Lead scoring Summary Report

In this assignement, we aimed to build a logistic regression model to predict the conversion probability of leads for X Education, an online education company. The model was developed based on a dataset containing various parameters, including variables such as "Total Time Spent on Website," "Last Notable Activity," and "Lead Source."

First we understood the data from the data dictionary provided with the problem statement. There were lot of binary columns, which included yes/no values, we converted those values into zero and ones. Then we handled a special case in our data where a lot of columns had “Select” values, These are the default dropdown values on the website and they were equivalent to null for us, so we handled all such columns.

Next we removed the columns with null values more approx. 50% because they could impact model performance.

Last step in data preparation, we Bucketed the categorical variables which had a lot of distinct values. Frequency for most of these values were too low to be considered so we bucketed such values as “Low frequency” Bucket. We wrote python functions to handled all these variables before proceeding to model building.

In model building process, First, we conducted a thorough analysis of the model's performance by examining the coefficients, p-values, and VIF values. The coefficients provided insights into the variables' impact on lead conversion probability, with "Total Time Spent on Website," "Last Notable Activity_SMS Sent," and "Last Notable Activity_LessFrequent" being the top contributors. To ensure the reliability of the model, we assessed the multicollinearity using VIF values. Fortunately, all variables showed VIF values below 2, indicating no significant issues of multicollinearity. Upon implementing the model, we achieved an accuracy of 76.31% in predicting lead conversion. This metric was measured using the metrics.accuracy_score function, comparing the predicted conversion outcomes with the actual conversion status. To improve lead conversion rates, we recommended focusing on potential leads, or "Hot Leads," which have a higher probability of conversion. This could be achieved by implementing a lead scoring system that assigns a lead score to each potential customer, allowing the sales team to prioritize their efforts accordingly.

In conclusion, the logistic regression model demonstrated promising results in predicting lead conversion probabilities for X Education. By leveraging the insights provided by the model, X Education can streamline their sales efforts, improve lead targeting, and enhance their overall conversion rate.