Lead Scoring Case Study Summary

Step 1 – Reading and understanding the data

Step 2 – Inspecting of the dataframe

Step 3 – Data Preparation: - Check the null values and its percentage,

Step 4 – Data Cleaning: -

- Dropped the irrelevant columns such as Prospect ID and Lead Number.
- ‘Select’ is the one of the values in the dataset hence replaced it with Nan and after replacement of it the null value percentage has changed.
- Now we have decided to drop the features that containing null values are more than 40% and those 7 features got dropped.
- Checking the imbalance of the dataset. Features are imbalanced, it means they are talking about only one side of the feature, so we dropped those 11 columns.
- Other columns like ‘Do Not Email’ and ‘Do Not Call’ and ‘Country’ were also decided to drop because they seem like irrelevant.
- After this we analyzed the remaining columns by value counts and also checked the relation with target variable. In this we found 3 columns which were not contributing much so we dropped them.
- Checked Outliers of numerical and categorical variables and outlier treatment was done.
- Imputed the null values with mode.
- Converted the ‘Yes’ and ‘No’ values with 0 and 1.
- Created the dummies of the variables. And concatenated the categorical and dummy variable dataframe.
- Converted the datatype of categorical variable to integer.

Step 5 – Train test Split: - Define the X and y and split the data into 70:30 ratio.

Step 6 – Feature Scaling: - Scale the features with standard scalar and check the conversion rate with converted columns, it is about 38%.

Step 7 – Model Building : -Run the first training model.

Step 8 – RFE: - 15 features were selected by RFE.

- Asses the models with stats models and build the 1st model and drop the high p-value columns one by one which are more than 0.05.

- Creating new column 'predicted' with if Converted_Prob and check the confusion matrix. overall accuracy is 80%.

- Checked VIF and dropped the high VIF columns. Checked the overall accuracy which is 80% (not dropped). and also checked confusion matrix. Check the correlation.

- Metrics beyond simply accuracy – Sensitivity 66.44%, Specificity 88.52%, False positive rate 11.42%, positive predictive 78.26% and negative predicted 80.93%.

Step 9 – Plotting ROC curve: - Area under the curve is 87%.

Step 10 – Finding the optimal Cut off point: – around 0.37 but we consider 0.3.

- Now Overall accuracy is 79.77%, sensitivity is 84%, specificity is 77%, false positive rate is 22%, Positive predicted rate 69.51%, negative predicted rate 88.67%.

- Precision and Recall – Precision is 78.62% and Recall is 66.44%. the precision and recall tradeoff are around 0.4.

Step 11 – Prediction on test set (X and y)

Calculating the Lead score and final prediction. Overall accuracy 78.5%, Sensitivity 83.82%, Specificity 75.41%, Precision score – 66.38%, Recall score 83.82%.

Determining the feature importance: - 8 features are positive in nature means that if value increases then the conversion rate also increases, negative impacted features are 5 means if value decreases the conversion rate increases.