Lead Scoring Case Study

MANIKYAMBA AKA

SUDERSHAN VASUDEVAN

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%.

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

Analyses Approach

1. Reading and understanding the data
2. Data Cleaning
3. EDA
4. Data Preparation
    i. Creating dummies
    ii. Test-Train split
    iii. Scaling
5. Data Modelling
    i. RFE technique for variable selection
    ii. Model Building
    iii. Model Evaluation – Accuracy, Specificity, Sensitivity
    iv. Predicting on test data

Data Understanding and Cleaning

- The dataset Leads.csv has around 9240 entries with 37 attributes.
- The given data has entries as “Select” which are blank values and are not selected by the user while filling the data. These values are treated as null and are handled
- Other missing values are imputed accordingly.
- Dropping columns/rows with high missing values.
- Dropping columns with high data imbalance for a single category
- Minority categories are clubbed into one category
- Outlier analysis
- EDA

Data Preparation

- Binary variables of Yes or No are mapped to 1 or 0
- Dummies are created for categorical variables.
- Data is split into Train and Test in the ratio 70:30
- Scaling is performed on this data

Model Building

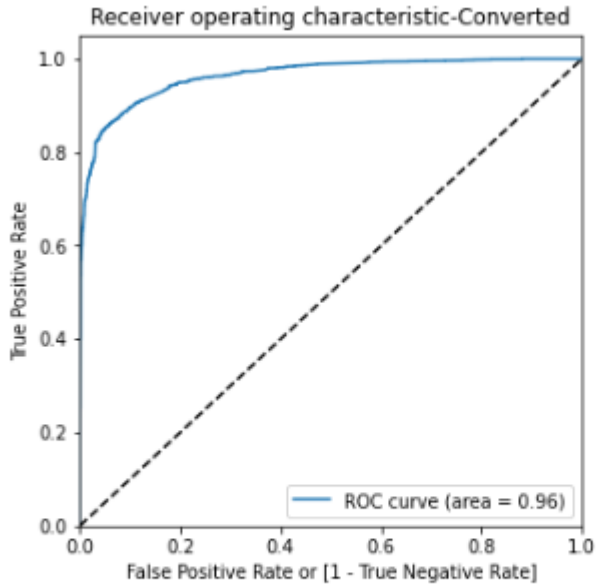
- Features are selected using RFE
- Model is formed using Logistic Regression
- By checking p value and vif values, features are dropped and optmal model is obtained
- Parameters like Accuracy, Specificity, Sensitivity are calculated
- Predictions are applied on Test data and evaluated

Variables impacting Conversion rate

- Tags_Closed by Horizzon
- Tags_Will revert after reading the email
- Lead Source_Welingak Website
- Lead Origin_Lead Add Form
- Tags_Others
- Lead Source_Olark Chat
- Total Time Spent on Website

Model evaluation –ROC Curve

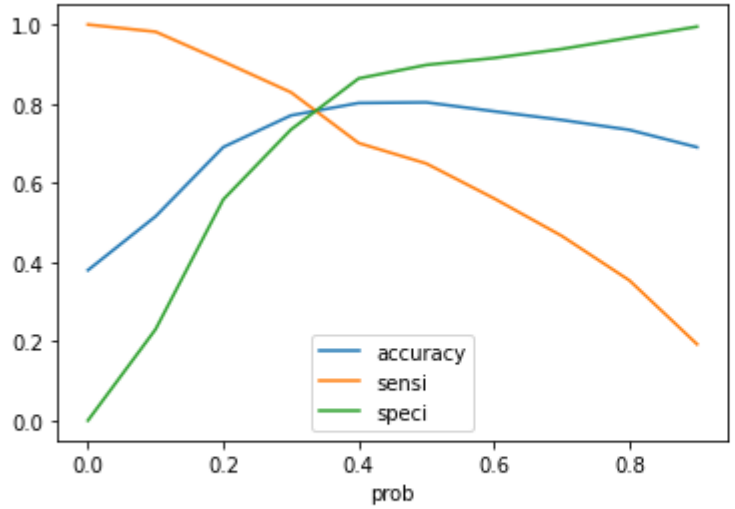
- ROC Curve



The ROC area from the above plot is 0.96 indcating good predictve model

Model Evaluation-Accuracy

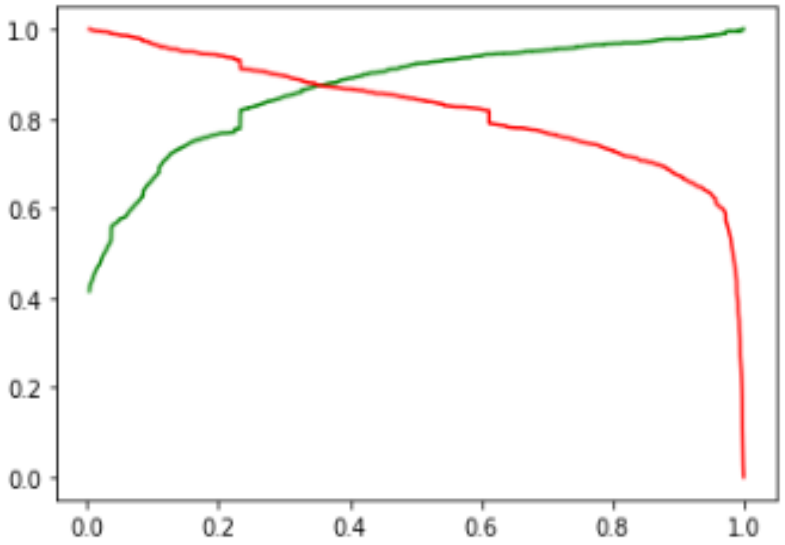
- Accuracy : 89.8
- Sensitivity :89.5
- Specificity : 90.06



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Model Evaluation-Precision and Recall

- Precision : 92.2
- Recall : 84.3



Model Evaluation on Test data- Accuracy

- Accuracy : 88.5
- Sensitivity :86.9
- Specificity : 89.4

Inferences

- Accuracy, Sensitivity and Specificity are in similar range for Train and Test data
- Precision and Recall are 92.2 and 84.3 which implies good model
- ROC Curve area is 0.96 which again implies a good model.
- Resultant Conversion rate for data is ~84.9 meeting the ballpark of lead conversion rate .