Project Proposal

Advanced Python for AI and ML Tools (AML – 2203)

# Rationale

The Information Technology Industry grew rapidly in the past few decades, and we have noticed people from all over the world, especially the Indian subcontinent have pursued advanced degrees in the discipline and ended up finding high rewarding jobs. In recent years, we have noticed a rapid growth in the sector which has added more domains like database engineering, android/iOS development, machine learning, etc.

Because of this exponential growth of the IT industry, it is obvious for the current youth to be inquisitive as to what specific field in IT they should pursue, which companies they should aim for, and what cities they should try to move into. Therefore, our ambition for this project is to apply the tools and techniques we learned in this course to the proposed dataset and gather insights that would help the aspiring IT professional to make an informed decision based on real data.

# Data Description

The data comprises more than 22,000 instances of data collected from software professionals in various companies, and includes the following features:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Type** | **Description** |
| Rating | Float | This field shows ratings provided to companies by employees. The ratings range between 1 (the lowest) and 5 (the highest) |
| Company Name | Object | This field states the name of the company |
| Job Title | Object | This shows the title of the job held by the professional |
| Salary | Integer | This field shows the monthly salaries reported by the professionals |
| Salary Reported | Integer | This field shows the number of professionals who reported their salary level from a specific company |
| Location | Object | This field provides the names of the cities where the company is located |
| Employment Status | Object | This field provides information on the type of contractual agreement between the professional and the company he works for. It describes whether the professional is working as a full-time employee, intern, contractor or just a trainee. |
| Job Roles | Object | This field shows the specific domains the professional is working in, which include Android, backend, database, frontend, IOS, Java, Python, etc. |

Figure : Data Description

# Proposed Methodology

The following is a chronological order of activities that the team will be carrying out to prepare the data for analysis:

## Data Collection

The team will download this data from Kaggle.com in raw format. This will ensure that the team members have the opportunity to practice the data cleaning skills acquired during this course.

## Data Cleaning

Data cleaning will involve the detection and treatment of missing/null values and outliers. The purpose of this exercise is to attain normalcy in the dataset as far as possible.

## Exploratory Data Analysis (EDA)

This exercise will entail checking the shape of the data, understanding the data type of each feature and checking for multicollinearity within the features. Furthermore, the section will include creating various charts to better understand the behaviour of various features. The team will make use of univariate, bivariate and multivariate graphs. Furthermore, the team will apply various statistical measures on the dataset to measure the central tendency and dispersion of data.

# Intended Experiments

The primary analysis, based on which the team intends to collect insights, will be to apply a multivariate regression on the proposed dataset and identify features that have the greatest influence on the salaries of Software Professionals. To measure the accuracy of the regression model the team will be using R2 score which shows the amount of variation in the data that the model can explain.

# Planning Milestones

The following figure shows how team members have decided to divide the tasks among themselves. It also shows a route map toward the final completion of the project.



Figure : Task Distribution and Timeline

# Group Members

* HARDIK SOLANKI: CXXXXXXX
* KATLEEN ORATA: CXXXXXXX
* MOHAMMAD IMRAN UDDIN: C0800487
* TOM JAMES: CXXXXXXX