LEAD SCORING CASE STUDY

Priyanka Neravati

Chakravarthy S L N Virivinti

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. 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. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. Although X Education gets a lot of leads, its lead conversion rate is very poor. The company wishes to identify the most potential leads, also known as 'Hot Leads'.
- Need 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.

- Goals of the 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.

Solution Methodology followed by us :

1. Data cleaning and data manipulation:

- Reading and understanding the data.
- Data pre-processing
- Missing value imputation
- Outlier analysis

2. Visualising the data:

- Performing EDA i.e univariate ,bi-variate and multi variate analysis on the data

3. Splitting the data into train-test split.
4. Feature Scaling & Dummy Variables.
5. Model building using - logistic regression.
6. Model Evaluation.
7. Making predications on training data and plotting ROC curve,knowing the threshold and cut off values, accuracy,sensitivity and specificity values etc.,
8. Making predications on test set.
9. Drawing Conclusions.

Step 1: Reading and Understanding the data.

a) Loading the data and understanding the data.

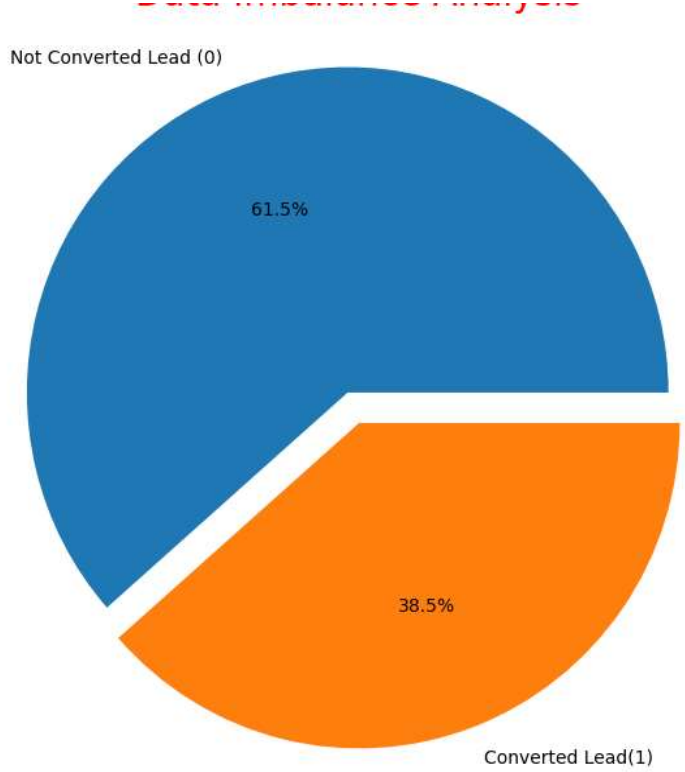
b) Data Cleaning which included data pre-processing,missing value imputation etc.,

- In Data pre-processing we removed the columns which represent the index as they are no way related to target varaiable. During missing value imputation and outlier analysis, we followed steps as follows:

- We dropped the columns having NULL values greater than 40%.
- Then, there were few columns with value ‘Select’ which means the leads did not choose any given option. We changed those values to Null values.
- Dropped columns which have single unique value as they do not affect the model.
- Dropping all the columns who have a single value weighing >95% in them , since keeping them makes the data skewed.
- Converting few columns with binary variables (Yes/No) to 0/1.
- We plotted box plots for numerical variables to see if there are any outliers in them. Also handled the outliers by doing a cutoff at 95% retaining most of the data.

Step 2 : Visualizing the data

- Performed univaraiate analysis for both categorical and numerical data
- Performed bi-variate and multi-variate analysis using pairplots and heatmaps.
- Also did data imbalance analysis.(as shown in figure here)



Step 3: Data Preparation for Modelling

- Dummy variable creation for categorical data.

Step 4: Splitting the data into Test Train Split.

- The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.
- Feature Rescaling.

a. We used the Min Max Scaling to scale the original numerical variables.

b. Then we did plot the a heatmap to check the correlations among the variables.

