University of
Strathclyde
Glasgow

Engineering Graduate Salary Prediction

CS 982: Big Data Technologies

Name: Raj Biswas

Contents

List of figures

Chapter 1

Introduction

India is the second-largest populous country having a population of about 1.355 billion. Among this huge number, about 2.451 crores people graduate each year which includes streams from arts, science, engineering, commerce, medical, management, and law. Most of the students apply for jobs after their graduation directly or indirectly to the company. It becomes a time-consuming and difficult task for the companies to find potential candidates and hire them as there is a chance that many of the applications from the students can be almost the same. So, in order to solve this problem, a company named Aspiring Minds from India organizes a test called AMCAT every month.

AMCAT, which is also known as Aspiring Minds Computer Adaptive Test is an employability test to assess new graduate students based on various skills. Companies filter out students based on their AMCAT marks which help them to hire the best potential candidate for the job role as well as saves time. About seventy thousand job seekers appear for the AMCAT test every month. Most of the companies in India, from large to medium level companies uses AMCAT scores for hiring.

The test carried out with two modules - a compulsory module and an optional module. The compulsory module includes English comprehension, quantitative ability, reasoning, and logical ability which test a candidate's aptitude and personality skills. Optional module tests skillset as per the job perspective.

Chapter 2

Dataset - Engineering Graduate Salary Prediction

Aim:

Analyze the dataset and predict the salary for a candidate by providing marks and other required data.

Source:

The dataset is provided by the Aspiring Minds Research team and it is freely available on Kaggle(https://www.kaggle.com/). It contains data for graduate students from the engineering stream. The dataset consists of and 34 columns 2999 rows including the column heading. The dataset contains the following information about the students:

- ID: A unique ID to identify a candidate
- Gender: Gender of the candidate
- DOB: Date of birth of the candidate
- 10percentage: Overall marks(percentage) obtained in 10th grade examinations
- 10board: The school board whose curriculum the candidate followed in grade 10
- 12graduation: Year of graduation - high school
- 12percentage: Overall marks(percentage) obtained in grade 12 examinations
- 12board: The school board whose curriculum the candidate followed
- CollegeID: Unique ID identifying the university/college which the candidate attended for his/her undergraduate
- CollegeTier: Each college has been annotated as 1 or 2. Colleges with an average score above a threshold are tagged as 1 and others as 2.
- Degree: Degree obtained/pursued by the candidate
- Specialization: Specialization pursued by the candidate
- CollegeGPA: Aggregate GPA(percentage) at graduation
- CollegeCityID: A unique ID to identify the city in which the college is located in.
- CollegeCityTier: The tier of the city in which the college is located in. This is annotated based on the population of the cities.
- CollegeState: Name of the state in which the college is located
- GraduationYear: Year of graduation (Bachelor's degree).
- English: Scores in the AMCAT English section(compulsory module).
- Logical: Score in AMCAT Logical ability section(compulsory module).
- Quant: Score in AMCAT's Quantitative ability section(compulsory module).
- Domain: Scores in AMCAT's domain module(optional module)

- ComputerProgramming: Score in AMCAT's Computer programming section(optional module)
- ElectronicsAndSemicon: Score in AMCAT's Electronics & Semiconductor Engineering section(optional module).
- ComputerScience: Score in AMCAT's Computer Science section(optional module).
- MechanicalEngg: Score in AMCAT's Mechanical Engineering section(optional module).
- ElectricalEngg: Score in AMCAT's Electrical Engineering section(optional module).
- TelecomEngg: Score in AMCAT's Telecommunication Engineering section(optional module).
- CivilEngg: Score in AMCAT's Civil Engineering section(optional module).
- conscientiousness: Scores in one of the sections of AMCAT's personality test(compulsory module).
- agreeableness: Scores in one of the sections of AMCAT's personality test(compulsory module).
- extraversion: Scores in one of the sections of AMCAT's personality test(compulsory module).
- nueroticism: Scores in one of the sections of AMCAT's personality test(compulsory module).
- openesstoexperience: Scores in one of the sections of AMCAT's personality test(compulsory module).
- Salary: Annual salary offered to the candidate (in INR)

Chapter 3

Analysis of the dataset:

The dataset contains some columns which are least essential to do our analysis. For this reason, some nonessential columns are dropped which are listed below with the reasons for dropping them.

ID, DOB, 10board, 12board, CollegeID: These columns are not required for analyzing and predicting the salary of a candidate.
CollegeCityID, CollegeCityTier, CollegeState: CollegeCityID, CollegeCityTier, CollegeState are provided but the dataset does not provide any information about how many companies are present in that state/city or the hiring percentage of candidates in that state/city.
12graduation, GraduationYear: As AMCAT examination date is not provided, 12graduation and GraduationYear cannot correlate and compare with the Salary field.

In order to prevent errors and wrong outputs, the dataset is been cleaned up before performing any analysis techniques. Columns that have string data are converted into numeric data to get a better result by analyzing the dataset. Three columns of data are converted to a numeric value which is “Gender”, “Degree” and “Specialization”. Also, the missing data are replaced with the mean value of the column.

