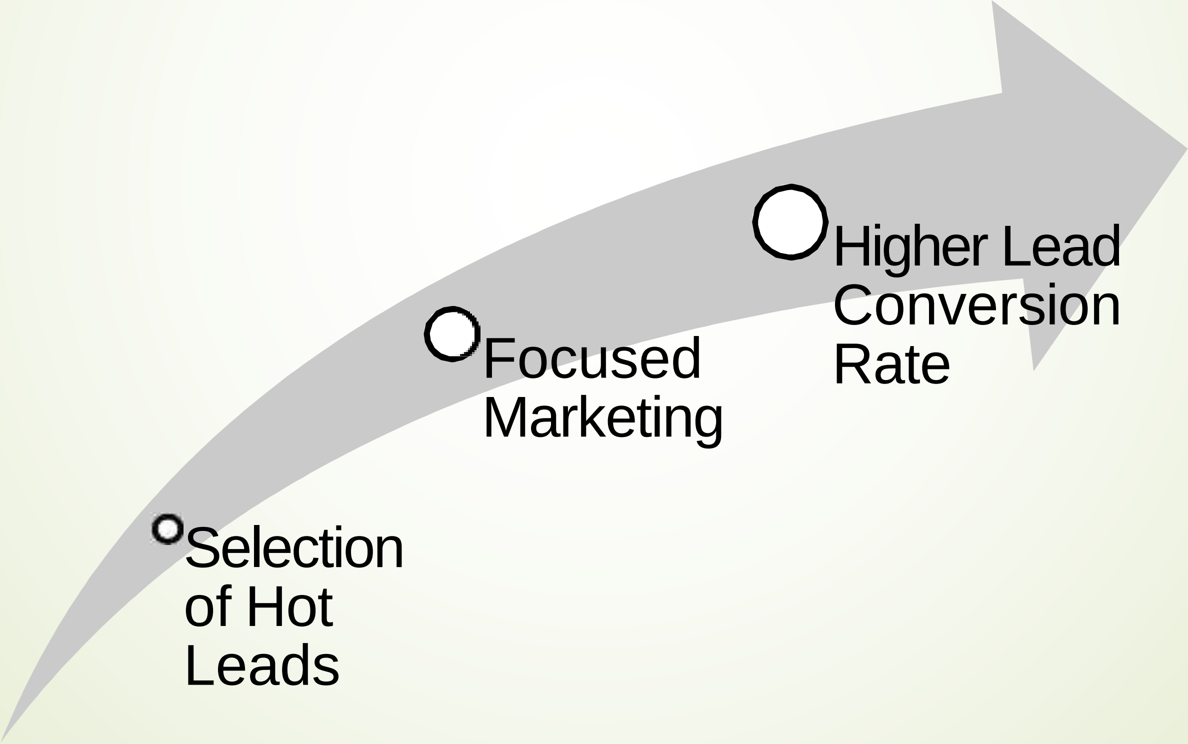
LEAD SCORING CASE STUDY

“Lead Scoring in Action: A Case Study in Precision Marketing"

Business Objective

Assisting **XEducation** in selecting high-potential leads that have the greatest chance of turning into customers.



Selection of Hot Leads

Focused Marketing

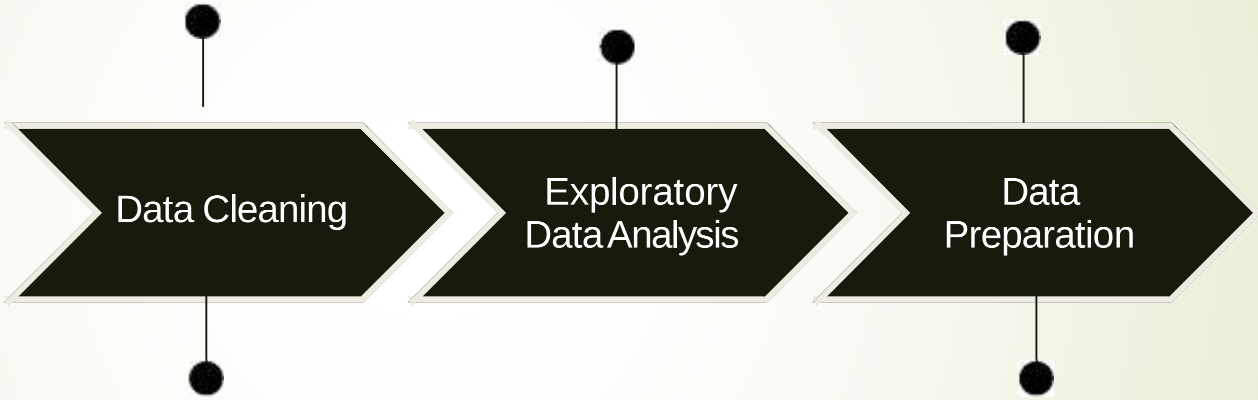
Higher Lead Conversion Rate

METHODOLOGY

Designing a Logistic Regression model to rank leads based on their likelihood of conversion, with the goal of achieving an approximate 80% lead conversion rate.

