Aspiring Minds' Employability Outcomes 2015: Data Analysis and Visualization

Raj Choudhary*
**School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127*
*Email: raj.choudhary2016@vitstudent.ac.in **

***Abstract* – Aspiring Minds' Employability Outcomes 2015 [AMEO 2015] [1] aims to discern the entry level labor market for fresh engineers in India. The dataset holds the employability outcomes in terms of salary, job title and job location. The predictors of employability comprise of standardized scores in three fundamental areas – personality, cognition and technical skills. Other factors such as location of the college, academic grades and specialization of the candidate are also accounted for.**

This paper attempts to study and understand the typical engineering graduate from a purely vocational perspective. The important factors that influences job success are investigated. The trends and patterns in the data are analysed and visualized to later aid the design and development of machine learning models that would predict the job title or salary of a graduate. Here, the interesting insights, underlying trends and other such findings discovered by mining the data is reported.

***Index Terms*—Employability, Entry-level Jobs, Data Visualization**

1. Introduction

In this paper, the Aspiring Minds' Employability Out- comes 2015 dataset [1] is analysed and pre-processed for the machine learning pipeline, which will be able to predict the salary of an engineering graduate given his curriculum vitae. Standard data analytics and visualizations are used to understand and prepare the data. The data is cleaned of erroneous and anomalous values. The categorical features are encoded using one-hot encoding scheme. For categorical classification, the target points are placed into class bins based on quantile distribution. Relevant questions are asked and hypotheses are designed to capture the features that determine the employability outcomes. This paves the way for feature engineering to enable a good enough prediction ability. The pre-processed data is then packaged and stored for future use down the pipeline.

2. Methodology

Industry standard tools and libraries like Pandas, Scikit Learn, and Numpy are used to carry out the data analytics. Seaborn and Matplotlib are used to generate insightful visualizations that shed light on the nature of the data. Most of the work is carried out on jupyter notebooks for ease of collaboration and reproducibility.
The data pre-processing pipeline is as follows:

1) Organize Column Labels
   a) targets = ['Salary', 'DOJ', 'DOL', 'Designation', 'JobCity']
   b) features = features = ['Gender', 'DOB', '10percentage', '10board', '12graduation', '12percentage', '12board', 'CollegeID', 'CollegeTier','Degree','Specialization', 'CollegeGPA','CollegeCityID', 'CollegeCityTier', 'CollegeState', 'GraduationYear', 'English', 'Logical', 'Quant', 'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism', 'openess_to_experience']

2) Drop unnecessary columns
   a) drop_features = ['DOB', '10board', '12graduation', '12board', 'CollegeID', 'CollegeCityID', 'CollegeCityTier']
   b) drop_targets = ['DOJ', 'DOL', 'Designation', 'JobCity']

3) Replacing Standardized Test score of -1 (not at-tempted) with 0

4) Scaling CGPA to a scale of 100

5) Mapping numerous redundant specializations to categories of core engineering disciplines

6) One-Hot encode categorical features

7) Make target data (salary) categorical

After pre-processing, the data is packaged and stored for further use down the pipeline. The plotted figures store the discovered trends and insights.

3. Dataset – AMEO 2015

For every engineer, AMEO [1] dataset provides anonymised bio data information along with their respective skill scores and employment outcome

information. Specifically, the following information is available for every engineer:

1) Scores on Aspiring Minds' AMCAT - a standardized test of job skills. The test includes cognitive, domain and personality assessments.
2) Personal information like gender and date of birth.
3) Pre-university information like 10th and 12th grade marks, board of education and 12th grade graduation year.
4) University information like GPA, college major, college reputation proxy, graduation year and college location.
5) The following employment outcome information is available for every engineer: First job annual salary First job title First job location Date of joining and leaving of first job.

AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

4. Analysis and Results

Here, dataset is visualized to infer the patterns and trends. Hypothesis are formed by asking questions that reveal the nature of the data.
Candidates in the dataset can be seen performing worst as they move up the education ladder. The distribution of marks in 10th year was skewed towards the right; most students scored better than their peers. The distribution of marks in 12th year moved to a bell curve; most students performed average as compared to their peers. The distribution of college grades is slightly skewed towards the left; most students were performing worse than their peers. And finally, the distribution of salary is heavily skewed towards the left.

