LEAD SCORING CASE STUDY

PRESENTED BY

Zainab Rahman
Vijaya Patil
Abhishek Madure

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- When person fills form with email ID, Phone number they are classified as lead. Some leads they get from referral as well.
- The typical lead conversion rate at X education is around 30%, means out of acquired leads only 30% gets convert to final and company feels that its lead conversion rate is very poor.
- Expectations from case study:

➢ Identify most potential leads and increase conversion rate

➢ Build model with assignment of lead score to selected leads, higher score means higher chances of conversion.

➢ Target Lead conversion rate to be around 80%.

APPROACH TO SOLUTION

STEP 1: DATA SOURCING, CLEANING AND PREPARATION

1. Handling Null values
2. Outlier treatment
3. Exploratory Data analysis
4. Duplicates removal
5. Dropping Insignificant columns

STEP 2: DATA PREPARATION

1. Dummy variables substation for Categorical columns
2. Feature scaling of Numeric data
3. Dataset split to Test Train

STEP 3: MODEL BUILDING

1. Finding most imp. Features using RFE
2. Build Logistic Regression model

STEP 5: RECOMMENDATIONS

Based on model provide recommendations to Business

STEP 4: MODEL EVALUATION

Calculate measuring metrics like Precision, Accuracy, Recall and check if model predicts 80% conversion rate or not

STEPS FOLLOWED…

STEP 1: DATA SOURCING, CLEANING AND PREPARATION

- Basic Metadata check
  a) .shape function: 9240 rows x 37 columns
  b) .info() function: Information
  c) .describe() function: Descriptive Statistics: To check for probable outliers
- Data wrangling
  a) Identify gaps in data and either fill them or delete them
  b) Deleting data that is either unnecessary or irrelevant
  c) Identify extreme outliers and deal with them
- Duplicate values:
  a) Check for Duplicate values/columns and drop it: Prospect ID, Lead number

STEPS FOLLOWED…

- Check for Null values and Handle the same
  a) Finding null values using .isnull().sum() function
  b) ‘Select’ replaced by NaN values
  c) Dropped columns with Null values ≥39%
  d) Imputation of Null values:
     - ✓ Categorical columns: Replaced null values by missing
     - ✓ Numeric columns: Imputation with median value e.g Total visits column
- Outlier treatment: By use of descriptive statistics analysis. Columns treated are Total visits and Total views per visit
- Column dropping criteria:
  a) Columns which has only one category are dropped as inferences cannot be done
  b) Removed Country column as 60% data is related to ‘India’ and rest data is missing.

STEPS FOLLOWED…

c) Checked for Data Imbalance and removed columns if that is insignificant. Dropped columns ‘Do not Email’, ‘ Digital Advertisement’, ‘Through Recommendations’ etc. as they found insignificant based on EDA analysis.

d) Last Notable activity and Last Activity columns are redundant to each other, hence dropped ‘Last Activity’ column.

STEP 2: DATA PREPARATION

- Prepared Dummy variables for Categorical columns and Dropped original columns
- Concatenated Dummy variables dataset and Lead dataset
- Shape of new dataset: 9240 x 57
- Test-Train data split: Test dataset: 2772 x 57, Train dataset: 6468 x 57
- Target variable = Converted

STEPS FOLLOWED…

- Feature scaling using MinMaxScaler() function
- fit_transform() applied to numerical columns of Train data

STEP 3: MODEL BUILDING AND FITTING

- Applied LogisticRegression() function to build model
- Automated Elimination: Used Recursive Feature Elimination (RFE) to eliminate least important features and select most important top 15 variables.
- Manual Elimination: Used Variance Inflation Factor (VIF) manual approach for further elimination of variables.
- Dropped some columns based on p-value>0.05 and VIF>5
- Checked Pearson-Chi square and Log-likelihood for overall goodness of fit of model

STEPS FOLLOWED…

- Predictions on Train dataset
- Residual analysis

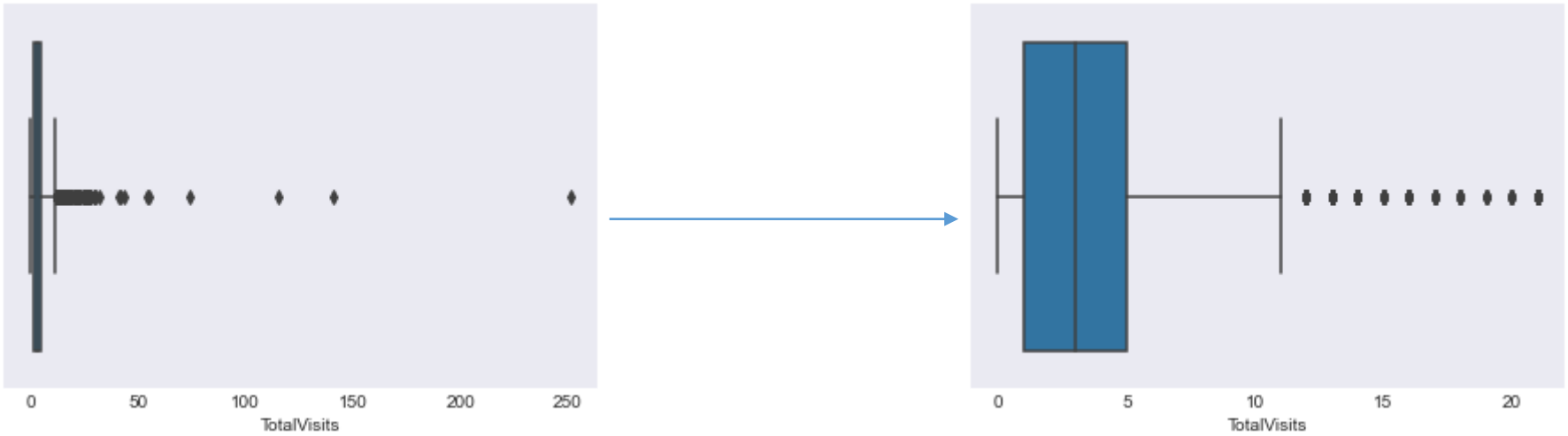
STEP 4: MODEL EVALUATION

- Confusion matrix, Classification report, accuracy, precision, recall
- Threshold value set at 0.5 initially and then Cut-off optimization done using Precision-Recall curve
- Checked AUC-ROC curve
- Model finalization and check fit for Test dataset
- Difference between Train and Test data should be <5%

