Lead Score Case Study

Presented By

Shyam Sundar S A

Problem Statement

- The X Education sells online courses to industry professionals. The Business of the company depends on the leads generated by the marketing team and the conversion rate.
- Currently the lead conversion rate is very poor. The conversion rate is only 30%. i.e., Out of 100 leads there are only 30 conversion which is very less for the business.
- The business facing the problem of identifying the potential leads or ‘Hot Leads’. The business may experience losses if they fail to identify the Hot Leads also called promising leads which can pay to the services.

Research Design

- Lead Score Data Set is used for the analysis which contains 9240 entries with 37 columns.
- Supervised Learning technique is used to understand the data and make prediction.
- The ‘Converted’ feature is set as target variable.
- Logistic Regression is employed for the analysis because targeted variable is a categorical which has binary values.
- Intensive Exploratory Data (EDA) Analysis has been done to understand the pattern of the data set also Recursive feature elimination (RFE) and Variance Inflation Factor (VIF) is used to identify the optimal features to build model and make prediction.
- Model Evaluation is done with the help of statistical Techniques such as Accuracy, Recall, Precision, Sensitivity and Specificity.

Methodology

- Data Inspection
- Data Cleaning
- Data Preparation
- Model Building
- Model Prediction
- Model Evaluation

Data Inspection and Cleaning

- The data set was containing 9240 entries with 37 columns,
- The Features which has Null rate less than 40% is considered for the analysis.
- The final data set has 7643 entries with 15 columns after dropping the unnecessary columns.
- The columns which were dropped was highly skewed towards ‘NO’ and found irrelevant for the studies.
- The sub-categories which has negligible entries has been merged and put under the Other Category.
- Few features was having ‘Select’ as the one of the sub-category. It is treated as Null Values and dropped such features based on the Null Value rate.

Data Preparation

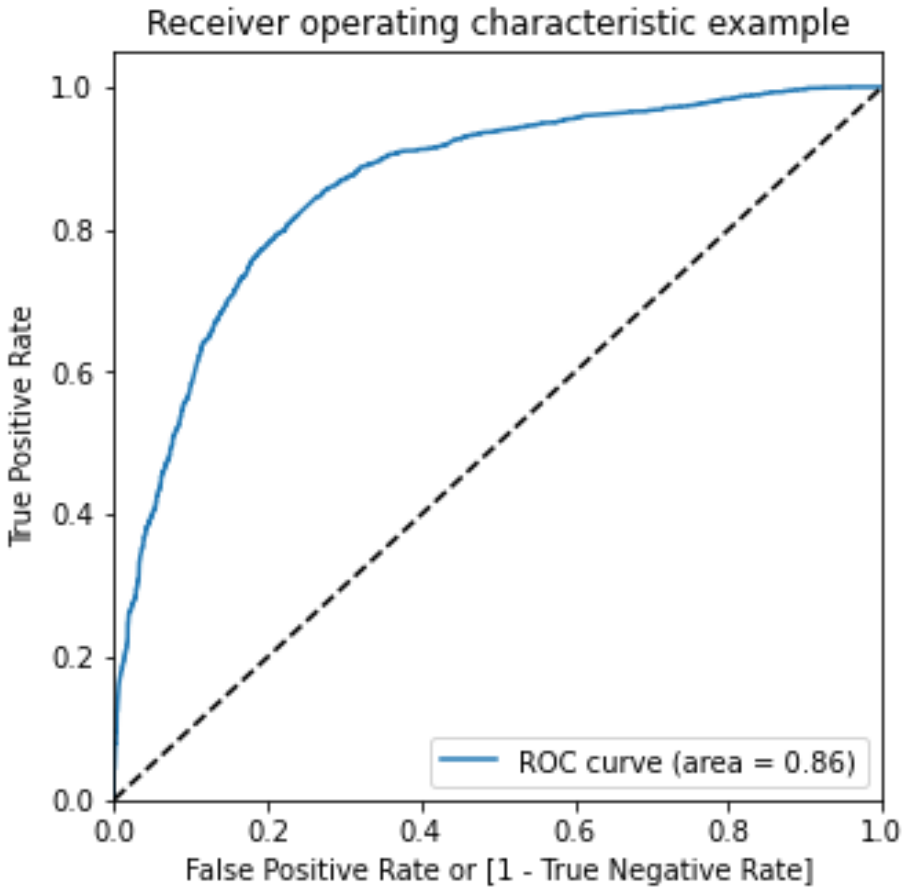
- Converting binary variables such as ‘Yes’ and ‘No’ to ‘0’ and ‘1’.
  - Example: Do Not Email feature has ‘Yes/No’
- Creating Dummy Variable for the categorical features.
  - The dummy columns created for 'Lead Origin’, 'Lead Source’, 'Last Activity’, 'Specialization’, 'City’, and 'Last Notable Activity’.
- Splitting the Data into Train and Test. Train size is of 70%.
- Standardised the Data set of Train and Test. MinMaxScaler is used for scaling the data set.

