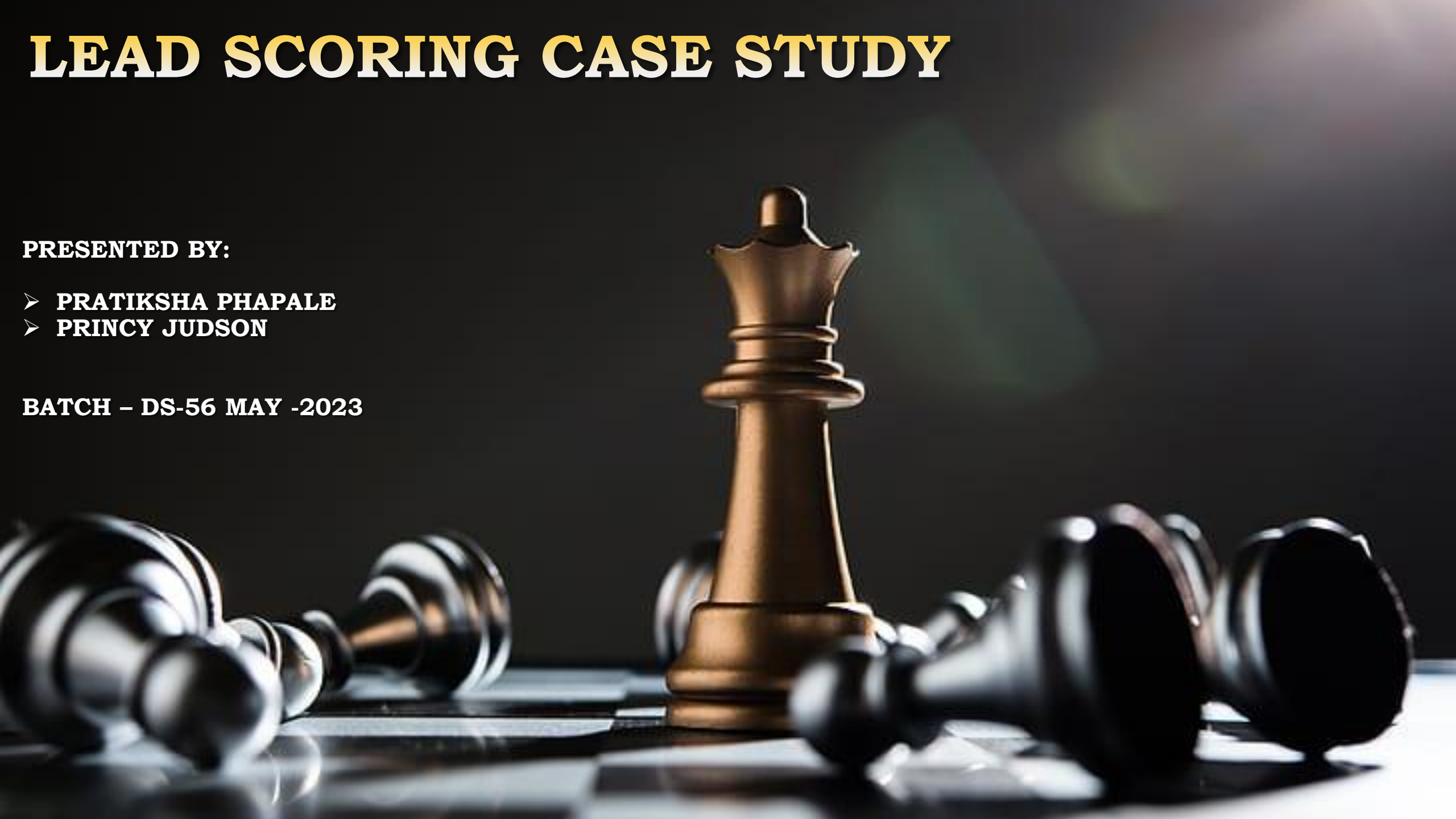
LEAD SCORING CASE STUDY

PRESENTED BY:

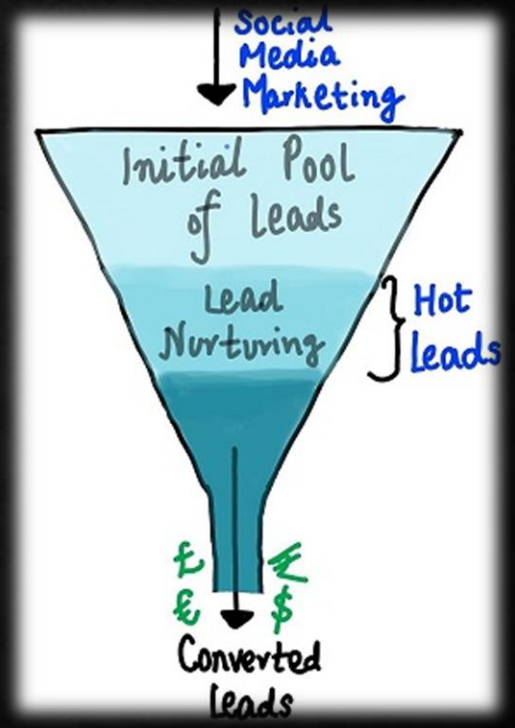
- **PRATIKSHA PHAPALE**
- **PRINCY JUDSON**

BATCH – DS-56 MAY -2023



CASE STUDY DESCRIPTION

- 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.
- X Education, an online course provider, experiences a low lead conversion rate despite a substantial influx of professionals to their website.
- The company seeks to enhance efficiency by identifying 'Hot Leads'—individuals with a higher likelihood of conversion.
- Currently achieving a 30% conversion rate, the company aims to optimize its lead conversion process, focusing efforts on potential leads to increase overall conversion rates.
- This strategic approach involves targeting individuals who express interest by filling forms or engaging with course content, ultimately refining the sales team's outreach efforts.



