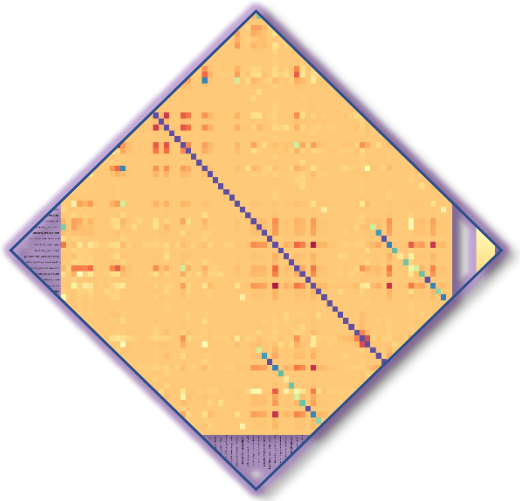
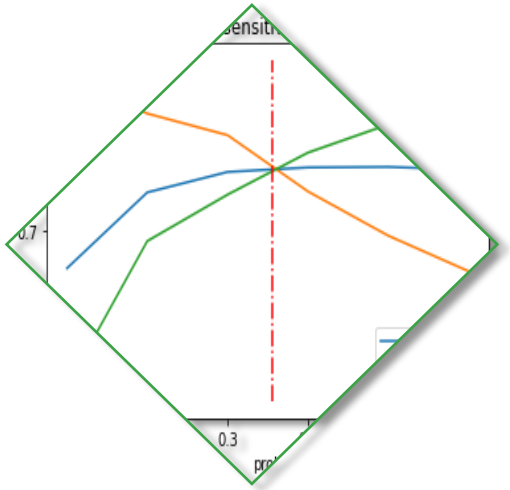
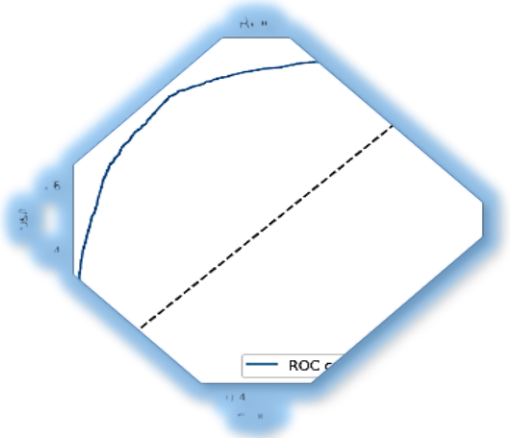
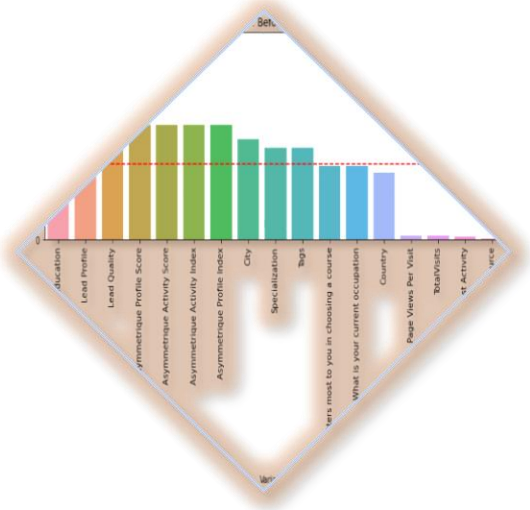




LEAD SCORE CASE STUDY





Submitted By
FAWAZ KHAN
SHARUN UTHARAN
NAMRATA PATEL

INDEX

Problem Statement

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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.
The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

In order to achieve that , we need to build a model and assign a lead score between 0 to 100 to each lead. Based on logistic regression model and the given requirement, a cut-off lead score will be selected to classify if the lead will be converted or not.

Overall Approach

1

EDA

Perform data pre-processing operations.

1. Reading and understanding data
2. Data Imbalance
3. Data Clean up
4. Impute Data
5. Univariate Analysis
6. Check correlation

