LEAD SCORING CASE STUDY

By – Vishnu Priya S

PROBLEM STATEMENT

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have a lower conversion chance.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion.
- Identify the outliers, if any, in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- 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 a lower lead score have a lower conversion chance.
- The CEO 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: Check imbalance, Univariate & Bivariate analysis

Data Preparation: Dummy variables, test-train split, feature scaling

Model Building: RFE for top 15 feature, Manual Feature Reduction & finalizing model

Model Evaluation: Confusion matrix, Cutoff Selection, assigning Lead Score

Predictions on Test Data: Compare train vs test metrics, Assign Lead Score and get top features

Recommendation: Suggest top 3 features to focus for higher conversion & areas for improvement

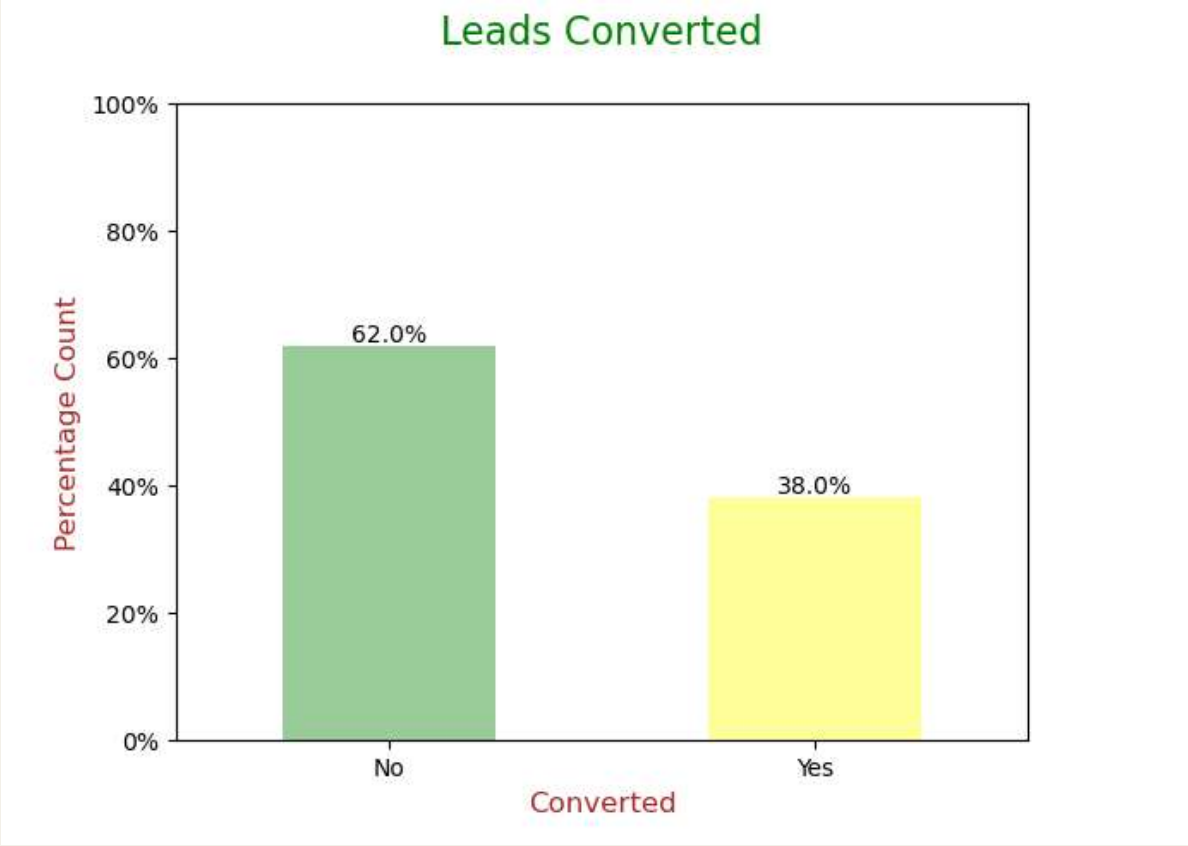
Data Cleaning

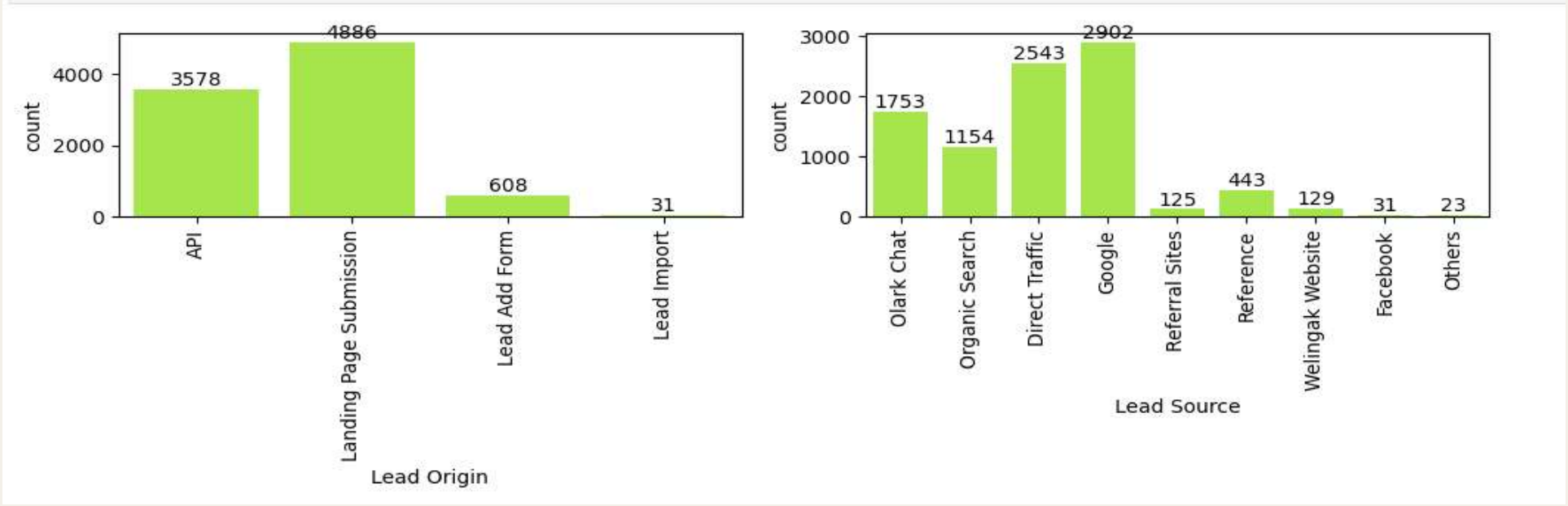
- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 35% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Imputation was used for some categorical variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in TotalVisits and Page Views Per Visit were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped

EDA

DATA IMBALANCE :

Target Variable has Data imbalnce of only about 38.5% Lead Converted Score





Univariant - Categorical Columns

Contribution of Various Variables are found:

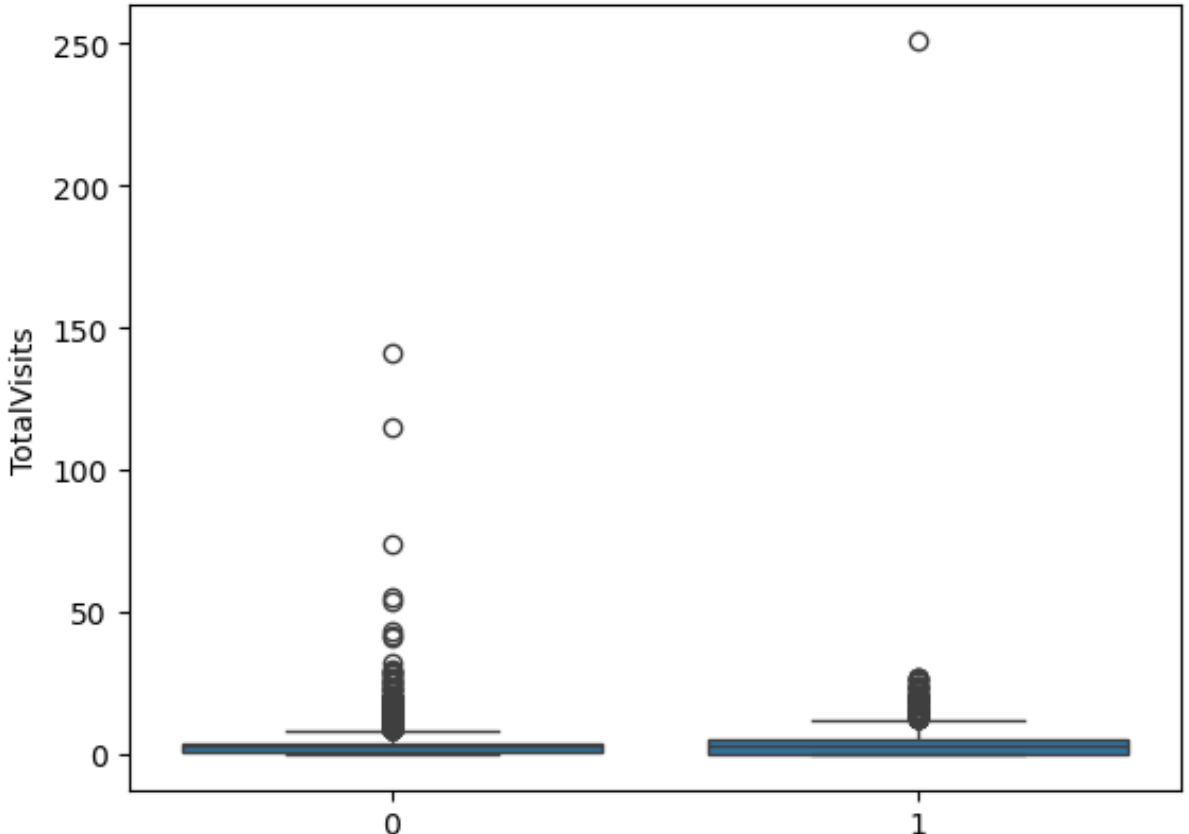
Lead Origin : Landing page submission has the highest contribution.

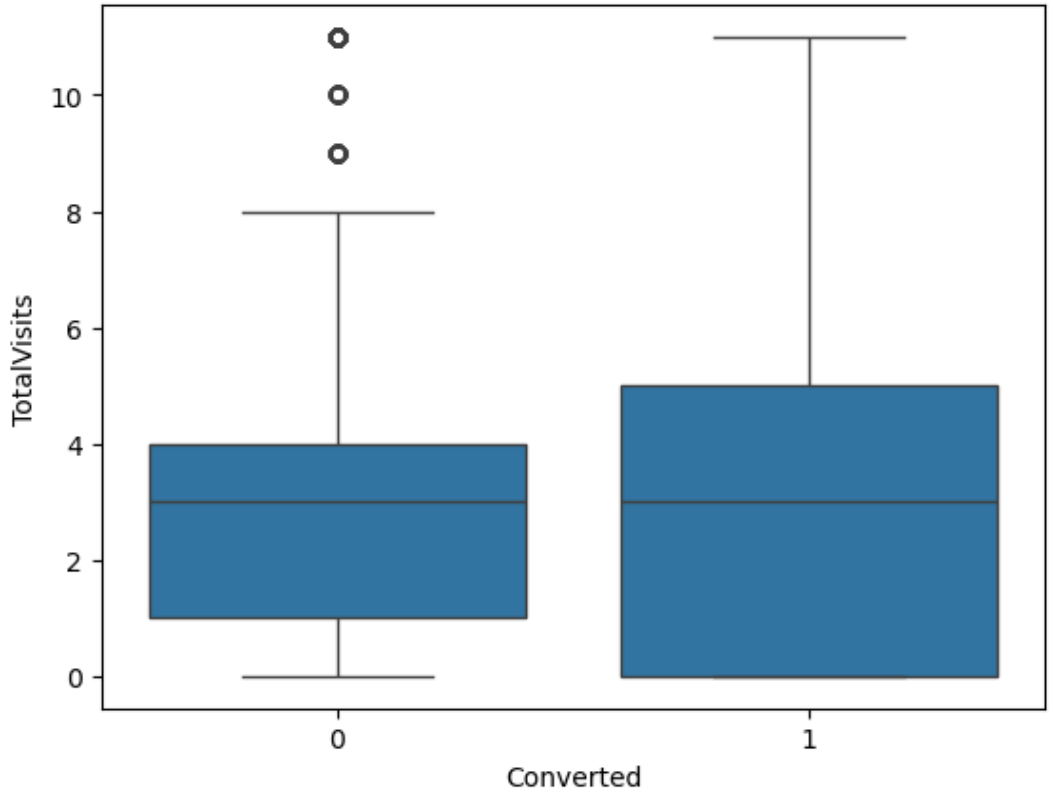
Lead Source : Google and Direct Traffic has highest Contribution.

