Lead Scoring Case Study

A Presentation by-

Anugrah Stanley **Amruta Gaurav** **Vishal Singla**

Problem Statement

X Education, an educational company provides online courses to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses. The company also markets its courses on several websites and search engines like Google.

The company gets leads through the people who land on their website browse the courses or fill up a form for the course or watch some videos. The company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

Now, although X Education gets a lot of leads, its lead conversion rate is around **30%** which is very poor. To make this process more efficient, the company 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.

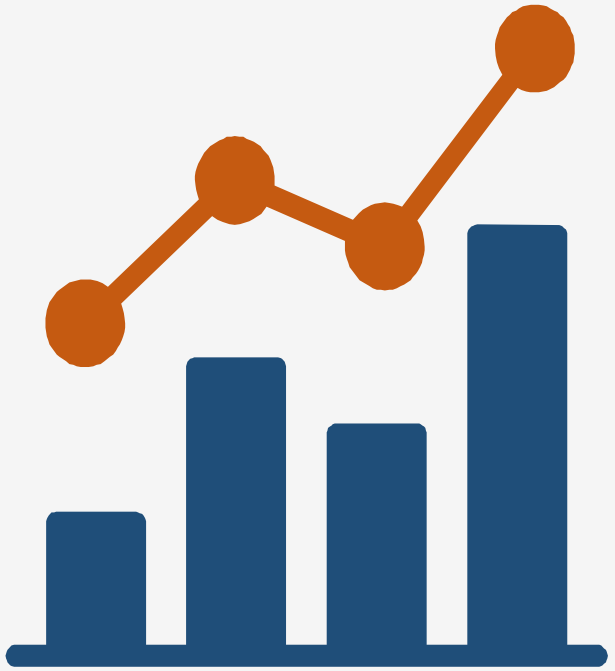
Business Goal

The company requires to build a model wherein a lead score is to be assigned 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%**.

Work Flow

1. Importing Libraries
2. Data Loading and Sanity Checks
3. Data Cleaning
4. Exploratory Data Analysis
5. Data Preparation for Modelling
6. Model Building
7. Prediction and Model Evaluation on Train Set
8. Prediction and Model Evaluation on Test Set



1 Importing Libraries

For Mathematical Functions

For Data Manipulation and Analysis

For Visualizations

For Visualizations

For Estimation of Statistical Model

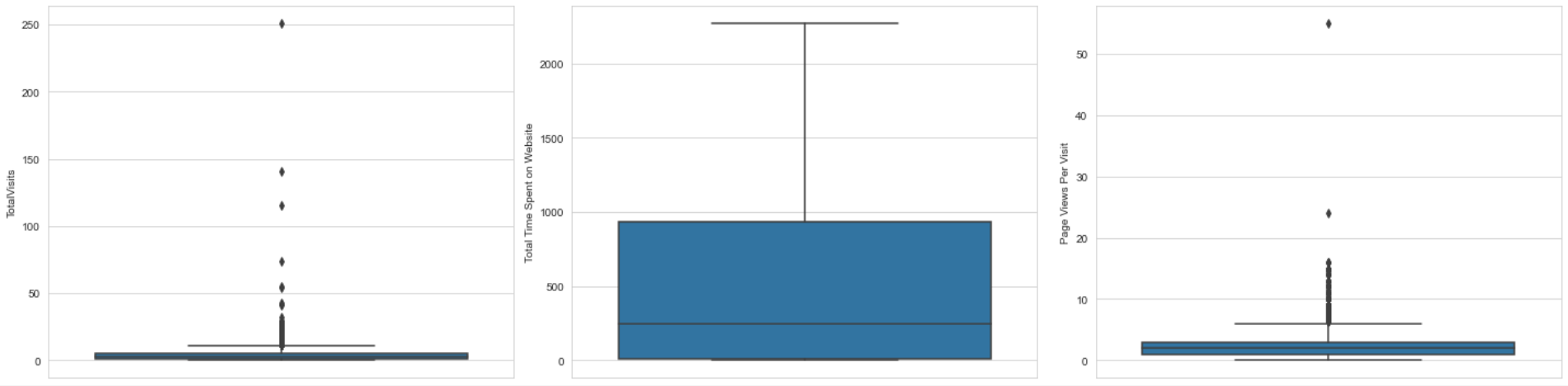
For Predictive Analysis

2 Data Loading and Sanity Checks

- Loaded the data set using pandas
- The data frame contains 36 features, 1 target variable and 9240 rows.
- The data types of features are as follows:
  - 4 features are float64
  - 3 features are integer
  - 30 features are object
- There are a lot of features with null values but datatypes of these features looks fine.

