LEAD SCORING CASE STUDY

Focused business approach using logistic regression technique

Submitted by: Dikshya Napit
Nirali Negi
Upeksh Naik

BUISNESS OBJECTIVE

To assist X Education in identifying the most promising leads (Hot Leads), meaning the leads with the highest likelihood of converting into paying customers.

METHODOLOGY

To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa.

Target Lead Conversion Rate ≈ 80%

1. Importing and Observing the past data provided by the Company

2. Data Cleaning

- Missing value imputation
- Removing duplicate data and other redundancies

3. Exploratory Data Analysis

- Univariate and Bivariate analysis

4. Data Preparation

- Outlier treatment
- Dropping unnecessary columns
- Dummy variable creation
- Feature standardization

1. Model Building

- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs

2. Model Evaluation

- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold

3. Comparison with PCA

Building another model using PCA

- Comparing the two models

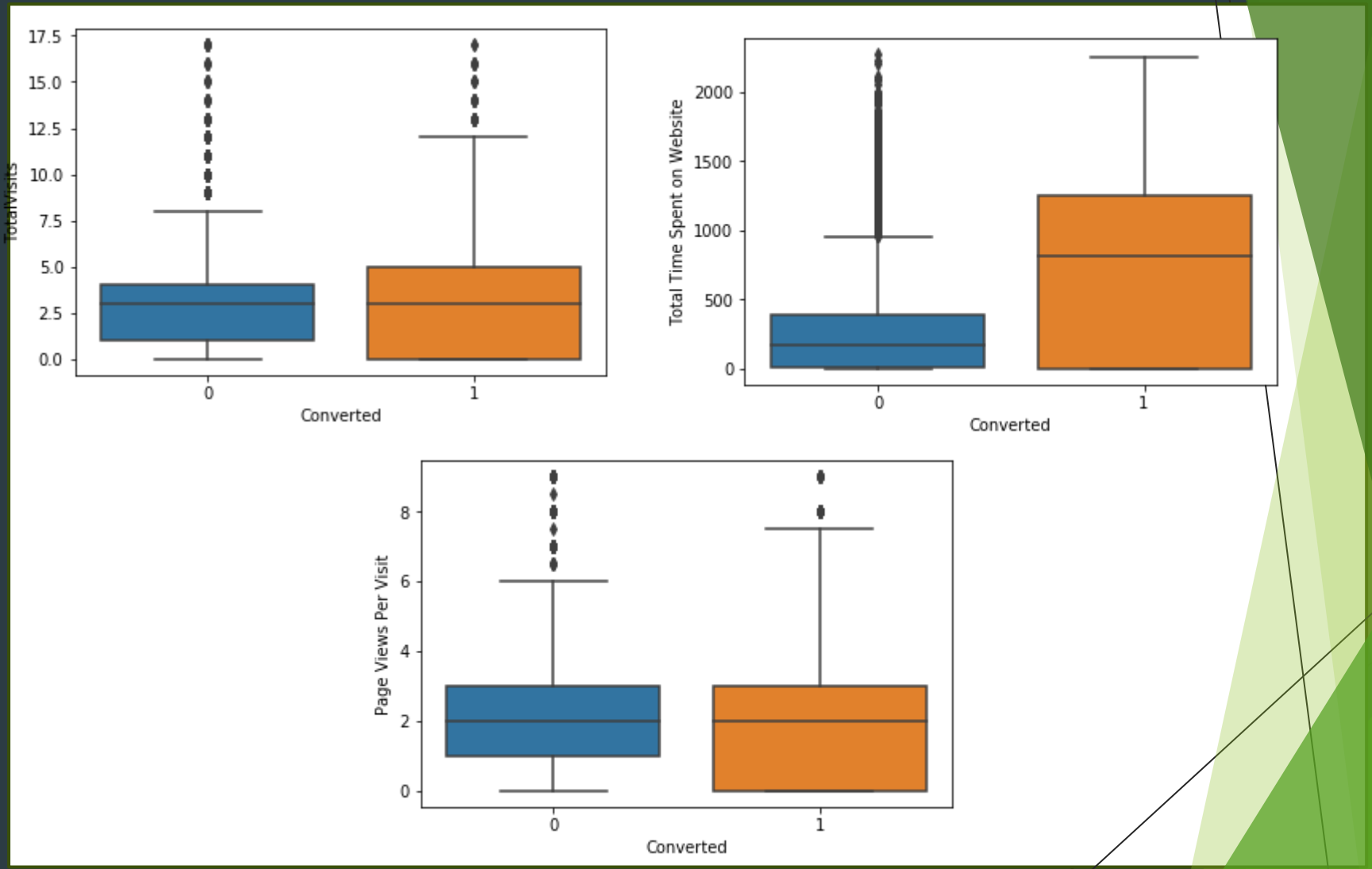
4. Assigning Lead Scores

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:
- Lead Score = Probability * 100

