

**“Predicting Graduate Engineers Employability”**

Submitted in Partial Fulfillment of requirements for the Award of certificate of

Post Graduate Program in Data Science

**Capstone Project Report**

Submitted to



**Submitted by**

Sumit Das

Priyanshu Shekhar Sinha

Aseem Kumar Bastia

Anish Shanbhogue

Rohit

**Under the guidance of**

Mr. Srikar Muppidi

Batch- PGPDSE March-2019

**CERTIFICATE**

This is to certify that the participants Rohit, Anish Shanbhogue, Aseem Kumar Bastia, Priyanshu Shekhar Sinha, Sumit Das who are the students of Great Lakes Institute of Management, have successfully completed their project on “Predicting Graduate Engineers Employability”

This project is the record of authentic work carried out by them during the academic year 2019.

Mentor’s Name

Mr. Srikar Muppidi

Date:

Place: Bengaluru

**ACKNOWLEDGEMENT**

We take this opportunity to express our profound gratitude and deep regards to our mentor Mr. Srikar Muppidi for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The help and guidance given by him time to time shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to the faulty and management office of Great Lakes Institute of Management for their support, valuable information and guidance, which helped us in completing this task through various stages. We are grateful for their cooperation during the period of our project.

Anish Shanbhogue

Sumit Das

Rohit

Aseem Kumar Bastia

Priyanshu Shekhar Sinha

**1. About the Project**

**1.1 Introduction**

Graduate employability is an increasingly major concern for academic institutions and assessing student employability provides a way of linking student skills and employer business requirements.

Student’s employability is a major concern for the institutions and predicting their employability beforehand can help in taking timely actions in order to increase institutional placement ratio. To know weakness before appearing for interview of any company can `help students to work in areas that they need to improve in order to best match the skillset required by company. Enhancing student assessment methods for employability can improve their understanding about companies in order to get suitable company for them.

Data mining and predictive modelling technique such as classification and regression is best suited for predicting the employability of students. The application of data mining in student employability is to search for significant relationships such as patterns, association and changes among variables in datasets. It provides classification methods to predict the level of employability for students.

**1.2 Objectives**

Under the project study, we are trying to utilize the dataset containing information about a set of engineering graduates and their employment outcomes to analyse the following few use cases –

* Given a new student profile, can we predict his/her annual salary from historic data?
* Can we understand what factors in the labor market determine one’s salary? Is it just one’s skills or there are other factors which influence the return in the labor market? What signals and biases enter the labor market?

**1.3 Statistical tools & techniques used and Limitations**

**Tools:**

1. Python was used throughout the course of the project
2. Tableau was used for data visualization and extracting exploratory data analysis

**Techniques Used:**

1. Technique used were simple linear regression………………….
2. For classification ………..

**Limitations:**

1. The publicly available dataset of AMCAT was not huge so the algorithms might not have the exact prediction.
2. Dataset was not optimal as some data points have values that logically doesn’t make sense.
3. Validation dataset was not available so the models that were built could not be tested on external datasets.

**1.4 Dataset Introduction**

The entire data is collected from Aspiring Minds' Employment Outcomes 2015. The dataset contains various information about a set of engineering candidates and their employment outcomes. For every candidate, the data contains both the profile information along with their employment outcome information. Candidate Profile Information includes:

Scores on Aspiring Minds’ AMCAT – a standardized test of job skills. The test includes cognitive, domain and personality assessments

* Personal information like gender, date of birth, etc.
* Pre-university information like high school grades, high school location
* University information like GPA, college major, college reputation proxy.
* Demographic information like location of college, candidates’ permanent location

Employment Outcome Information includes:

* First job annual salary
* First job title
* First job location

Random AMCAT takers were surveyed via email wherein they provided information on the dependent variables in this dataset – the jobs they are in and their corresponding annual salaries. Corresponding independent information about the candidates was recorded at the time of them taking AMCAT.