The dataset contains data of 2282 male candidates and 716 female candidates. Though the date of AMCAT examination is not provided in the dataset, figure 3.1 shows that it contains 76% of males and 24% of female candidates appeared for the examination.

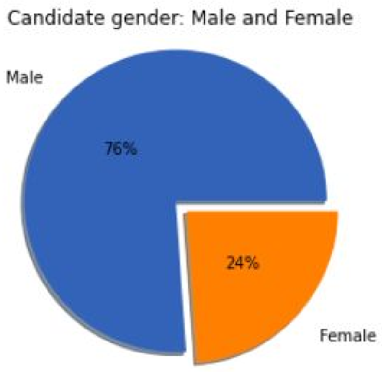


Figure 3.1: Percentage of male and female

Although the percentage of males is much greater than females, the average salary package for both genders is found almost close to each other. The average annual salary of males and females is ₹309804 and ₹290418 respectively which have a 6.45% difference. The below bar graph(see fig 3.2) is set up to get a better understanding.

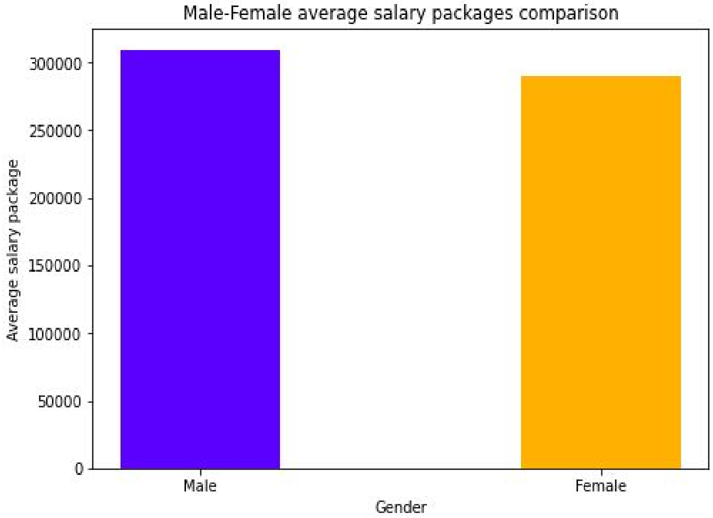


Figure 3.2: Male-Female average salary package

At the same time, it is also interesting to look at the 10th, 12th standard marks, and university/college GPA counts and their salary package offered to them.

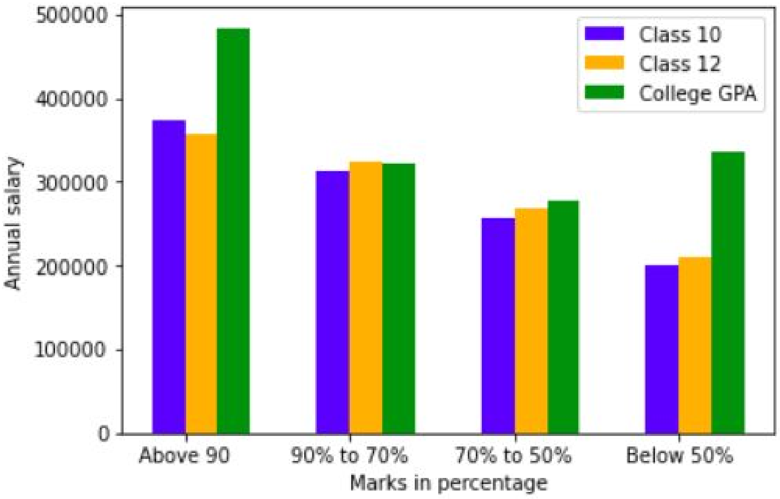


Figure 3.3: Class 10-12, College GPA marks counts and their salary package

From the above figure(see fig 3.3) it can be said that the candidates who have got marks above 90% in their class 10 and 12 examinations, got a higher salary package, and candidates who have got below 50% marks were offered the least salary package. It can also be seen from the above figure that candidates who have got more than 90% marks in their class 10 examination have received a higher salary range if compared with candidate got above 90% in their class 12 examination. But, it is totally opposite for the candidate below got 90% marks in their class 10 and 12 examinations. While the salary package is almost similar if compared with class 10 and 12 marks, there is a lot of difference in the salary package if it is compared with a college GPA for above 90% and below 50% marks candidates which means that the candidate's college/university GPA is preferred more than class 10 and 12 marks i most of the cases.

Accordingly, to analyze more on college GPA it will thought-provoking to look at the college tier. From figure 3.4 it states that the number of candidates in tier 2 college is much higher than the number of candidates from tier 1 colleges which have about 169.8% difference.

