**Executive Summary**

**Business Problem**

The business, which is a contracting firm which offers technology and scientific solutions has expressed the need to be more competitive in the hiring market. It has highlighted two specific goals:

1. To determine the industry factors that are most important in predicting the salary amounts for these data.
2. To determine the factors that distinguish job categories and titles from each other.

**My Objectives**

As a student in this Data Science Immersive Course in General Assembly, I have looked into data related listings in order to extract the necessary information needed to address the business goals. I have set a few objectives which are:

1. Collect a set of job listings and information; process and clean it into a useful data set
2. To predict estimated salary levels using features generated from the job listings
3. To understand what factors set apart a data scientist or analyst from a data-related job

**My Workflow**

This project was separated into four notebook files. The first two notebook files performed web scraping activities while the later two was responsible for data cleaning & analysis need to answer the main objectives.

1. **STEP 1: Run 01-webscrape-pages.ipynb** - This extracts details from the Search Results Page.
2. **STEP 2: Run 02-webscrape-details.ipyn**b - This uses outputs from STEP 1 and extracts details from the Job Application webpage.
3. **STEP 3:** Move all output CSV files to data directory.
4. **STEP 4: Run 03-data-val-and-cleaning.ipynb** – This Reads all output files from STEP 2, cleans the data and creates final output final\_webscrape.csv in the data directory
5. STEP 5: **Run 04-data-modelling.ipynb** – This notebook file will Read the output from STEP3, run the models needed to answer questions 1 & 2 of the project.

**Data Background & Information**

To limit our scope we will be focusing on extracting information from www.mycareersfuture.sg.

We have limited our scope to finding job listings based on Singapore only.

**DATA DICTIONARY**

|  |  |  |
| --- | --- | --- |
| **Feature Column** | **Data Type** | **Description** |
| Job Title | Categorical | Job Titles of the Job Application |
| Company | Categorical | The company that created the Job Application |
| Employment Type | Categorical | Type of Employment (e.g. Permanent, Contractual, etc) |
| Seniority | Categorical | e.g. Professional, Executive, Internship, etc |
| Industry | Categorical | Industry the Job Application belongs to (e.g. Information Technology, Finance, etc) |
| Minimum Salary | Numeric | Minimum Salary found in Salary Range |
| Maximum Salary | Numeric | Maximum Salary found in Salary Range |
| Salary | Numeric | This is the Average of the Minimum and Maximum Salary |
| Job Description | Categorical | Description of general tasks, duties & responsibilities of a position |
| Job Requirements | Categorical | Qualification and skills needed for a certain position |

**Data Validation & Data Cleaning**

1. **Removing Duplicate Rows**. Since a lot of key words used to search job applications are similar (e.g. DATA ENGINEER VS DATA SCIENTIST VS DATA), removal of duplicate records were done. Used the python command drop\_duplicates and those records that have the same links.
2. **Cleaning Fields with Null Values.** Feature columns with blank values were cleaned, or dropped depending on the data assumptions.
3. **Job Posts with Undisclosed Salary.** These records were dropped as well.

**Feature Engineering.**

The most important Variable that was feature engineered is the Salary Amount. We were only able to extract a salary range which contains a minimum and maximum salary and thus had to extrapolate the salary amount by computing the average between the minimum and maximum salary. Please note that records that had blank salary range as well as records that undisclosed salary were not included in this project.

**Question 1: Predicting Salary Amounts**

Several Models were executed to help answer this question. We tested our data using Logistic Regression, Random Forest Classifier (RFC), Decision Tree Classifier, SVM as well as Regression Models. After comparing results, RFC was found to have the best result among the models.

**Question 2: Predicting Job Category Factors**

Several Models were executed to help answer this question. We tested our data using Logistic Regression, Random Forest Classifier (RFC) & Decision Tree Classifier. After comparing results, RFC was also found to have the best result among the models.