SUMMARY ON LEAD SCORE CASE STUDY

Problem Statement

The case study focuses on a real-life scenario faced by an education company named X Education, which sells online courses to industry professionals.

The company markets its courses on several websites and search engines like Google. Once potential customers land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these potential customers fill up a form providing their email address or phone number, they are classified as 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%.

Our task was to help X Education identify the most potential leads, also known as ‘Hot Leads’. The company wanted to build a model to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, had given a ballpark of the target lead conversion rate to be around 80%.

We were provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Solution Summary

1. Reading and Understanding Data

Read and analyze the Data.

2. Data Cleaning

We dropped the variables that had high percentage of Null values in them. This step also included imputing the missing values as and when 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

3. Data Analysis

Then 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.

4. Creating Dummy Variables

We went on with creating dummy data for the categorical variables.

5. Test Train Split

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

6. Feature Rescaling

We used Standard Scaler 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

7. Feature Selection using RFE

Using the Recursive Feature Elimination we went ahead and selected the 15 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 11 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 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 **‘Specificity’** matrices to understand how reliable the model is.

8. 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.

9. 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 was found out to be 0.3

Based on the new value we could observe that close to 83% values were rightly predicted by the model.

We could also observe the new values of the ‘Accuracy=77.05%,’Sensitivity=82.8%’,’Specificity=73.49%’.

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

10. Computing the Precision and Recall metrics

We also found out the Precision and Recall metrics values came out to be 65.6% and 82.8% respectively on the train data set.

11. 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 77.52%,

Sensitivity=83.01%; Specificity=74.13%.