

LEAD SCORING CASE STUDY

TEAM MEMBERS:- Abhinav Sharma and Tina Pawar

PROBLEM STATEMENT

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

Background of X Education Company

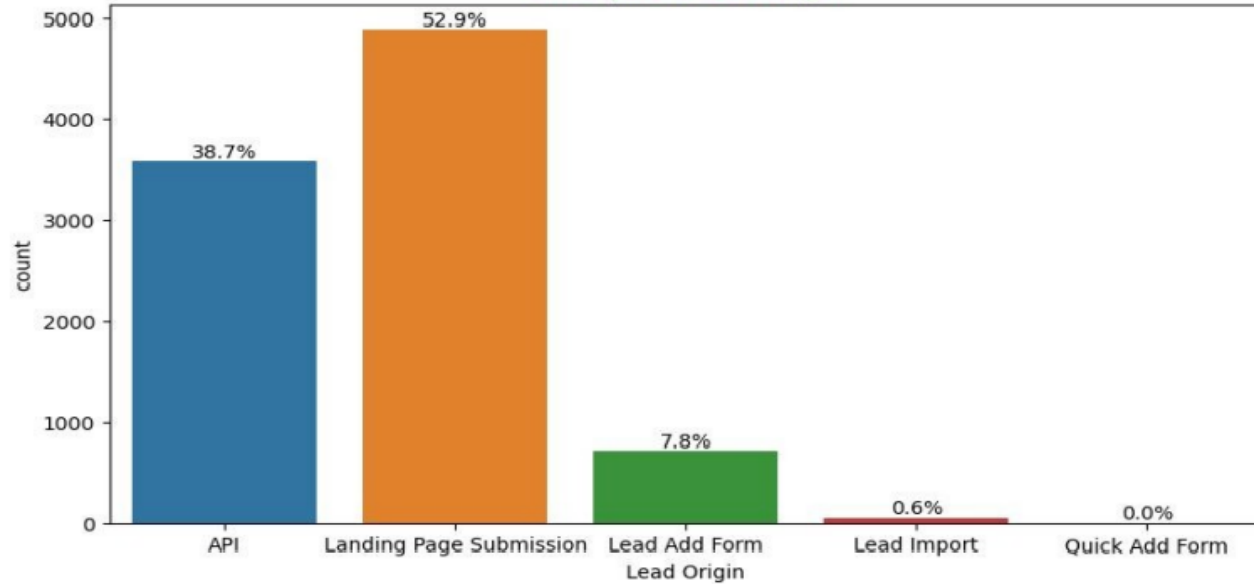
X Education is an online education company catering to industry professionals. Daily, many interested professionals visit their website, exploring courses. The company advertises on various platforms like Google. Visitors may view courses, complete forms, or watch videos. When forms are submitted with contact details, they become leads. The sales team contacts leads via calls or emails. While some leads convert to customers, most do not. X Education's average conversion rate is 30%.

Data Cleaning Process Summary:

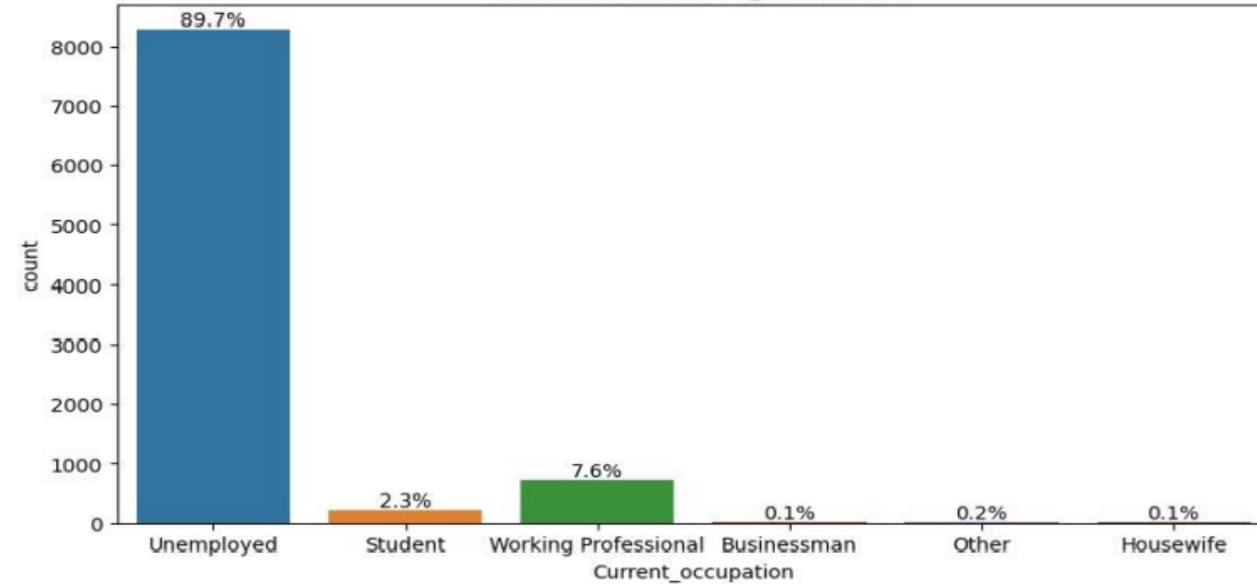
- ❑ Null values in categorical variables were represented as "Select" level, indicating non-selection.
- ❑ Columns with over 40% null values were dropped.
- ❑ Missing categorical values were managed based on value counts and considerations.
- ❑ Irrelevant columns (tags, country) were removed.
- ❑ Imputation was applied to certain categorical variables.
- ❑ New categories were created for specific variables.
- ❑ Columns with no modeling utility (Prospect ID, Lead Number) or just one response category were dropped.
- ❑ Numeric data was imputed with mode after assessing distribution.
- ❑ Skewed categorical columns were discarded to prevent logistic regression bias.
- ❑ Outliers in Total Visits and Page Views Per Visit were treated and capped.
- ❑ Invalid values were corrected, and data standardization performed (e.g., lead source).
- ❑ Infrequent values were grouped into an "Others" category.
- ❑ Binary categorical variables were mapped.
- ❑ Additional data quality checks and standardization steps were executed (e.g., fixing casing inconsistencies).

