LEAD SCORE CASE STUDY

Presented by

Priyanka

Alnawaz

Sai Kiran

BUSINESS PROBLEM & OBJECTIVE

- *X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines also gets leads through past referrals.*
- *X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.*
- *X Education wishes to identify the most potential leads, also known as 'Hot Leads', If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.*

OBJECTIVE

- *Company wants Model that Identify most promising lead knows as Hot Lead. So that it help CEO to achieve conversion rate around 80%*
- *Model should able to adjust change with the future requirement of the company .*

STEPS AND STRATEGY FOR MODEL BUILDING

- Importing and understanding data
- Data Cleaning
- Handling missing values
- Identifying & handling Outlier
- EDA - Univariate and Bivariate analyses
- Data Preparation - Creating dummy for categorical variable
- Features Scaling
- Splitting Data in to Train and Test
- Building logistic Regression model checking P-value and VIF and finalising Model
- Model evaluation on different metrics - Accuracy , Sensitivity, Specificity , Precision & Recall.
- Final Model Prediction on Test set
- Analysing the metric score for train and test dataset ,
- Conclusion and Recommendation

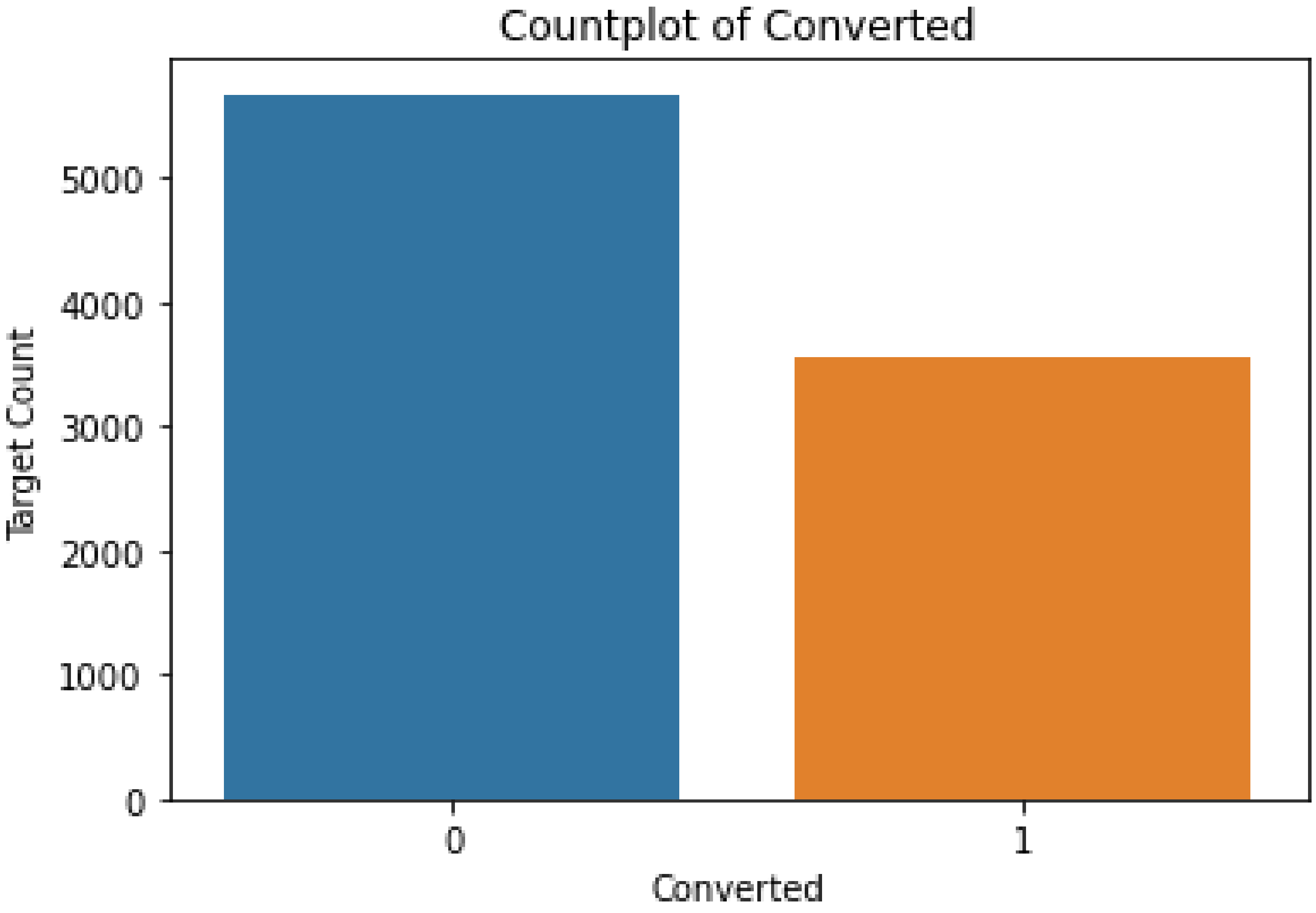
ASSUMPTIONS AND DATA CLEANING

- Data received in csv file with 9240 rows and 37 columns.
- Dropped the column having more than 39 % missing value.
- Categorical feature showing "Select" value considered it as nan value since the customer not provided any details.
- For missing values in categorical variable Mode considered to impute values , for Specialization we considered missing values as Other
- For missing Numerical variable we used Median since the numerical feature having outlier.
- For "Lead Source" there few levels showing less frequent so created a new level Other
- We find some of categorical variable not having enough variance so not able to contribute any information to our analyses hence dropped.
- We dropped "Lead Number" as ID not necessary for our analyses.

