

LEAD SCORING CASE STUDY

Using logistic regression technique

METHODOLOGY

To built 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%

Receiving and Understanding the Data

- Importing and Observing the past data provided by the Company

Data Cleaning

- Missing value imputation
- Removing duplicate data and other redundancies

Exploratory Data Analysis

- Univariate and Bivariate analysis

Data Preparation

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

Model Building

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

Model Evaluation

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

Assigning Lead Scores

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

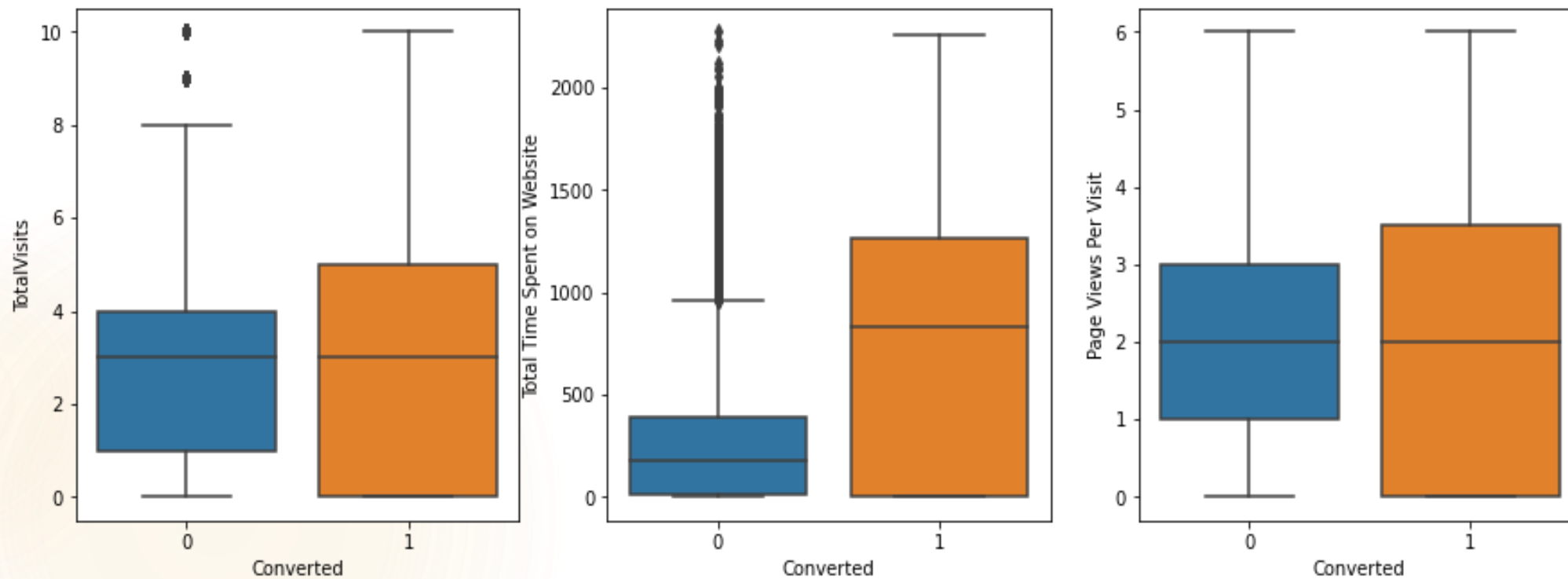
Conclusion

- Final Prediction

DATA VISUALIZATION

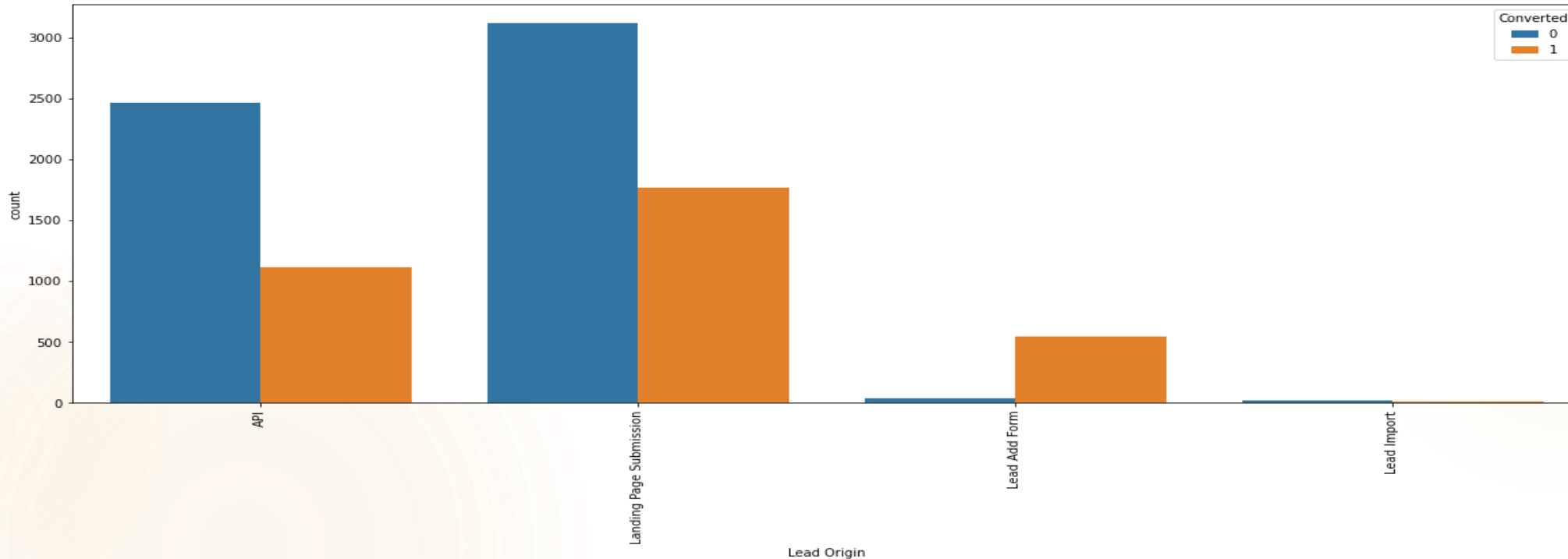
- ▶ To identify important features
- ▶ To get insights

