Leading Score Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. 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. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if they acquire 100 leads in a day, only about 30 of them get converted.

Objective:

- 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 we 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%.

Analysis Approach

- Data Cleaning
  - Replace Column value ‘Select’ with ‘NaN’
  - Plot the columns and decide the values to replace ‘NaN’
  - Perform Univariate Analysis on each column to find out appropriate values and replace ‘NaN’ with them
  - Check % of missing values and drop those columns having maximum missing values
  - Check the outliers if any. For huge outliers, drop those columns
  - Perform Univariate Analysis on each column and decide whether that column can be considered for lead conversion
  - Drop other columns which are not required for lead conversion

Analysis Approach

- Data Preparation
  - Covert Boolean variables to Binary variables
  - Create dummy variables for categorical columns
- Model Building
  - Split data to training and test sets
  - As part of feature scaling, use StandardScaler() to scale the columns except Binary/dummy variables
  - Run Training Model
  - Feature Selection using RFE
  - Build model using StatsModel
  - Making Predictions on Train set
  - Evaluate Model by checking VIFs, specificity, sensitivity, by plotting ROC, Optimal Cutoff point etc
  - Check Model performance over data
  - Generate Score variable
  - Making Predictions over test set

Result

- Top most columns which can be considered for conversion

  Tags_Lost to EINS

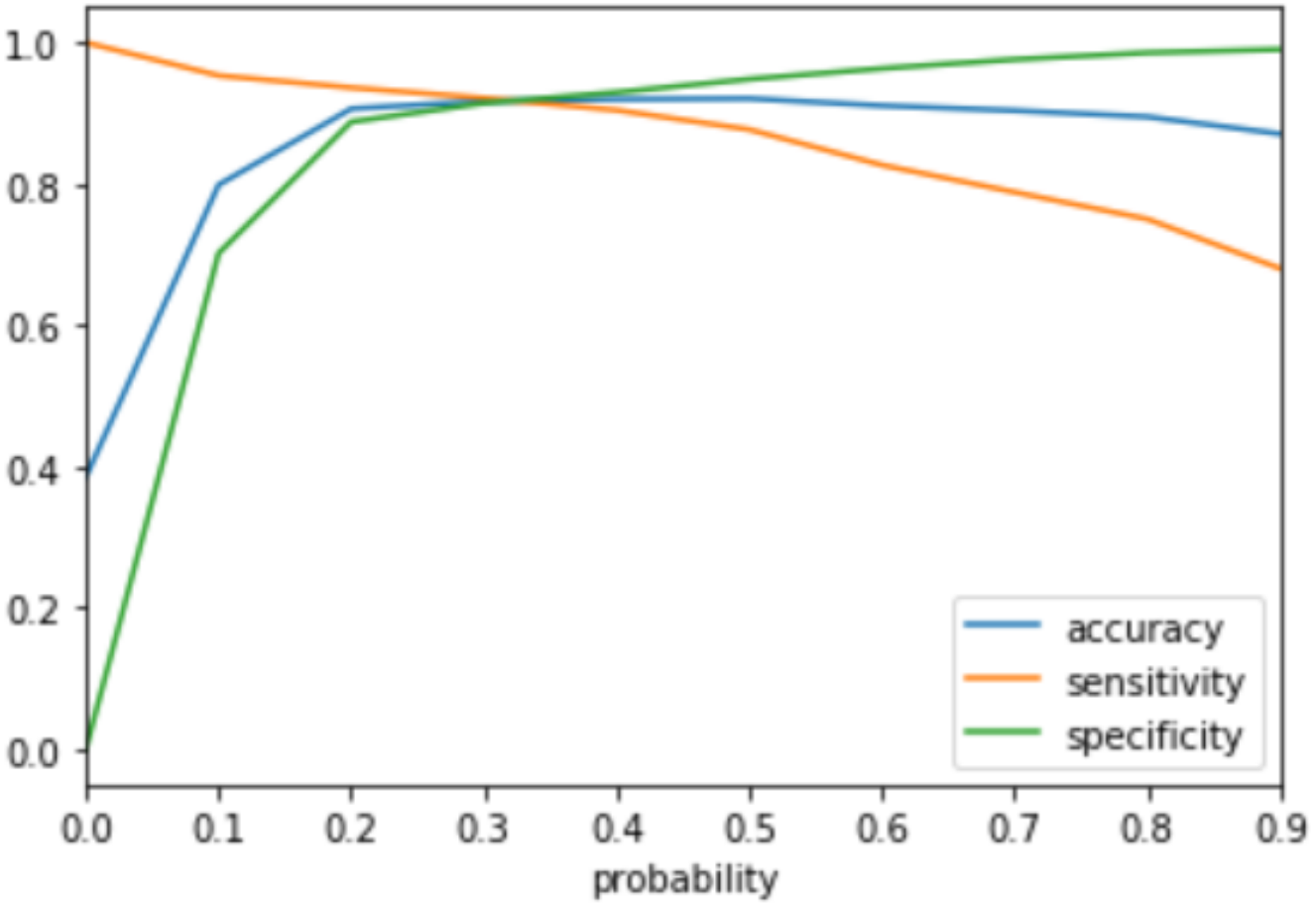
  Tags_Closed by Horizzon

  Tags_Will revert after reading the email

- Sensitivity is 92.5% which is quite high. It implies that there is high chance of identifying leads who wants to convert

Result

ROC Curve



Result

Precision_Recall Curve