EXPLORATORY DATA ANALYSES

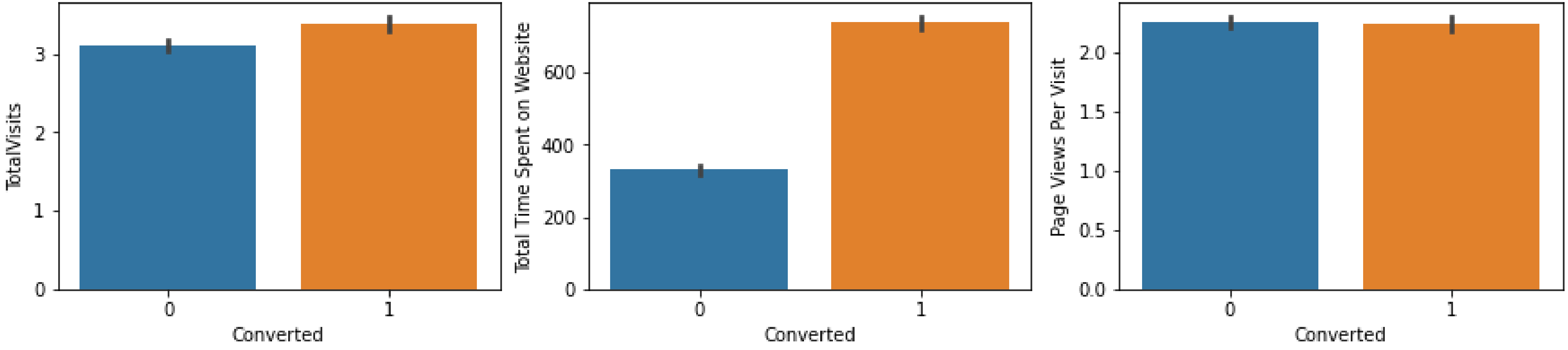
TARGET DATA

We found current Lead conversation rate as 38.54%

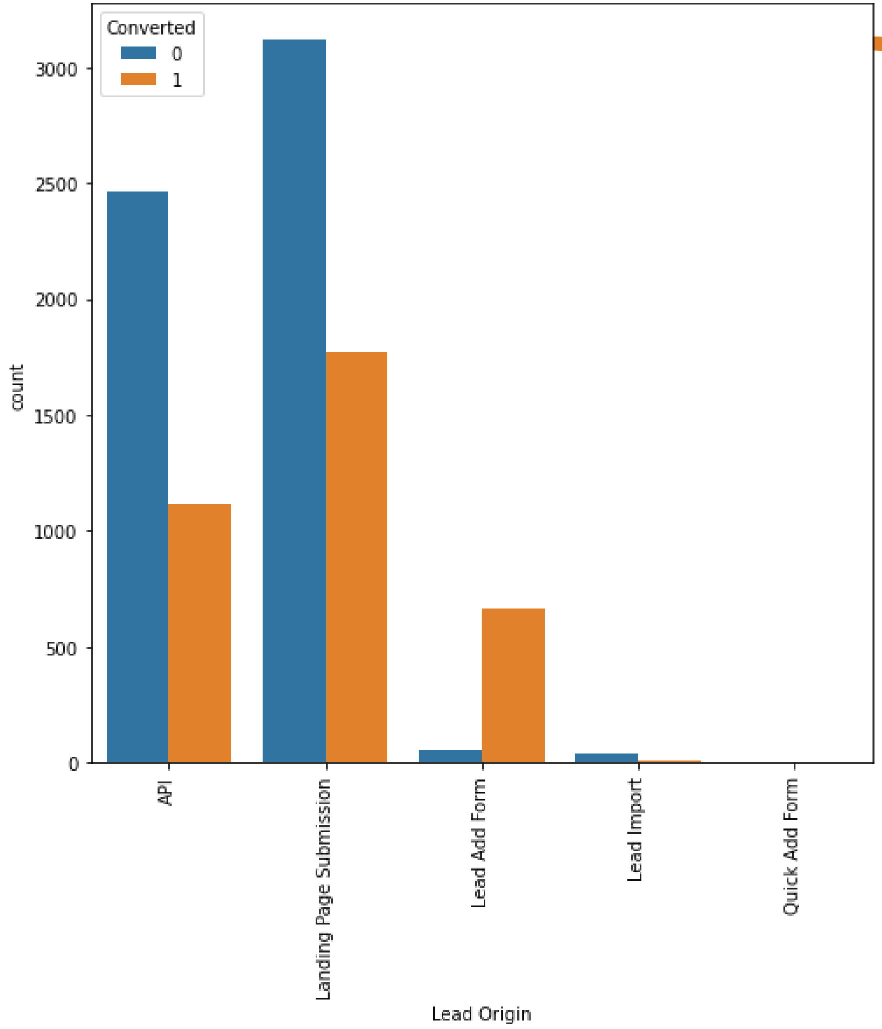


