Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving conversion rates for X Education.

TEAM MEMBERS – CHINMAY BISWAL & ASHISH PAWAR

BATCH - DS C60

Problem Statement

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. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, 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. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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. A typical lead conversion process can be represented using the following funnel.

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them 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 higher lead score have a higher conversion chance and the customers with 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%.

Analysis Approach

- **Data Cleaning**: Loading Data Set, understanding &cleaning data.
- **EDA** : Checking co-relation between numeric column, checking for outliers, Outlier treatment - capping & Categorical Analysis.
- **Data Preparation**: Dummy variables, test-train split & feature scaling.
- **Model Building**: RFE for top feature, Manual Feature Reduction & finalizing model.
- **Model Evaluation**: Confusion matrix, Cut-off Selection, ROC Curve, Finding Optimal Cut off Point, assigning Lead, Score.
- **Predictions on Test Data**: Compare train vs test metrics, Assign Lead Score and get top features,
- **Recommendation**: Suggest top features to focus for higher conversion & areas for improvement

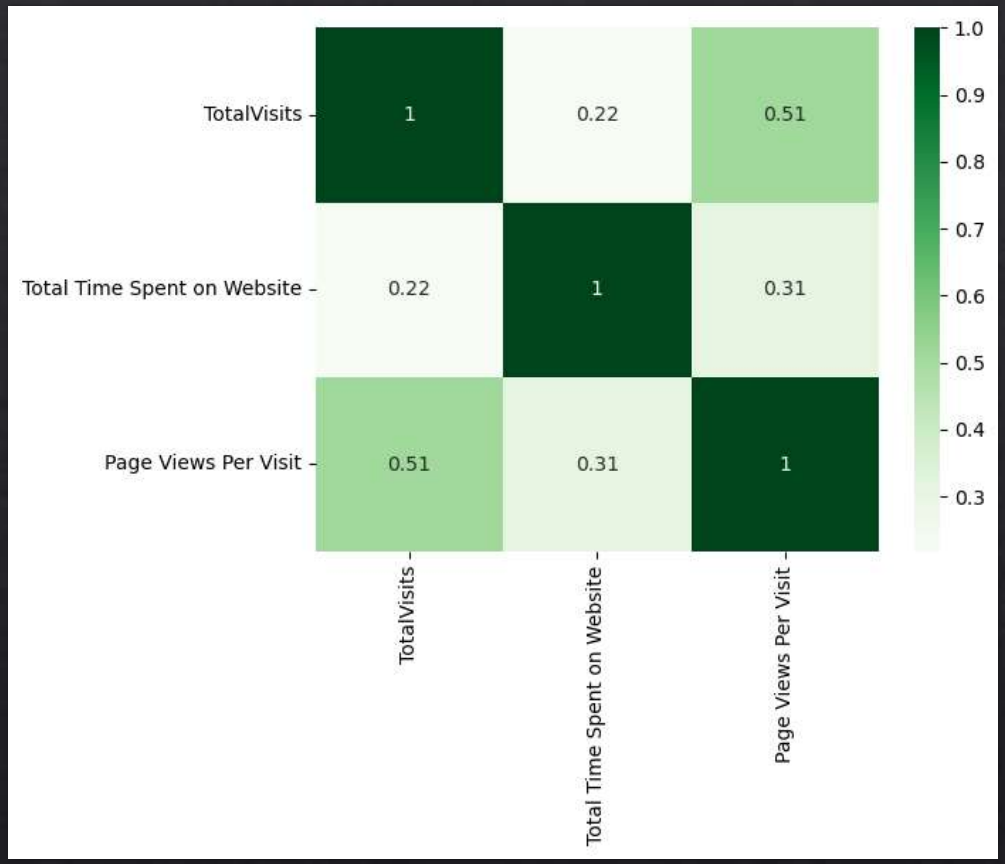
Data Cleaning

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Check the percentage of Null Values and dropping columns with more than 40% null values
- Impute numerical columns with median and categorical columns with mode in place of Null Values
- *Check unique value counts*
- *Check the value counts in each column for the data distribution and data imbalance*

