

Employability Outcomes of Engineering Graduates

Palak Kothari

PES1UG19CS321

Section: E

Summary:

In this paper, I have elaborated our approach for data understanding, data preparation, predictive modelling and model evaluation for predicting salary of new graduates entering the labour market.

Data analysis is conducted on a dataset that provides employment outcomes and salaries of engineering graduates from different cities of India. The given dataset is pre-processed and cleaned. Exploratory data analysis is performed to study the various aspects related to salary outcomes. This is paired with appropriate graphs that help to visualise these outcomes. A suitable population parameter is chosen and hypothesis testing is performed.

The correlation and dependence of certain parameters are studied which could provide an insight on the factors that decide success after graduating college.

Introduction:

As per the data shared by the HRD Ministry, 2.9 million students enroll into Engineering and Technology institutions every year, and on average 1.5 million students get their degree in engineering. But less than 20 percent get employment and their job description would not include their core domain. Companies are unable to recruit merited talent in spite of new engineering colleges being set up new year. There is no dearth of B.E./B.Tech degree seats, but the quality of such colleges is quite poor and graduates seem to lack the necessary skills required to succeed in the competition intensive IT market.

To get a deeper understanding of this problem, we have chosen the Aspiring Mind's Employability Outcomes 2015 (AMEO 2015) dataset which contains engineering graduates' employment outcomes. A relevant question is what determines the salary and the jobs these engineers are offered right after graduation.

Various factors such as college grades, candidate skills, proximity of the college to industrial hubs, the specialization one has, market conditions for specific industries determine this.

Dataset:

Aspiring Minds' Employability Outcomes 2015 (AMEO 2015) is a unique dataset which contains standardized assessment (AMCAT) scores in three fundamental areas - cognitive skills, technical skills and personality and the salaries and job titles offered right after graduation. College grades, candidate skills, specialization done by the candidate are studied in order to understand how they weigh into a candidate's employability.

It includes the **Train Dataset** which contains:

5 dependent variables: Salary, Designation, date of joining, date of leaving and city in which the candidate is offered the job.

32 independent variables: Candidate biodata, AMCAT assessment scores, college GPA, 10th Board and 12th Board percentage, etc.

AMCAT is a computer adaptive test that measures job applicants on critical areas like communication skills and logical reasoning, thus helping recruiters identify the suitability of a candidate.

Data preparation and cleaning:

Data cleaning or cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. It also refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Improved data quality leads to better decision making across an organization. Good data decreases risk and can result in consistent improvements in results. Listwise deletion technique can be adopted to drop observation containing missing values. If the missing data rows are informative then we can conduct missing data imputation.

Typos, inconsistencies and redundant data:

- The dataset contains some free form self-reported text information containing inconsistent or incorrect data (e.g., 10th and 12th board names, Job city, designation, specialization). Such discrepancies and errors are handled manually.
- Inconsistent capitalization is handled. Some columns (College ID, College City ID) were considered to not contribute in our study and were thus dropped.

Handling missing data:

- NaN values in Salary column are replaced with the mean salary.
- Similar method is employed for numeric values in 10th and 12th Percentage.
- 10th and 12th Boards specify the Education Board from which the student has graduated. As it contains categorical data, the missing values can be imputed by the mode (most frequent value).

Exploratory Data Analysis:

Data analysis deals with the process of discovering useful data from the given data set to draw conclusions and to make choices. This can be done using data visualization methods.

Distribution patterns of values influence choices of appropriate statistical tests. Graphs are the most common ways to arrive at answers for statistical tests since they summarize the data visually.

In our project, we have chosen the following three graphs as means of data visualization-

Bar Graph:

A bar graph is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. One axis of the chart shows the specific genders being compared, and the other axis represents the measured frequency.

We conclude from the graph that the labour market is rooted in beliefs about average gender differences in jobs. We also observe the favouritism for one group of gender compared to the other.

