Lead Scoring - Case Study

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need 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, has given a ballpark of the target lead conversion rate to be around 80%.

Goals of Case Study

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Solution Summary

Step 1: Reading and Understanding Data

In the initial phase of our analysis, we performed essential data inspection steps, including reviewing the dataset using the head () function, determining its shape with the shape function, checking for data types and missing values using the info () function, and conducting a duplicate check.

Step 2: Data Cleaning

a) In the data cleaning process, we dropped variables with unique values as they did not provide any meaningful information for our analysis.

b) We converted the 'Select' values in certain columns, indicating that leads did not choose any option, to null values as part of the data cleaning process. This adjustment allowed us to handle missing data appropriately in those columns.

c) We dropped the columns having NULL values greater than 35%.

d) Afterwards, we addressed imbalanced and redundant variables by removing them from the dataset. Additionally, we imputed missing values with median values for numerical variables and created new classification variables for categorical variables as needed. We also identified and removed outliers. Furthermore, we resolved an issue in one column where the label had

inconsistent cases by converting the label with a lowercase first letter to uppercase for consistency. These steps ensured data integrity and improved the quality of the dataset for further analysis.

e) Sales team generated variables were eliminated to ensure a clear and unambiguous final solution.

Step 3: Data Transformation

Changed the binary variables into '0' and '1'

Step 4: Dummy Variable Creation

We created dummy variables for categorical variables and removed any repeated or redundant variables.

Step 5: Test Train Split

Subsequently, we divided the dataset into training and testing subsets, with a split ratio of 70% for training data and 30% for testing data.

Step 6: Feature Re-scaling

a. To standardize the numerical variables, we applied Min-Max Scaling.

b. We visualized the variable correlations using a heatmap.

c. Highly correlated dummy variables were dropped to avoid multicollinearity issues in the analysis.

Step 7: Model Building

a. Utilizing Recursive Feature Elimination, we selected the top 15 important features.

b. By analyzing the statistics and P-values, we iteratively identified and dropped insignificant variables.

c. We arrived at 11 significant variables, and their Variance Inflation Factors (VIFs) were found to be satisfactory.

d. To determine the optimal probability cutoff for our final model, we assessed accuracy, sensitivity, and specificity at various points.

e. We plotted the ROC curve, which exhibited a respectable area coverage of 86%, further validating the model's performance.

f. We verified if 80% of the cases were correctly predicted based on the converted column.

g. Precision, recall, accuracy, sensitivity, and specificity were assessed for our final model on the training set.

h. Considering the trade-off between precision and recall, we established a cutoff value of approximately 0.3.

i. We applied the findings to the test model, calculated conversion probability using sensitivity and specificity metrics, and obtained an accuracy of 77.52%, sensitivity of 83.01%, and specificity of 74.13%.

Step 8: Inference / Conclusion

Accuracy: 77.52%

Sensitivity :83.01%

Specificity: 74.13%

While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.

Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%

Hence overall this model seems to be good.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- Lead Origin Lead Add Form
- What is your current occupation_Working Professional
- Total Time Spent on Website