Univariate Analysis – Categorical Variables

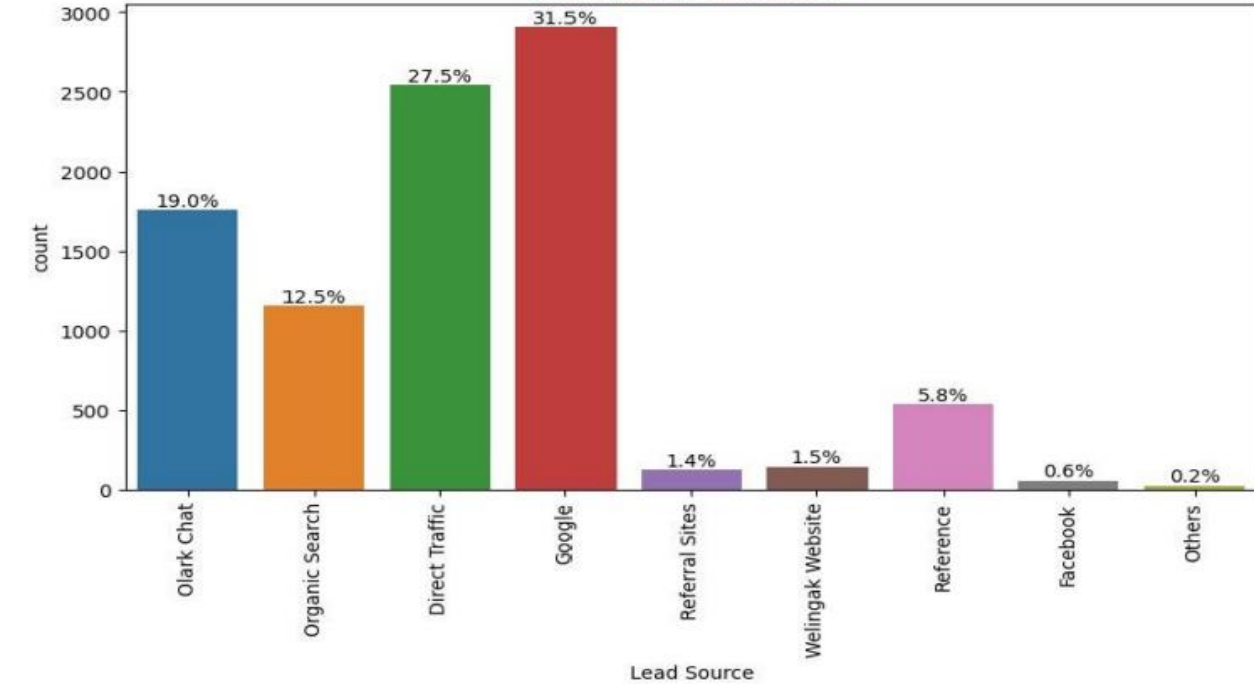