Dataset Source: <http://research.aspiringminds.com/resources/#ameo>

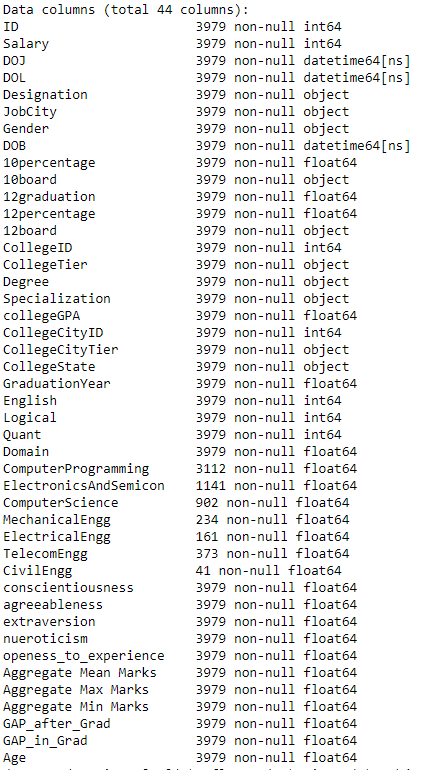
**1.5 Variables in the Dataset**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| ID | AMCAT ID |
| Salary | Annual Salary of AMCAT Students |
| DOJ | Date of Joining the job |
| DOL | Date of Leaving the job |
| Designation | Job profile of the candidates |
| JobCity | City where the candidates secured job |
| Gender | Gender |
| DOB | Date of Birth |
| Age | Age of candidates |
| 10percentage | Percentage of candidates in 10th |
| 10board | Board of education (High School) |
| 12graduation | Percentage of candidates in 12th |
| 12board | Board of education (12) |
| CollegeTier | Status of college |
| CollegeID | College ID |
| Degree | Degree pursued by candidate |
| Specialization | Specialization in degree |
| collegeGPA | GPA of candidates in Degree |
| CollegeCityID | ID of the college |
| CollegeState | State where the college is situated |
| CollegeCityTier | Status of city |
| GraduationYear | Year when the candidates graduated |
| English | AMCAT Score in English |
| Logical | AMCAT Score in Logical |
| Quant | AMCAT Score in Quant |
| Domain | AMCAT Domain Percentile |
| ComputerProgramming | AMCAT Score in Computer Programming |
| ElectronicsAndSemicon | AMCAT Score in Electronics and Semiconductor |
| ComputerScience | AMCAT Score in Computer Science |
| MechanicalEngg | AMCAT Score in Mechanical Engineering |
| ElectricalEngg | AMCAT Score in Electrical Engineering |
| TelecomEngg | AMCAT Score in Telecom Engineering |
| CivilEngg | AMCAT Score in Civil Engineering |
| conscientiousness | Personality traits |
| agreeableness | Personality traits |
| extraversion | Personality traits |
| nueroticism | Personality traits |

**2. Data Cleaning**

**2.1 Data Information**

* The dataset has a lot of null values in ComputerScience, MechanicalEngg, ElectricalEng, TelecomEngg, CivilEngg, 'ComputerProgramming','ElectronicsAndSemicon' ,’JobCitiy’.
* To treat above null values columns we have merged them into three new columns named as 'Aggregate Mean Marks', 'Aggregate Max Marks', 'Aggregate Min Marks'.
* We have changed ‘DOL’ data type as datetime64[ns]
* We have extracted new columns of ‘Age’ from the feature ‘DOB’.
* Extracted 'GAP\_after\_Grad', 'GAP\_in\_Grad' from the gaps taken by students in their academics.



**2.2 Data Cleaning column wise**

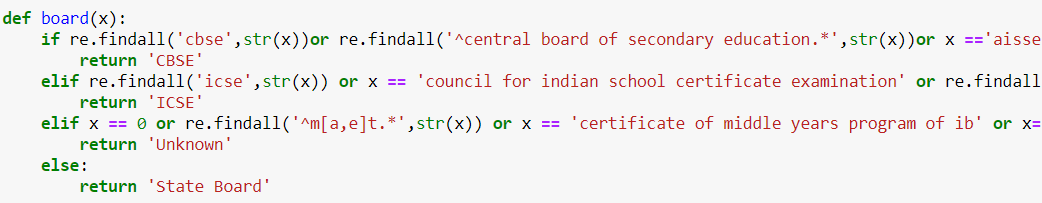
**2.2.1 JobCity Column:**

* City’s were named differently like Bangalore, Bengaluru.etc so we have changed them into a single City for better EDA.
* There were many spelling mistakes, spaces and upper and lower cases in the city names.
* To get it rectified, we have used **Regular Expression** using Python.
* Below are few of the cities which we have changed the names.



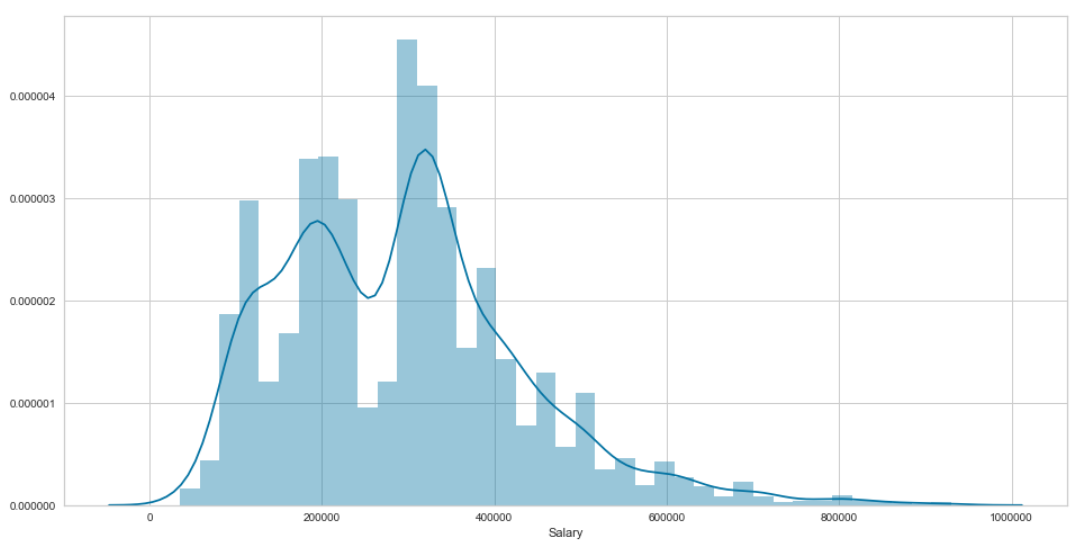
**2.2.2 10th Board and 12th Board Column:**