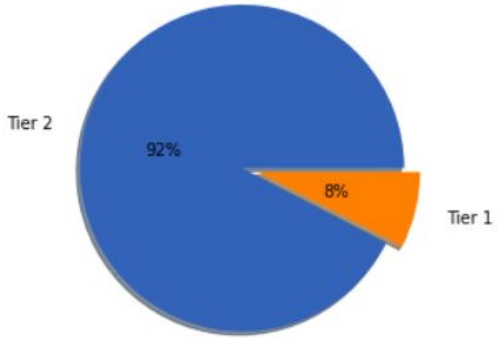


Figure 3.4: Number of students based on college tier

Although the percentage of the candidate belongs to tier 1 college is much lower than tier 2 college, the salary package of the tier 1 college candidates is greater than tier 2 candidates which can say that the tier 1 colleges are better compared to tier 2 colleges. But for the insufficient data, it can not predict which fields/areas of college from tier 1 are better than tier 2. The average salary package comparison of tier 1 and tier 2 college can be seen in figure 3.5

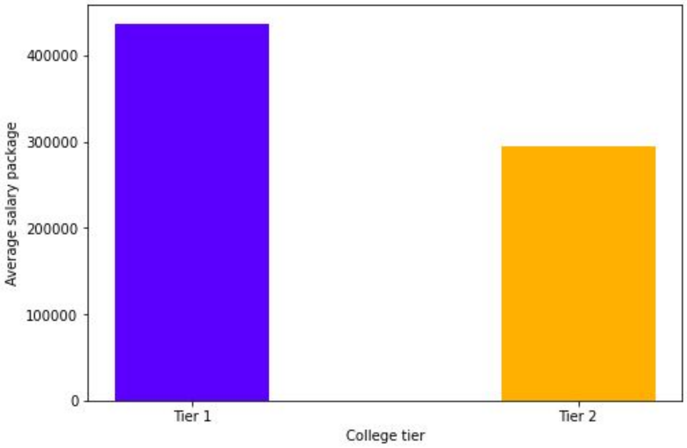


Figure 3.5: Average salary package for college tier 1 and 2 candidates

Apart from class 10,12 standard scores and college GPA, the most important factor which can influence salary is the AMCAT module scores. As AMCAT consists of compulsory and optional modules from where the candidate can take multiple or single module, it is better to proceed with the compulsory module and analyze it.

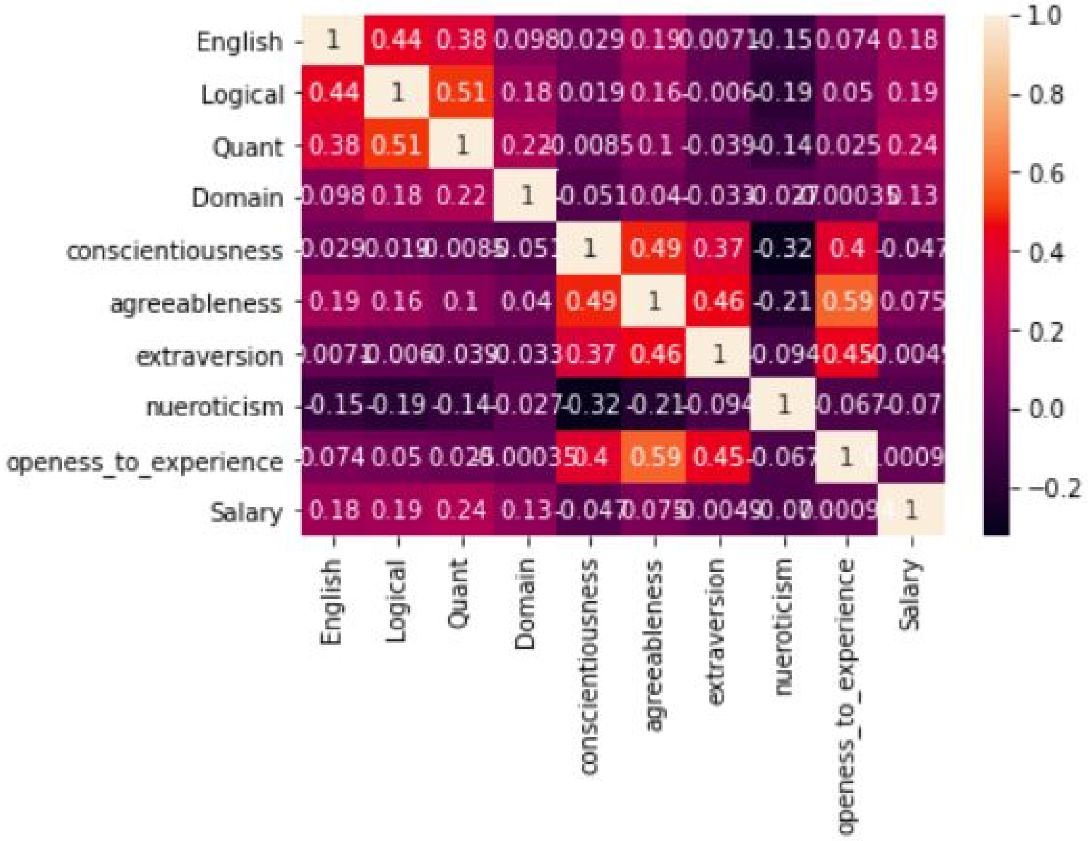


Figure 3.6: Heatmap of all AMCAT modules with Salary

The heatmap(see fig 3.6) indicates that Quant, Logical and English module are the top three modules which are strongly related to the Salary if compared with other modules. So it can be said that there there is a chance of getting a higher salary package if the candidate scores good marks in these three modules. But it is not guaranteed that performing well in these three modules can lead to a better salary package as there may be the chance where the candidate perform very bad on other modules which can decrease the chance of getting the higher package as other modules, though weakly but correlated with the salary.