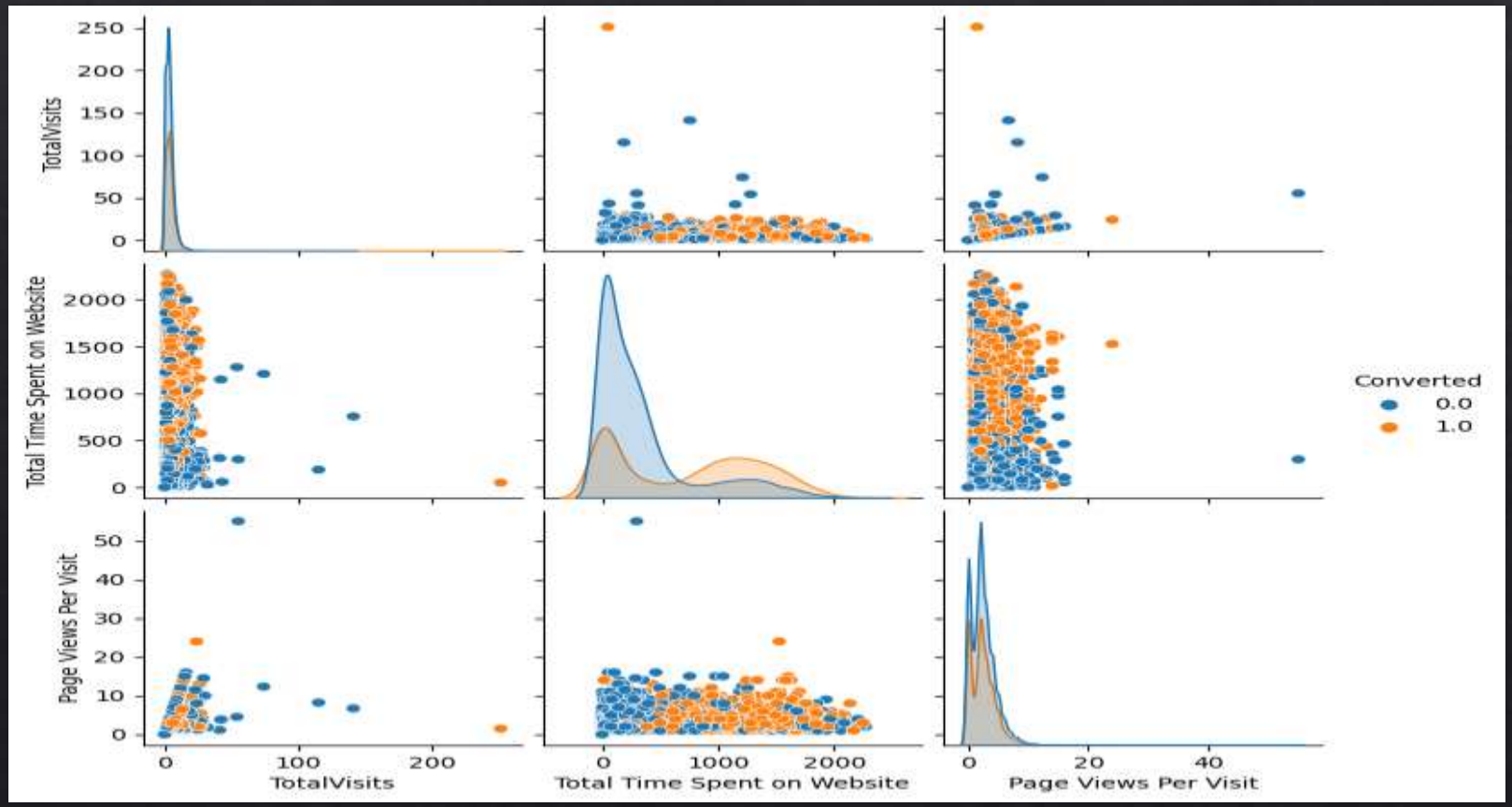
EDA

- **Co-relation between the numeric columns**: Checking the co-relation using Heat map and Pair-plot.
- **Check for outliers** : We can see that there are few outliers in "TotalVisits" and "Page Views Per Visit" columns. There are no outliers in "Total Time Spent on Website".
- **Outlier treatment** - capping : Removed the outliers of those columns.
- **Categorical Analysis**: Maximum number of leads are generated by Google and Direct traffic, API and Landing Page Submission bring higher number of leads as well as conversion.

