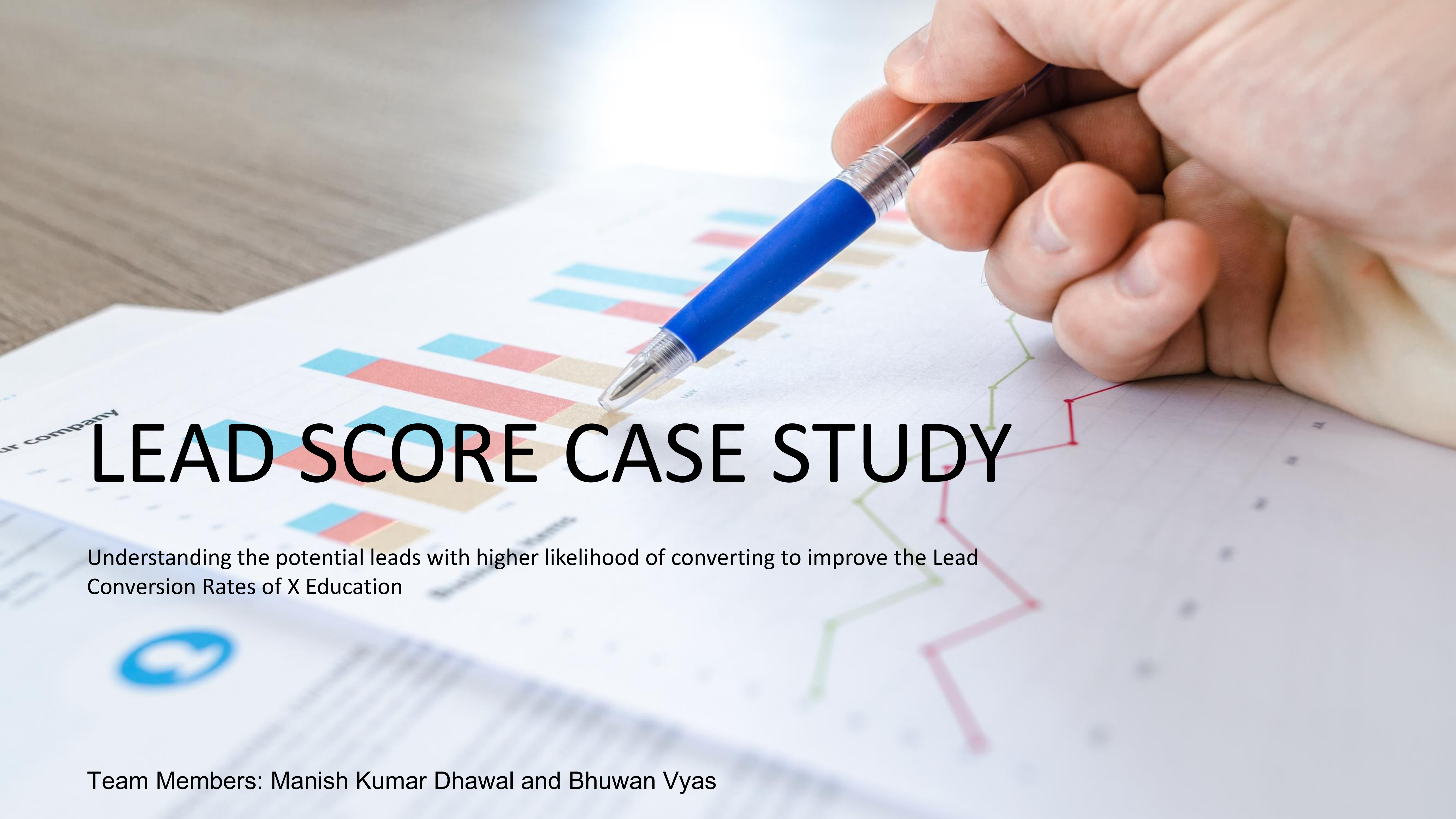


LEAD SCORE CASE STUDY

Understanding the potential leads with higher likelihood of converting to improve the Lead Conversion Rates of X Education

Team Members: Manish Kumar Dhawal and Bhuwan Vyas

Index:

- Problem Statement and Goals of the Case Study
- Suggested Ideas for Lead Conversion
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE and model optimization)
- Model Evaluation
- Insights and Conclusion

Problem Statement and Goal of the Case Study

Problem Statement:

- The company gets a lots of leads but the conversion rate is very poor
- The company wants to make this leads identification process more efficient to generate more 'Hot Leads'
- Their sales team wants to understand these set of potential clients on whom they will focus more rather than making useless calls

Goals of the Case Study:

- To help the company in identifying the most potential leads because their chances of converting are high
- The company requires us to build a model wherein we 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 lower lead score have a lower conversion chance
- The CEO has given a threshold of the target lead conversion to be around 80%

Suggested Ideas for Lead Conversion

- Leads are groups based on their likelihood of getting converted
- This results in focus on groups with potential leads

Leads Grouping

- We could have a smaller pool of leads to communicate with, which would allow us to target potential leads in an efficient way