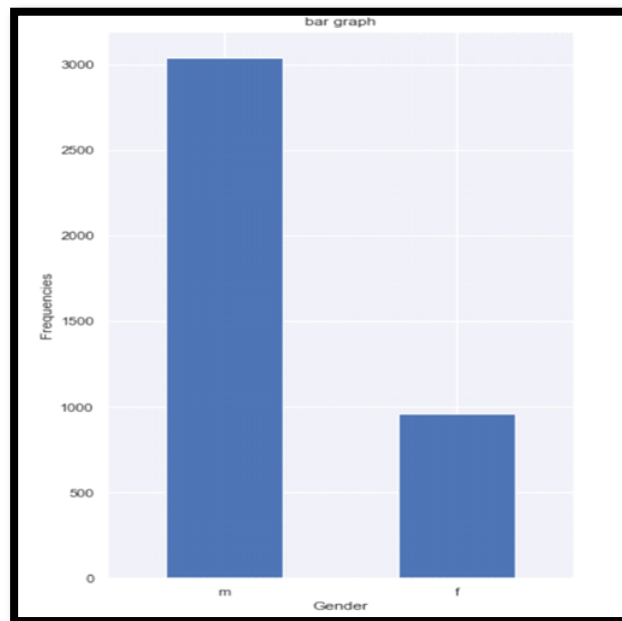


Figure 1. Bar graph of Gender vs. Frequency of higher salaries

Scatter plot:

Scatter plot is a graph of plotted points that show a relationship between two sets of data. We have taken into consideration the salaries based on a person's experience in the field._

From this graph we concluded that as the experience of the person increases their salary also increases. we can see that a cluster is formed at the bottom, and the points we see outside the cluster are some outliers where even with the same experience people earn a lot.

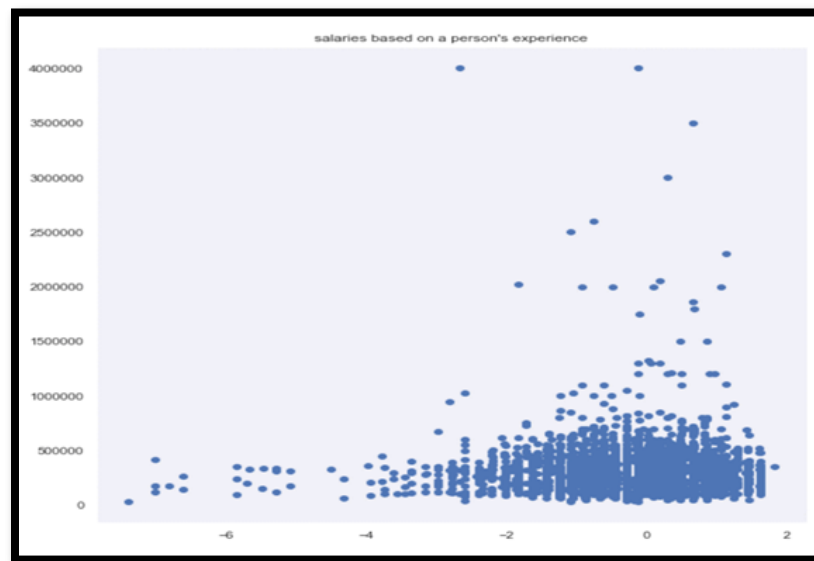


Figure 2. Scatter plot of salaries based on a candidate's experience

Violin plot:

A violin plot is a method of plotting numeric data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side. Here the data distribution is multimodal (more than one peak). Here we have compared the distribution across various personality traits: consciousness, agreeability, extraversion, neuroticism and openness to experience respectively.