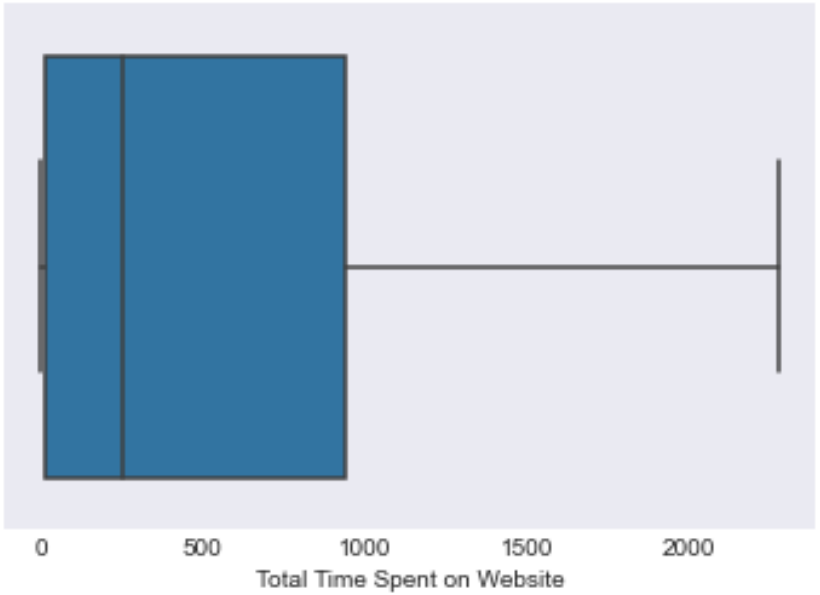
STEP 5: RECOMMENDATIONS

EXPLORATORY DATA ANALYSIS

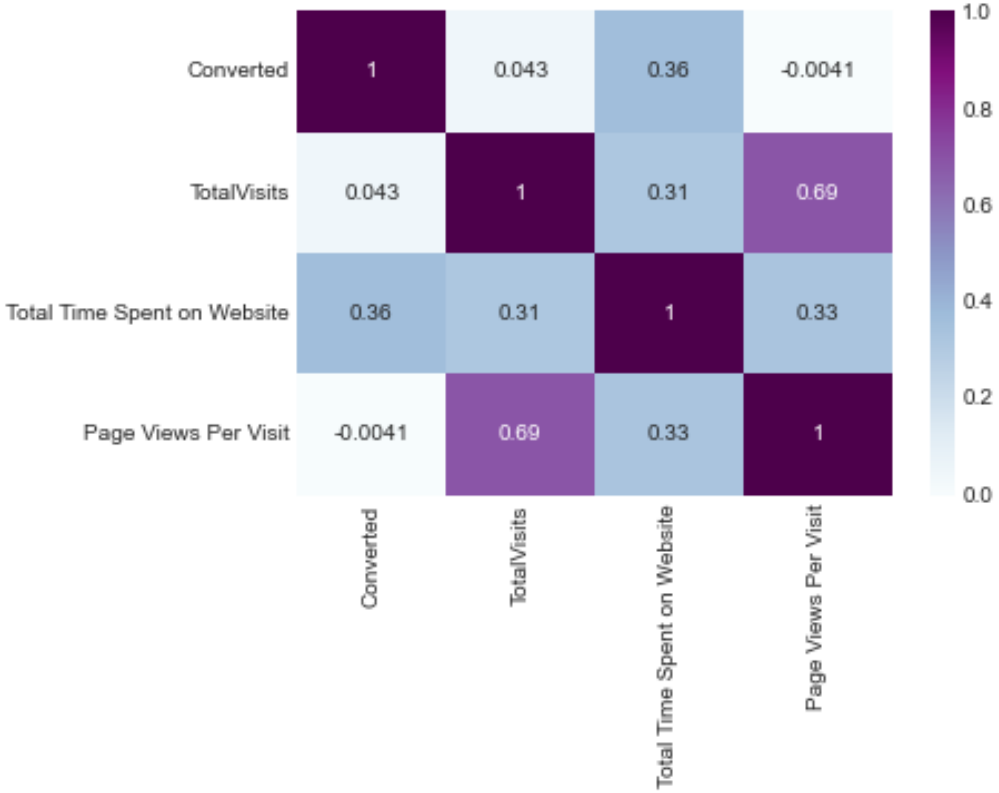
EDA - OUTLIER TREATMENT



EDA – NUMERICAL COLUMNS

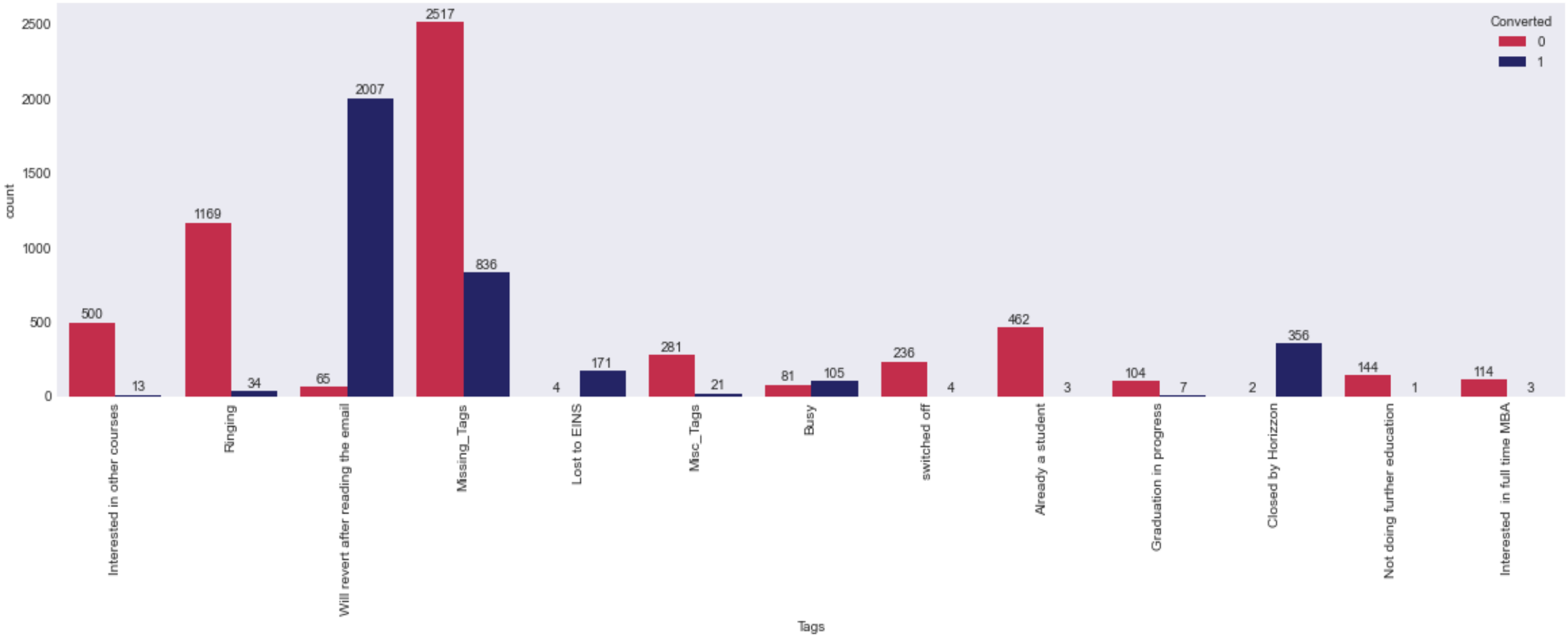


No visible outliers in Total time spent on website



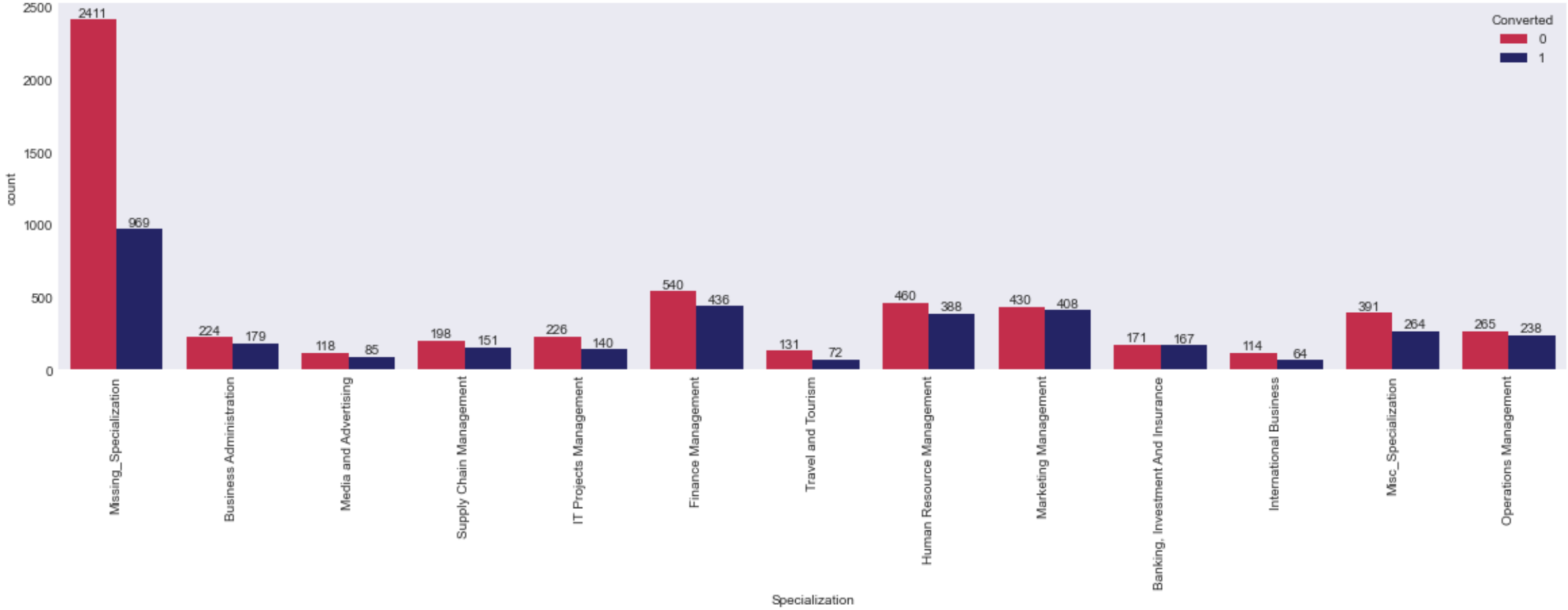
Positive correlation between Total Visits and Page views per Visit

EDA: BIVARIATE ANALYSIS OF CATEGORICAL COLUMNS



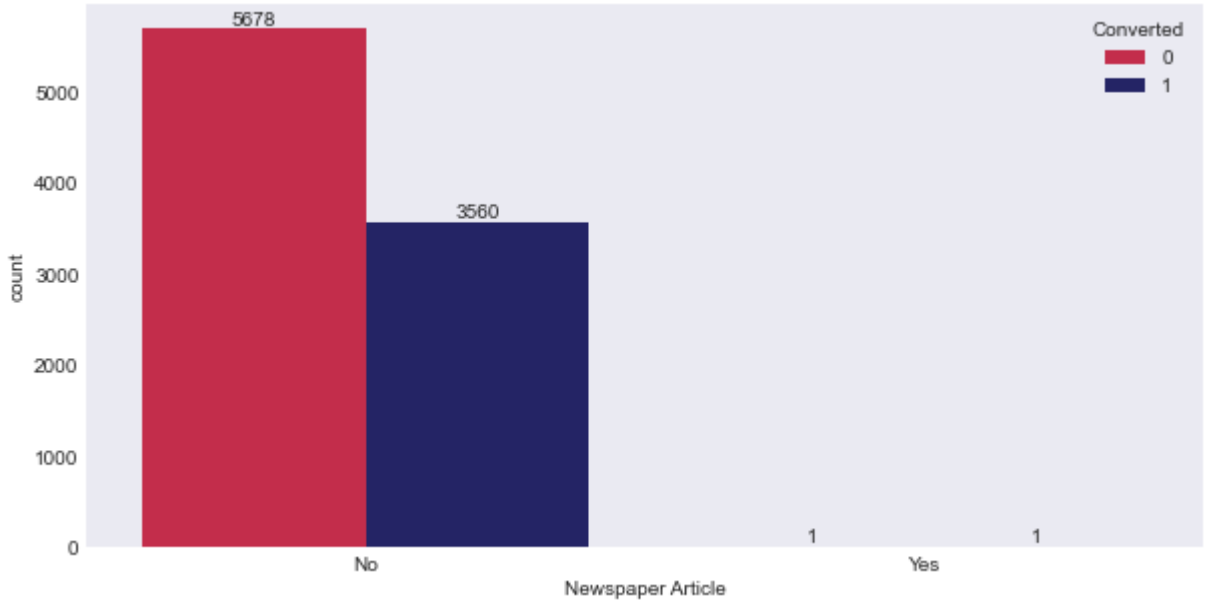
- Groped some categories as Misc.
- 'Will revert after reading the mail' has highest conversion rate followed by 'Closed by horizon'

