

# Lead Scoring Case Study Summary

## **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. Once the 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%.

## **Solution Summary:**

### **Step1: Reading and Understanding Data.**

Read and analyze the data from the .csv file.

### **Step2: Data Cleaning.**

We dropped the variables that had high percentage of NULL values in them. And included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

### **Step3: Data Analysis**

We started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

### **Step4: Creating Dummy Variables**

We created dummy data for the categorical variables.

### **Step5: Test Train Split:**

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

### **Step6: Feature Rescaling**

We used the Min-Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

### **Step7: Feature selection using RFE:**

Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values to select the most significant values that should be present and dropped the insignificant values.

Finally, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good.

We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

#### **Step8: Plotting the ROC Curve**

We then tried plotting the ROC curve for the features and the curve came out decent with an area coverage of 89% which further solidified the model.

#### **Step9: Finding the Optimal Cutoff Point**

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.37

Based on the new value we could observe that close to 80% values were rightly predicted by the model.

We could also observe the new values of the 'accuracy=81%', 'sensitivity=79.8%', 'specificity=81.9%'.

Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

#### **Step10: Computing the Precision and Recall metrics.**

We found out the Precision and Recall metrics values came out to be 79% and 70.5% respectively on the train data set.

Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

#### **Step11: Making Predictions on Test Set**

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out that-> Accuracy= 80.8%; Sensitivity=78.5%; Specificity= 82.2%.