Summary Report:

Problem statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30% which is very poor rate.

Business Goal:

Our job is 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%.

Solution Methodology:

Below are the steps as follow:

Step 1: Reading and Understanding the Data.

- Reading the Lead file, and inspect the like shape, info, describe, index, etc.

Step 2: Data Cleaning.

- Check for Columns having select field categorical variables.
  - Found the few categorical variables with ‘select’ value. So, imputed them with null values.

- Clean the data if any missing values.
    - There are many columns with above 40% of missing values. So, dropped the missing high percentage missing variables which cannot impute due to high missing value percentage.
- Check for count of unique categories for all the categorical variables.
    - We observed that there are few variables with high amount if unique categories for all the categorical variables, so dropped those columns because those variables are not used for analysis.
- Check for columns having so many categories but with less percentage of rows.
    - We found such variables and combine categories and named it as 'Others'.
- Impute/remove the less percentage of missing variables.
    - Imputed/remove the variables with mean/mode based on the type of variable.

Step 3: Exploratory Data Analysis (EDA).

- Handling outliers.
    - We observed that there are outliers for few variables and we do have the flexibility of not removing the outliers, so used capping technique to treat the outliers.
- Univariate analysis.
    - Visualizing single/individual variables.
- Bivariate analysis.
    - Visualizing multiple variables.

Step 4:Dummy Variables.

- For numeric values we used the MinMaxScaler.

Step 5:Train-Test split.

- The split was done at 70% and 30% for train and test data respectively.

Step 6:Looking for Correlations.

- Check for correlations and visualize the heat map.

Step 7: Feature Scaling.

- The MinMaxScaler scaling transforms the data to have a min of zero and a max of one.

Step 8:Model Building.

- Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

Step 9:Model Evaluation.

- A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 88% each.

Step 10:Prediction.

- Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 88%.

Step 10: Accuracy,sensitivity,specificity, Precision and Recall on train set:

- Final Model Accuracy is 0.888699161751009
- Sensitivity/Recall is 0.8886703383162864
- Specificity is 0.8887179487179487
- Precision (Positive Prediction) : 0.8388414407723728

Step 11:Precision and recall trade-off:

- Check the precision and recall curve.

Step 12:Making predictions on the test seton test set:

- Final Model Accuracy is 0.888699161751009
- Sensitivity/Recall is 0.8886703383162864
- Specificity is 0.8887179487179487
- Precision (Positive Prediction) : 0.8388414407723728

Recommendations:

Since the model has resulted high accuracy results in predicting the leads who can be converted. So the marketing team can leverage this to make their operations more efficient by reducing the number customer interactions thereby improving the conversions as well.

The top three variables that contribute towards the probability of a lead getting converted are:

- Leads who tag with Lost to EINS
- Leads tag with Closed by Horizon
- Total Time Spent on Website

Phone calls should be done for the following people:

- They spend a lot of time in the website and this can be done by making the website interesting and thus bringing them back to the site
- They are seen coming back to the website repeatedly
- Their last activity is through SMS or through Olark chat conversation
- They are working professionals

Thank You