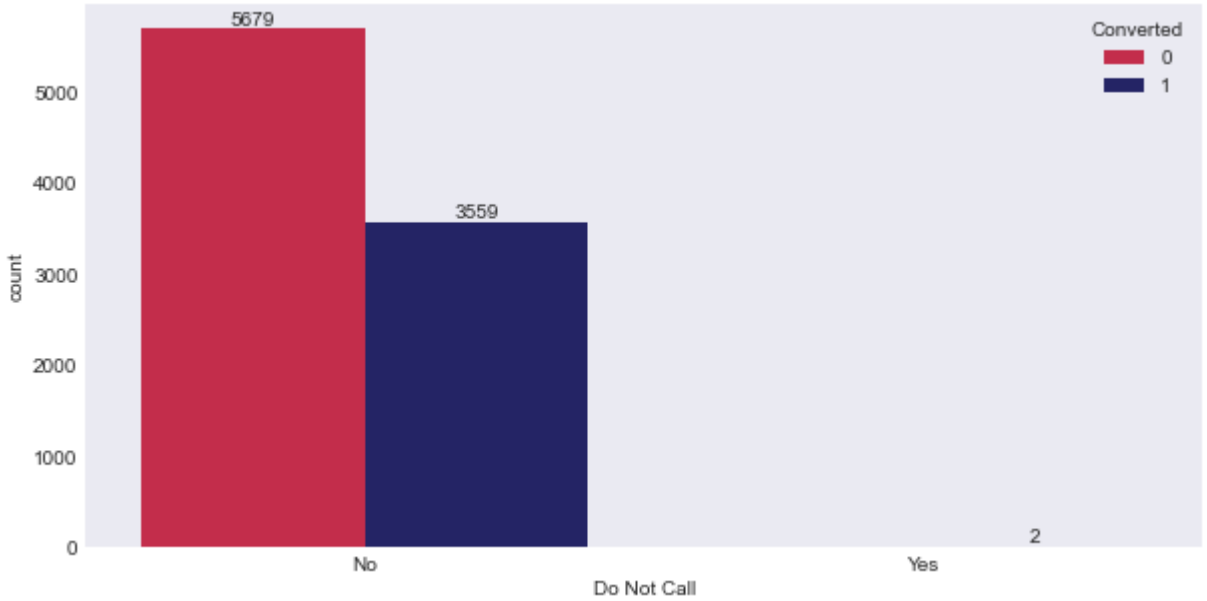
EDA: BIVARIATE ANALYSIS OF CATEGORICAL COLUMNS

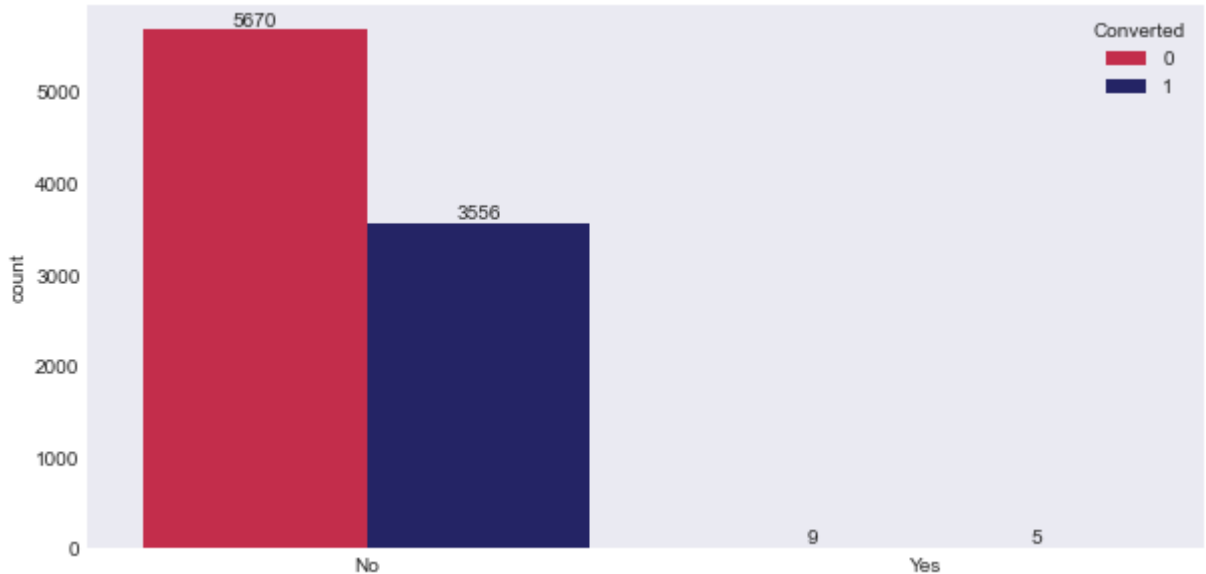


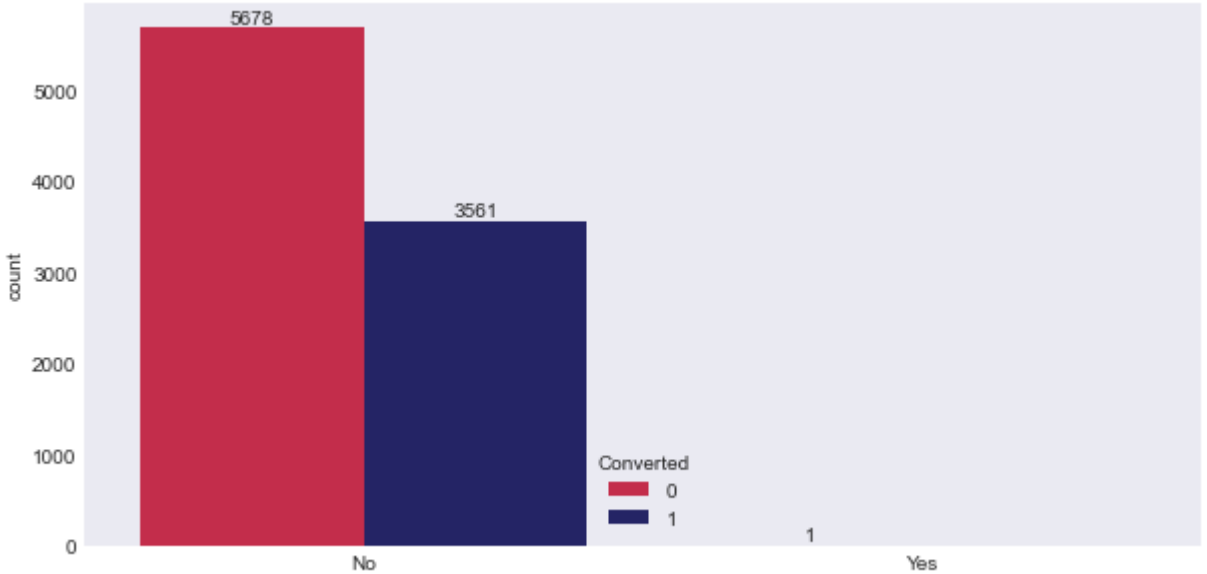
- Positive conversion rate observed for ‘Finance Management’, ‘ Human Resource Management’, ‘Marketing management’, ‘Operations Management’ variables.