Columns like Do Not Call, Search, Newspaper Article, X education Forums, NewsPaper, Digital Advertiment, Through Recommendations has negligible Contribution towards Leads so are not useful for Further analysis and Model building thus **can be dropped.**

Univariant – Numerical columns

Outliners are found and handled using IQR in columns Total Visits and Page Views per Visit

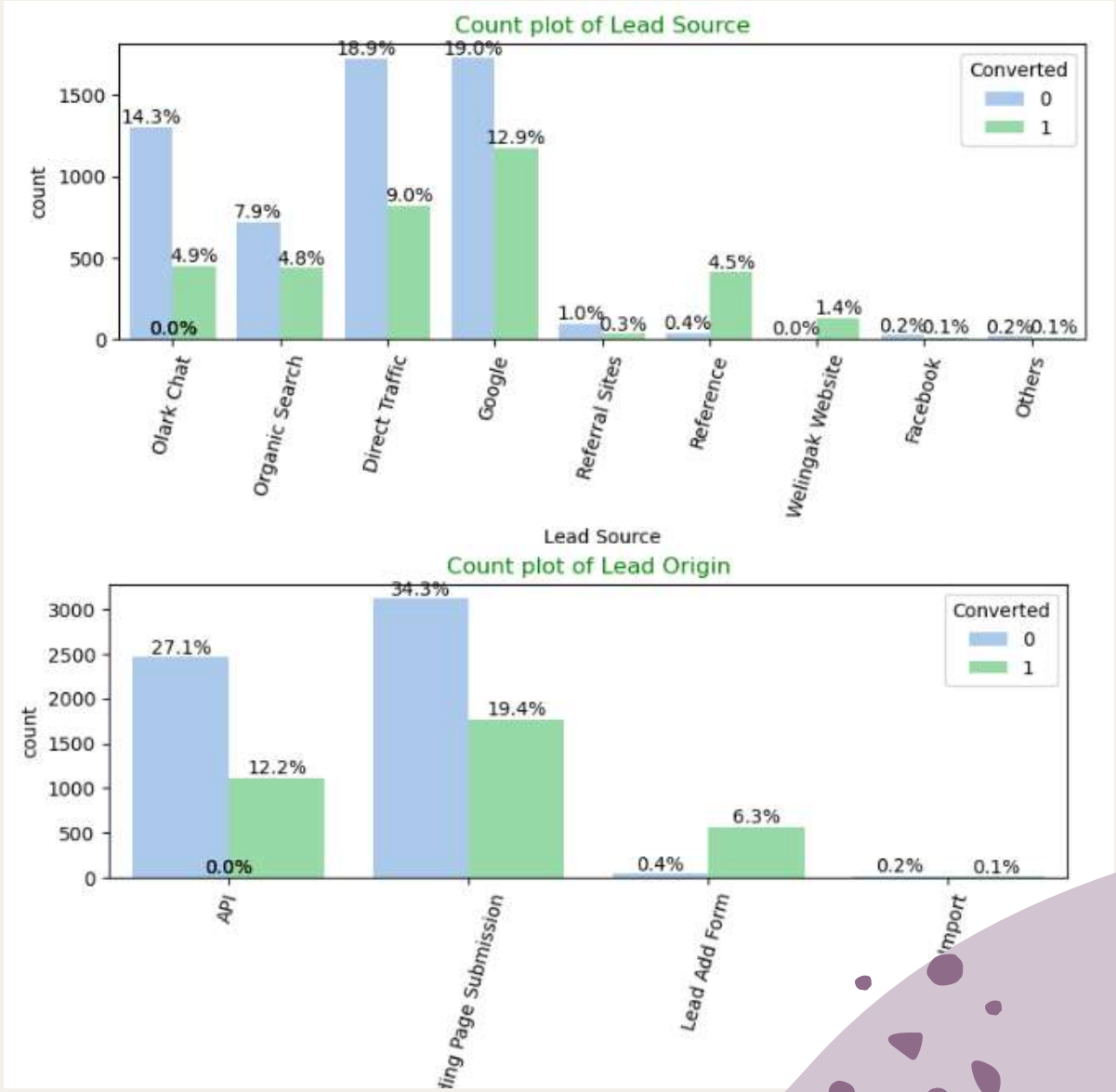




Bivariant Analysis – Categorical Variables

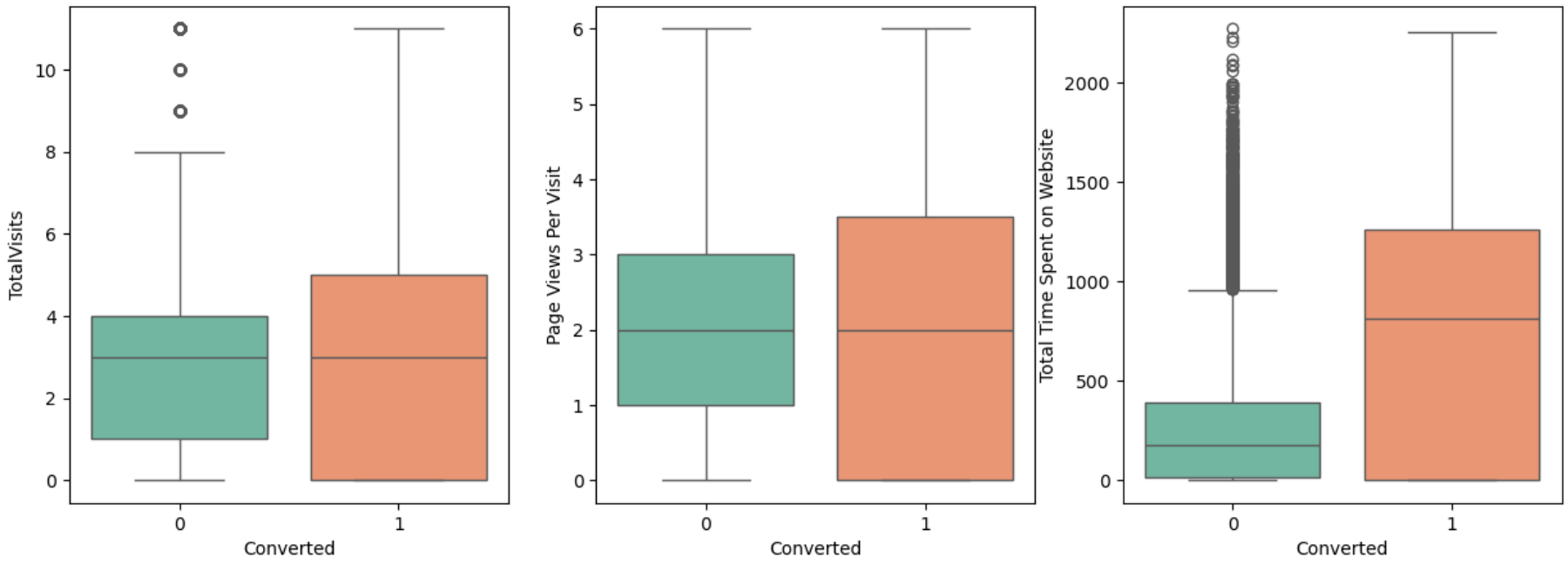
We found the following Variable important in univariant Analysis:

1. Lead Source : Google, Direct Traffic and Reference(highest Lead Conversion Rate)
2. Lead Origin : Lead Add Form(Highest Lead Conversion Rate) and Landing Submission Pages,API
3. Last Activity : SMS sent has highest Conversion Rate
4. Occupation : Working Professionals followed by Businessman has High Lead Conversion Rate while unemployed are highest contributors



Bivariant Analysis – Numerical Variables

Past Leads who spends more time on the Website have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot



Past Leads who spends more time on Website are successfully converted than those who spends less as seen in the boxplot

Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity,Current_occupation
- Splitting Train & Test Sets

  70:30 % ratio was chosen for the split
- Feature scaling ○ Standardization method was used to scale the features
- Checking the correlations

  ○ Predictor variables which were highly correlated with each other were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form has high Lead Conversion rate thus not dropping them irrespective of high Correlation).

Model Building

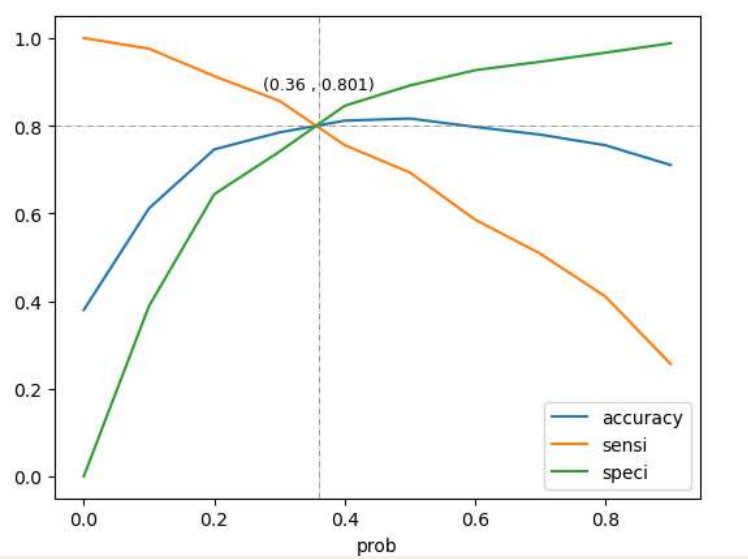
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome
    - Pre RFE – 31 columns & Post RFE – 15 columns
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 2 looks stable after Two iteration with:
    - significant p-values within the threshold (p-values < 0.05)
    - No sign of multicollinearity with VIFs less than 5
- Hence, logm2 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Model Evaluation