Step 5: Building a logistic regression model:

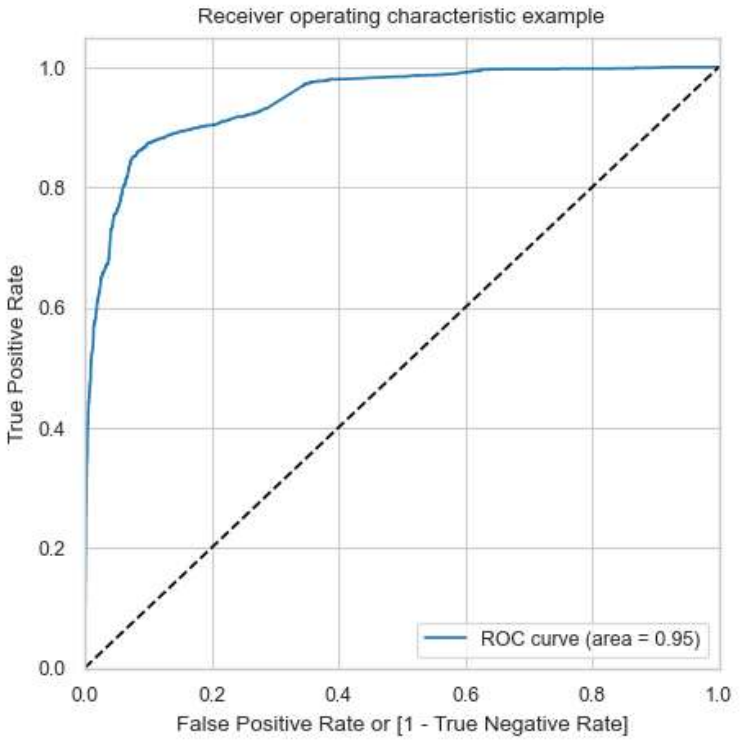
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5.

Step 6:

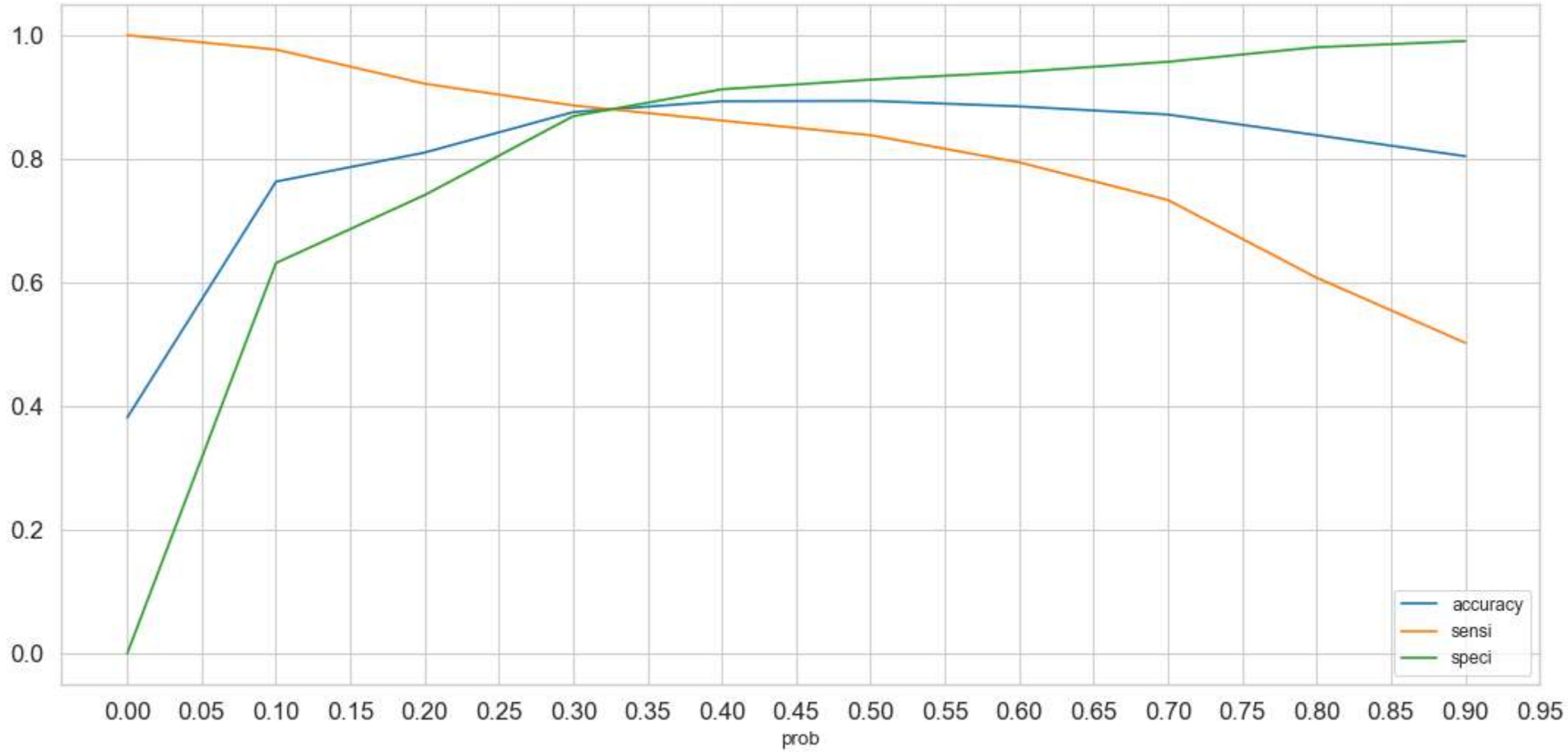
- Making predictions using the final model(I.e model 6)
- Predicting the probabilities (of the "converted" value being 1) on the train set.