2

Model Building

1. Data Preparation
2. Train-Test Split
3. Building initial model
4. Building models using RFE

3

Model Prediction

1. Test and train prediction
2. Check Confusion Metrics

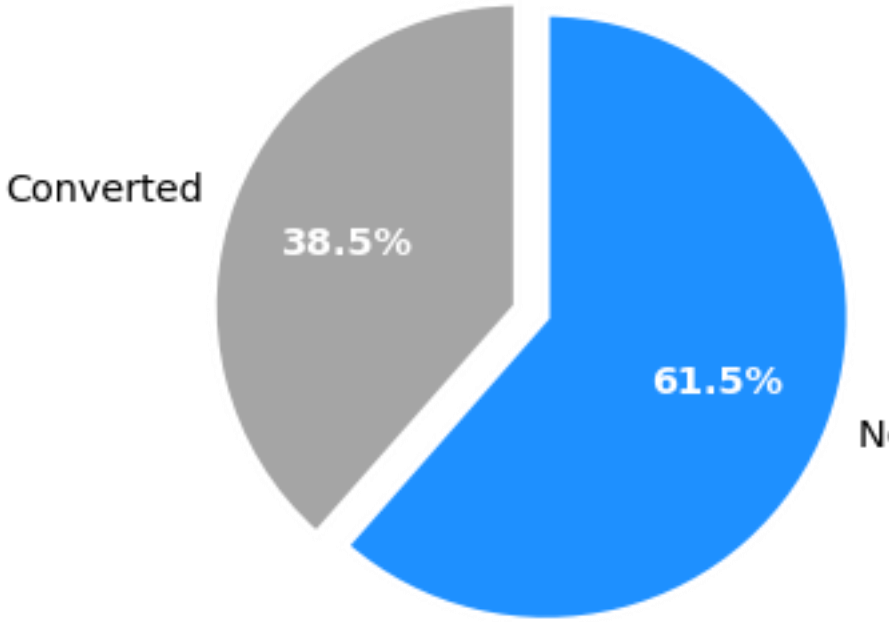
4

Model Evaluation

1. ROC Curve
2. Confusion Metrics and Scores
3. Train Dataset on Cut off
4. Precision Recall Plot
5. Prediction on Test Set
6. Merge Dataset with Main Data

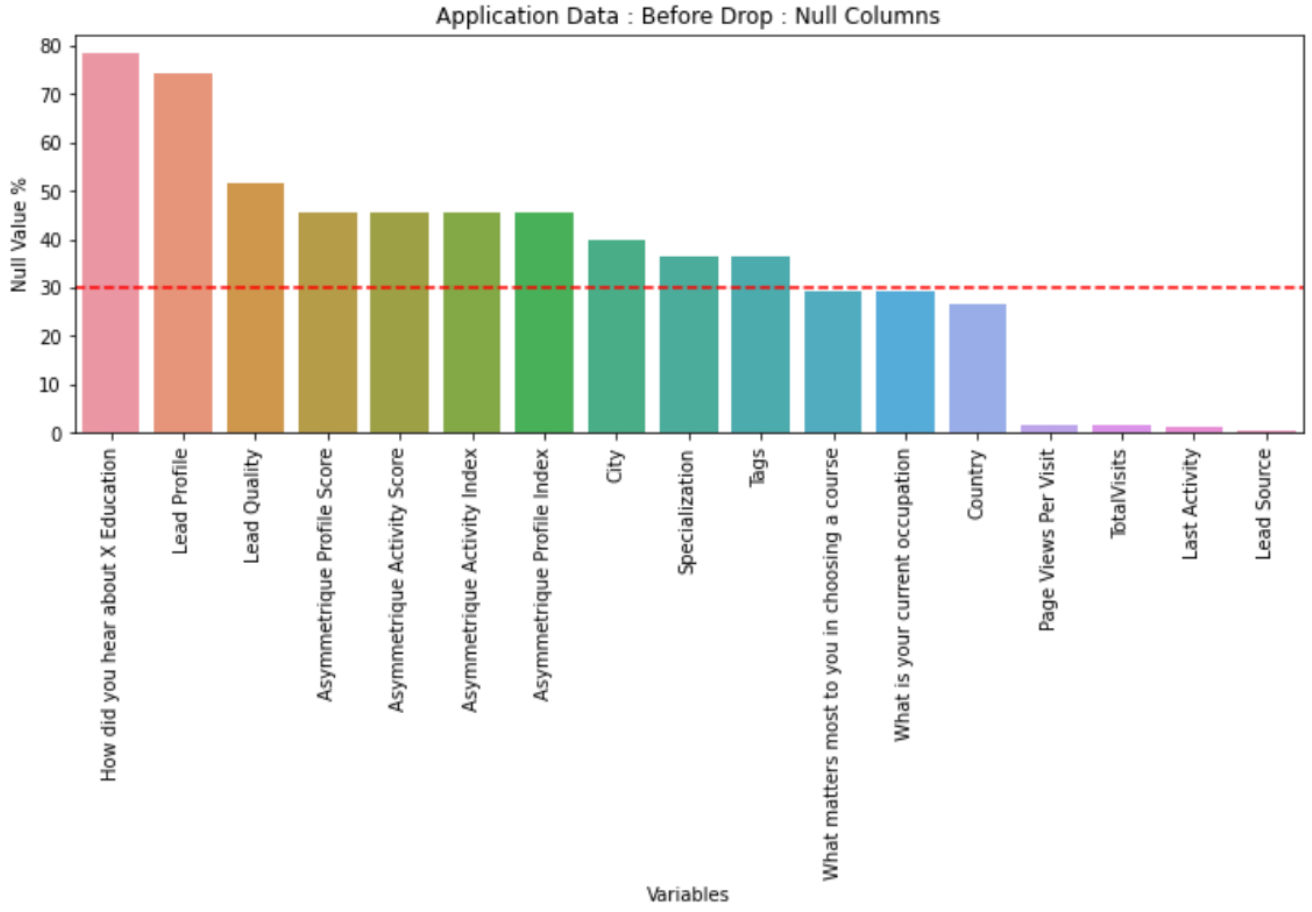
Exploratory Data Analysis(EDA)