Importing and Observing the past data provided by the Company

Univariate and Bivariate analysis



Data Cleaning

Exploratory Data Analysis

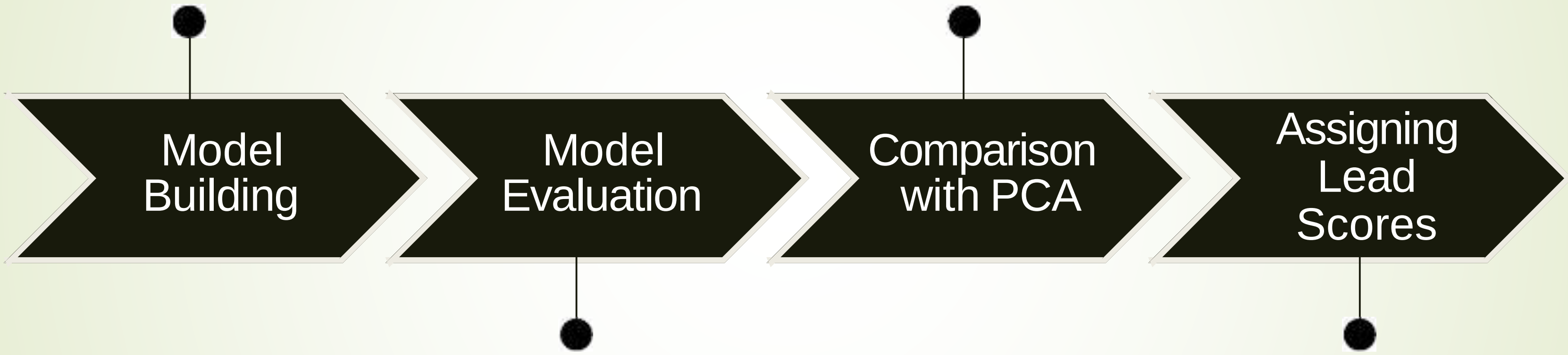
Data Preparation

- Missing value imputation
- Removing duplicate data and other redundancies

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

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

- Building another model using PCA
- Comparing the two models



Model Building

Model Evaluation

Comparison with PCA

Assigning Lead Scores

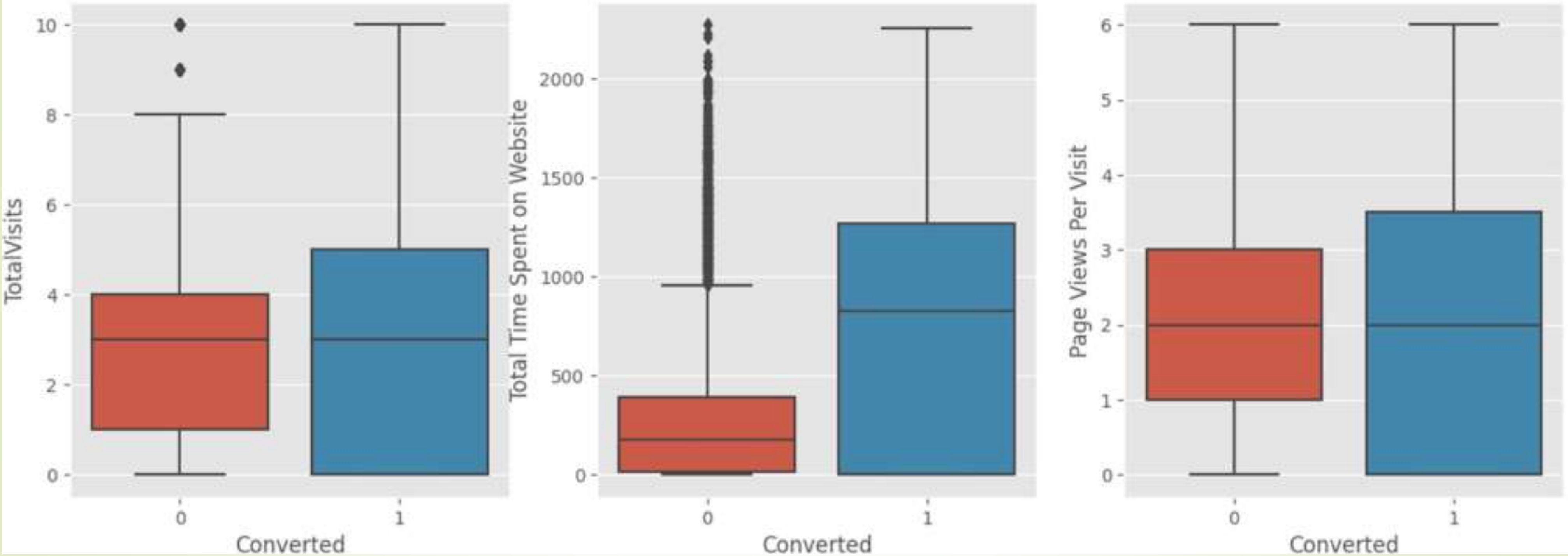
- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold

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

DATA VISUALIZATION

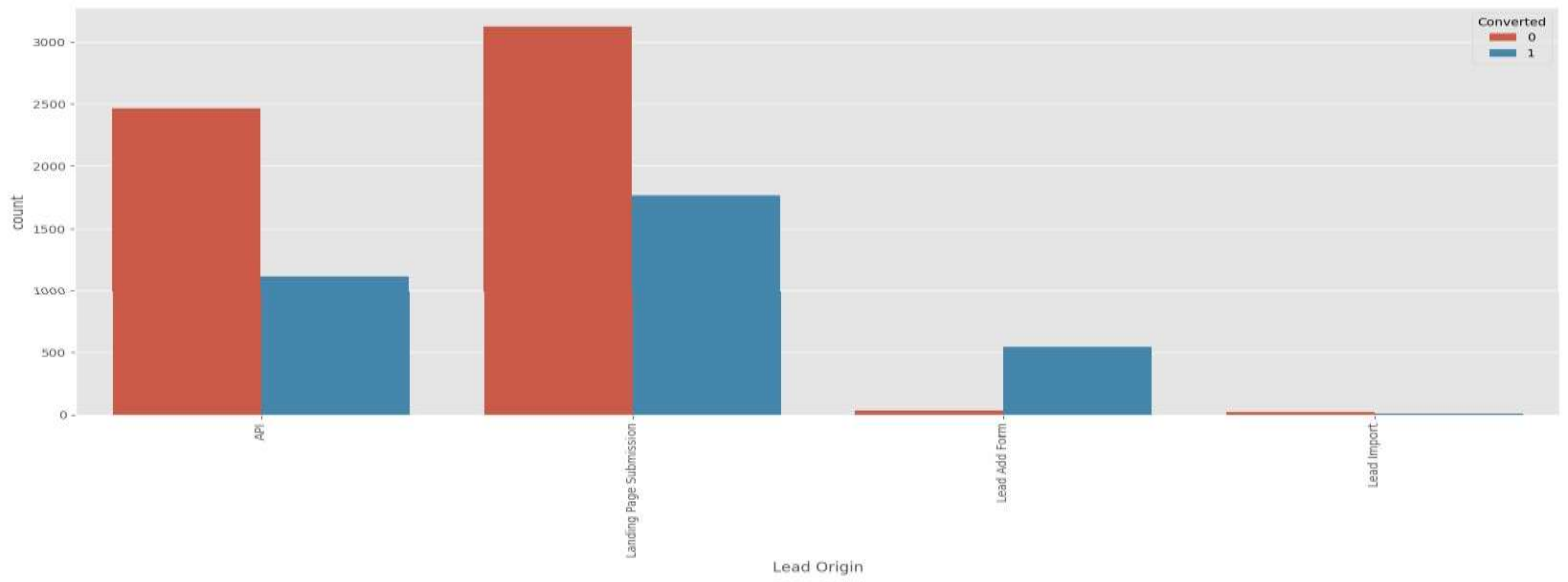
- To identify important features
- To get insights