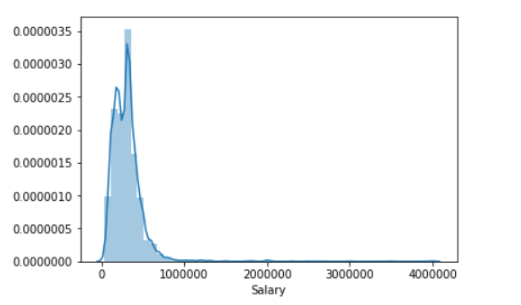
* There were many different types of state boards in our dataset.
* Few of them have made spelling mistakes, short form of their boards and few of them have not mentioned their boards.
* We have divided these boards as CBSE, ICSE, State Board and Unknown.
* To divide these boards we have used **Regular Expression** using Python.

****

**2.3 Treating Outliers**

* The Salary distribution consist of a lot of outliers.
* Nearly 40 Students have gained packages more than 10,00,000 are basically extreme values for our dataset since the data features we have received are very less so these salaries will be treated as outliers in our dataset.
* The Z scores of above students are more than 3. So we have normalized our salary distribution by removing the outliers.

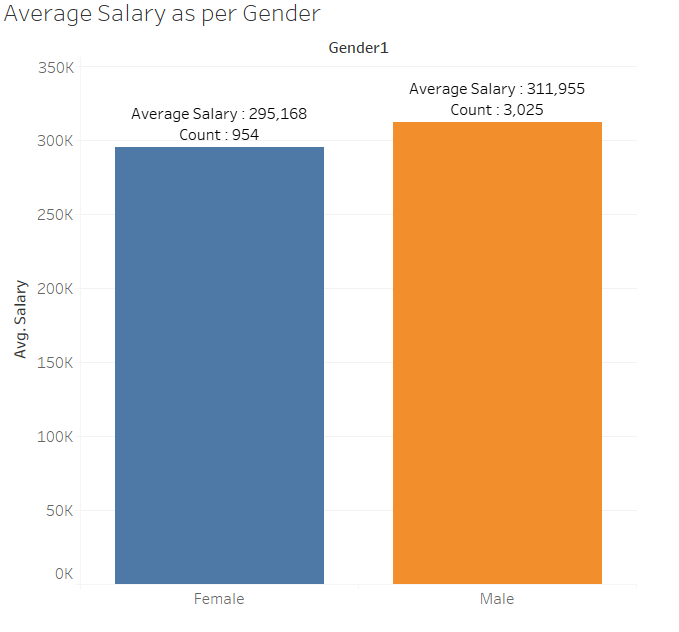


****

**3. Exploratory Data Analysis (EDA)**

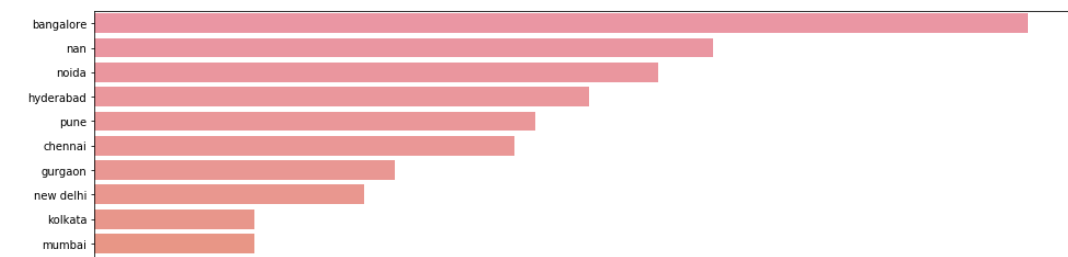
**3.1 Count of Students according to gender**

* The Below graph states that the count of male students are more as compared to females.
* The mean salaries received by the male students are more than females.



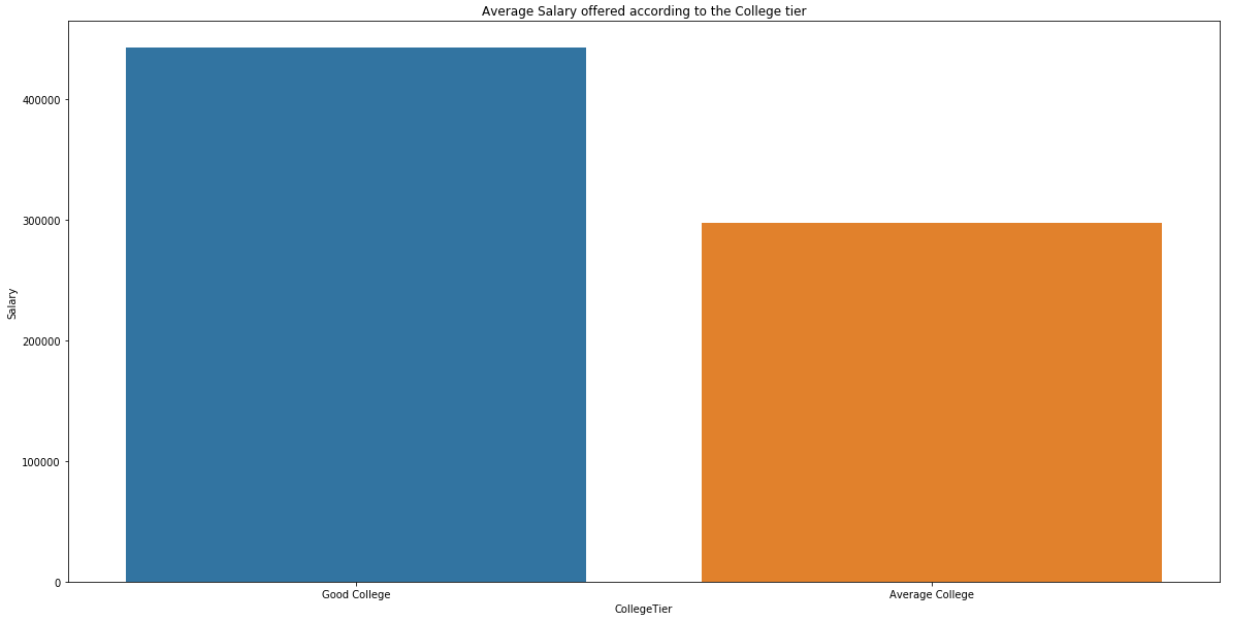
**3.2 Top 10 Cities where Students got Placed**