Numerical Variables



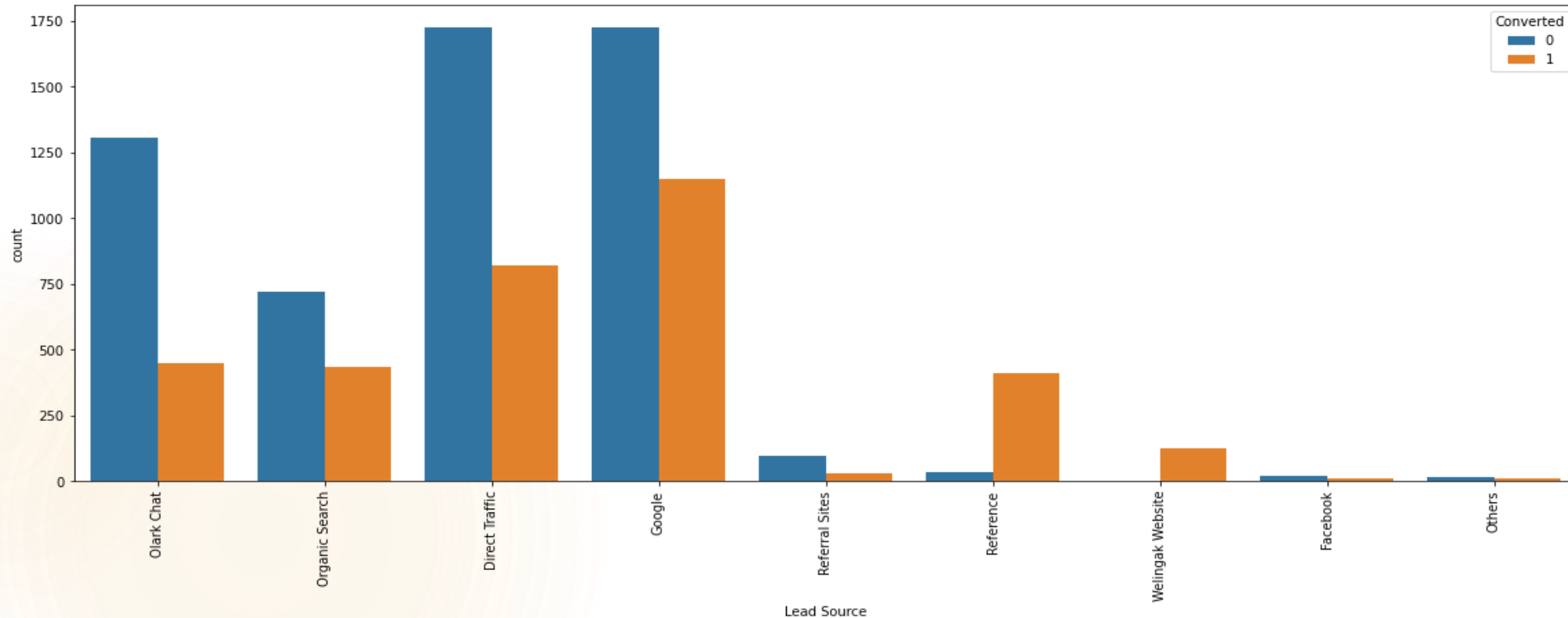
- ▶ People spending more time on website are more likely to get converted.

