Lead Score Case Study

Group Members
K Kishore
Ramanuj Singh Yadav
Kumari Vidya

Problem Statement

- 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. 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%.

Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Problem solving methodology

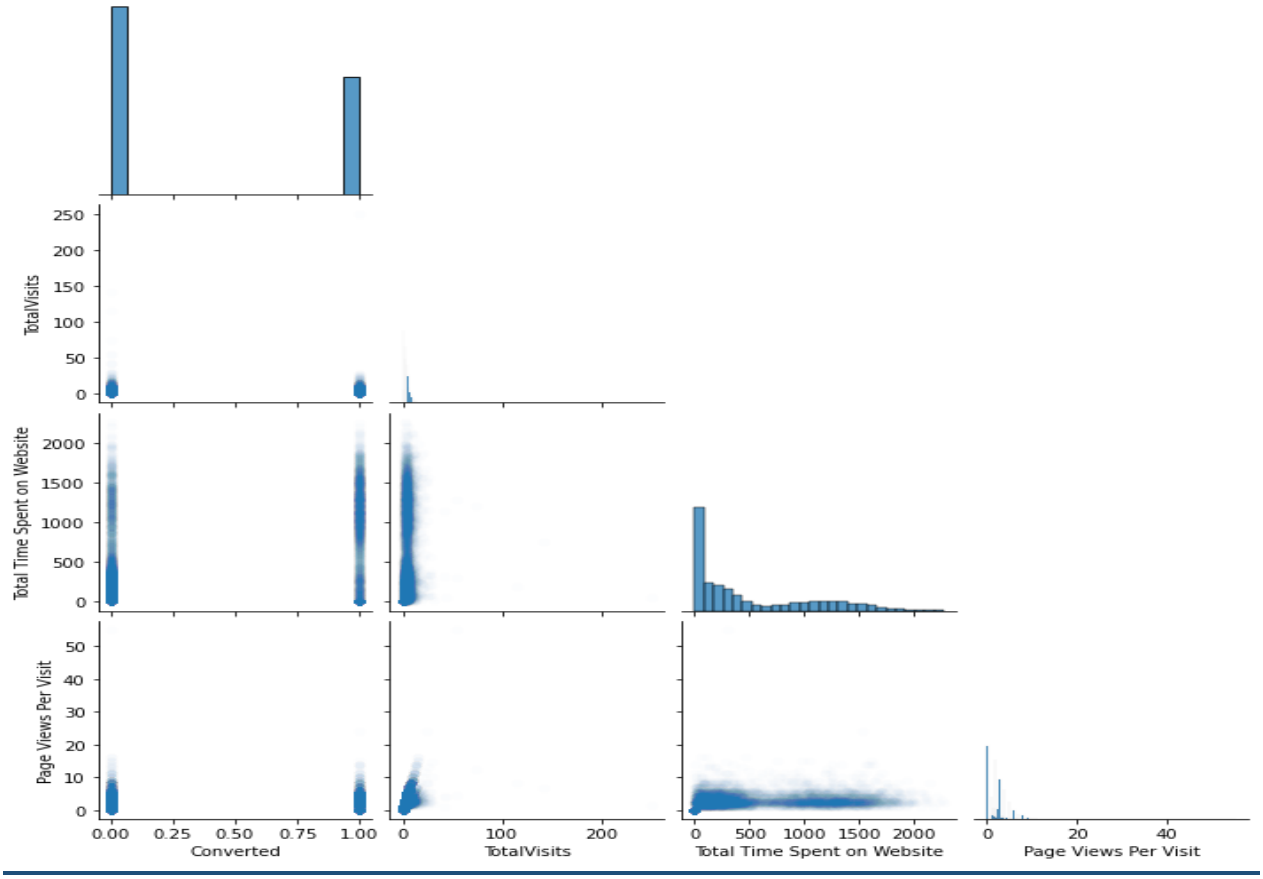
- Importing Libraries and Data File
- Cleaning the Data before proceeding to EDA
- Performing Analysis on the Data
- Preparing the data for Model Building
- Splitting the data into Train-Test datasets
- Splitting the data into Train-Test datasets
- Model Building
- Feature Selection using RFE
- Making Predictions on the Train Data
- Checking metrics using a confusion Matrix, Sensitivity and Specificity
- Plotting the ROC Curve to check for AOC
- Finding the Optimal cut off Point
- Making Predictions on the Test Set
- Model Evaluation

