Multilinear Regression Analysis on the Salary Dataset

WMASDS Batch 09, Class Id: 20229011, 20229021, 20229028, 20229030

Objectives:

- To analyze the relationship between the independent variables (Years of Experience, Education Level, Position) and the dependent variable (Salary) in the Salary dataset using multilinear regression.
- To evaluate the performance of the multilinear regression model on the test data and make predictions on new samples.

Data:

The Salary dataset contains information about the salaries of employees in a company. The dataset includes three independent variables (Years of Experience, Education Level, Position) and one dependent variable (Salary).

Methodology:

We used Python and various libraries such as pandas, numpy, seaborn, and scikit-learn to perform multilinear regression analysis on the Salary dataset. We began by loading the dataset and performing exploratory data analysis (EDA) to gain a better understanding of the data. We then preprocessed the data by encoding the categorical variables using onehot encoding, splitting the data into training and testing sets, and scaling the independent variables using standardization.

Next, we fit a multilinear regression model on the training data and made predictions on the test data. We evaluated the performance of the model on the test data using mean sqared error (MSE) and R-squared (R^2) scores. The MSE measures the average squared difference between the predicted and actual values, while the R^2 score measures the proportion of variance in the dependent variable that is explained by the independent variables.

Findings:

Our analysis revealed that there is a strong linear relationship between the independent variables (Years of Experience, Education Level, Position) and the dependent variable (Salary) in the Salary dataset. The multilinear regression model achieved an MSE of ‘12823412.298’ and an R^2 score of 0.988 on the test data, indicating that the model has a good fit and is able to explain 98.8% of the variance in the dependent variable.

We used the final trained model to make a prediction for a new employee with 10 years of experience, a Master's degree, and a position of Manager, and obtained a predicted salary of $120,292. This prediction is consistent with our expectations, as an employee with higher years of experience, a higher level of education, and a higher position would typically receive a higher salary.

Conclusion:

In conclusion, our analysis demonstrates that multilinear regression is a useful tool for analyzing the relationship between multiple independent variables and a dependent variable. By applying multilinear regression to the Salary dataset, we were able to gain insights into the relationship between the independent variables (Years of Experience, Education Level, Position) and the dependent variable (Salary), and develop a predictive model that can accurately predict the salary of new employees based on their characteristics. This model can be useful for companies in making hiring and salary decisions.