Data Visualization

- To identify important features
- To get insights

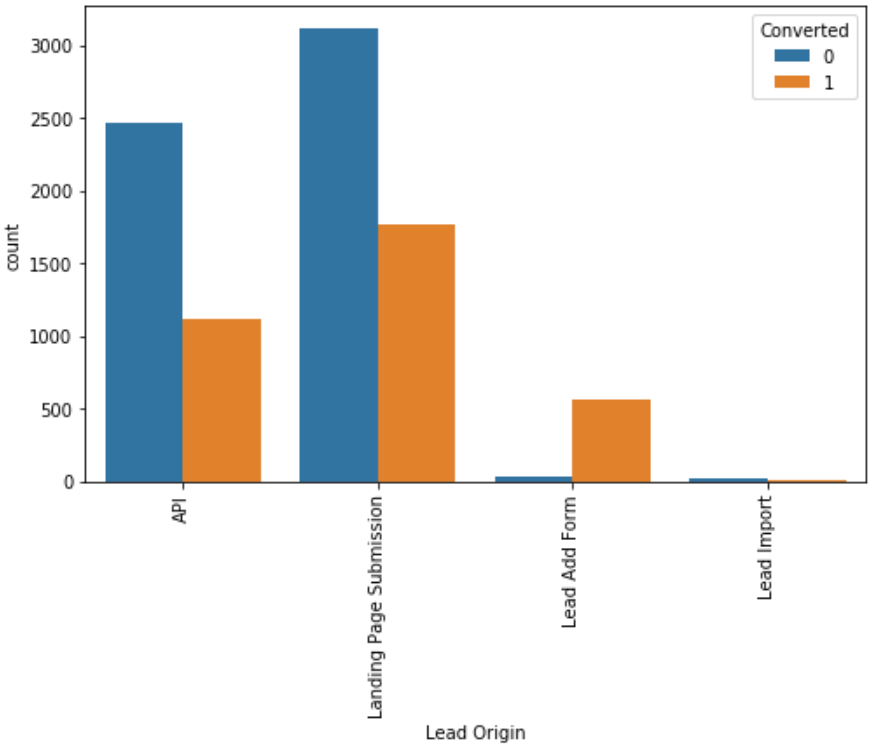
Numerical Variables



The longer people stay on websites, the higher their chances of converting

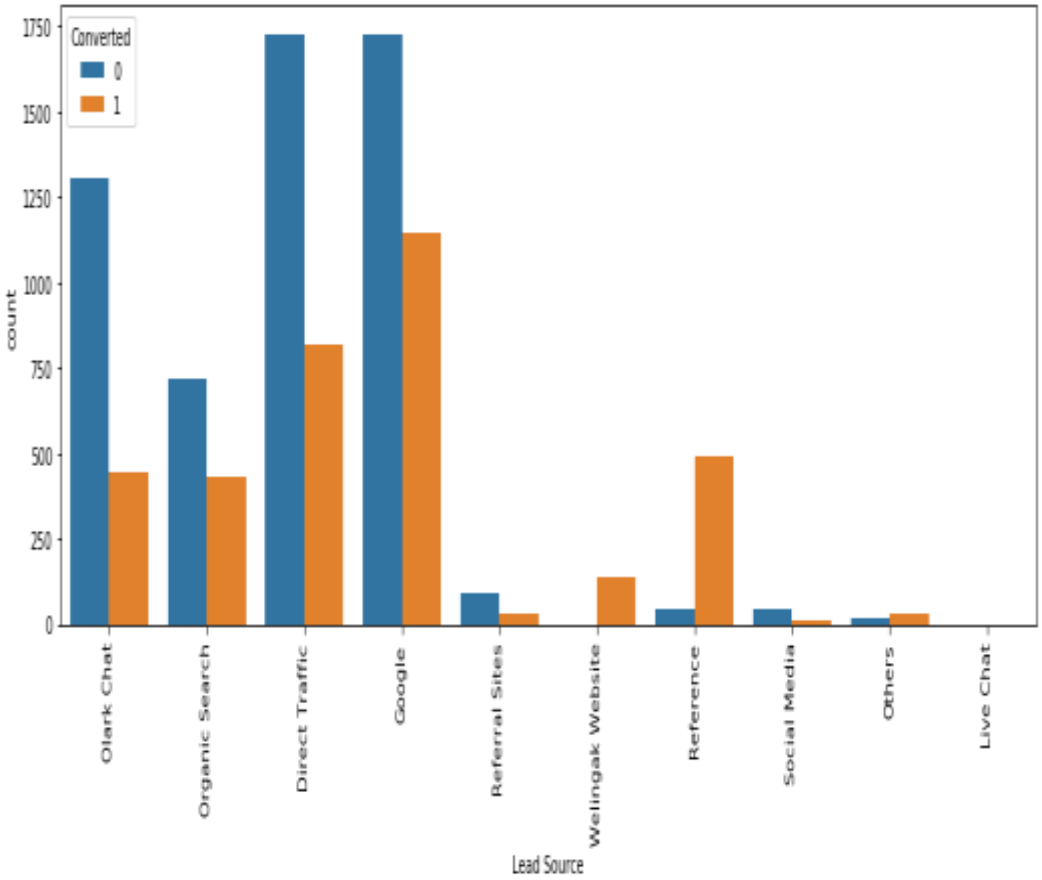
Lead Origin

- ‘API’ and ‘Landing Page Submission’ generate the highest number of leads but have lower conversion rates, while ‘ ‘Lead Add Form’ produces fewer leads but achieves a higher conversion rate.
- Focus on improving the conversion rates for ‘API’ and ‘Landing Page Submission’ while boosting lead generation for the ‘Lead Add Form.’”