Data Imbalance



We can observe that approximately only 38.5% people gets converted

Data Cleaning



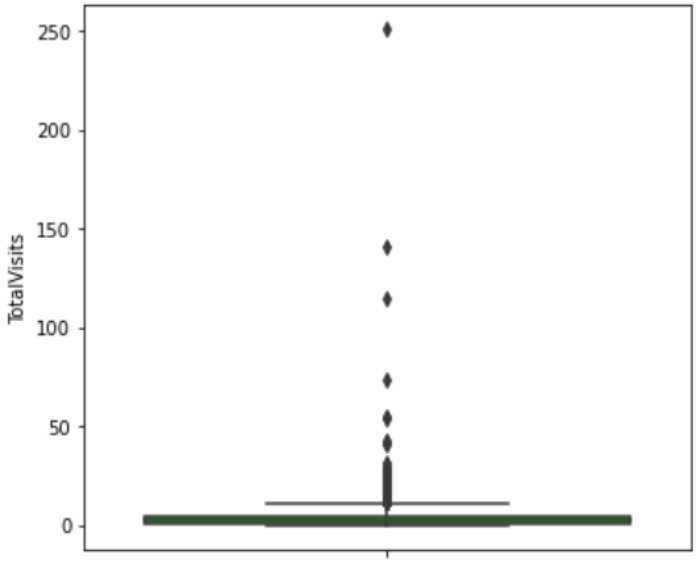
Data cleaning process starts with dropping columns, rows who has large amount of NULL/NAN values. After than impute data to replace null with mean/median or whatever is feasible according to Numerical/Categorical columns.

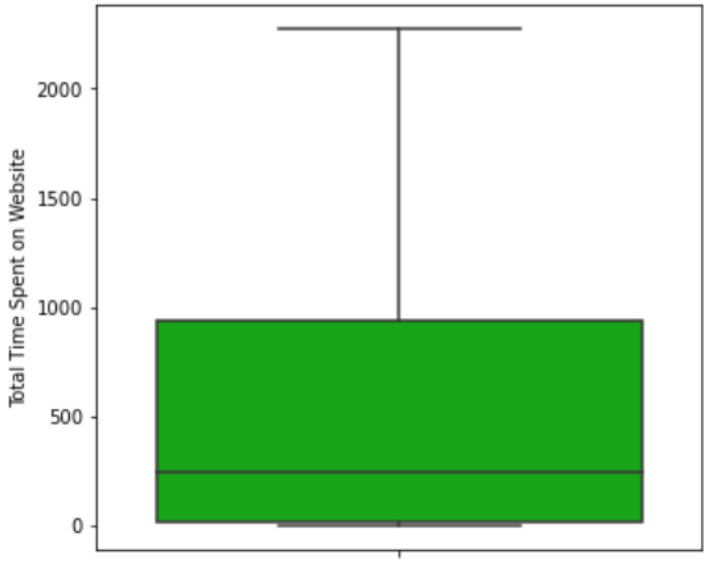
To drop null columns,30% as the threshold value is used. We can observe that there are many columns which contains NULL values more than threshold.