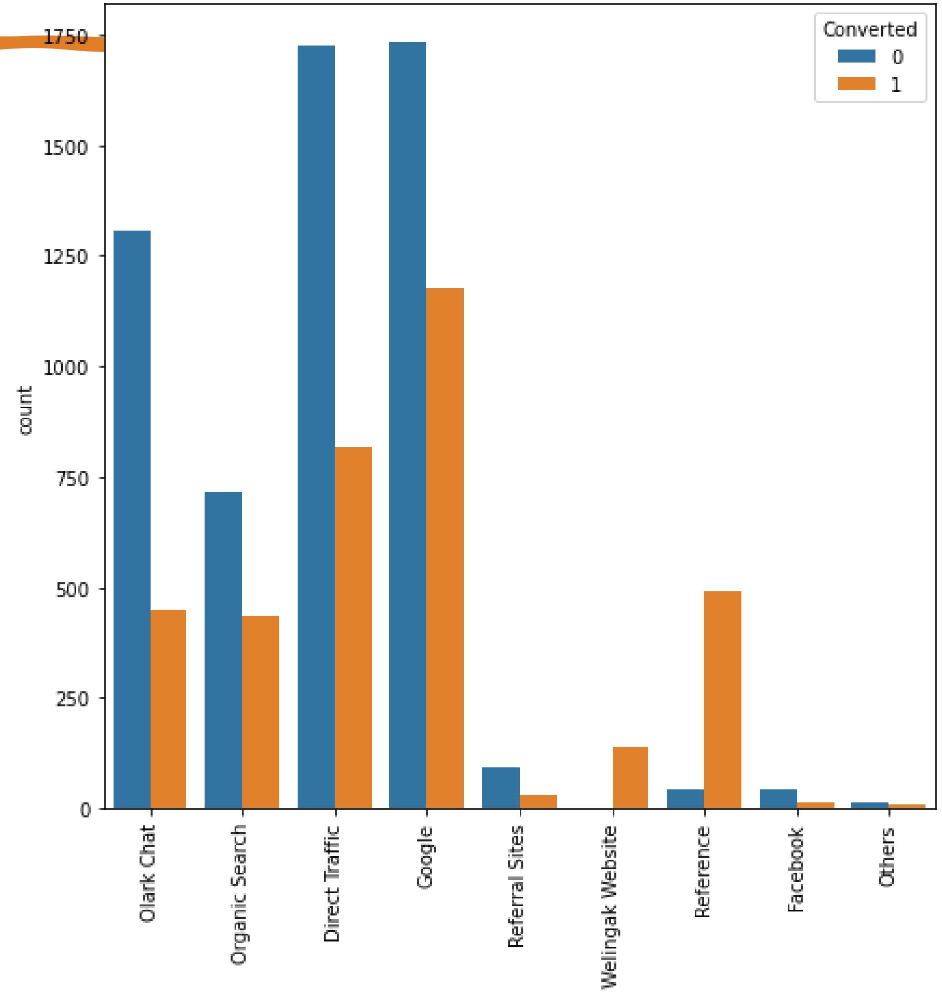
NUMERICAL VARIABLE VS CONVERTED

- "TotalVisit" and "Page Views per visit show " are showing 50% conversion, whereas "Total time spent on website" showing high conversion



