INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

About Me

I am a first-year student embarking on a dual journey in both computer science, driven by a fervent passion for understanding and solving complex problems. With a penchant for delving into the intricacies of data analysis, I have eagerly undertaken numerous courses to hone my skills in this domain. However, it was the opportunity to tackle my first major project that truly ignited my enthusiasm and propelled me further into the realms of data and machine learning.

Thanks to Innomatics Research Labs

1. Introduction:

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, curated by Aspiring Minds, serves as a treasure trove of information regarding the employment trajectories of engineering graduates. In this report, we delve into an extensive analysis of the dataset, aiming to unearth crucial insights into the relationship between various features and the pivotal determinant, salary. Our exploration is driven by a multifaceted objective, encompassing the comprehensive description of dataset attributes, detection of discernible patterns and trends, elucidation of inter-variable relationships, and identification of outliers or anomalies.

2. Data Overview:

The AMEO dataset comprises a myriad of pertinent information pertaining to engineering graduates, encompassing crucial variables such as Salary, Job Titles, and Job Locations, alongside standardised scores reflecting cognitive, technical, and personality skills. Boasting around 39 independent variables and a voluminous corpus of 3998 data points, the dataset traverses the realms of both continuous and categorical data, inclusive of demographic particulars and unique identifiers for each candidate.

3. Data Cleaning and Preprocessing:

3.1 Handling Missing Values:

- Preliminary scrutiny reveals a dearth of missing values within the dataset, affirming its integrity for subsequent analysis.

3.2 Columns Removal:

- A discerning eye identifies extraneous columns, including ID, CollegeID, CollegeCityID, and Unnamed: 0, which are expeditiously excised from the dataset to streamline the analysis process.

3.3 Data Type Conversion:

- An imperative step towards homogenizing the dataset involves the conversion of date columns (DOJ, DOL, DOB) from object to datetime data types, ensuring uniformity and consistency.

3.4 Handling Contradicting Dates:

- Scrutinizing the temporal aspect of the dataset reveals incongruities, where instances of DOL predating DOJ are encountered. To rectify this anomaly, such entries are judiciously expunged from the dataset to maintain coherence.

3.5 Handling Illogical Values:

- Certain columns exhibit illogical values such as 0 or -1, necessitating remedial action to restore data integrity before delving deeper into analysis.

3.6 Removing Columns with High Missing Values:

- Columns plagued by an overwhelming proportion of missing values (75-80%) are deemed inconsequential and hence are systematically purged from the dataset to mitigate their potential confounding influence.

3.7 Imputation of Missing Values:

- The process of imputation is diligently undertaken, with missing values in categorical columns being replaced with the mode, while those in numerical columns are imputed with the median to preserve the statistical robustness of the dataset.

4. Exploratory Data Analysis (EDA):

4.1 Univariate Analysis:

- In-depth scrutiny of individual features unfolds their distributional characteristics, unearthing nuances within continuous and categorical variables through an array of visualizations and summary statistics.

4.2 Bivariate Analysis:

- The exploration of relationships between variables unveils intricate correlations and associations, offering insights into potential patterns and trends lurking beneath the surface of the dataset.

5. Research Questions:

5.1 Salary Claims Analysis:

- Rigorous statistical tests are wielded to scrutinize the veracity of salary claims associated with different job roles, with the outcome shedding light on the divergence between actual salary distributions and anticipated benchmarks.

5.2 Gender and Specialization Association:

- The association between gender and specialization preferences is meticulously scrutinized through sophisticated statistical analyses, unraveling the intricate interplay between gender dynamics and career preferences.

6. Conclusion:

This exhaustive analysis casts a spotlight on the multifaceted landscape of employment outcomes for engineering graduates encapsulated within the AMEO dataset. Through a judicious amalgamation of data cleaning, preprocessing, and exploratory analyses, invaluable insights have been gleaned into the intricate web of factors shaping salary distributions, career trajectories, and gender-specific specialization preferences.

7. Recommendations:

7.1 Further Analysis:

- Future endeavors could delve deeper into the temporal dynamics of employment outcomes, conducting longitudinal studies to trace the evolution of career trajectories and salary trajectories over time.

7.2 Policy Implications:

- Policymakers could leverage the insights gleaned from this analysis to craft targeted interventions aimed at fostering gender diversity and inclusivity across specialized domains, thereby engendering a more equitable and inclusive professional landscape.