Team members:

Aditya Yogesh Nair

Alan Birchler De Allende

Qin Zheng

Engineering Salary Predictor

The purpose of our project will be to perform predictions on what the yearly average salary of an engineer living in the United States will be based on multiple features. This model would be integrated in a job data platform and used for student engineers who are performing research on what field they would like to specialize in.

A large number of e-recruitment websites are helping job seekers set their expectations through salary estimating systems. However, it is difficult to know the methods they are using and the accuracy of their predictions.

From a research perspective for this problem we’re tackling, in 2016 scholars from Thailand made an experiment to compare five data mining techniques including Decision trees, Naive Bayes, K-Nearest neighbor, Support vector machines, and Neural networks to find the suitable technique for salary prediction. In addition, in 2018 scientists from India proposed a prediction engine that provided a suitable and justified salary for any job by applying decision tree and ensemble models.

The feature vector, denoted as **x**, will include attributes of post-graduated engineers (whose data will be scraped from the aforementioned datasets) such as (but not limited to): the highest degree level (i.e. bachelor’s) that was obtained and the institution that an engineer obtained it from, the city that they live in, the field of engineering that they are in, what specific job they have (i.e. project manager or technical lead), and how long they have been working at said job. The output label, denoted as **y**, will be the predicted value of the yearly average salary that an engineer can expect based on the input data given in feature matrix **X**.

For the model itself, we frame it as a supervised learning problem and we will be implementing Linear Regression, Decision Trees and Support Vector Machines. Besides predicting an average salary, other experiments we’d like to conduct are testing each parameter to see which ones have a higher correlation and determining if either root mean squared (RMSE) or mean absolute error (MAE) would be the better evaluation metric. If given time, we would like to venture to see if we can predict income in fields beyond engineering/tech as well.

For the data that we will be using to train, validate, and then test our model, we are planning to use datasets from a research study conducted by Deloitte as a basis for what our primary dataset will constitute while integrating the raw data of the study’s parent datasets.

Citations

**Previous research:**

*[Implement of salary prediction system to improve student motivation using data mining technique] (*[*https://ieeexplore.ieee.org/abstract/document/7951419*](https://ieeexplore.ieee.org/abstract/document/7951419)*)*

*[Design of a novel Prediction Engine for predicting suitable salary for a job] (*[*https://ieeexplore.ieee.org/abstract/document/8718711/*](https://ieeexplore.ieee.org/abstract/document/8718711/) *)*

**Example datasets:**

[Engineering] (<https://datausa.io/profile/cip/engineering>)

[United States Census Bureau] (<https://data.census.gov/mdat/#/search?ds=ACSPUMS5Y2020>)