Data cleaning and data manipulation

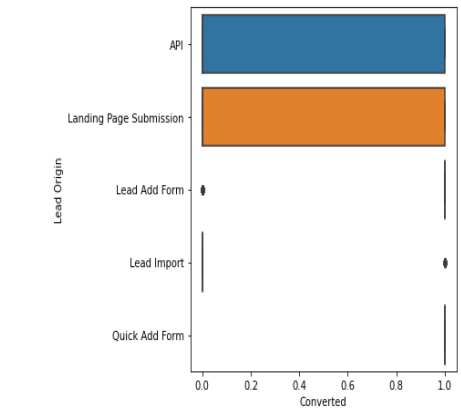
- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

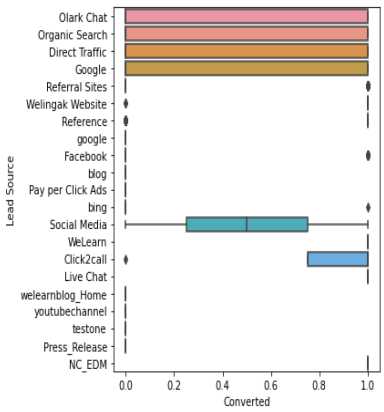
Performing Analysis on the Data

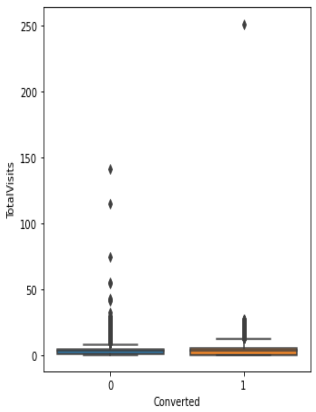
Pair Plot for numerical variables

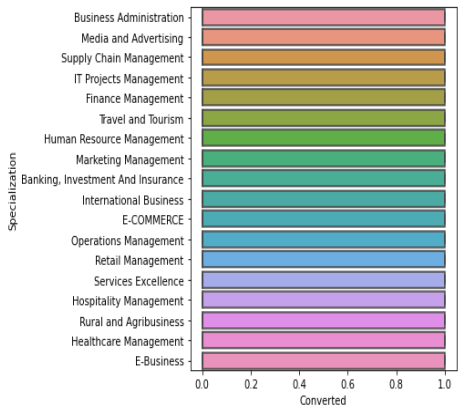


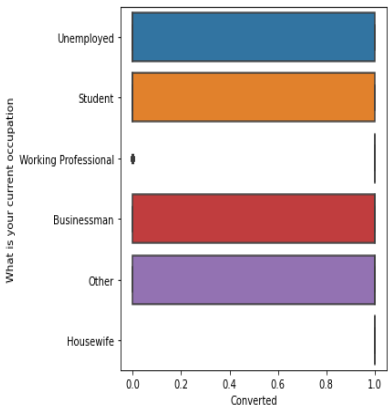
Box Plots for categorical variables

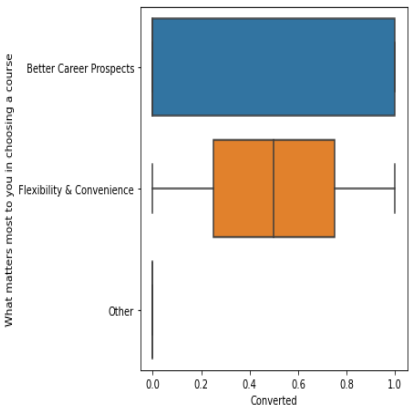


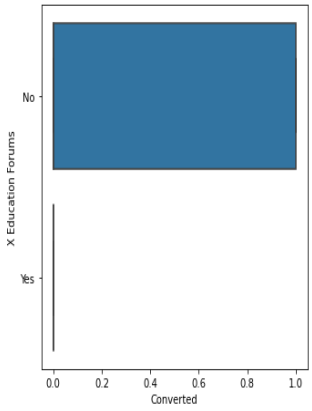


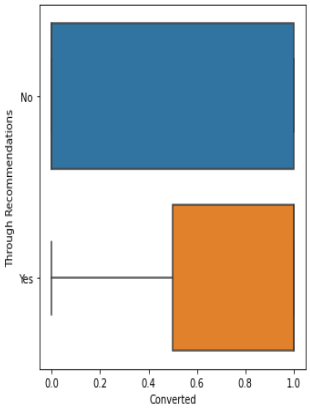


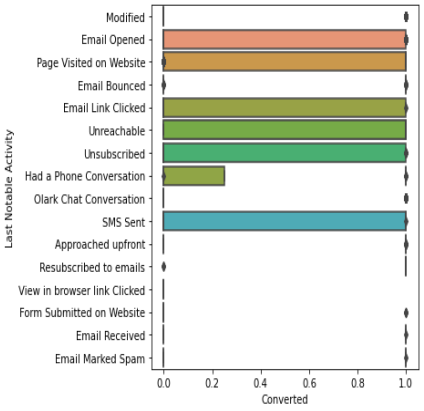












Data Conversion

- Converting Yes/No to 0/1
- Dummy Variables are created for object type variables
- Dropping Original columns for which dummies were created
- Converting all dummy variables for all categorical variables in the data

