Summary

Lead Score Case Study

-by Shikhar Verma & Uttkarsh Mishra

Problem Statement: An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

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

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e., educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you 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%.

Goal: 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 this step, we import all the useful libraries, then we import our dataset and then we use some other features like info (), shape, describe () to get better understanding of the dataframe.

Step 2: Inspecting the DataFrame and EDA:

In this step, at first, we will check the null values in each column, then dropping those columns which have more than 30% null values. Then there were many values as ‘Select’, we have converted those values to null. Then we have observed each column and imputed their null values either by mean or mode. We have also used count plot to visualize data distribution in each column.

Step 3: Data Preparation (Dummy Variable Creation):

In this step we will convert binary categorical variable with Yes/No value to 0/1. Then we will create dummy variables for other categorical columns and drop repeated columns.

Step 4: Train- Test Split:

In this step we will divide our data into train data and test data. We kept 70% data in train data and 30% data in test data.

Step 5: Feature Scaling:

In this step we will do feature scaling for numerical variables. For this we have used Standard Scaler. Moreover, we will also check conversion rate.

Step 6: Looking at correlations:

Here we will draw correlations between variables. Then we will draw heat map and will drop some columns.

Step 7: Model Building:

Here we will start with our model building process. First, we will build model with all the columns using statsmodel library.

Step 8: Feature Selection using RFE:

In this step we will use RFE to select top 15 variables then we will build model using those variables. We will again check p value and VIF to remove other variables which are of no use in model building. We will remove above step till we get p value less than 5% and VIF less than 5.

Finally, we got 13 variables which we will use to build our final model.

We will now create the dataframe having the converted probability values, initially we will assume that values with probability less than 0.5 is considered as 0 and values with probability less than 0.5 is considered as 1.

Based on above assumptions, we will build the Confusion Metrics and will calculate the overall ‘Accuracy’ of the model.

We will now calculate ‘Sensitivity’ and ‘Specificity’ and other metrics to understand how reliable the model is.

Step 9: Plotting the ROC curve:

We will now build the ROC curve. We found out that the area that comes under ROC curve is 0.88.

Step 10: Finding Optimal Cutoff Point:

In this step, we will plot the probability graph for ‘Accuracy’, ‘Sensitivity’ and ‘Specificity’ for different probability values. The intersecting point of the graph will be considered as the optimal probability cutoff. Here we got cutoff value as 0.35.

Now based on this cutoff value, we will make final predicted column where probability below cutoff value is considered as 0 and value above cutoff value is considered as 1.

We will now calculate the new ‘Accuracy’, ‘Sensitivity’ and ‘Specificity’.

Step 11: Precision and Recall:

In this step, we will calculate Precision and Recall metrices for the dataset. Based on these Precision and Recall value we will make Precision and Recall tradeoff, we got a cutoff value of 0.41

Step 12: Making predictions on Test Set:

We will now make predictions on test model based on training data. We will again create final predicted column using cutoff value of 0.35.

Using this final predicted value, we will again make confusion matrix and will calculate ‘Accuracy’= 81.64%, ‘Sensitivity’= 80.55%, ‘Specificity’= 82.35%, ‘Precision’= 74.87% and ‘Recall’= 80.55% of test data.

Step 13: Calculating Lead score for the entire dataset:

In this step we will make a new column ‘Lead Score’ after merging train and test dataset and then multiplying Converted_Probability by 100.