Exploring the Data Job Market: An Analysis and Prediction of Job Listings and Salary Trends

Milestone: Final Project Proposal

Group 7
Valli Meenaa Vellaiyan
Hrithik Sarda

857-832-0123
978-654-0445

vellaiyan.v@northeastern.edu
sarda.h@northeastern.edu

Percentage of effort contributed by student 1: _____50_____

Percentage of effort contributed by student 2: _____50_____

Signature of student 1: _____Valli Meenaa Vellaiyan_____

Signature of student 2: _____Hrithik Sarda_____

Submission date: _____3rd February 2023_____

Problem Setting

The problem setting is as follows: analyze and understand the current job market for Data Scientists, Data Engineers, and Data Analysts, such as identifying:

- The most in-demand skills and qualifications
- The most common industries hiring data scientists/engineers/analysts etc.
- The regions or cities with the highest concentration of data job openings
- The average salary range for various positions
- The most common job titles
- The most common words and phrases used in job descriptions to identify specific tasks and responsibilities that are most sought after by employers.
- Most importantly, to use the job descriptions to develop a model that can predict the job's salary range, given the job description.
- Additionally, it can also be used to build a model that can predict the job industry, given the job description.

A variety of techniques such as data cleaning, data visualization, statistical analysis, and machine learning will be used to understand and provide necessary insights from the dataset.

Problem Definition

The project's goal is to solve the issue of unreported salaries for data scientists, data analysts, and data engineers. Companies can also use this methodology to determine the appropriate starting compensation for new hires. A model is developed based on various criteria to group firms for staff transfers. The projected pay from the model might help employees choose the next ideal position. The project uses job descriptions as input to forecast job attributes such as income range, industry, job title, and other aspects of the work to study and understand the employment market for data scientists, engineers, and analysts. In summary, this model uses machine learning techniques to analyze and understand the job market for data scientists/analysts/engineers by predicting prospective job characteristics.

Data Sources

The datasets for “Data Analyst Jobs”, “Data Scientist Jobs”, and “Data Engineer Jobs” have been taken from https://www.kaggle.com/, an open-source, secure online community, which allows users to browse through numerous datasets. The links for the three datasets are provided below:

i) https://www.kaggle.com/datasets/durgeshrao9993/data-analyst-jobs-datset
ii) https://www.kaggle.com/datasets/andrewmvd/data-scientist-jobs
iii) https://www.kaggle.com/datasets/andrewmvd/data-engineer-jobs

Data Description

After combining the three datasets, the final dataset contains 16 columns (15 attributes and 1 target variable - “Salary Estimate” and 8676 rows. The dataset contains job listings for data scientist roles, data analyst roles, and data engineer roles. It includes several fields such as job title, company name, location, salary, and job description. All the attributes and their descriptions are given below:

No.	Attribute	Description
1.	ID	To identify each record uniquely
2.	Job Title	The title of the job listing
3.	Salary Estimate	The salary range for the job, and this is our response variable
4.	Job Description	A text description of the job responsibilities and qualifications
5.	Rating	Job rating out of 5
6.	Company Name	The name of the company that posted the job listing
7.	Location	The location of the job
8.	Headquarters	Headquarters of the company providing the respective job role
9.	Size	Number of employees working at the company currently
10.	Founded	The year in which the company was founded
11.	Type of Ownership	Type of business [Public, Private, Sports, etc.]
12.	Industry	The industry in which the company operates
13.	Sector	The sector to which the industry belongs to
14.	Revenue	Total income of the company.
15.	Competitors	Other companies that are current competitors to the listed company
16.	Easy Apply	Tells if it is easy to apply or requires specific reference/connections