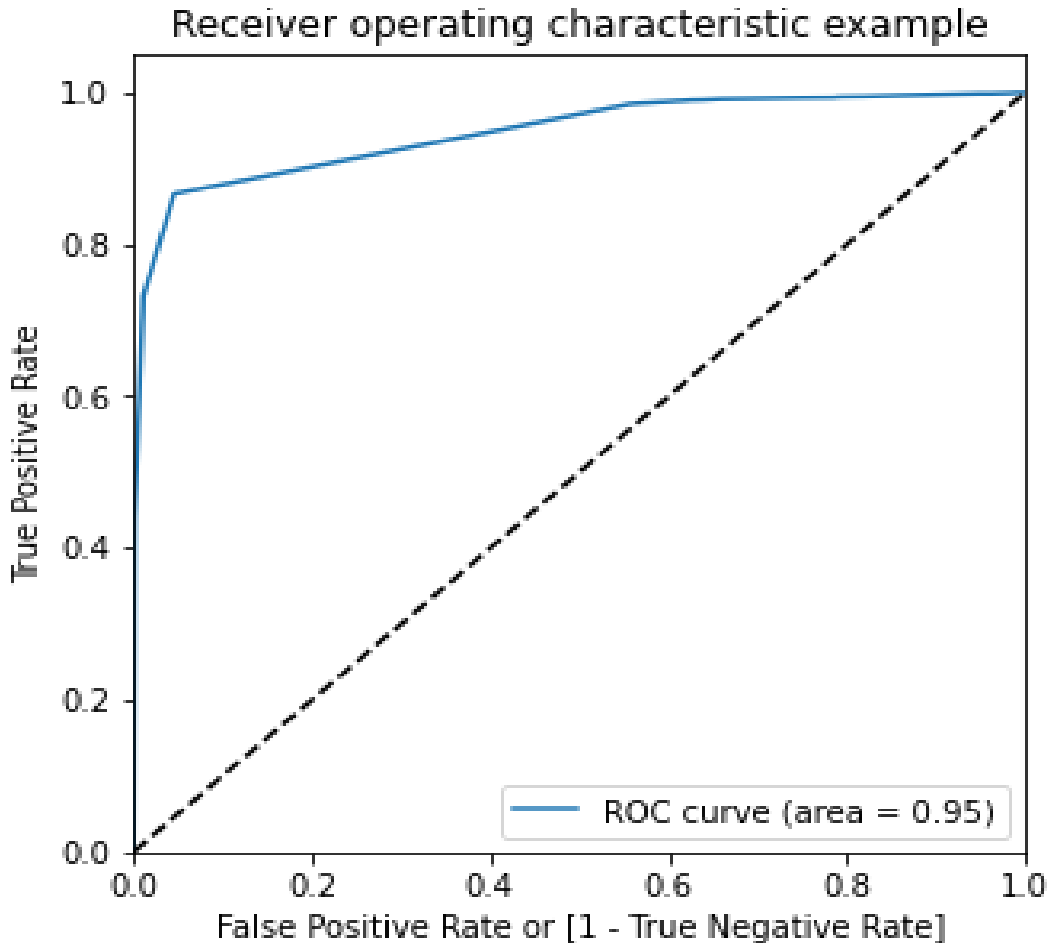
Model Building

- 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 variables as output
- Dropping the most insignificant values and constant
- Selecting a threshold of 0.5 to make predictions
- Predictions on test data set
- Overall accuracy 92%

