Predictive Model for Salary Expectation

I started the preliminary steps of analysis by data preprocessing and exploratory data analysis on the TechSales dataset. I conducted quality checks to ensure data integrity. Specifically, identifying and removing any null values or duplicate entries. Utilizing summary statistics and visualizations revealed insights into the distribution and characteristics of key variables such as age, salary, and Net Promoter Score. Box plots were employed to visually inspect for potential outliers in these variables. To enhance the quality of the TechSales dataset, it was decided to use logical conditions to filter out outliers. These conditions focused on plausible ranges for age, salary, and NPS scores, aiming to exclude extreme values that could skew our analysis. Subsequently, a cleaned version of the dataset was generated. Summary statistics were conducted to understand the clean data. For further feature understanding, two linear regression models were created. We ran two basic linear regression models on the clean dataset to gain insights into basic features like salary, age, and NPS.

The linear regression model was constructed to predict salaries based on various features, excluding 'Years' and 'Personality'. Among the features, 'Age' emerges as a significant factor positively influencing salary, with older individuals generally commanding higher earnings. Conversely, 'Female' gender is associated with lower salaries, reflected in its negative coefficient. Additionally, expertise in 'BusinessSoftware' is linked to lower salaries, while possessing 'Certificates' and receiving positive 'Feedback' and 'NPS' ratings correlate with higher earnings. The model demonstrates a reasonably good fit, capturing around 51.1% of salary variance, supported by a significant F-statistic. The F-statistic assesses the collective significance of the model, and a notably significant p-value implies the model's effectiveness in predicting Salary. However, the residual standard error, measuring the average discrepancy between predicted and actual Salary values, stands at $15,300. This relatively high error could reflect limitations in both the quantity and quality of our data. Despite this, the model offers a comprehensive understanding of Salary determinants in the Tech Sales.

The model, Training2.lm, provides valuable insights into salary prediction, explaining around 51.1% of salary variance with significant features like age and education. However, its reliability for direct use relies on further validation, additional features, and testing to address residual variance. While promising, caution and ongoing refinement are needed before deploying it for strategic decisions.