* The Below graph has given those cities where students get placed.
* The Count of students placed in Bangalore is at top.
* There are many students who have not mentioned there Cities where they have got placed.
* The Below Bar Graph also states that the major students have got placed in Metropolitan city.



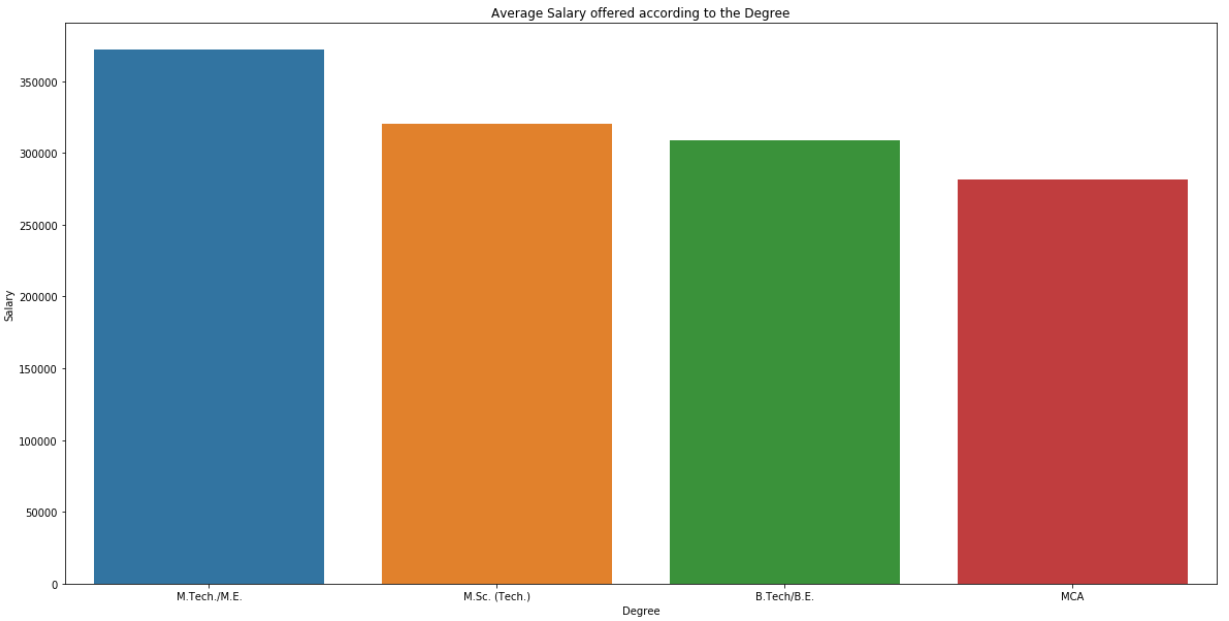
**3.3 Salary based on College Tier**

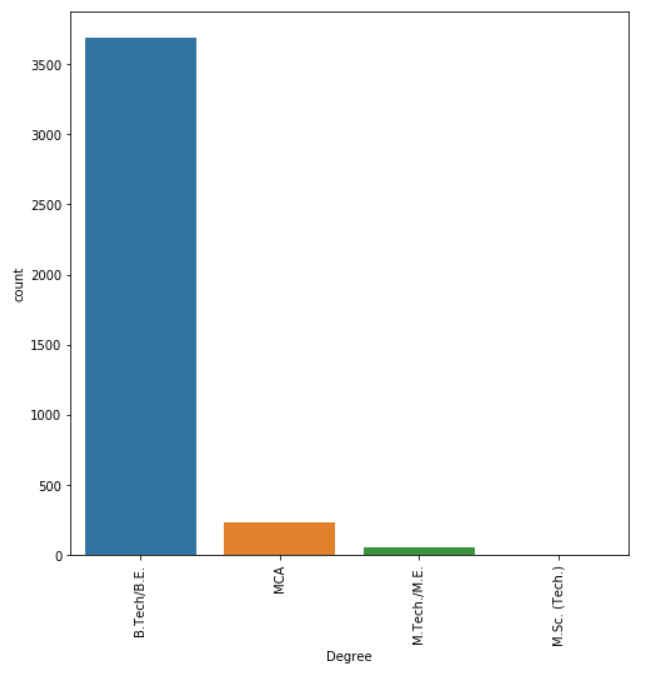
* The Below graph states that if the student belongs to the Good college then they have received an average package more than Average college.
* This inference will be used for feature selection for this particular feature since we can find a clear trend.