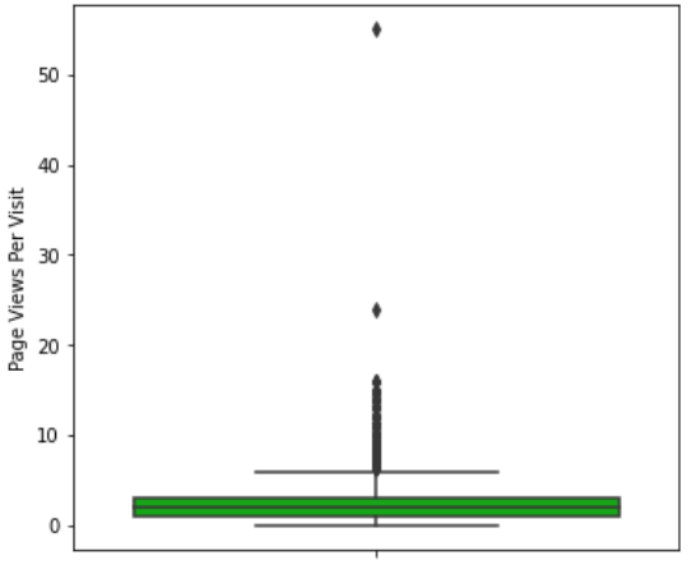
Impute Data

- Variables which doesn't add much value and can be dropped
- Categorical variables which cannot be dropped can be imputed with the mode
- Numerical variables which cannot be dropped can be imputed with the median
- Since "India" dominates the Country variable, we can drop the column
- We replace “Select” with np.nan as so many columns has this value indicating user hasn’t selected anything
- Since "Google" has the major chunk of data, we can impute the null values in “Lead Source ” with Google
- Since we do not have any information of what the last activity of the customer would have been, we can add a new category called 'Not Sure' for the null values in “Last Activity”
- Since no information has been provided ”Current Occupation”, we can add a new category called No Information and set as value for the null columns
- 'A free copy of Mastering The Interview', 'Through Recommendations', 'Digital Advertisement', 'Search','Do Not Call','Do Not Email‘ are binary value columns, so we replace “no” with 0 and “yes” with 1.
- We drop few columns which has same data in all the rows like 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'