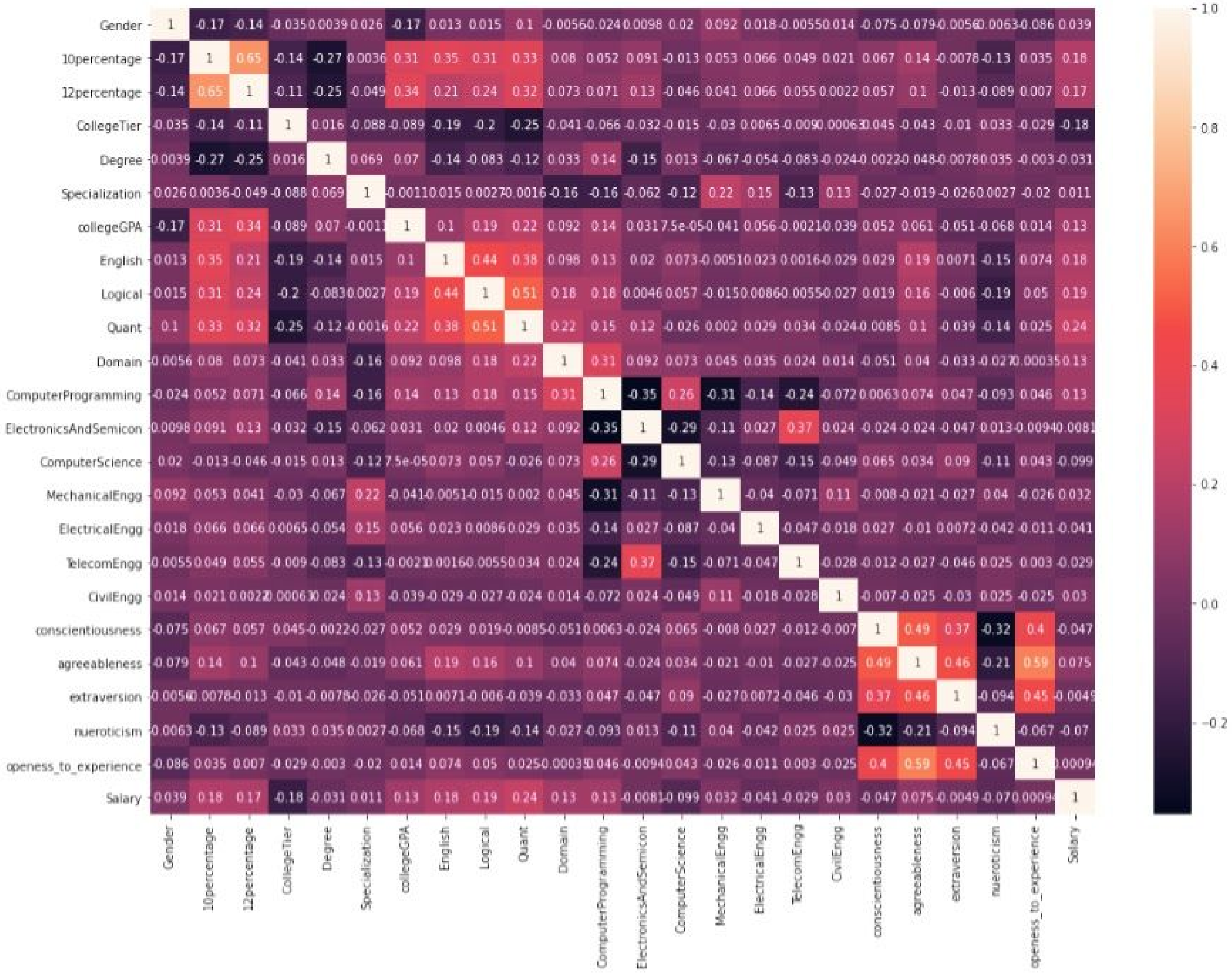


Figure 3.7: Heatmap of all variables

From the above heatmap(see fig 3.7) some observations can be made about the factors correlating with the annual salary package offered to a candidate. Quant, Logical from AMCAT modules along with 10th and 12th standard marks may be preferred first than the other modules marks by most of the companies while hiring a candidate. It can also be seen from the above analysis that college GPA percentage can be more preferred if it is compared with the marks obtained in personality test modules that are conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience.