**3.4 Salary based on Degree**

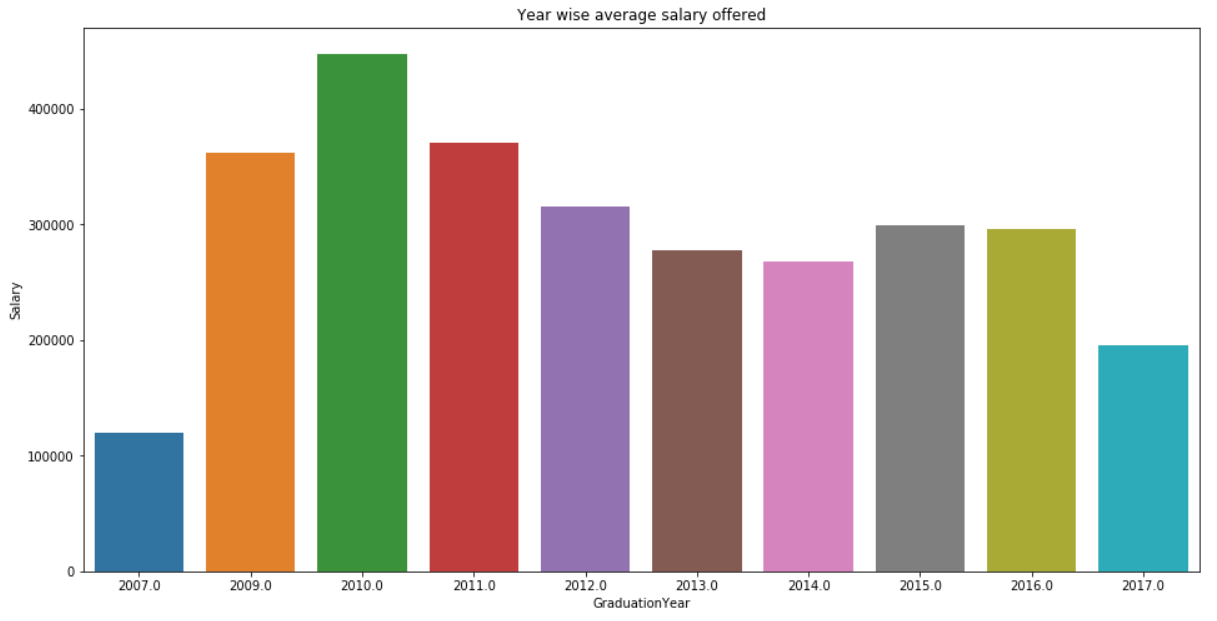
* The Below graph states that the students from M.Tech are most probably to get more average package than other degree students.
* Mean salaries from M.Sc students are less since the count of students are very less.
* B.Tech Students are more in count but still the average package is better than MCA students.