PROBLEM STATEMENT

An Education company named X Education sells online courses to industry Professionals

Now although X education gets a lot of leads , its leads conversion rate is very poor of about 30%

The company wants to increase it to 80%

GOAL

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

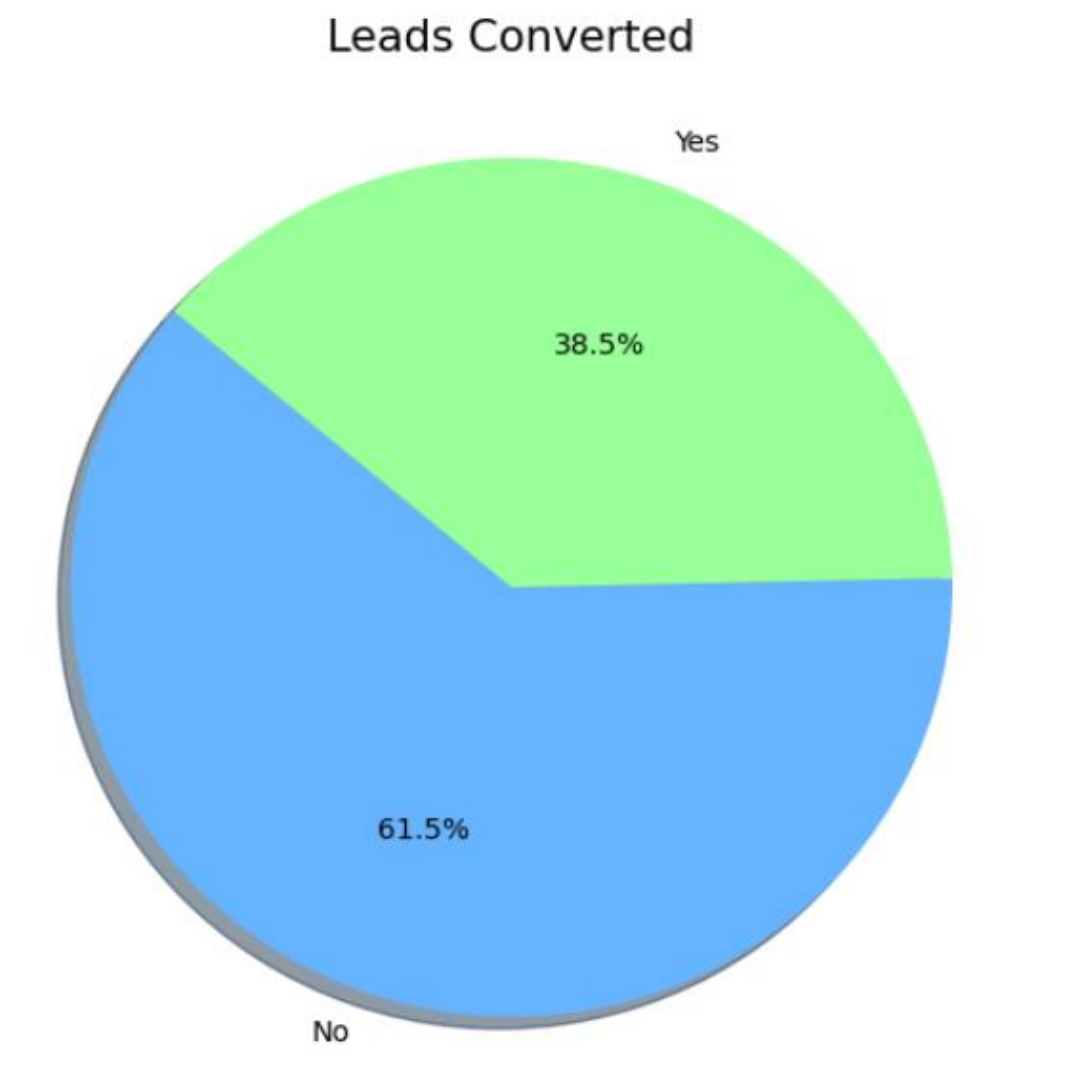
APPROACH

- To improve the lead conversion rate to be around 80%, Logistic Regression model is created to identify the important variables and derive insight on how to improve the lead conversion count
- Below Steps are performed in the case study for the outcome:
  - Reading Understanding Data
  - EDA and Visualizing the Data
  - Data Cleaning and Preparation
  - Preparing the data for modelling (train-test-split , rescaling , accuracy , precision , recall etc.)
  - Training the model
  - Predictions and evaluation on the test set
  - Model Evaluation

ASSUMPTIONS

- Dropped the columns where missing values percentage is greater than 40 %, Lead quality had 51% nan values, replaced with a new category, since this feature was relevant to the problem statement.
- Category columns, null values has been replaced with a new Category to segregate the data.
- Dropped few unnecessary columns where data was heavily skewed to not impact the overall model building. Eg – Search Column

EXPLORATORY DATA ANALYSIS

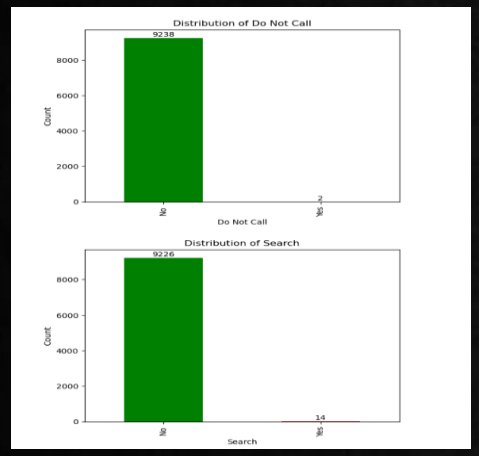


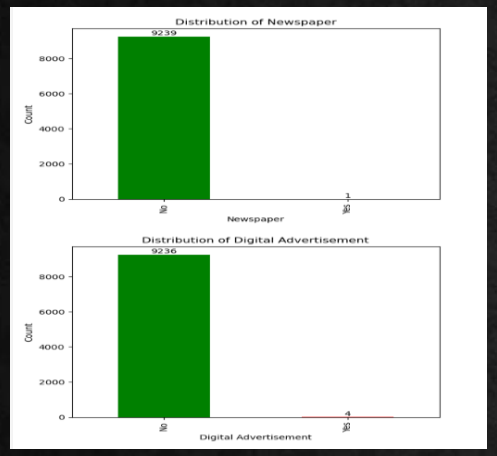
Insights:

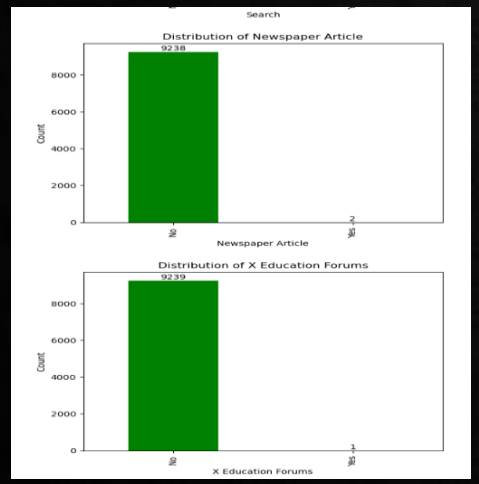
The majority, comprising 61.5% of individuals, did not convert to leads, signifying a substantial portion. In contrast, only 38.5% of the people successfully converted, representing a minority in the dataset.

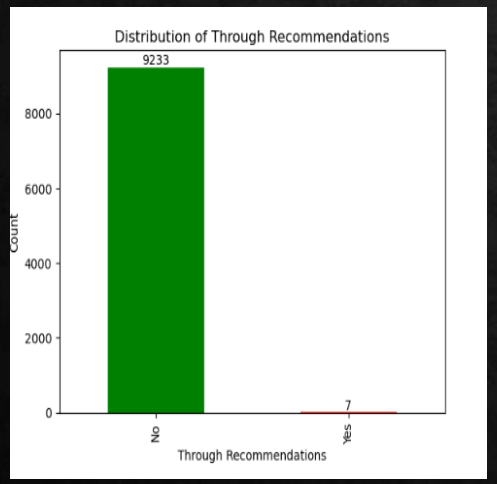
EDA- UNIVERIATE ANALYSIS

Categorical Columns





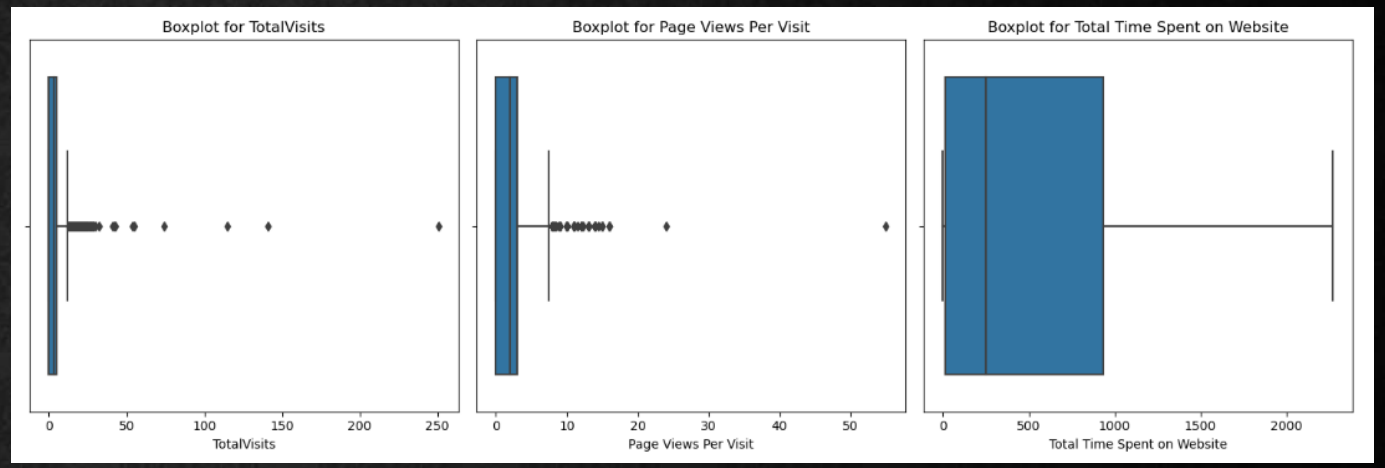


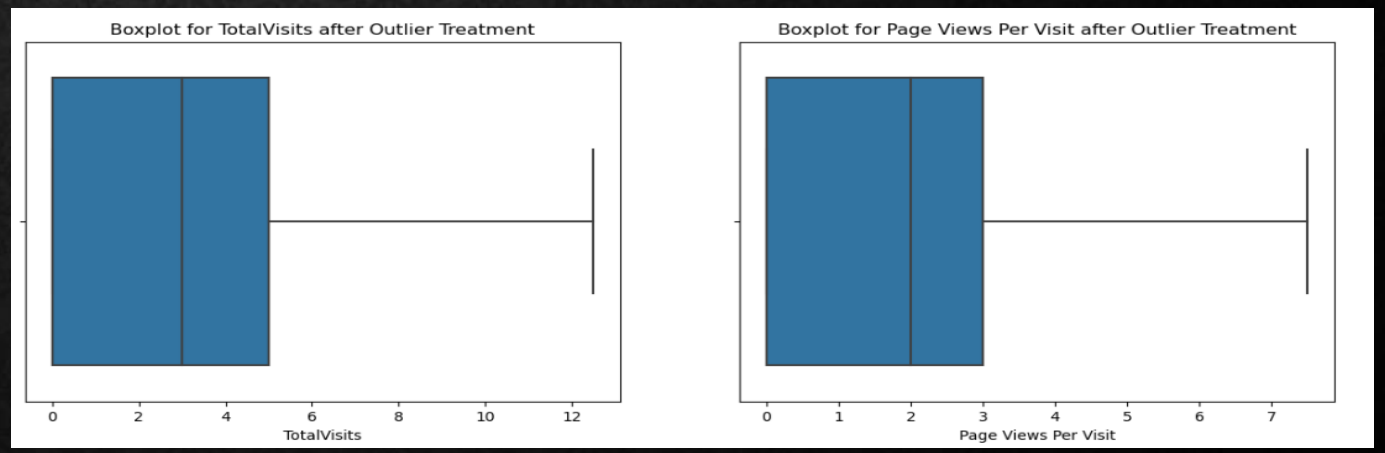


Insight:

All these flag columns have 100% no, hence, these columns were dropped

Numerical Columns



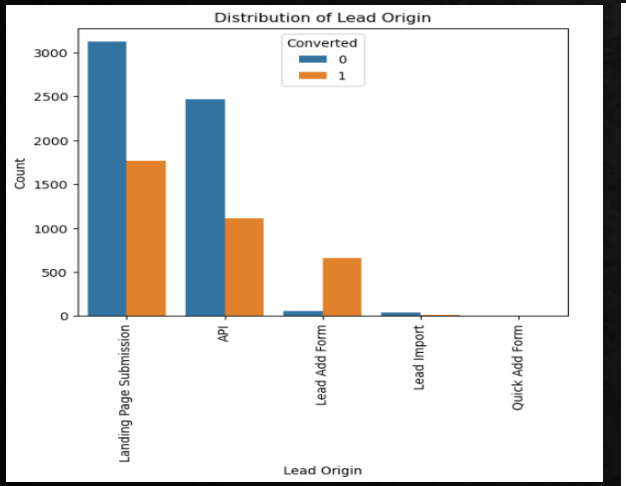


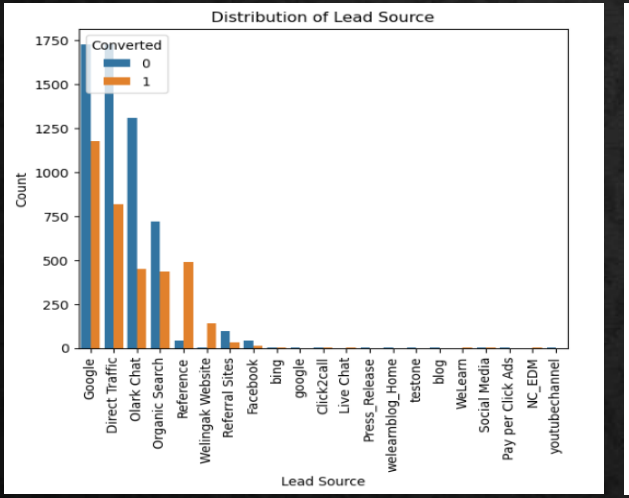
Insight:

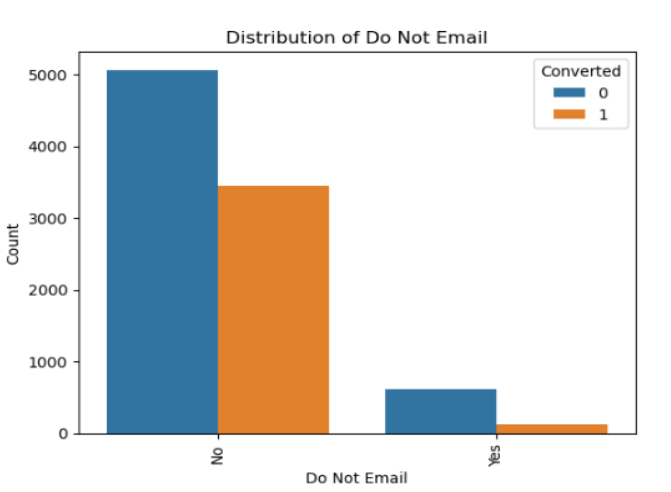
- The columns TotalVisit , Page Views Per Visit contain outliers as we seen in the boxplot.

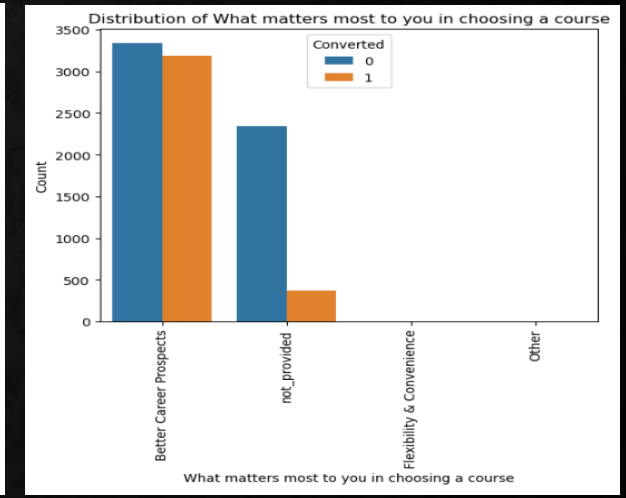
EDA- BIVERIATE ANALYSIS

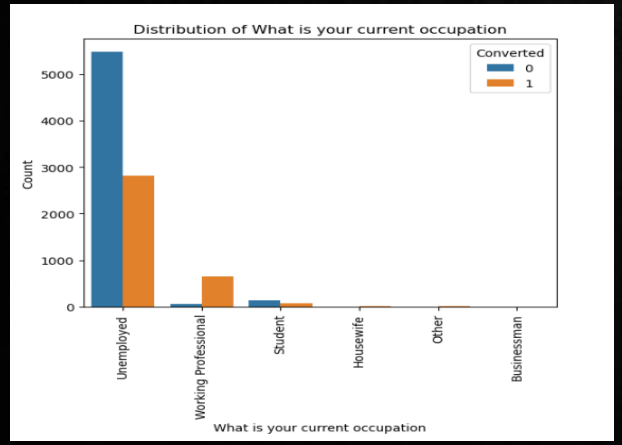
Distribution of categorical columns with the "Converted" column