EXAMPLES OF DATA IMBALANCE AND INSIGNIFICANT COLUMNS

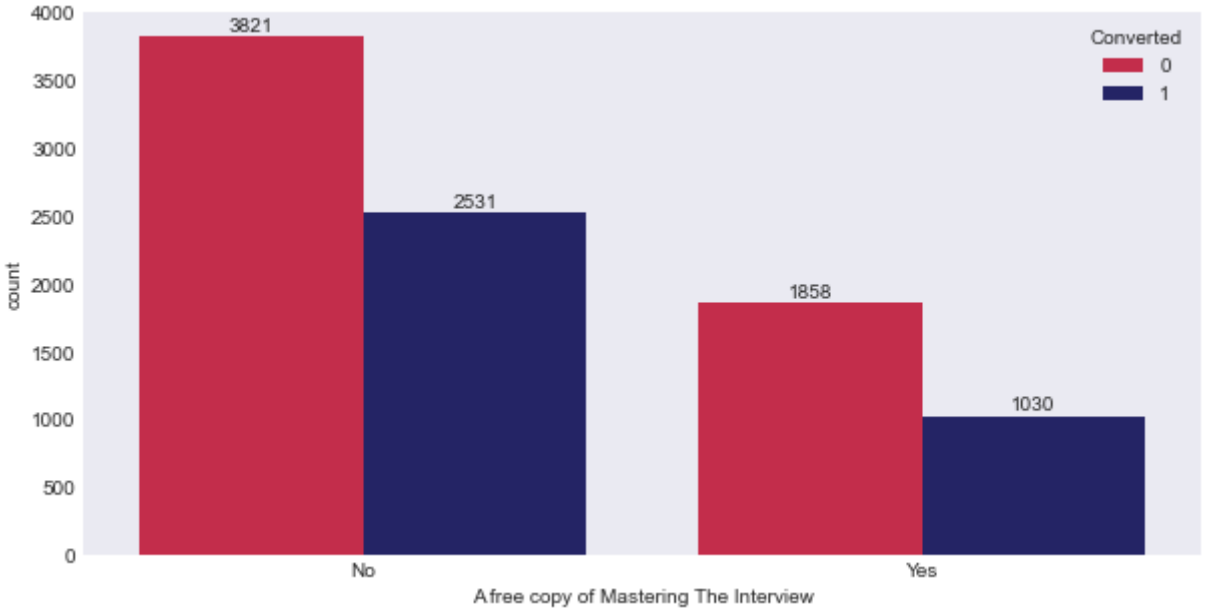


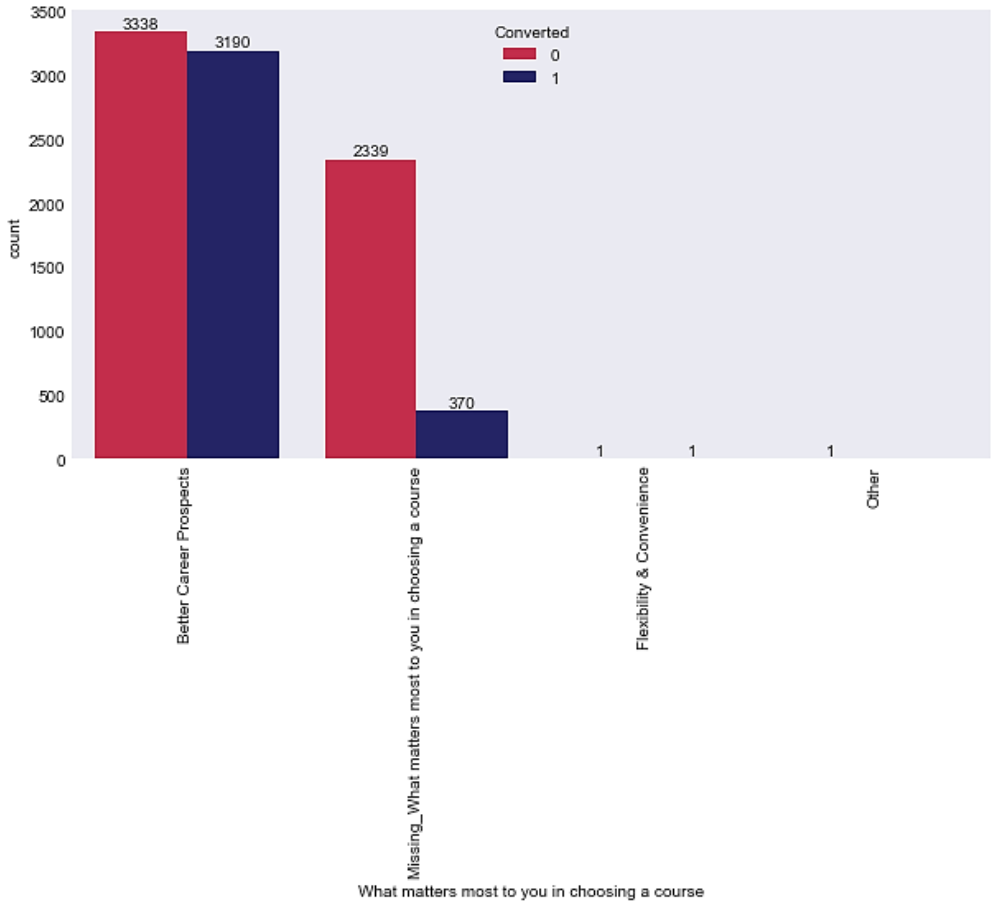






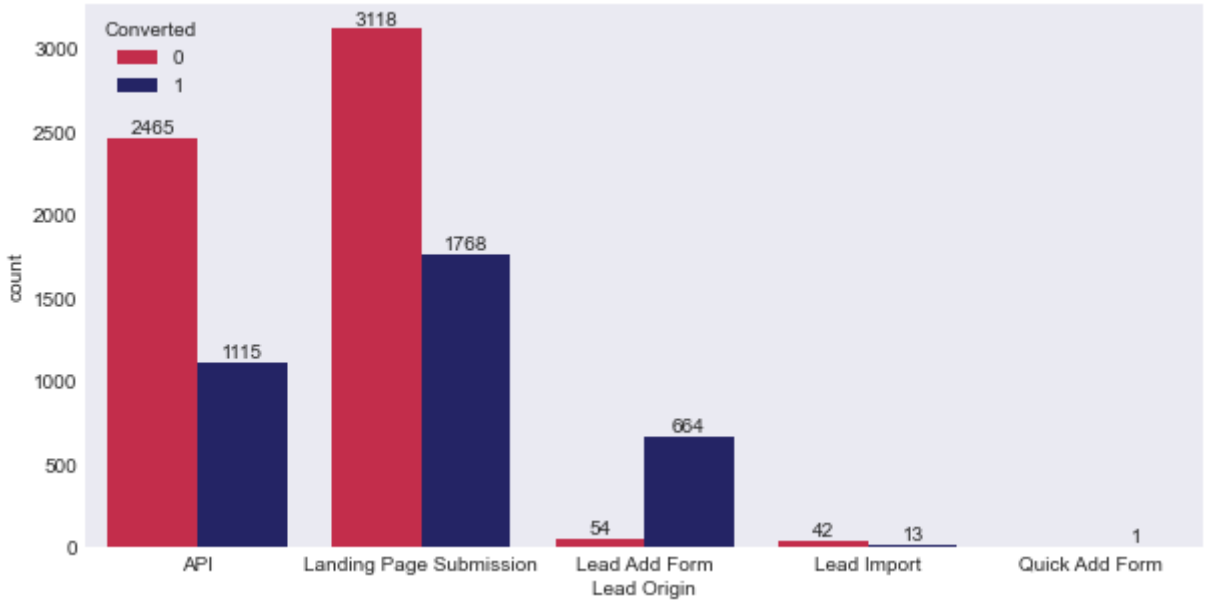
SIGNIFICANT COLUMNS

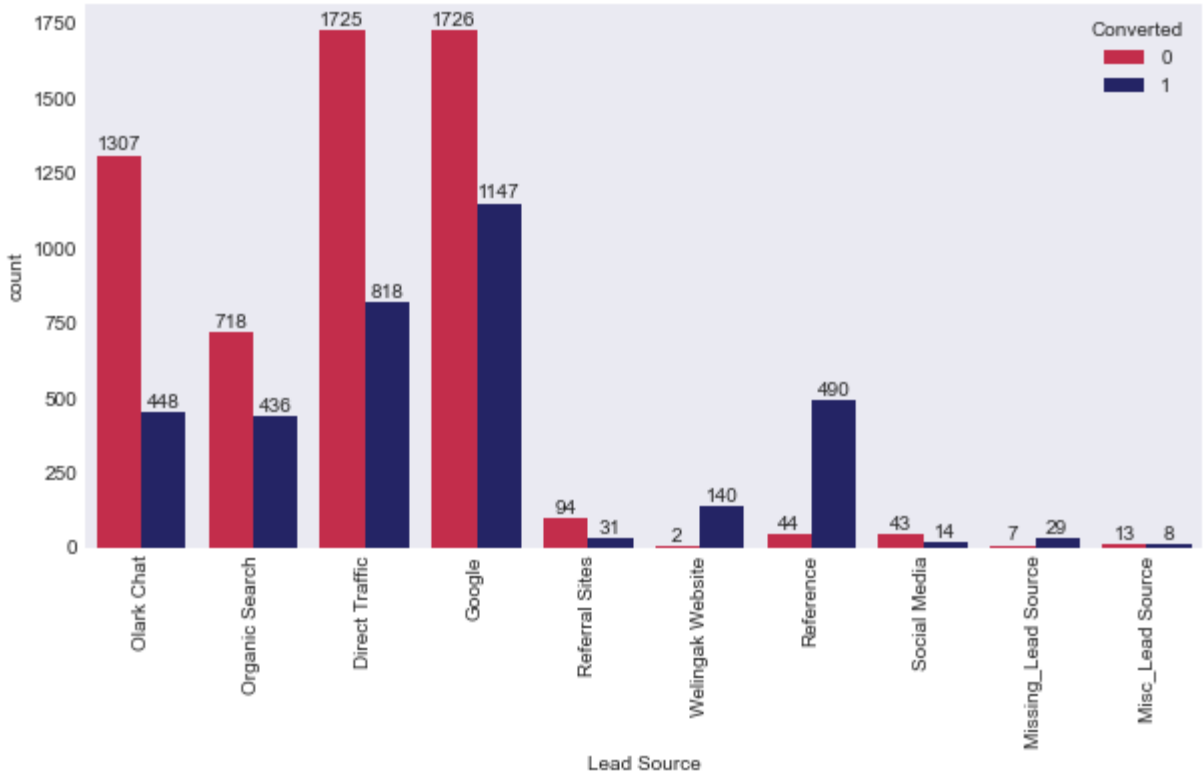




- Can not withdraw any conclusion towards Positive conversion rate.
- ‘Better Career Prospects’ among highest reason for Positive Conversion Rate.