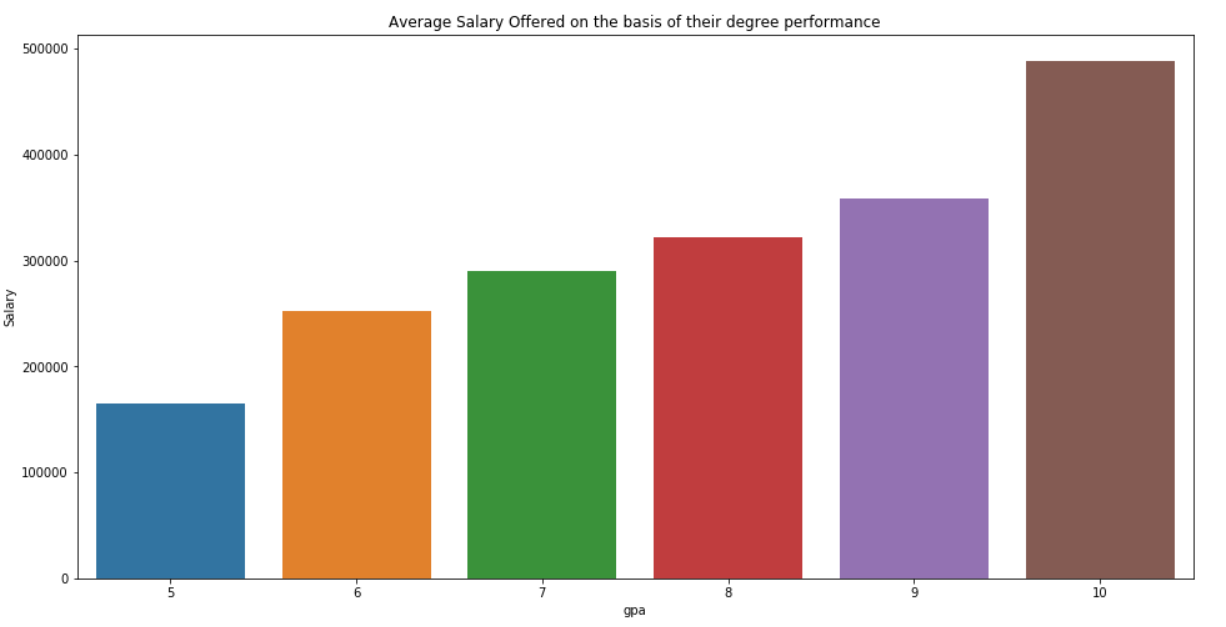
**3.5 Salary based on Graduation Year**

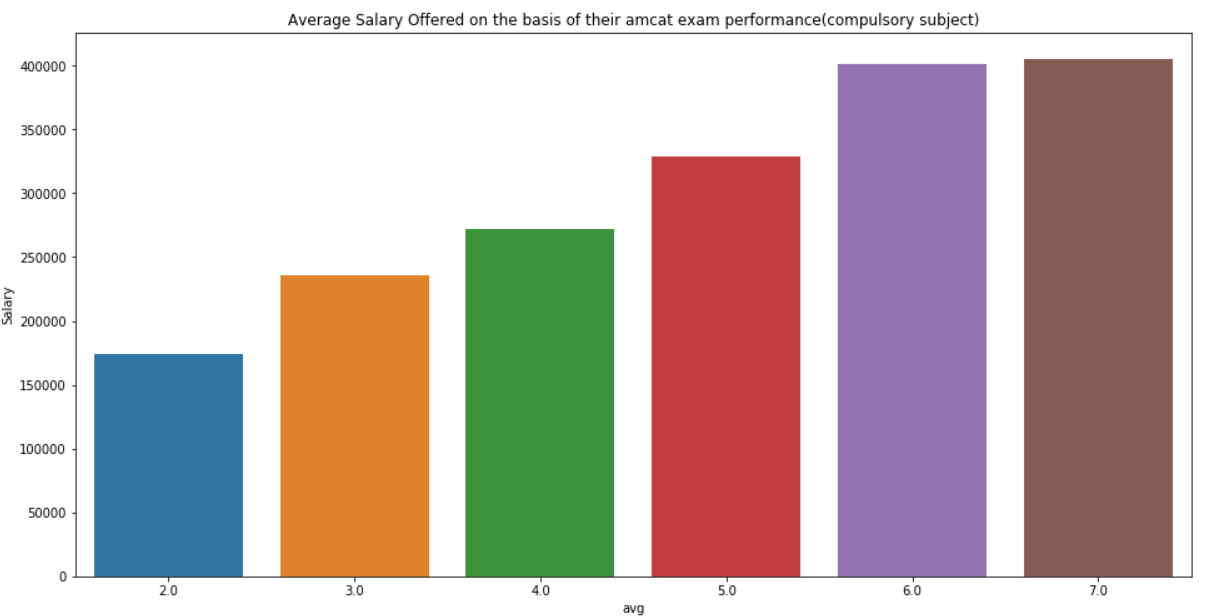
* The Below graph states that the students those who have graduated in 2010 have got better average package as compared to the other year of graduation.

