Summary Report

Problem Statement -

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

What we need to do -

● X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
● The company requires you 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.

Overall approach for the model

Step-1 : Data Importing and Cleaning

We go through the problem statement and the data provided. After importing the data into jupyter notebook for analysis we inspect the data for missing values, data types, rows, columns and build our approach to solve the problem statement.
We began the data cleaning process by analysing each and every column and treating missing values in them one by one. Also we have grouped and distinguished the variables having highly skewness properties in a column by visualizing boxplot and distplot. We also dropped those columns which have the most weight of only one variable.

Step-2 : Data Preparation

Now in the data preparation process we prepared the data for modelling by creating dummy variables and then we splitted the data into the train and test set for further process. After that we have performed feature scaling on the numeric values present in the data set. We looked into the correlations among the variables and drop the categorical variables accordingly which have high correlation.

Step-3 : Model Building

We are now moving forward with the Model Building process using statsmodels and scikit learn modules. During this process we used a combined approach for selecting the variables. We first selected the features using RFE(Recursive Feature Elimination) and then with manual approach by looking into the p-values and VIFs and based on that started dropping the variables which have high p-values and/or high VIFs one by one. We were also predicting the accuracy of our model after dropping every variable. We used 0.5 optimal cut off point for the probability.

Step 4: Making Predictions

Then we calculated the optimal cut off point of 0.3 by plotting accuracy, sensitivity and specificity. But using 0.3 optimal cut off point we can't recall 80% of the converted variable correctly so because to achieve that we lowered our cut off point to 0.272 and achieve the recall ability of 80%. In this process we also evaluated the other matrices and choose the one which is best suited to solve the problem in hand.