Numerical Variables



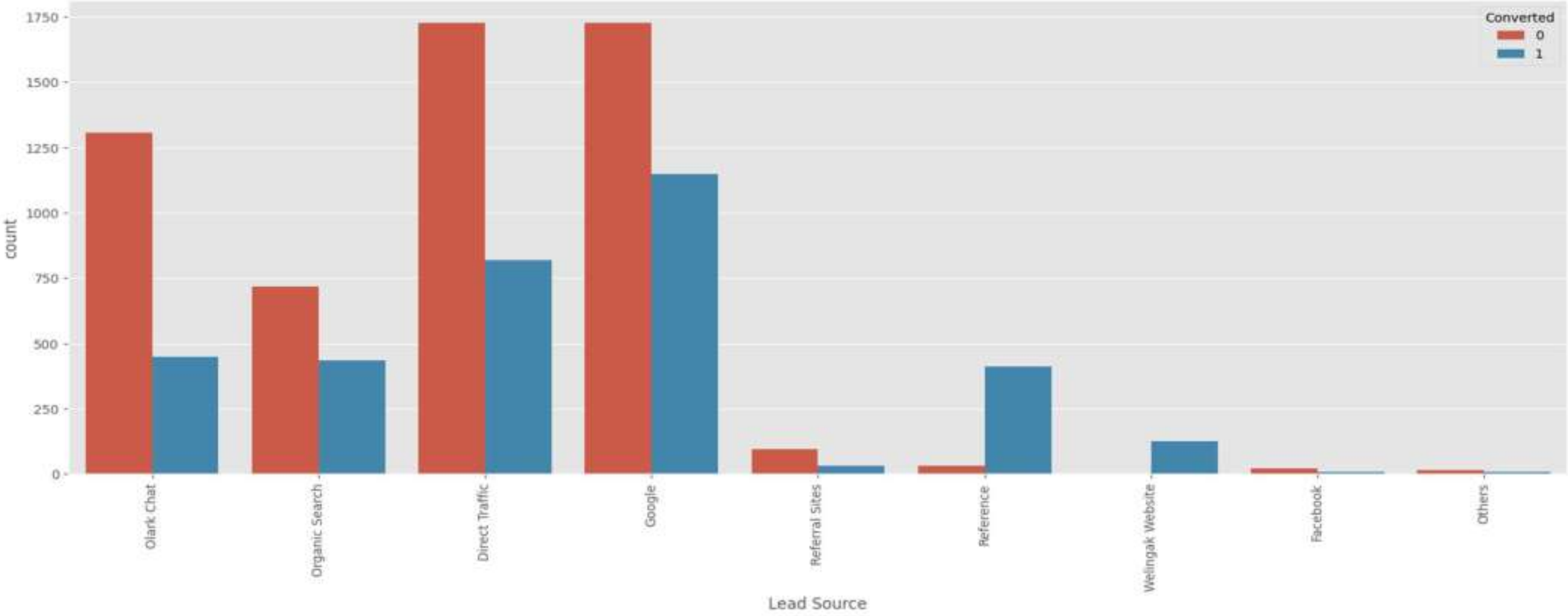
People spending more time on website are more likely to get converted.

Lead Origin



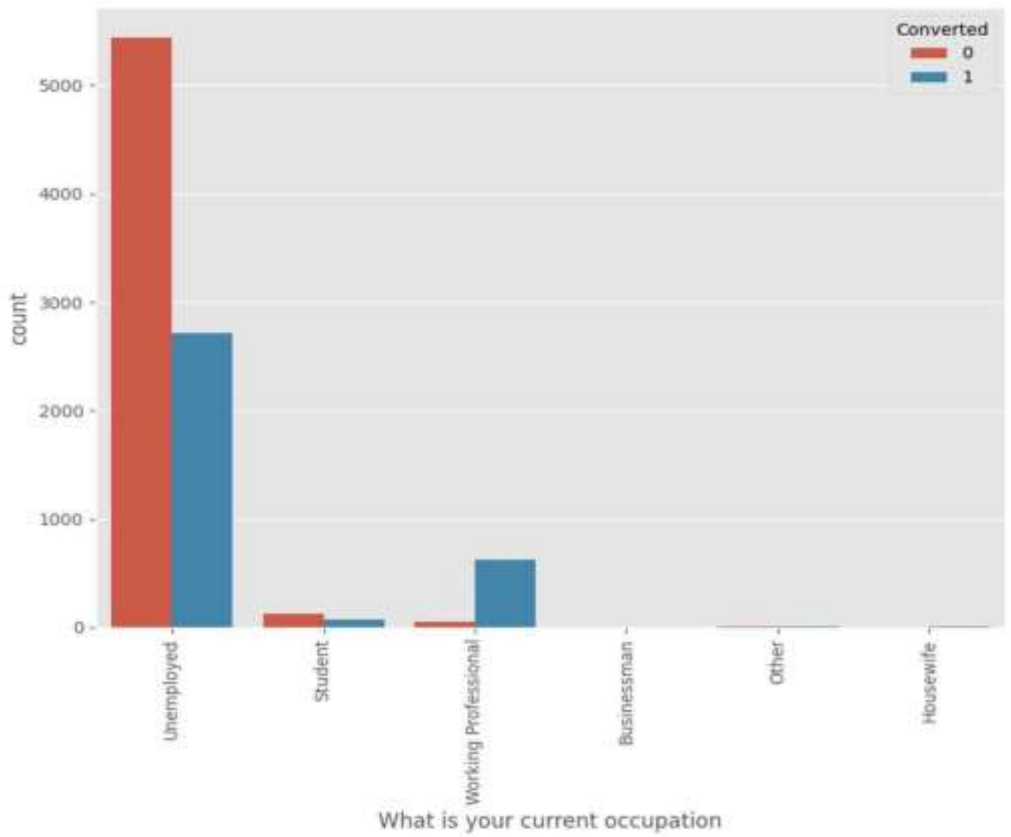
- ***'API'*** and '***Landing Page Submission***' generate the most leads but have less conversion rates, whereas '***Lead AddForm***' generates less leads but conversion rate is great.
- **Tryto increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'**.