Lead Origin



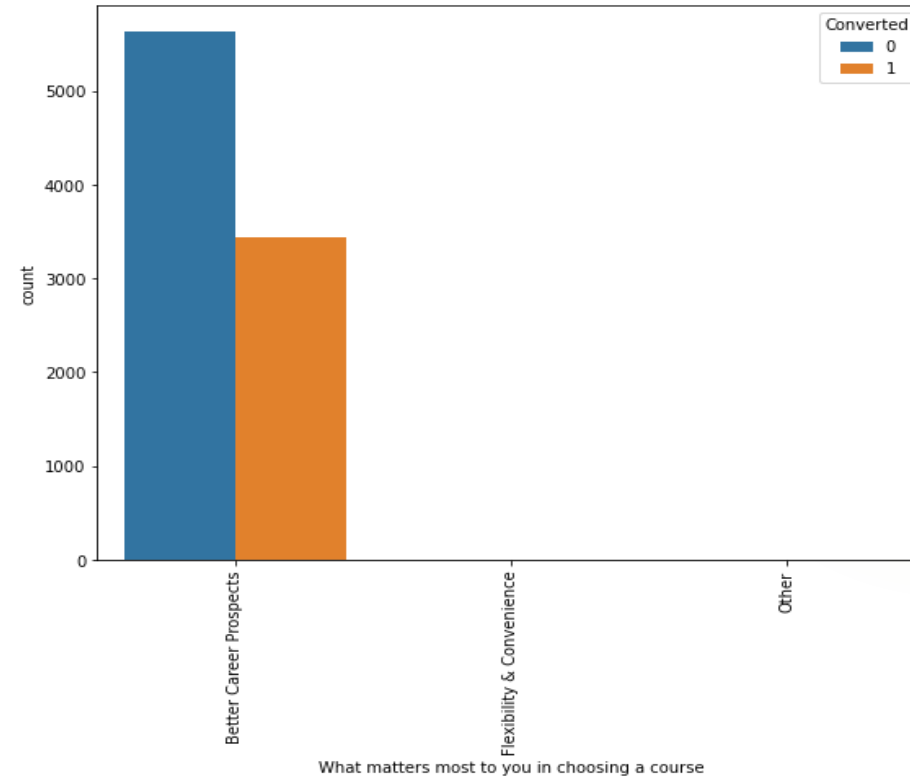
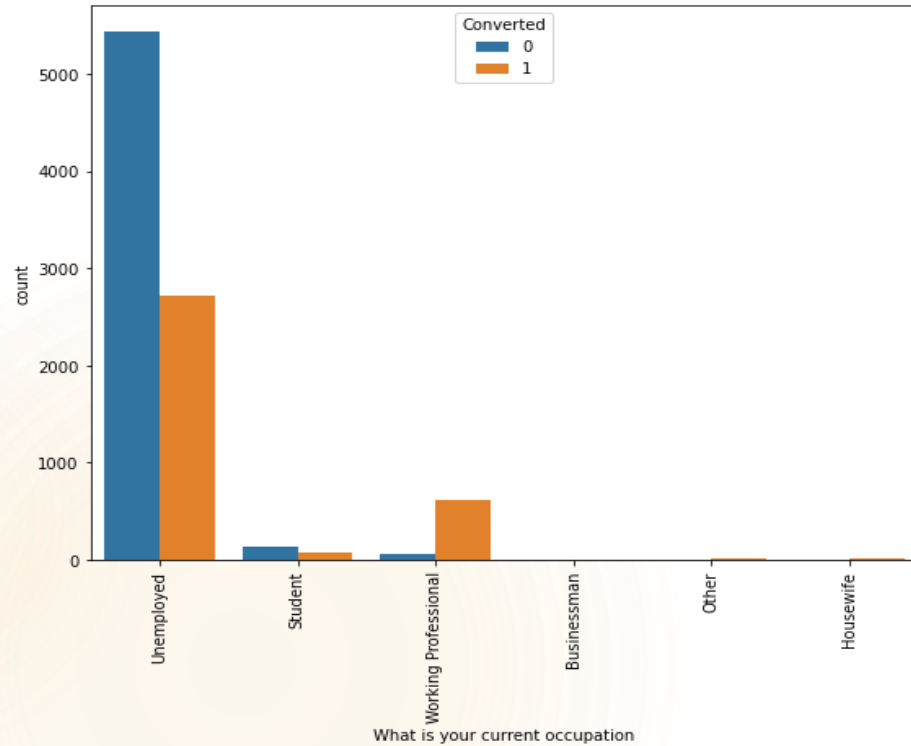
- 'API' and 'Landing Page Submission' generate the most leads but have less conversion rates, whereas 'Lead Add Form' generates less leads but conversion rate is great.
- Try to Increase conversion rate for 'API' and 'Landing Page Submission', and increase leads generation using 'Lead Add Form'