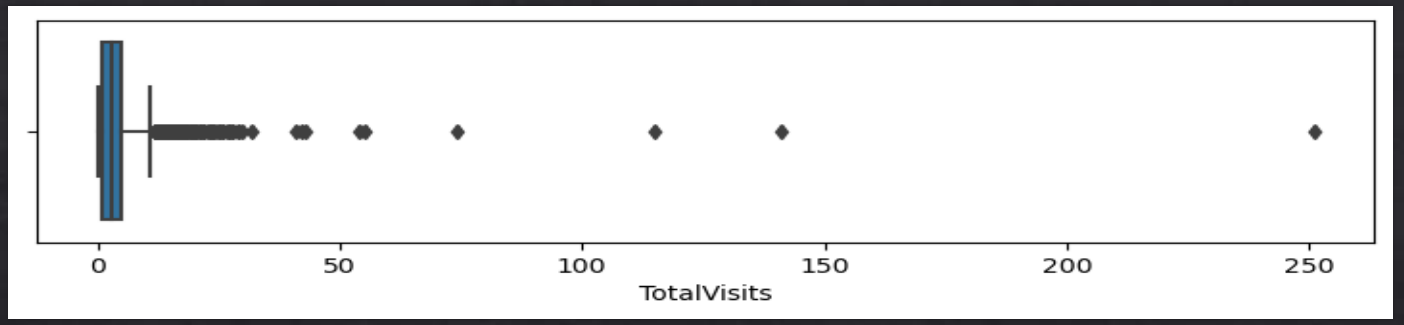
EDA - Co-relation between the numeric columns using Heat Map.

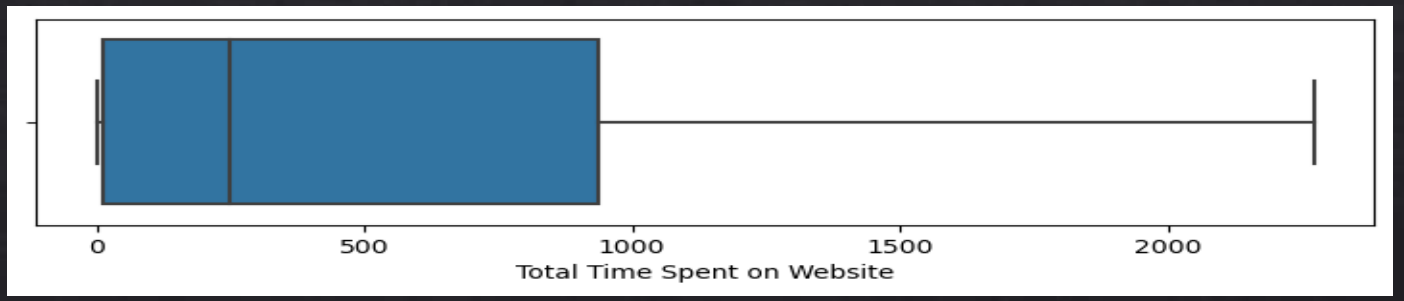


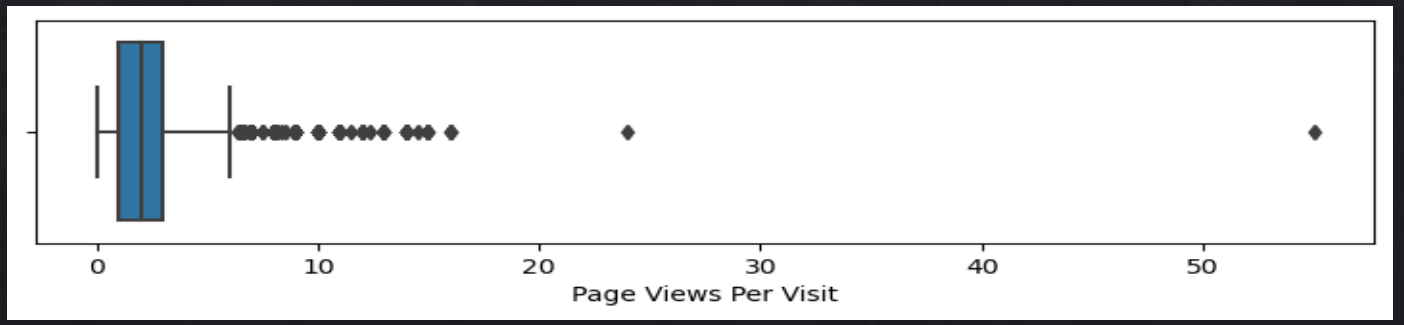
EDA - Co-relation between the numeric columns using pairplot.