Step 7,8: Model Evaluation

- Plotting ROC curve for tain dataset.(as shown in fig here)
- Accuracy : 88.54%, Sensitivity : 87.76% and Specificity : 89.02%

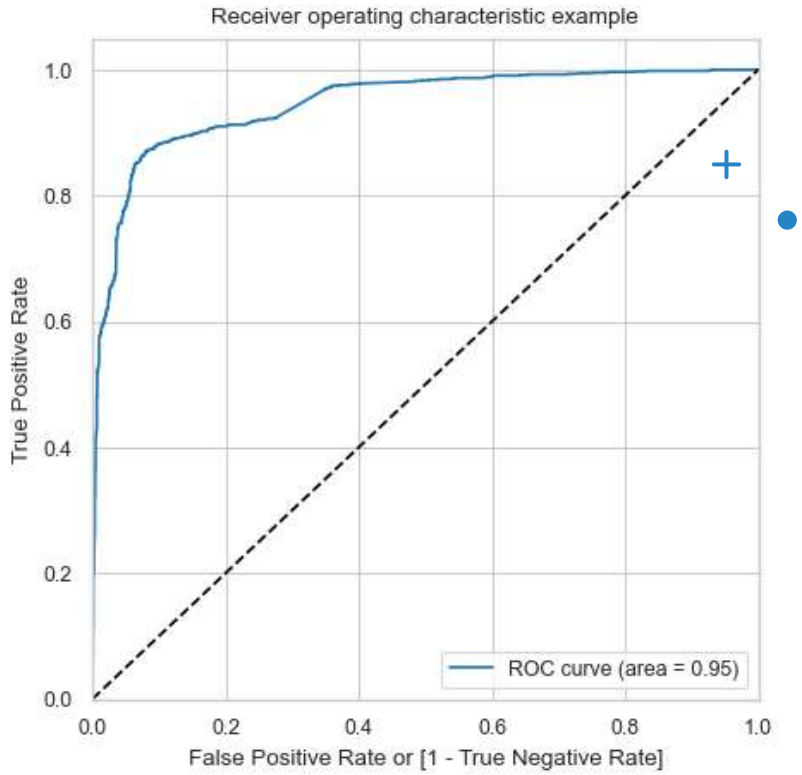


- Evaluating the model with optimal probability cutoff as 0.33 (as seen from the graph here.)



Model Evaluation for test dataset

- ROC curve for test dataset(as showin in figure here)
- Accuracy : 89.4%, Sensitivity : 87.89% and Specificity : 90.03%



Step 9:CONCLUSION:

- After trying several models, we finally chose a model no 6 with the following characteristics:
- -All variables have p-value < 0.05, showing significant features contributing towards Lead Conversion.
- -All the features have very low VIF values, means hardly there is any muliticollinearity among the features. This can be seen from the heat map.
- -The ROC curve has a value of 1, which is very good!
- -The overall accuracy of Around 89% at a probability threshold of 0.3 on the test dataset is also acceptable.
- The optimal threshold for the model is 0.33 which is calculated based on tradeoff between sensitivity, specificity and accuracy. According to business needs, this threshold can be changed to increase or decrease a specific metric. High sensitivity ensures that most of the leads who are likely to convert are correctly predicted, while high specificity ensures that most of the leads who are not likely to convert are correctly predicted.