Lead Source



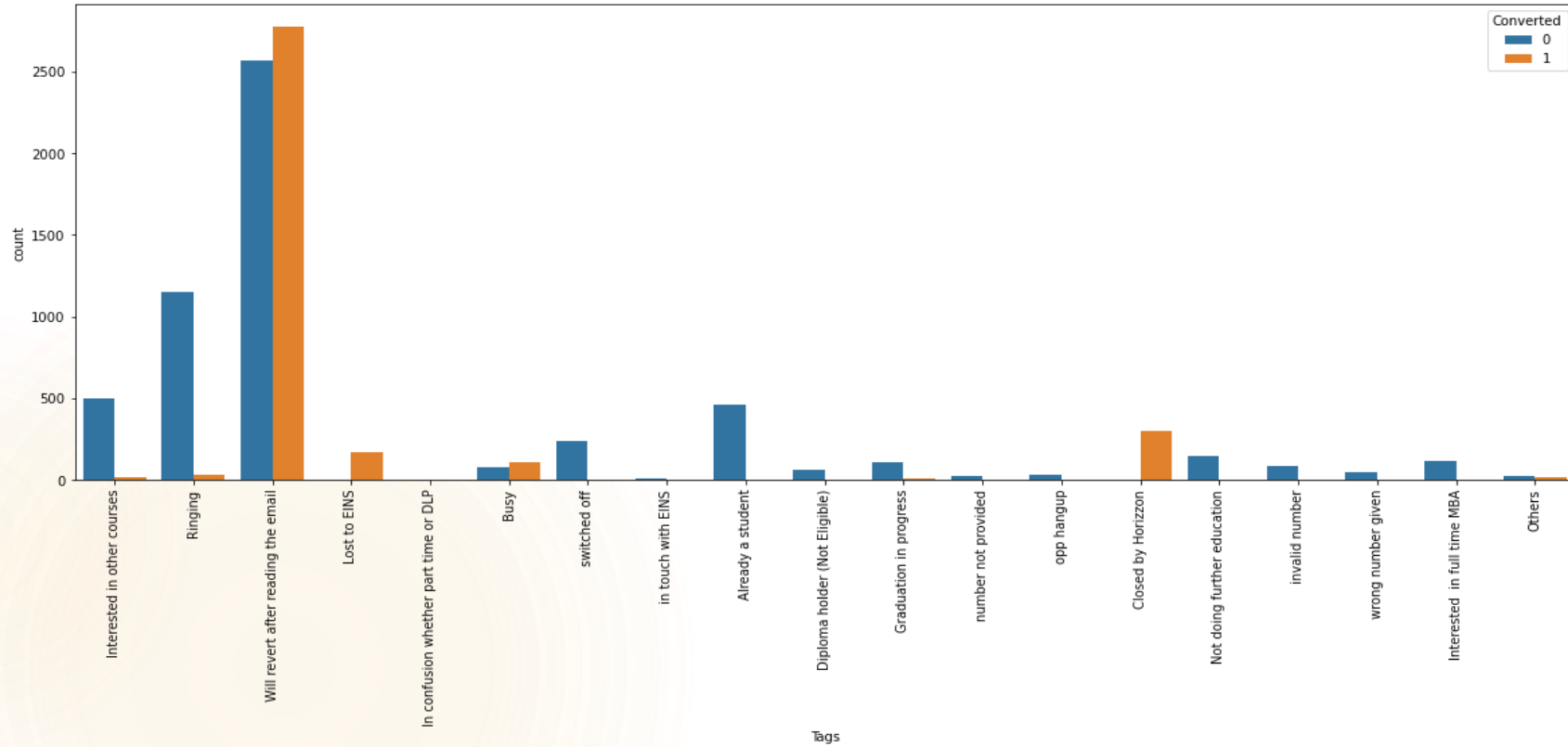
- Very high conversion rates for lead sources 'Reference' and 'Wellngak website'.
- Most leads are generated through 'Direct Traffic' and 'Google'