Count plot of Lead Origin



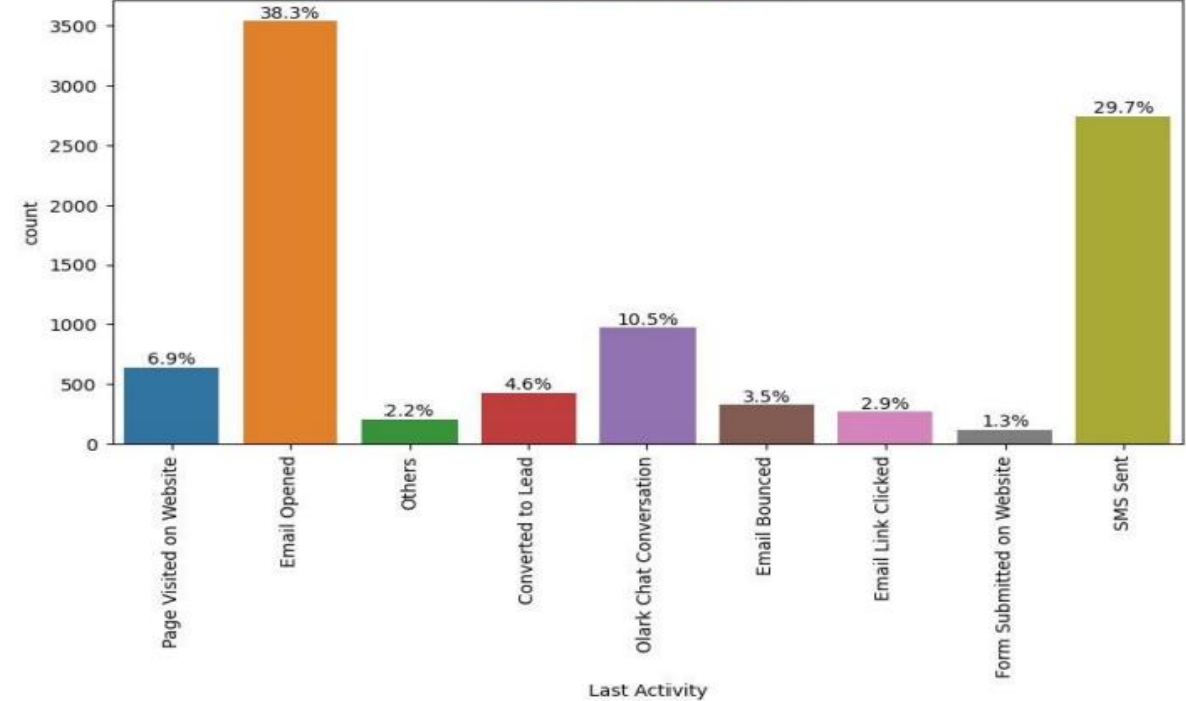
Count plot of Current_occupation