3 Data Cleaning : Univariate Analysis

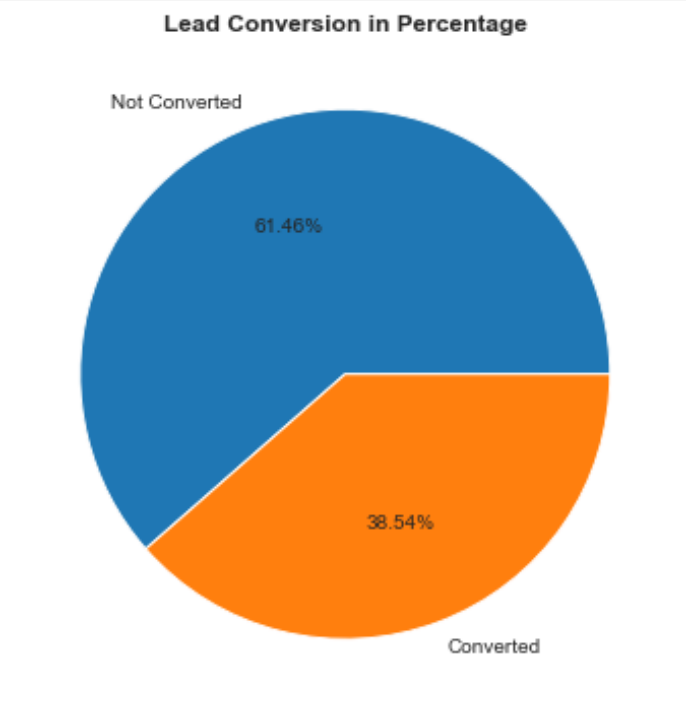
- Dropped features having null values above 45%
- Imputed the rest of the missing values in the best possible way after analysis.
- Dropped some irrelevant/unnecessary features after thorough analysis.
- Fixed incorrect data values and modified some features for ease of analysis.
- Identified and handled the outliers.



4 Exploratory Data Analysis:

❖ Conversion Rate

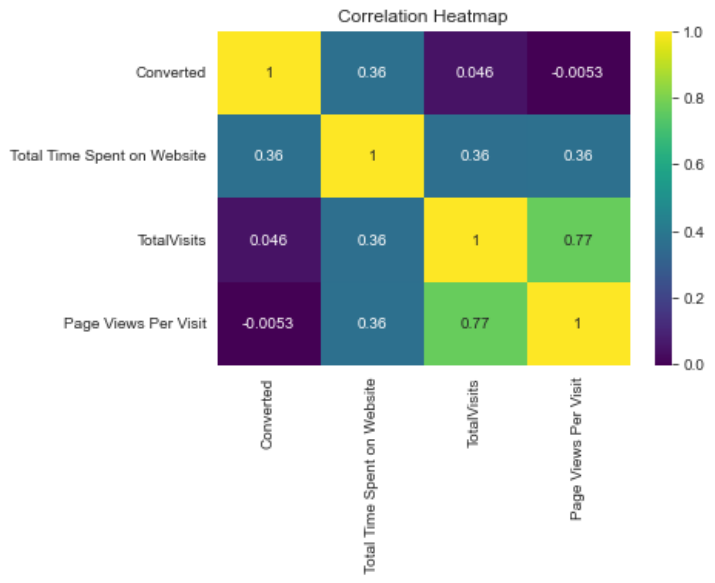
✓ Lead Conversion rate is 38.54 %



❖ Multivariate Anaylsis

✓ Total Visits and Page Views Per Visit have strong correlation

✓ Total Visits and Total Time Spent on Website have low correlation.

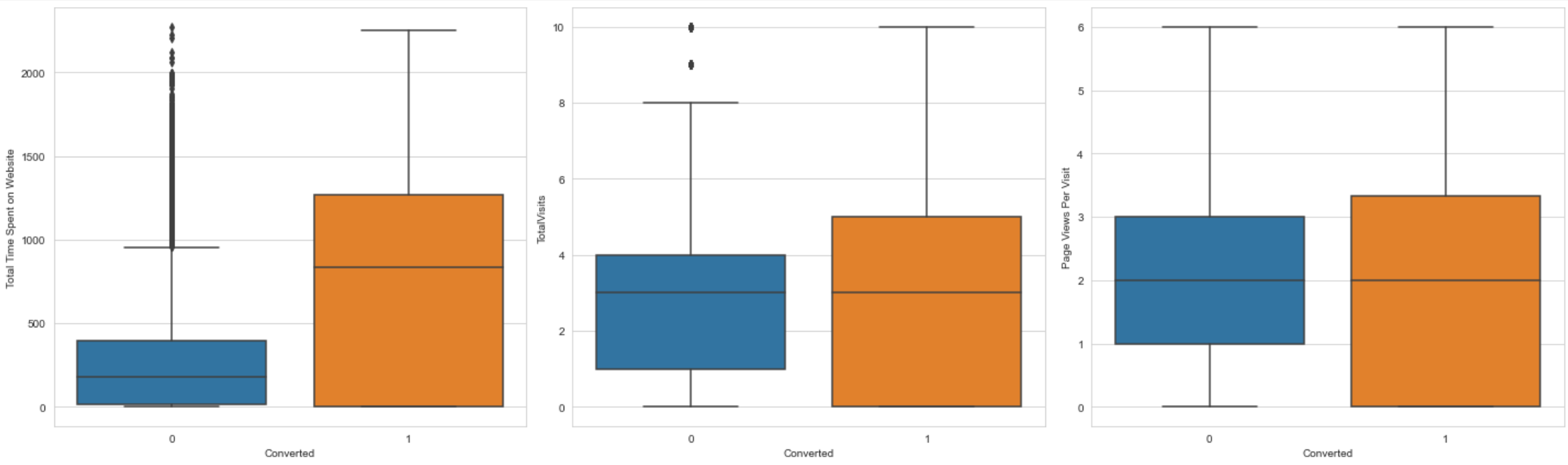


4 Exploratory Data Analysis: Bivariate Analysis (Numeric Features)

❖ Insights

✓ Leads spending more time on the website are more likely to be converted.

