LEAD SCORE CASE STUDY

Assignment Team Members
1) Vinayak Jadhav
2) Sahil Soni
3) Abinaya K

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ ₹
€ $
Converted Leads

Problem Statement

- X Education is an organization which provides online course for industry professional. The company marks its courses on several popular websites like google.
- The company generates leads through a variety of channels, including website visits, email marketing, and social media.
- X Education wants to select most promising leads that can be converted.
- Once a lead is generated, it is assigned a lead score based on a number of factors, such as the lead's industry, job title, company size, and level of engagement with X Education's content.
- The lead score is used to prioritize leads and determine which ones should be contacted by the sales team.
- The current lead scoring system at X Education is not very effective. The lead conversion rate is only around 30%, and the sales team is spending a lot of time on leads that are unlikely to convert.

Business Goals

- The goal of this case study is to build a new lead scoring system that will help X Education improve its lead conversion rate.

- Company wishes to identify the most potential leads, also known as “Hot Leads”

- The new system should be more effective than the current system in identifying leads that are most likely to convert. This will allow the sales team to focus their time and resources on the most promising leads, resulting in a higher conversion rate.

- The CEO of X Education has set a target lead conversion rate of 80%.

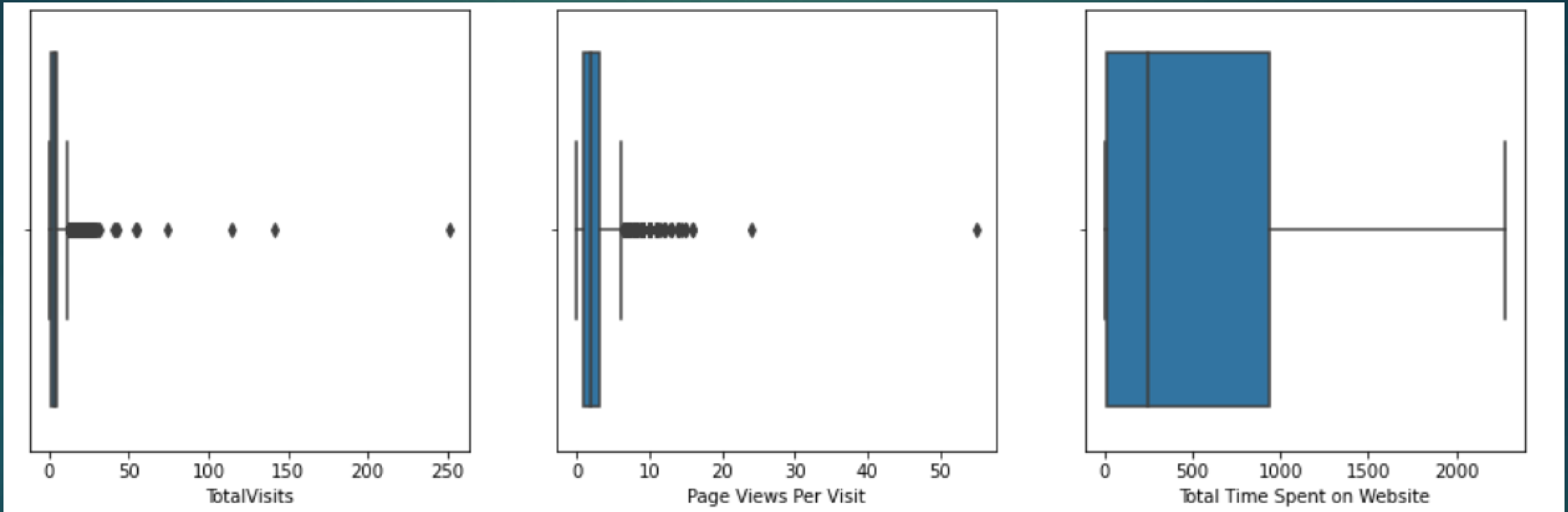
Approach for Analysis

- Understanding the domain of given data frame, variables mentioned in the file.
- Access and reading the csv formatted datasets for analysis.
- Identify the number of columns and rows present in each data frame
- Understanding the relevance of columns by its description.
- Checking the missing values from each columns
- Clean and prepare the acquired data for further analysis
- EDA for figuring out most helpful attributes for conversion
- Scaling features
- Data preparation for model building
- Build a logistic regression model
- Test the model on train set
- Evaluate model by different measures and metrices
- Test the model on test set
- Measure accuracy of model and other metrices