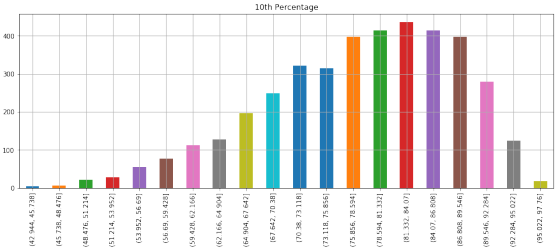

Figure 1. 10th percentage distribution

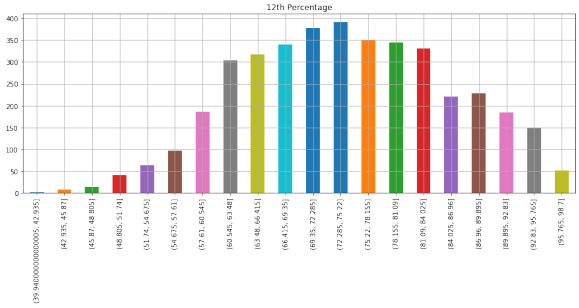

Figure 2. 12th percentage distribution

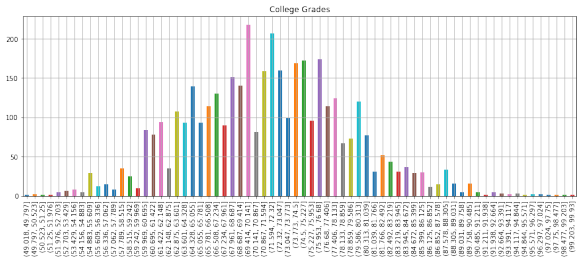

Figure 3. College grades distribution

Figure 4. Salary distribution

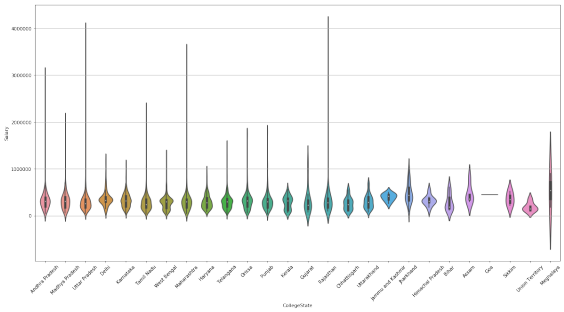
Figure 5.Does location affect salary?

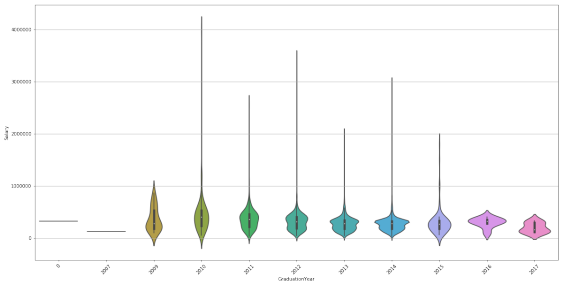
Figure 6. Does year of graduation affect salary?

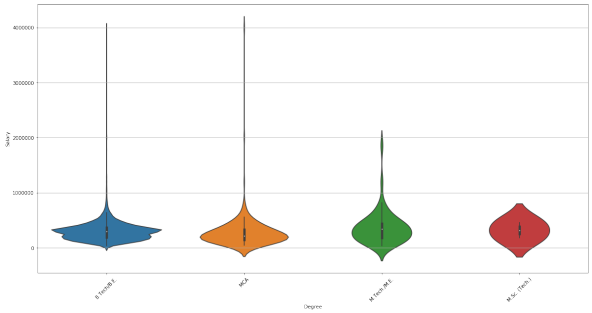

Figure 7. Does degree affect salary?

5. Conclusion

Hence, we have discovered some interesting insights into a typical engineering graduate and his/her job prospects. One of the curious observations is the large overlap of data points which will prove to be a challenge during the model development phase. The analytics can also suggest graduates methods of improving their job prospects.

6. References

[1] V. Aggarwal, S. Srikant, and H. Nisar, "Ameo 2015: A dataset com- prising amcat test scores, biodata details and employment outcomes of job seekers," in *AMEO 2015: A Dataset Comprising AMCAT Test Scores, Biodata Details and Employment Outcomes of Job Seekers*. ACM, 2016.