Outlier Treatment

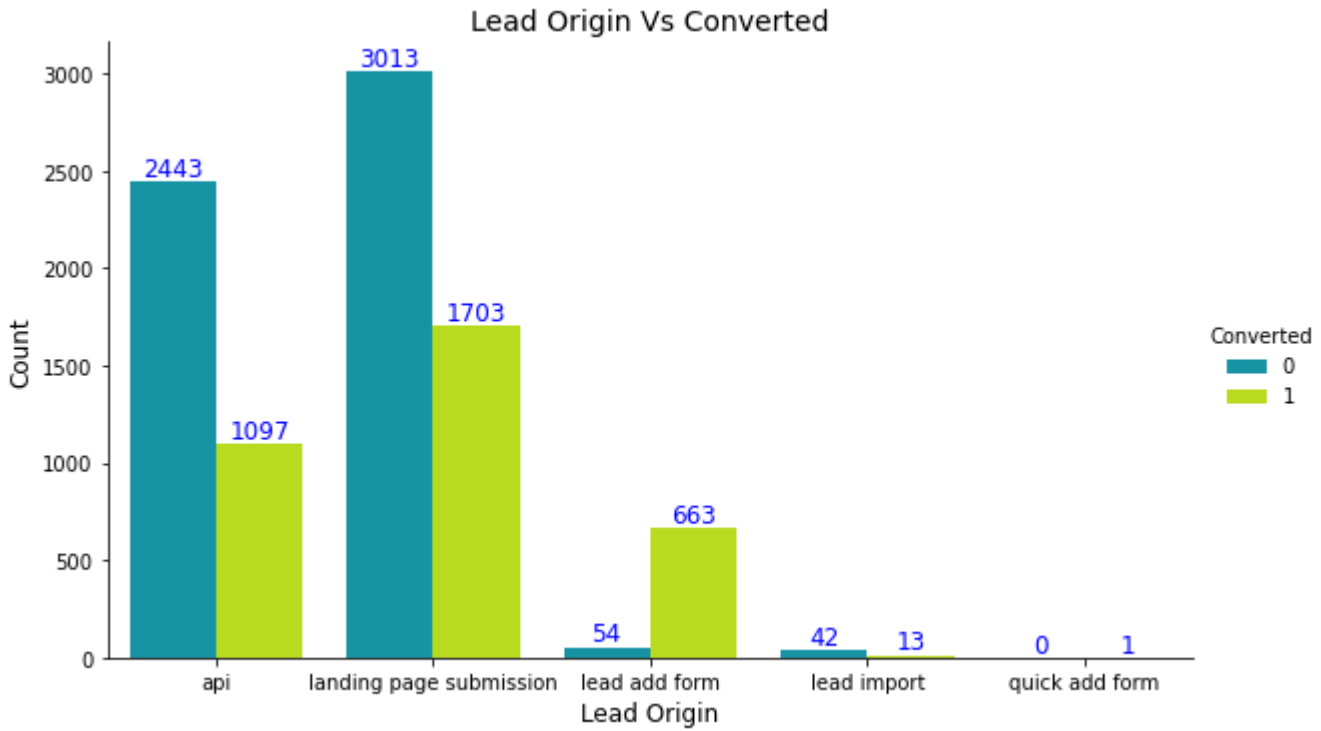






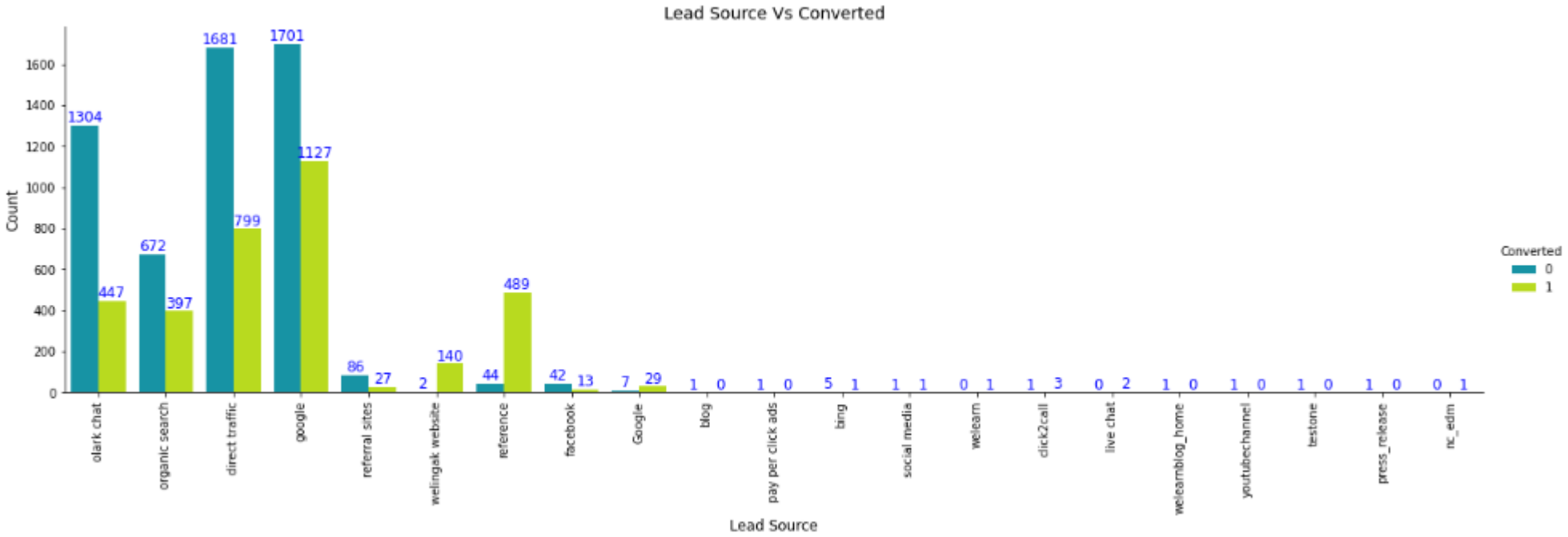
- We can see there are no outliers in Total time spent on website
- There are few outliers for other two features