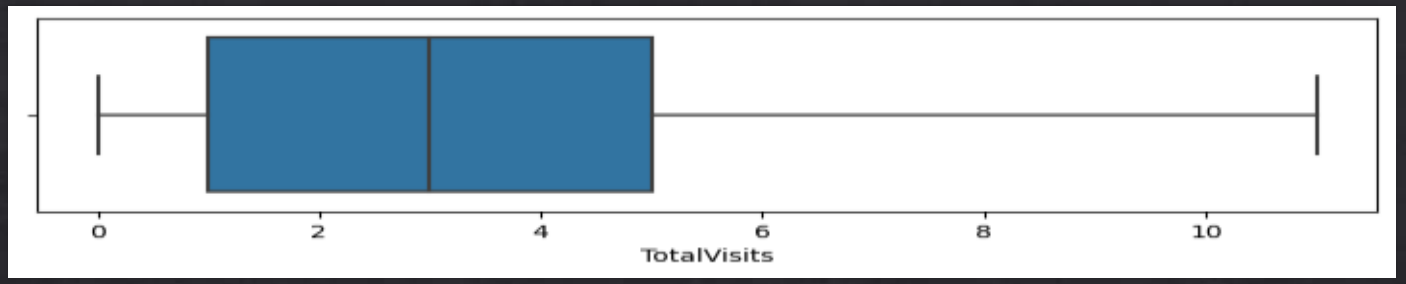
EDA - Check for outliers

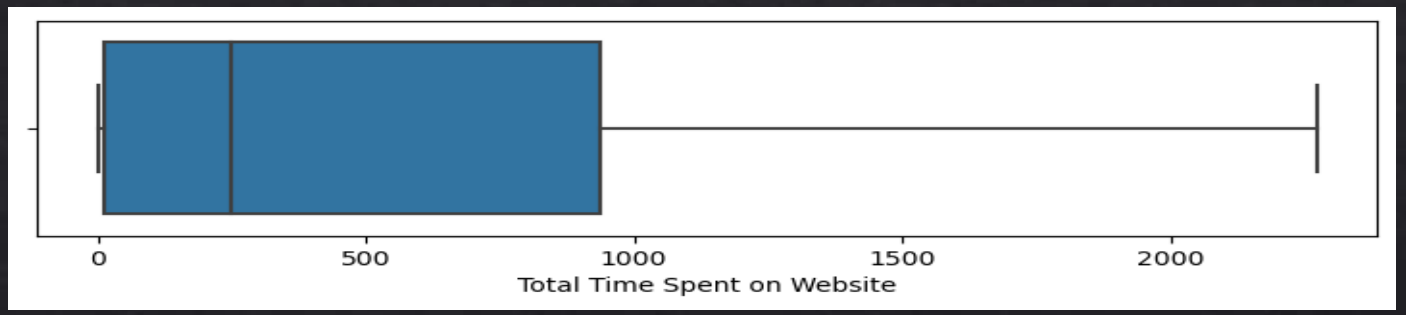


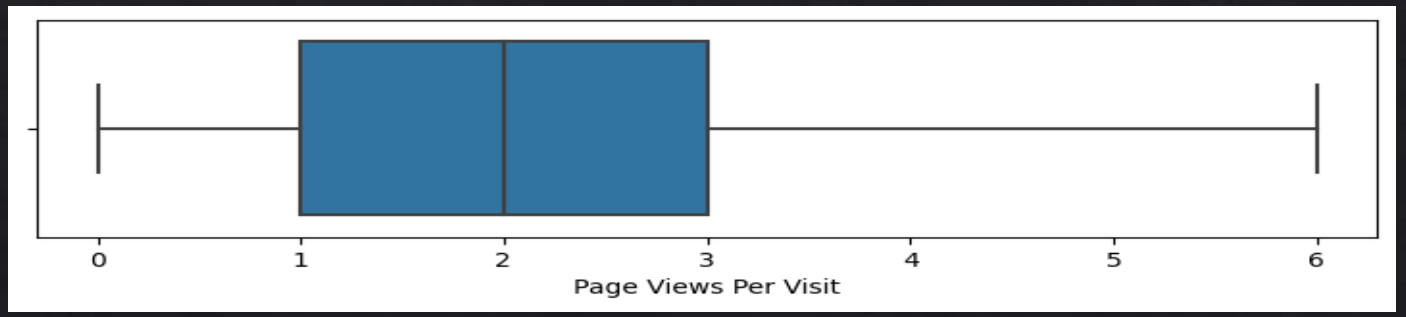




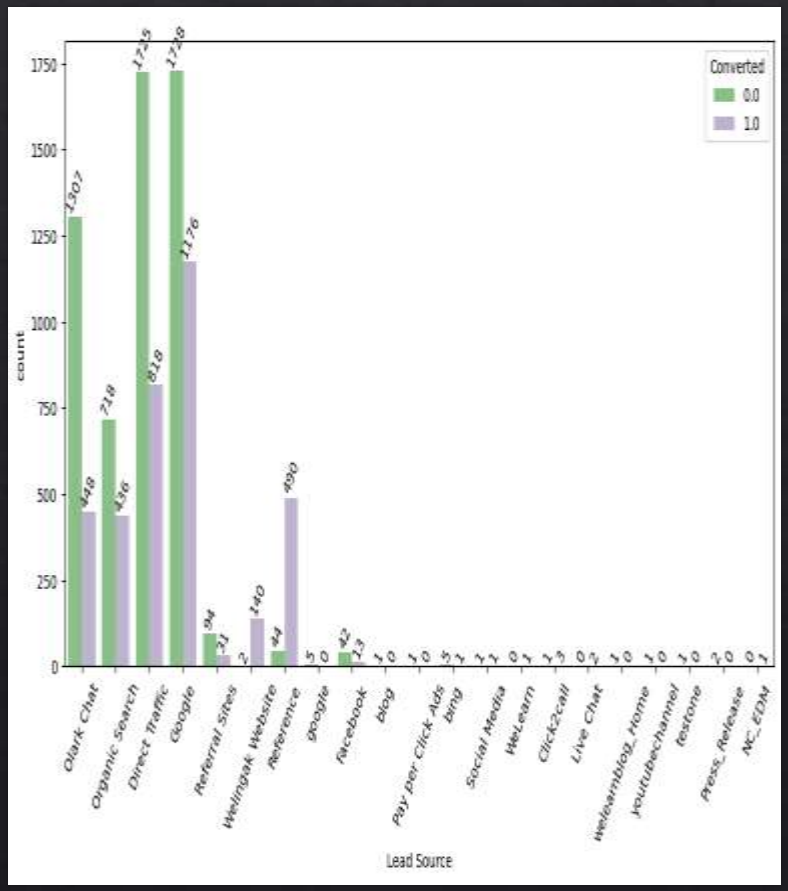
EDA - Outlier treatment - capping

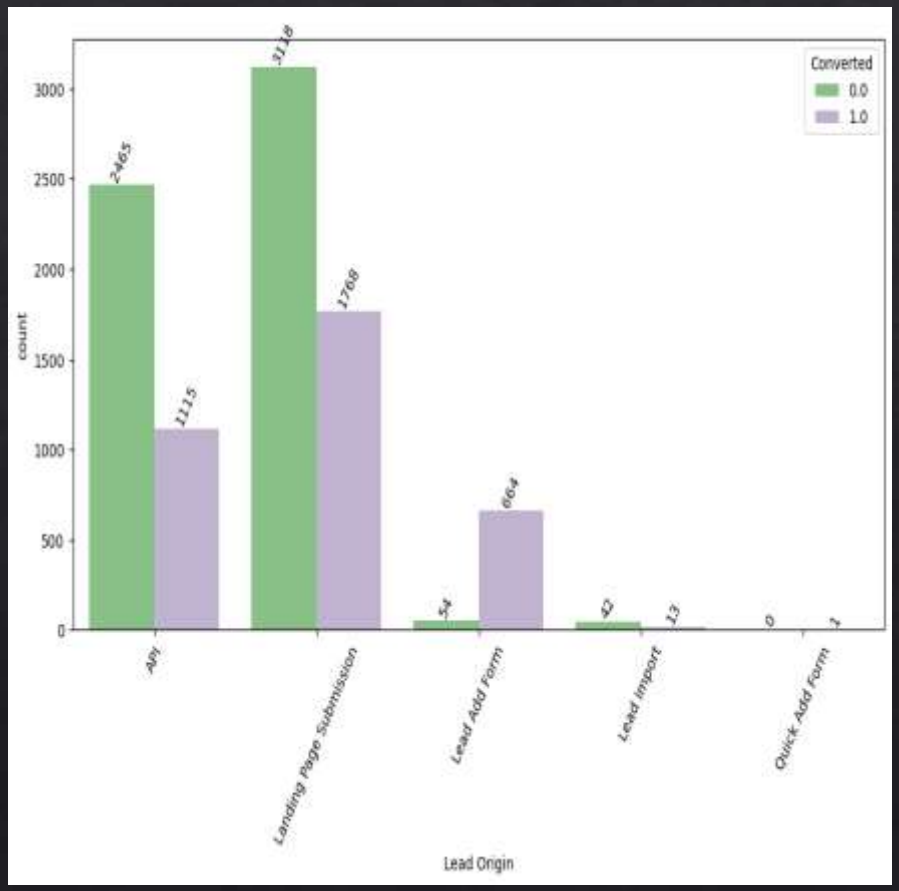






EDA – Categorical Analysis





Model Building

Feature Selection

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only theimportant columns.
- Then we can manually fine tune the model.
- RFE outcome:

Pre RFE – 30 columns & Post RFE – 18 columns

Model Building

Feature Selection

- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 13 looks stable after four iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
- Hence, **Model 13** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Model Evaluation

It was decided to go ahead with 0.299 as cutoff after checking evaluation metrics coming from both plots

```
**************************************************

Confusion Matrix
[[3490  512]
 [ 319 2147]]

**************************************************

True Negative                   :  3490
True Positive                   :  2147
False Negative                  :  319
False Positve                   :  512
Model Accuracy                  :  0.8715
Model Sensitivity               :  0.8706
Model Specificity               :  0.8721
Model Precision                 :  0.8074
Model Recall                    :  0.8706
Model True Positive Rate (TPR)  :  0.8706
Model False Positive Rate (FPR) :  0.1279

**************************************************
```

