Case Study: Lead Scoring Model Building

- By Devanshu Bhardwaj

Case Study - Leading Scoring Model

Introduction

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 a higher lead score have a higher conversion chance and the customers with a 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%.

Objective of the Study

In this study, we built a logistic regression model using a leads dataset from the past with around 9000 data points and create a model that assigns 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.

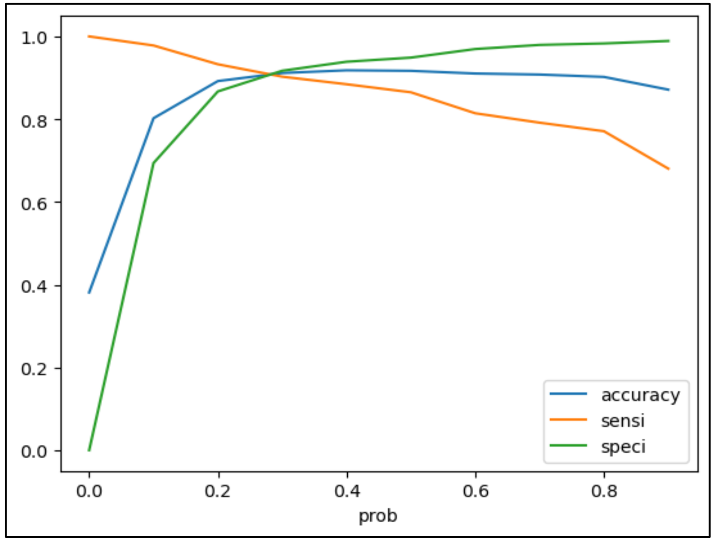
Approach –

We performed the following to build the Logistic Regression (LR) model for lead scoring –

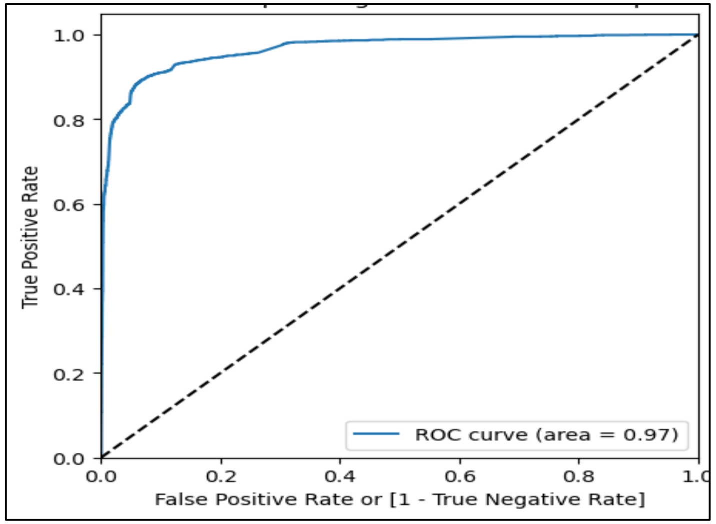
i. **Data Cleaning Steps** – Fixing Rows & Columns, Missing Value Handling, Variable Transformations, etc
ii. **EDA** - **Categorical & Numerical Variables**: Findings Variables with Good Predictive Power
iii. **Preprocessing –** Applied Normalization Scaling for Numerical & One-Hot Encoding for Categorical Variables, Divided data into test and train sets, & selected best variables using RFE
iv. **Model Building & Evaluation** – Created the LR Model using StatsModel.api Library following an iterative approach. Generated confusion matrix to find the accuracy% of the predictions made for both train & test sets.

Findings & Results of the Study

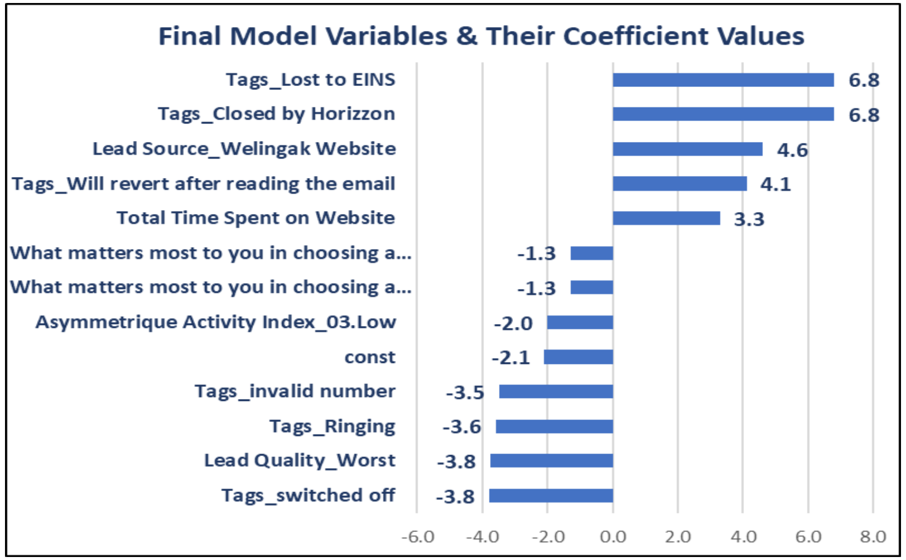
Accuracy vs Sensitivity vs Specificity of Model



ROC Curve



Final Model Variables



Confusion Matrix: Train & Test Set

Dev Set	Predicted	
Actual	Dropout	Converted
Dropout	3670	332
Converted	240	2226

Test Set	Predicted	
Actual	Dropout	Converted
Dropout	1571	106
Converted	121	974

	Dev Set		Test Set
Accuracy %	91.2%	**Accuracy%**	91.8%
Specificity	91.7%	**Specificity**	93.7%
Sensitivity	90.3%	**Sensitivity**	88.9%

Findings -

- **Area under Curve (AUC) or Receiver operating characteristic (ROC) curve** was used to evaluate and compare the performance of Logistic Regression model
- **Higher the AUC score, better the model** – ROC with area of 0.97 suggests that model is able to distinguish (separates) events and non-events well
- The point at where the **Accuracy, Sensitivity, and Specificity Curves** cross is considered the best probability cut-off value. Based on this technique, the model's **optimal probability cut-off is 0.3.**
- Using the 0.3 cut-offs, the **Accuracy Score for both the Train and Test Sets is in the 91%-92%** range, indicating good model performance.

Recommendations -

- X Education can transform the model's likelihood to lead scores and should target possible leads with scores more than 30 or probabilities greater than 0.3.
- A few variables improve the likelihood of lead conversion based on the Model Outcome. As a result, the business should prioritise consumers with favourable variables, i.e. variables with positive coefficient values in the model.