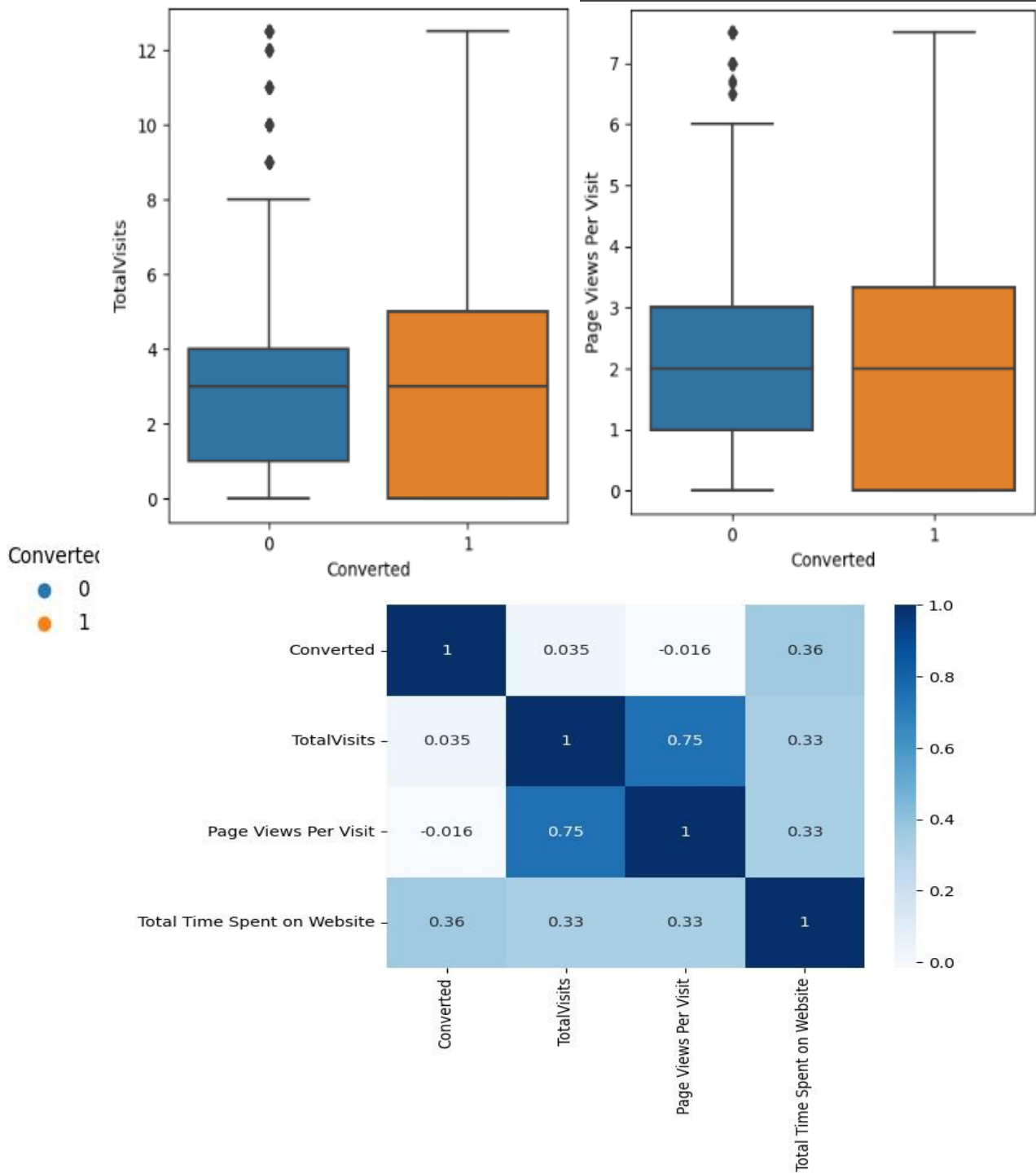
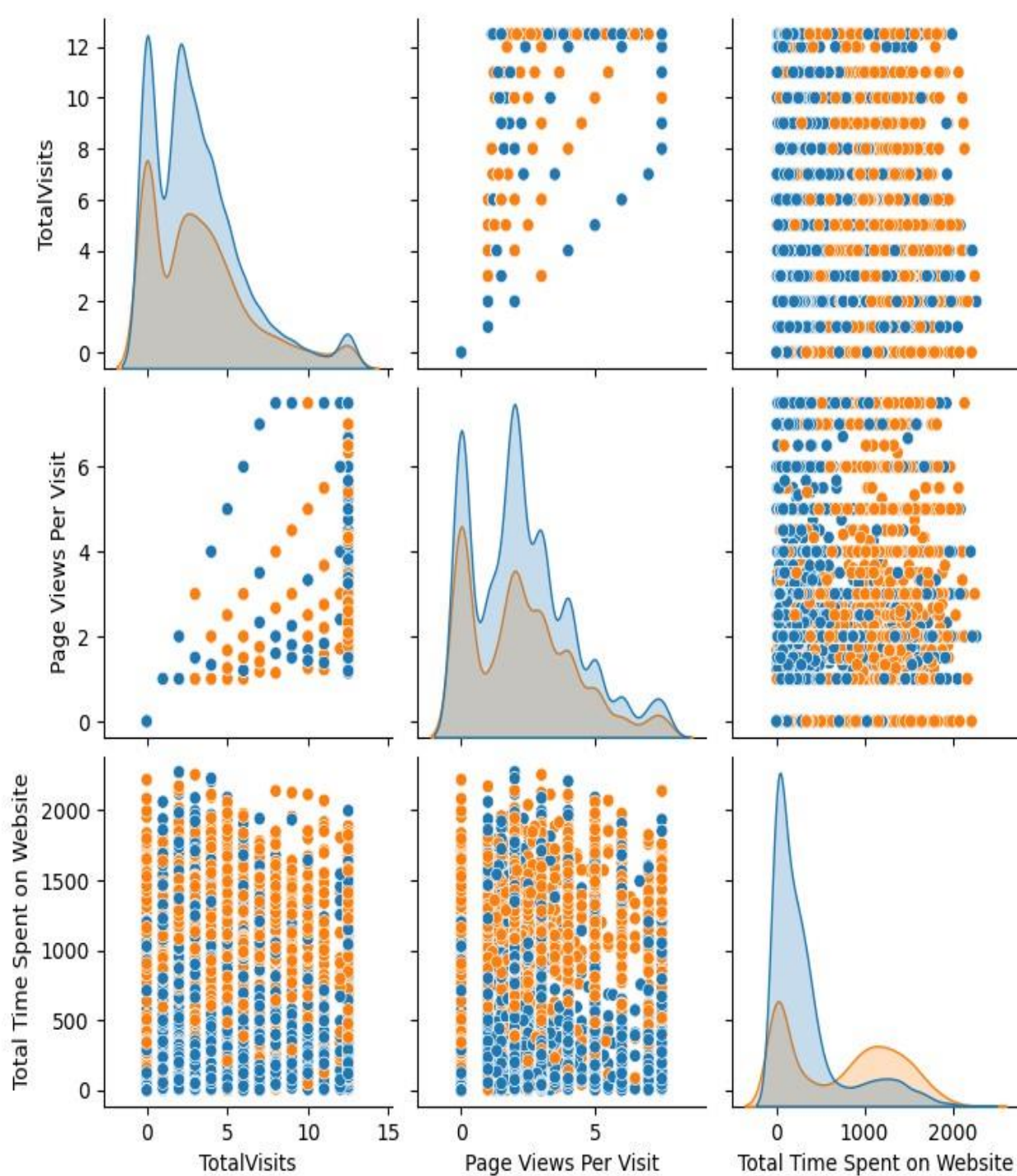


