greatlearning

"Predicting Graduate Engineers Employability"

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

Post Graduate Program in Business Analytics and Business Intelligence

Capstone Project Report

Submitted to

GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

Submitted by

Aamir Rather (BABIGFEB1801)

Kushal Bansal (BABIGFEB1821)

Sudhakar Chaudhary (BABIGFEB1845)

Saurabh Gupta (BABIGFEB1840)

Vipul Jain (BABIGFEB1850)

Under the guidance of

Mr. Neelesh Singh

Batch- PGPBABI.G.Feb'18 January 2019

Abstract: The purpose of this report is to predict employability of the graduate engineers and to propose an implementation of sector specific job readiness tests for job seekers that (1) assess their employability; (2) are voluntary initially, with a longer term view to universal coverage; (3) are championed by a group of employers that graduates aspire to work for; and (4) use independent assessment agencies for implementation.

Tools and Techniques: R, R Studio, Tableau, Simple Linear Regression, Random Forest, Gradient Boosting, Decision Trees, Naive Bayes

Domain: Education

CERTIFICATE

This is to certify that the participants Aamir Rather, Kushal Bansal, Sudhakar Chaudhary, Saurabh Gupta, Vipul Jain 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 2018- 2019.

Mentor's Name & Sign

Mr. Neelesh Singh

Program Director

Dr. Bappaditya Mukhopadhyay

Date:

Place: Gurugram

ACKNOWLEDGEMENT

We take this opportunity to express our profound gratitude and deep regards to our mentor Mr. Neelesh Singh 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.

Lastly, we thank almighty, our family and friends for their constant encouragement without which this course would not be possible.

Aamir Rather

Kushal Bansal

Sudhakar Chaudhary

Saurabh Gupta

Vipul Jain

Contents

1. 7
2. 8
    - Introduction 8
    - Objectives 8
    - Dataset Introduction 8
    - Statistical tools & techniques used and Limitations 9
3. 10
    - Exploratory Data Analysis (EDA) 10
    - Key Observations from EDA 12
4. 15
    - Data Cleaning 16
    - Missing Values Treatment 16
    - Outlier Treatment 18
6. 19
7. 20
8. 21
    - Annexure 1 - Data Description 21
    - Annexure 2 – Summary Stats 24

List of Figures

Figure 1 – Frequency Plots of Variables 11
Figure 2 – Correlation plot between variables 12
Figure 3 – Bar Chart for Salary by City 13
Figure 4 – Salary Trend over the last 10 years 13
Figure 5 – Missing value assessment after preliminary cleaning 17

List of Tables

Table 1 – Measures of Central Tendency and Dispersion for Salary 12
Table 2 – Measures of Central Tendency and Dispersion for Academic Parameters 12
Table 3 – Salary comparison based on gender 14
Table 4 – Salary comparison based on Specialization 14
Table 5 – Assessment of Missing Values in data 16

Executive Summary

India has seen unprecedented economic growth since 1995, which is roughly more than 2.5 decades earlier. But even now, more than 60% of the urban Indian workers don't have formal sector jobs. This number might look unalarming, but actually it is. Most of this population is isolated from the country's progress, has much lower salaries, don't have labour protection laws and don't have access to formal credit services. If this catastrophic problem is not looked into seriously, the much expected demographic dividends will nightmarishly turn into a demographic disaster.

Regulatory reforms, government policies such as MNREGA, National Skill Development Corporation etc.usually take a lot of time to implement. Hence, the most effective measure would be to upgrade the skill levels of India's labor force. Recruiters are struggling to find skilled candidates to fill formal sector jobs, paying a hefty premium for the right talent. Job seekers, in turn, are rushing to get degrees to improve their likelihood of getting jobs, fueling a dramatic boom in higher education. A staggering 20,000 colleges have opened in the last decade. But the fact of the matter is, most of these institutes are remarkably low in terms of quality of education imparted.

The recent surge of low quality colleges is leading employers to question the value of degrees. Facing employer skepticism, job seekers are finding it harder to demonstrate their ability and earn higher wages.An intervention that generates reliable information about the quality of graduates can help firms hire better workers, reward skilled workers with higher wages, and reveal the quality of colleges to both employers and students. We, hereby, propose implementing sector specific Job Readiness Tests for job seekers that (1) assess their employability; (2) are voluntary initially, with a longer term view to universal coverage; (3) are championed by a group of employers that graduates aspire to work for; and (4) use independent assessment agencies for implementation.

Job Readiness Tests should be implemented by domain level Skill Development Councils (SDCs), led by employers motivated by self-interest in recruiting better talent. Importantly, SDCs must remain firmly independent from government to ensure industry trust of the tests. The autonomous nature of these tests will empower both.

1. About the Project

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.

In the last four years, there is no significant improvement in employability of engineers. Recent study by Aspiring Minds NRE Report shows that only 17.91% of engineers were employable for the software services sector, 3.67% for software products and 40.57% for a non-functional role such as Business Process Outsourcing.

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.

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?

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

Statistical tools & techniques used and Limitations

Tools:

1) R/R Studio 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, Random Forest, Gradient Boosting algorithms for feature selection
2) For classification, decision trees and Naive Bayes classification techniques were used

Limitations:

1) The publicly available dataset of AMCAT was not huge so the algorithms might not have the predictive
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.
4) New emerging skills such as AI, IoT, Cloud Computing etc. are not a part of the dataset. Hence, the important skills for predicting salary may differ over time.