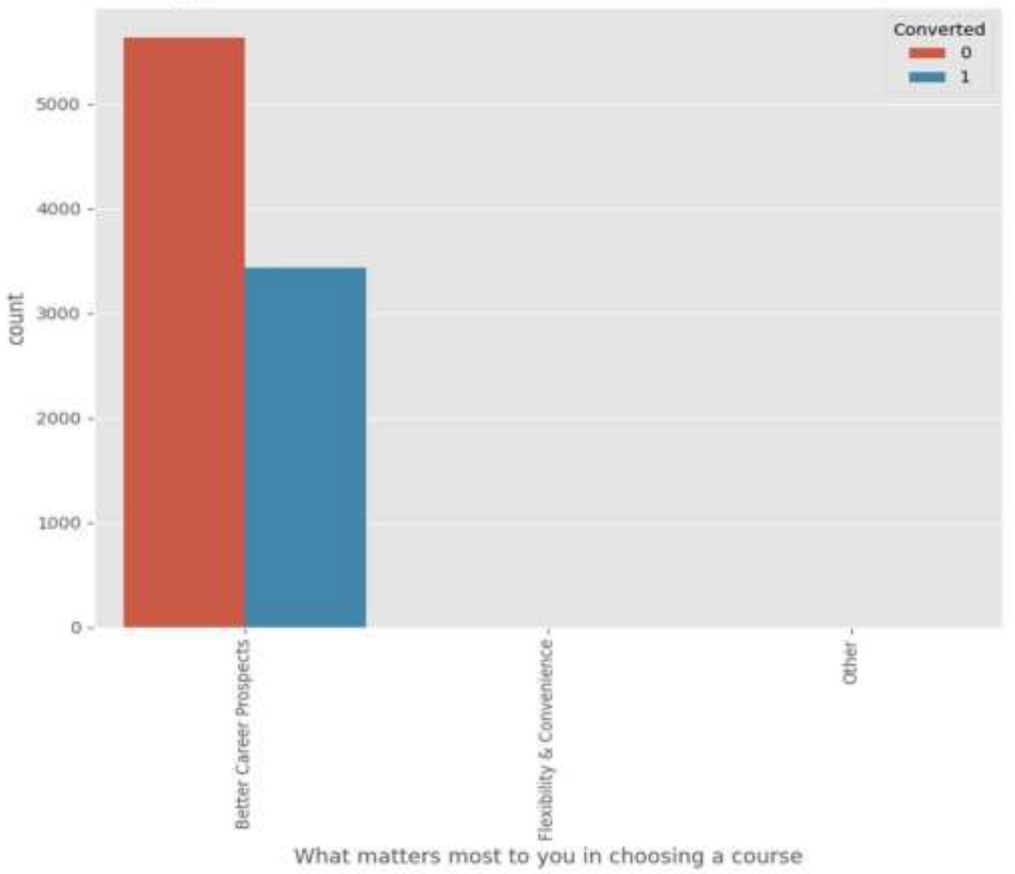
Lead Source



- Very high conversion rates for lead sources ‘**Reference**’ and ‘**Welingak Website’**
- Most leads are generated through **‘Direct Traffic**’ and ‘**Google**’.

