Lead Scoring Case Study

BY: HRITIK YADAV
MANISHA RAJPUT
POOJA AHER

Problem Statement

An online education company named X Education wants to know its potential lead (unpaid customers) which can convert into paid customers. The company wants a model which assign a lead score to each lead based on chances of its converting to paid customer. Higher the lead score, higher the chance of conversion.

Approach for the analysis

- Reading the data
- Data cleaning
- Data transformation
- Train-Test split and feature scaling
- Model building and feature selection using RFE
- Calculating the model evaluation metrics
- Plotting ROC and finding optimal cutoff point
- Calculating the precision and recall metrics
- Prediction on test set

Correlation

From the correlation heat map, we can see that ‘page views per visit’ has high correlation with ‘total visits’.

Variables Impacting the conversion rate

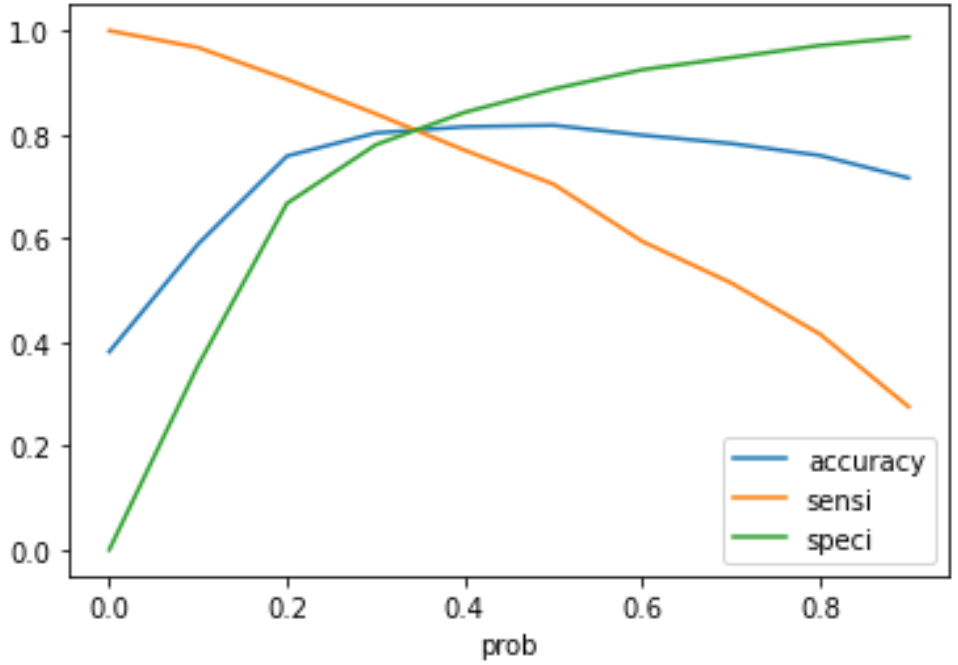
- Do Not Email
- Total Visits
- Total Time Spent on Website
- Page Views Per Visit
- Lead Origin-Lead Add Form
- Lead Source-Olark Chat
- Lead Source-Welingak Website
- Last Activity-Converted to Lead
- Last Activity-Email Bounced
- Last Activity-Olark Chat Conversation
- What is your current occupation-Working Professional
- Last Notable Activity-Email Link Clicked
- Last Notable Activity-Email Opened
- Last Notable Activity-Modified
- Last Notable Activity-Olark Chat Conversation
- Last Notable Activity-Page Visited on Website

Model Building

We first build the model with all dummy variables but after that we used RFE to select 20 variables and calculated p value. We then eliminate variables one by one which has p value more that 0.05. After 6 iterations we got our final model where every variable has p value less than 0.05 and VIF value less than 3.

Model Evaluation - Sensitivity and Specificity on Train Data Set

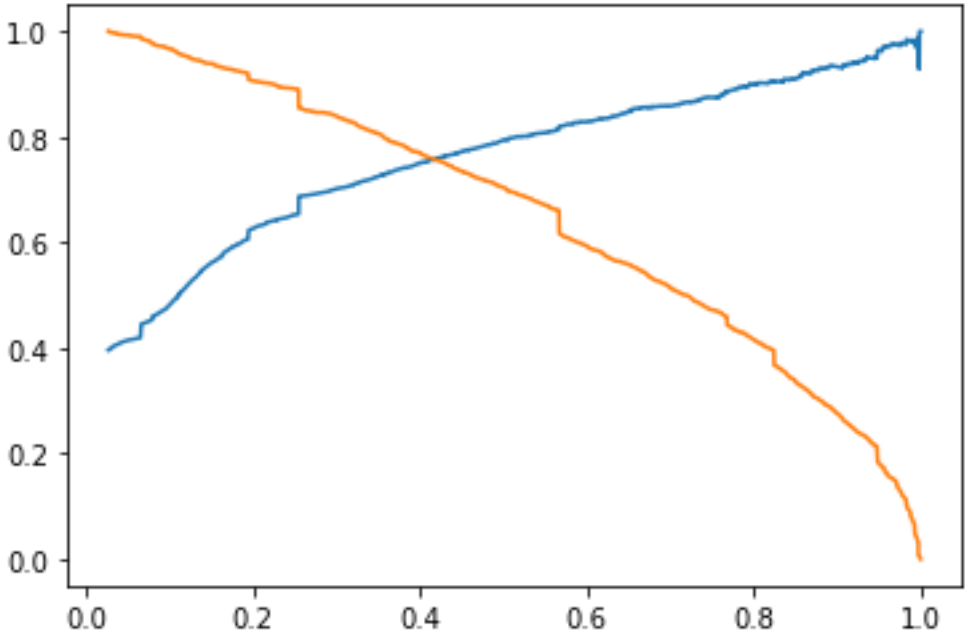
The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



- Accuracy - 81%
- Sensitivity – 70.4 %
- Specificity – 88.7 %
- False Positive Rate – 11.2 %
- Positive Predictive Value – 79.4 %
- Negative Predictive Value – 82.9%

Model Evaluation - Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



- Precision – 79.4 %
- Recall – 70.4 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Accuracy – 81.9 %
- Sensitivity – 79.6 %
- Specificity – 83.4%

Conclusion

- 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%.
- The top 3 variables that contribute for lead getting converted in the model are:
  - Total Visits
  - Total Time Spent on Website
  - Lead Origin-Lead Add Form
- In business terms, this model has an ability to go along with the company's requirements in coming future.
- Hence overall this model seems to be good.

Thank You!!