LEAD SCORING CASE STUDY

BY

Preeti Kumari, Chetna Panchal, Nelisa Sebastian

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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the case study

- There are quite a few goals for this 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Approach

- Source the data for analysis
- Reading and Understanding the data
- Data cleaning
- Exploratory Data Analysis
- Feature scaling
- Splitting the data into train and test dataset
- Prepare the data for modelling
- Model building
- Model evaluation- specificity & sensitivity or precision recall
- Making prediction on test set

Data cleaning and analysis

- Read the data from CSV file
- Data cleaning and null values treatment
- Removing unwanted columns from data
- Drop null values columns from the data
- Feature scaling

Data Preparation

- Converted binary variables into 0 and 1
- Created dummy variables for categorical variables.

Feature Scaling and Splitting Train & Test Sets

- Feature scaling of numerical data
- Splitting data into Train & Test Set
- Checking correlation

Model Building

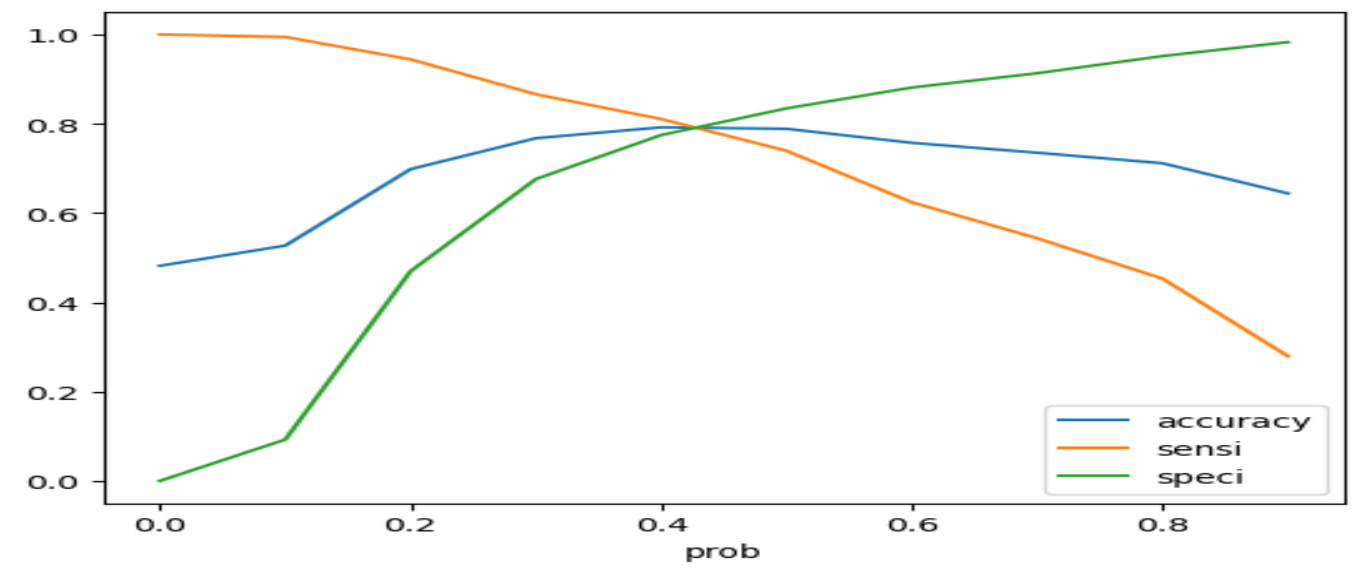
- Feature selection using RFE
- Determined optimal Model using Logistic Regression
- Checking VIFs
- Calculated accuracy, sensitivity, specificity, precision & recall and evaluate model

Variables impacting the conversion rate

- Total visits
- Total time spent on website
- Lead Source_Olark chat
- The number of visits
- Lead source with google
- Lead source_Referral Sites