Better Communication

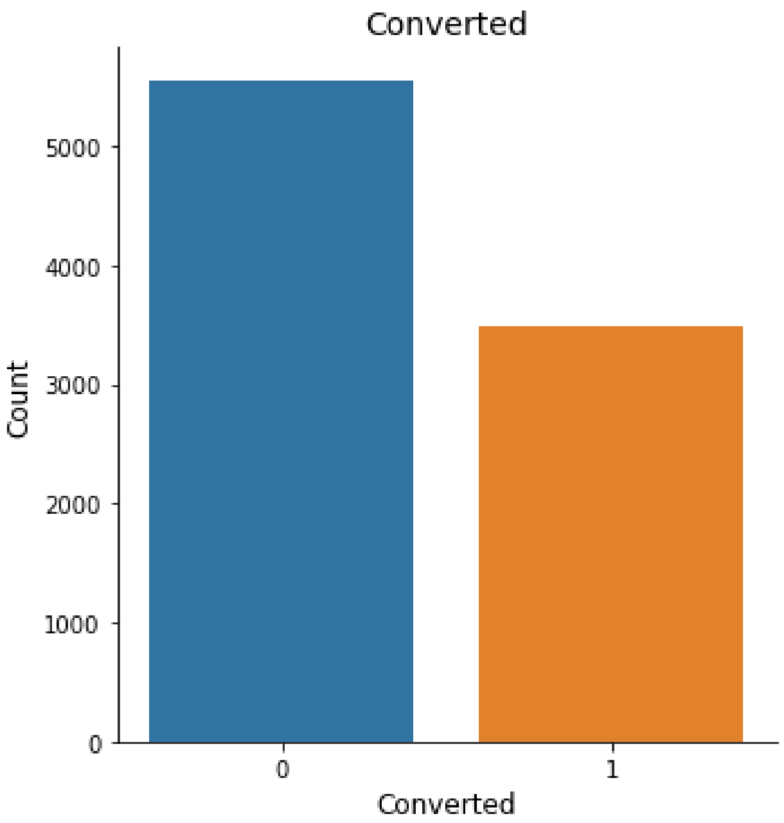
- By following a proper strategy, we could cross the threshold of 80% as proposed by the CEO and concentrate on other things to improve the model and client relationships

Boost Conversions

Data Cleaning

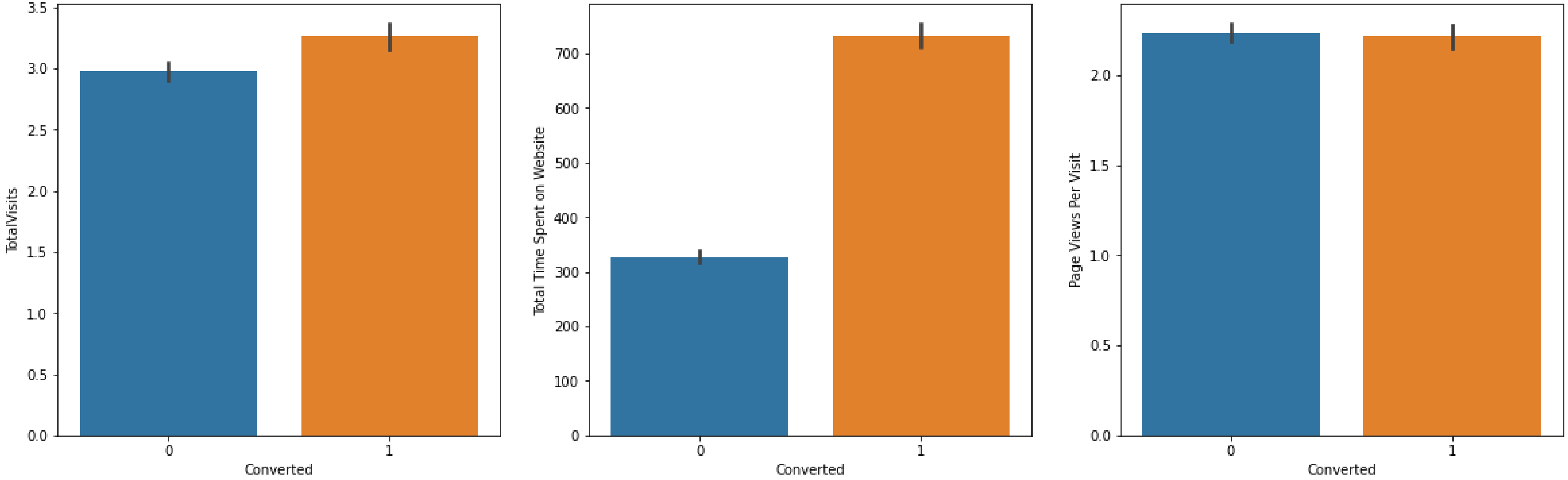
- Data is first checked for any duplicated or null values. We can notice no duplicate values but we do have several columns with missing values.
- There another value named ‘Select’ which represents missing value but considered as an input. It should be replaced with NaN, in whichever column it is present.
- Now, columns with 35% or more missing values must be dropped as it won’t be feasible to impute values in them because they won’t add any value during model building.
- Further dropping columns with highly skewed values and columns which don’t add any meaningful value to the model building 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’, etc.
- We will create additional categories within columns where one variable has majority of the values and rest of the variables has some or no values. For example ‘Lead score’ and ‘Last Activity’.
- Numerical data was imputed with mode method.
- Outliers in numerical columns are capped.