SIGNIFICANT COLUMNS

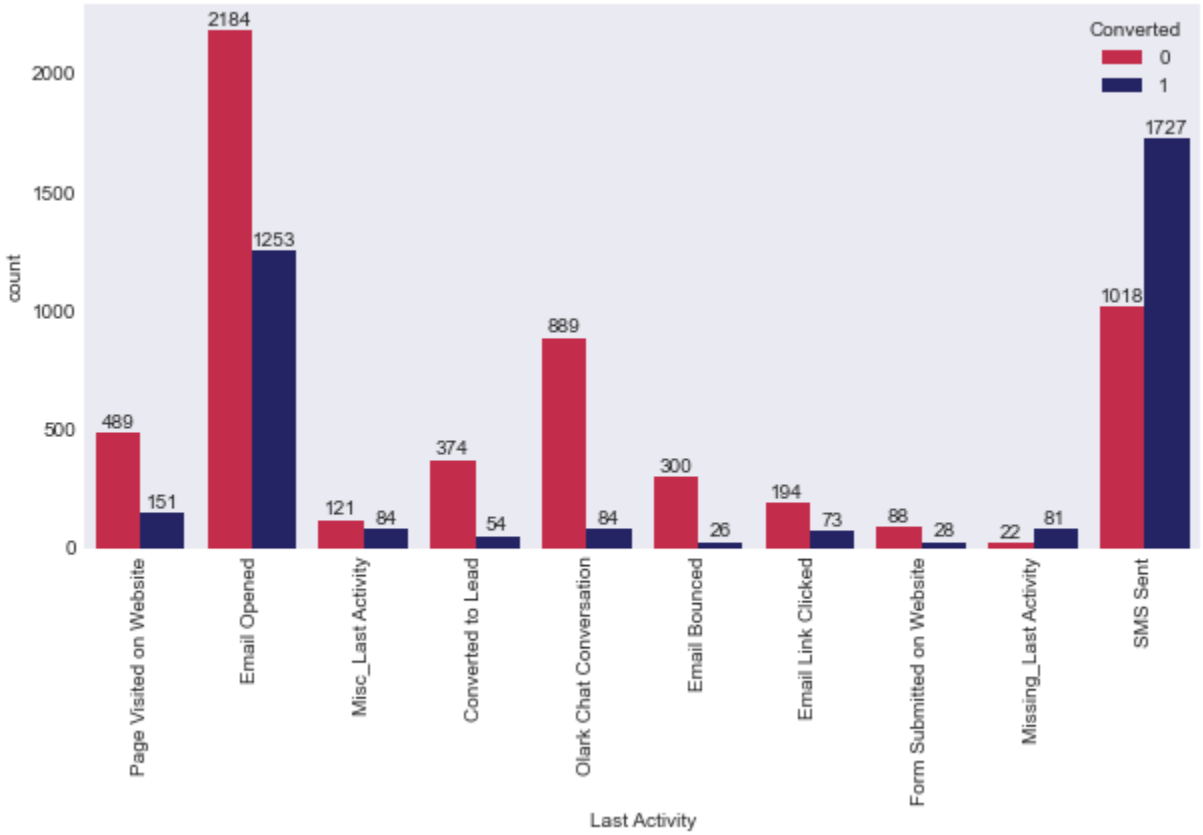


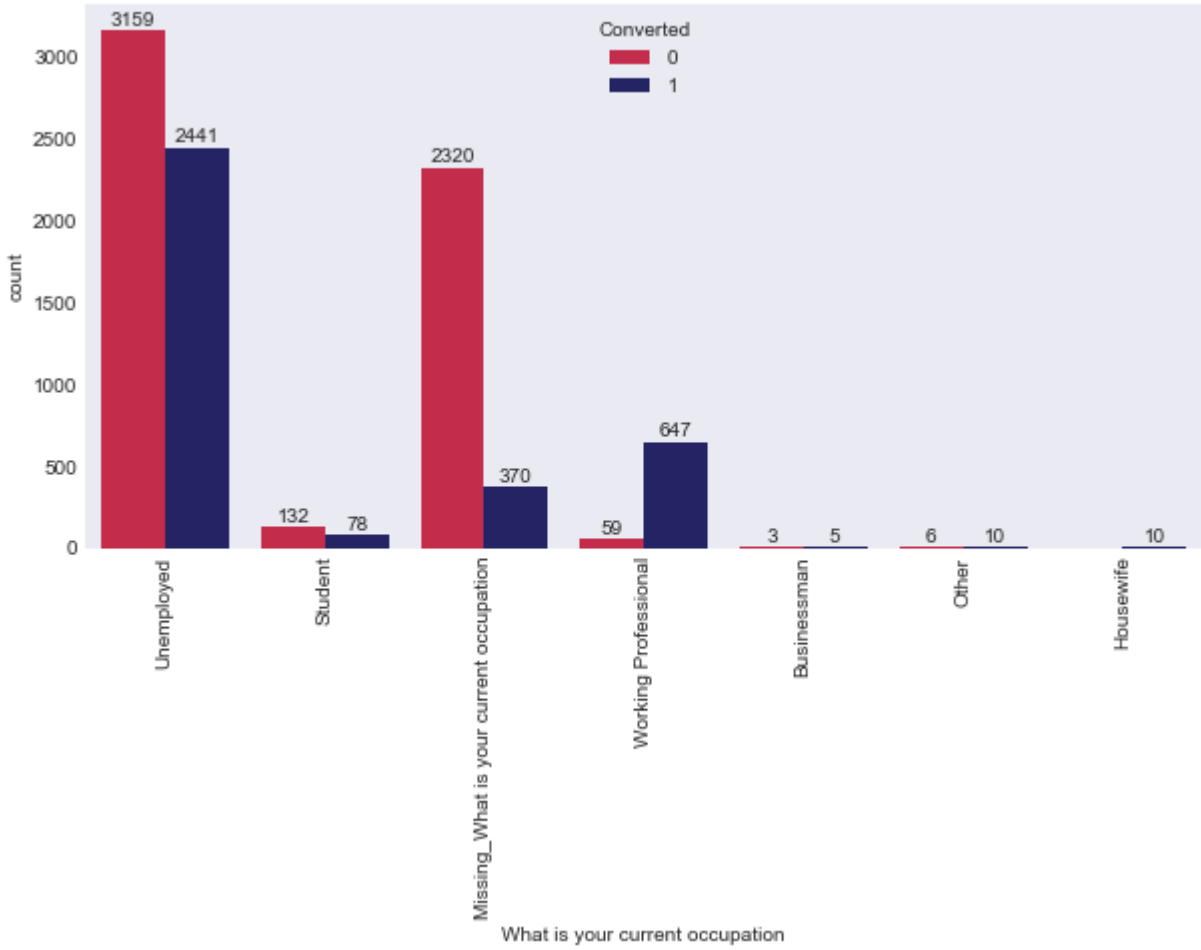


‘Landing Page submission’ and ‘API’ variables show good conversion rate

In Lead Source categories ‘Google’ and ‘Direct Traffic’ variables show good conversion rate