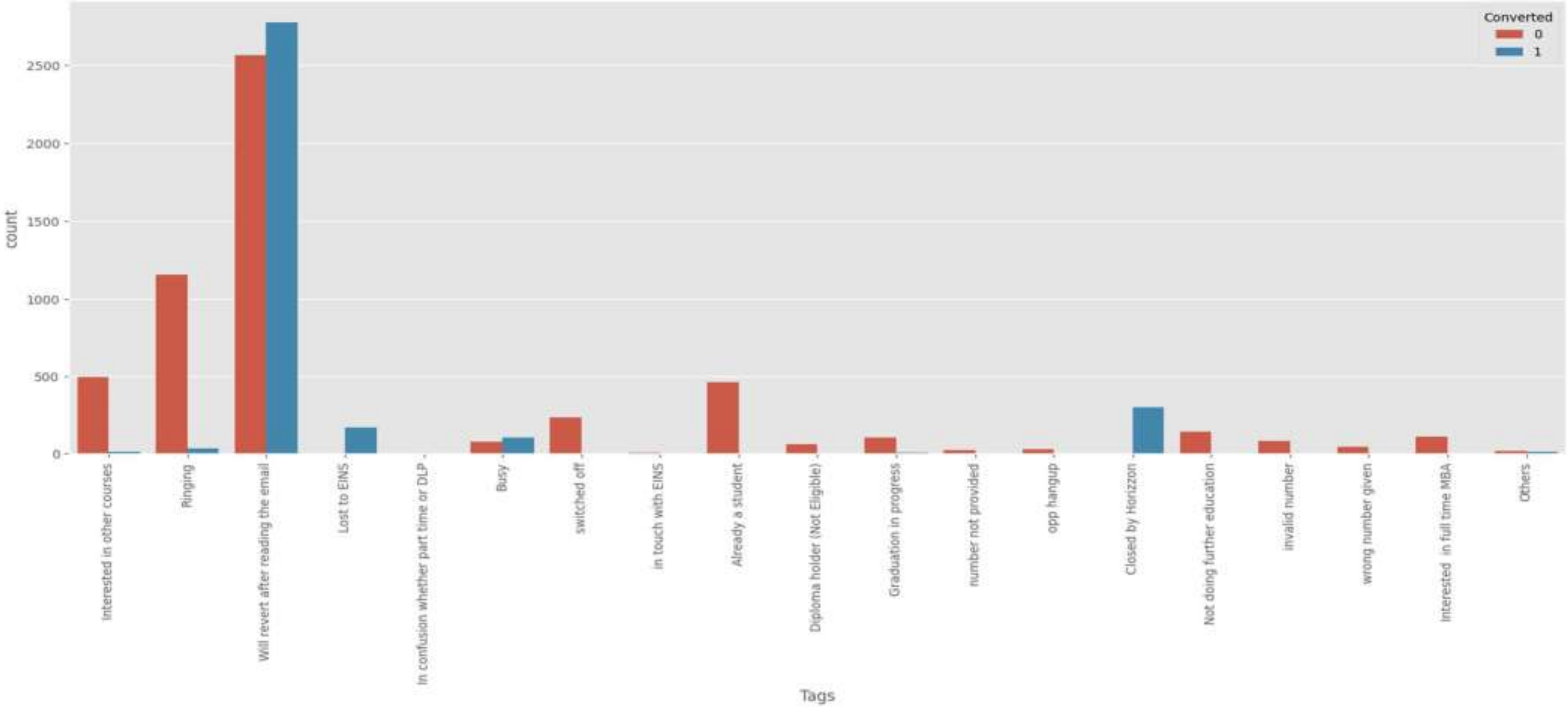
Current Occupation





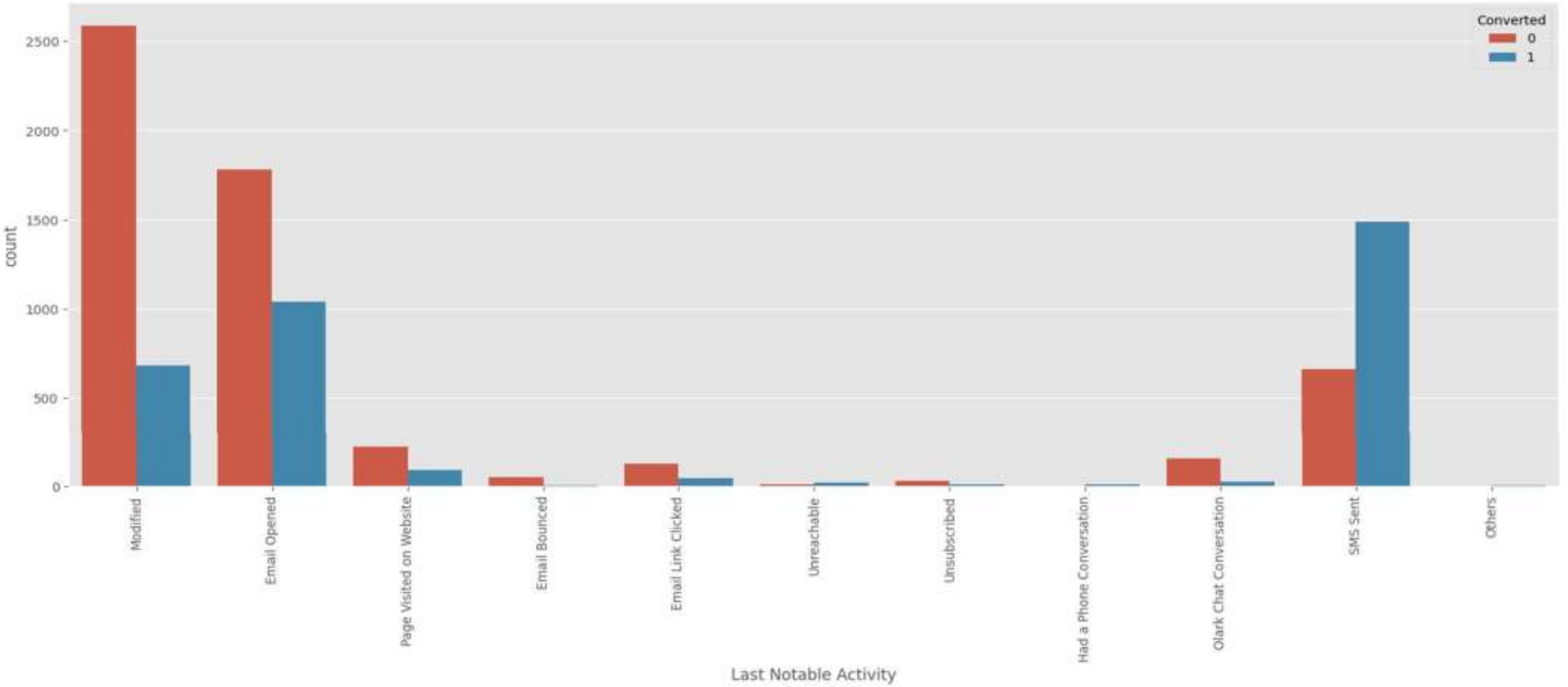
Working Professionals are most likely to get converted.

Tags