Current Occupation



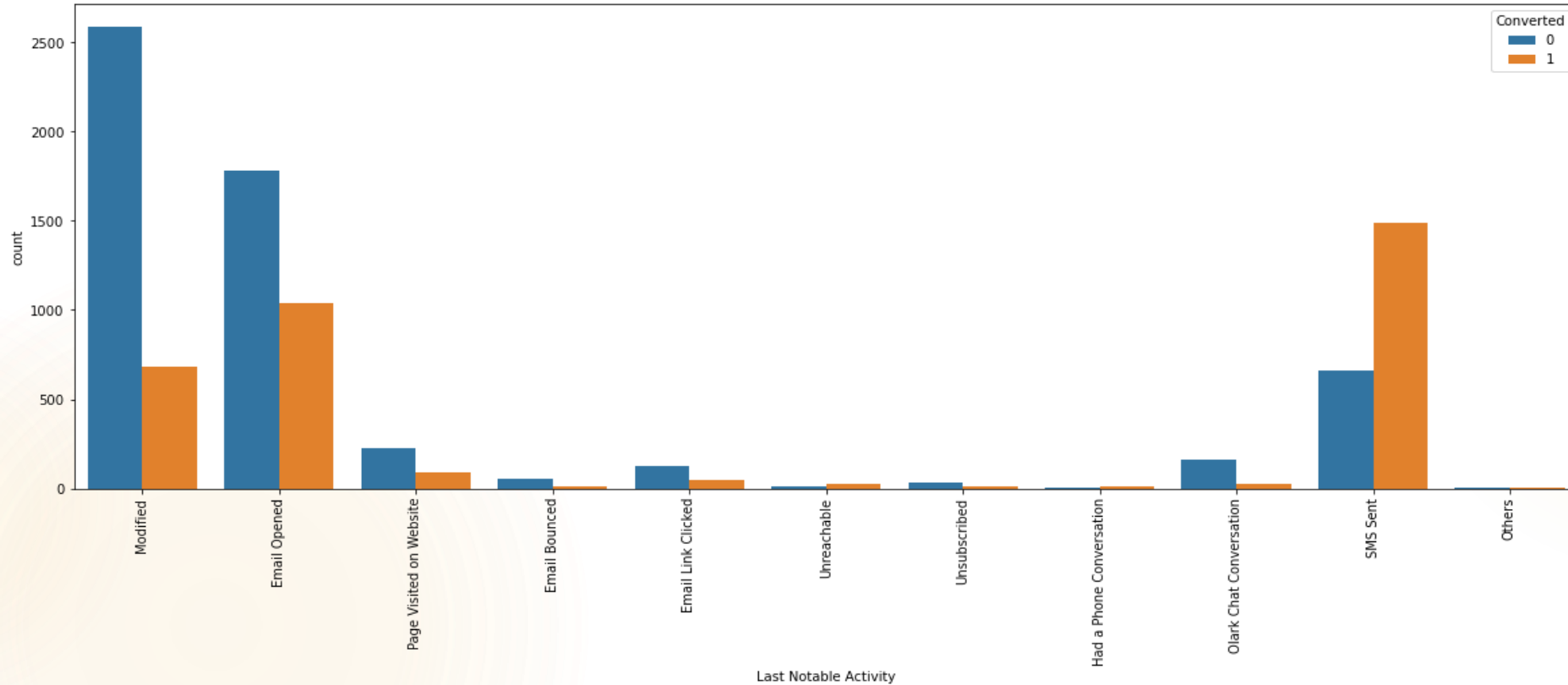
- Working Professionals are most likely to get converted.

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 'SMS Sent'.

MODEL EVALUATION

Generalized Linear Model Regression Results

```

=====
Dep. Variable:          Converted    No. Observations:          6351
Model:                  GLM         Df Residuals:              6338
Model Family:          Binomial    Df Model:                  12
Link Function:         logit       Scale:                    1.0000
Method:                IRLS        Log-Likelihood:           -1601.0
Date:                  Mon, 17 Oct 2022    Deviance:                3202.0
Time:                  12:09:05    Pearson chi2:            3.48e+04
No. Iterations:        8
Covariance Type:       nonrobust
=====