Data Conversion

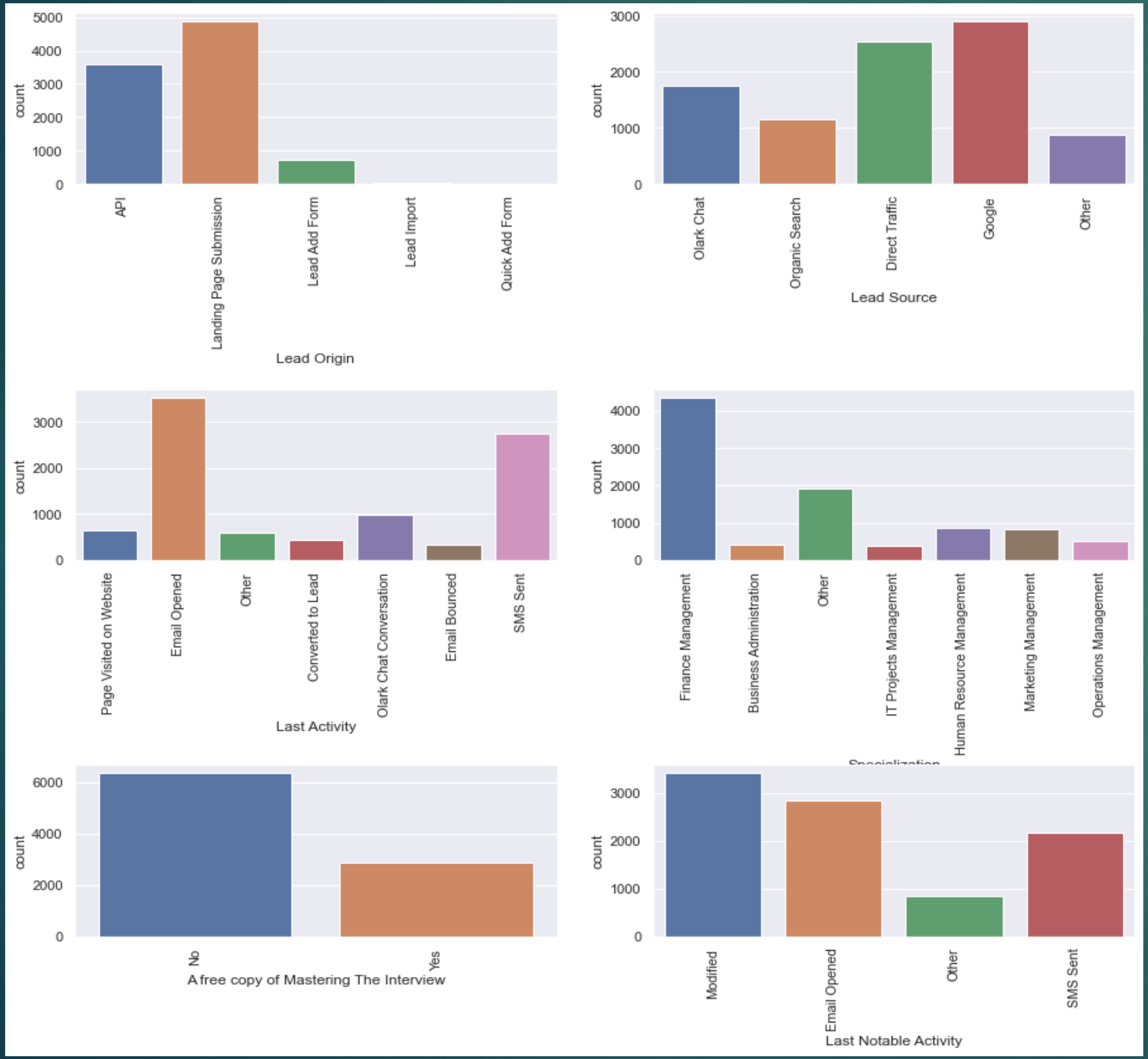
- Converting the variable with values Yes/No to 1/0s
- Converting the ‘Select’ values with NaNs
- Dropping the columns having > 40% of Null values
- Dropping unnecessary columns
- Dropping the rows as the Null Values were <2%