✓ Total Visits and Page Views Per Visit bears no impact on lead conversion as their medians are same.



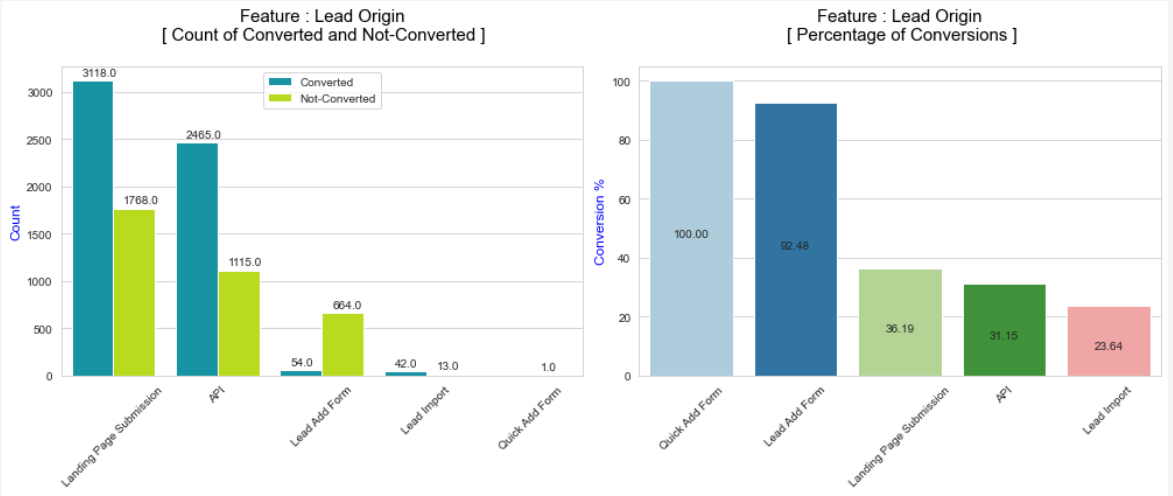
4 Exploratory Data Analysis: Bivariate Analysis (Categorical Features)

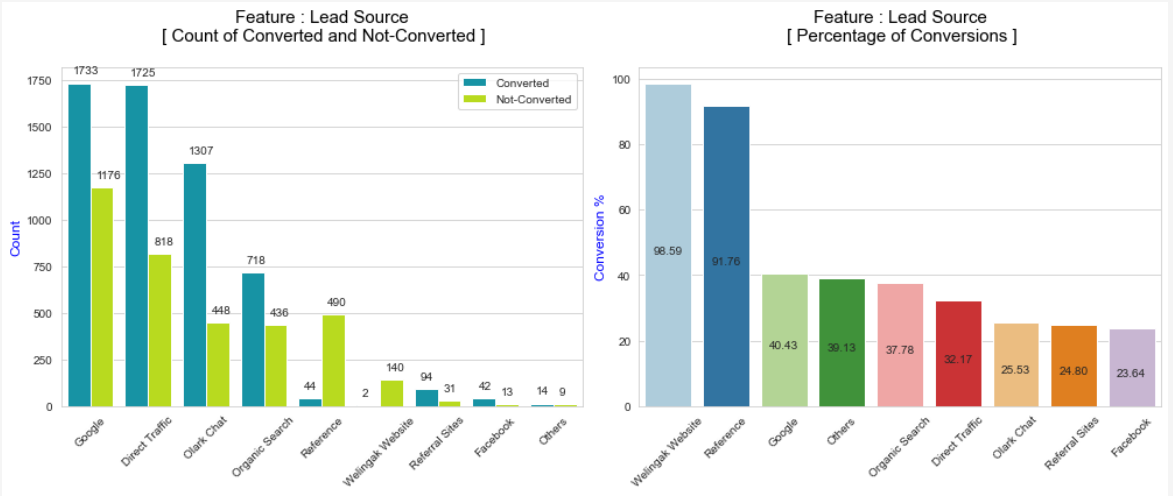
❖ Insights from Lead Origin

- ✓ Quick Add Form and Lead Add Form have 100% and ~92% conversion rate but number of leads are very less.
- ✓ Landing Page Submission and API have considerable number of leads and have a conversion rate of 30-35%.
- ✓ Lead import has around 29% conversion rate but number of leads are very less.

❖ Insights from Lead Source

- ✓ Google and Direct Traffic generates maximum number of leads.
- ✓ Conversion Rate of leads through Reference and Welingak Website is high.