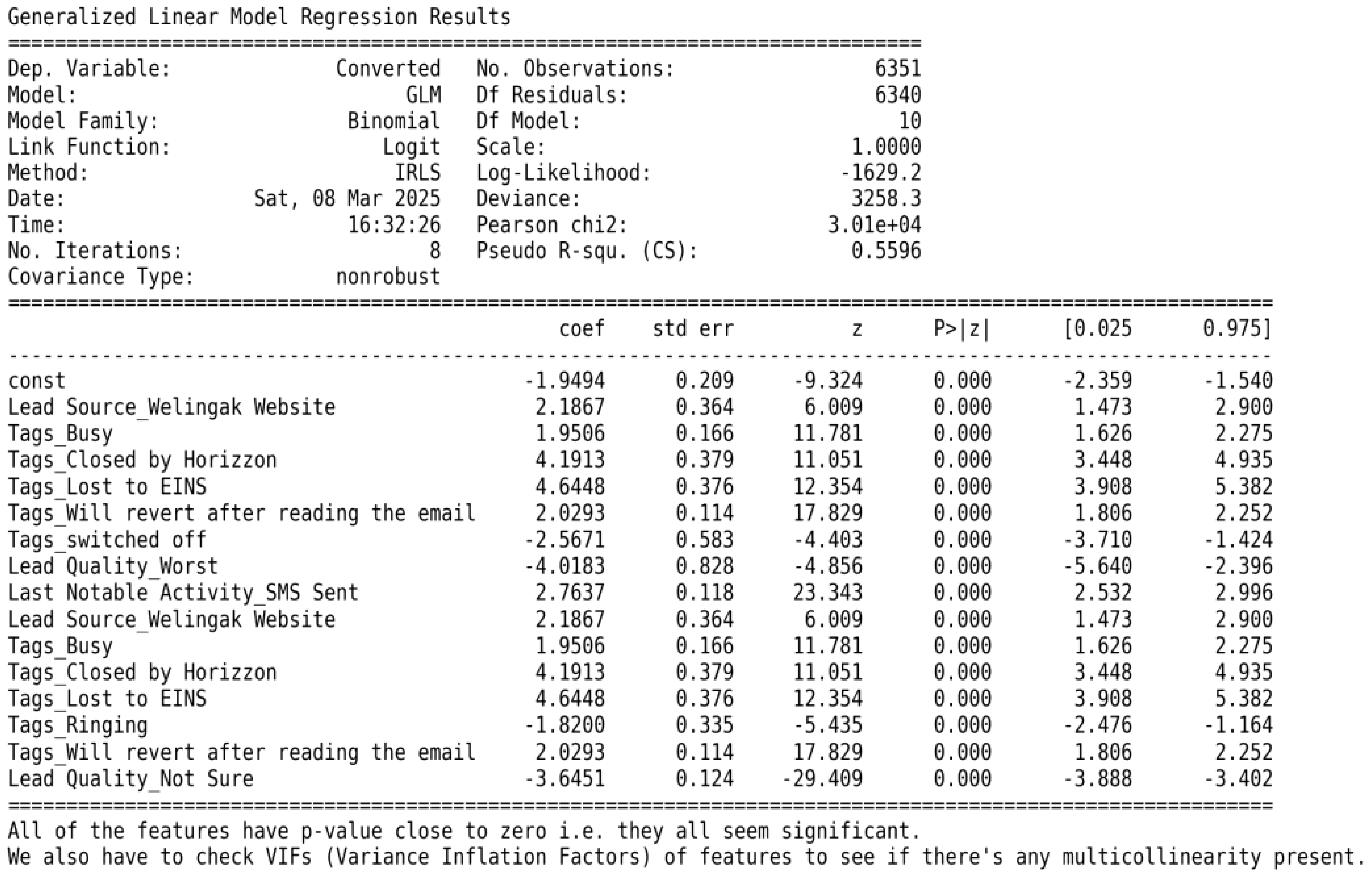
High conversion rates for tags ‘**Will revert after reading the email**’, ‘**Closed by Horizon**’, ‘**Lost to EINS**’, and ‘**Busy**’.