Univariate Analysis



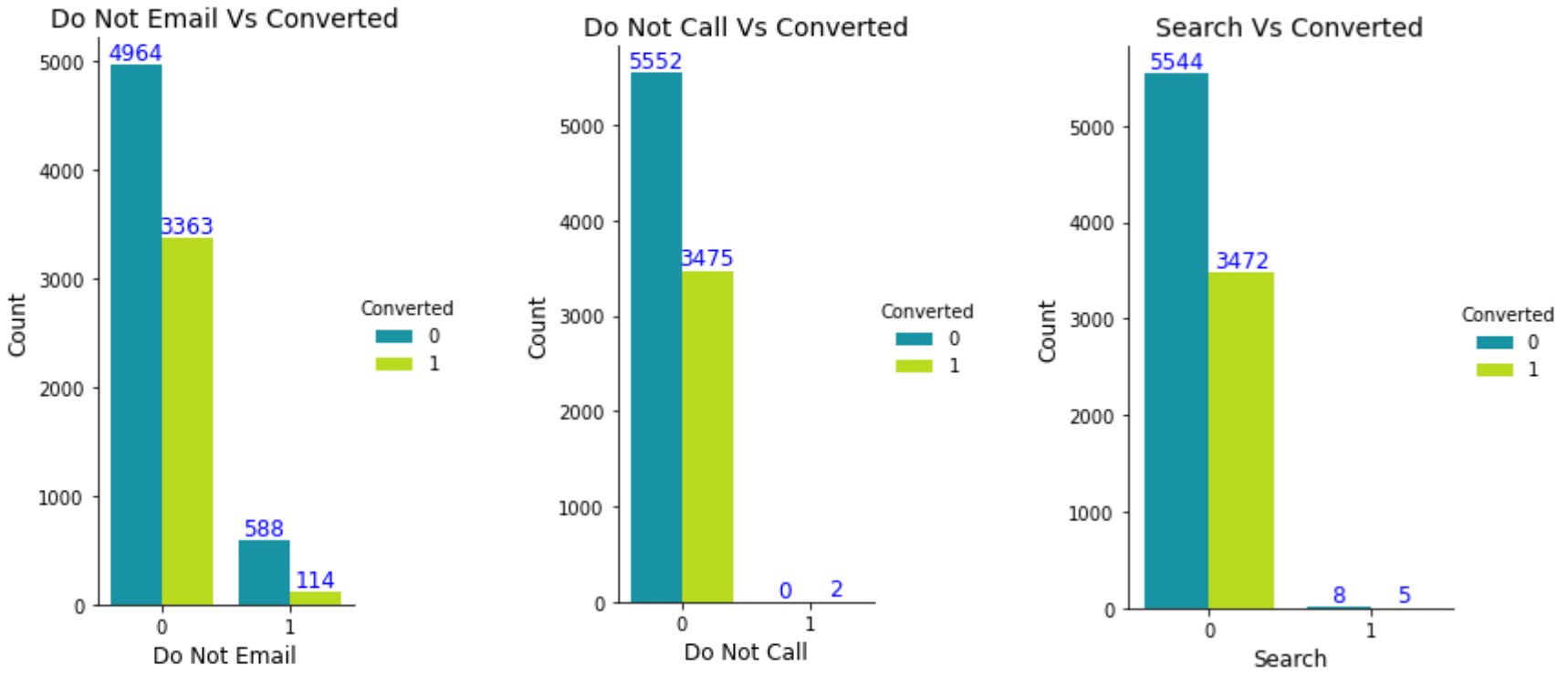
- Majority of leads come from API and Landing Page Submission, but they have conversion rate less than 20%
- Total of leads from "lead add form" are less but the conversion rate is high

Univariate Analysis



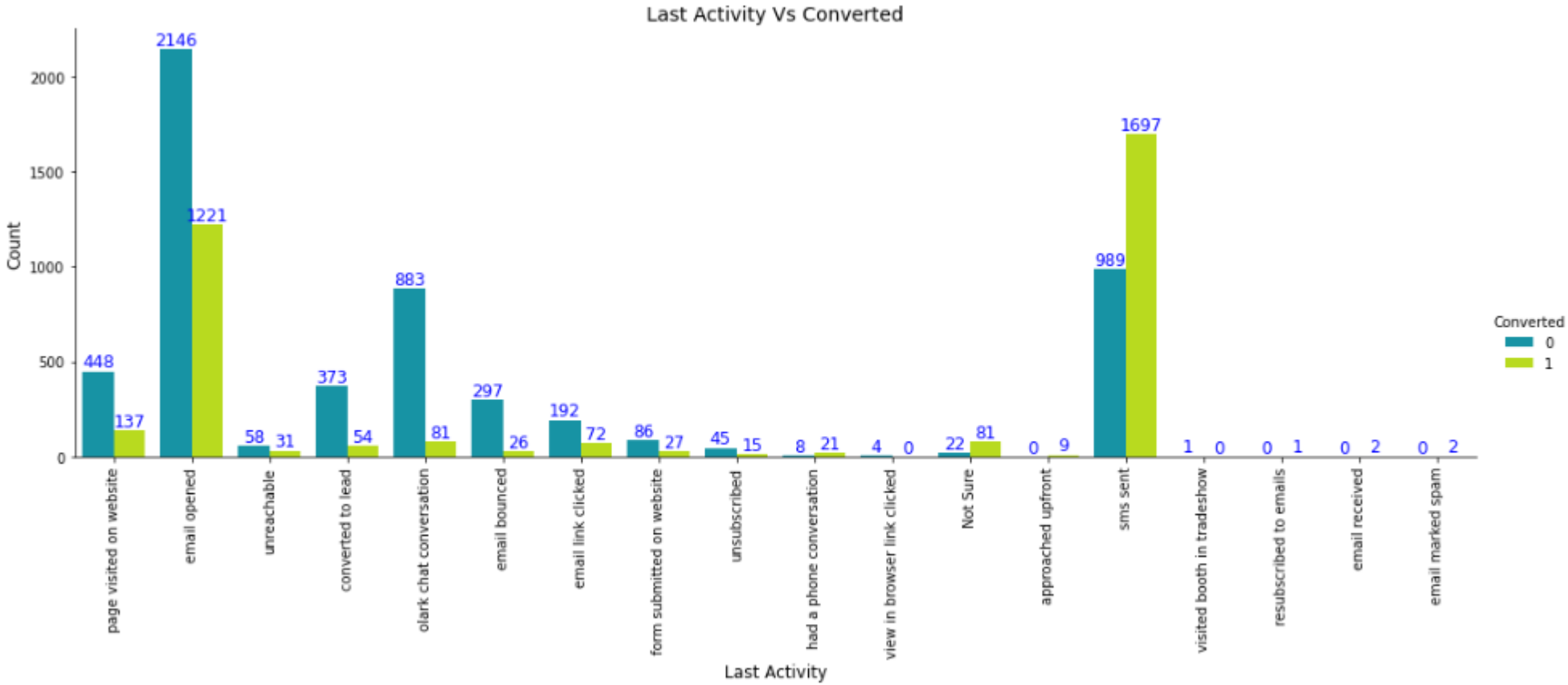
- Leads from Direct Traffic and Google have almost equal amount but the conversion rate of Google is high
- Conversion rate from reference and welingak website is maximum