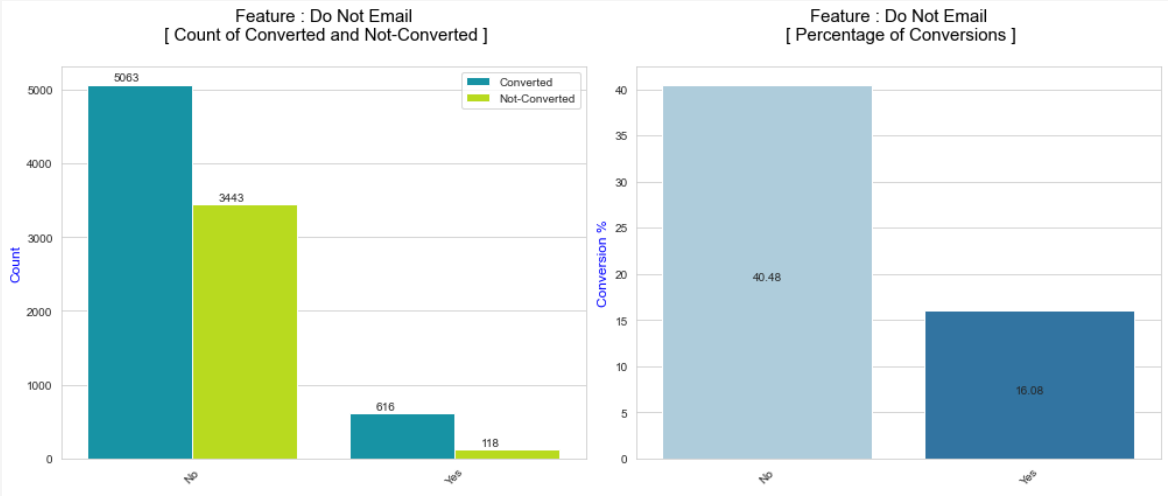




4 Exploratory Data Analysis: Bivariate Analysis (Categorical Features)

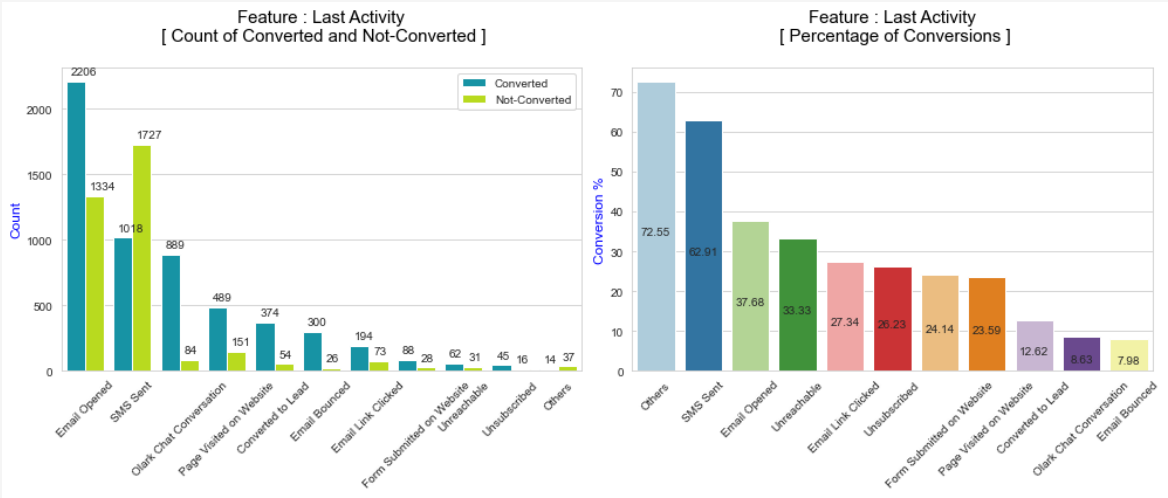
❖ Insights from Do Not Email

- ✓ Visitors who choose to get an email are more likely to be converted and have higher number of leads.



❖ Insights from Last Activity

- ✓ Most of the leads have Email opened as their last activity.
- ✓ Conversion rate for leads with last activity as SMS Sent is around 63%.
- ✓ Insights from Last Notable Activity are similar