EDA



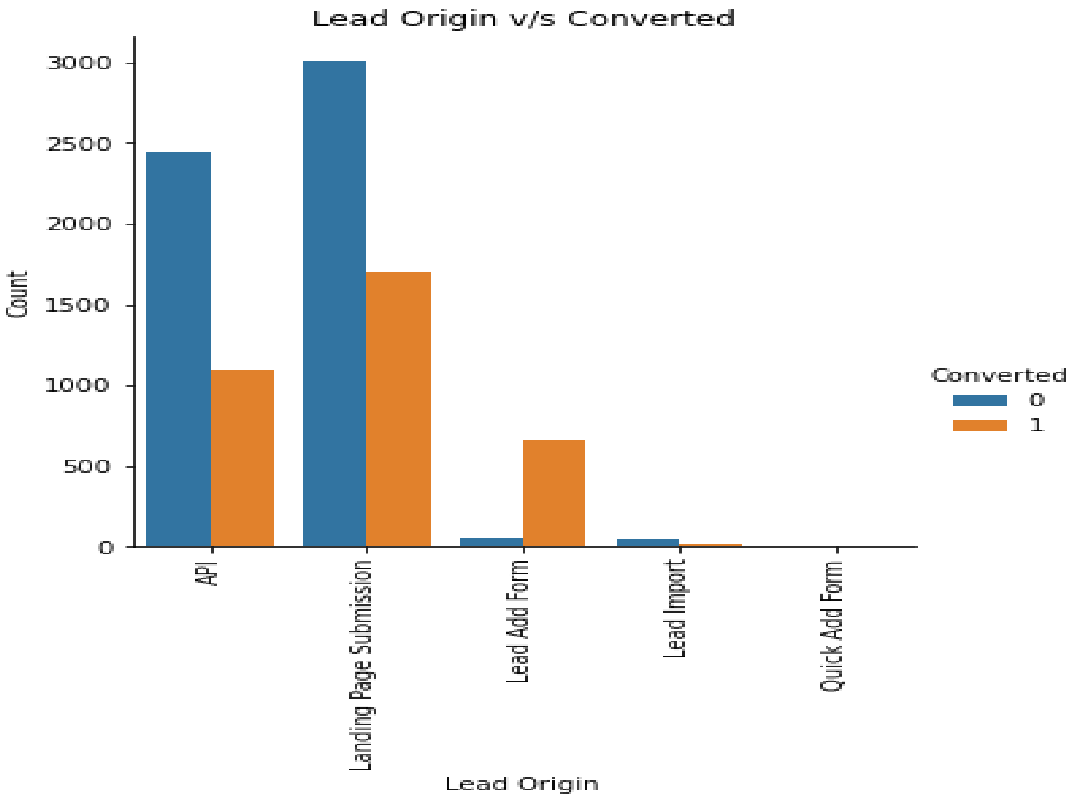
- We can see the data is quite imbalance
- Conversion rate is around 38.5%, which means that only 38.5% potential clients have converted.
- While 61.5% of the potential clients did not convert.

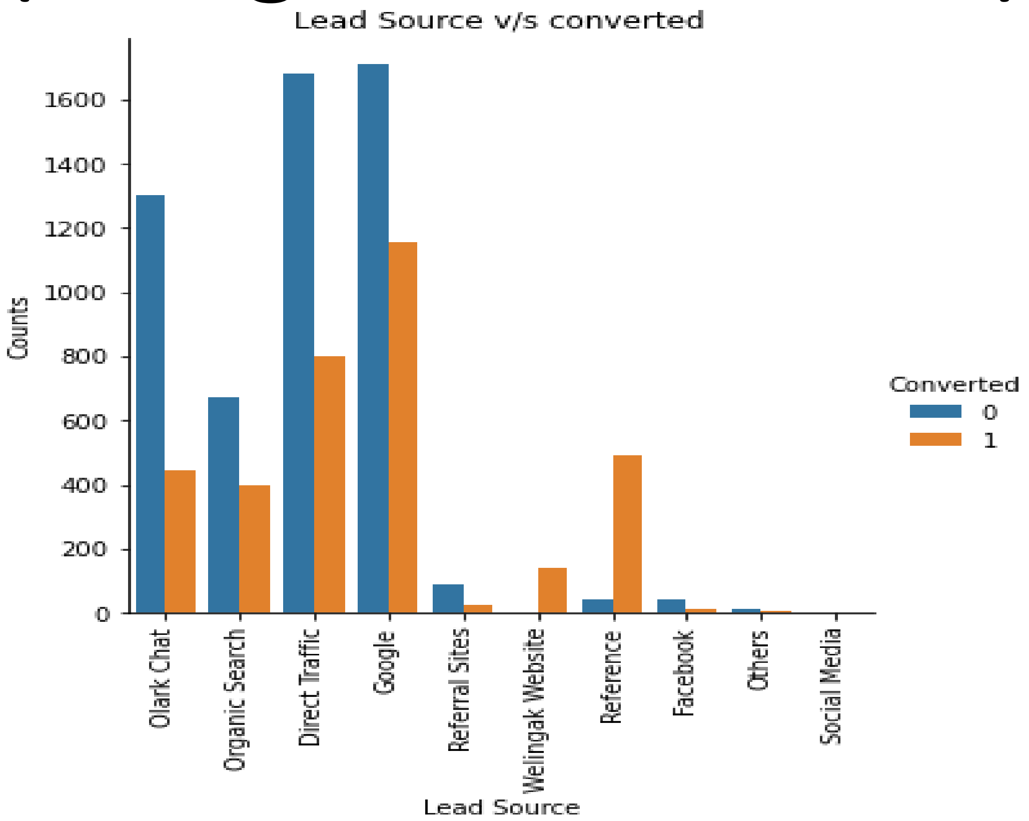
EDA – Bivariate Analysis of Numerical Variables



