Summary Report

The Problem Statement:

An Education company sells online courses to industry professionals, and needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach and methodology:

We started off with reading and understanding the data, followed by cleaning the data which includes inspecting and treating null values by imputing or dropping them depending on the number of missing values and their importance. Next step was performing EDA which includes univariate and bivariate analysis of the categorical and continuous variables, this helped us in gaining in depth knowledge about the data and about the factors which play a major role in lead conversion. Followed by treating outliers as Logistic Regression is heavily affected by outliers.

Next step was to actually build the Logistic regression model for which we had to prepare the data by converting binary variables into numbers and creating dummy variables for all the categorical columns. The data had to be split into train and test data and we also performed scaling on it. This was followed by building the first model, we used Recursive Feature Elimination(RFE) to eliminate the columns which are not very important to our model. We then proceeded to remove more variables from the model due to high p value. We calculated the VIF to ensure that there is no multicollinearity between the variables.

Next step was to evaluate the model based on various metrics like accuracy, sensitivity, precision,recall etc, followed by finding the optimal cut-off point. Using this optimal cut-off point we arrived with good accuracy(81.31%) and sensitivity(84.05%) values.

The last step was to evaluate the model with the test data and see if it is able to predict well of the test data as well. And our model was used to predict on the test data which resulted in an accuracy of 80.83% and sensitivity of 81.49%, therefore the model is working well and is not over trained.

The learnings and conclusions from this case study are:

- The leads who are tagged as 'Closed By Horizon' have a very high conversion rate.

- The leads who are tagged as 'Will revert after reading email' have a good conversion rate and the number of customers tagged to this are very high and therefore these leads need to be focused more.
- The leads with lead source 'Welingak website' have a high conversion rate.
- The leads with Last notable activity 'SMS sent' have a high conversion rate.
- The total time spent on the Website has an impact on the conversion rate.
- Leads who are working professionals have a high conversion rate.