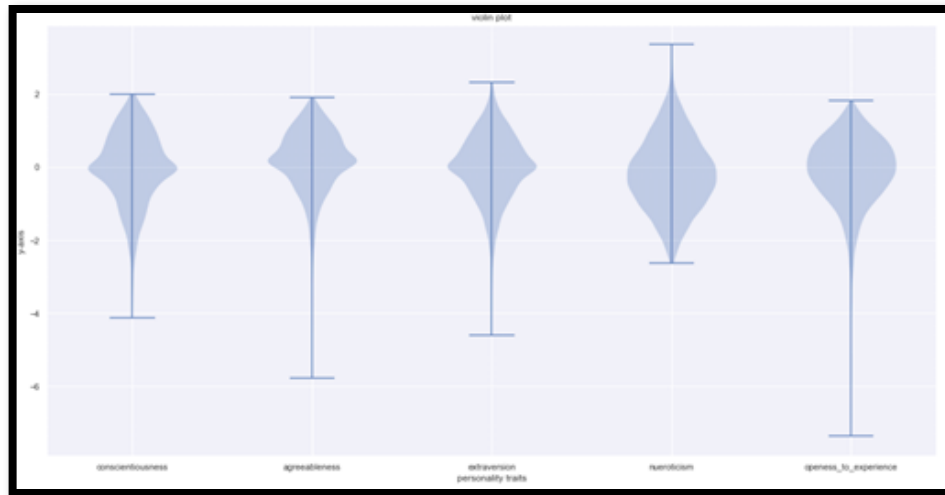


Figure 3. Violin plot showing the distribution across various personality traits

REMOVING OUTLIERS:

Outliers increase the variability in the data and hence by removing the outliers, output can be statistically significant.

Box plot shows the shape of the distribution, its Central value and its variability. With the help of a box plot we can remove the outliers.

From the box plot we see that means of both genders around the same but the inter-quartile ranges differ by a little, and hence we could say that men and women are earning around the same salary with each level of experience.

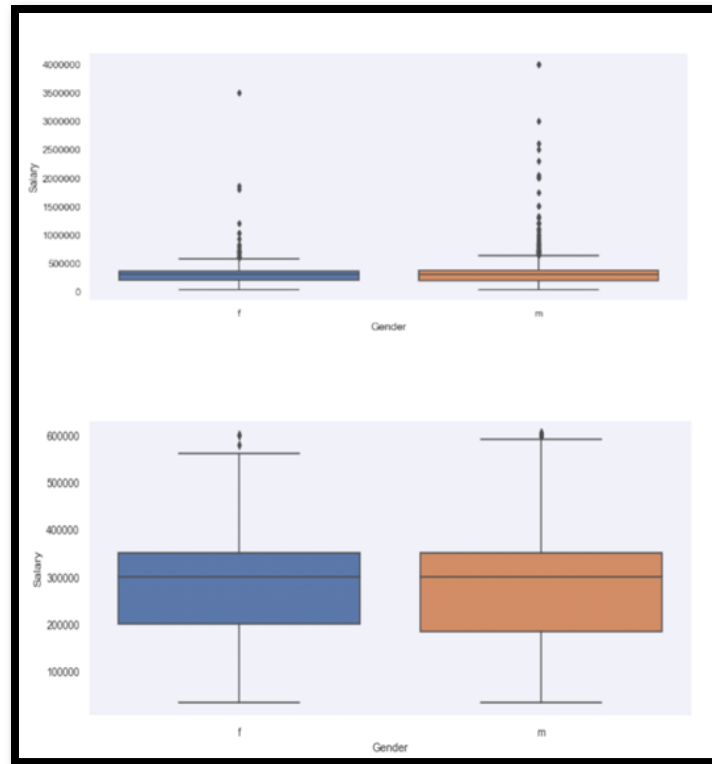


Figure 4. Comparison of Boxplots with and without outliers

STANDARDIZATION:

Standardizing a vector means subtracting a measure of location and dividing by a measure of scale. Standardizing the features around the centre and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.

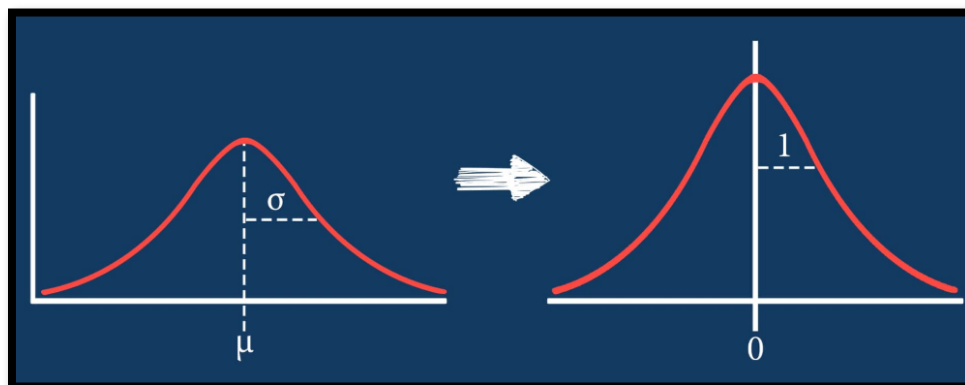


Figure 5. Standardization of a distribution with Mean = 1 and SD = 0

NORMALIZATION:

Normalizing a vector means dividing by a norm of the vector. It also often refers to rescaling by the minimum and range of the vector, to make all the elements lie between 0 and 1 thus bringing all the values of numeric columns in the dataset to a common scale.