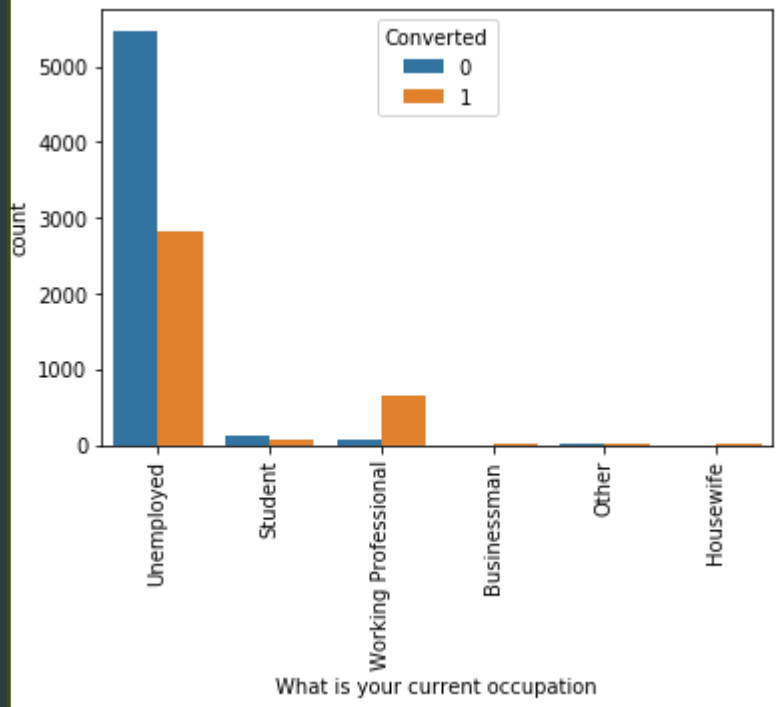


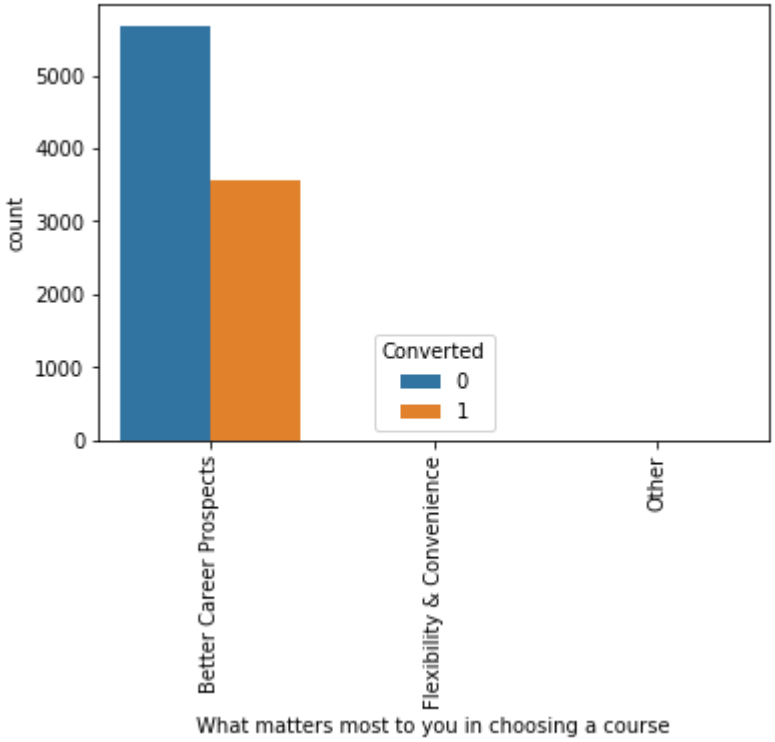
Lead Source

- Lead sources like 'Reference' and 'Welingak Website' have exceptionally high conversion rates, while most leads come from 'Direct Traffic' and 'Google'."



