An analysis of starting career salary and mid career salary by Matt Xi and Nick Frasco

Introduction

The goal of this project is to assess trends in starting median salary vs mid-career salary. As we are beginning our careers, it is important to consider how your starting salary impacts your salary in the future. Do you take a 6 figure job in California with a higher living cost or a 5 figure job in Chicago with a lower living cost. How does your starting salary impact your salary mid career? This problem is important to consider as a college graduate during one of the toughest economic times in the past decade. We will be analyzing the dataset from Kaggle, provided by the Wall Street Journal. It contains three different spreadsheets; 'degrees that payback', 'salaries by college type', and 'salaries by religion'. We'll be using the 'salaries by college type,' spreadsheet for our analysis. This spreadsheet contains an array of different colleges as well as the type of college that it is, starting salary, mid-career salary, and different percentiles of the mid salary pay.

We decided to use linear regression for our machine learning model. Intuitively, the higher the starting salary, the higher the mid career salary, so this was the best option in terms of predicting other values. Furthermore, we decided to see how the linear regression model performed against other types of classification models. For this, we constructed an MLP regressor with a relu activation function and an adam solver, as well as a decision tree model. For all three of our models, we tested with and without 5 fold cross-validation error to see which would perform the best.

Analysis of preliminary data

Although our data was easy to read, it still needed some preprocessing. We had a few columns with 'NaN' that we had to fix. Being as it is a salary, we just made our best guess at these and kept them all consistent so we didn't lose any validity in the data. There were also dollar signs next to each of the numeric values, which we had to get rid of. Just from looking at the data, it is easy to tell why linear regression would most likely be the best choice in terms of a learning model. This is because most salaries go up as you reach your mid-career. We wanted to find out just how much, but more specifically if we could accurately predict this rise in salary.

Evaluation of model

After evaluating our three models (both with and without 5 fold cross-validation) we found that the decision tree with 5 fold cross-validation performed the best with an accuracy of 95%. Linear regression performed a bit worse than we had expected, with an accuracy of 78% (71% with 5 fold validation). And the MLP performed the worst at 77% (69% with 5 fold validation). We tested an array of different hyperparameters to find which would give us the best results given our dataset. We believe the reason why the decision tree outperformed the other models is because of the few outliers in the data. When there are substantial outliers in the data, linear regressors don't perform as well while decision trees aren't affected quite as much.

Contribution

Matthew Xi: created mlp model,preprocessed data, wrote half report, created linear regression model

Nick Frasco: created decision tree model,found the data, wrote other half report, created 5 fold accuracy