- Past leads who **spends more time on the Website** have a higher chance of getting converted than those who spends less time as seen in the above **barplots.**

EDA – Bivariate Analysis (Categorical Variables)

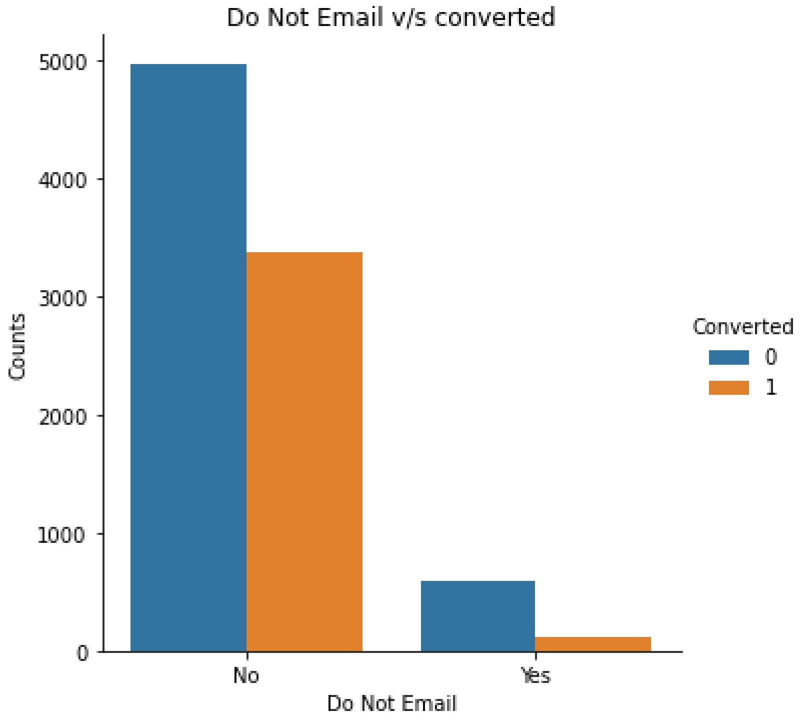


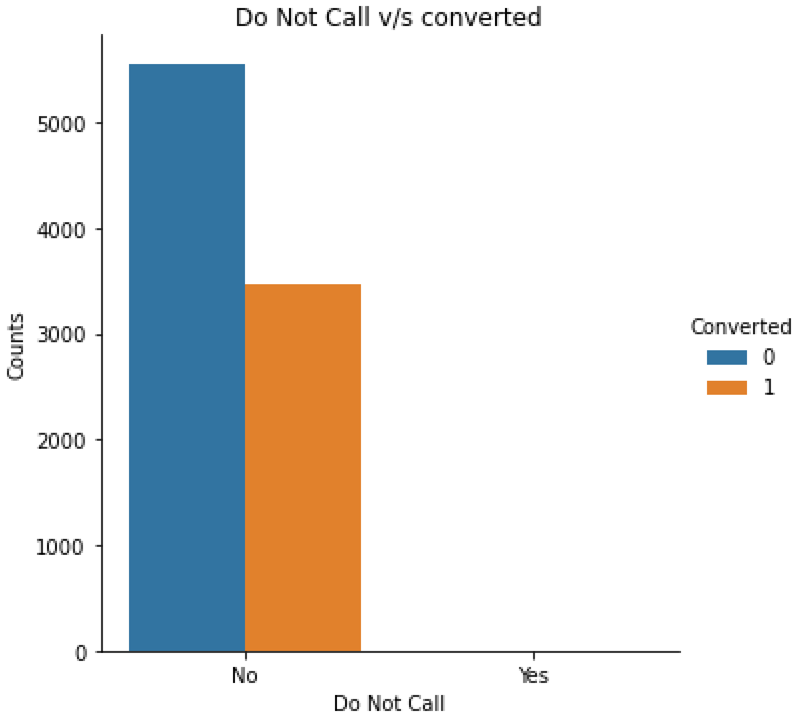


- Lead Origin has major leads coming from ‘Landing Page Submission and followed by ‘API’

- Lead Source has 58% of leads coming from ‘Google’ and Direct Traffic
- We can also see that ‘Google’ has created a majority class, hence we can club all the remaining minority classes into a new separate variable ‘Others’