The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

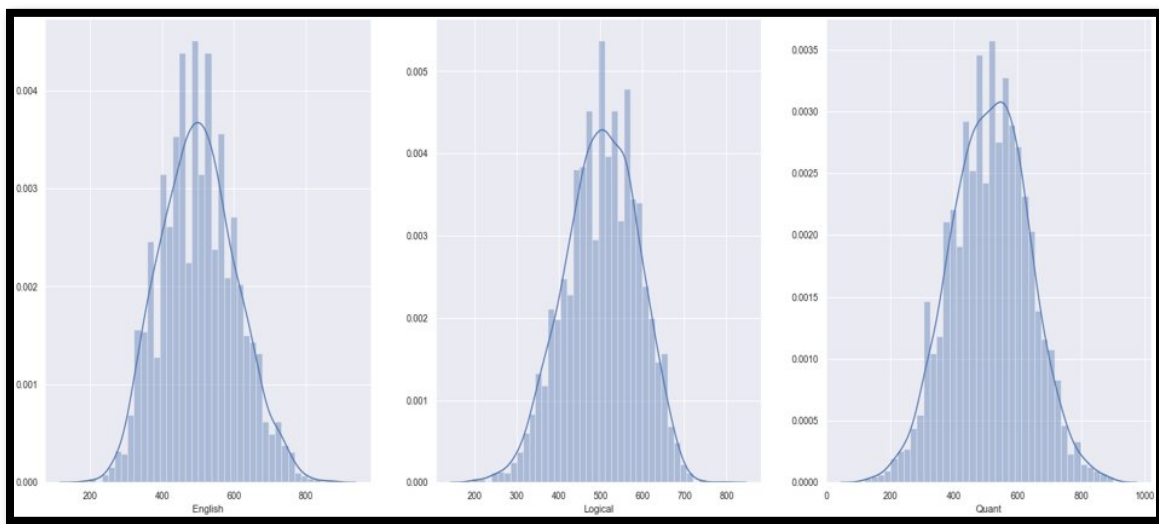


Figure 6. Normalization graphs for English, Logical and Quantitative ability scores.

Hypothesis Testing:

Involves testing an assumption about some parameter. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. In hypothesis testing, one tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.

Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analysed. One uses a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis.

The null hypothesis, denoted by **H_0** , is usually the hypothesis that sample observations result purely from chance.

The alternative hypothesis, denoted by ***H1 or Ha***, is the hypothesis that sample observations are influenced by some non-random cause.

Four Steps of Hypothesis Testing

All hypotheses are tested using a five-step process:

1. Define H_0 and H_a
2. Assume H_0 is true
3. Compute the test statistic
4. Compute the p-value of the test statistic.
5. State a conclusion about the strength of the evidence against H_0 .

The mean score of 70 scores in Computer Programming section of the AMCAT test is 366.4 with standard deviation of 205.32. A new passing grade will be introduced if it can be shown that the mean score is greater than the existing passing grade of 420

1. Defining H_0 and H_a

a. **Null Hypothesis:**

$$H_0: \mu \leq 420$$

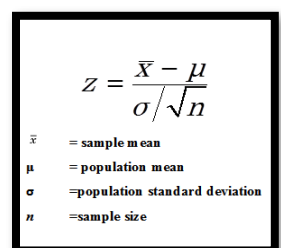
b. **Alternate Hypothesis:**

$$H_1: \mu > 420$$

2. Assume H_0 is true

Since it is a large sample and standard deviation is known, we assume the distribution is normal and use Z-Test. We assume the Null Hypothesis is true and consider $\mu_0 = 420$. \bar{X} will follow a normal distribution with mean = 420 and variance = $205.32^2 / 70$.