```

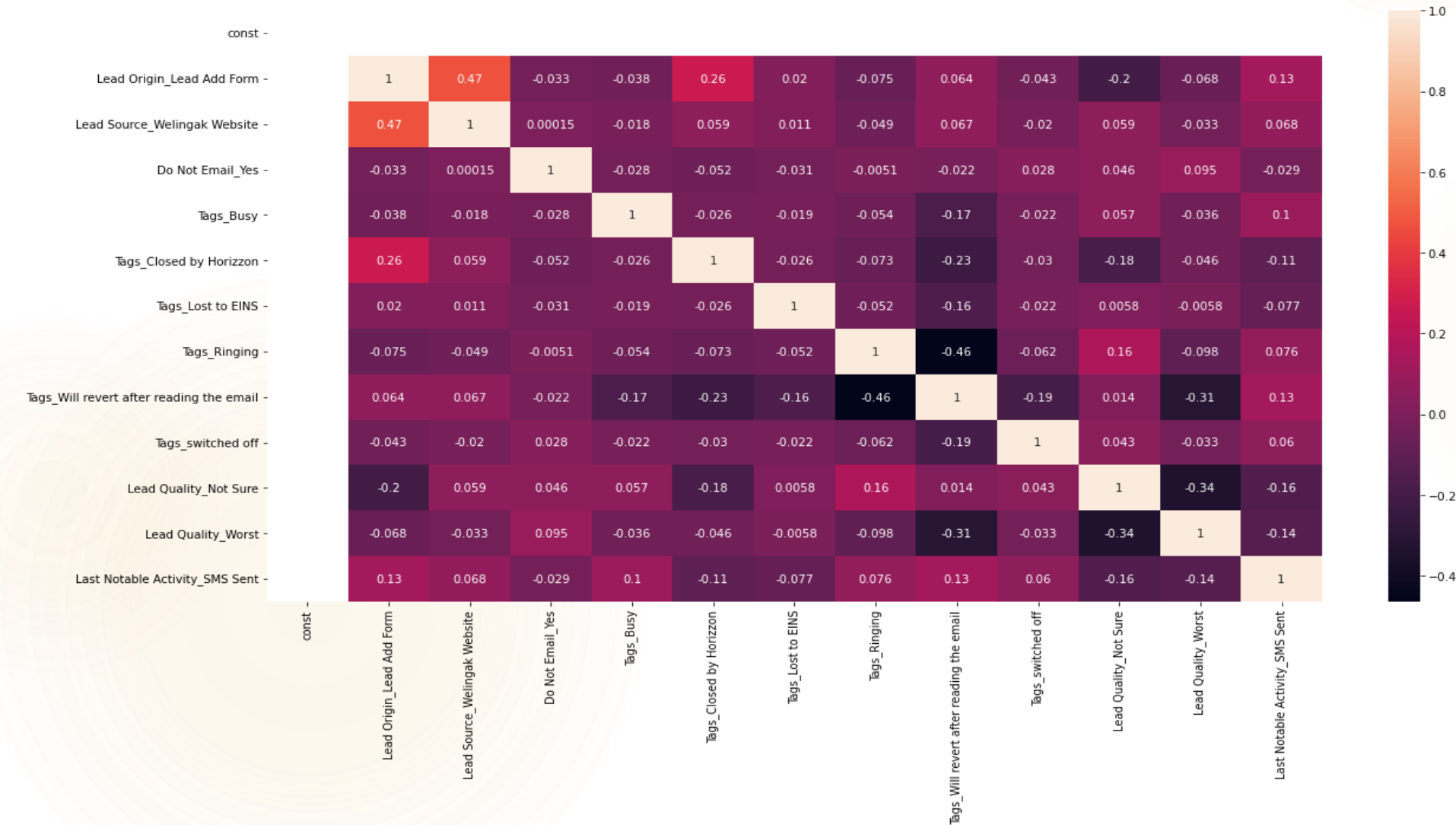
```

=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const              -1.9192      0.211     -9.080      0.000     -2.333     -1.505
Lead Origin_Lead Add Form      1.2035      0.368      3.267      0.001      0.482      1.925
Lead Source_Welingak Website   3.2825      0.820      4.002      0.000      1.675      4.890
Do Not Email_Yes              -1.2835      0.212     -6.062      0.000     -1.698     -0.868
Tags_Busy                   3.8043      0.330     11.525      0.000      3.157      4.451
Tags_Closed by Horizzon       7.9789      0.762     10.467      0.000      6.485      9.473
Tags_Lost to EINS             9.1948      0.753     12.209      0.000      7.719     10.671
Tags_Ringing                -1.8121      0.336     -5.401      0.000     -2.470     -1.154
Tags_Will revert after reading the email   3.9906      0.228     17.508      0.000      3.544      4.437
Tags_switched off            -2.4456      0.586     -4.171      0.000     -3.595     -1.297
Lead Quality_Not Sure        -3.5218      0.126    -28.036      0.000     -3.768     -3.276
Lead Quality_Worst          -3.9106      0.856     -4.567      0.000     -5.589     -2.232
Last Notable Activity_SMS Sent   2.7395      0.120     22.907      0.000      2.505      2.974
=====

