Lead Score Case Study

Problem Statement:

X Education sells online courses to industry professionals. The business advertises its classes via a number of websites and search engines, including Google. Upon accessing the website, these individuals may pursue the available courses, complete the course registration form, or view some videos. When someone fills out a form with their phone number or email address, they are categorized as leads. Additionally, the business receives leads from previous recommendations. After obtaining these leads, sales team members begin calling, emailing, and so on. While most leads do not convert during this process, others do. At X education, the lead conversion rate is typically 30%.

Summary:

Step 1- Reading the Data:

Reading and analysing the given dataset

Step 2- Data Cleaning:

We cleaned the data by dropping the variables which had more than 30% of null values. In this step we also imputed some missing values with creation of new variables in case of categorical variables or by median values in case of numerical variables. The outliers were also identified and removed.

Step 3- Exploratory Data Analysis:

In this step we did exploratory data analysis where we got a feel of how the data is oriented. Also dropped some variables which were insignificant in the analysis.

Step 4- Creating dummy variables:

Created dummy data for categorical variables.

Step 5- Train-Test data split:

Then we divided the dataset into train and test data with 70%-30% proportion respectively.

Step 6- 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 gave us a complete statistical view of all the parameters of our model.

Step 7-Feature Selection using RFE:

Using the Recursive Feature Elimination we selected top 20 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. After several models, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good. Then we 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 the reliability of the model.

Step 8- Plotting the ROC Curve:

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

Step 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 point was 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=80%', 'specificity=82%'. Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step 10- Computing the Precision and Recall metrics:

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

Step 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 81%; Sensitivity=80%; Specificity= 82%.