Summary

The aim of the case study was to help X Education select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needed a model wherein higher lead score assigned would have a higher conversion chance and the customers with lower lead score would have a lower conversion chance. Finally a good conversion rate of 80% was desired.

The following are the steps we underwent:

# Data Cleaning and Preparation:

The data had 37 columns. I removed many columns which higher number of null values. Some columns had values like “Asymmetrique Profile Index, Asymmetrique Activity Index, Asymmetrique Activity Score, Asymmetrique Profile Score, Lead Profile, Tags, Lead Quality, How did you hear about X Education, City, Lead Number” which were having more than 35% of nulls. They were treated as nulls and the respective columns were removed. Then we removed rows where certain columns had small percentage of null rows.

# EDA:

# I have done Univariate analysis of each feature and the target variable and found out the inferences from the features given. Also plotted boxplot to find the outliers and then imputed them in appropriate ways.

# Dummy Variables:

The dummy variables were created for the categorical variables like Lead Origin, Lead Source, Last Activity, Specialization, What is your current occupation, Last Notable Activity.

# Data split and Scaling:

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

We used MinMax scaler to scale the numerical features.

# Model Building:

At first, we used Regressive Feature Elimination to select 15 significant variables and eliminate insignificant variables. Later some more variables were removed one by one in few iterations after checking p value (>0.05) and higher VIF (>5). Finally, our model is left with 13 features.

# Model Evaluation:

I built a confusion matrix considering True Positives, False Positives, True Negatives, False Negatives with the arbitrary cut off at 0.5. We analyzed the performance metrics of the model with ROC curve.

Later I plotted sensitivity and specificity trade off to find the optimal threshold. I saw that around the probability value of 0.40 the sensitivity and specificity were optimal on train dataset. Specificity was 0.80 and Sensitivity was 0.80 approximately.

# Precision – Recall:

We checked the precision and recall metrics and model was stable at precision of 0.77 and recall of 0.80.

# Evaluation on Test dataset

# The sensitivity was 0.81 and specificity was 0.80.

We checked the precision and recall metrics also and model was stable at precision of 0.78 and recall of 0.76.

These are the top three variables from the model which had a greater impact on the model were the below.

• TotalVisits

• Total Time Spent on Website

• Lead Origin\_Lead Add Form

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

The total time spend on the Website.

Total number of visits.

When the lead source was:

a. Google

b. Direct traffic

c. Organic search

d. Welingak website