Outlier Detection



- Outliers are observed in both the variables of Total Visits and Page Views Per visit. It is required to be treated.
- The value spreaded above median highly in Total time spent on website

EDA – Univariate Analysis



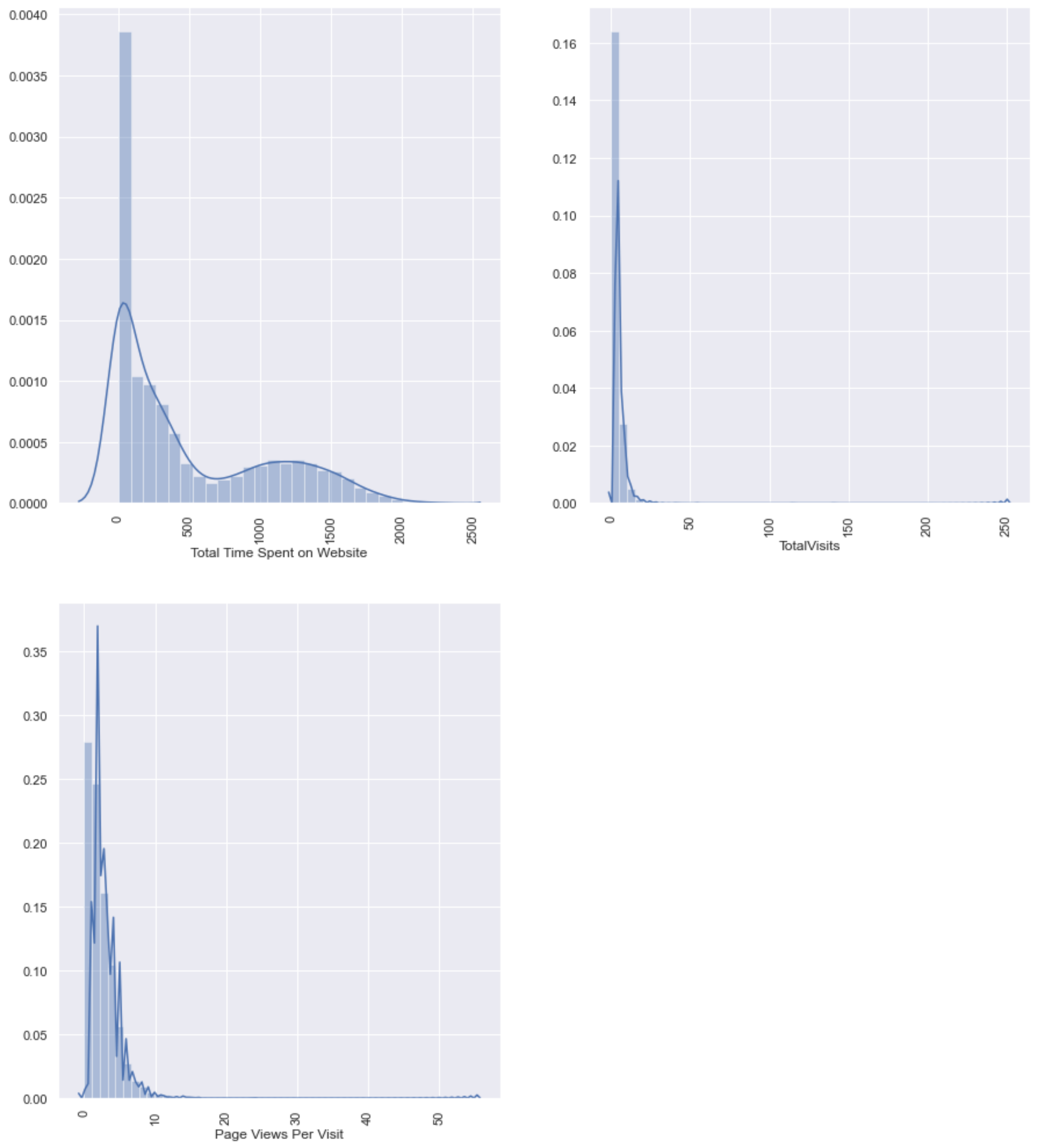
- In Lead Source Direct Traffic and Google are the two main source for Leads