2. Data Understanding

The dataset being used for this study has a total of 38 fields and the detailed data description is submitted in Annexure 1 of this report.

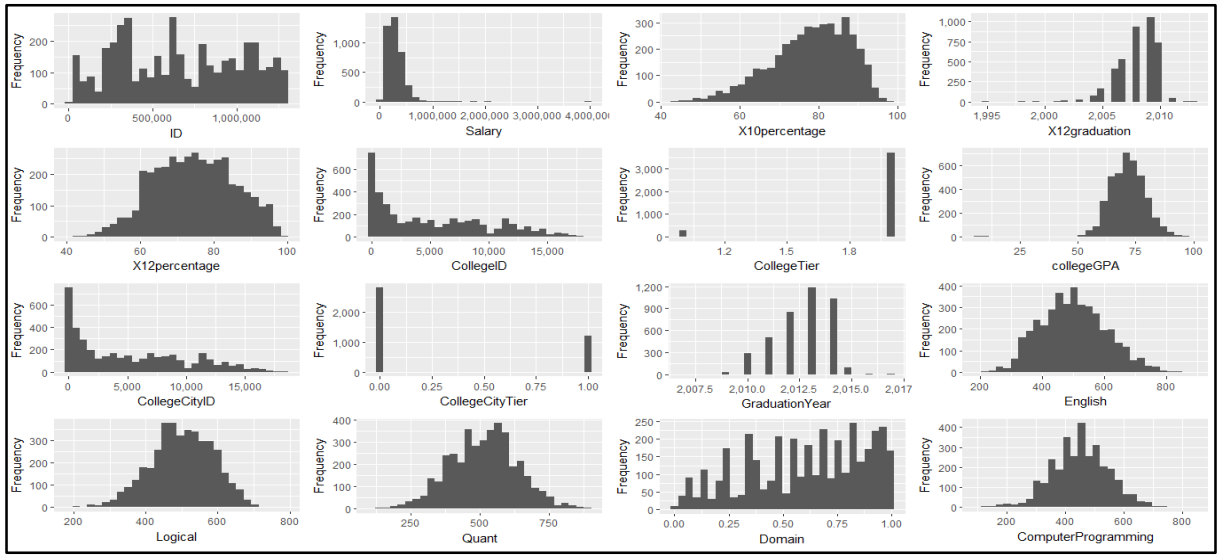
Exploratory Data Analysis (EDA)

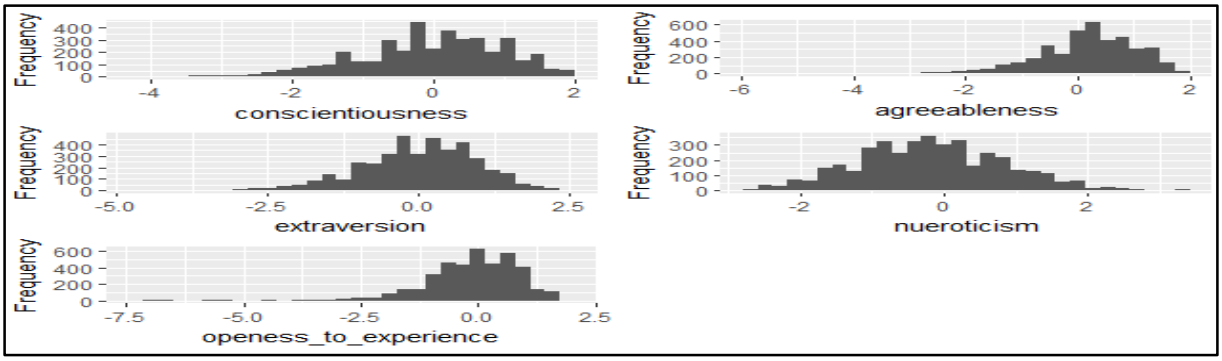
a) Summary stats for all the fields as calculated in RStudio are submitted in Annexure 2