EDA – Bivariate Analysis (Categorical Variables)

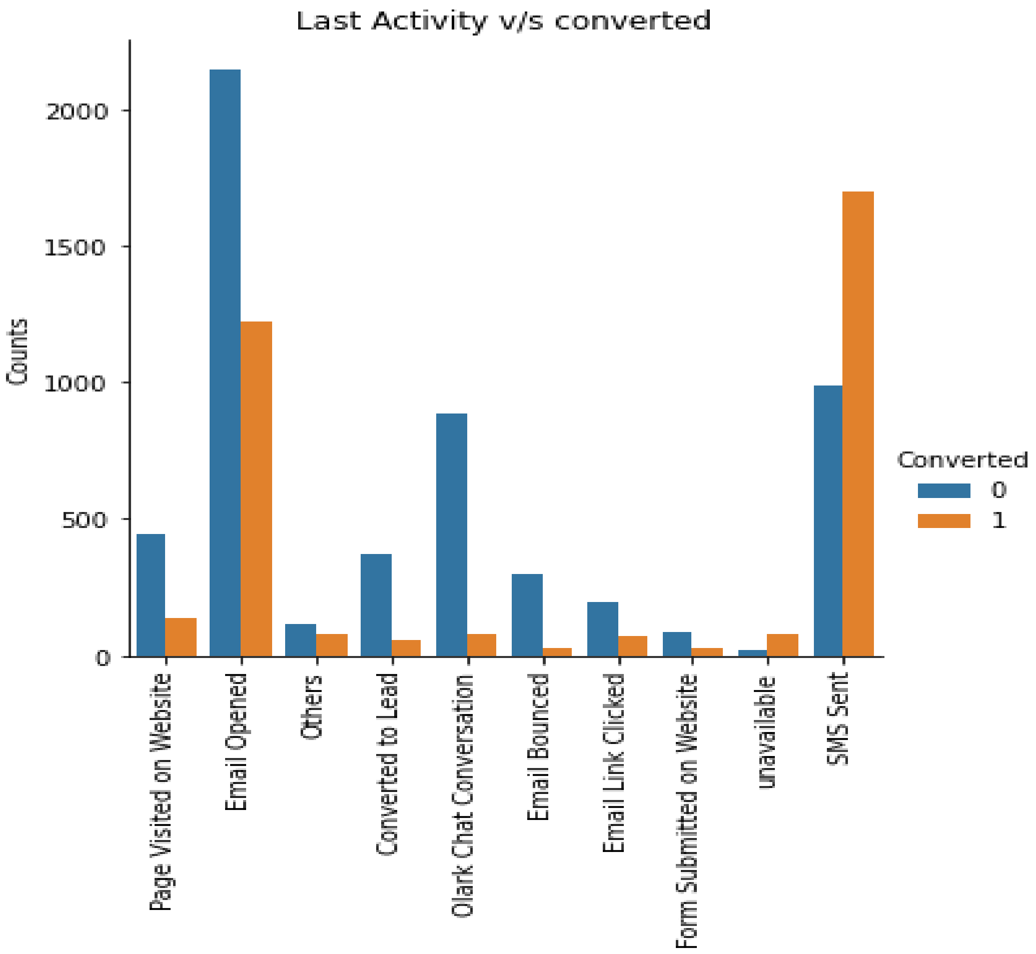




- Do Not Email: We can notice that majority of leads come when Email was sent.

- Do Not Call: We can notice that this data is highly skewed so we can drop this column.

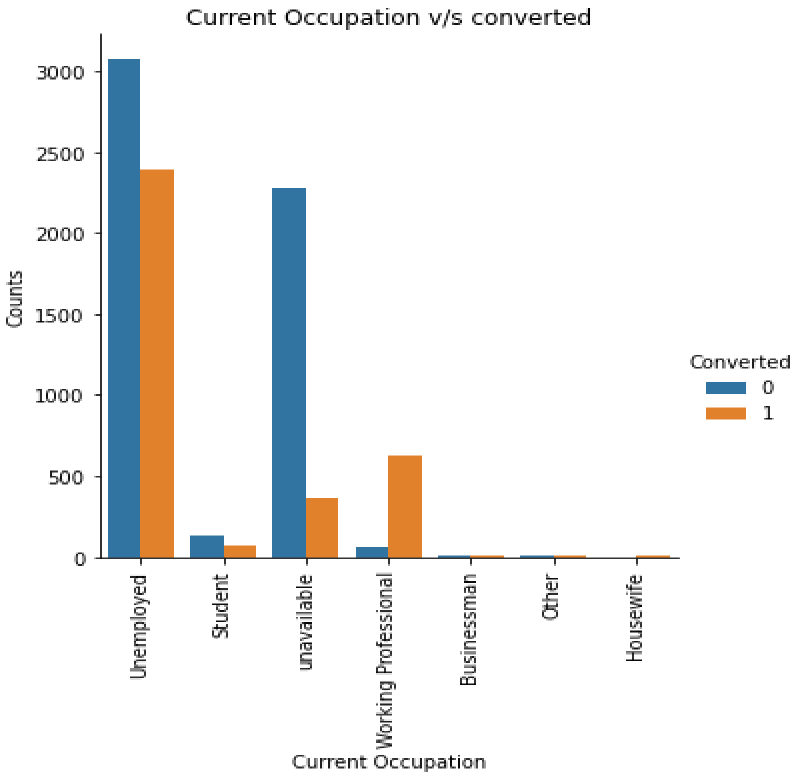
EDA – Bivariate Analysis (Categorical Variables)