Confusion Matrix & Evaluation Metrics with 0.299 as cut-off

```
**************************************************

Confusion Matrix
[[3602  400]
 [ 402 2064]]

**************************************************

True Negative                   :  3602
True Positive                   :  2064
False Negative                  :  402
False Positve                   :  400
Model Accuracy                  :  0.876
Model Sensitivity               :  0.837
Model Specificity               :  0.9
Model Precision                 :  0.8377
Model Recall                    :  0.837
Model True Positive Rate (TPR)  :  0.837
Model False Positive Rate (FPR) :  0.1

**************************************************
```

Confusion Matrix & Evaluation Metrics with 0.42 as cut-off

Predictions on Test Data

Comparing train vs test metrics

```
----------------------------------------

Confusion Matrix
[[3490  512]
 [ 319 2147]]

----------------------------------------
True Negative          : 3490
True Positive          : 2147
False Negative         : 319
False Positve          : 512
Model Accuracy         : 0.8715
Model Sensitivity      : 0.8706
Model Specificity      : 0.8721
Model Precision        : 0.8074
Model Recall           : 0.8706
Model True Positive Rate (TPR)  : 0.8706
Model False Positive Rate (FPR) : 0.1279

----------------------------------------
```

Observation for Train set

```
----------------------------------------

Confusion Matrix

([1450,227]
[129,966])

----------------------------------------

- True Negative          : 1450
- True Positive          : 966
