Spencer Mui - sm3061

Data Science Salary Prediction Final Report

Objective

The goal of this project was to forecast and predict future salaries in the data science field using historical data from recent years. This in turn can predict the growth or decline of the field using this prediction model, if proven to be accurate.

Importance

This dataset contains information about a trending career field with many more opportunities popping up today. In particular to the class, the data gives the class insight on the potential future careers they may have. The overall project also contains many elements used in the class, creating an all encompassing review for the class.

Outcomes

From the conception of the project, the target was to create an accurate model for data science salaries. This prediction model could be used by many prospecting data science students or people who are seeking a career change. For most people, the field is only important for work and the biggest thing is the salary involved with the job. If the salaries are predicted to be strong in the field and the amount of jobs is increasing, more people would be interested in moving to data science. This predictor was expected to show if the outlook is positive. Ideally the predictions would show job growth and salary growth. (Job growth was not added as a predictor in this model as relevant data would have to be sourced and for many job boards, scrapping is against the terms of service, therefore the structure for this is included but no real data is added.)

Overview

Data science salary predictions is a very difficult real-world task. There are nearly countless factors that can affect job growth and stability in this field. This predictor was set out to slim down the factors and keep it simple, will salaries seemingly increase and how likely are the included factors going to change the outcome? The biggest factor included in the project is experience. As with many industries, experience is key to a single person’s salary and position. From the original dataset, it showed a slight amount of variation year over year. After completing the project, the outlook is pretty similar to the other datasets. Running the prediction a few times shows that the trends stay true. In this project, the data set is uploaded to an sql database. The data is then reread into

Python and interpreted. Two graphs are shown that give a visual representation of salary information. The prediction algorithm is then created using data from the prior dataset. It uses linear regression to interpret the income based on the position. After the algorithm is created, the data then creates prediction test cases and potential error scores are calculated. The prediction test cases are then drawn into graphs that are of the same type as the other graphs. The user can then interpret all of the results to understand where the data science career field may take them.

The dataset used is (https://www.kaggle.com/datasets/yusufdelikkaya/datascience-salaries-2024). Scraping job boards goes against their terms of service so it will not actually be included in the dataset, however I will add it as a part of steps for data collection and handling to simulate what the process would look like but will not implement it in the actual project. Some manual entry of data from these boards will be used to add more usable data.

* - *Denotes pieces of the project that will not be implemented but explained*

Steps

1. Data Collection

- Extract data from Kaggle data set
- Inject random unrelated variables/missing variables into dataset (for data cleaning)
- Gather data from job boards *
    - Including job board listings would add more validity to the data set and add more data to interpret

2. Data Handling

- Error handle potential defects from csv file
- Load data into Pandas dataframe
- Remove extraneous data
- Handle missing values in Pandas by removing data that cannot be interpreted
- Delete duplicate data from job boards*
    - Job boards have many incorrect values and variables, many of which are not important for interpretation or completely invalidate the dataset. For example, job postings for internships that are actually full-time jobs and vice-versa can be incorrectly listed, leading to incorrect data and models.

3. Databases

- Design SQL database to perpetually store collected data

    - Create Job_Posting to store data of job market, contains data such as posted date and job type * (Will be entered into the SQL database but will not contain any data inside)
    - Create Historic_Jobs to store data of disclosed salaries and relevant jobs information
- Create structure to store collected data into SQL
- Load existing data into the new database
- Verify correctness of loaded data
    - Checks to make sure all the data entered into SQL is accurate, to test this, query SQL and check the first 3 lines are entered correctly

3. Modeling

- Calculate mean and variance for data
- Remove data outliers
- Create Z-Score normalization of data
- Split dataset into testing and prediction sets
- Use linear regression to calculate prediction model
- Visualize current data
    - Show the job markets with data science

4. Analysis

- Use model to predict salaries over the next 5 years
- Visualize prediction showing both historic and new data points
    - Create graphs to visualize the growth of jobs and salaries in the industry
    - Show the countries with a high demand for data science employees
- Produce R2 Scores
- Calculate the Mean Squared Error
- Compare the historical data to the current prediction model

Conclusion

Based on the original hypothesis, the salaries are expected to grow over the next few years. After completing the algorithm, salaries are stagnant. One thing that may affect this is the data volume. Inside the used dataset, the set only goes back 3 years. While the growth is not significant, this is also a sign of a slowing job market and salaries due to many other external factors. Varying the prediction set yields many similar results to what was found in the original data set. As for a project, the project was very successful. It was designed to show solid statistics of salaries in the data science field, and was able to help discern the potential growth of the career field as a whole. Overall, the while

the results did not line up with the hopeful hypothesis, the project was successful and was able to help users draw conclusions about data science as a career path.