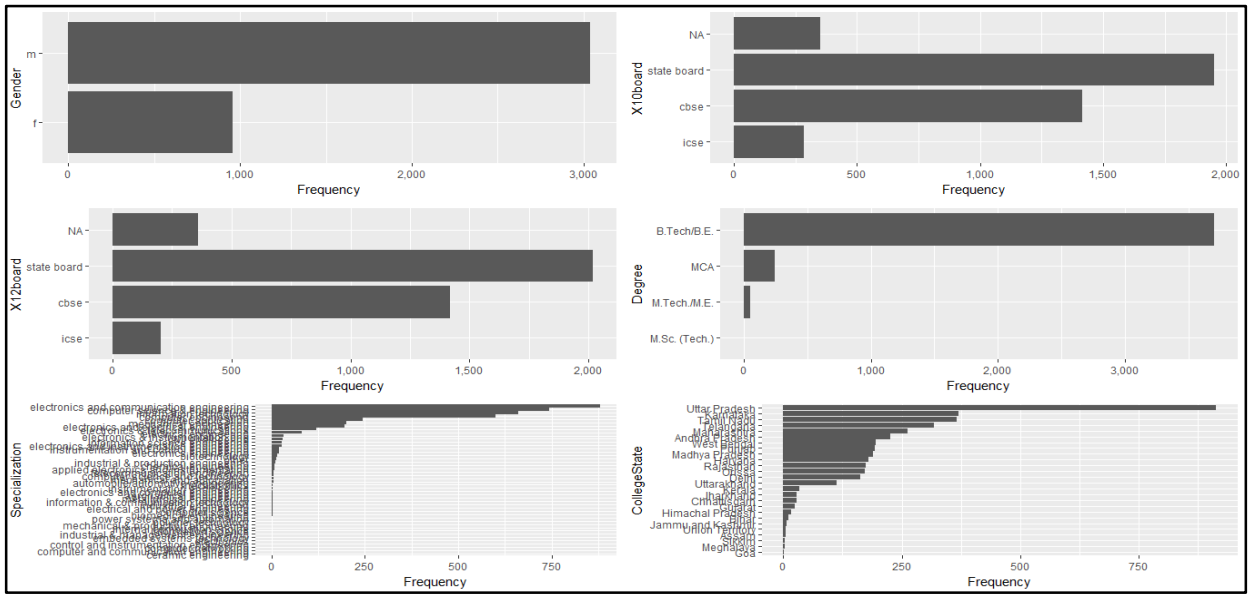
b) Checking the dimension of the input dataset and the type of variables (continuous or categorical)

c) Plotting the histograms for each of the variables for deriving insights

Figure 2 – Frequency Plots of Variables

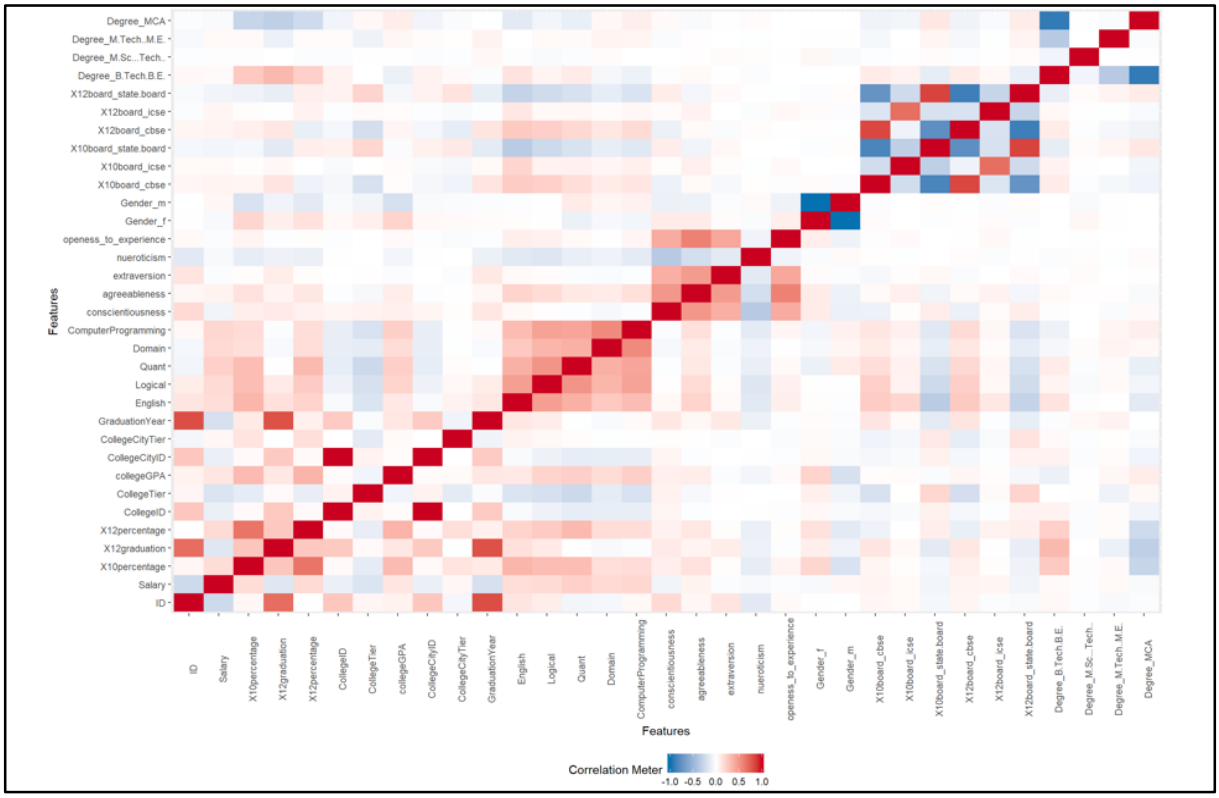






Correlation plot between continuous variables (categorical to be ignored):

Figure 3 – Correlation plot between variables



Key Observations from EDA

a) Salary is the target variable for the project. The unit of Salary is INR (Indian Rupee). The histogram shows the distribution of Salary. The data is slightly skewed on the right. Correlation coefficients show a weak positive correlation between the Salary and academic variables: 10percentage, 12percentage and collegeGPA.

