Leading Score Case Study Summary

By:
Sayan Dutta

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 (top) but only a few of them come out as paying customers from the bottom. 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%.

Our Goals of the Case Study:

1. To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads that can be used by the company to target potential leads.
2. To adjust to if the company's requirement changes in the future so you will need to handle these as well.

Solution Summary:

Step1: Reading and analyzing the Data.

- To proceed with the understanding of data, we first imported all the required packages and the dataset.
- Using ".info", we figured out the dimensions and types of data we had. There were 9240 rows and 37 columns in the dataset including null values.

  To gain statistical insights from the dataset we used ".describe" which works on numerical variables.

Step2: Data Cleaning:

- In this step, we figured out that the following columns had null values : Country, Lead Source, Total Visits, Page Views Per Visit, Last Activity, What is your current occupation, What matters most to you in choosing a course
- To start with data cleaning we dropped the variables with high percentage of null values, more than 30% as they are not useful and may mislead with model building and predictions. Whereas, variables with null values less than 30% can be treated separately.
- After the treatment of columns, we imputed the missing values as and where 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.
- -For Example: columns having only one value "No" in all the rows were eliminated: Magazine, Receive More Updates about Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque.
- After the complete process around 98% of data was retained.

Step3: Data Analysis

- We started our Data analysis by understanding the column effects on the conversion rates and found out that, there has been an overall conversion rate of around 39%.
- Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.
- From the Exploratory data analysis we had useful insights such as, the maximum conversion happened from Landing Page Submission and the major conversion in the lead source is from google, etc.
- While performing this step, there were variables that were identified to have only one value in all rows. These variables were dropped.

Step4: Data Preparation - Creating Dummy Variables

- In this step, we created dummy data for the categorical variables such as Lead Origin, Lead Origin, Last Activity, What is your current occupation, Last Notable Activity and such others.

Step5: Test-Train Split:

- In this step we divided the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling with MinMax Scaling

We used the Min Max Scaling 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.

Step7: Model Building

- After performing all the above steps, we concluded that Logistic Regression would be the best fit model.

Step8: Feature selection using RFE

- In this step we used the Recursive Feature Elimination and selected the 20 top important features.
- Using the generated results, we looked at the P-values in order to select the most significant values that should be present.
- All the insignificant values were dropped. We removed columns such as, LastActivity_Approached_Upfront,CurrentOccupation_Housewife, LastActivity_Had, LeadSource_Reference, LeadOrigin_API
- 15 most significant variables were found with good VIF's.
- We then 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 how reliable the model is.

Step9: Plotting the ROC Curve

- An ROC curve demonstrates 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.
- Hence in this step after plotting the ROC curve for the features we found out that the curve came out be pretty decent with an area coverage of 89% which further solidified the of the model.

Step10: Finding the Optimal Cutoff Point

- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- Hence in this step we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values.

-

The intersecting point of the graphs was found out to be 0.37. Here, it will be considered as the optimal probability cutoff point.

- The final prediction of conversions had a target of 80% (79.8%) conversion as per the X Educations CEO's requirement. Hence this is a good model.
- Also, we found out new values of the ‘accuracy=81%, ‘sensitivity=79.8%’, ’specificity=81.9%’.

Step11: Computing the Precision and Recall metrics

- We also found out the Precision and Recall metrics values came out to be 79% and 70.5% respectively on the train data set.
- Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

Step12: Making Predictions on Test Set

- Then we implemented the learnings to the test model and calculated the conversion probability.
- While we have checked 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 81%, 79% and 82% 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.