Chapter 4

4.1 Unsupervised Analysis

Clustering analysis using k-means:

Another, significant factor in exploring how much AMCAT modules, that are Quantitive and Logical ability is dependent on the Salary of a candidate. For this, values of Quant and Domain are first scaled so that we get a proper result from our analysis. The scaled values are used in k-means and clustering is formed between 2 to 30 clusters. The performance is stored in variables for each time and the graph(see fig 4.1) is plotted in order to visualize the performance of k-mean clustering with our data. While iterating through the loop inertia is also calculated which is the sum of squares of all value points to their closest centroid.

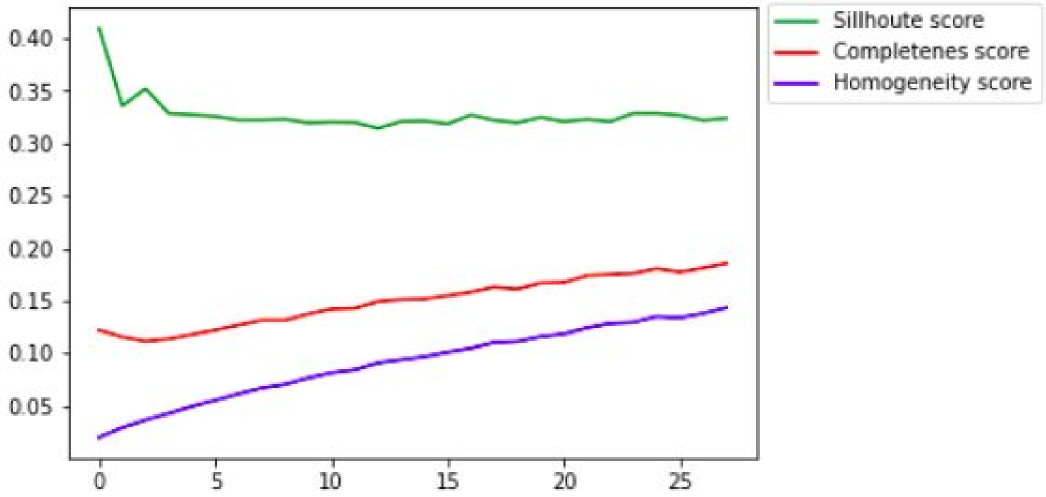


Figure 4.1: Silhouette, Completeness, and Homogeneity score

In order to choose the optimal value of the cluster in which the data may be clustered, the elbow method is used. From figure 4.2, the elbow which is the point at which inertia started decreasing can be said 6. Therefore, we can conclude that the most appropriate number for the cluster of data is 6.

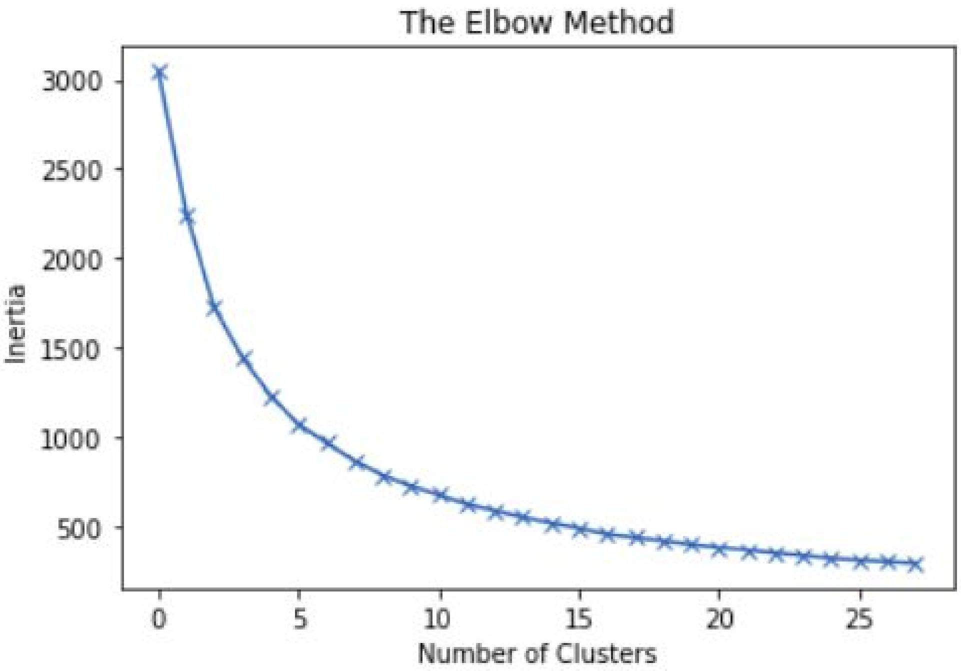
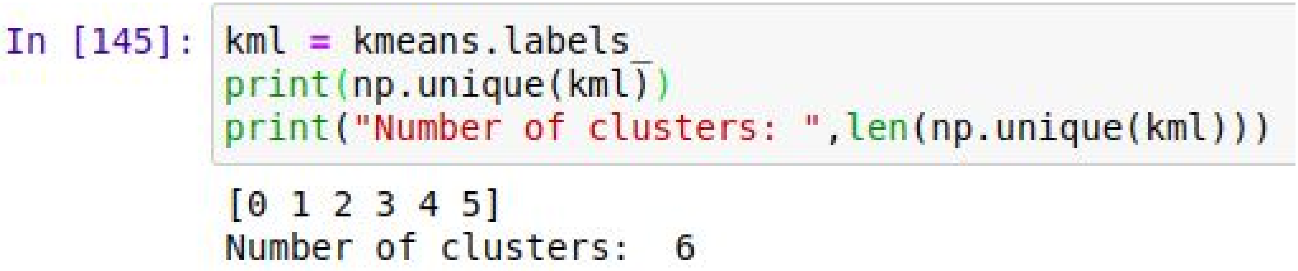