Last Notable Activity



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

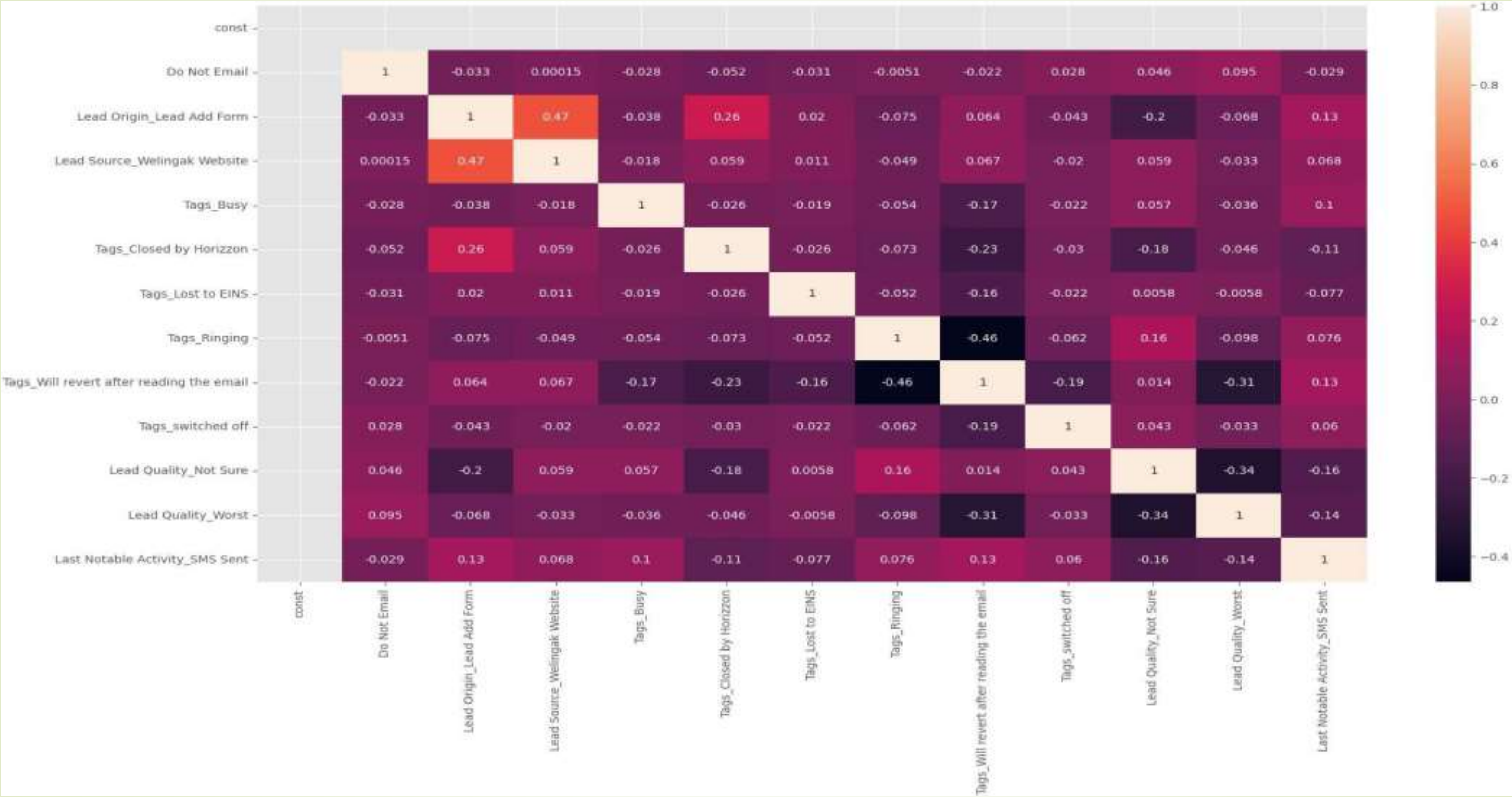
MODEL EVALUATION



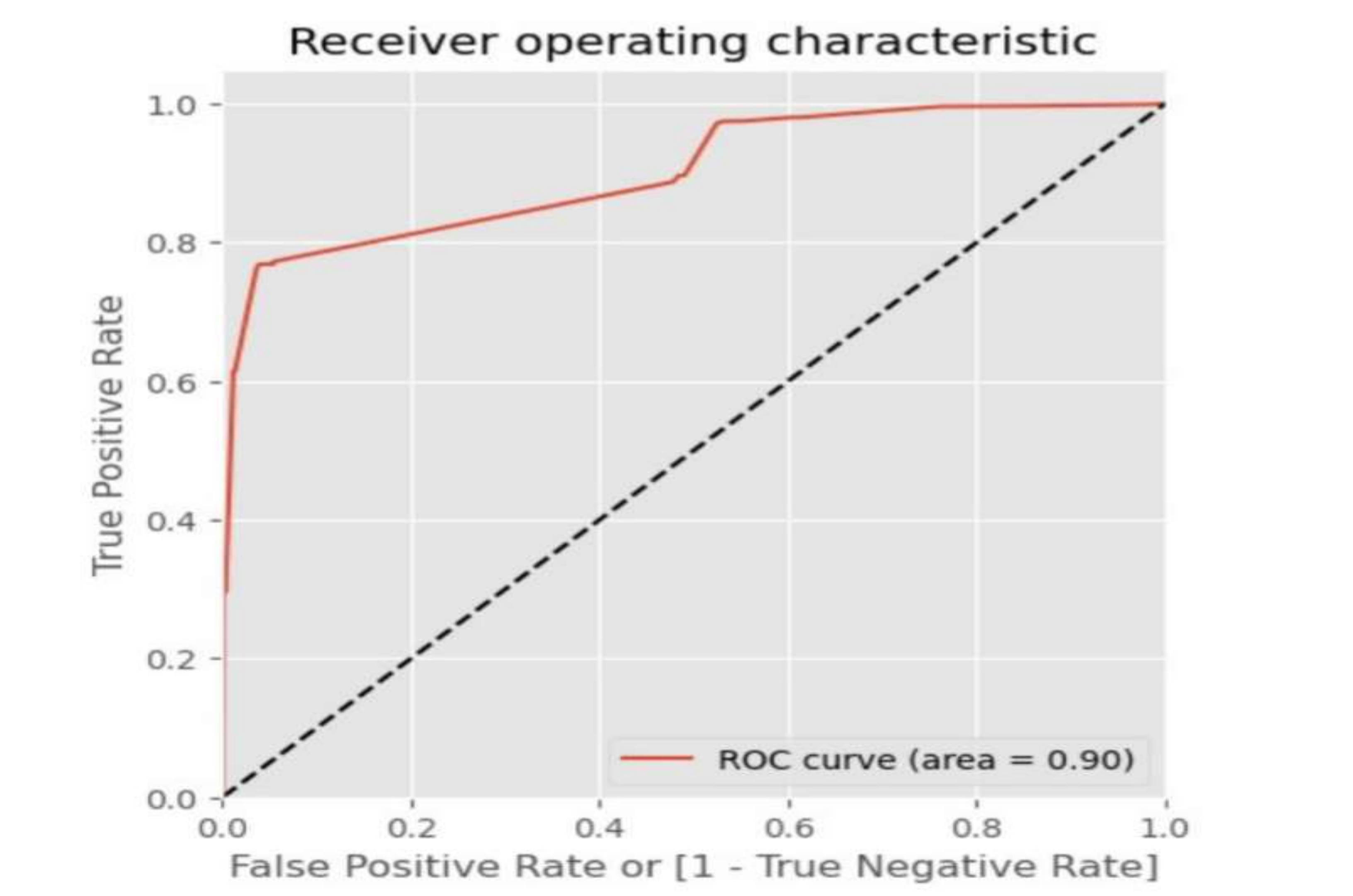
```
Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6340
Model Family:                Binomial   Df Model:                           10
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1629.2
Date:                Sat, 08 Mar 2025   Deviance:                       3258.3
Time:                        16:32:26   Pearson chi2:                 3.01e+04
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5596
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       -1.9494      0.209     -9.324      0.000      -2.359      -1.540
Lead Source_Welingak Website                 2.1867      0.364      6.009      0.000       1.473       2.900
Tags_Busy                                    1.9506      0.166     11.781      0.000       1.626       2.275
Tags_Closed by Horizzon                      4.1913      0.379     11.051      0.000       3.448       4.935
Tags_Lost to EINS                            4.6448      0.376     12.354      0.000       3.908       5.382
Tags_Will revert after reading the email     2.0293      0.114     17.829      0.000       1.806       2.252
Tags_switched off                           -2.5671      0.583     -4.403      0.000      -3.710      -1.424
Lead Quality_Worst                          -4.0183      0.828     -4.856      0.000      -5.640      -2.396
Last Notable Activity_SMS Sent               2.7637      0.118     23.343      0.000       2.532       2.996
Lead Source_Welingak Website                 2.1867      0.364      6.009      0.000       1.473       2.900
Tags_Busy                                    1.9506      0.166     11.781      0.000       1.626       2.275
Tags_Closed by Horizzon                      4.1913      0.379     11.051      0.000       3.448       4.935
Tags_Lost to EINS                            4.6448      0.376     12.354      0.000       3.908       5.382
Tags_Ringing                                -1.8200      0.335     -5.435      0.000      -2.476      -1.164
Tags_Will revert after reading the email     2.0293      0.114     17.829      0.000       1.806       2.252
Lead Quality_Not Sure                       -3.6451      0.124    -29.409      0.000      -3.888      -3.402
============================================================================================================
```

All of the features have p-value close to zero i.e. they all seem significant.

We also have to check VIFs (Variance Inflation Factors) of features to see if there's any multicollinearity present.