Model Building

- Recursive feature elimination (RFE) was used to eliminate the unnecessary features and kept only 20 features to build the model.
- GLM() function is used to build the regression model.
- Five iteration was done to get the Optimal Features with P-value less than 0.05.
- Variance Inflation factor (VIF) was checked for all the features and found that VIF is less than 5.
- 15 features were selected at the end for the prediction.

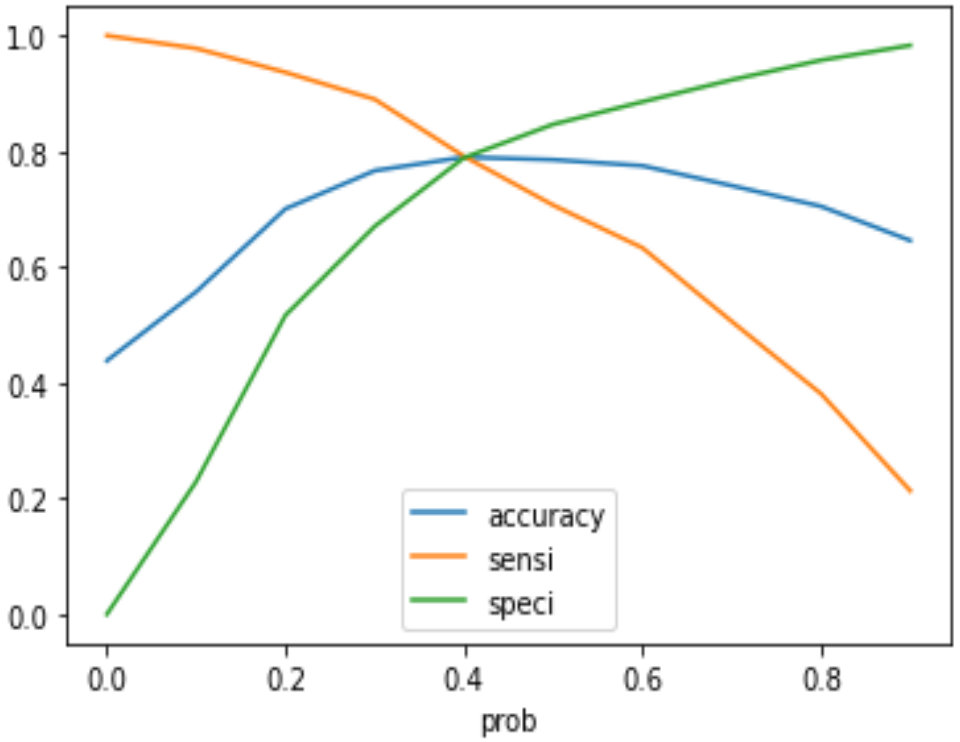
Model Prediction

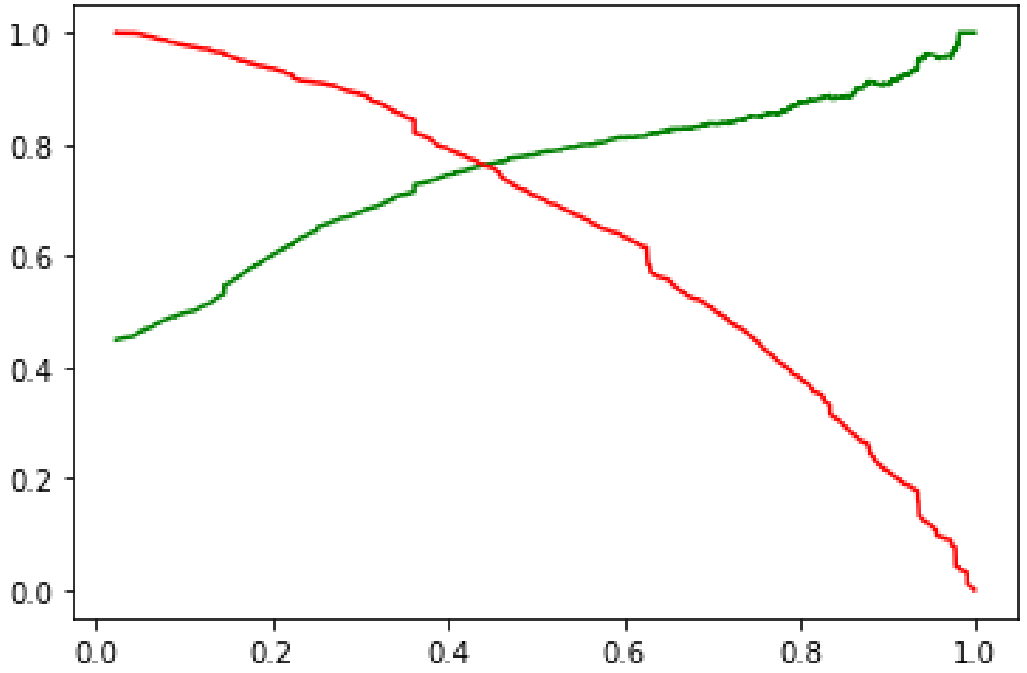
- Keeping some assumed probability of 0.5, the prediction has done to obtain the predicted values.
- ROC cure was developed and found the area under the curve is 0.86.
- It shows True Positive Rate is higher than the False Positive Rate.
- Confusion matrix was developed to understand the Accuracy, Recall, Precision, Sensitivity and Specificity for 0.5 probability



Optimal Cut-Off Point

- The cut-off point was found to be 0.4.
- It is clear that we can keep 0.4 as the probability rate for final prediction values.
- Even threshold survey conducted and found that the probability is 0.4.





Model Evaluation

- Model Prediction is done even for ‘Test’ data set.

Result Obtain on Train and Test Data

Train Data

- Accuracy = 78.99
- Recall(True Positive Rate) = 79.14
- Precision(Positive Predictive Value) = 74.50
- Sensitivity(True Positive Rate) = 79.14
- Specificity(True Negative Rate) = 78.87

Test Data

- Accuracy = 77.76
- Recall(True Positive Rate) = 78.19
- Precision(Positive Predictive Value) = 70.67
- Sensitivity(True Positive Rate) = 78.19
- Specificity(True Negative Rate) = 77.45

‘Hot-Leads’

- Based on the results of the Model Evaluation, probability of more than 80% of the lead can be converted very easily provided few strategies need to adopted.
- Lead Score was calculated by keeping standard probability rate of 85%.
- 336 Hot leads was found. They are the promising customer.
- 15 important features that impacted the conversion rate.