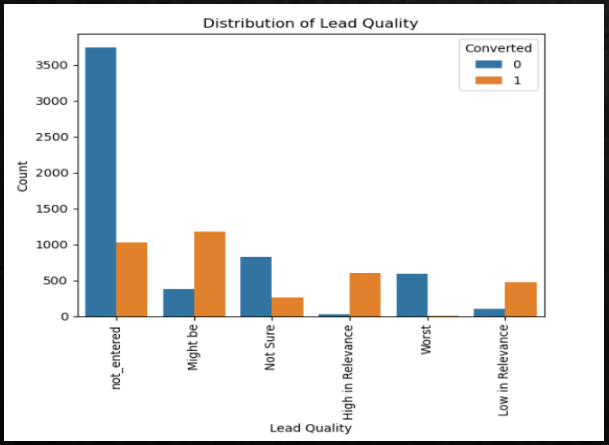


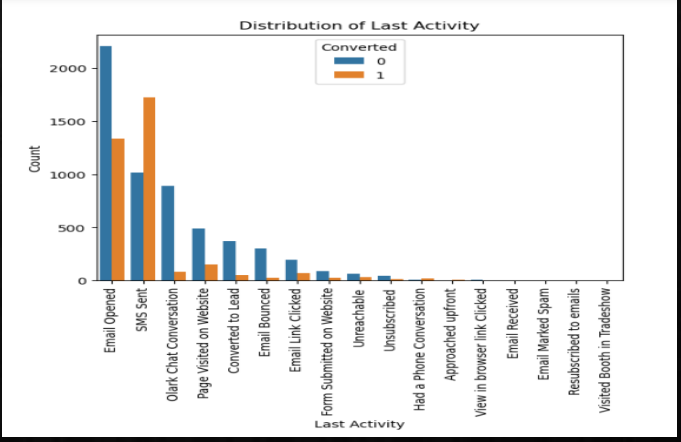


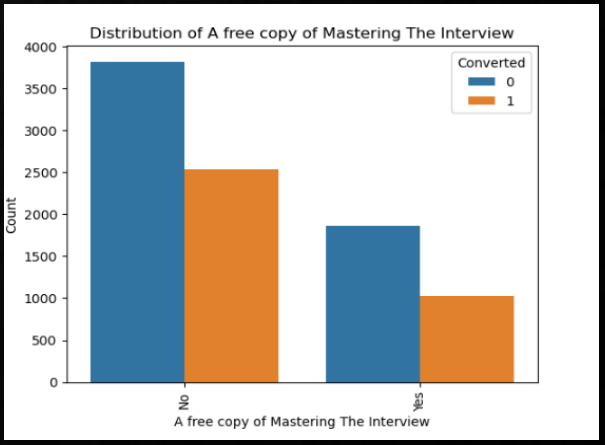






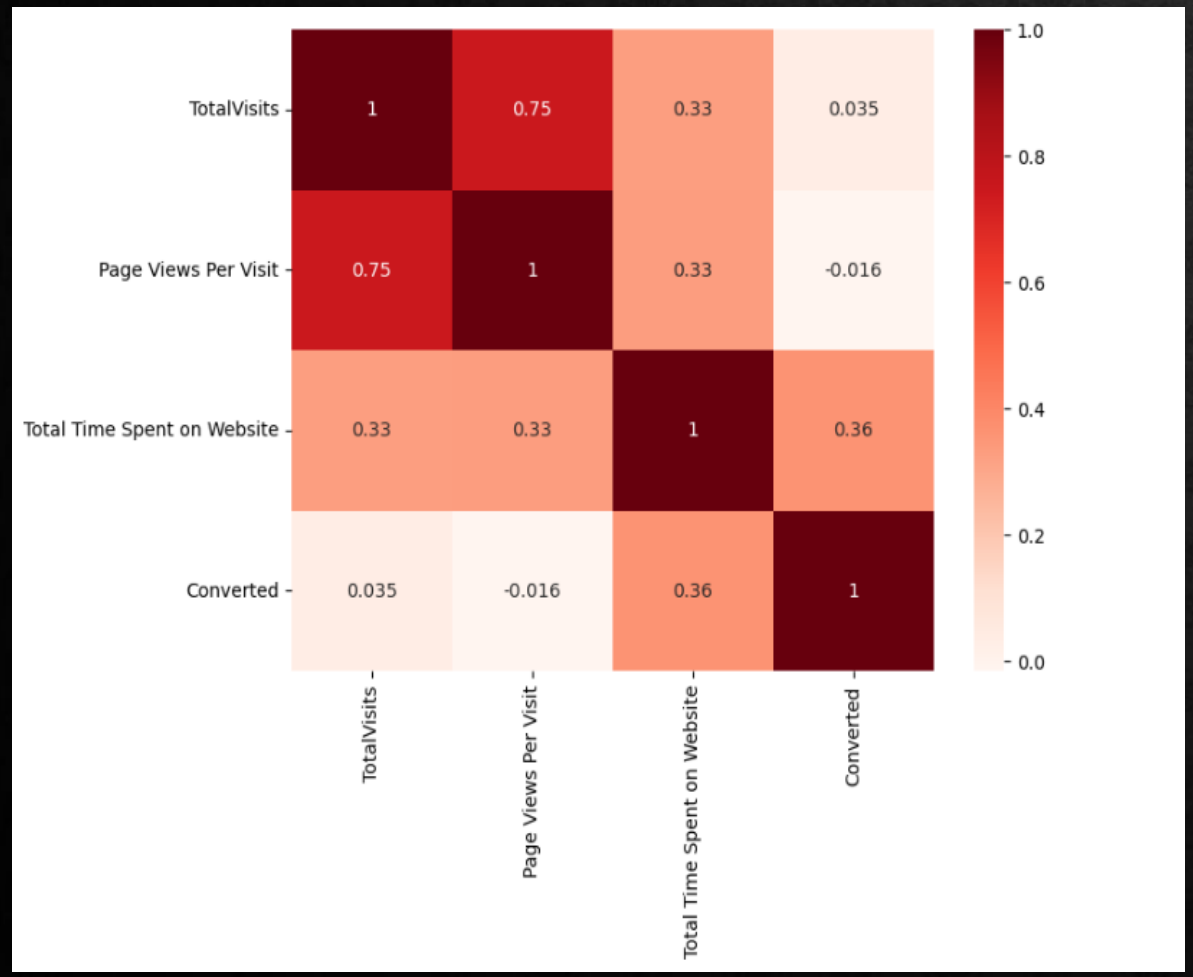


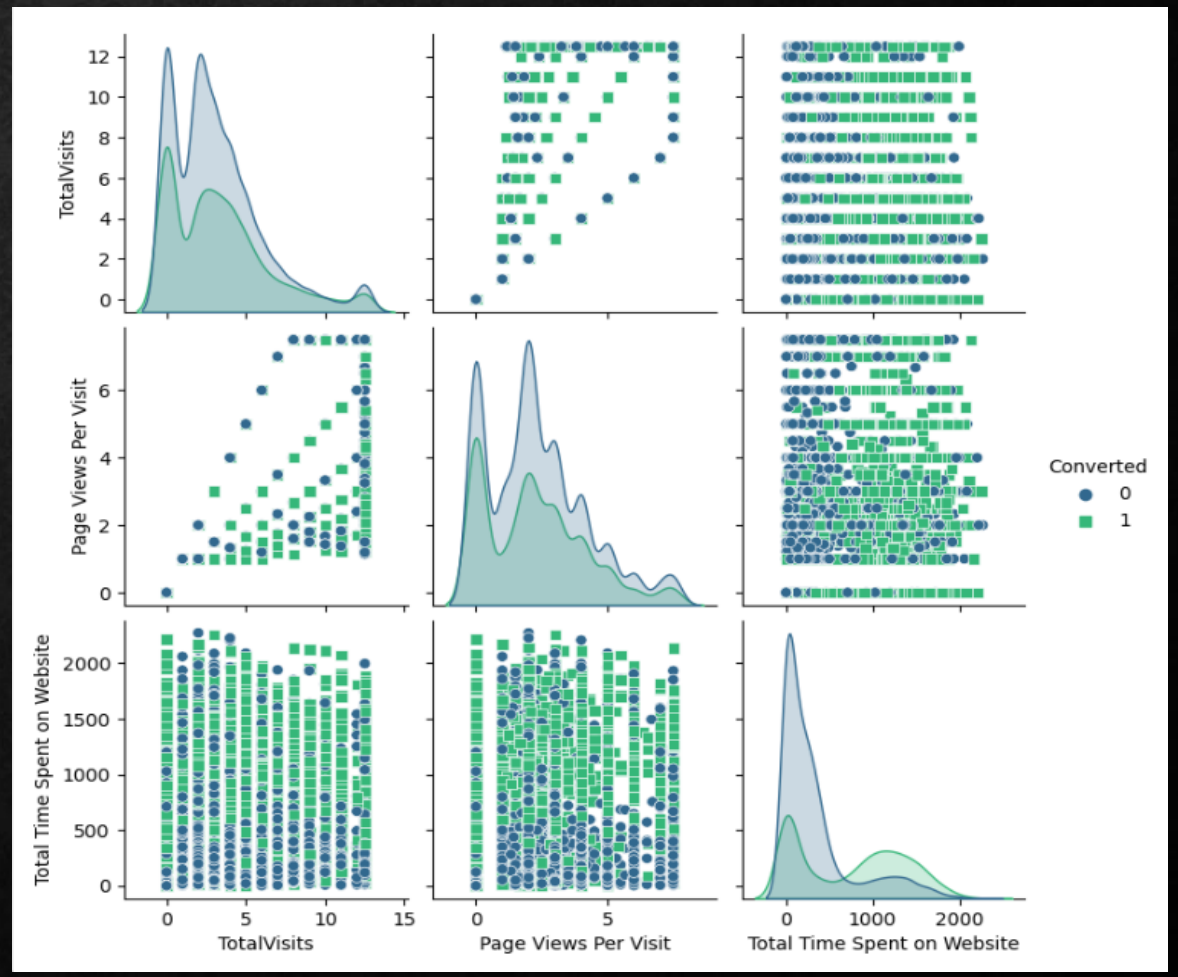




EDA- MULTIVARIATE ANALYSIS

Correlation metrics of the Numerical Columns





DATA PREPARATION

- Columns has more skewed to a particular category is dropped. Example Country and City
- Columns with nulls closer to 40% and imputation is not straight forward, like tags column is dropped.
- Outliers in numeric data were treated using the Interquartile Range (IQR) method to identify and cap/floor values beyond a certain threshold.
- Page views column is imputed with mode values.
- Lead_quality and specialization categorical column a new category was created for null values, since that provided more information.

PRE MODELLING STEPS

- Categorical features are converted to dummy variables
- The data was split into train 70% and test 30%
- Numerical columns where standardized using MinMax scaler method.
- Logistic model is built using all 74 features, accuracy was 77% and features had high p values and high VIF
- RFE is used for feature selection, top 15 features are selected

MODEL 1

- Used RFE to select only top 15 features, built a logistic regression model using. VIF was high for 'PageViewsPerVisit'
- Train Accuracy – 85%

```
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6452
Model Family:                Binomial   Df Model:                           15
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2244.7
Date:                Tue, 21 Nov 2023   Deviance:                       4489.3
Time:                        00:18:27   Pearson chi2:                 6.73e+03
No. Iterations:                     7   Covariance Type:             nonrobust
==========================================================================================================