- The Number of values is High in Email Opened and SMS Sent in Last Activity

- Most of the people chooses Finance Management Specialization rather than other Specialization

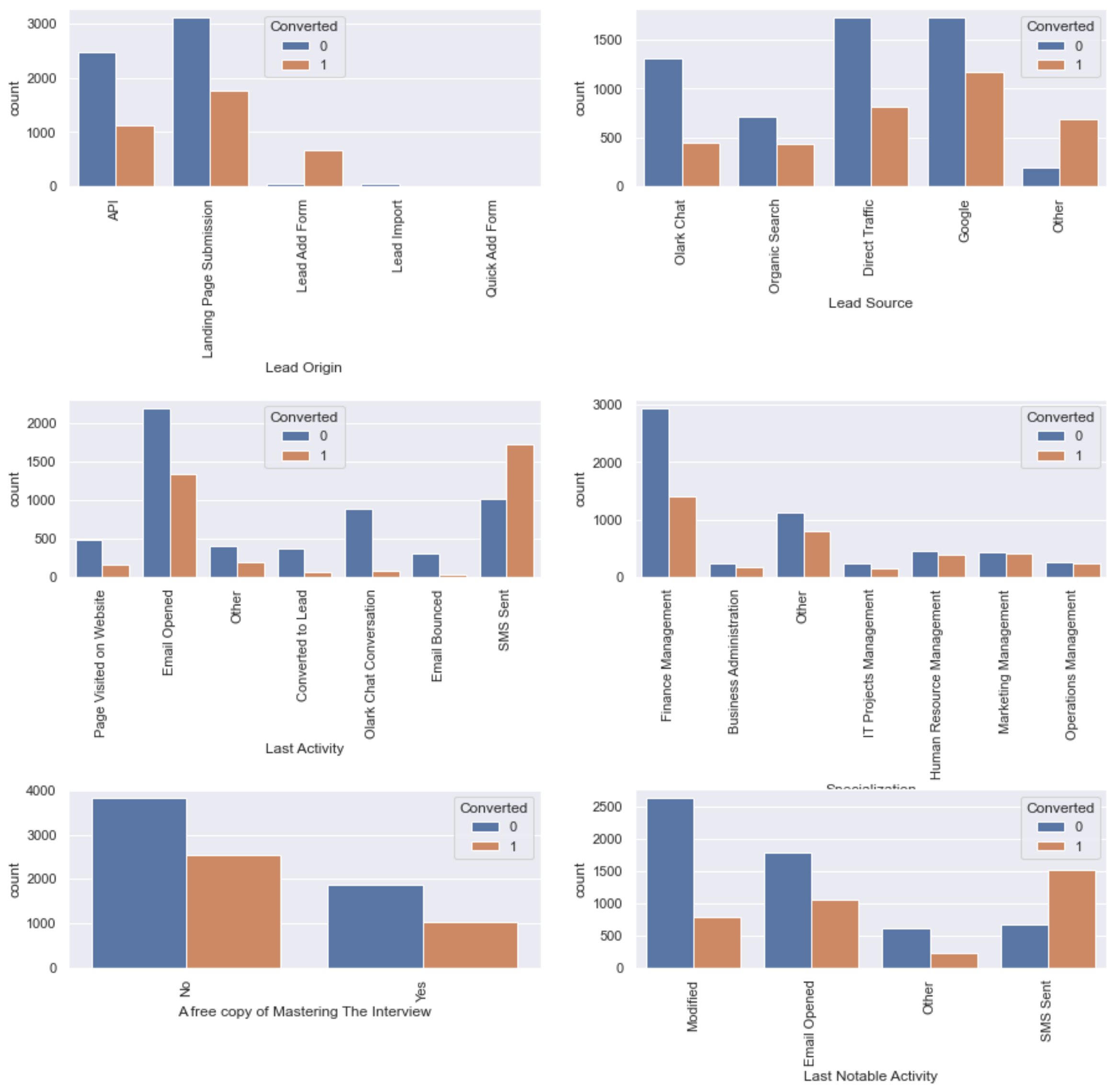
- The IT Project management have very lees so that most of the People not prefered this Specialization

