



INNOMATICS[®]
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Aspiring Mind Employment Outcome 2015

About me

Background:

I hold an MSc in Data Science, with a strong foundation in machine learning, statistical analysis, and data visualization. I have hands-on experience in tools like Python, SQL, and Power BI.

Why I Want to Learn Data Science:

I want to leverage data science to solve real-world problems, gain insights from data, and contribute to advancements in AI, particularly in areas like Natural Language Processing (NLP) and automation.

Work Experience:

I have interned as a data scientist at YBI Foundation, where I worked on data analysis and model building. Additionally, I've completed various projects in forecasting, fraud detection, and e-commerce analysis.

LinkedIn Profile: [Gaurav Tailor's LinkedIn](#)

GitHub Profile: [Gaurav Tailor's GitHub](#)

Business Problem and Use Case Domain Understanding

Business Problem:

The dataset pertains to the employment outcomes of engineering graduates, including various demographic and educational features, standardized test scores, and salary information. Understanding the factors that influence the salary of engineering graduates can help universities and recruitment companies tailor their programs to improve job placement outcomes.

Use Case Domain:

The domain of this project is the employment and education sector, focusing on analyzing the outcomes of engineering students in terms of their salaries and job placements, while taking into account their cognitive, technical, and personality skills.

Objective of the Project

Objective:

The primary objective of this project is to perform Exploratory Data Analysis (EDA) on the dataset, with Salary as the target variable. The goal is to:

- Understand the factors influencing salary.
- Discover any patterns or trends between skills, demographics, and salaries.
- Validate claims related to job roles and salary levels.
- Provide actionable insights that can help improve the employability of engineering graduates.

Summary of the Data

- Dataset: Aspiring Minds Employment Outcome Dataset (2015).
- Size: Around 4000 data points and 40 independent variables.
- **Types of Variables:**
 - Categorical: Gender, Specialization, Designation, College Tier, etc.
 - Continuous: Salary, 10th percentage, 12th percentage, GPA, AMCAT scores in English, Logical, and other skill-based scores.
 - Target Variable: Salary (continuous variable).

Exploratory Data Analysis - Data Cleaning Steps

- **Handling Missing Data:**
 - Checked for missing values in the dataset.
 - For continuous variables with a few missing values, applied imputation (e.g., mean or median).
 - Dropped or filled missing values for categorical variables based on their relevance.
- **Handling Date Columns:**
 - Converted date columns like DOJ, DOL, DOB to appropriate datetime format for further analysis.
- **Outlier Detection:**
 - Used boxplots to detect outliers in continuous variables like Salary and CollegeGPA.
 - Outliers were addressed by either transformation or removal, based on their impact on the analysis.

Exploratory Data Analysis - Data Manipulation Steps

- **Feature Engineering:**
 - Created new features such as:
 - Experience (DOL - DOJ) for candidates who left jobs.
 - Age (DOB) based on current date or joining date.
- **Categorical Encoding:**
 - Converted categorical variables like Gender, CollegeTier, Specialization into numerical format using techniques like One-Hot Encoding for analysis.
- **Normalization:**
 - Applied normalization to continuous variables like AMCAT scores, GPA, and salary for better model performance and visualization.

Exploratory Data Analysis - Univariate Analysis Steps

- **Univariate Analysis:**

- Histogram and PDF: Analyzed the distribution of continuous variables such as Salary, 10 percentage, CollegeGPA, and AMCAT scores.
 - Observation: Most continuous variables were skewed.
- Boxplots: Identified outliers in salary, GPA, and AMCAT scores.
 - Observation: Significant outliers were present in the Salary column, which were carefully addressed.
- Countplots (Categorical Variables):
 - Countplots were used for categorical columns like Gender, Specialization, and JobCity.
 - Observation: Males outnumbered females in this dataset. Most graduates specialized in Computer Science.

Exploratory Data Analysis - Bivariate Analysis Steps

- **Bivariate Analysis:**

- Scatter Plots: Analyzed the relationship between Salary and continuous variables like CollegeGPA, AMCAT scores.
 - Observation: Positive correlation between GPA and salary, AMCAT scores also showed some correlation.
- Swarmplot/Boxplot: Explored relationships between categorical variables (Specialization, Designation) and salary.
 - Observation: Graduates with Computer Science specialization or Software Engineer designation had relatively higher salaries.
- Bar Plots: Compared categorical variables like Gender and Specialization.
 - Observation: No significant gender bias in terms of specialization preference, but males had slightly higher salaries.

Conclusion (Key Findings Overall)

- Salary Predictors: Higher AMCAT scores, better GPA, and specialization in fields like Computer Science positively impacted salary.
- No Strong Gender Bias: Gender did not significantly impact salary or specialization preference in this dataset.
- Experience vs Salary: Those with higher experience tended to have better salaries, but the impact of experience was lower than that of GPA and AMCAT scores.

THANK
YOU