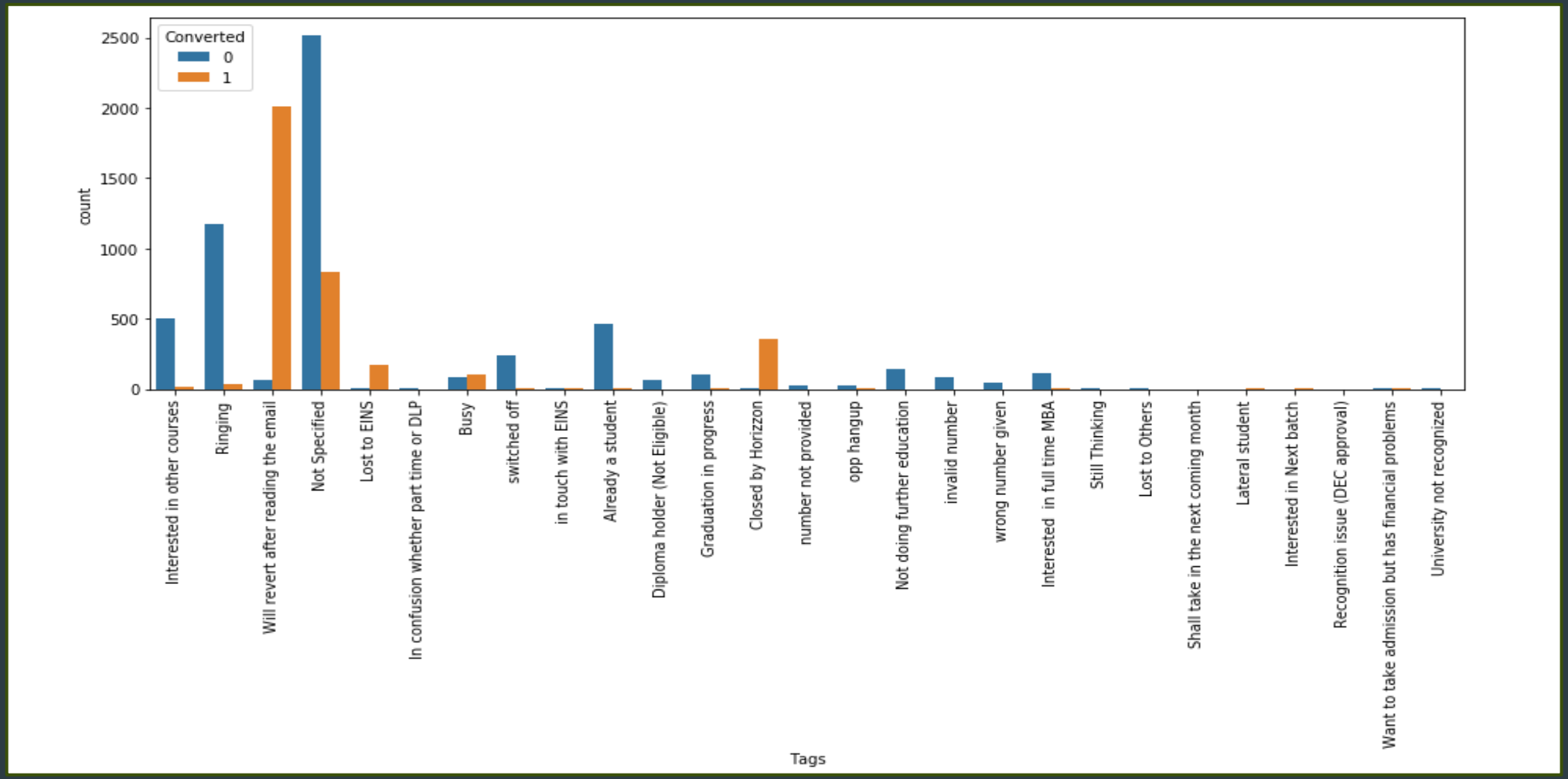
Active Professional





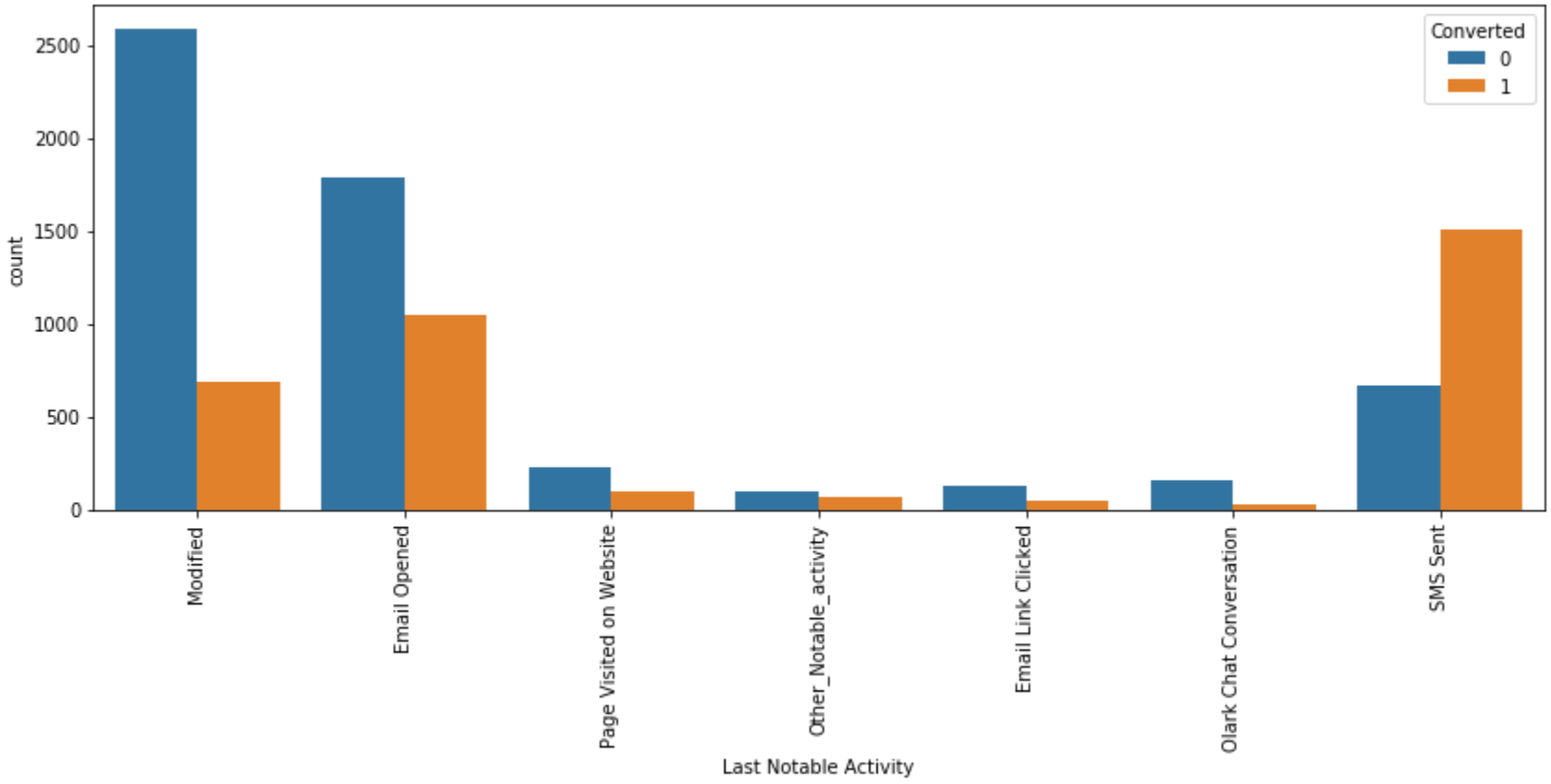
Professionals currently employed have the highest likelihood of successfully converting.

Tags



Tags like 'Will respond after reviewing the email,' 'Closed by Horizon,' 'Lost to EINS,' and 'Busy' show high conversion rates.