Table 1 – Measures of Central Tendency and Dispersion for Salary

Dependent Variable 'Salary' in INR					
Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
0%	25%	50%		75%	100%
35000	180000	300000	307700	370000	4000000

Table 2 – Measures of Central Tendency and Dispersion for Academic Parameters

	Variables related to Academic Performance					
	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
	0%	25%	50%		75%	100%
10percentage	43	71.68	79.15	77.92	85.67	97.76
12percentage	40	66	74.4	74.47	82.6	98.7
collegeGPA	6.45	66.4	71.72	71.49	76.33	99.93

b) Salary mentioned in the dataset is current salary drawn by the candidate (participant of the AMCAT test)

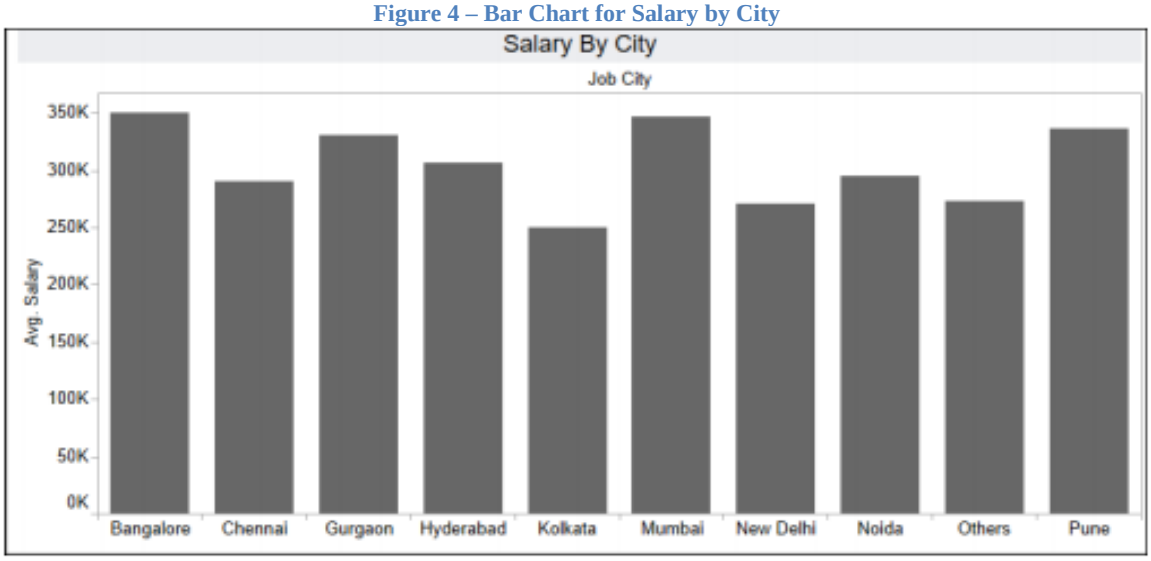
c) A positive correlation between the Salary and Cognitive Skill variables: English, Logical, and Quant. It is also evident that there is a correlation between the cognitive variables

d) The engineering domain scores have a weak positive correlation with Salary. The personality scores also seem to have a very low positive correlation with Salary.

e) Below barplot of Salary Vs JobCity shows that cities with highest mean salaries are Bangalore, Mumbai and Pune followed by Gurgaon & Hyderabad. New Delhi, despite being the national capital and Kolkata despite being a metro city, don't fare well in terms of salaries.

f) Salaries less than Rs. 100000 were capped at Rs. 100000 due to the reason that it might be incorrectly entered as monthly salary and not per annum salary.

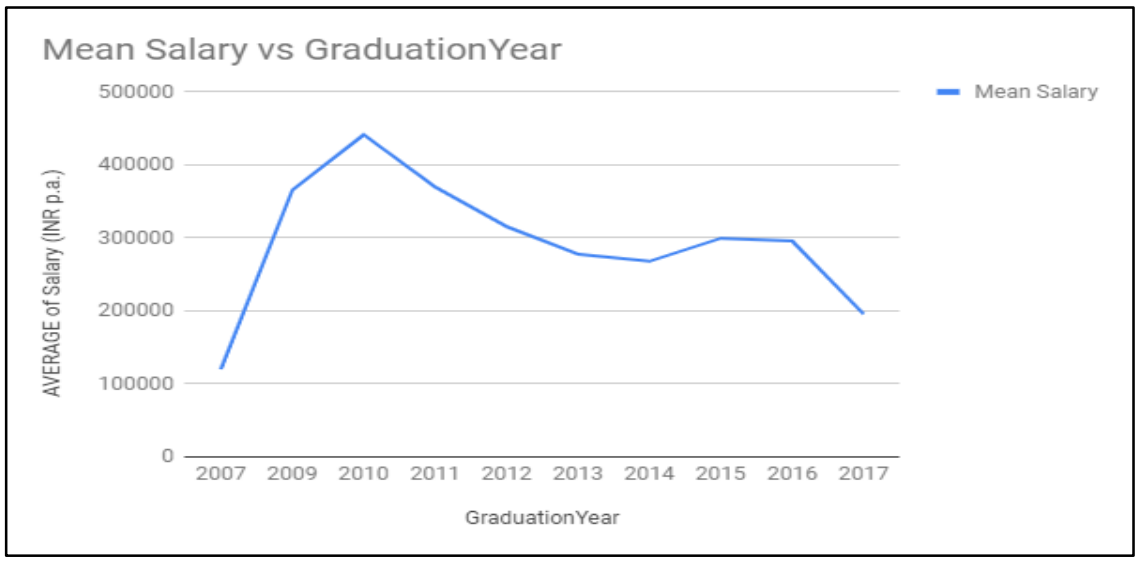
Figure 4 – Bar Chart for Salary by City



g) Salary trends have not remained constant over the years of graduation. For instance, there is a gradual increase in Salary per annum for candidates passing out since 2007 till 2010 (the peak salary year). Then onwards, there is a gradual dip in annual salaries offered to passed candidates till 2014, after which again the salaries have started increasing, but not to the previous levels.

