity in ‘West’ exceeds that of the rest of the regions. Each region, however, may include states with very high paying jobs along with states with low paying jobs.

Figure 1b

A colorful pie chart with numbers and text with Crust in the background

Description automatically generated

There are 24 **sectors** listed in the dataset.

Figure 2 is a bar chart that shows the maximum and minimum salaries averaged for each sector. The top 4 maximum salary bars show large differences between maximum and minimum averages of salaries. Information Technology and Biotech & Pharmaceuticals are among the large income sectors. The last 2 bars show the average minimum as greater than average maximum, indicating that taking the average, rather than the differences between maximum and minimum, is more insightful than the maximum and minimum salaries.

Figure 2

A graph of a number of people

Description automatically generated with medium confidence

Figure 3 follows Figure 2 in showing the number of jobs available in each sector.

‘Information Technology’ tops the list followed by ‘Biotech & Pharmaceuticals’. Both of these sectors are in the top 4, both in terms of job numbers and salary. ‘Agriculture & Forestry’ has the least number of jobs.

Figure 3

A graph with green and orange bars

Description automatically generated

Figure 4 combines Figure 3 and Figure 4, to create a bubble chart to assess the number of jobs and the average salary for a sector at a glance. The height of the bubble from the x-axis shows the salary and the size of the bubble represents the number of jobs posted for a particular sector.

The chart clearly shows that ‘Information Technology’ (IT) and ‘Biotech & Pharmaceuticals’ (Biotech) are two of the highest and largest bubbles.

Figure 4

A graph showing different colored circles

Description automatically generated

A similar exploration was performed on Job Titles. There are 7 **Job Titles**.

Figure 5 is a violin plot, showing the salary distribution for each title. For the data-related job titles: ‘Data Scientist’, ‘Data Analyst’, ‘Data Engineer’, ‘Machine Learning Engineer’, there are a large number of outliers toward higher salaries.

Therefore, the Logistic Modeling for salary prediction is based on the salaries of these job titles.

Figure 5

A graph of different colored diamonds

Description automatically generated with medium confidence

Figure 6 is a bubble chart that shows the average salary for each job title. The size of each bubble is the number of jobs for each title. The largest number of jobs are for title ‘Data Scientist’.

Figure 6

A graph with blue and green circles

Description automatically generated

Figure 7 is a bar chart that is a combination of job numbers, sectors and job titles. Figure 7 shows that the number of job types like ‘Data Scientist’, ‘Data Analyst’ and ‘Data Engineer’ are available in most sectors, the greatest numbers being in ‘IT’ (gray). ‘Research Scientist’ is most needed in ‘Biotech’ (violet) sector. Title ‘Machine Learning Engineer’ (MLE) is available in sectors, ‘Education’, ‘IT’, ‘Aerospace & Defense’ and ‘Finance’.

Figure 7

A graph showing different colored lines

Description automatically generated

Figure 8 Shows the Glassdoor user ratings data versus the age of the company. The mean /median is as user-rating of 3.7. Well established companies older than 100 years received user rating close to the median value of 3.7. This could be due to the large number of people employed by a company over the long period of existence of the company resulting in a large number of reviews and ratings leading to the average rating close to 3.7.

Younger companies less than 100 years old may not have a large set of review/rating data, hence the wide range of ratings ranging from 1.9 to 5, for companies less than 25 years old.

The size of the bubbles is proportionate to average salaries offered by companies. A general trend of lower ratings for lower salaries and higher ratings for higher salaries can be observed.

Figure 8

A graph of a number of people

Description automatically generated with medium confidence

**PART II**

**Linear Regression Models**

Salary distribution is shown in Figure 9. The distribution is a normal distribution skewed to the right, with outliers toward higher salaries. The distribution also shows extreme outliers (salaries > salary(mean) + 3(standard deviation)).

Figure 9

A graph of a salary distribution

Description automatically generated

**Ordinary Least Squares (OLS) model**

Since the salary histogram exhibits a normal distribution, the first model chosen to predict salaries is Ordinary Least Squares (OLS) model, which uses Linear Regression to represent salaries. The principle of OLS is to find a line that minimizes the sum of the squared differences between the **observed** and **predicted** values of the dependent variable (salary), with other features as independent variables.

Only numerical variables can be included in OLS. Therefore, all categorical variables were changed into dummy variables with Boolean values 0 or 1.

The main assumption of OLS model is that the independent variables are truly independent and show low or no collinearity with other features. The following is the correlation matrix which shows **low** collinearity between variables considered in this OLS model.

Figure 10

A close-up of a chart

Description automatically generated

Considering only the variables showing low collinearity the OLS model shows the following results:

Figure 11

A screenshot of a computer

Description automatically generated

The 2 significant numbers in this result are the R2 value and the Durbin-Watson number.

R2 value (**R2**) represents the measure of variance in the dependent variable that is explained by the independent features considered in the OLS model. The higher the variance the, the better the model for the given data. The value obtained in Figure 11 is 0.42 or 42%, is in the ‘**moderate**’ category between the extreme values 0.00 (‘extremely bad’) or 1.00 (‘extremely good’).

0.42 indicates a ‘moderate’ degree of variation in salary is explained by the independent features included in this model.

**Durbin-Watson** statistical test detects the presence of auto-correlation in the residual of OLS analysis. The number ranges between 0 and 4. The value 2 indicates no autocorrelation; the value 2.07 indicates the presence of mild negative autocorrelation, implying that a pattern in the residuals cannot be explained by the model.

The data is split into training and testing of the model in the ratio **0.75 : 0.25**, for **Training : Testing**.

The results are as follows:

**Table 1**

**Error Comparison Using OLS Linear Regression Model**

|  |  |  |
| --- | --- | --- |
|  | With Training Data | With Testing Data |
| MAE | 21.7430 | 24.0752 |
| MSE | 736.4403 | 895.6746 |
| RMSE | 27.1374 | 29.9278 |
| R2 | 41.39% | 39.81% |
| Durbin-Watson | 2.03 | 2.03 |

* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* R2 -Coefficient of Determination (R2-Square)

The testing data shows R2 = 0.398.

Possible Reasons

* The distribution of salaries in Figure 9 shows a large number of outliers creating a **bias toward the right** indicated by **kurtosis: 3.11** (Figure 11). The moderately low R2 value could be attributed to the salary distribution being skewed to the right.
* The salary distribution may not be strictly linear and may be explained by a non-linear expression.
* There may be collinearity between features, that affects the model’s performance.

Therefore, a second model, which does not require normal distribution, is also considered.

**Decision Tree Model**

Tree Regression Model does not require any specific distribution. The residuals in the result of the Tree Regression Model do not necessarily follow a normal distribution.

This is because trees can capture complex relationships in the data that a model like OLS does not.

The results are in Table 2.

**Table 2**

**Error Comparison Using Tree Regression Model**

|  |  |  |
| --- | --- | --- |
|  | With Training Data | With Testing Data |
| MAE | 9.3584 | 18.8957 |
| MSE | 250.6310 | 739.4727 |
| RMSE | 15.8313 | 27.1932 |
| R2 | 80.05% | 50.31% |

The R2 values for testing data in particular is 0.503, a significant improvement from that for the OLS model. Tree Regression model is able to capture more complex patterns between variables, leading to a higher R2 value.

The much larger value of 0.801 with training data may be due to overfitting the model for training data. This is expected with most Decision Tree models.

An advantage of using the Tree model is that the model can automatically handle interactions between features. OLS model is unable to capture non-linear relationships, while decision trees can discover such relationships during the training process.

**CONCLUSION**

Of the two regression models considered for Salary Prediction, **Tree Regression Model** is a better predictor of salaries compared to the OLS model

**PART III**

**Logistic Regression Models**

From **Part I** it can be inferred that the highest paying jobs as well as the largest number of jobs are in the sector: **Information Technology**. All ‘IT’ jobs require the knowledge of one or more of the following Programming Languages or Application Packages listed in the Glassdoor dataset:

* Python,
* R,
* Spark,
* Amazon Web Services (AWS)
* Excel

In Figure 12, each plot is a histogram of job-count versus salary for each programming skill. All plots except one for ‘R’, resemble a ‘normal’ curve skewed right. This implies that the average value of salary moves to the right and salary range is extended to the right.

The jobs requiring such skills offer much more than the median value of salary listed here:

* Spark: $108k
* Python: $107k
* AWS: $107k
* Excel: $92k
* R: $70k

Figure 12

A group of graphs showing different colored lines

Description automatically generated with medium confidence

For Logistic Regression, the threshold salary is the median value of $ 107k. Salaries ≥ $ 107k are represented by ‘1’ and salaries < $ 107k are represented by ‘0’.

For each of the three models considered here, Confusion matrix, Classification Report, and Accuracy Score are analyzed. The following is a representation of a Confusion Matrix:

A diagram of a graph

Description automatically generated with medium confidence

The goals of the models are

* High Accuracy Score,
* Low False Negatives (FN).

**Decision Tree Classification Model**

A screenshot of a graph

Description automatically generated

The Classification Report shows that Training data shows less accuracy compared to the testing data, when the contrary is expected. Testing data is ¼ the size of the training data. Since Decision Tree model is prone to overfitting the training data, accuracy should be less than that with testing data.

**Random Forest Classification Model**

The Random Forest (RF) model uses an ensemble learning method that builds **multiple decision trees** and merges their predictions to improve overall performance. Random subsets (bootstrap samples) of the data is used to generate each tree, introducing diversity.

RF is a more robust model compared to Decision Tree and performs well when there are complex interactions between features, as might be the case in the present dataset.

The subsets of data used to create multiple trees can detect biases, that are not detected in the Decision Tree model leading to high variance. RF model attempts to reduce variance while taking biases into consideration.

The following are the results of RF modeling:

A screenshot of a graph

Description automatically generated

The Accuracy Score with Training data is 0.96, whereas with Testing data it is 0.79. This result is to be expected, as the model is fit to the training data.

The ratio of FN/FP = 0.92 for training data and 0.77 for testing data. This indicates lowering the FN value compared to FP, which is desirable.

**LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno)**

Although, it is not a classification model, it is often used as an optimization algorithm to train classification models by minimizing their ‘loss function’ during the training process.

In the context of the given data, the LBFGS algorithm is used for optimizing the parameters of logistic regression.

A screenshot of a graph

Description automatically generated

The results show that Accuracy score is 0.77 for training and 0.67 for testing.

These values are lower than the RF model results.

Secondly, FN/FP = 1.35 >1 for training data and FN/FP = 1.03 > 1 for testing data as well.

Both factors above indicate that **‘LBFGS’ is not suitable** for the present data.

**CONCLUSION**

**Table 3**

**Comparison Between Logistic Models**

A table with numbers and text

Description automatically generated with medium confidence

Table 3 illustrates the performance of each classification model/algorithm.

The model that satisfies both objectives viz., high accuracy and low FN number, is the Random Tree Classification Model. Therefore, the RF model is the best among the three to classify salaries for data similar to the given Glassdoor data.