LAST NOTABLE ACTIVITY



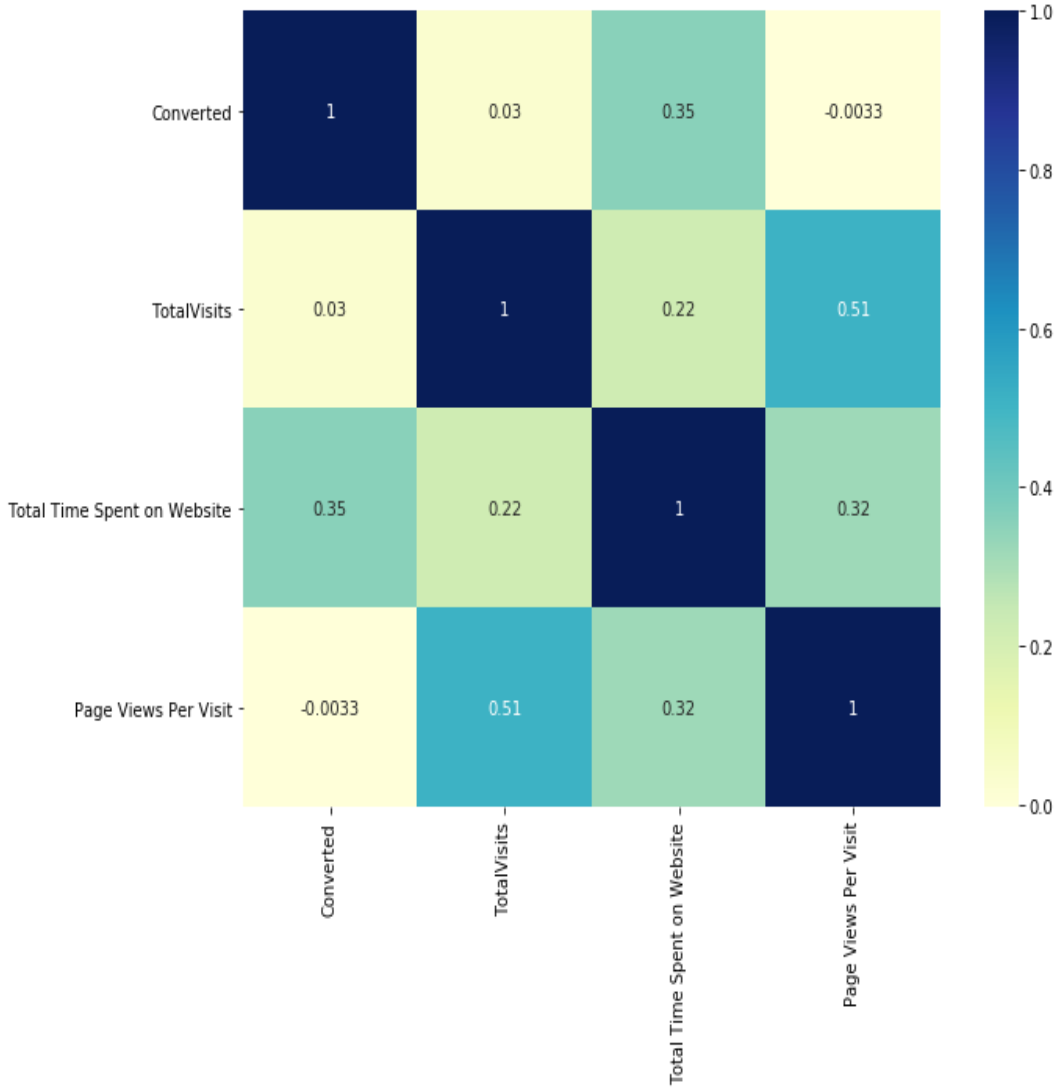
Highest conversion rate is for the last notable activity '**SMSSent**'.