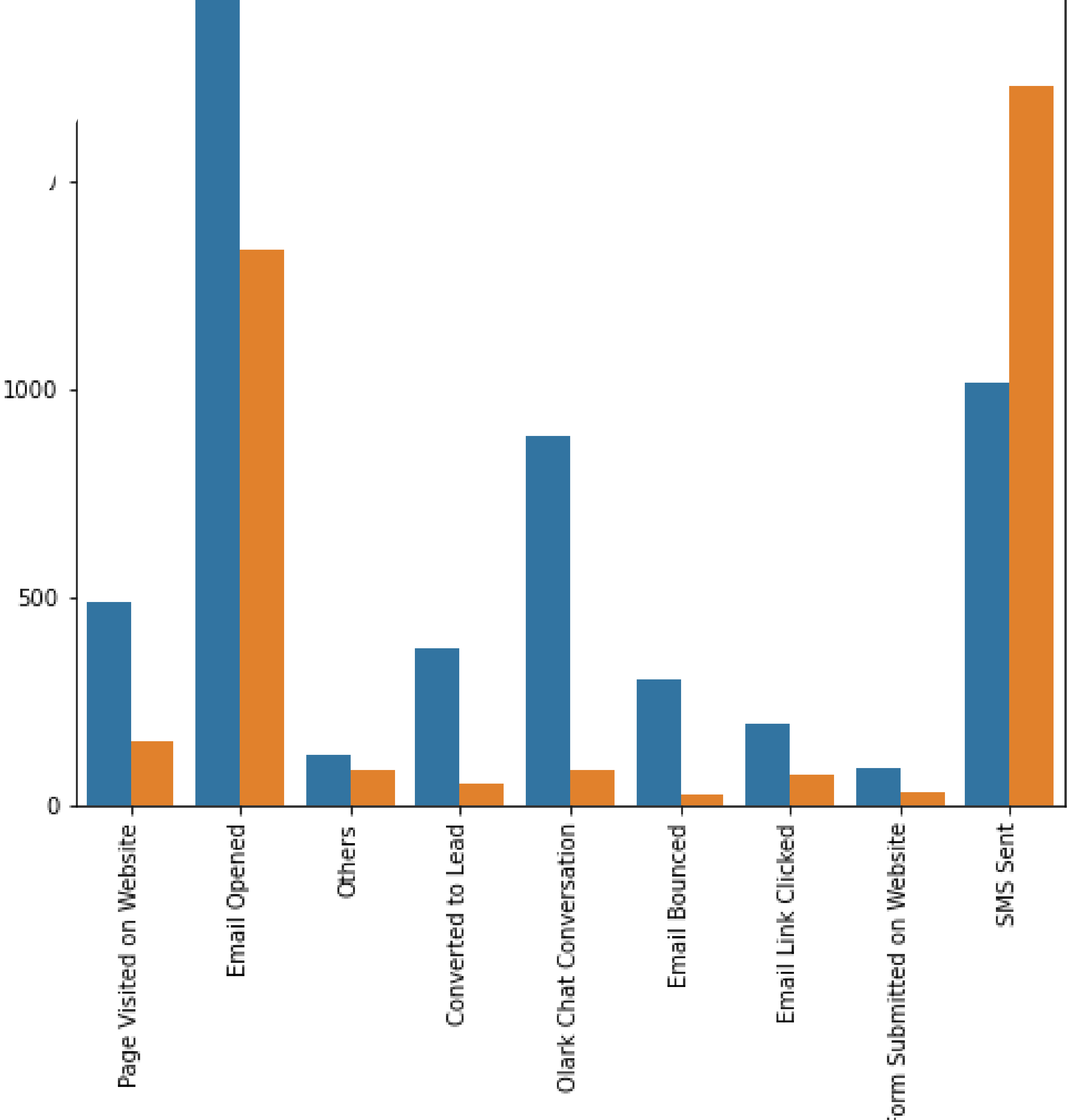
CATEGORICAL VS CONVERTED





LAST ACTIVITY VS CONVERTED

- We find SMS sent showing high conversion



DATA PREPARATION

- Dummy variable created for categorical variable with multiple levels
- Binary variable are encoded with 0/1
- Numerical features are standardized
- Data shape is 9240 rows and 64 columns
- Checked correlation to drop some of highly correlated features

MODEL BUILDING

- We split dataset into train and test using 70:30 ratio
- Used RFE method for feature selection and selected 20 features.
- We build 6 model and selected final model with low p value ie; below 0.05 and low VIF ie; below 5.
- From model we find the top 3 variable as below