                                                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                            0.0967      0.162      0.596      0.551      -0.221       0.415
TotalVisits                                      1.3547      0.231      5.876      0.000       0.903       1.806
TotalTimeSpentonWebsite                          4.3419      0.181     23.976      0.000       3.987       4.697
PageViewsPerVisit                               -1.1990      0.247     -4.857      0.000      -1.683      -0.715
DoNotEmail                                      -1.1191      0.180     -6.203      0.000      -1.473      -0.766
LastNotableActivity_Had a Phone Conversation     2.5322      1.242      2.038      0.042       0.097       4.967
LastNotableActivity_SMS Sent                     1.7534      0.089     19.701      0.000       1.579       1.928
LastNotableActivity_Unreachable                  1.6265      0.617      2.635      0.008       0.417       2.836
Whatisyourcurrentoccupation_Working Professional 1.7423      0.213      8.195      0.000       1.326       2.159
LeadSource_olark chat                            1.1101      0.134      8.265      0.000       0.847       1.373
LeadSource_welingak website                      3.5661      0.752      4.741      0.000       2.092       5.040
LeadOrigin_Lead Add Form                         2.5707      0.227     11.345      0.000       2.127       3.015
LeadQuality_Might be                            -1.5060      0.154     -9.806      0.000      -1.807      -1.205
LeadQuality_Not Sure                            -3.4150      0.167    -20.393      0.000      -3.743      -3.087
LeadQuality_Worst                               -5.0570      0.361    -14.018      0.000      -5.764      -4.350
LeadQuality_not_entered                         -3.1318      0.137    -22.866      0.000      -3.400      -2.863
==========================================================================================================
```

