MA678 Midterm Project Proposal

Shicong Wang

11/4/2021

Personal Statement

After grasuation, I plan to work in a tech company as a data scientist.This project is closely related to my future work.

The data set I select is "Data Science and STEM Salaries", based on which I plan to build a "A two-way selection system for graduates and companies" model. As every year plenty of graduates eager to obtain STEM job offers after graduation, this data analysis will hopefully serve as a reference for graduates' future career options, since many of the companies in the data set are their wished-for companies and job occupations. The model helps for either choosing companies that satisfies with various motivation and preferences by figuring out the distribution of salaries and positions, or making graduates more in line with the requirements of these companies and some of the job positions by analyzing the preferences of company employments. As a result, it's a good chance for me to learn more about these companies, identifying jobs that are high-probability matches to me.

Questions

Given on the information in the data set, I raise following questions:

1. What are the factors affecting salary?(Analyze the relationship between salary and other variables, such as education level, rank, position, or even gender and race?)
2. Whether there exist differences in the salaries of each top company? What is the salaries distribution?
3. Are there any characteristics of the location distribution of the companies in the data set? What enlightenment will this give future job seekers who graduate?
4. What are the differences in salaries among positions? Does each position have a preference when recruiting personnel? (Analyze the relationship between positions and gender, education level etc.)

The data source

Data science and STEM salaries

This data set contains useful information such as education level, compensation (base salary, bonus, stock grants), race, and more. These 62,000 salary records are all from top companies.

Proposed Timeline of work

As it's a tough work, the time arrangement exerts vital importance in this project.

- EDA: Nov.1st to Nov.7th
- Data Processing: Nov.8th to Nov.15th
- Modeling and Validation: Nov.16th to Nov. 23rd
- Write up: Nov.24th to Nov.30th