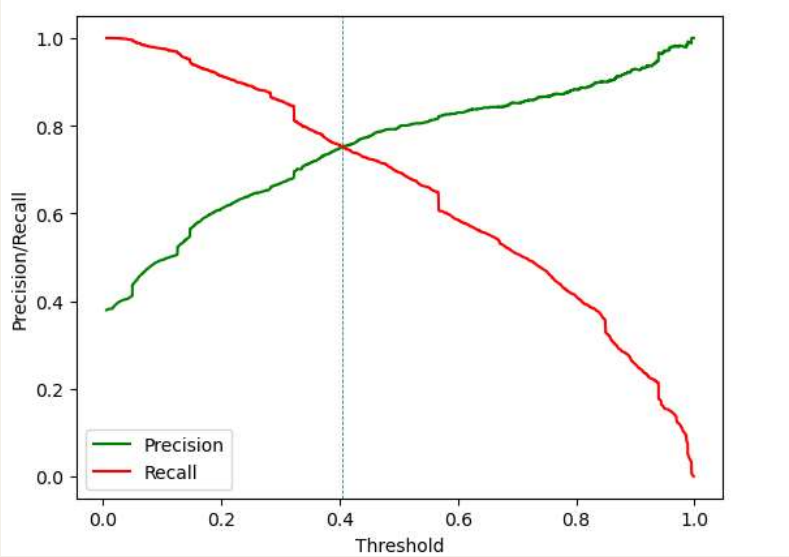
Train Data Set :

It was decided to go ahead with 0.36 as cutoff after checking evaluation metrics coming from both plots

*C*onfusion Matrix & Evaluation Metrics with 0.36 as cutoff *using Accuracy, Specificity, Sensitivity*



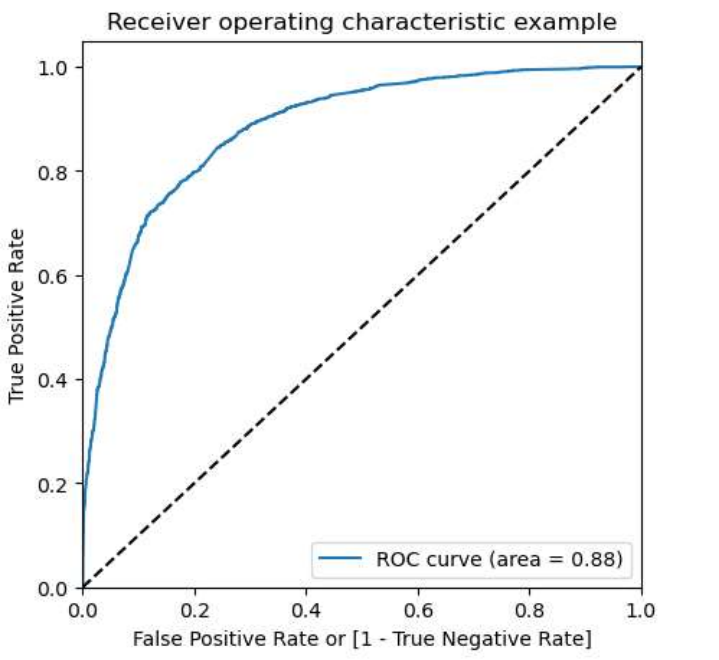
*C*onfusion Matrix & Evaluation Metrics with 0.405 as cutoff using Precision and Recall



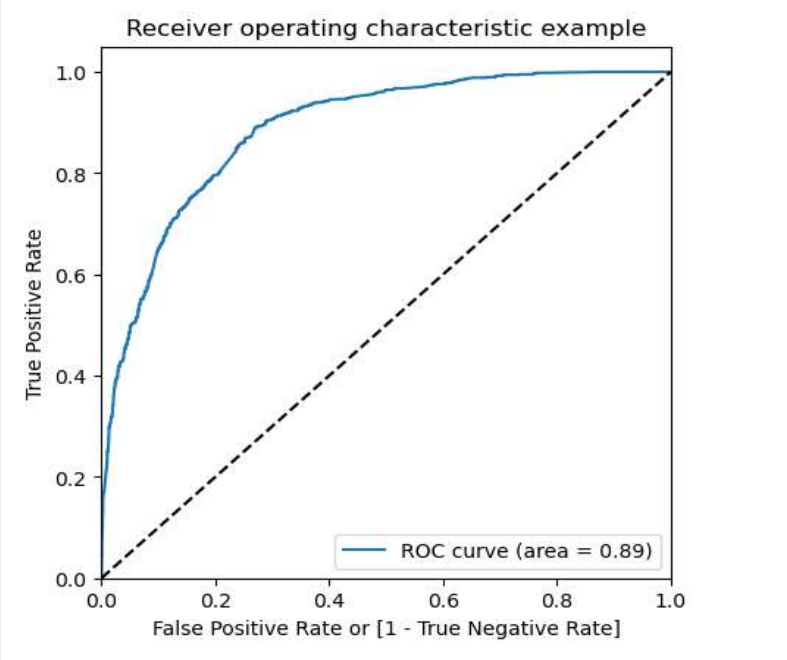
ROC CURVES :