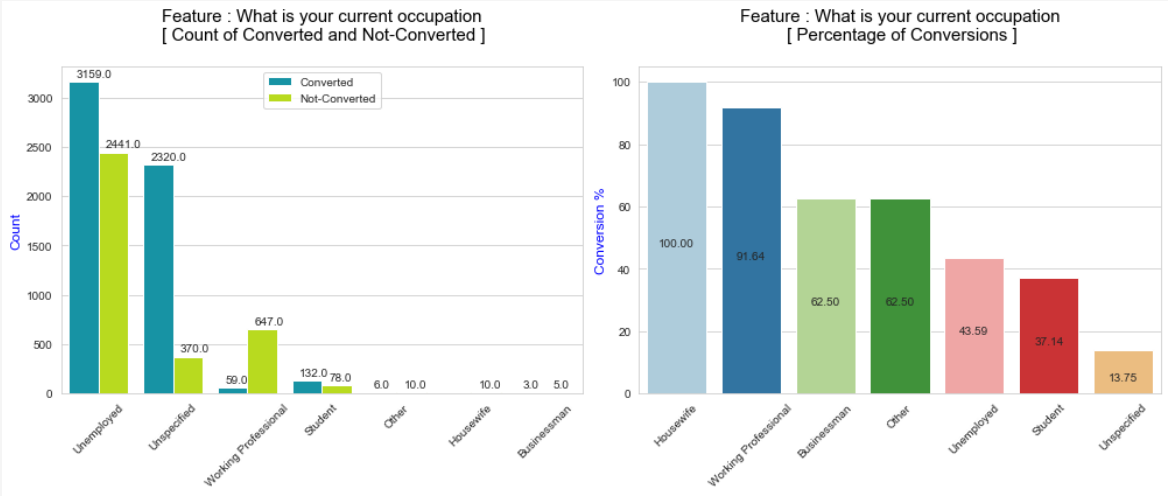
4 Exploratory Data Analysis: Bivariate Analysis (Categorical Features)

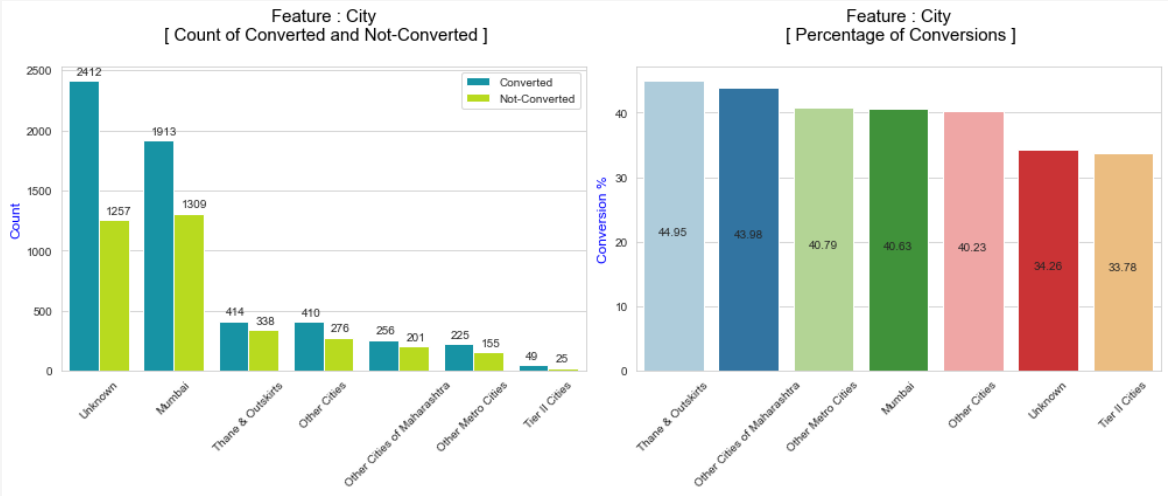
❖ Insights from What is your current occupation

- ✓ Working Professionals have a strong conversion rate of around 92%.
- ✓ Unemployed leads are the most in numbers but have around 43% conversion rate.