Count plot of Lead Source



Count plot of Last Activity





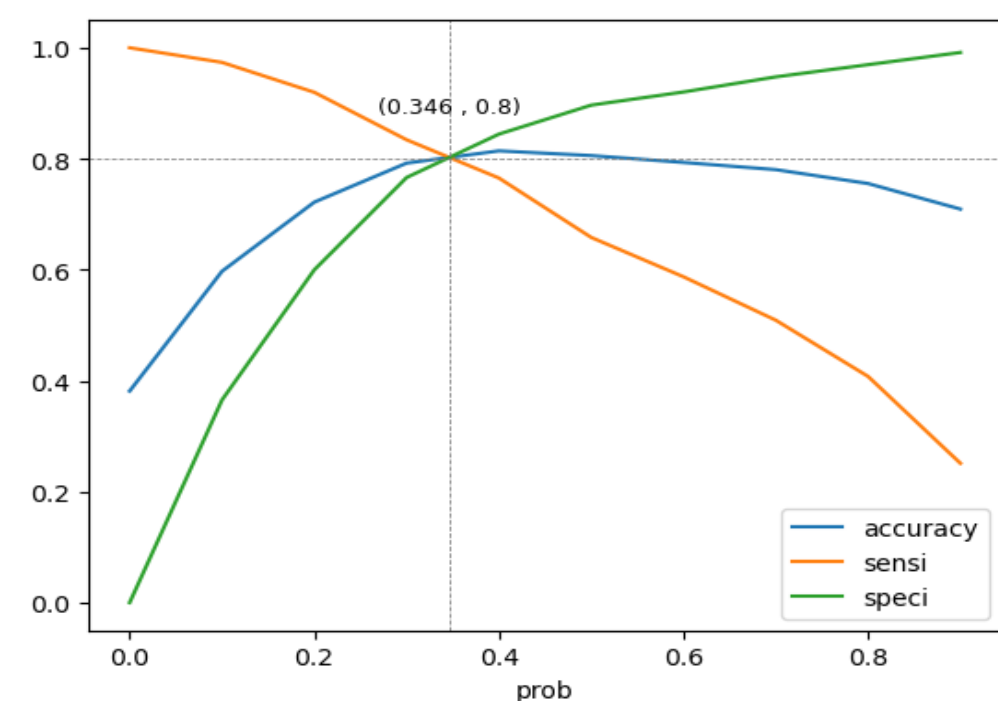
Data Preparation Summary

Before Model Building:

- Dummy features (one-hot encoding) were created for categorical variables: Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation.
- The dataset was split into Train and Test sets using a 70:30 ratio.
- Feature scaling was performed using standardization to ensure consistent scales for model training.
- Correlations were checked, and highly correlated predictor variables were dropped (e.g., Lead Origin_Lead Import and Lead Origin_Lead Add Form).