MODEL EVALUATION

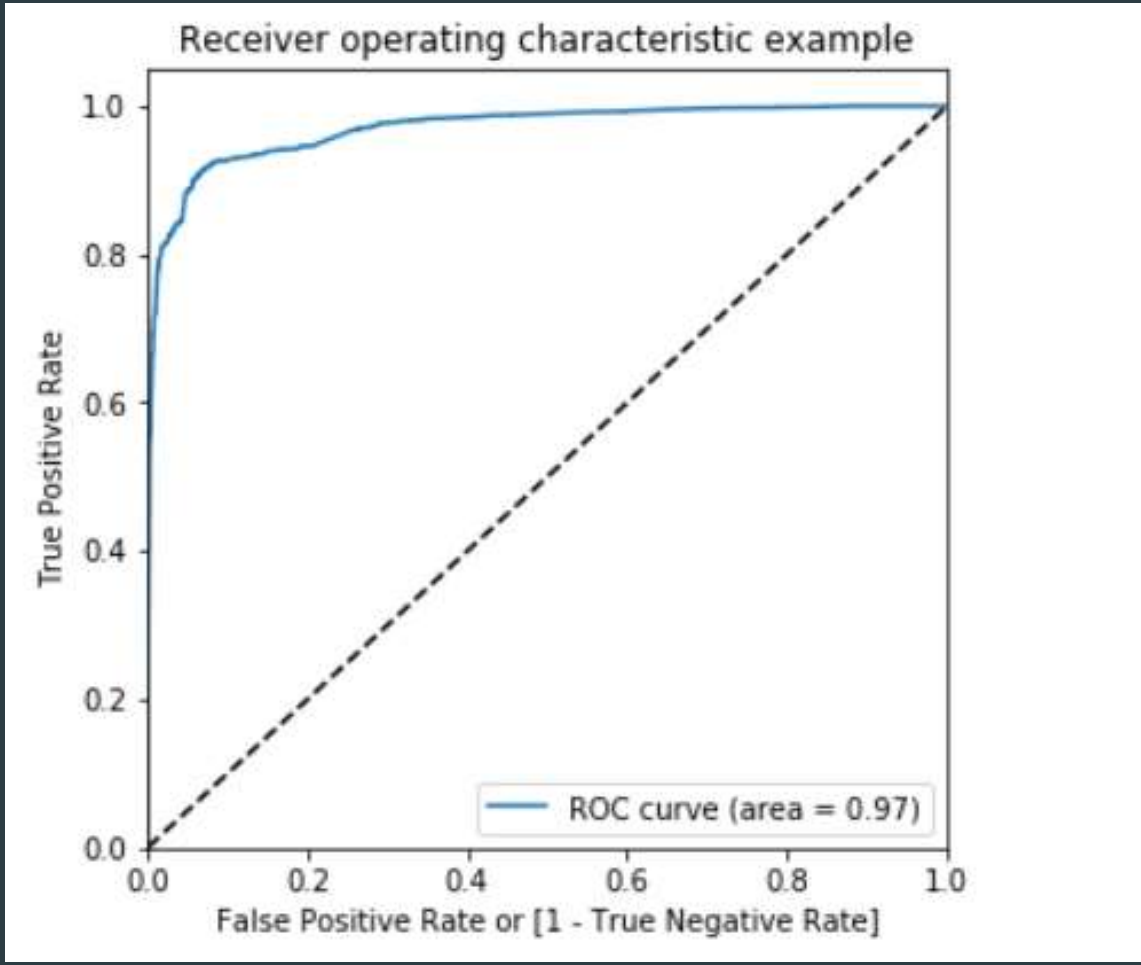
Assessment of the model

CORRELATION OF NUMERIC VALUES

- The final model's features exhibit minimal correlations.