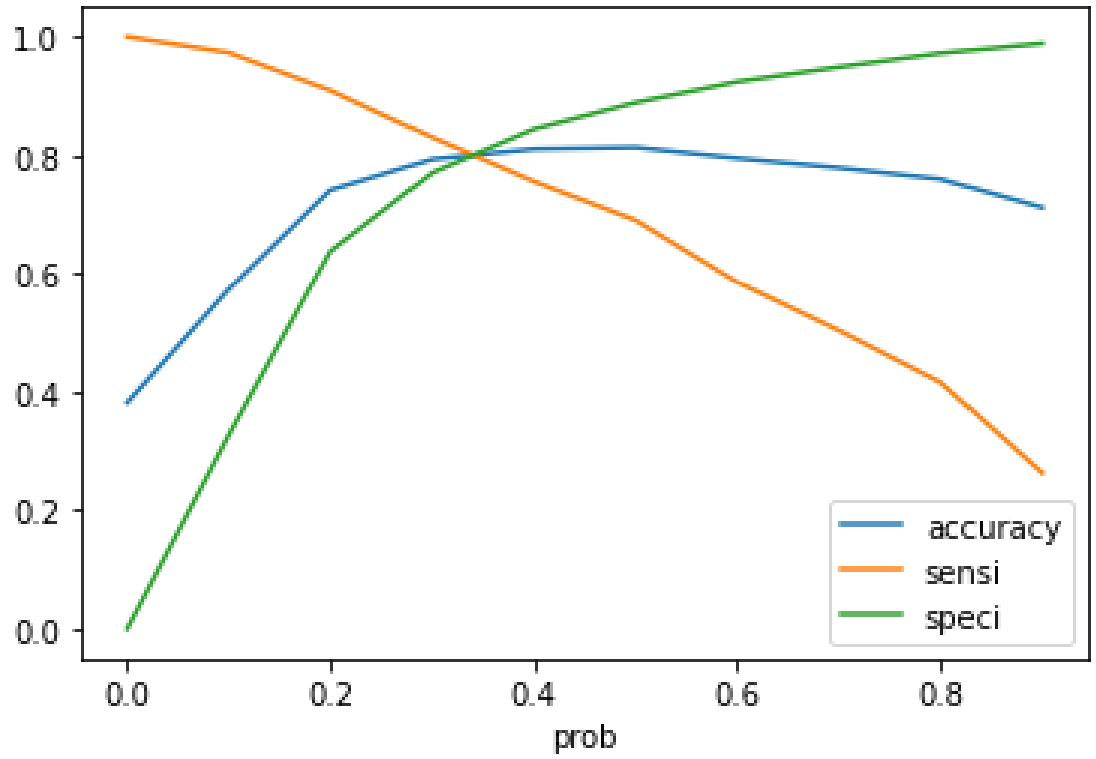
1. Lead Source_Welingak Website
2. Lead Source_Reference
3. Last Notable Activity_Had a Phone Conversation

- **Important Variable with coefficient**

Lead Source_Welingak Website	5.588836
Lead Source_Reference	3.773473
Last Notable Activity_Had a Phone Conversation	2.552070
const	1.549072
Lead Source_Others	1.294941
Last Activity_SMS Sent	1.277861
Lead Source_Olark Chat	1.108470
Total Time Spent on Website	1.074644
Last Activity_Others	0.879437
Last Activity_Olark Chat Conversation	-0.883814
Last Notable Activity_Modified	-0.891432
Specialization_Hospitality Management	-0.903375
Do Not Email	-1.229556
What is your current occupation_Student	-2.361526