Univariate Analysis



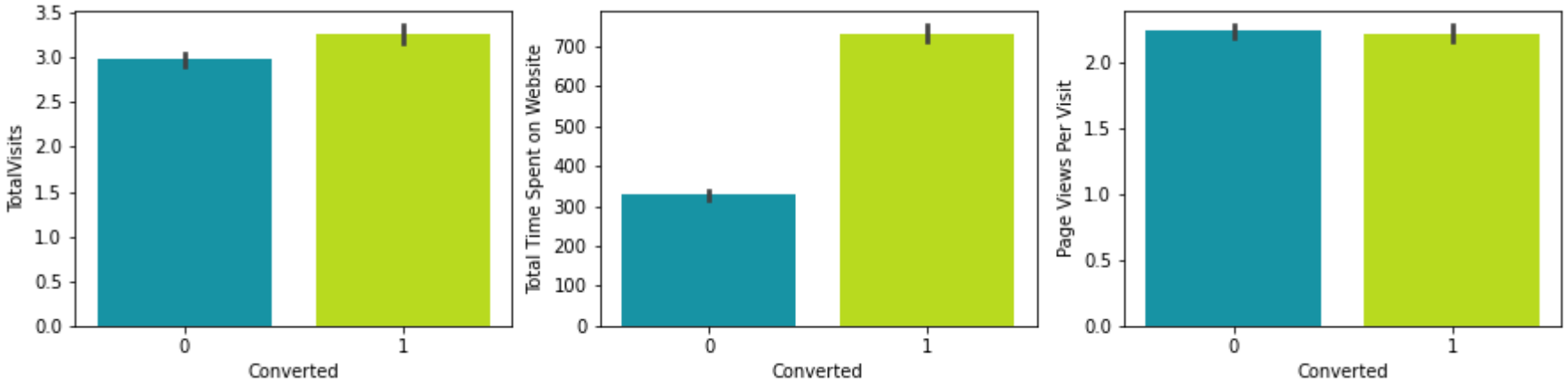
- This shows majority of the conversion were made when emails were sent
- Even though few customers has selected whether or not they want to be emailed about the course or not, small amount have been benefited and converted by emails.
- This shows that majority were converted when call were made
- 2 customers who selected 'Do Not Call' also got converted
- This shows customers who has searched has been converted.

Univariate Analysis



'Last Activity' with SMS Sent have the highest conversion rate followed by Email Opened

Univariate Analysis



This shows that the continuous variables has good conversion rate

Model Building

- Dummy variables were created for 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization' etc
- Divide data set in to Train and Test data using train_test_split from sklearn
- StandardScaler was performed on X_train and X_test for futher process
- Initial training model is built upon train data to see co-relation
- Features are selected using RFE (Recursive feature elimination), total 6 models were generated
- Feature elimination was done by criteria of p-value < 0.05 and VIF <= 5