EDA – Univariate Analysis



- None of the Continueous Variables are in Normal distribution
- Presence of Outliers in Total Visits and Page Views Per Visit
- In total visits more values is between 0-50 and page views per visits 0-20

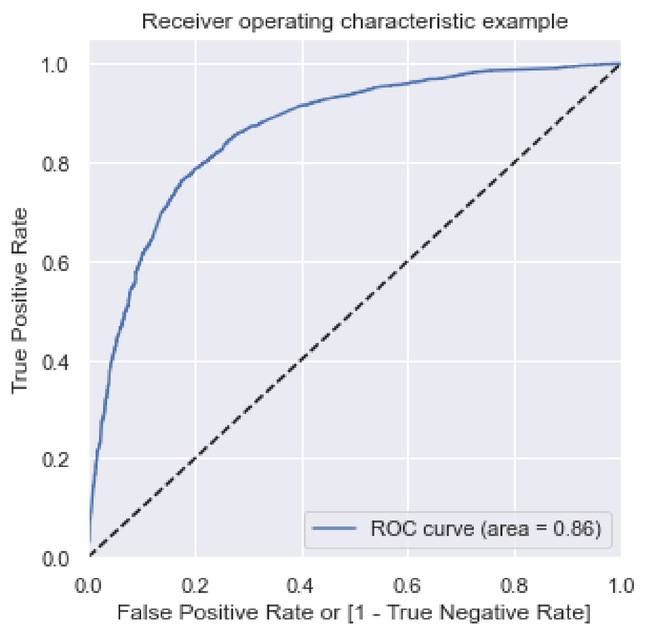
EDA – Bivariate Analysis



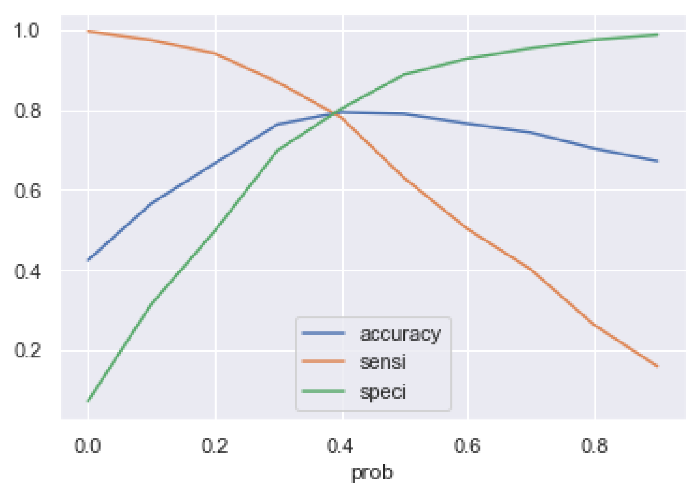
- In Lead Source The number of Hot leads is higher in Direct Traffic and Google less in Other Category
- In Last Activity the number of Hot leads is higher in SMS and in EMAIL cold leads is higher than hot leads.
- In Last Notable Activity it's mostly same as Last Activity.
- In Specialization the most of the leads are comes from Finance management but here Hot leads are lesser than Cold leads.