Figure 4.2: Elbow method

Again the clusters are formed by using k-means but by providing the number of clusters that need to be formed, which is 6. The silhouette score, homogeneity score, and completeness score are calculated from it. In order to verify the number of clusters formed, k-means labels are printed(see fig. 4.3) to count the unique numbers from which we can verify the number of clusters formed.



```
In [145]: kml = kmeans.labels_
          print(np.unique(kml))
          print("Number of clusters: ",len(np.unique(kml)))

          [0 1 2 3 4 5]
          Number of clusters:  6
```

Figure 4.3: Verify the number of clusters

To visualize how the data has been clustered, a scatter plot graph is plotted which can be seen in the below figure(fig 4.4).

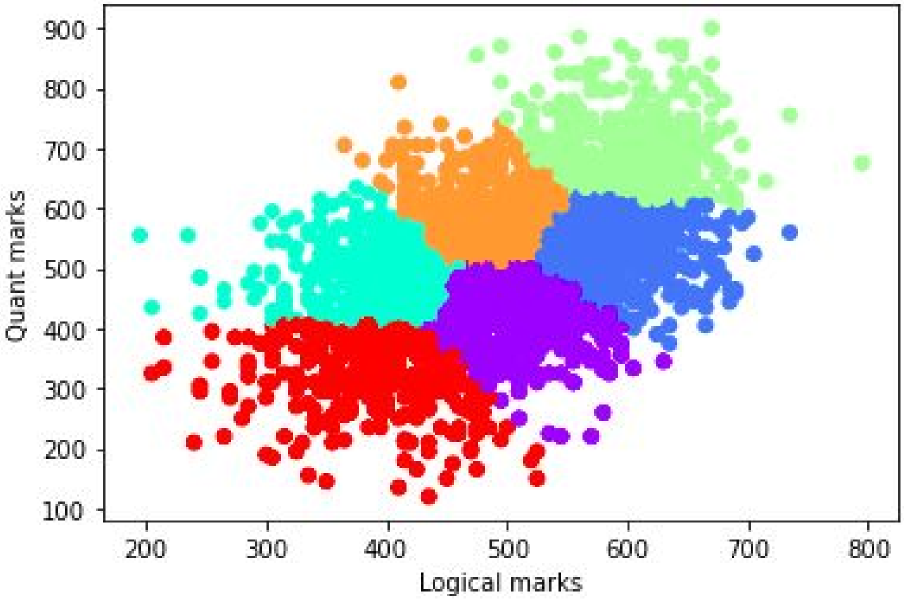


Figure 4.4: Data clusters

From the above graph(see fig. 4.4), we get the range for plotting a new candidate in the cluster by determining their AMCAT logical and quantitive module marks. For example, suppose a candidate got 500 in the logical module and 600 in the quantitive module, then we can assign the candidate on the yellow cluster in the scatter plot graph(see fig. 4.4).

4.2 Supervised Analysis

Linear Regression with a single variable:

In order to predict the salary of a candidate by using the random number as a logical module marks, the linear regression method is used as it is used to predict the value of a variable based on another variable. After performing the regression method, using a loop actual salary and predicted salary of a candidate has been stored in a variable and to compare how similar or different the predicted salary is with the actual one for 5 random candidates, a bar graph is plotted(see fig 4.5)

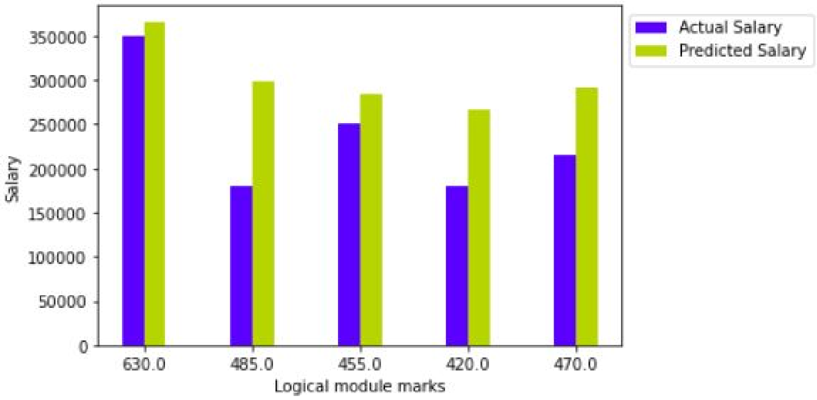


Figure 4.5: Actual and predicted salary for Logical module

It is possible to see that in most cases there is a huge difference between the actual and predicted salary. But in some cases, the prediction is almost similar. The same procedure for predicting salary by providing quantitive marks is been done and the bar graph(see fig 4.6) is plotted with the actual and predicted salary by providing marks of 5 random candidates. Here most of the cases predicted salary is almost to the actual salary but for minimum cases, it showed a huge difference.