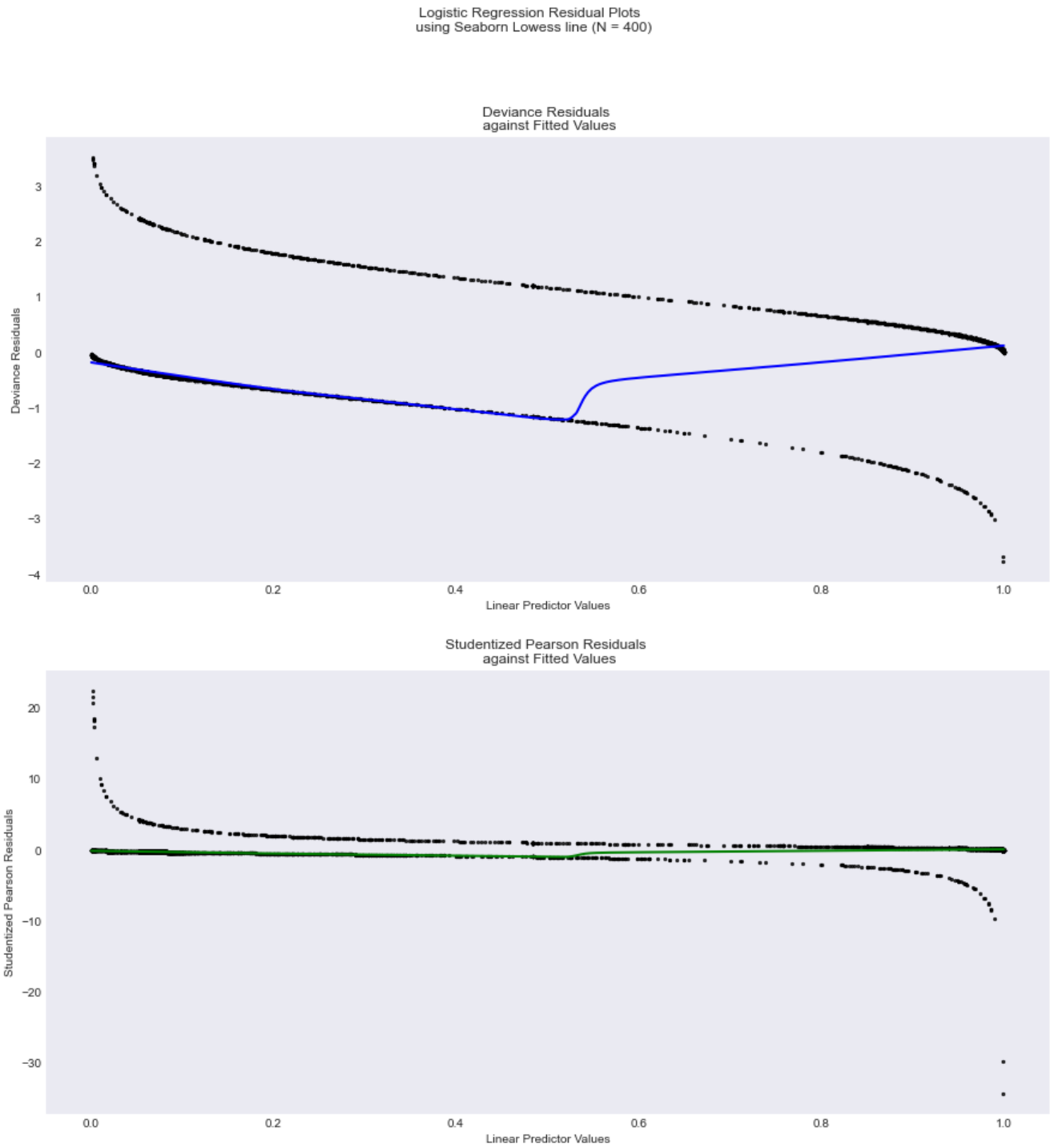
SIGNIFICANT COLUMNS





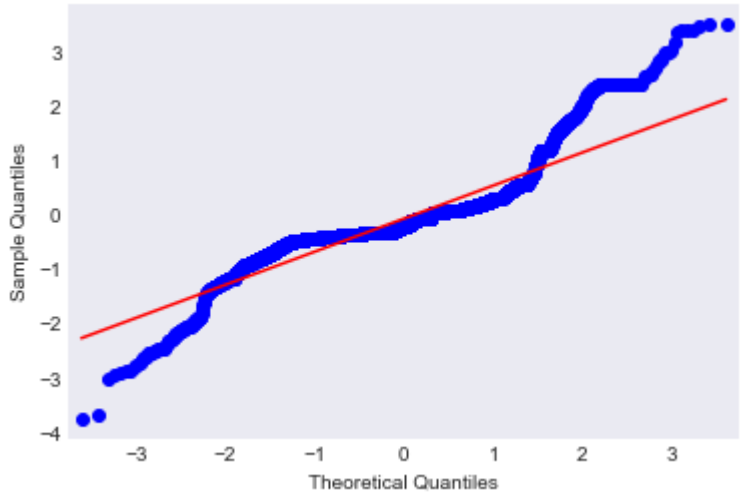
- Positive Conversion rate found for ‘Email open’ and ‘SMS sent’ of Last activity
- Positive Conversion rate found for ‘Unemployed’ category of Current Occupation

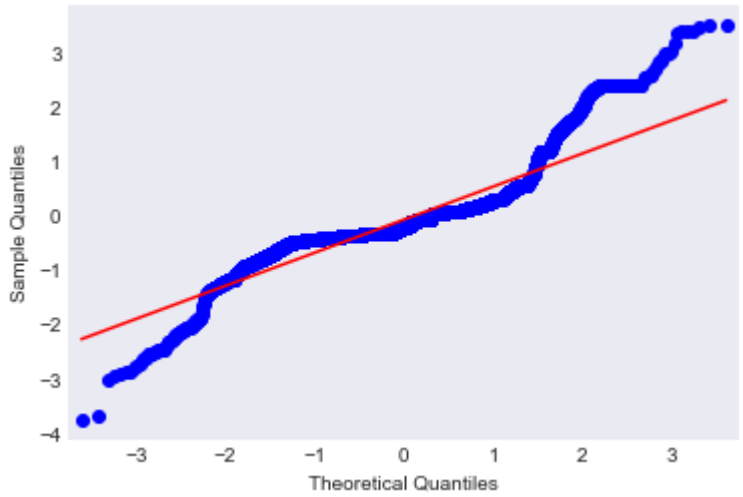
MODLE EVALUATION



RESIDUAL ANALYSIS

The plot has parallel lines with zero intercept which indicates that there is no significant model inadequacy.

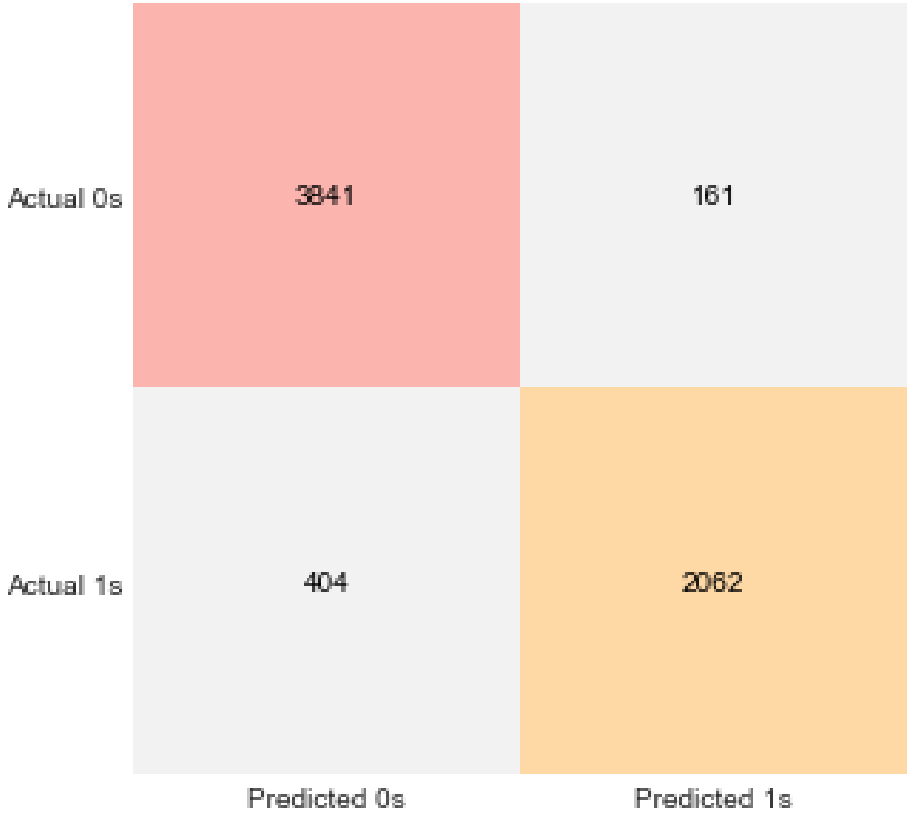




MODEL EVALUATION – TRAIN DATASET

CONFUSION MATRIX

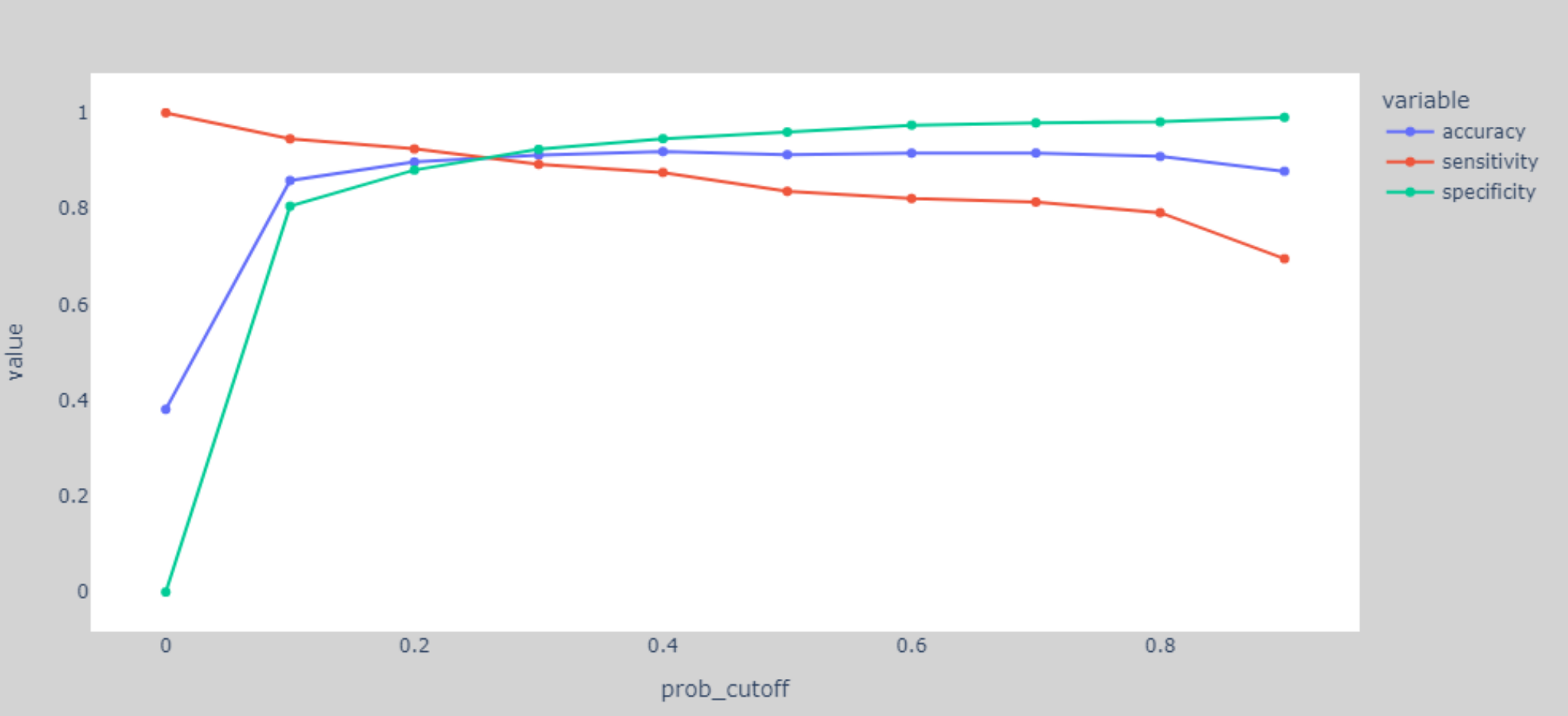
Threshold value = 0.5



PARAMETER	THRESHOLD = 0.5
Accuracy	0.9126
Precision	--
Recall	0.8361
Sensitivity	0.8361
Specificity	0.9597
False Positive Rate (FPR)	0.0402
Positive Predictive Value	0.9275
Negative Predictive Value	0.9048

