Summary

Lead Score Case Study

Submitted by :
Amit Kulkarni
Asuthosh Kumar
Swarangi Arvikar

Problem Statement:

X Education provides online courses tailored for industry professionals.

Although X Education generates a significant number of leads, its conversion rate is low; for instance, out of 100 daily leads, only about 30 convert.

To enhance efficiency, the company aims to pinpoint the leads with the highest potential, referred to as 'Hot Leads'.

By accurately identifying these 'Hot Leads,' the conversion rate is expected to improve, allowing the sales team to focus their efforts on engaging with these promising leads instead of contacting every prospect.

Solution Summary:

Step1:

Reading and Understanding Data. Read and analyse the data.

Step2:

Data Cleaning: We dropped the variables that had high percentage of NULL values in them. The outliers were identified and removed.

Step3:

Data Analysis Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.

Step4:

Creating Dummy Variables we went on with creating 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 top important features. Using the statistics generated, we recursively tried looking at the P-values in order 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 be pretty decent with an area coverage of 86% which further solidified the of 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 0.42

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=79.2%, 'sensitivity=79.8%', 'specificity=78.6%'.

Step10:

Computing the Precision and Recall metrics we also found out the Precision and Recall metrics values came out to be 78.3%and 78.2% 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 the accuracy value to be 78.5%; Sensitivity=77.4%; Specificity= 79.6%.