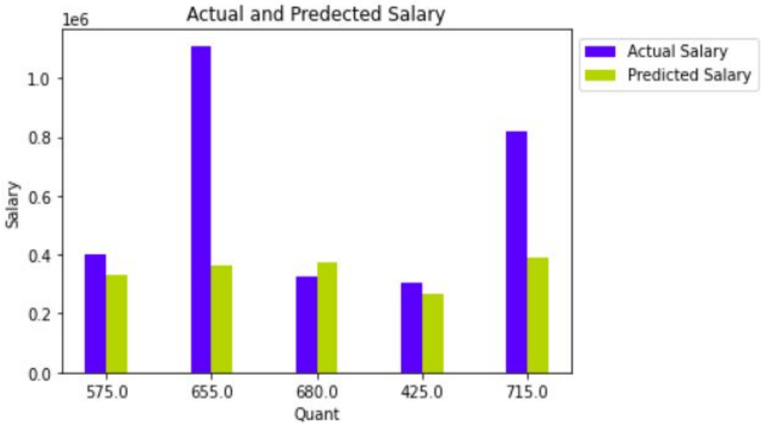


Figure 4.6: Actual and predicted salary for Quantitive module

Linear Regression with multiple variables:

In order to predict the salary of a candidate by using the random number as a logical module marks, the linear regression with multiple variables is been used so that it can verify whether it could generate a more accurate result than the previous regression method performed. In order to proceed with it, columns or fields which have a higher number of correlation(see fig. 3.7) with the salary is been used for this regression technique. The columns are 10percentage, 12percentage, collegeGPA, English, Logical, Quant, and Domain which is used for multiple variable regression.

Five random rows are selected from the dataset and using the values from the respective rows the data is provided to predict salary and then the predicted and the actual salary is stored in the variables. A bar graph(see fig 4.7) is plotted to visualize and compare the difference between the actual and the predicted salary.

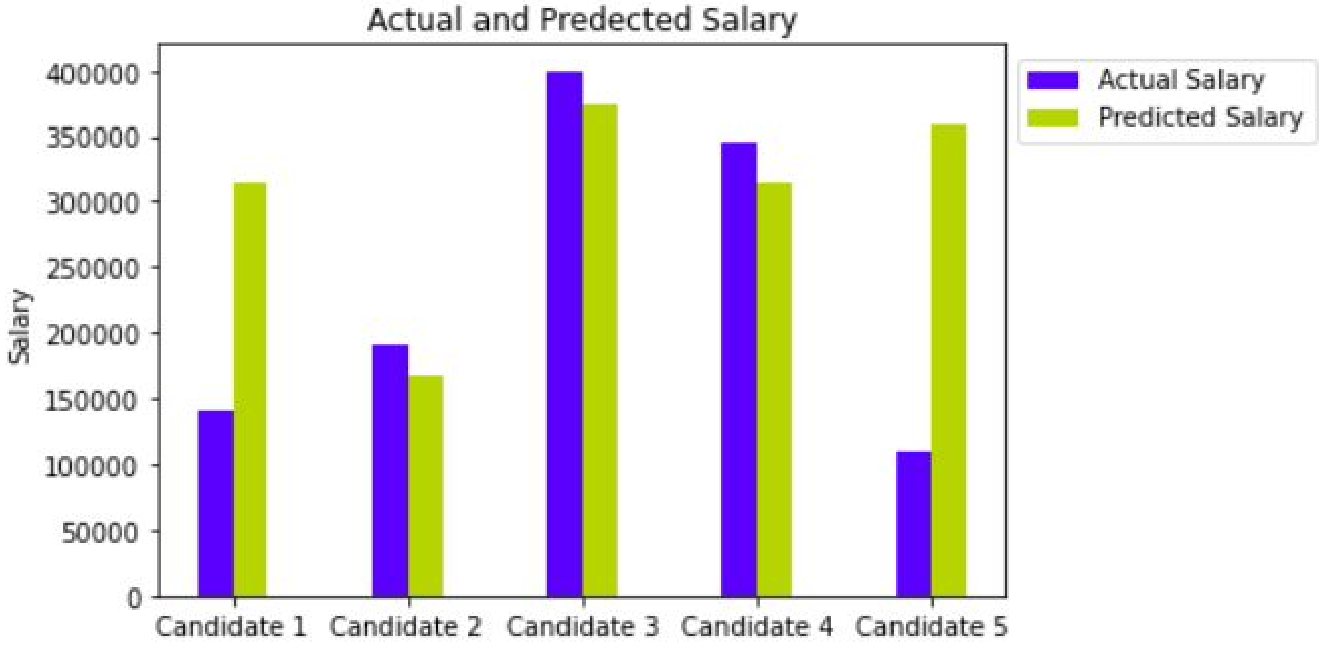
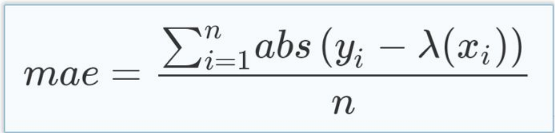