Model Building Process:

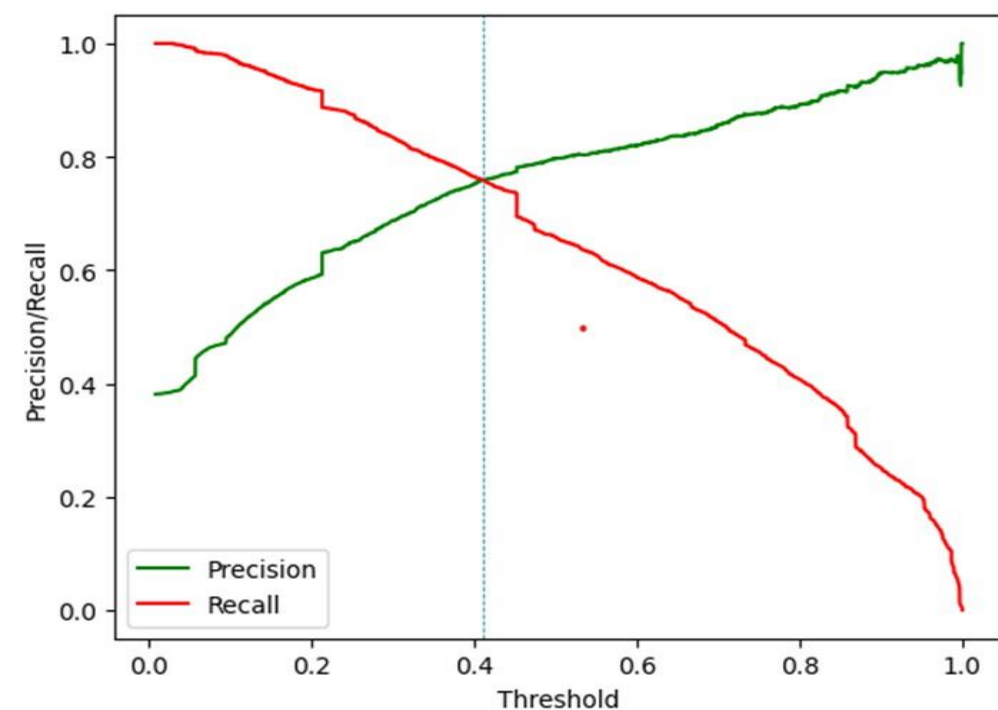
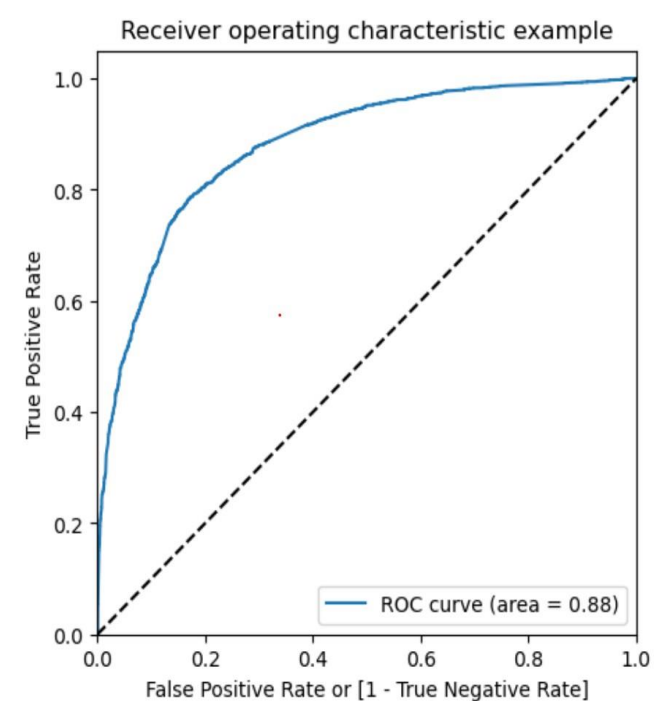
- Feature selection aimed to address the high dimensionality of the dataset.
- Recursive Feature Elimination (RFE) was employed to select important columns, reducing the number of features from 48 to 15.
- The final model, logm4, was selected for Model Evaluation and prediction. The data preparation and model building phases focused on optimizing feature selection and ensuring model stability, setting the stage for accurate evaluation and prediction in subsequent steps.