Model Prediction

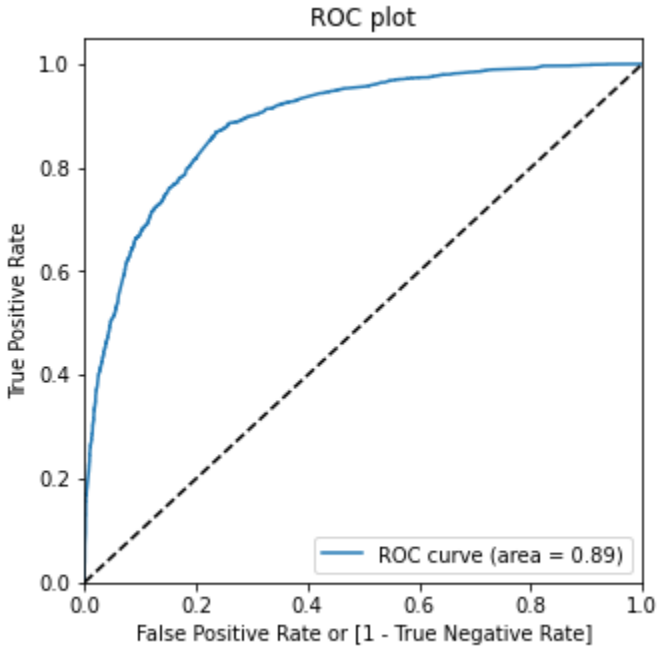
- Define the arbitrary cutoff to 0.5
  - Since the logistic curve gives us the probabilities and not the actual classification of 'Converted' and 'Not-Converted', we need to find a threshold probability to classify leads
  - Let's choose 0.5 as arbitrary cutoff where if probability of the lead to be 'Converted' is less than 0.5, we would classify them as 'Not-Converted' and if greater than 0.5, then 'Converted‘
- Check confusion matrix

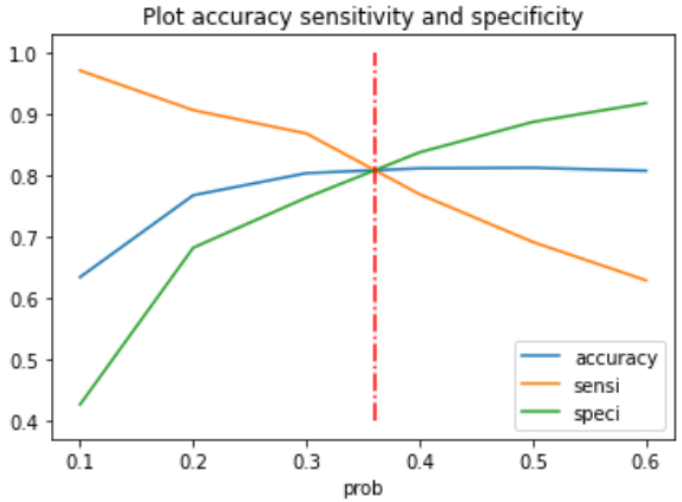
Actual	Predicted	
	No(Not-Converted)	Yes(Converted)
No(Not-Converted)	TN	FP
Yes(Converted)	FN	TP

- Accuracy Score turns out to be approximately 81% which is good and close to the expected score

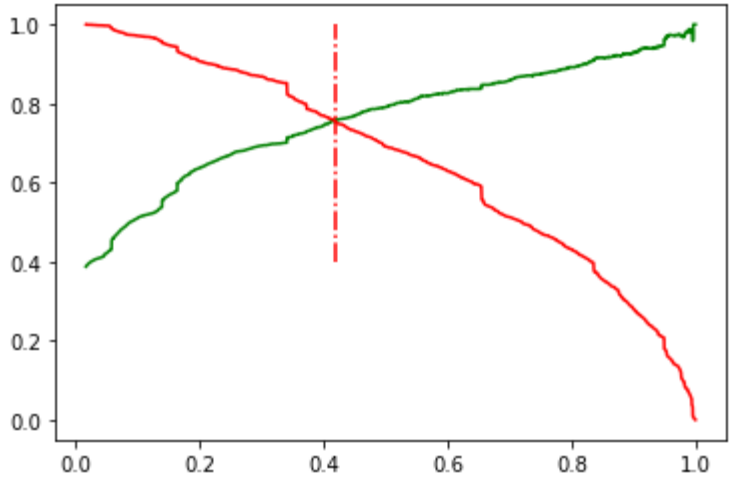
Model Evaluation

- We plotted the ROC(Receiver Operating Characteristic) curve to check our prediction
- Calculate accuracy sensitivity and specificity for various probability cutoffs an plot it in graph
- We plotted precision-recall
- Then we applied prediction on test data set to compare with train data set





Accuracy Sensitivity and Specificity cutoff 0.36



Precision recall cutoff 0.42

Conclusions

- 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.
- Accuracy, Sensitivity, Specificity and Precision values of train set are 80.81%, 80.55%, 80.97% and 72.30% respectively
- Accuracy, Sensitivity, Specificity and Precision values of test set are 81.36%, 82.08%, 80.89% and 73.59% respectively
- The values of Accuracy, Sensitivity and Specificity for both train and test are closer and hence we can conclude that this is a good model.

Thank You.