Important Features

- Total Time Spent on Website - **4.392359**
- Lead Origin_Lead Add Form - **3.379660**
- Last Activity_Had a Phone Conversation - **2.505693**
- Last Notable Activity_Unreachable - **2.270459**
- Last Activity_Other Activity - **2.111941**
- Lead Source_Welingak Website - **1.975983**
- Last Activity_SMS Sent - **1.605957**
- Lead Source_Olark Chat - **1.247731**
- TotalVisits - **1.170842**
- Last Activity_Email Opened - **0.522848**
- Last Notable Activity_Modified - **0.679392**
- Page Views Per Visit - **1.026165**
- Lead Origin_Landing Page Submission - **1.097524**
- Specialization_Select - **1.100788**
- Const - **1.239000**
- Do Not Email - **1.324706**

Recommendations

- Since Accuracy, Recall and Precision has Positive rate around 80%, the company can convert those customer who has lead score more than 80.
- The company should make a phone call for those customer who spent time on website, lead origin who has lead add form, customer had a phone conversation in the last activity, Last notable activity even though unreachable, Others Activity under last activity, Customer who sourced Welingak Website under lead source, Customer those who received the SMS and Olark chat.
- The company should not make phone calls to those customer comes under Last Notable Activity – Modified, Page Views Per Visit, Lead Origin – Landing Page Submission, Specilalization_select and Do Not Email.

Conclusion

- The X Education can better improvise the visibility by developing the website content and adding new features init. As the model recommend, most of the conversion had happened is because of the company website. The company need to monitor closely the time spent by each of their customer on their website also phone conversation made and follow up done is the integral part of the conversion rate.
- The company can improvise its conversion rate by segregating the lead score and identifying the ‘Hot Leads’ based on the probability rate.

Thank You