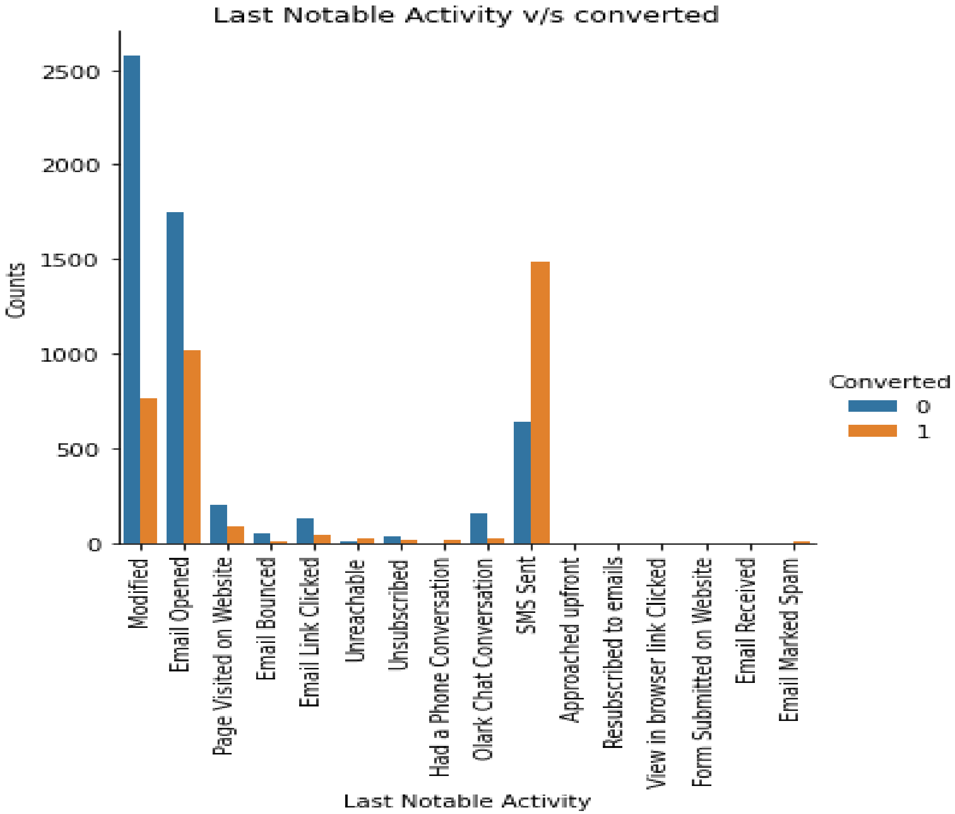


Last Activity:

- We can notice that the conversion rate is highest when ‘SMS Sent’.
- This is followed by ‘Email opened’.
- We can also recall that this is a sales team generated data, so we can drop this.

EDA – Bivariate Analysis (Categorical Variables)





- Current Occupation: We can see that majority of the leads are coming from 'Unemployed', which is followed by 'Working Professional'

- Last Notable Activity: We can observe that majority of the leads are coming from 'SMS Sent'.
- We can also recall that this is also a sales generated data, so we can choose not to use it further

Data Preparation before Model Building

- We have to convert the binary variables from Yes/No to (0/1)
- Create dummy features for variables ‘Lead Origin’, ‘Lead Source’ and ‘What is your Current Occupation’
- Splitting the data into Train and Test sets
- 70:30 ratio was chosen to create the split
- For Feature Scaling we have used MinMaxScaler()
- We have also checked for highly correlated dummy variables with the help of heatmap
- After finding the correlation, we proceed to drop columns ‘Lead Origin_API’ and ‘Lead Origin_Landing Page Submission’

Model Building

Feature Selection

- The data has lots of columns and features
- This might hinder the model evaluation process and also increase the computation time
- Thus, it is important to do **Recursive Feature Elimination** (RFE) and only selecting the important columns
- After that, we can manually refine the model
- Outcome of RFE:
  - Columns before RFE: 22
  - Columns after RFE: 15