The following formula is used to calculate Z- score:


$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

\bar{x} = sample mean
 μ = population mean
 σ = population standard deviation
 n = sample size

3. Computing test statistic and p-value

p-value is calculated using the Standard Z-Table:


```
Sample Size = 70
CS Mean = 366.4
Standard Deviation = 205.32983552770708
Z-score = -2.184045845406384
p-value = 0.014479439716114192
```

Figure 8. Calculation of p-value

4. The result is tested against the significance value, $\alpha = 5\%$

Test the p-value for significance level of 0.05

alpha = 0.05

```
alpha = 0.05
if(pval < alpha):
    print('Null Hypothesis is rejected')
else:
    print('Null Hypothesis is plausible')
```

Null Hypothesis is rejected

Figure 9. Testing p-value against 5%

Conclusion: The result is appropriately rejected at 5%.

- 4.1. Is the test statistically significant at $\alpha = 1\%$?

Check if the test is statistically significant at the 1% level

alpha = 0.01

```
alpha = 0.01
if(pval < alpha):
    print('Null Hypothesis is rejected')
else:
    print('Null Hypothesis is plausible')
```

Null Hypothesis is plausible

Figure 10. Testing p-value against 1%

Conclusion: We failed to reject the Null Hypothesis at 1% implying that the null hypothesis is plausible.

Correlation and Dependence of bivariate data:

Correlation denotes an association(linear) between two quantitative variables. The degree of association is measured by a correlation coefficient, denoted by r . The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increase as the other increases the correlation is positive; when one decreases as the other increases it is negative.

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

In our study, we will employ a heatmap to visualize the correlation matrix. A heatmap is a tool in which values are depicted using colour.

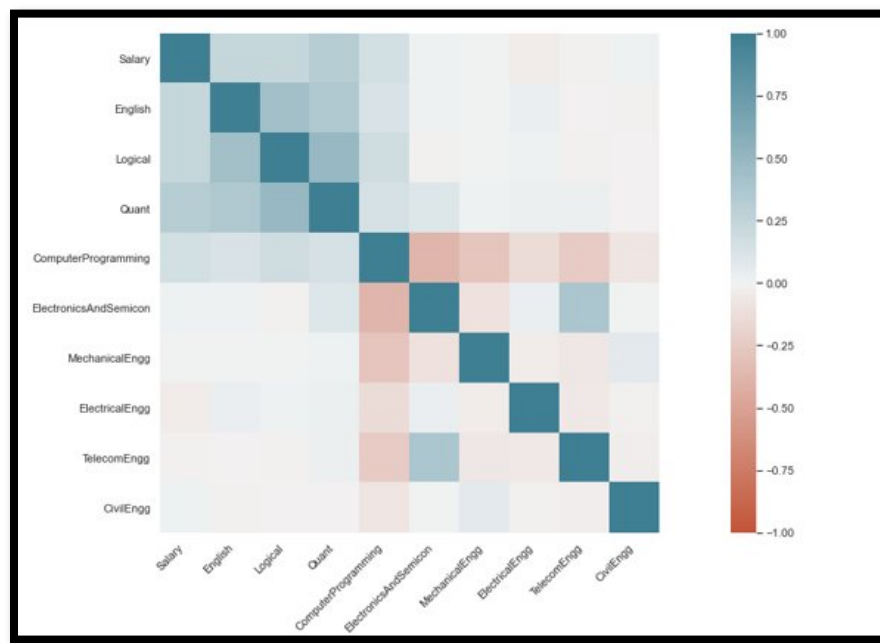


Figure 11. Heatmap representing the correlation between scores and salary

The above graph shows the relation between salaries and scores in the technical field of the AMCAT test. Stronger the colour, larger the correlation magnitude. Observation of this graph leads one to believe that higher Computer Programming scores directly relate to higher salaries.

Results:

Analysing this dataset provides vital information that can find numerous applications. Career counsellors can guide students in choosing specializations that are relevant in the current market trends. The AMCAT scores in Quantitative ability and Logical field prove to be the best

predictors for salary. Education boards can emphasize higher order thinking skills from early schooling to develop quantitative and logical aptitude.