- It is to be noted that for 2017, only 8 records are present in the dataset, thus, hardly any conclusion can be made for year 2017.

Figure 5 – Salary Trend over the last 10 years



h) Salaries of Male and Female candidates were almost similar, although the male candidates outnumbered female by a ratio of 4:1.

Table 3 – Salary comparison based on gender

Salary (INR)	Male	Female
Mean	311716.2	294937.3
Min	35000	35000
Max	4000000	3500000

i) Table below shows the different engineering domains, with their min, mean and max salaries. Evidently, there's no much difference in mean salaries across various engineering branches except Biotech which yields a little lower salary than others.

Civil and Biotech both have a high lower quartile range as compared to other engineering domains.

Table 4 – Salary comparison based on Specialization

Specialization Domain	Min Salary	Avg Salary	Max Salary
Biotech	100000	258529	450000
Civil Engineering	110000	381207	800000
Computer Science	35000	314161	4000000
Electrical Engineering	40000	291786	1860000
Electronics & Communication	40000	298562	3000000
Information Technology	35000	307308	2000000
Mechanical Engineering	60000	315019	1300000
Others	100000	315714	730000

j) Higher the Tier of the college, higher are the salaries offered to their students. Mean salaries of Tier 1 colleges is INR 442356 versus salaries of Tier 2 colleges which is INR 296893 only.

k) 10th percentages are clustered most in the range 70%-90% while 12th percentages are clustered in the range 60%-80%.

l) Out of five given personality traits, candidates scored lower in Neuroticism (degree of emotional stability and impulse control) as compared to other 4 traits.

m) Electronics & Comm. and Computer Science are the 2 topmost specialization domains

3. Data Preparation

Data Cleaning

a) Conversion of categorical variables to factors– All the categorical variables like designation, jobcity, gender, 10th board, 12th board, 12graduation etc. were converted to factor variables

b) Cleaning of 10th board and 12th board field – The original data set had specific names of state examination boards and other boards like Kerala board, up board etc.. For the sake of determining if students coming from state boards have significant impact on salary, all the different state boards were clubbed into one broad category of "state board" and data cleaning is done for bringing in uniformity amongst student from cbse and icse boards.

c) Removing 6 variables which have more 70% missing values and also not significant and dropping 1 variable which has same value for every row.

Table 5 – Assessment of Missing Values in data

Field Name	Percentage of Missing values
ElectronicsAndSemicon	71%
ComputerScience	77%
MechanicalEngg	94%
ElectricalEngg	96%
TelecomEngg	91%
CivilEngg	99%

d) Created a new variable called Experience (by months) by calculating the difference of DOJ and DOL. We can use this variable for better analysis.

e) Also the DOJ variable has value "present" in it which signifies that the employee is still working in that company. We have replaced "present" with the date when the data was collected.

f) DOB variable was converted to Age to have a better understanding in the conclusions..

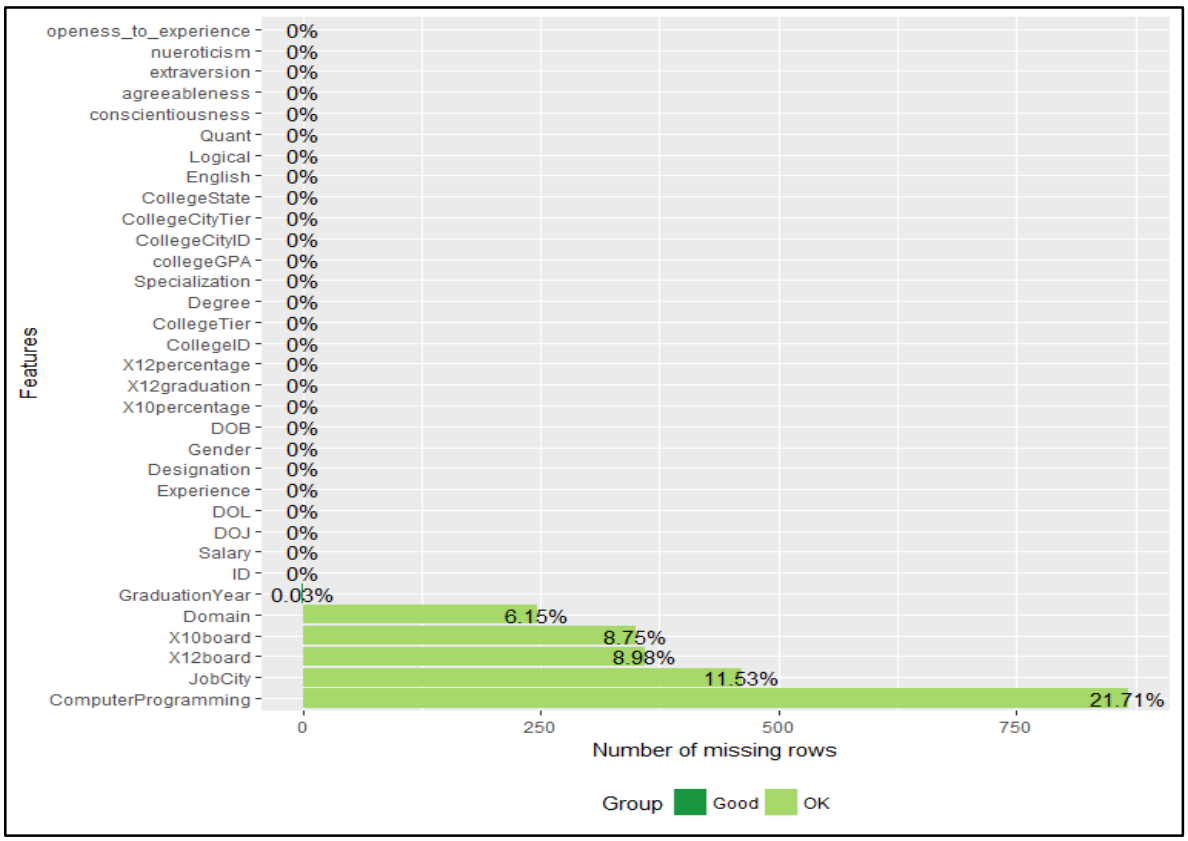
g) CollegeGPA - This is the raw information submitted by candidates. Some have submitted percentages while others have posted on a 10-point scale. Some of these GPAs might be relative while others can be absolute. We have used percentage as a metric and converted GPA's into percentage.