Confusion Matrix

```
[[3233 769]
 [ 493 1973]]
```

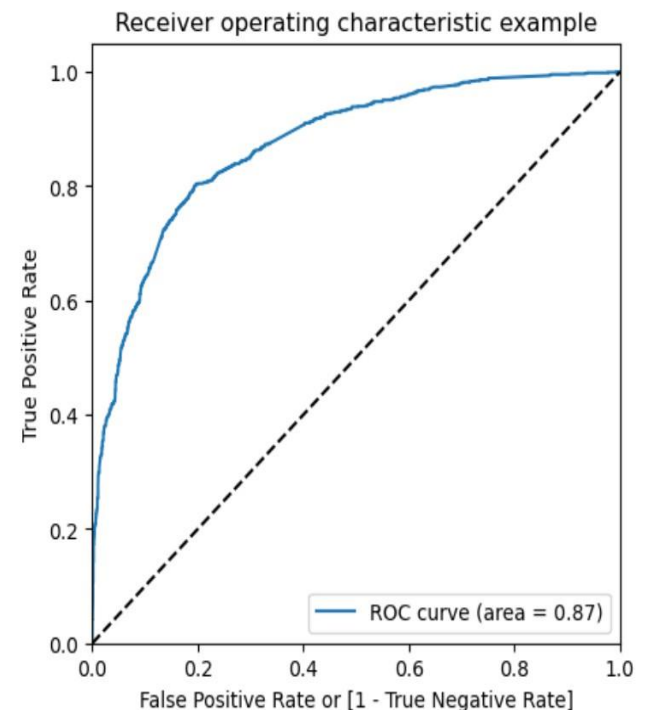
| | | |
|---------------------------------|---|--------|
| True Negative | : | 3233 |
| True Positive | : | 1973 |
| False Negative | : | 493 |
| False Positive | : | 769 |
| Model Accuracy | : | 0.8049 |
| Model Sensitivity | : | 0.8001 |
| Model Specificity | : | 0.8078 |
| Model Precision | : | 0.7195 |
| Model Recall | : | 0.8001 |
| Model True Positive Rate (TPR) | : | 0.8001 |
| Model False Positive Rate (FPR) | : | 0.1922 |



Confusion Matrix

```
[[3407 595]
 [ 593 1873]]
```

| | | |
|---------------------------------|---|--------|
| True Negative | : | 3407 |
| True Positive | : | 1873 |
| False Negative | : | 593 |
| False Positive | : | 595 |
| Model Accuracy | : | 0.8163 |
| Model Sensitivity | : | 0.7595 |
| Model Specificity | : | 0.8513 |
| Model Precision | : | 0.7589 |
| Model Recall | : | 0.7595 |
| Model True Positive Rate (TPR) | : | 0.7595 |
| Model False Positive Rate (FPR) | : | 0.1487 |



Recommendations Based on Final Model:

- To Improve Lead efforts towards features with positive coefficients, as they have the most iConversion Rates:
- Prioritize marketing impact on lead conversion.
- Develop targeted marketing strategies to attract leads from high- performing sources like Welingak Website and Reference.
- Optimize communication channels based on the engagement impact of leads. For example, focus more on channels like SMS Sent, Email Opened, and Olark Chat.
- Engage working professionals through personalized messaging, considering their higher conversion potential.
- Allocate a higher budget for Welingak Website advertising to increase lead acquisition from that source.
- Encourage customers to provide references by offering incentives or discounts, as these references have a positive effect on conversion.

Areas for Improvement:

- Analyze specialization offerings with negative coefficients, such as Hospitality Management and Others, to identify opportunities for enhancement.
- Review the landing page submission process for potential improvements, as the Lead Origin of Landing Page Submission has a negative coefficient.
- In conclusion, the recommendations are geared towards leveraging the identified positive coefficients to optimize marketing strategies and increase lead conversion rates. Additionally, analyzing negative coefficients offers insights into areas that may need improvement for further enhancement of the lead conversion process

THANKYOU!