Area 0.89 and 0.88 is good ROC Score

Test Set



Train Set



Accuracy, Specificity and Sensitivity of the models

Train

```
Confusion Matrix
[[3214  739]
 [ 520 1899]]

True Negative                    :  3214
True Positive                    :  1899
False Negative                   :  520
False Positve                    :  739
Model Accuracy                   :  80.24 %
Model Sensitivity                :  78.5 %
Model Specificity                :  81.31 %
Model Precision                  :  71.99 %
Model Recall                     :  78.5 %
Model True Positive Rate (TPR)   :  0.785
Model False Positive Rate (FPR)  :  0.1869
```

Test

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

Confusion Matrix
[[1378  311]
 [ 228  814]]

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

True Negative                    :  1378
True Positive                    :  814
False Negative                   :  228
False Positve                    :  311
Model Accuracy                   :  80.26 %
Model Sensitivity                :  78.12 %
Model Specificity                :  81.59 %
Model Precision                  :  72.36 %
Model Recall                     :  78.12 %
Model True Positive Rate (TPR)   :  0.7812
Model False Positive Rate (FPR)  :  0.1841
```

Using a cut-off value of 0.36, the model achieved a sensitivity of 80.24% in the train set and 80.26% in the test set.

- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model achieved an sensitivity of 78.5% and accurancy > 80%

*R*ecommendation based on Final Model

1.Prioritize High-Impact Features:

1. **Occupation_Working Professional**: With the highest coefficient (3.7478), leads identified as working professionals should be prioritized as they have the highest likelihood of conversion.
2. **Lead Origin_Lead Add Form**: This feature also has a strong positive impact (3.6829). Focus on leads originating from the Lead Add Form.
3. **Occupation_Businessman**: Another significant feature (2.1988). Leads identified as businessmen should be given priority.

2.Enhance Website Engagement:

1. **Total Time Spent on Website**: This variable shows a positive impact on lead conversion (1.1180). Encourage potential leads to spend more time on the website through engaging content and interactive features.

3.Utilize Effective Communication Channels:

1. **Last Activity_SMS Sent**: This has a very high positive impact (2.0291). Continue sending SMS messages to leads as it significantly boosts conversion rates.
2. **Last Activity_Email Opened**: Also shows a strong positive impact (1.0177). Ensure that email campaigns are effective and track email opens to identify engaged leads.

Recommendation based on Final Model

4. **Leverage Multiple Lead Sources**:
    1. **Lead Source_Olark Chat**: This source has a significant positive impact (1.2122). Utilize the Olark Chat feature on the website to engage with potential leads.
    2. **Lead Source_Welingak Website**: Although less impactful than others, it still shows a positive effect (1.7369). Continue to optimize this source for lead generation.
5. **Address Negative Indicators**:
    1. **Last Activity_Email Bounced**: This has a negative impact (-1.0726). Minimize email bounces by maintaining a clean email list and verifying email addresses.
6. **Focus on Diverse Occupations**:
    1. **Occupation_Other, Occupation_Student, and Occupation_Unemployed**: These categories also show positive impacts. Tailor marketing strategies to effectively engage with these groups.
7. **Monitor and Adjust Strategies**:
    1. Regularly review the model's performance and adjust strategies based on new data and changing market conditions. This will help in maintaining high conversion rates and adapting to new challenges.

Thank You