ROC Curve

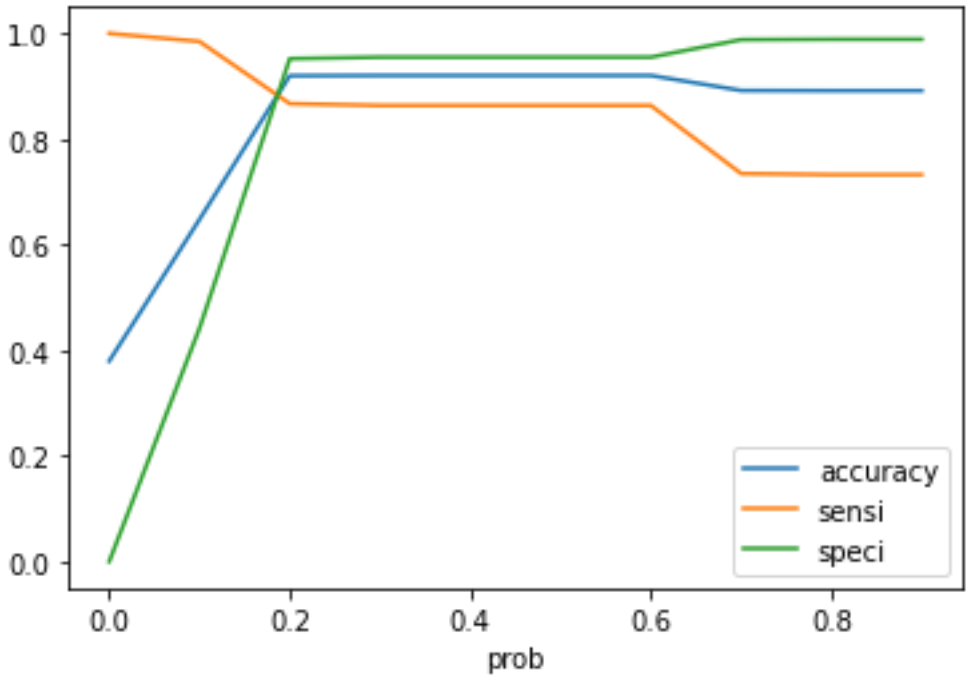
We plot ROC curve to check the trade off between Sensitivity and Specificity

Plotting the ROC curve:



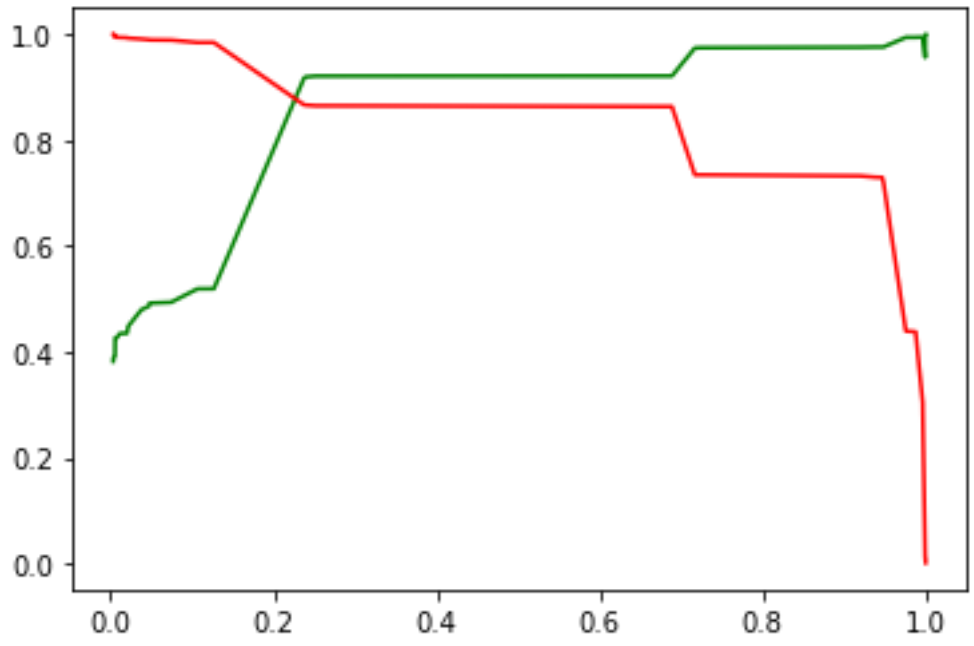
Finding Optimal Cut off Point

- Optimal cut off probability is the probability where we get balanced sensitivity and specificity.
- From the below graph it is visible that the optimal cut off is at 0.2.



Checking Precission and Recall

Scores show we have Precission of 92% and Recall of 86% for the Train data



Conclusion

While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.

Scores show we got a precision of 93% and a recall rate of 87% on the test data.

This value is almost identical to the values we got on the train data (92% Precission and 86% recall).

So, we can conclude to recommend this model for making good predictions.