Lead Score Case Study

Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. Importing Libraries:
   Imported necessary library to perform the analysis
2. Reading and Understanding the Data
   Lading the data and for better understanding performing some sanity checks.
   a. Data Inspection and Cleaning
   In this step, we found that some of the data provided had null values, and missing values. In the presence of null values in the data set we can't do any analysis if we do so we may predict wrong results or Analysis. So, we have done the missing value treatment in this step. And cleaned the data set to make it missing value-free.
   b. Outlier treatment
   In this step we have taken care of all outlier in our features. As we can see them in the Jupyter Notebook File.
3. Exploratory Data Analysis
   After cleaning the data set we perform the EDA which gives us some inference about the data set what is going inside the data and we have handled the data imbalance. In the data set. The inferences are also mentioned in the Presentation file and also in the Jupyter Notebook File.
4. Data Preparation
   a. Dummy Variable Creation
   We have converted categorical columns to the numeric column by creating the dummy variable. After Dummy variable creation we check the corelation to do further analysis.
5. Test-Train Split
   Here we have splited our data set into to parts to build our model and also to evaluate it. The split was done at 70% and 30% for train and test data respectively
6. Rescaling the features with MinMax Scaling
   We have used the min-max scaler to scale our data as you can see that there have most of the columns in 0 and 1. Some of them have numeric values of more than 1 so we scale them to maintain standards.

Submitted by:
Mr. Muna Sahu & Mr. Pushkar Rajwadikar
DS C43 | Lead Scoring Case Study | 2022

7. Model Building using Stats Model & RFE

We have used the hybrid feature selection using RFE and manual selection by keeping eye on the P value and VIF. Which is mostly preferred by the analytic industry.

8. Plotting the ROC Curve

We have plotted the ROC curve because ROC curve give several information like:

- ❖ It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- ❖ The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- ❖ The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

9. Finding Optimal Cutoff Point

We have plotted the accuracy, sensitivity, and specificity curve to find the optimal Cutoff Point we have the cutoff point here is 0.31.

10. Making predictions on the test set

Here we have made the prediction on the test data.

11. Precision – Recall

This method was also used to recheck and a cut off of 0.31 was found with Precision around 73% and recall around 75% on the test data frame.

12. Inferences from the model

- The Sensitivity, Accuracy and Specificity of the model is 82%, 77 and 73% respectively.
- For this model we have considered higher sensitivity value to achieve better lead converion rates. Due to higher sensitivity chances of missing the Hot leads is lower.
- The model achieves the target of 80% by predicting 82% of Hot leads. Hence this is a good model for Education X company to improve their conversion rate.

As per our regression model, below are the features that influence the conversion rate

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Occupation_Working Professional
- Occupation_Other
- Occupation_Unemployed
- Lead Source_Olark Chat
- Occupation_Student
- TotalVisits
- Do Not Email

All the features except 'Do Not Email' improve the chances of lead conversion whereas 'Do Not Email' reduces it. It can be inferred from the negative coefficient value for the variable.

Submitted by:
Mr. Muna Sahu & Mr. Pushkar Rajwadikar