Cole Biehl, Johnny Huang, Jonathan Fong

14 February 2024

CSCI 183, Data Science

Project Proposal: Machine Learning for Salary Prediction

Data Set:

This “Salary Prediction” dataset contains a list of job postings from Glassdoor.com from 2017-2018. The dataset includes information such as job title, salary estimate, job description, rating, company name, location, headquarters, size, industry, sector, revenue, competitors, python_yn, spark, and various related features.

Project Idea:

In today's competitive job market, understanding the determinants of salary can empower both employers and job seekers. Employers can set competitive salary ranges to attract top talent, while job seekers can negotiate better compensation packages. By leveraging the "Jobs Dataset from Glassdoor," we aim to uncover the patterns and trends that influence salary levels across various industries and positions.

Our primary objective is to develop a predictive model that estimates the salary for a given job listing. By analyzing the relationships between job characteristics (such as industry, location, and company size) and their associated salaries, we will identify the key factors contributing to salary variations. This insight could guide job seekers in targeting their job search and help employers in benchmarking salary offerings.

Proposed Methodology:

Software We Will Be Using

Python: As our primary programming language, given its extensive support for data science and machine learning.

Pandas and NumPy: For data manipulation and numerical computations.

Scikit-learn: To implement various machine learning models, including KNN, as well as for data preprocessing and model evaluation.

Matplotlib and Seaborn: For data visualization and exploratory data analysis.

Jupyter Notebooks: For interactive development and documentation of our analysis.

Data Pre-processing:

- Normalization: Essential for KNN, as this algorithm relies on distance calculations. We will normalize our features to ensure that all features contribute equally to the distance computation.
- Feature Engineering: We will explore creating new features that might improve our model's predictive capabilities. For instance, converting company location into coordinate to better calculate the difference for KNN algorithm.

Analysis:

- K-Nearest Neighbors (KNN): We will introduce KNN as one of our primary predictive models. KNN works by finding the 'k' closest training examples to a given test point and making predictions based on the majority label of these neighbors. For salary prediction, we will use a regression variant of KNN, which predicts the output based on the average of the 'k' nearest neighbors' values.

- Optimizing 'k': The choice of 'k' (number of neighbors) is crucial for the model's performance. We will use cross-validation to find the optimal 'k' value that minimizes our error metric.
- Feature Selection: Given that KNN is sensitive to irrelevant features, we will perform feature selection to identify the most significant predictors of salary.

Evaluation:

- Mean Squared Error (MSE): To evaluate the average squared differences between our predictions and actual values.

Papers to Read:

1. A Data-Driven Approach to Salary Prediction: This paper explores how machine learning techniques can extract meaningful patterns from job descriptions and company information to predict salaries accurately.
2. Skills-Based Salary Prediction for Freshers: This paper presents a model using linear regression and sentiment analysis to predict salaries for new graduates in India, highlighting its utility for both job seekers and companies.
3. Impact of firm characteristics on wages: This paper examines the influence of industry and firm size on wages, finding that these factors significantly affect wages, with human capital factors only partially explaining the differentials.