I began by analysing the various features provided in the dataset and their relationship to the problem statement. Next, I focused on cleaning the data and handling missing values. Additionally, I addressed outliers in the numerical data. Moving forward, I conducted analysis such as Univariate and Bivariate to gain a better understanding of the relationship between the features and the target variables. To proceed, I converted categorical variables into dummy variables and initiated the development of my model.

To select the initial set of features, I employed Recursive Feature Elimination (RFE) and iteratively built the model by manually eliminating features based on their p-values and Variance Inflation Factor (VIF). After identifying the significant features, the next step involved determining a suitable cutoff point for the model to predict the conversion flag. To achieve this, I utilised the ROC curve and considered the intersection of Accuracy, Sensitivity, and Specificity.

Finally, I evaluated the performance of my model on the test data, specifically examining its accuracy, precision, and recall.

Throughout this assignment, I acquired valuable knowledge and skills related to Exploratory Data Analysis (EDA). This included techniques for handling missing data and outliers. Furthermore, I learned about converting columns to their appropriate data types. Additionally, I gained insights into conducting univariate and bivariate analysis to assess the relationship between features and the target variable.

Moreover, I discovered how to identify multicollinearity among variables using a heat map, as well as the importance of eliminating features with high multicollinearity. The process of converting categorical variables into dummy variables using the sklearn library was also explored. Through RFE, I gained an understanding of how to select relevant features and discard irrelevant ones. Furthermore, I learned to use the stats model to analyse the model summary and make informed decisions based on p-values.

Additionally, I grasped the concept of eliminating features with high multicollinearity using the VIF, where an ideal VIF score is typically less than 5. Subsequently, I successfully built a logistic linear model and assessed its accuracy. Importantly, I discovered how to derive meaningful insights from the model results.

Within this context, I developed an understanding of important metrics such as sensitivity and specificity. A high sensitivity value indicates a low rate of missed positive instances (false negatives), while a high specificity value suggests a low rate of false positives, indicating that the test is effective at avoiding misclassifying negative cases as positive.