- False Negative         : 129
- False Positve          : 227
- Model Accuracy         : 0.8715
- Model Sensitivity      : 0.8821
- Model Specificity      : 0.8646
- Model Precision        : 0.8097
- Model Recall           : 0.8821
- Model True Positive Rate (TPR)  : 0.8821
- Model False Positive Rate (FPR) : 0.1353

----------------------------------------
```

Observation for Test set

Recommendation based on Final Model

- The company should make calls to the leads coming from the `Tags_Closed by Horizzon` and `Tags_Lost to EINS` as these are more likely to get converted.
- The company should make calls to the leads coming from the `Last Notable Activity_Had a Phone Conversation`, `Lead Origin_Lead Add Form `, `Last Notable Activity_Email Bounced ` and `Last Notable Activity_SMS Sent` as these are more likely to get converted.
- The company should make calls to the leads who are the `working professionals` as they are more likely to get converted.
- The company should make calls to the leads coming from `Lead Source_Welingak Website` and who spent `more time on the websites` as these are more likely to get converted.

- The company should not make calls to the leads whose `last activity` was `Olark Chat Conversation`, `Converted to Lead` and `Email Bounced` as they are not likely to get converted.
- The company should not make calls to the leads whose `Tags` is `Graduation in progress`, `switched off`, `Ringing`, `invalid number` as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of `"Do not Email" as "yes"` as they are not likely to get converted.