Model Building

- Splitting data in train and test sets
- Chosen the train test split ratio as 70:30
- Using RFE to choose top 20 variables
- Build model by removing the variables where p-Value >0.05 and VIF >5
- Predictions on Test data set



ROC Curve

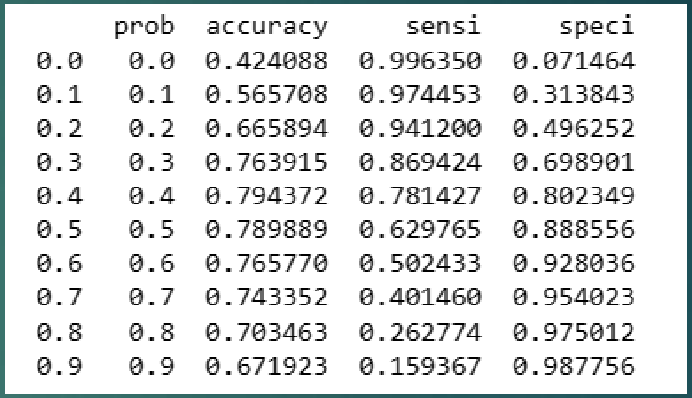


Optimal Cut-Off

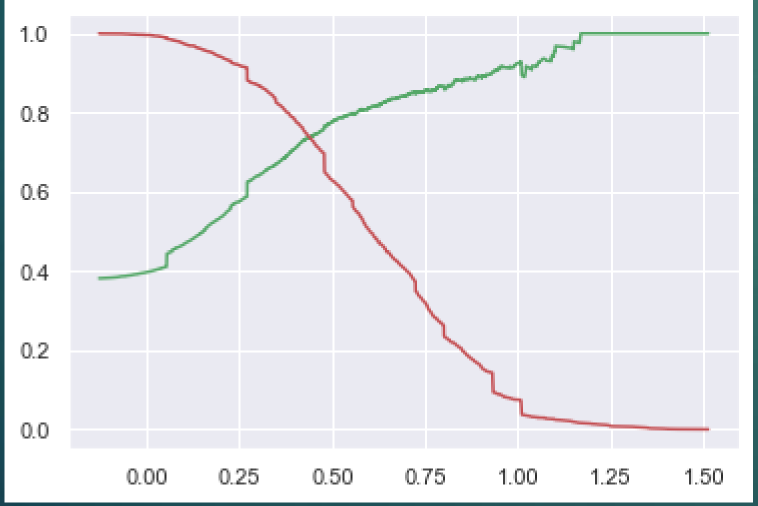
Model Evaluation (Train & Test)

- Calculated accuracy, sensitivity, specificity for various probability cutoffs from 0.1 to 0.9
- As per the graph, cutoff is 0.37
- Confusion Matrix for train dataset

  [3556, 446], [913,1553]



```
     prob  accuracy     sensi     speci
0.0   0.0  0.424088  0.996350  0.071464
0.1   0.1  0.565708  0.974453  0.313843
0.2   0.2  0.665894  0.941200  0.496252
0.3   0.3  0.763915  0.869424  0.698901
0.4   0.4  0.794372  0.781427  0.802349
0.5   0.5  0.789889  0.629765  0.888556
0.6   0.6  0.765770  0.502433  0.928036
0.7   0.7  0.743352  0.401460  0.954023
0.8   0.8  0.703463  0.262774  0.975012
0.9   0.9  0.671923  0.159367  0.987756
```



0.37 as the Cut-off as Precesion-Recall Thresholdm

	Train Data	Test Data
Accuracy	79%	70%
Sensitivity	81%	28%
Specificity	77%	97%

Conclusion

We have noted that the variables that important the most in the potential buyers are:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was: a. Google b. Direct traffic c. Organic search d. Olark Chat
- When the last activity was: a. SMS b. Olark chat conversation
- When the lead origin is Lead add format.