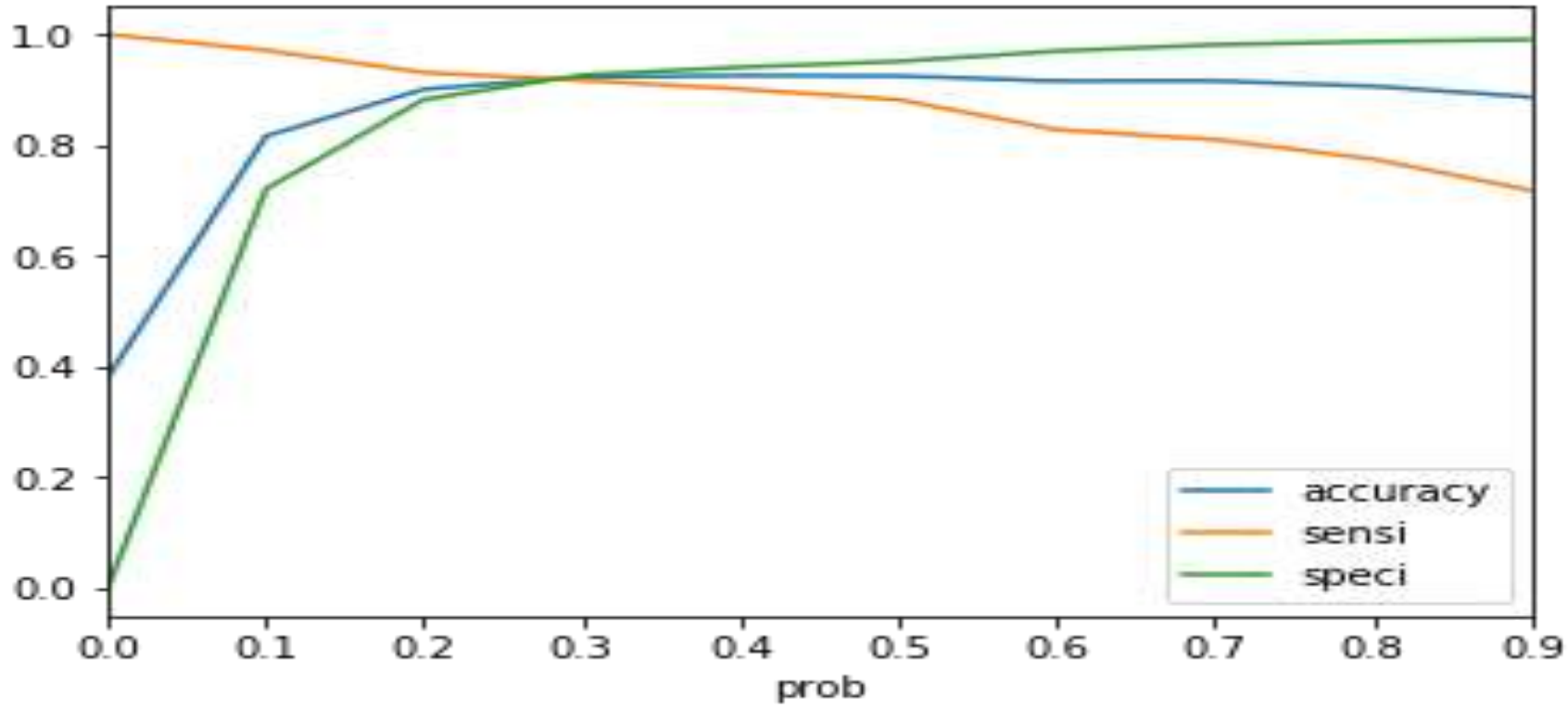


AREA UNDER CURVE



Area under curve = 0.95

DETERMINING IDEAL THRESHOLD



A graph illustrating how Sensitivity, Specificity, and Accuracy vary with different probability threshold values. The optimal cutoff is set at 0.20.

FINAL OUTPUT

Final Observation:

Let us compare the values obtained for Train & Test:

Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

Test Data:

- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%

FEATURE SIGNIFICANCE

The top three variables influencing the probability of lead conversion, in descending order of impact, are:

1.Tags_Lost to EINS

2.Tags_Closed by Horizon

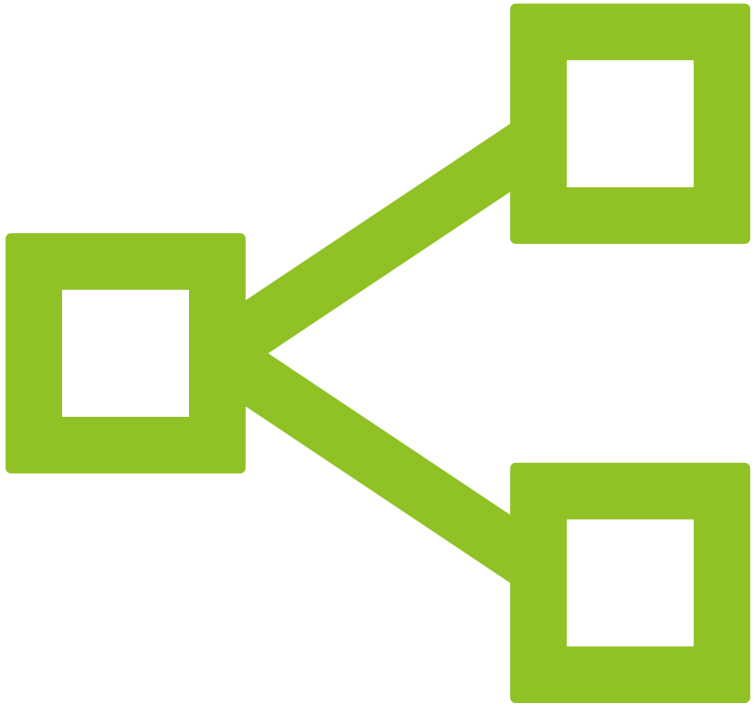
3.Tags_Will revert after reading the email.

These are binary features derived from the categorical variable Tags.

- All three factors positively influence the likelihood of lead conversion.
- These findings suggest that the company should prioritize leads associated with these three tags.

COURSE OF ACTION

- Analyze data visualizations to improve conversion rates for categories with more leads.
- Generate additional leads for categories with higher conversion rates.
- Consider the relative importance of model features and their positive or negative impact on conversion probability.
- Adjust the probability threshold value based on evolving business needs to better identify potential leads.



THANK YOU