❖ Insights from City

- ✓ Most leads are from Mumbai with around 41% conversion rate

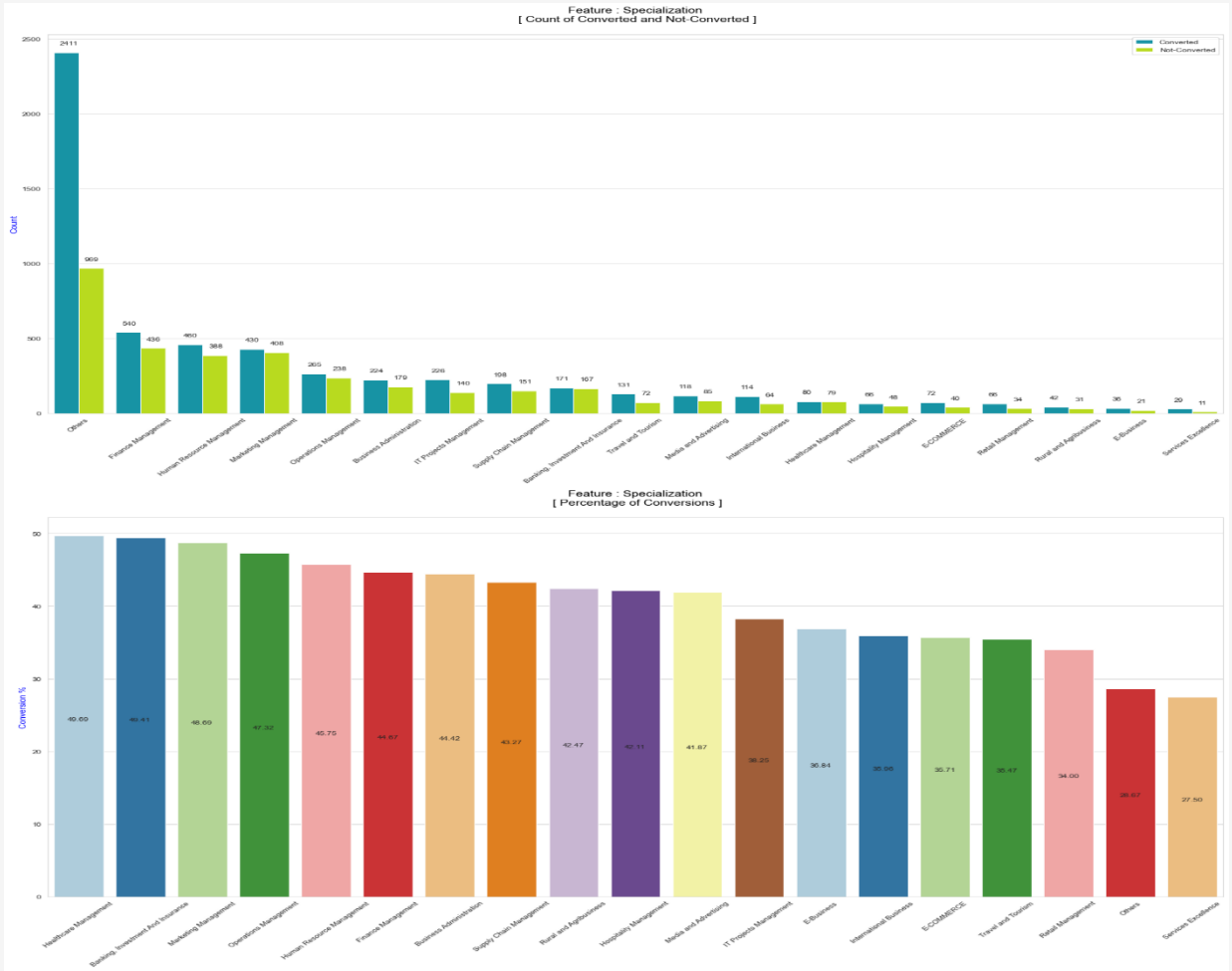




4 Exploratory Data Analysis: Bivariate Analysis (Categorical Features)

❖ Insights from Specialization

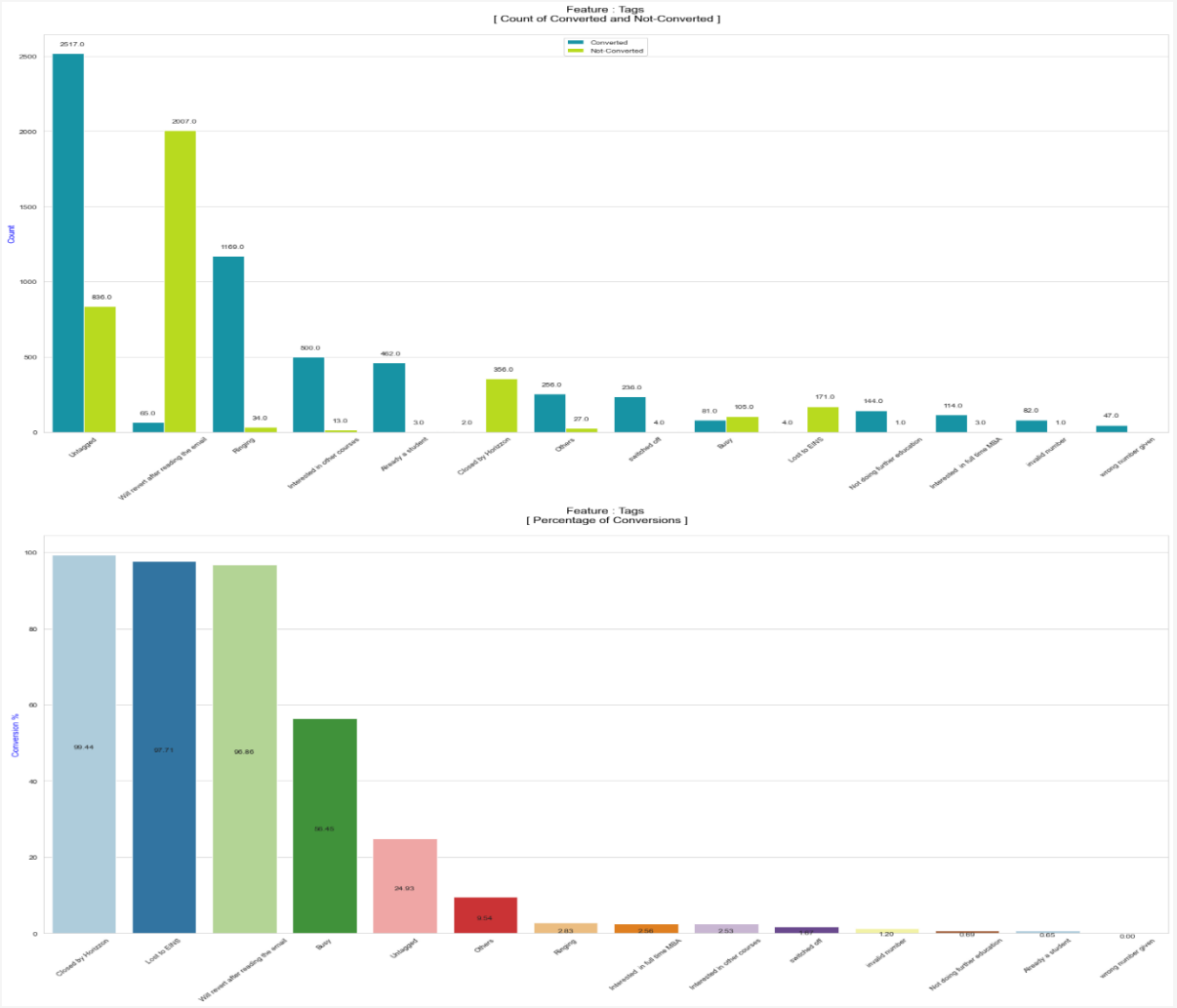
- ✓ Finance Management, Human Resource Management and Marketing Management generate higher number of leads.
- ✓ Conversion rate is highest for Healthcare Management followed by Banking, Investment and Insurance.