h) After finalizing the regressions models, 3 levels were created for the dependent variable Salary. These are low, mid and high. Further classification algorithms are run over the complete dataset.

Missing Values Treatment

- Further checking the missing values in the dataset after removing the 6 fields:

Figure 6 – Missing value assessment after preliminary cleaning



- Only 6 variables are now present with significant missing values
- Domain and ComputerProgramming

Field Name	Percentage of missing values
Domain	6.15%
ComputerProgramming	21.71%

- Since both these fields are continuous, we are imputing these with median value. The correlation of these fields was also measured against our dependent variable "Salary" and the correlation coefficients pre and post imputation are given below:

Field Name	Correlation pre-imputation	Correlation post-imputation
Domain	0.178	0.169
ComputerProgramming	0.17	0.16

- X, XII board, Graduation Year and Job City – Imputation of categorical data has been done by using KNN with 5 nearest neighbors

Outlier Treatment

- Checking all the continuous variables for outlier and capping them. We have used **Quantile distribution** for handling Outliers. We can also try **(1.5*IQR)** method but as we have only one outlier it will not be needed

Figure 6 – Quantile values for Outlier assessment

```
> quantile(Salary,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
    1%    2%    3%   10%   20%   30%   40%   50%   60%   70%   80%   90%   95%
 75000 95000 100000 120000 180000 200000 240000 300000 325000 350000 400000 480000 570000
   99%  100%
930600 4000000
```

```
> quantile(`10percentage`,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
    1%     2%     3%    10%    20%    30%    40%    50%    60%    70%    80%    90%    95%
52.0000 54.8000 57.0000 64.0000 69.7120 73.0820 76.3040 79.1500 82.0000 84.4000 87.0000 89.8000 91.6000
   99%   100%
94.2012 97.7600
> quantile(`12percentage`,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
   1%    2%    3%   10%   20%   30%   40%   50%   60%   70%   80%   90%   95%   99%  100%
50.000 52.894 54.200 60.400 64.330 68.000 71.000 74.400 77.500 81.000 84.688 89.600 92.900 96.000 98.700
```

```
> quantile(collegeGPA,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
    1%     2%     3%    10%    20%    30%    40%    50%    60%    70%    80%    90%    95%
54.7500 56.0964 58.0000 62.0000 65.0140 67.7230 70.0000 71.7200 73.4000 75.3490 77.7120 81.0000 84.0000
   99%   100%
90.0000 99.9300
> quantile(Logical,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
  1%   2%   3%  10%  20%  30%  40%  50%  60%  70%  80%  90%  95%  99% 100%
 295  315  335  385  425  455  485  505  525  555  580  610  640  680  795
> quantile(Quant,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
   1%    2%    3%   10%   20%   30%   40%   50%   60%   70%   80%   90%   95%   99%  100%
234.70 265.00 285.00 355.00 405.00 445.00 485.00 515.00 545.00 575.00 615.00 665.00 715.00 795.15 900.00

> quantile(Domain,c(0.01,0.02,0.03,0.1,0.2,0.3,0.4,0.50,0.6,0.7,0.8,0.9,0.95,0.99,1))
       1%        2%        3%       10%       20%       30%       40%       50%       60%
-1.0000000 -1.0000000 -1.0000000  0.1121394  0.2454557  0.3760596  0.5259226  0.6226429  0.7040904
      70%       80%       90%       95%       99%      100%
 0.7935806  0.8727970  0.9522456  0.9787993  0.9967445  0.9999104
```

- Outlier treatment of Salary - there is an outlier at 100 percentile and the difference between 99 percentile & 100 percentile is huge. So we have padded the salary at 99 percentile removing the outlier. We have capped the highest salary as 925400 which is the 99th percentile of salary field. Also flooring has been done at 100000.
- No significant outliers found in other continuous variables

Modelling

The detailed R script along with screenshots of outputs are submitted in Annexure 3.