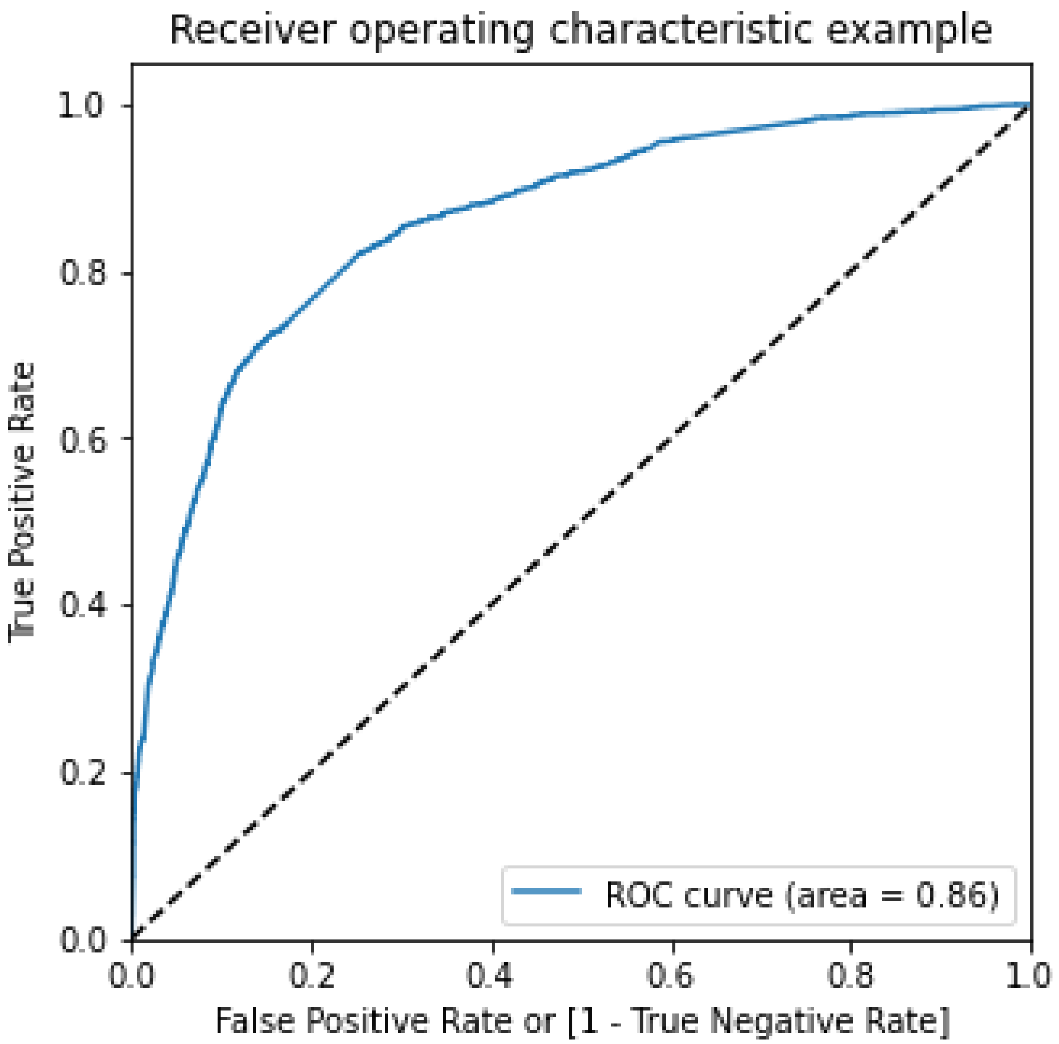
Model Building

- Now we will manually refine the model by iterating a process
- We will remove variables with high p-value and refine the model again
- This process will continue until the p-value is under 0.05 and there is no multicollinearity and VIF values is also under 5
- For us, Model 7 seems to have a significantly low p-value and low VIF value as well.
- Hence, Model 7 will be our final model and we will go ahead with model for evaluation to make predictions on train and test sets

Model Evaluation

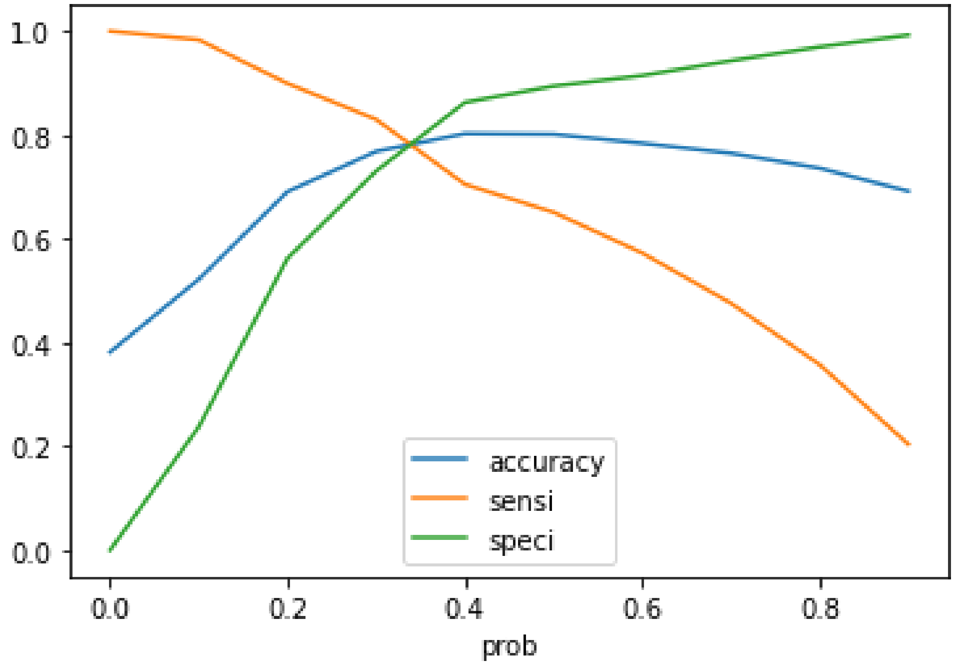
ROC Curve -Train Data Set

- We can notice that the area under the ROC Curve is 0.86 out of 1, which is a sign of a good predictive model.
- The curve is as close to the top left corner of the plot and this represent a model with high true positive rate and a low false positive rate at all thresholds.