Features	VIF
PageViewsPerVisit	5.95
TotalVisits	4.79
LeadQuality_not_entered	2.89
TotalTimeSpentonWebsite	2.00
LeadQuality_Might be	1.94
LeadSource_olark chat	1.75
LastNotableActivity_SMS Sent	1.59
LeadQuality_Not Sure	1.47
LeadOrigin_Lead Add Form	1.43
Whatisyourcurrentoccupation_Working Professional	1.30
LeadSource_welingak website	1.27
LeadQuality_Worst	1.20
DoNotEmail	1.11
LastNotableActivity_Had a Phone Conversation	1.01
LastNotableActivity_Unreachable	1.01

MODEL 2

- Removed ‘PageViewsPerVisit’ as the VIF was above 5, all features have VIF less than 5.
- Accuracy -85% not affected by removing ‘PageViewsPerVisit

```
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6453
Model Family:                Binomial   Df Model:                           14
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2256.6
Date:                Tue, 21 Nov 2023   Deviance:                       4513.3
Time:                        00:18:28   Pearson chi2:                 6.79e+03
No. Iterations:                     7   Covariance Type:             nonrobust
==================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                            -0.1735      0.152     -1.142      0.254      -0.471       0.124
TotalVisits                                       0.7281      0.192      3.795      0.000       0.352       1.104
TotalTimeSpentonWebsite                           4.3252      0.181     23.951      0.000       3.971       4.679
DoNotEmail                                       -1.1105      0.180     -6.186      0.000      -1.462      -0.759
LastNotableActivity_Had a Phone Conversation      2.5145      1.247      2.016      0.044       0.070       4.959
LastNotableActivity_SMS Sent                      1.7117      0.088     19.427      0.000       1.539       1.884
LastNotableActivity_Unreachable                   1.6107      0.619      2.602      0.009       0.397       2.824
Whatisyourcurrentoccupation_Working Professional  1.7332      0.213      8.133      0.000       1.316       2.151
LeadSource_olark chat                             1.3597      0.125     10.878      0.000       1.115       1.605
LeadSource_welingak website                       3.5469      0.752      4.718      0.000       2.073       5.021
LeadOrigin_Lead Add Form                          2.8367      0.220     12.909      0.000       2.406       3.267
LeadQuality_Might be                             -1.5053      0.153     -9.827      0.000      -1.806      -1.205
LeadQuality_Not Sure                             -3.4150      0.167    -20.424      0.000      -3.743      -3.087
LeadQuality_Worst                                -5.0252      0.361    -13.928      0.000      -5.732      -4.318
LeadQuality_not_entered                          -3.1031      0.136    -22.771      0.000      -3.370      -2.836
==================================================================================================
```

Features	VIF
TotalVisits	2.64
LeadQuality_not_entered	2.55
TotalTimeSpentonWebsite	1.96
LeadQuality_Might be	1.87
LeadSource_olark chat	1.62
LastNotableActivity_SMS Sent	1.57
LeadOrigin_Lead Add Form	1.41
LeadQuality_Not Sure	1.38
Whatisyourcurrentoccupation_Working Professional	1.30
LeadSource_welingak website	1.26
LeadQuality_Worst	1.16
DoNotEmail	1.11
LastNotableActivity_Had a Phone Conversation	1.01
LastNotableActivity_Unreachable	1.00

MODEL 3- FINAL MODEL

- Removed 'LastNotableActivity_Had a Phone Conversationt' to reduce p-value , all features have VIF less than 5 and low p-value
- Accuracy -85% not affected