MODEL EVALUATION – TRAIN DATASET

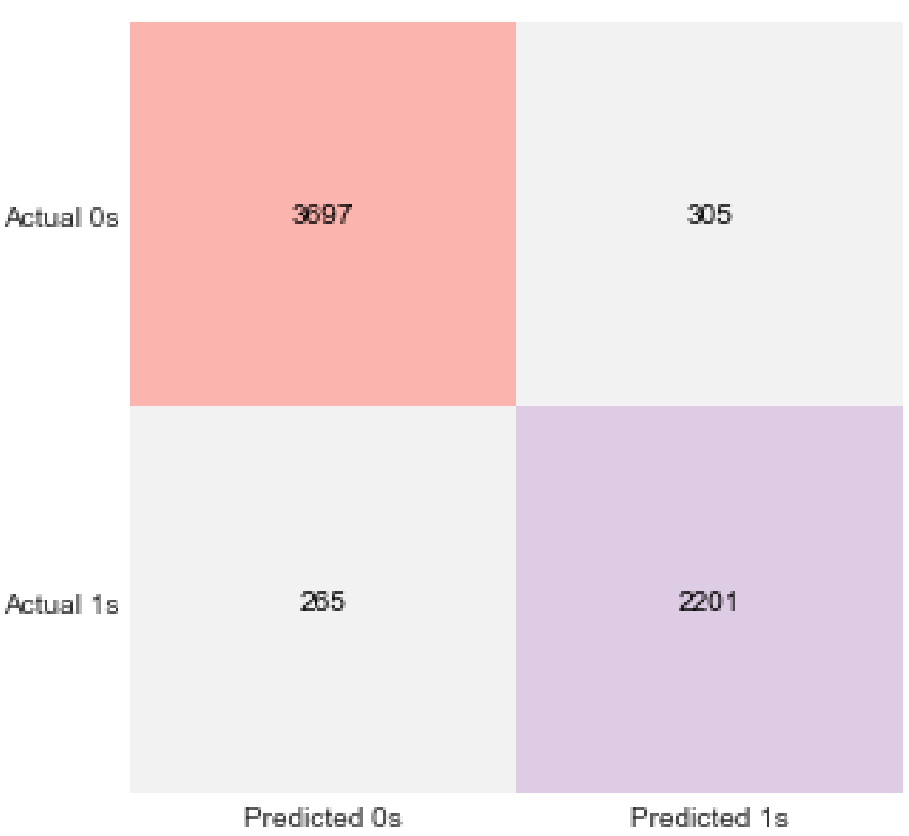
OPTIMAL PROBABILITY CUT-OFF FINDING = 0.3