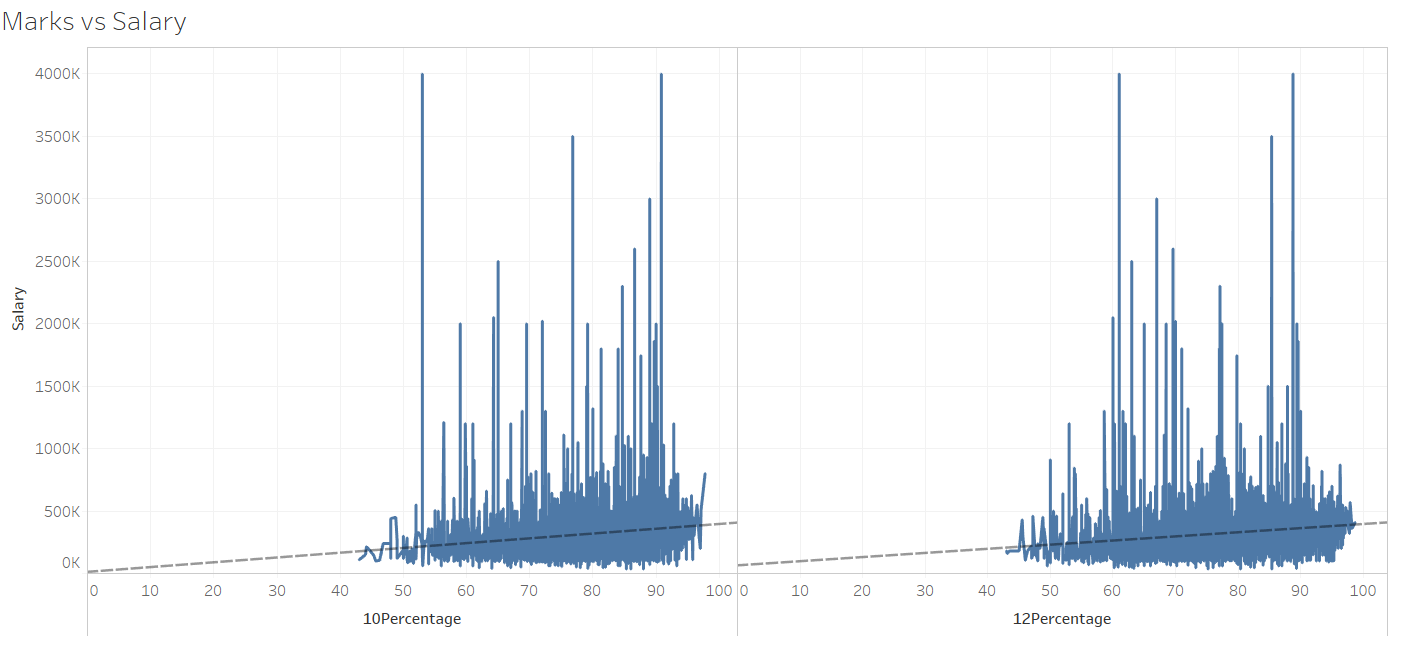


**3.6 Salary based on Marks and Perfrmance in Academics**

* The Below three graph states that if a student gets good marks in his 10th, 12th, College CGPA or AMCAT marks then it is likely to get good salary.
* This trend of marks are directly proportional to the salary.







**3.7 Correlation Plot**

This is the correlation plot between the variables after applying data cleaning. This are all continuous features from the dataset. The correlation matrix for all continuous variable the shows that many features are multi-collinear with each other. So we must fix this by using VIF(Variance Inflation Factor) or PCA(Principal component analysis).



**4. Data Preparation**

## **4.1 Introduction**

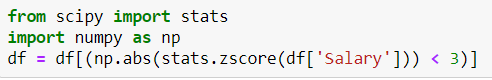
In this part of the project we are trying to prepare our data to build a Regression model. Below are the Models we are going to implement and we will be preparing our data to implement those models

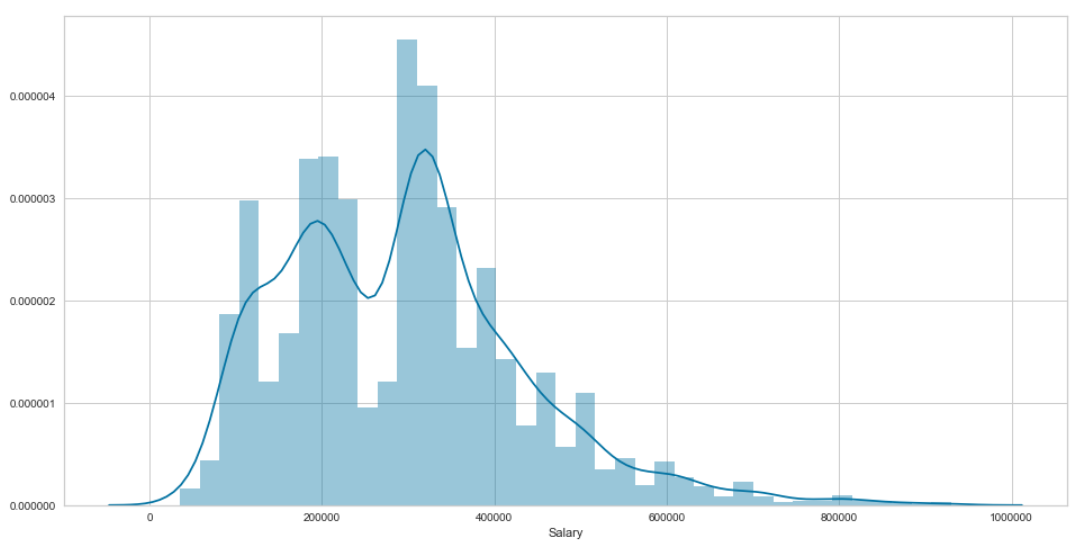
* Regression Model
* Classification model (if Regression model fails)

**4.2 Preparation for Regression model**

The Data preparation process are as follows:-

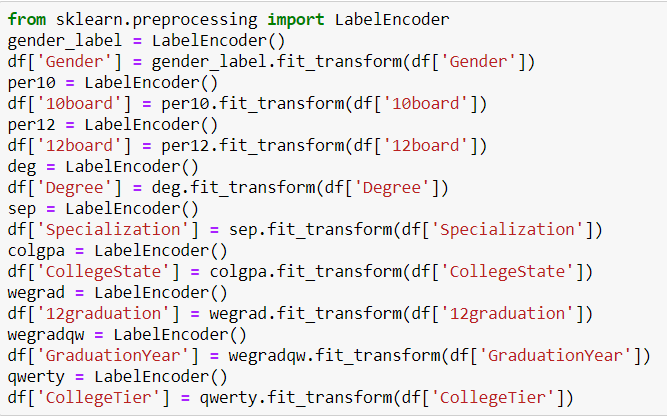
* After data cleaning and EDA we have exported a cleaned data and we will use the same for further model building.
* After importing the data we have removed all the outliers using Z-score.
* We chose those z-score values which are in the range of -3 to 3.