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6454
Model Family:                Binomial   Df Model:                           13
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2259.5
Date:                Tue, 21 Nov 2023   Deviance:                       4519.1
Time:                        00:18:28   Pearson chi2:                 6.79e+03
No. Iterations:                     7   Covariance Type:             nonrobust
==================================================================================================================
                                                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
const                                             -0.1553      0.152     -1.022      0.307      -0.453       0.142
TotalVisits                                        0.7332      0.191      3.834      0.000       0.358       1.108
TotalTimeSpentonWebsite                            4.3157      0.180     23.937      0.000       3.962       4.669
DoNotEmail                                        -1.1144      0.180     -6.207      0.000      -1.466      -0.762
LastNotableActivity_SMS Sent                       1.7060      0.088     19.375      0.000       1.533       1.879
LastNotableActivity_Unreachable                    1.6025      0.620      2.586      0.010       0.388       2.817
Whatisyourcurrentoccupation_Working Professional   1.7324      0.213      8.134      0.000       1.315       2.150
LeadSource_olark chat                              1.3543      0.125     10.842      0.000       1.109       1.599
LeadSource_welingak website                        3.5520      0.752      4.724      0.000       2.078       5.026
LeadOrigin_Lead Add Form                           2.8279      0.220     12.865      0.000       2.397       3.259
LeadQuality_Might be                              -1.5096      0.153     -9.865      0.000      -1.810      -1.210
LeadQuality_Not Sure                              -3.4281      0.167    -20.518      0.000      -3.756      -3.101
LeadQuality_Worst                                 -5.0395      0.361    -13.972      0.000      -5.746      -4.333
LeadQuality_not_entered                           -3.1159      0.136    -22.881      0.000      -3.383      -2.849
==================================================================================================================
```

Features	VIF
TotalVisits	2.63
LeadQuality_not_entered	2.54
TotalTimeSpentonWebsite	1.96
LeadQuality_Might be	1.87
LeadSource_olark chat	1.62
LastNotableActivity_SMS Sent	1.57
LeadOrigin_Lead Add Form	1.41
LeadQuality_Not Sure	1.38
Whatisyourcurrentoccupation_Working Professional	1.30
LeadSource_welingak website	1.26
LeadQuality_Worst	1.16
DoNotEmail	1.11
LastNotableActivity_Unreachable	1.00

METRICS OF FINAL MODEL

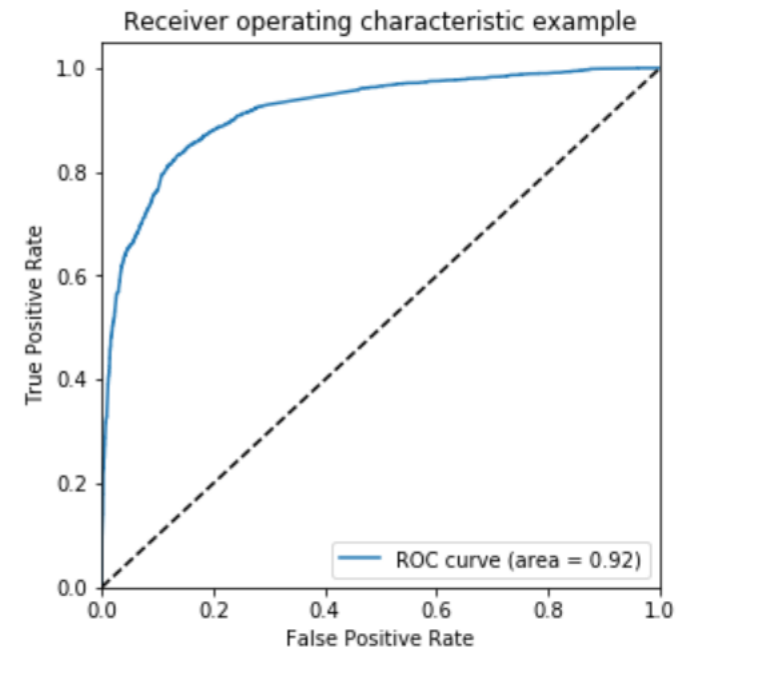
These train metrics are calculated with default cutoff 0.5

- Train accuracy =85%
- confusion matrix for training data

  [[TN=3638 FP=364]

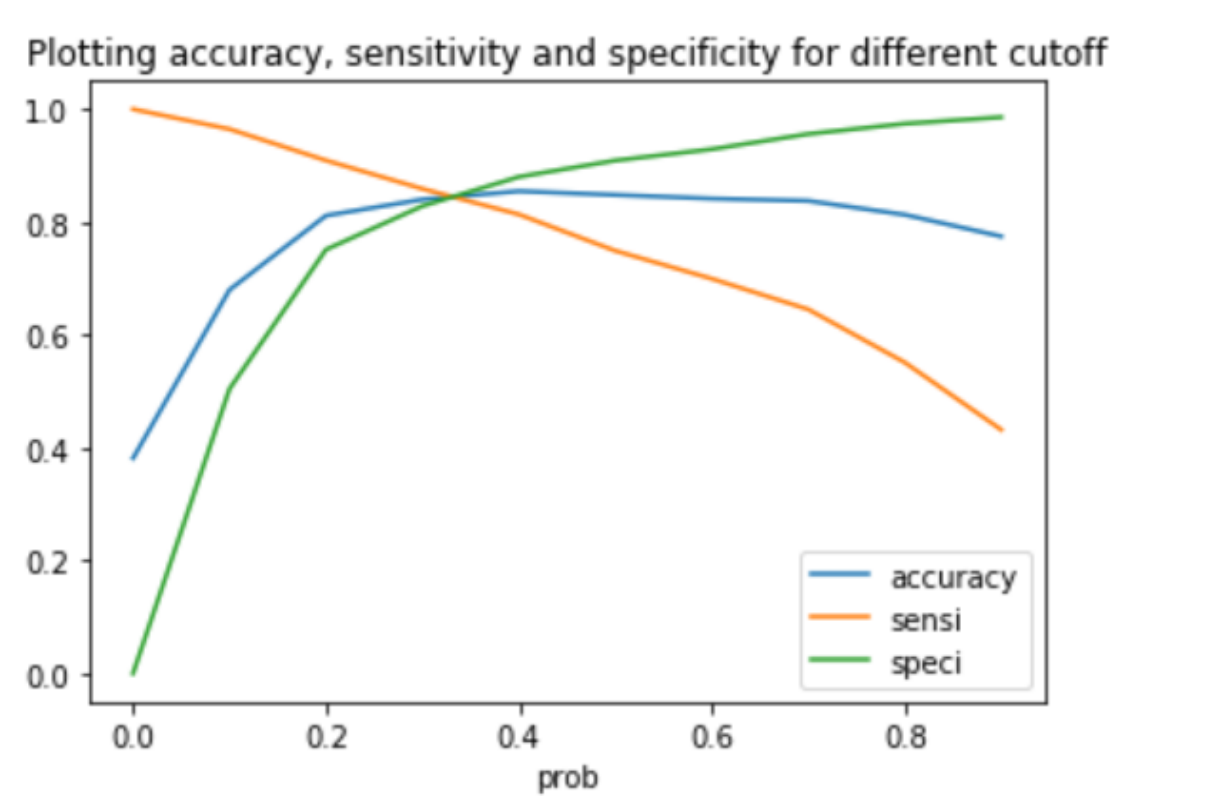
  [FN=619 TP=1847]]

- sensitivity for train = 0.74
- specificity for train = 0.9090454772613693
- AOC = 0.92, which is very good



CALCULATING THE OPTIMAL CUTOFF

- Accuracy, sensitivity and specificity are calculated for different cutoff.
- This is plotted with cutoff on x-axis and metric on y-axis
- The intersection is taken as the optimum cutoff which will give better results for all 3 metrics.
- The cutoff obtained is 0.39



TRAIN AND TEST METRICS WITH OPTIMAL CUTOFF

Train metrics:

- Train Accuracy 85%
- Confusion Matrix

  [[TN= 3496, FP= 506],

  [FN= 445, TP= 2021]]
- Sensitivity 0.81
- Specificity 0.87

Test metrics:

- Train Accuracy 85%
- Confusion Matrix

  [[TN= 1476, FP= 201],

  [FN= 202, TP= 893]]
- Sensitivity 0.81
- Specificity 0.88

SUMMARY

- We have a good train and test **accuracy of 85%**
- Sensitivity and specificity has more than **80%** for both test and train
- The 0's are prediction with high specificity, which will reduce the sales cost by focusing on the hot leads.
- More budget can be done on Welingak Website in terms of advertising, and focus more on olark chat too.
- Employee Intuitions are important if the lead will get converted or not, the lead quality should be entered by employees.
- Working professionals should be targeted as they have high conversion rate and will have better financial situation to pay higher fees.