MODEL EVALUATION
TRAIN - ACCURACY , SENSITIVITY , SPECIFICITY

ROC curve showing optimal cutoff 0.35



Confusion Metrix

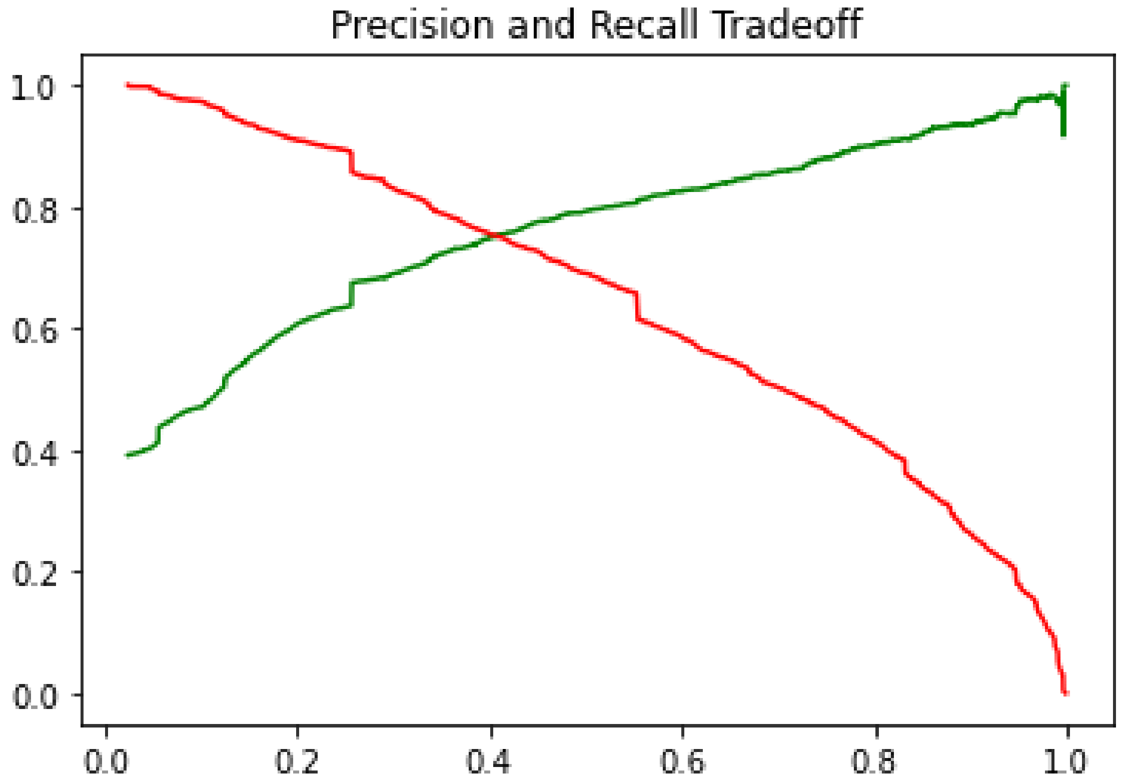
3257	745
521	1945

- Accuracy : 0.80
- Sensitivity :0.78
- Specificity :0.81
- False Positive rate : 0.18
- Positive predictive value :0.72
- Negative predictive value : 0.86

PRECISION AND **RECALL** SCORE FOR TRAIN

Precision score : 0.79

Recall score : 0.69



MODEL EVALUATION ON TEST DATA SET

- Accuracy :0.81
- Sensitivity :0.80
- Specificity :0.82

CONCLUSION AND RECOMMENDATIONS

- *We evaluated our model on both Sensitivity , Specificity and Precision, Recall metrics and used optimal cutoff based on ROC curve for final prediction.*
- *We calculated and analyses Accuracy ,Sensitivity & Specificity for both Train and Test data set and found it is approximately closer which indicate that model is good.*
- *Top5 Important variables from the model are as below which we recommend to the company to consider while contacting the leads.*

1. *Lead Source_Welingak Website*
2. *Lead Source_Reference*
3. *Last Notable Activity_Had a Phone Conversation*
4. *Last Activity_SMS Sent*
5. *Lead Source_Olark Chat*

- *Model seems good to enhance the conversion rate around 80%*