Model Evaluation-Sensitivity & Specificity on Train dataset

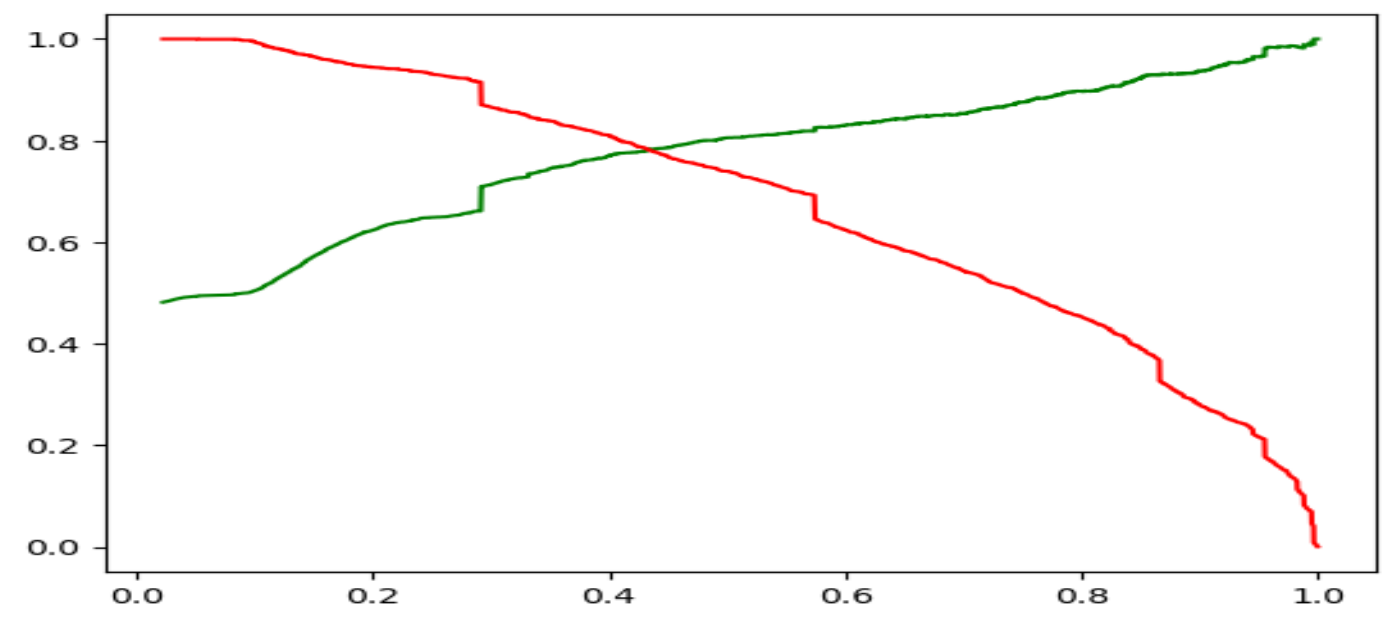
Graph depicts an optimal cut off of 0.42 bases on Accuracy, Sensitivity and Specificity



- Accuracy- 79%
- Sensitivity- 79.3%
- Specificity- 78.8%

Precision and Recall on Train data

Graph depicts an optimal cut off of 0.42 based on Precision and Recall



- Precision- 78.4%
- Recall- 77.7%

Model Evaluation- Sensitivity & Specificity on Test dataset

- Accuracy- 75%
- Sensitivity- 83.5%
- Specificity- 69%

Result

- Accuracy, Sensitivity and Specificity values of training and test set are close to training set
- Accuracy, Sensitivity and Specificity values of training set are 79.0%, 79.3%, 78.4%
- Accuracy, Sensitivity and Specificity values of test set are 78.9%, 78.4%,77.7%
- Conversion rate for Train and Test dataset is 82.3% and 71.8%
- We have done the prediction on the test set using cut off threshold from Sensitivity and Specificity metrics.

Conclusion

- While we have checked both sensitivity-specificity as well as Precision and Recall metrics, we have considered the optimal cut off based on sensitivity-specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 78.9%, 78.4%,77.7%
- which are approximately closer to values calculated using Trained data set
- Lead score calculated for the conversion rate final model on Train and Test dataset is 82.3% and 71.8% respectively
- Hence, overall model seems to be good.

Summary

There are a lot of leads generated in the initial stage but only a few come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads as well in order to get a higher lead conversion. First, sort out the best prospects from the leads you have generated. 'Total Visits', Total time spent on website, Page views per visit which contribute most towards the probability of a lead getting converted

Thank You