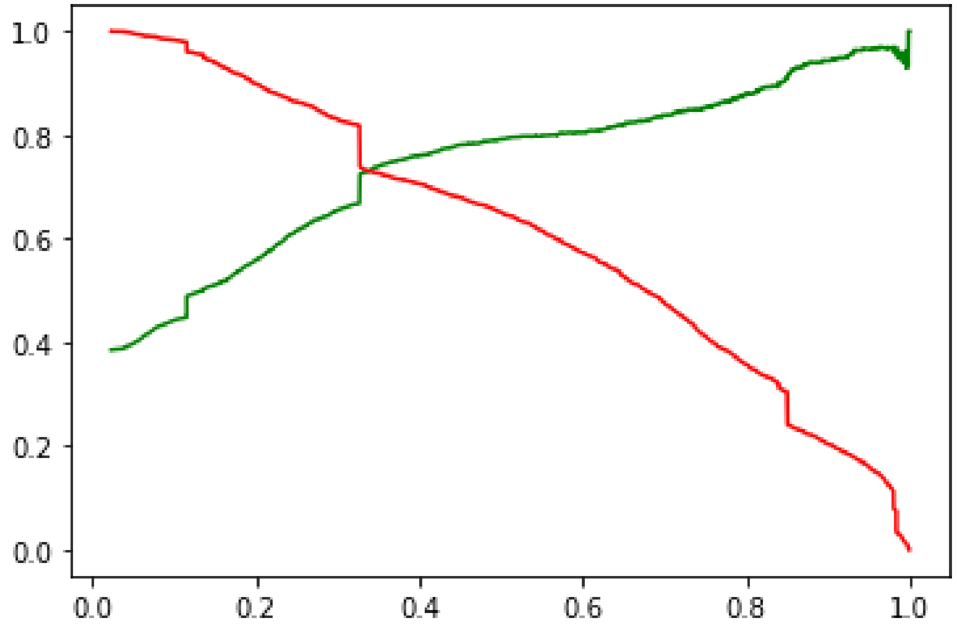
Model Evaluation

Optimal Cut-Off (Train Data Set)



- We can observe the optimal cut-off point to be around **0.3** and we can take this as the cut off probability

Precision and Recall Trade-off (Train Data Set)

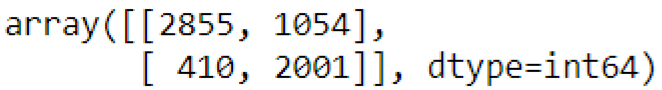


- We can observe the precision- recall trade-off point to be around 0.3 again.

Model Evaluation

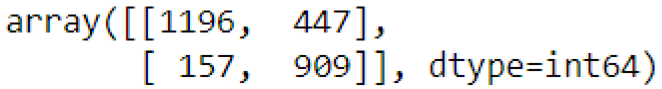
Confusion Matrix

Train Data Set



```
array([[2855, 1054],
       [ 410, 2001]], dtype=int64)
```

Test Data Set



```
array([[1196,  447],
       [ 157,  909]], dtype=int64)
```

Other Metrics

- Using the cut-off probability (0.3), we have achieved a sensitivity of 83% in the Train Set and 85.27% in the Test Set.
- Here sensitivity defines how many leads that the model identifies correctly out of all the leads that are converting.
- The ballpark set by the X Education’s CEO was a **sensitivity of 80%.**
- The model has also achieved an accuracy of 77% which is in line with the Case Studies objectives.

Insights and Conclusion

- As given in the problem statement, we had to develop a regression model that can help us identify the most crucial factors that would affect the lead conversions, which will help X Education to grow.
- Through the model, we have outlined the following features which has the highest positive coefficients and they can be prioritized to increase lead conversion rates:
  - Total Time Spent on Website: 4.5220
  - Lead Origin_Lead Add Form: 3.5797
  - Current Occupation_Working Professional: 2.3810
  - Lead Source_Welingak Website: 2.1843
  - Lead Source_Olark Chat: 1.1205
  - TotalVisits: 0.9084

- We have also Identified features that have a high negative coefficient and these could be potential areas for improvements:
  - Do Not Email: -1.1349
  - Current Occupation_Unavailable: -1.3174
  - Lead Source_Direct Traffic: -0.2601

Insights and Conclusion

- Focus on features that have a positive coefficient for targeted marketing strategy
- Prioritize leads coming from top-performing lead sources to curate high quality leads
- Use a balanced approach by utilizing communication channels like SMS and Email as they have a higher LCR
- Incentivize and create custom offers for Working Professionals as they have a higher chance of getting converted

Areas for Improvement:

- Analyze variables with high negative coefficients as they can help in creating better models
- Review emailing process for improved catering of potential clients