Objective

To predict the Salary based on the student's academic & other information.

Step –wise Approach

1. Creation of dummy variables for each of the categorical variables and applied one-hot encoding to categorical variables
2. Binding variables
3. Checking for outliers for all numeric variables
4. Treating outliers
5. Converting CGPA into percentage grades
6. Partitioning the data in 70:30 for train and test
7. Creation of a linear regression model considering all the independent variables and building a regression model for predicting the salary knowing all other variables
8. Derived the regression equation with coefficients of the variables and the intercept figures
9. Derived the correlation between salary and predicted salary
10. Creation of a regression model with only significant variables

11. Running the random forest algorithm for another check of significant variables
12. Running the regression model with only significant variables from random forest
13. Running the random forest model with only significant variables from random forest
14. Applying Gradient boosting for regression. Summary gives a table of variable importance and a plot of variable importance
15. Running the regression model on selected 22 variables including Age
16. VIF for all the variables comes out to be less than 2 hence we derive that the variables are not collinear.
17. Classifications on Salary
18. Using cut function to derive the low, mid and high salary slab
19. Checking for class imbalance in the salary slab
20. Partitioning the data in 70:30 for train and test
21. Checking the dimensions for test and train datasets
22. Applied a rule based C5.0 classification model
23. Predict method was used to get the hard class prediction
24. Created a confusion matrix to calculate cross-tabulation of observed and predicted class with associated statistics

4. Recommendations and Conclusions

5. References and Bibliography

[1] Aspiring minds. http://www.aspiringminds.com

[2] http://ikdd.acm.org/Site/CoDS2016/datachallenge.html

[3] ACM IKDD CODS. Data challenge, March 2016. http://ikdd.acm.org/Site/CoDS2016/

[4] Aspiring Minds. National employablity report engineers annual report, 2015. http://www.aspiringminds.com/research-reports.

6. Annexures

Annexure 1 - Data Description

Input	Description	Remarks
ID	A unique ID to identify a candidate	
Salary	Annual CTC offered to the candidate (in INR)	
DOJ	Date of joining the company	
DOL	Date of leaving the company	"present" means the candidate continues to work at the company at the time of collecting this information
Designation	Designation offered in the job	
JobCity	City in which the candidate is offered the job	
Gender	Candidate's gender	m denotes Males and f denotes Females
DOB	Date of birth of candidate	
10percentage	Overall marks obtained in grade 10 examinations	
10board	The school board whose curriculum the candidate followed in grade 10	
12graduation	Year of graduation - senior year high school	
12percentage	Overall marks 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 her/his undergraduate	
CollegeTier	Each college has been annotated as 1 or 2. The annotations have been computed from the average AMCAT scores obtained by the students in the college/university. Colleges with an average score above a threshold as tagged as 1 and others as 2.	
Degree	Degree obtained/pursued by the candidate	
Specialization	Specialization pursued by the candidate	

CollegeGPA	Aggregate GPA 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 AMCAT English section	
Logical	Score in AMCAT Logical ability section	
Quant	Score in AMCAT's Quantitative ability section	
Domain	Scores in AMCAT's domain module	Since different candidates give different domain-specific tests, this field captures the percentile of the candidates in their respective tests. The scores are reported on a scale of 0-1. This is an optional section for the candidates. Those opting out of it get a score of -1.
ComputerProgramming	Score in AMCAT's Computer programming section	
ElectronicsAndSemicon	Score in AMCAT's Electronics & Semiconductor Engineering section	This is an optional section for the candidates. Those opting out of it get a score of -1.
ComputerScience	Score in AMCAT's Computer Science section	This is an optional section for the candidates. Those opting out of it get a score of -1.
MechanicalEngg	Score in AMCAT's Mechanical Engineering section	This is an optional section for the candidates. Those opting out of it get a score of -1.
ElectricalEngg	Score in AMCAT's Electrical Engineering section	This is an optional section for the candidates. Those

		opting out of it get a score of -1.
TelecomEngg	Score in AMCAT's Telecommunication Engineering section	This is an optional section for the candidates. Those opting out of it get a score of -1.
CivilEngg	Score in AMCAT's Civil Engineering section	This is an optional section for the candidates. Those opting out of it get a score of -1.
conscientiousness	Scores in one of the sections of AMCAT's personality test	Normalized score with mean 0 and SD 1
agreeableness	Scores in one of the sections of AMCAT's personality test	Normalized score with mean 0 and SD 1
extraversion	Scores in one of the sections of AMCAT's personality test	Normalized score with mean 0 and SD 1
nueroticism	Scores in one of the sections of AMCAT's personality test	Normalized score with mean 0 and SD 1
openess_to_experience	Scores in one of the sections of AMCAT's personality test	Normalized score with mean 0 and SD 1

