**** 

**Salary Prediction of Graduates**

**PGP Data Science and Engineering| Great Lakes Institute of Management June’2020**

**Under Guidance of**

Mr. Shashank Shirude

**Under-Taken By:**

Abhishek Prasoon

Apoorv Jain

Harshit Bharrdwaj

Sudeep Kadyan

Urvashi Panwar

**Domain: Salary Growth Prediction**

**DATASET SOURCE: -**

This data was first released as part of the IKDD CoDS 2016 Data Challenge.

**DATASET NAME: -**

Aspiring Minds’ Employability Outcomes 2015 (AMEO 2015)

**PROBLEM STATEMENT: -**

We have to predict annual salaries of engineering graduate and factors affecting them.

**DOMAIN/BACKGROUND DETAILS: -**

This dataset will provide ample opportunities to forecast the salary of engineering graduate. 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.

**OBJECTIVE: -**

To predict annual salaries of engineering graduate and factors affecting them with extensive use of EDA tools and Machine Learning Algorithm to find out inferences from dataset.

**ABSTRACT: -**

More than a million engineers enter the global workforce every year. A relevant question is what determines the jobs and salaries these engineers are offered right after graduation. Previous studies have shown the influence of various factors such as college reputation, grades, the field one specializes in and market conditions for specific industries. An important input which such analyses do not have is a standardized measures of job skills done at the time of completion of studies. We present here Aspiring Minds' Employability Outcomes 2015 (AMEO 2015), a unique dataset which provides engineering graduates' employment outcomes (salaries, job titles and job locations) together with standardized assessment scores in three fundamental areas - cognitive skills, technical skills and personality. Coupled with biodata information, AMEO 2015 provides an opportunity for a unique and comprehensive study of the entry level labor market. The data could be used to make an accurate salary predictor, but also understand what influences salary and job titles in the labor market. In this paper we describe the details of the dataset and discuss a spectrum of questions around meritocracy in labor markets, biases in labor selection and other prevalent market forces it can help uncover and answer.

**DATA DESCRIPTION: -**

**Numerical Features in the Dataset:**

|  |  |  |
| --- | --- | --- |
|  | | |
| **Input** | **Description** | **Comments** |
| Train/Test | Whether the data belongs to the train set or test set |  |
| ID | A unique ID to identify a candidate |  |
| **DEPENDENT VARIABLES** | | |
| Salary | Annual CTC offered to the candidate (in INR) |  |
| DOJ | Date of joining the company | Candidates may not have accurately entered information. |
| DOL | Date of leaving the company | A value of "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 |  |
| **INDEPENDENT VARIABLES** | | |
| 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 | Domain of values: [0,100] |
| 10board | The school board whose curriculum the candidate followed in grade 10 | India has several boards of education which follow their own course curricula. Schools are affiliated to one of these boards. More information here - https://en.wikipedia.org/wiki/Boards\_of\_Education\_in\_India |
| 12graduation | Year of graduation - senior year high school |  |
| 12percentage | Overall marks obtained in grade 12 examinations | Domain of values: [0,100] |
| 12board | The school board whose curriculum the candidate followed |  |
| CollegeID | Unique ID identifying the university/college which the candidate attended for her/his undergraduate | Multiple candidates may belong to the same university/college in this dataset. College names have not been disclosed to maintain privacy. |
| 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 | Relevant information here -  \* https://en.wikipedia.org/wiki/Bachelor%27s\_degree#India  \* https://en.wikipedia.org/wiki/Bachelor\_of\_Engineering#India  \* https://en.wikipedia.org/wiki/Bachelor\_of\_Technology#India |
| Specialization | Specialization pursued by the candidate | Some information here - https://en.wikipedia.org/wiki/Bachelor\_of\_Engineering#India |
| CollegeGPA | Aggregate GPA at graduation | 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. |
| 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 | The provided data set does not accurately capture state-wise distributions. |
| GraduationYear | Year of graduation (Bachelor's degree) |  |
| English | Scores in AMCAT English section | AMCAT or Aspiring Minds Computer Adaptive Test is an employability assessment test to assess the aptitude, reasoning and technical skills of students and new graduates. It is a statistical assessment platform to provide the skill measurements for each student who takes up the test. The assessments focus on three primary skills – cognitive (which include tests like English ability, Logical ability, Quantitative ability, Attention to detail), domain (which includes a test on the specific branch of studies taken up by the student) and personality (based on the big-5 personality traits).  The scores are measured and reported on a scale of 100-900 (barring the personality assessments).  For more details, see:  Link1: aspiringminds.com  Link2: research.aspiringminds.com |
| 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, we report here 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. One may consider this as missing data.  Aspiring Minds provides an internal map which relates a student's specialization to particular AMCAT tests.  In case the score is -1 for this test, you could make use of the scores in the tests mentioned below to get more information. If the scores in the tests below are -1 as well, then it can be concluded that the candidate has not given a module listed here. |
| ComputerProgramming | Score in AMCAT's Computer programming section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| 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. One may consider this as missing data. |
| 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. One may consider this as missing data. |
| 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. One may consider this as missing data. |
| 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. One may consider this as missing data. |
| 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. One may consider this as missing data. |
| 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. One may consider this as missing data. |
| conscientiousness | Scores in one of the sections of AMCAT's personality test | For a clarity on what the traits mean on which the scores are reported, see https://en.wikipedia.org/wiki/Big\_Five\_personality\_traits  The scores are sampled from a distribution with mean 0 and standard deviation 1 |
| agreeableness | Scores in one of the sections of AMCAT's personality test |
| extraversion | Scores in one of the sections of AMCAT's personality test |
| nueroticism | Scores in one of the sections of AMCAT's personality test |
| openess\_to\_experience | Scores in one of the sections of AMCAT's personality test |

**FUTURE WORK: -**

1. Data Collection
2. Basic Data Exploration
3. Data Visualization
4. Treating Anomalies (like missing values, outliers, bogus data, etc)
5. Dealing with Data Imbalance
6. Evaluating Feature Importance
7. Building Base Models
8. Tuning hyper-parameters / improving model performance/reducing over fit / under fit (if any)
9. Evaluating Performance based on various performance metrics
10. Using Ensemble Techniques to further improving model performance
11. Final Model Building
12. Drawing Inferences and Providing Recommendations.

**DECLARATION:**

This is to declare that the dataset that we are using for our capstone project is publicly available and can be used to showcase the work we do on it as a presentation in Great Learning.