MODEL EVALUATION – CONFUSION MATRIX

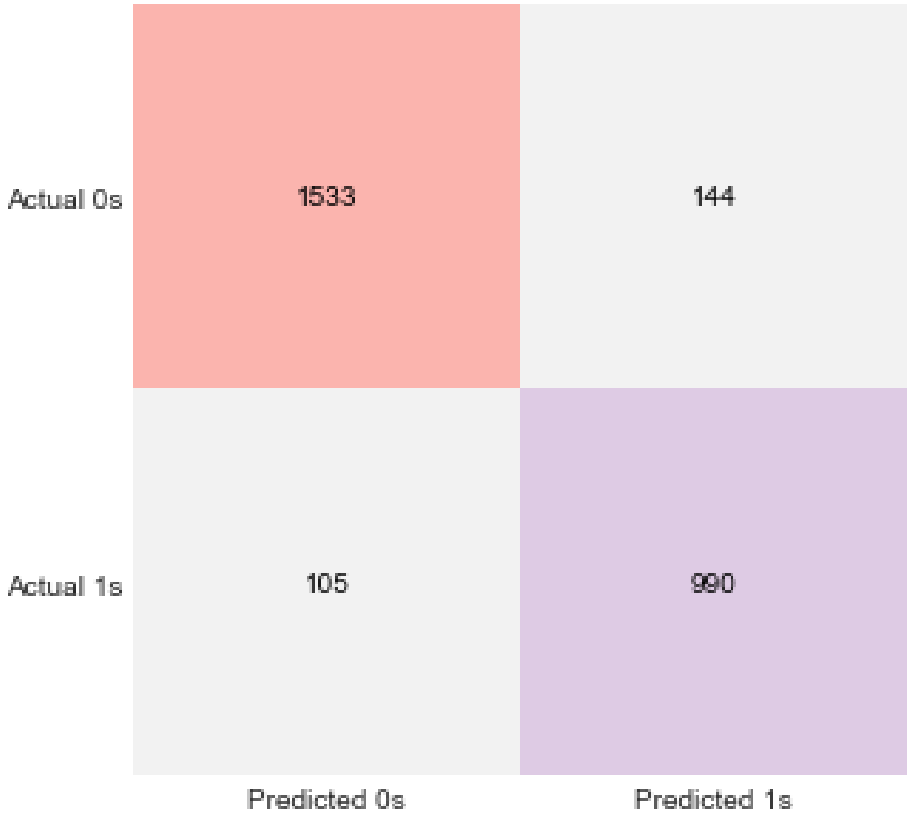
TRAIN DATASET

Threshold value = 0.3



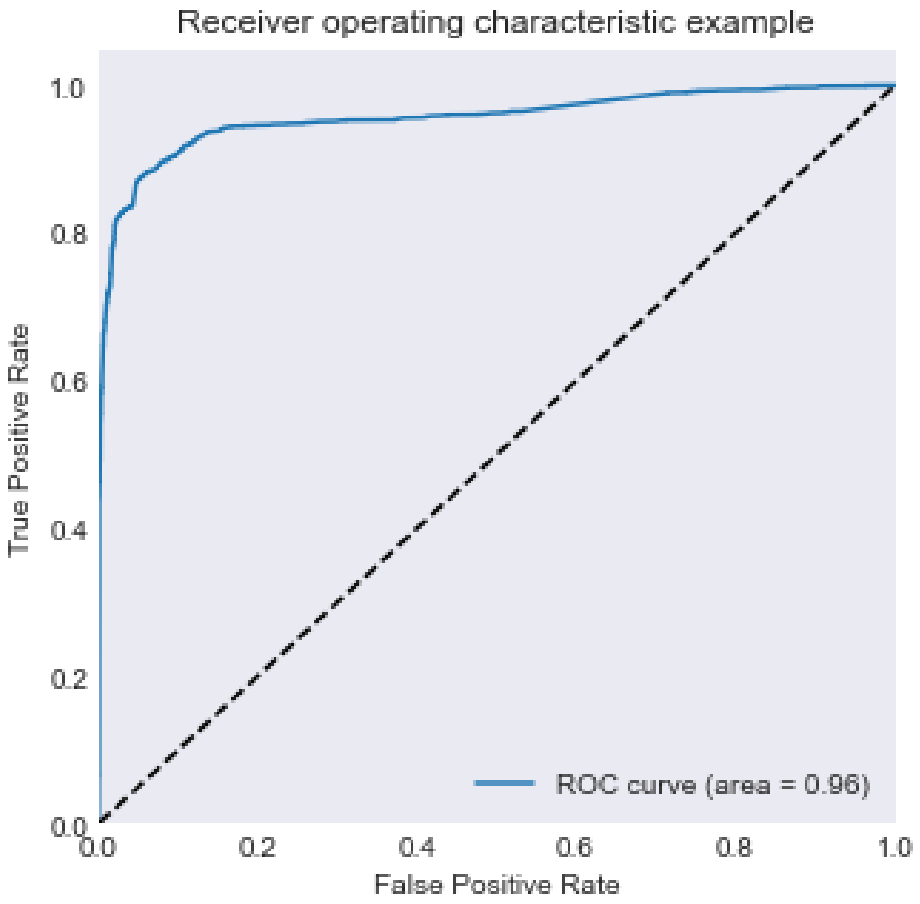
TEST DATASET

Threshold value = 0.3

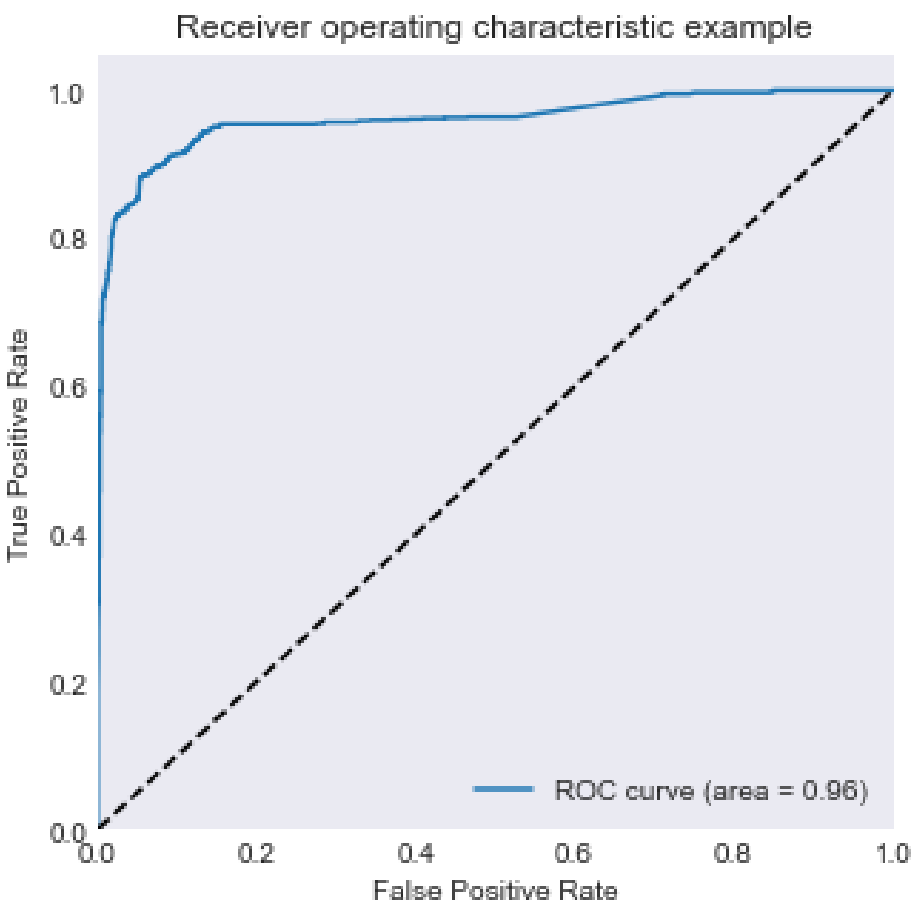


MODEL EVALUATION: AUC –ROC CURVE

TRAIN DATASET

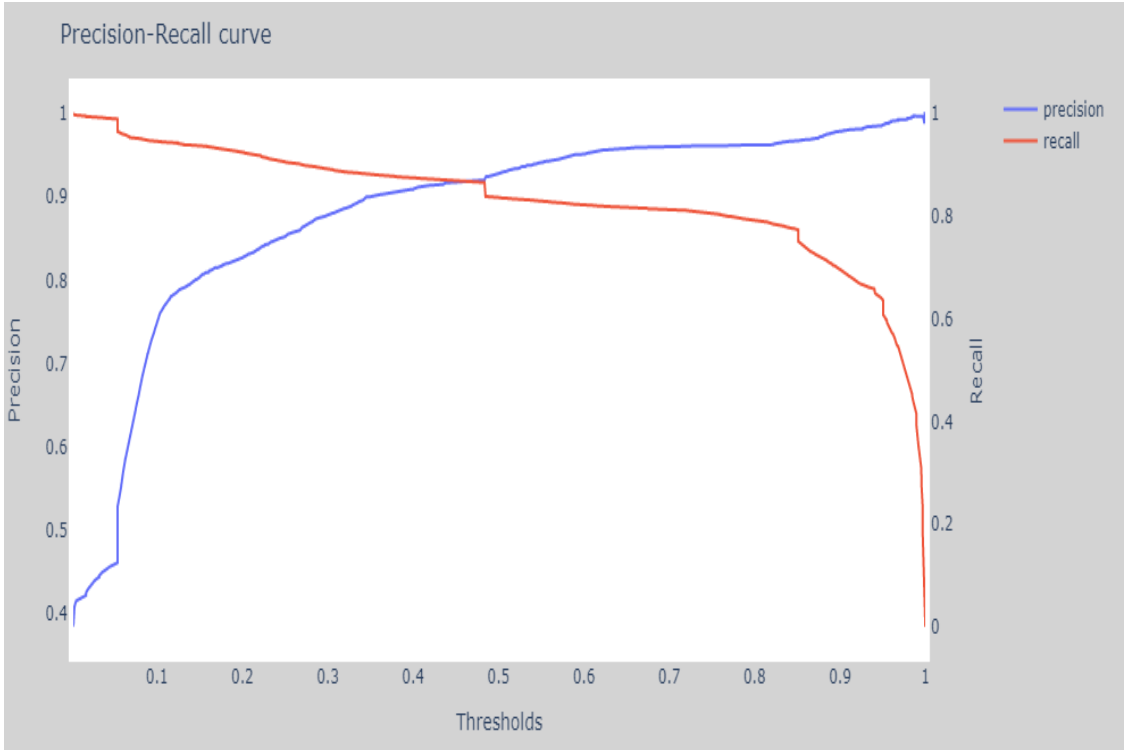


TEST DATASET

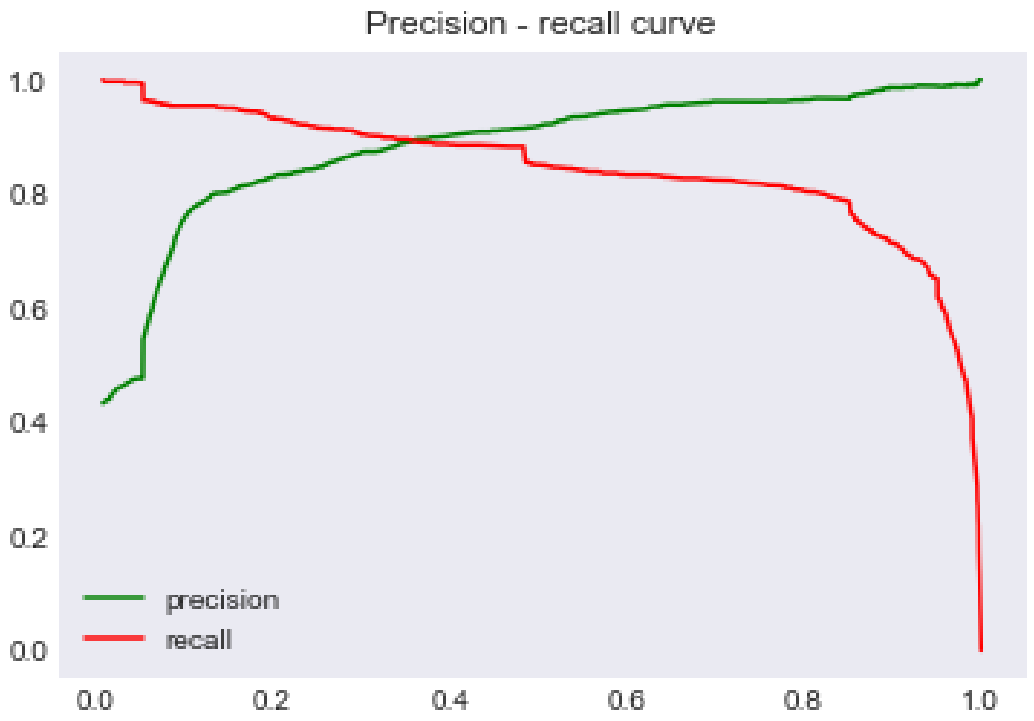


MODEL EVALUATION: PRECISION-RECALL CURVE

TRAIN DATASET



TEST DATASET



MODEL EVALUATION

	TRAIN DATASET	TEST DATASET
PARAMETER	THRESHOLD = 0.3	THRESHOLD = 0.3
Precision	0.8782	0.8730
Recall	0.8925	0.9041
Specificity	0.9237	0.9141

- The predictors were statistically significant i.e, VIFs less than 3 and p-value less than 0.05 and relevant to the target variable.
- The Pearson chi-square and Log-Likelihood of the model signifies goodness of fit i.e, 9.99e+03 and -1453.1 respectively.
- The Pseudo R-squared validates the goodness of fit of model i.e, 0.58

MODEL EVALUATION – COMPARISON

- CLASSIFICATION REPORT - Threshold value = 0.3

TRAIN DATASET	precision	recall	f1-score	Support
0	0.93	0.92	0.93	4002
1	0.88	0.89	0.89	2466
Accuracy			0.91	6468
Macro avg.	0.91	0.91	0.91	6468
Weighted avg.	0.91	0.91	0.91	6468

TEST DATASET	precision	recall	f1-score	Support
0	0.94	0.91	0.92	1677
1	0.87	0.90	0.89	1095
Accuracy			0.91	2772
Macro avg.	0.90	0.91	0.91	2772
Weighted avg.	0.91	0.91	0.91	2772

BUSINESS REQ. AND RECOMMENDATIONS

- Business requirement: Conversion rate ~80%
- Business Recommendations:

Following categories should be focused more for maximum conversion rate:

1. Total Time Spent on Website
2. Lead Origin_Lead Add Form
3. Lead Source_Welingak Website
4. Tags_Closed by Horizzon
5. Tags_Will revert after reading the email

BUSINESS RECOMMENDATIONS

- Considering a stage to make the lead conversion aggressive can be attained by setting a threshold less than 0.3, which aims to increase the count of potential leads by predicting customers as 1 who holds a conversion probability less than 0.3.
- Accounting a scenario, where the company needn’t focus to make unnecessary phone calls and consider only those customers who have a high probability of conversion positively, can be achieved by setting a threshold higher than 0.5. This contributes in predicting customers as 1 who holds a conversion probability greater than 0.5.

THANK YOU