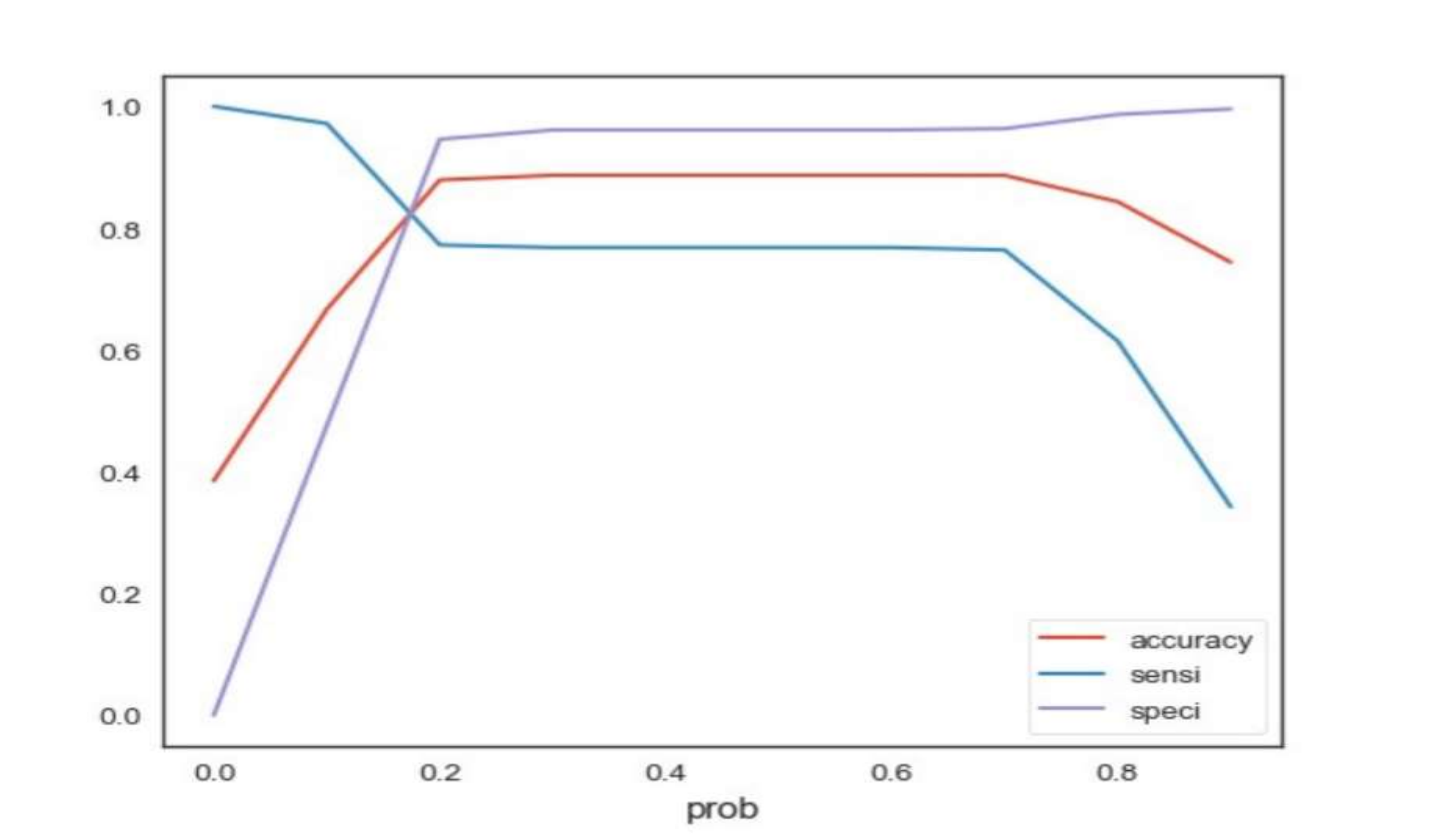


Correlations between features in the final model are **negligible.**



Area under curve = 0.90

Finding Optimal Threshold



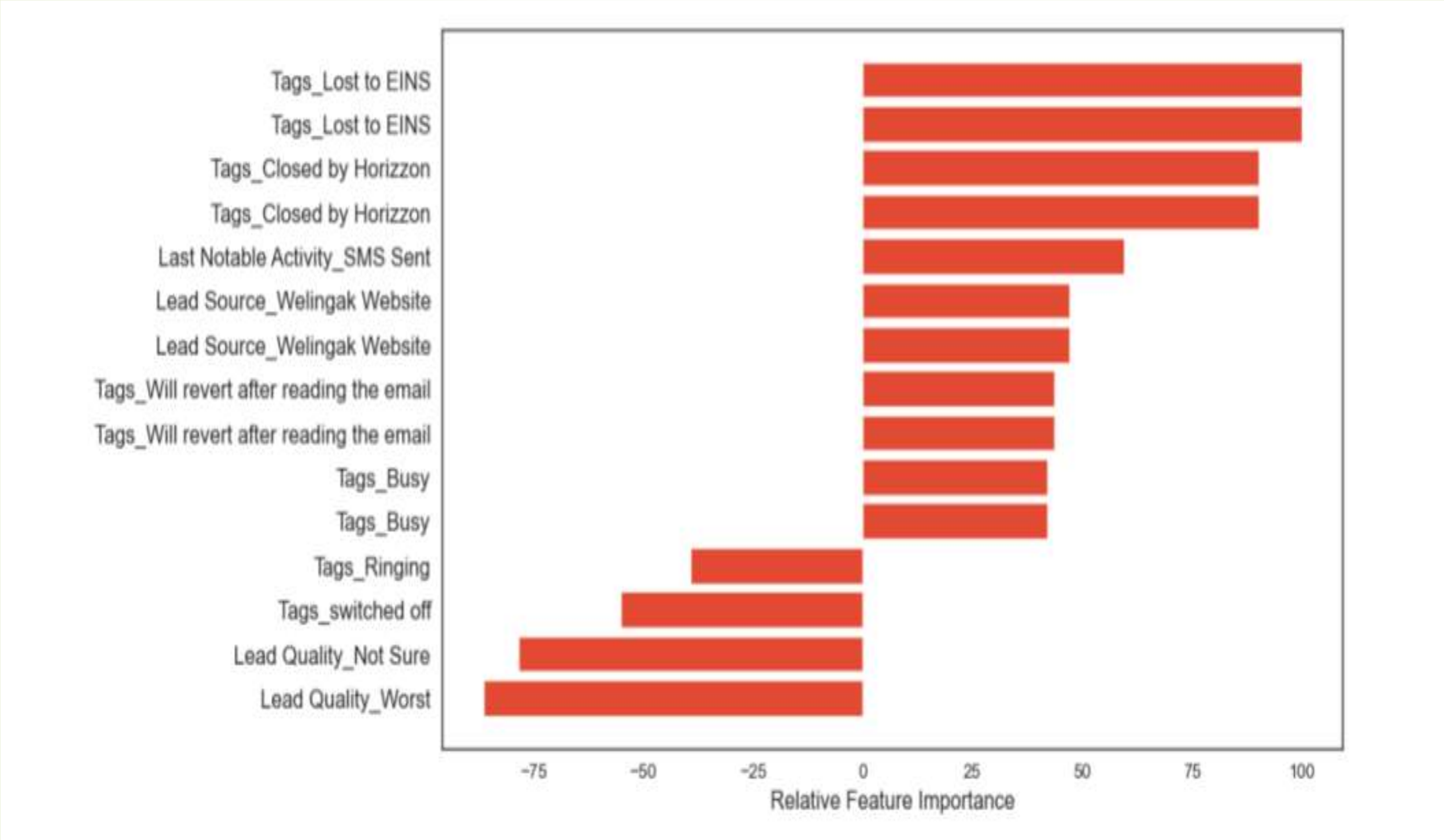
Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values

Optimal cutoff = 0.20

Final Results

Data	Train set	Test set
Accuracy	**0.8776**	**0.9070**
Sensitivity	0.7686	0.8402
Specificity	0.9460	0.9452
False Positive Rate	0.0540	0.0541
Positive Predictive Value	0.8991	0.8974
Negative Predictive Value	0.8671	0.9120
AUC	0.9009	0.9362

Relative Importance Of Features



INFERENCES

Feature Importance

- Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:
    - ***Tags_Lost to EINS***
    - **Tags_Closed by Horizzon**
    - **Tags_Wil revert after reading the email**
- These are dummy features created from the categorical variable Tags.
- All three **contribute positively** towards the probability of a lead conversion.
- These results indicate that the company should **focus more on the leads with these three tags**.

Recommendations

- By referring to the data visualizations, focus on
  - *Increasing the conversion rates for th generating categories more leads and conversion.*
  - *Generating more leads for categories having high rates.*

- Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.
- Based on varying business needs, modify the probability threshold value for identifying potential leads.

THANK YOU