4 Exploratory Data Analysis: Bivariate Analysis (Categorical Features)

❖ Insights from Tags

- ✓ Closed by Horizons or Lost to Competitors have the highest conversion rate, but have limited leads.



4 Data Preparation for Modelling

- Mapped binary values of the features to 0 and 1.
- Created dummy variables for categorical features
- Performed the test-train split
- Fitted and transformed the numerical features with Standard Scaler

5 Model Building

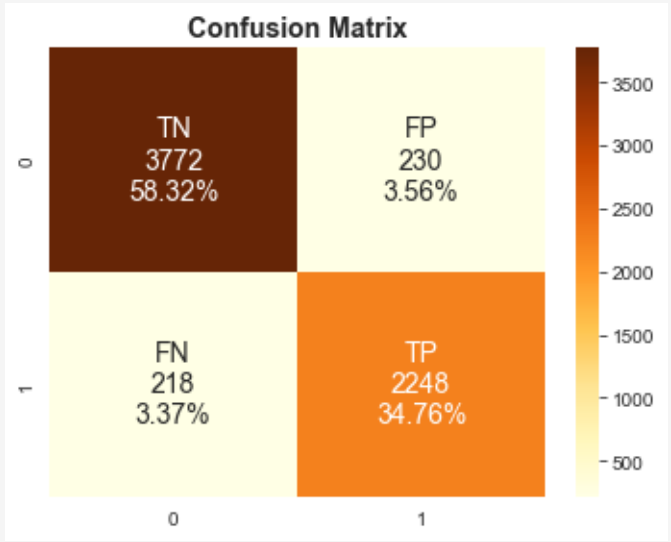
- Check correlation coefficients to find Top-15 correlation pairs among independent variables.
- Found the data is fit for building logistic regression model.
- Split the train set in X and y train.
- Selected features through feature ranking using Recursive Function Elimination(RFE).
- Manually eliminated independent variables on the basis of high p-value(>0.05) and high VIF(>5).
- Chose Model 5(glm5) as the final model, since it had all important statistics high , along with no insignificant variables and no multi collinear (high VIF) variables.

6 Prediction and Model Evaluation on Train Set-I

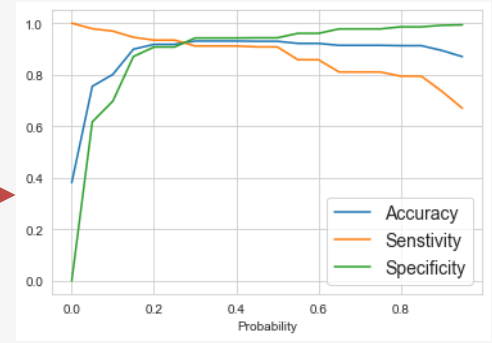
- Made predictions using Model 5(glm5) by setting cut-off as 0.5 and evaluated the model using various evaluation metrics.
- Found the optimal cut-off as 0.3 using various probabilities and estimated the final predicted values.
- Revaluated metrics for final predicted values and assigned lead score.
- Since the scores of evaluation metrics were as desired, decided to go ahead with the prediction on the test set.

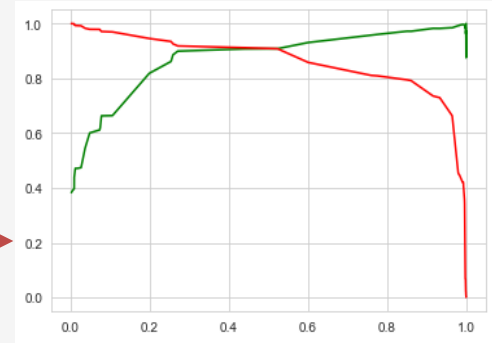
6 Prediction and Model Evaluation on Train Set-II

