Lead Scoring Case Study
Summary Report

Problem Statement:

X Education Company sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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 Summary:

This case study is completed by building a Logistic Model which will predict if a given lead would get converted or not.

Following are the steps followed to build the model:

Step1: Reading and Understanding Data

- The historical data in the form of csv file is imported into a data frame in Python.
- The data set is analyzed by checking the number of rows and columns and also the data type of each column.

Step2: Data Cleaning

- As a first step into data cleaning, we dropped the rows with duplicate values.
- There were a few columns with value 'Select' which means no values were selected. Since this is equivalent to null, we changed these values to Null values.
- We dropped the columns having NULL values greater than 35%.
- Missing values in categorical columns were imputed with Mode (i.e., Highest occurring value)
- We also removed those rows where certain columns had negligible number of null values (less than 2%).

Step3: Data Visualization and EDA

- We visualized categorical variables using count plot and removed those columns with low variance or imbalanced data.
- We also created a new category – Others for certain categorical columns to group together less frequent categories.
- Outliers in Numerical columns we handled by capping the value at 95th percentile.
- Correlations among Numerical columns were checked using Heat Map.

Step4: Data Preparation:

- Binary columns with Yes/No values were converted to '1' and '0'
- Dummy Variables Creation:
  - We created dummy variables for the categorical variables with appropriate names.
  - Removed all the repeated and redundant variables for avoiding ambiguity & time consumption.

Step5: Model Building

- **Test - Train Split:** The next step was to divide the data set into test and train data with a proportion of Train set 70% - Test set 30% values.
- **Feature Scaling**: We used the Min Max Scaling to scale the original numerical variables of Train data.
- **Correlation check:** We plotted a heatmap to check the correlations among the variables and dropped highly correlated dummy variables.
- **Model Building:**
  - We followed a balanced approach for model building. i.e., we selected top 20 features for model building using RFE - Recursive Feature selection.
  - Using the summary statistics generated by the model, we recursively tried looking at the p-values in order to select the most significant featured that should be present and dropping the insignificant features (Dropped features with p-value > 0.05)
  - We also checked VIF – Variance Inflation Factor of the features and dropped those with value greater than 5 as this means these features have multi-colinear relationship with rest of the features.
  - Finally, we arrived at the 13 most significant variables with model 8. The P- values and VIF's for these variables were also found to be good.

Step6: Model Evaluation:

- The final model was used to predict the probability on train data which was then used to get the Lead Conversion with an initial cut-off of 0.5.
- The model was evaluated by calculating their Confusion Matrix, Accuracy, Sensitivity, Specificity, Precision, F1-Score, etc.
- ROC curve was plotted for the model and the curve came out to be pretty decent with Area Under curve of 88% which further solidified the predictive capability of the model.
- We calculated the optimal probability cut off where we get balanced Accuracy, Sensitivity and Specificity which came out to be 0.34.
- This Optimal Cut Off is used to predict the conversion again and recalculated the evaluation metrics. The Training Accuracy of our model turned out to be 80.38%.
- Precision-Recall Trade Off was plotted which gave a cut off of 0.425

Step 7: Prediction on Test Data:

- The Test data was scaled using the Min-Max scaler object created using Train data
- Final model was used to predict on the test data and cut-off of 0.34 is used to calculate Lead Conversion.
- Evaluation Metrics were calculated for the test data predictions and Accuracy came out to be 80.32%

Step 8: Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 80.32% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.

Features which contribute more towards the probability of a lead getting converted are:

- Total Visits
- Total Time Spent on Website
- Lead Source
- Occupation:
  - Working Professional
- Last Notable Activity
  - SMS Sent
  - Others

Features having negative impact on conversion probability

- Do Not Email
- Last Activity
  - Email Bounced
  - Olark Chat Conversation
  - Page Visited on Website