Annexure 2 – Summary Stats

```
       ID              Salary            DOJ
 Min.   :  11244   Min.   :  35000   Min.   :1991-06-01 00:00:00
 1st Qu.: 334284   1st Qu.: 180000   1st Qu.:2012-10-01 00:00:00
 Median : 639600   Median : 300000   Median :2013-11-01 00:00:00
 Mean   : 663795   Mean   : 307700   Mean   :2013-07-02 11:04:10
 3rd Qu.: 990480   3rd Qu.: 370000   3rd Qu.:2014-07-01 00:00:00
 Max.   :1298275   Max.   :4000000   Max.   :2015-12-01 00:00:00

    DOL            Designation          JobCity             Gender
Length:3998        Length:3998        Length:3998        Length:3998
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character

      DOB                          10percentage     10board           12graduation   12percentage
 Min.   :1977-10-30 00:00:00   Min.   :43.00   Length:3998        Min.   :1995   Min.   :40.00
 1st Qu.:1989-11-16 06:00:00   1st Qu.:71.68   Class :character   1st Qu.:2007   1st Qu.:66.00
 Median :1991-03-07 12:00:00   Median :79.15   Mode  :character   Median :2008   Median :74.40
 Mean   :1990-12-06 06:01:15   Mean   :77.93                      Mean   :2008   Mean   :74.47
 3rd Qu.:1992-03-13 18:00:00   3rd Qu.:85.67                      3rd Qu.:2009   3rd Qu.:82.60
 Max.   :1997-05-27 00:00:00   Max.   :97.76                      Max.   :2013   Max.   :98.70
  12board            CollegeID      CollegeTier       Degree          Specialization
Length:3998        Min.   :    2   Min.   :1.000   Length:3998        Length:3998
Class :character   1st Qu.:  494   1st Qu.:2.000   Class :character   Class :character
Mode  :character   Median : 3879   Median :2.000   Mode  :character   Mode  :character
                   Mean   : 5157   Mean   :1.926
                   3rd Qu.: 8818   3rd Qu.:2.000
                   Max.   :18409   Max.   :2.000

  collegeGPA     CollegeCityID   CollegeCityTier  CollegeState       GraduationYear    English
 Min.   : 6.45   Min.   :    2   Min.   :0.0000   Length:3998        Min.   :   0   Min.   :180.0
 1st Qu.:66.41   1st Qu.:  494   1st Qu.:0.0000   Class :character   1st Qu.:2012   1st Qu.:425.0
 Median :71.72   Median : 3879   Median :0.0000   Mode  :character   Median :2013   Median :500.0
 Mean   :71.49   Mean   : 5157   Mean   :0.3004                      Mean   :2012   Mean   :501.6
 3rd Qu.:76.33   3rd Qu.: 8818   3rd Qu.:1.0000                      3rd Qu.:2014   3rd Qu.:570.0
 Max.   :99.93   Max.   :18409   Max.   :1.0000                      Max.   :2017   Max.   :875.0
    Logical          Quant           Domain        ComputerProgramming ElectronicsAndSemicon
 Min.   :195.0   Min.   :120.0   Min.   :-1.0000   Min.   : -1.0       Min.   : -1.00
 1st Qu.:445.0   1st Qu.:430.0   1st Qu.: 0.3423   1st Qu.:295.0       1st Qu.: -1.00
 Median :505.0   Median :515.0   Median : 0.6226   Median :415.0       Median : -1.00
 Mean   :501.6   Mean   :513.4   Mean   : 0.5105   Mean   :353.1       Mean   : 95.33
 3rd Qu.:565.0   3rd Qu.:595.0   3rd Qu.: 0.8422   3rd Qu.:495.0       3rd Qu.:233.00
 Max.   :795.0   Max.   :900.0   Max.   : 0.9999   Max.   :840.0       Max.   :612.00
ComputerScience  MechanicalEngg   ElectricalEngg    TelecomEngg       CivilEngg
 Min.   : -1.00   Min.   : -1.00   Min.   : -1.00   Min.   : -1.00   Min.   : -1.000
 1st Qu.: -1.00   1st Qu.: -1.00   1st Qu.: -1.00   1st Qu.: -1.00   1st Qu.: -1.000
 Median : -1.00   Median : -1.00   Median : -1.00   Median : -1.00   Median : -1.000
 Mean   : 90.74   Mean   : 22.97   Mean   : 16.48   Mean   : 31.85   Mean   :  2.684
 3rd Qu.: -1.00   3rd Qu.: -1.00   3rd Qu.: -1.00   3rd Qu.: -1.00   3rd Qu.: -1.000
 Max.   :715.00   Max.   :623.00   Max.   :676.00   Max.   :548.00   Max.   :516.000
conscientiousness  agreeableness      extraversion       nueroticism      openess_to_experience
 Min.   :-4.12670   Min.   :-5.7816   Min.   :-4.600900   Min.   :-2.6430   Min.   :-7.3757
 1st Qu.:-0.71352   1st Qu.:-0.2871   1st Qu.:-0.604800   1st Qu.:-0.8682   1st Qu.:-0.6692
 Median : 0.04640   Median : 0.2124   Median : 0.091400   Median :-0.2344   Median :-0.0943
 Mean   :-0.03783   Mean   : 0.1465   Mean   : 0.002763   Mean   :-0.1690   Mean   :-0.1381
 3rd Qu.: 0.70270   3rd Qu.: 0.8128   3rd Qu.: 0.672000   3rd Qu.: 0.5262   3rd Qu.: 0.5024
 Max.   : 1.99530   Max.   : 1.9048   Max.   : 2.535400   Max.   : 3.3525   Max.   : 1.8224
```

Annexure 3 – R Script with outputs