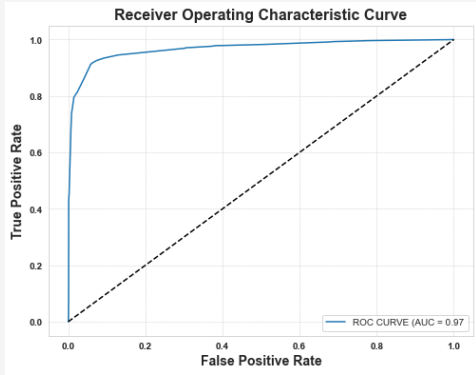
❖ Evaluation Metrics



	Score
Accuracy	92.981000
Senstivity	91.160000
Specificity	94.253000
FPR	5.747000
FNR	8.840000
Recall	90.794809
Precision	90.718000
AUC	96.805288





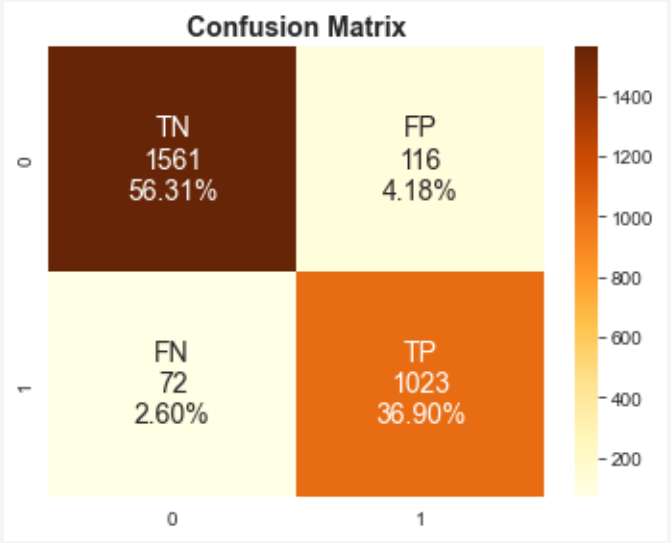


7 Prediction and Model Evaluation on Test Set-I

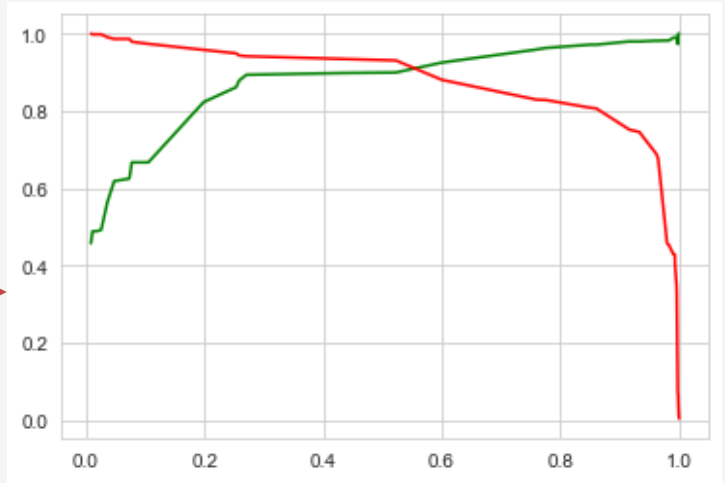
- Transformed the numeric features of the test data with Standard Scaler.
- Split the test data into X and y test and dropped features in X test to align with X train.
- Made predications on the test set using the final model (glm5).
- Set cut-off as 0.3(optimal cut-off found earlier) and evaluated the model using various evaluation metrics.
- The scores of evaluation metrics were found to as desired.

7 Prediction and Model Evaluation on Test Set-II

❖ Evaluation Metrics



	Score
Accuracy	93.218000
Senstivity	93.425000
Specificity	93.083000
FPR	6.917000
FNR	6.575000
Recall	93.424658
Precision	89.816000
AUC	97.316473



Variables impacting Lead Conversion Rate

Significant Variables	Coefficients
Lead Source_Welingak Website	2.815890
Last Activity_SMS Sent	2.229848
What is your current occupation_Unspecified	-2.545756
Tags_Busy	2.292987
Tags_Closed by Horizzon	9.743327
Tags_Lost to EINS	9.706759
Tags_Ringing	-1.468581
Tags_Untagged	3.691144
Tags_Will revert after reading the email	6.590905
Tags_switched off	-1.938484
Last Notable Activity_Modified	-1.486368

Recommendations

❖ To improve overall lead conversion rate, focus should be on:

- ✓ Improving lead conversion of Landing Page Submission and API and generating more leads from Lead Add Form.
- ✓ Improving lead conversion of Olark Chat, Organic Search, Direct Traffic and Google and generating more leads from Reference and Welingak Website.
- ✓ Generating more leads from Healthcare Management followed by Banking, Investment and Insurance.
- ✓ Generating more leads from Working Professionals.
- ✓ Generating more leads from Thane & Outskirts, Other Cities of Maharashtra and Other Metro Cities as they have a good conversion rate.
- ✓ Generating more leads from Closed by Horizzon or Lost to Competitors since they have good conversion rate.

❖ Additional Recommendations:

- ✓ The organization must do way with giving a free copy of "Mastering The Interview" to save cost as it has minimal effect on lead conversion.
- ✓ Leads spending more time on the website are more likely to be converted, hence the organization must work on building a more engaging website.

Thank You