Lead Score Case Study - Summary

To identify the potential leads of X Education and increase their revenue, the following steps were implemented and the inferences were made.

Step1: Importing the Data

Imported data and the necessary libraries.

Step 2: Inspecting the Dataframe

Initially we had 9240 rows and 37 columns in the dataset. Out of 37 variables 30-object,3-integerand 4-float datatypes.

Step 3: Data Cleaning and Preparation

The columns which had more than 20 % null values were dropped and the rows which had some null values were dropped too.

Outliers were identified, few were capped to 95 % for further analysis.As part of univariate and bivariate analysis, it was found that few columns were not crucial to the model.Those columns were dropped.

Categorical columns with value' Select ' where treated by dropping columns with huge counts and others we just removed that values from the columns. After cleaning the data we had 26 columns with 6373 rows.

Dummy variables were created for the categorical features.

Step 4: Test-Train Split

Prepared X and y variables by considering 'Converted' column as target variable and all other variables as independent variables. Data was split into train and test data with 60% and 40% .

Step 5: Feature Selection using RFE

20 variables were considered for RFE Selection.

Step 6: Model Building

A logistic regression model was built using the function GLM() under statsmodel library. Hence, some of these variables were removed first based on an automated approach, i.e. RFE and then a manual approach based on the VIFs and p-values are used for further process.We have fixed final model with 14 variables with pvalue <0.05 and VIF<3.

Step 7: Plotting the ROC Curve

When we plotted the true positive rate against the false positive rate, and got a graph which showed the trade-off.

• ROC Curve area is 0.88, which is good.

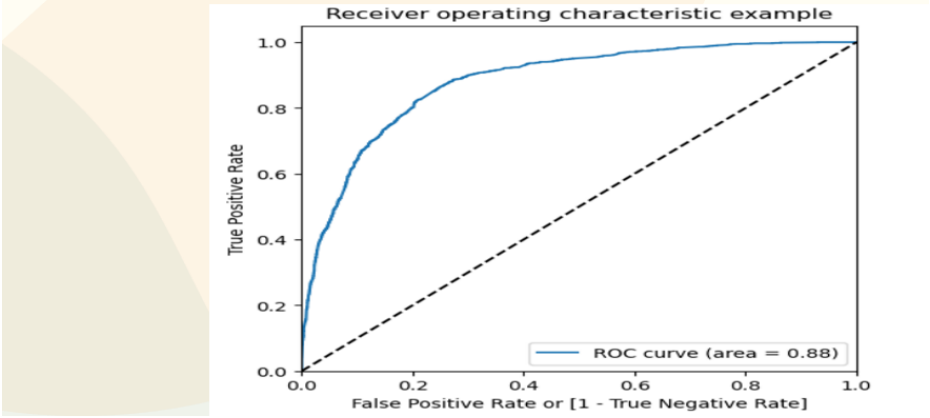

Fig1: ROC Curve

Step 8: Finding Optimal Cutoff Point

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.The optimal cut-off for the model was around 0.38.

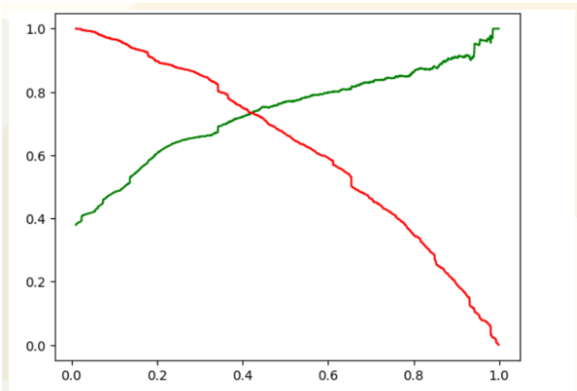

Fig 1. Precision Recall Trade off curve

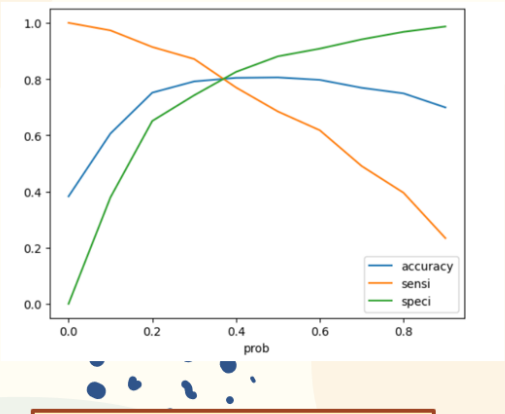

Fig 1. Cut off taking accuracy, sensitivity and specificity

Step 9: Making predictions on the test set

We also have checked precision and recall which was another pair of industry-relevant metric used to evaluate the performance of a logistic regression module. We have chosen a cut-off point of 0.38, made predictions on the test dataset and have obtained values for all three metrics such as

1.Accuracy 79.58%

2.Sensitivity 77.67%

3.Specificity 80.72%

The following variables were found to be useful for the conversion of the leads.

Do Not Email
Total Time Spent on Website
Lead Origin_Landing Page Submission
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Last Activity_Olark Chat Conversation
Last Activity_Other_Activity
Last Activity_SMS Sent
Last Activity_Unsubscribed
Specialization_Others
Specialization_Select
Last Notable Activity_Modified
Last Notable Activity_Unreachable

The company need to focus on these parameters for a better lead conversion rate.