IE 7275: DATA MINING IN ENGINEERING

PREDICTION OF DATA BASED JOB ROLES IN THE USA

ADITI NAMDEO

SNEHA SUBRAMANIAN

namdeo.a@northeastern.edu

subramanian.sn@northeastern.edu

Percentage of efforts by Aditi: 50%

Percentage of efforts by Sneha: 50%

PROBLEM SETTING:

Data Science has gained popularity over the past few years but has grown even more during the pandemic. With millions of gigabytes of data being generated every day, companies are now taking the data driven decision making approach to make decisions for the company. Data Scientists, Data Analyst, Data Engineers and Business Analysts are in huge demand to organize, work and pull-out insights from this large data.

PROBLEM DEFINITION:

The objective of our project is to predict salary based on information collected from these job roles 2 years ago. With this data set we aim to help those like myself and Aditi who are looking for one of these job roles and try and bring all the needed information in one place so that others could use it too. We also aim to find the trends on how the pay changes based on the company size, location, etc. We are looking to answer the following questions:

- The best jobs and salaries in the desired location across the US
- What kind of companies best fit my skillset?
- What are the companies looking for in an ideal candidate?
- How many job roles are currently available in the market?

DATA SOURCE: This data has been obtained from GitHub and was created by picklesueat.

DATA DESCRIPTION:

The data set consists of 4 CSV files which consist of information related to Data Scientists, Data Analyst, Data Engineers and Business Analyst. Each of these CSV files contain the job role, salary, job description and location for each company. Using matplotlib and seaborn we would be creating useful insights on these profiles and using NumPy, pandas, sklearn and a couple of other packages we would analyze and predict the salary.