* After removing the outliers the salary distribution is normalized.
* Now we will drop few unnecessary features. Those are as follows:-

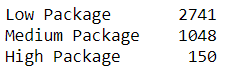
|  |  |
| --- | --- |
| **Variables** | **Reason to Drop the Variables** |
| ID | ID is Unique value. |
| DOJ | It is a dependent variable |
| DOL | It is a dependent variable |
| Designation | It is a dependent variable |
| JobCity | It is a dependent variable |
| DOB | Instead of using DOB we will use Age |
| College ID | It is unique value |
| College City ID | It is unique value |
| College City Tier | Not getting a proper inference from it |
| ComputerProgramming | Contains many null values |
| ElectronicsAndSemicon | Contains many null values |
| ComputerScience | Contains many null values |
| MechanicalEngg | Contains many null values |
| ElectricalEngg | Contains many null values |
| TelecomEngg | Contains many null values |
| CivilEngg | Contains many null values |

* Next we will factorize the categorical values using LabelEncoder()

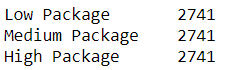


**4.3 Preparation for Classification model**

* The data preparation for Classification model is similar to regression model but here we have made some changes on features.
* We will classify the salary into three types i.e High Package, Medium Package and Low Package.
* Our Data is highly imbalanced and the count of multi classes are as follows:-

****

* So the count of Student who has got less package is more so our model may not predict High and Medium Package.
* To remove imbalance in our dataset we have used **SMOTE** in python
* **SMOTE** basicaly stands for Synthetic Minority Oversampling Technique. This is a statistical technique for increasing the number of cases in the dataset in a balanced way. The module works by generating new instances from existing minority cases that we supply as input. This implementation of SMOTE does **not** change the number of majority cases.
* After Using SMOTE, the dataset is balanced.

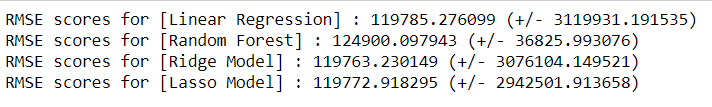


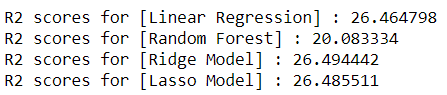
* According to our correlation matrix, we have a lot of multi collinearity in our dataset. Since we cannot remove few features due to lack of important continous features.
* So we will use PCA(Principle Component Analysis) in Continuous features only.
* We have used n\_components = 12 since the explained variance ratio is 96% at 12.
* We will merge 12 Principle components with the remaining categorical features followed by dropping all continous features.

**5. Building Model**

**5.1 Regression Model**

* Modelling using the listed algorithms was done under multiple scenarios and results compiled for overview.
* For Hyper parameter tuning, we have used Grid-SearchCV.
* The List of tuned models with the RMSE Scores and variance are as follows:-

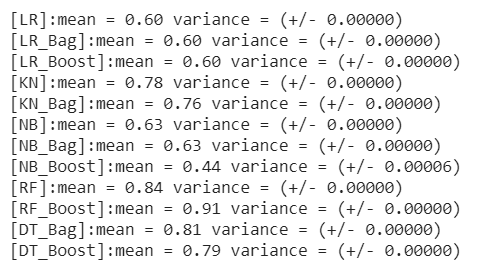


****

* The Above RMSE’s is very high with the very less accuracy of each model.
* This is because of lack of sufficient data in our dataset. So our model is not able to predict the salary.
* So that’s why we must go with multiclass problem since it is possible to provide a range of salary.

**5.2 Classification Model**

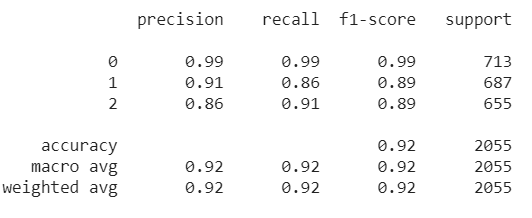
* While building Classification model, we have used the above model features with dependent Multi-class Salary.
* For Hyper parameter tuning, we have used Grid-SearchCV.
* The List of tuned models with the Mean F1 Scores and variance are as follows:-

****

* From the above different models Random Forest Ada-Boost model is performing very well with proper hyper-parameter tuning. So we will use the same as our main Algorithm.

**5.3 ADA-Boost Random Forest Algorithm Performance**

* Below is the Precision Recall and F1-Score of the **ADA-Boost Random Forest Algorithm**
* We will check multiclass performance using F1-Score.



|  |  |  |  |
| --- | --- | --- | --- |
| High Package | 706 | 0 | 7 |
| Low Package | 3 | 594 | 90 |
| Medium Package | 2 | 56 | 597 |

* The Above confusion matrix and Precison score for High Package is very high, this basically states that our model will predict High Package precisely. So It may reduce false promising data.
* This model may vary while predicting the Low Package.
* For predicting medium Package our model is struggling more but Still the F1-Score is similar to the Low Package.

**6. References**

* <https://www.researchgate.net/publication/317348739_An_Univariate_Feature_Elimination_Strategy_for_Clustering_Based_on_Metafeatures>
* <https://stattrek.com/chi-square-test/independence.aspx>
* <https://www.saedsayad.com/logistic_regression.htm>
* <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
* <https://en.wikipedia.org/wiki/Random_forest>
* <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
* <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>
* <https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/>
* <https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba>