SALARY PREDICTION SYSTEM

FOR INFORMATION TECHNOLOGY PROFESSIONALS

TERM PROJECT – ADVANCED PYTHON AI & ML TOOLS (AML 2203); SECTION 3

**SUPERVISED BY: DR. VAHID HADAVI, PhD.**

**PREPARED BY:**

|  |  |  |  |
| --- | --- | --- | --- |
| KATLEEN EZEKEIL ORATA  ID: C0848019 | TOM MADONIL JAMES  ID: c0850503 | MOHAMMAD IMRAN UDDIN  ID: C0800487 | HARDIK SOLANKI  ID: c0852302 |

# RATIONANLE

The Information Technology Industry grew rapidly in the past few decades, and we have noticed people from all over the world, especially the Indian subcontinent pursuing advanced degrees in the discipline and finding highly rewarding jobs. In recent years, we have noticed a rapid growth in the sector which has added more domains like database engineering, android/iOS development, machine learning, etc.

Because of this exponential growth of the IT industry, it is obvious for the current youth to be inquisitive as to what specific field in IT they should pursue, which companies they should aim for, and what cities they should try to move into. Therefore, our ambition for this project is to apply the tools and techniques we learned in this course to the proposed dataset and gather insights that would help the aspiring IT professional to make an informed decision based on real data

# Data Description

Table

Description automatically generated with medium confidence

Figure : Dataset Overview

After importing the dataset into a dataframe format using pandas read\_csv function, we used the dataframe.head function to get the overview of the first 5 instances in the dataset (see Figure 1 above). It is obvious to us that some of features are numerical while other categorical. However, we instead of eyeballing the characteristics of features, we used the pandas’ info function to identify the datatype of each feature.

Graphical user interface, application

Description automatically generated

Figure : Data Type

We can gather some useful characteristics from the output in Figure 2 above. Firstly, none of the features have any null or missing values which is always desirable in any dataset. Secondly, we can observe that none of the features has any datatype that is counterintuitive. Company Name, Job Title, Location, Employment Status and Job Roles are all saved as object (string) data types. Similarly, Rating, Salary and Salary Reported are saved as either float or integers which is understandable given their numeric nature.

We can also see that the dataset comprises more than 22,770 observations collected from software professionals in various companies and places in India (BANERJEE, 2022). The following is a brief description of the features:

|  |  |  |  |
| --- | --- | --- | --- |
| **Serial** | **Feature** | **Data Type** | **Description** |
| 1. | Rating | Float | This field shows ratings provided to companies by employees. The ratings range between 1 (the lowest) and 5 (the highest) |
| 2. | Company Name | Object | This field states the name of the company |
| 3. | Job Title | Object | This shows the title of the job held by the professional |
| 4. | Salary | Integer | This field shows the monthly salaries reported by the professionals |
| 5. | Salary Reported | Integer | This field shows the number of professionals who reported their salary level from a specific company |
| 6. | Location | Object | This field provides the names of the cities where the company is located |
| 7. | Employment Status | Object | This field provides information on the type of contractual agreement between the professional and the company he works for. It describes whether the professional is working as a full-time employee, intern, contractor or just a trainee. |
| 8. | Job Roles | Object | This field shows the specific domains the professional is working in, which include Android, backend, database, frontend, IOS, Java, Python, etc. |

Table : Feature Description

# Exploratory Data Analysis

Graphical user interface

Description automatically generated with low confidence

Figure : Basis Descriptive Statistics

At this stage, we attempted to develop a deeper understanding of each of the features in terms of the effect they may have on further analysis. As such, we applied descriptive statistics on the features. We used the pandas’ describe function to apply descriptive statistics on the features. The results are presented in Figure 3 above. Let us analyse the results for each of the features in some more details.

The range of the feature Rating is between 1.0 and 5.0 as can be seen in Figure 3. Both the mean and median value appears to be quite similar around 3.9, but they both tend to veer a little towards the higher end of the values. This indicates that the distribution of Rating maybe a bit skewed with the possibility of some outliers.

The feature Company Name is a categorial one and comprise 11,261 unique company names. The company Tata Consultancy Services appears to have the largest number of observations, 271, to be precise. Similarly, Job Title is also categorical with 1080 unique instances. The title Software Developer Engineer tends to appear most frequency in the dataset.

The Salary feature is of utmost interest to us because in the proposal we suggested creating regression model that will attempt at predicting salaries based on other features in the dataset. Figure 3 shows that salaries vary vastly across the professionals with the minimum being 2,112.0 (perhaps a stipend being paid to an intern or trainee) and the maximum being 90 million. Since the mean salary is nearly 700,000, we suspect the feature is highly skewed to the right and there are massive outliers.

Salary Reported also shows that there are vast differences in the number of salaries being reported by companies. The minimum shows that from some companies, just one employee has reported the salary. On the other hand, the maximum shows that in some company/companies, more than 350 employees have reported their salary information. As such, when it comes to representation of companies in the dataset, there is definitely a bias. One should note that this bias is likely to have a bearing on the company rating discussed above. Companies that have been rated by more employees would reflect their ratings better than those with fewer employees providing rating.

The location features have 10 unique values that represent the names of cities in India from where the data has been reported. We notice that Bangalore has the highest frequency of data which is quite conceivable since the city has been known the Silicon Valley of India (Canton, 2012).

The Employment status has 4 unique values, which we will explore in the subsequent sections when we conduct univariate analyses. At this moment, we can observe that the employee with a full-time status represent the largest fraction of the dataset.

Lastly, the feature Job Roles appear to have 11 unique values and the most frequent of those is SDE (Software Development Engineer). It is interesting to note that this is consistent with the feature Job Title, which we discussed earlier. This indicates that there maybe some overlap across both features. In most cases, the title and role of employees generally imply the same thing.

## Univariate Exploration

At this stage we attempted to look at some of the features individually to make greater sense of their distribution. We believe this analysis will help us understand what kind of pre-processing/cleaning the feature may need to fit the model better.

Graphical user interface

Description automatically generated

Figure : Distribution of Rating

As we can see from the boxplot and histogram in Figure 4, the distribution of Rating appears to be skewed towards the left (negatively skewed) and there are some outliers towards the lower data values of the distribution.

|  |  |
| --- | --- |
|  |  |

Figure : Distribution of Salary

From Figure 5, we can observe that the distribution of Salary is highly skewed to the right (positively skewed). Similar deductions can be made from the boxplot in Figure 5. While discussion the features under Exploratory Data Analysis above, we did note that the maximum salary value is 90 million, which has very little meaning considering the data is collected from India where the possibility of someone receiving such remuneration is extremely rare.

|  |  |
| --- | --- |
|  |  |

Figure : Distribution of Salary Reported

Figure 6 (see above) portrays the distribution of Salary Reported in histogram and boxplot format. As discussed under EDA, there are massively differences in the number of people reporting their salary from their companies. As such, the data is massively right-tailed (positively skewed), with quite a bit of outliers towards the higher end of the data values.

Chart

Description automatically generated

Figure : Location

Figure 7 portrays the frequency by location (city names) for the dataset. As we can see, most software professionals have reported their salaries from Bangalore, which is consistent with our earlier assumption of the city being the Silicon Valley of India. Hyderabad and New Delhi are significantly behind Bangalore but still represent a sizable fraction of software professionals in the regions. On the other hand, Jaipur and Kerala represents a very small fraction of all the software professionals in the dataset.

|  |  |
| --- | --- |
|  |  |

Figure : Employment Status and Job Roles

Figure 8 depicts the compositions of software professionals by their Employment Status and Job Roles. As we can see the largest majority of professionals work as fulltime workers. The proportion of interns, contractors and trainees are significantly lower than full-time employees. Considering their Job roles, we observe that most of the professionals work as software development engineers, which is followed by−albeit in much smaller proportions−Android developers, frontend developers, Java developers and Testers.

## Bivariate exploration

At this stage, we intended to explore relationships across features to identify any association with them. The purpose behind exploring correlation is to avoid multicollinearity across the features. It is worth mentioning at this stage is that the dataset originally has three numerical features expressed as either integer or float values, and five categorical features expressed as strings. We intended to use the dataframe.corr function to identify any possible correlation across numerical variables.

|  |  |
| --- | --- |
| Chart  Description automatically generated |  |

Figure : Correlations

As we can see at the left of Figure 9 above, there seems to be no significant correlation among the numerical variables.

|  |  |
| --- | --- |
|  |  |

Figure : Mean Salaries by Job Roles and Employment Status

To develop a better understanding of Salary relationship with Job Roles or Employment Status, we used the pandas’ groupby function and chained it with pandas’ mean function to see how average salary varies across various Job Roles and Employment Statuses. The graph on the lefthand side of Figure 10 shows average salaries by Job Roles. We can see that the Database developers make the highest average salaries followed by SDE (Software Development Engineers) and Mobile Developers. The web developers seem to make the lowest. The figure on the righthand side shows the average salary by Employment Status. Not surprisingly, the fulltime employees make the highest average salary followed by contractors. The interns and trainees seem to be making pretty much the same amount of average monthly salary.

Chart, bar chart

Description automatically generated

Figure : Mean Salaries by Locations

Chart, pie chart

Description automatically generatedFigure 11 portrays the average salaries by locations (or cities). It is interesting to observe that although Bangalore is the Silicon Valley of India, IT professionals in Mumbai makes more on average. The Silicon Valley city is ranked at 4th place when it comes to average salaries.

Figure : Salary Reported by Location & Job Roles

Chart, pie chart

Description automatically generatedLastly, we were curios with regards to the feature Salary Reported. So, we tried to see the distribution of Salary Reported by Location to see if there are any association between these two features.

Although the results in Figure 12 (left) shows that Bangalore reports the highest proportion of salaries. There are not major differences in the remaining cities. We should also consider the fact that Bangalore has highest number of jobs, hence it makes sense to have more salary reports from the location.

We also tried to look at the association between the features “Salary Reported” and “Job Roles”. We found, in Figure 12 (right) that SDE (Software Development Engineers) tend to report their salaries most often, followed by Testers. This is consistent with our findings in Figure 8, where most jobs are occupied by SDEs.

# Putting it all together

So far, we have explored the features individually and assessed at their associations with other features. We learned that there is no correlation among the numerical features, but all the numerical features have outliers that must be handled. In the following stages, we will be pre-processing the data before attempting to feed it into a regression model and test its predictability.

# Data pre-processing

## Outlier Detection and Removal

As discussed in Figure 5, the “Salary” feature had massive outliers. Therefore, we applied the quantile-based method to drop the outliers.

Graphical user interface, text, application

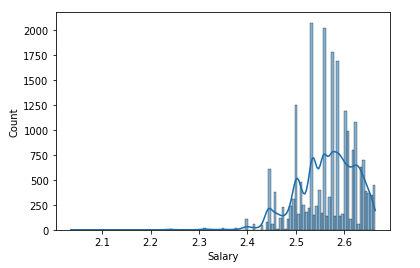
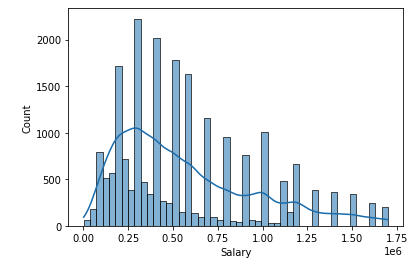
Description automatically generated

Figure 13: Treatment of Outliers

As we can see in Figure 13, we have calculated the Interquartile Range (IQR) for the Salary feature and then created the upper and lower whisker by adding and subtraction 1.5 times of IQR to 75th and 25th percentile. The figure shows that after removing the outliers, the length of the data reduced from 22,770 to 21,459.

## log transformation

Figure : Before and After Log Transformation of Salary



In Figure 14 (left) we noted that even after removing the outliers, the Salary feature was still skewed, so we applied log-transformation to improve the distribution a bit and ensure that the regression model is not affected. The salary distribution after log-transformation is depicted in Figure 14 (right) which shows a bit better distribution than before.

## Data cleaning

We realized that the salaries reporting feature does not add any value to the model because we have individual salaries from the professionals. Therefore, we dropped the feature from further analysis.

# Data Encodiong

Machine Learning models, especially regression models, cannot work directly with text data. As such, we had to encode the data to transform them into some form of meaningful numerical features that linear models can use. To accomplish this goal, we used the Target Encoder Module which expresses categorical features not in terms of 1 and 0 but in terms of a probability values calculated on the basis of the target variable.

Table

Description automatically generated

Figure 15: Encoding Categorical Variables

As we can see in Figure 15 above, all the categorical variables now appear as numerical values.

# Feature Enriching

We were unable to get a good score on the first attempt. Therefore, we tried to enrich the features by creating categories. We applied the K Means clustering method to create clusters.

Chart, line chart

Description automatically generated

Figure 16: Application of KMeans Clustering

From Figure 16, we can see that the optimal number of clusters is 4. Therefore, in the next stage we created an additional feature “groups” and appended that to the dataset.

Table

Description automatically generated

Figure 17: Enriched Dataset

In Figure 17, we can see that the revised dataset has an additional feature called “group”.

# Model Development

As narrated in the project proposal, we intended to establish a linear relationship between the features to predict the amount of salary. As such, we assigned “Salary” as the target and remaining features as independent variables. We used the train-test-split module to divide the dataset into training set (80%) and test set (20%). We also assigned random state to 4, to ensure reproducibility of the split.

# Model Evaluation

Graphical user interface, text, application

Description automatically generated

Figure 18: Model Evaluation

We used the R-square metrics to measure the accuracy of the model. The R-square score is 0.73, which can be interpreted as the model being able to predict 73% of the variation in salaries of software professionals in India.

# Conclusion and Learning

This project gave us the opportunity to work with real life dataset. We learned that real life datasets are often very messy and includes massive outliers, as we encountered in this exercise. We also noticed that real life data may have features that do not add value to certain analysis. As such we used to this opportunity to gain some domain experience and alter the data accordingly. As discussed in the lecture, treatment of outliers does not have a standard technique, so data practitioners are required to address this issue with several iterations of trials and errors. We had to treat the outliers in this case in two stages, first by applying the quantile-based method and secondly by using log transformation. We also learned that even after doing everything right, the model may not perform as expected. To this end, we applied the K Means clustering method to enrich the data by creating clusters. We noticed that after clustering the model is performing sufficient well with a R-square value of 73%.