Figure 4.7: Actual and predicted salary for multiple modules

For multiple linear regression, the predicted salary is near to the actual salary for most of the cases. But similar to the single value regressing result, this also has a huge difference between the original and the predicted salary for some output which can be seen for candidate 1 and candidate 5.

So in both cases, there are some errors in predicting the salary of a candidate. To measure the accuracy of the predicted variables the most common metrics used is are MAE which is known as Mean Absolute Error.



$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

Figure 4.8: Formula of MAE

Mean Absolute Error (MAE): MAE measures the average percentage of the errors in a set of predictions. The formula of MAE can be seen in figure 4.8

For the graph presented in figure 4.7, the Mean Absolute Error value is calculated as 124012.8. The value is generated by stooring all the predicted salary and actual salary to a variable while iterating through the loop to predict salary for random 5 candidates. Then each predicted value is subtracted from its actual value which will be an absolute value and then added all the absolute values. The added number is divided by the total numbers of predicted items.

Chapter 5

Reflections and Conclusion

Based on the result produced by the analysis techniques, the results have minimum accuracy. The intent of the analysis was to find the factors which can influence the salary of a candidate mostly and use them in regression algorithms to predict the salary of a new candidate who has all his scores ready. This would have helped the candidates in analyzing what are the factors that are most important to get a better package and using this candidate can prepare themselves. For analyzing the data firstly bar graphs and pie charts were used to compare data. Then, heatmaps are used to find the connection between various fields. The K-means clustering method is used for clustering with the most common factors that can have an impact on the salary. Following to that linear regressing techniques is used for predicting salary by providing mostly correlated factor with salary and other factors to verify which method can produce a more accurate result.

Using the dataset provided by the Aspiring Minds Research team, it is not possible to predict the actual salary of candidates as salary cannot completely depend on the AMCAT module scores and other details provided in the dataset. After taking the AMCAT examination candidates are called for an interview in which the performance of the candidate will surely be measured and can have a huge impact on the salary package they get. But the dataset does not provide the scores of the candidate in the interview. Along with that which city has more number of companies and in which sector the company operates is not provided as the salary can also depend on the city. If it would have provided then it would allow me to analyze it by comparing it with the candidate's specialization and college city.

Appendix A

Software versions, data, and included packages

Python version: Python 3.8.3

Jupyter notebook: 6.0.3

Dataset: https://www.kaggle.com/manishkc06/engineering-graduate-salary-prediction

Packages used:

- pandas
- matplotlib.pyplot
- numpy
- seaborn
- scale from sklearn.preprocessing
- metrics from sklearn
- DBSCAN from sklearn.cluster
- metrics from sklearn
- make_blobs from sklearn.datasets
- StandardScaler from sklearn.preprocessing
- MinMaxScaler from sklearn.preprocessing
- KMeans from sklearn.cluster
- linear_model from sklearn import
- randrange from random
- cluster from sklearn
- LabelEncoder from sklearn.preprocessing
- model_selection from sklearn

Bibliography

- **URL:** https://www.myamcat.com/about-amcat

- **URL:** https://everipedia.org/wiki/lang_en/AMCAT

- **URL:** https://github.com/jakkcoder/k-means-with-elbow/blob/master/Clustering%20with%20K-Means%20and%20Elbow%20Method%20.ipynb

- **URL:** https://www.quora.com/How-many-people-graduate-in-India-every-year-with-at-least-a-bachelors-degree

- **URL:** https://www.myamcat.com/amcat-syllabus?source=helpcenter

- **URL:** https://stackabuse.com/k-means-clustering-with-scikit-learn/

- **URL:** https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/

- **URL:** https://github.com/codebasics/py/blob/master/ML/13_kmeans/13_kmeans_tutorial.ipynb

- **URL:** https://github.com/codebasics/py/blob/master/ML/2_linear_reg_multivariate/2_linear_regression_multivariate.ipynb

- **URL:** https://stackoverflow.com/questions/41690905/python-get-random-ten-values-from-a-pandas-dataframe

- **URL:** https://www.datasciencemadesimple.com/bar-plot-bar-chart-in-python-legend-using-matplotlib/

- **URL:** https://stackoverflow.com/questions/47432224/pandas-sample-based-on-criteria