```

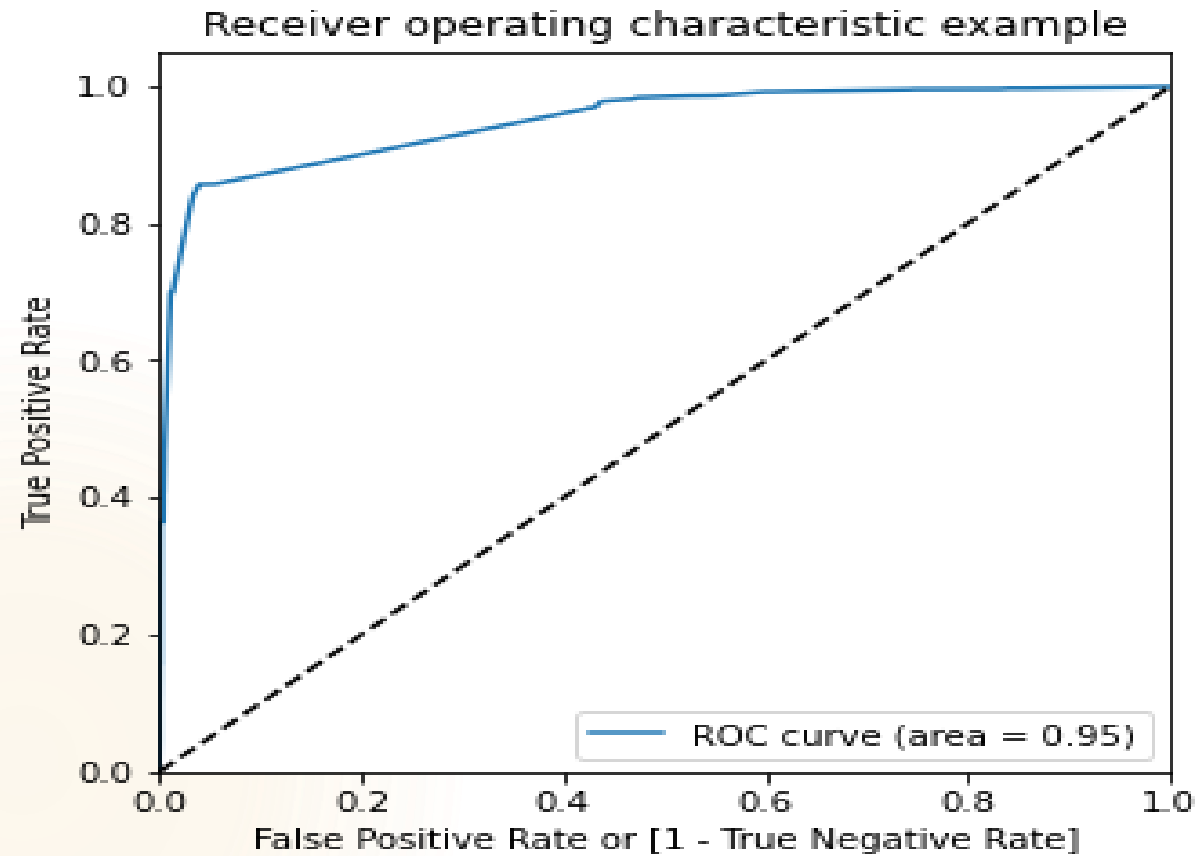
Final Model Summary : All p-values are zero

Heatmap



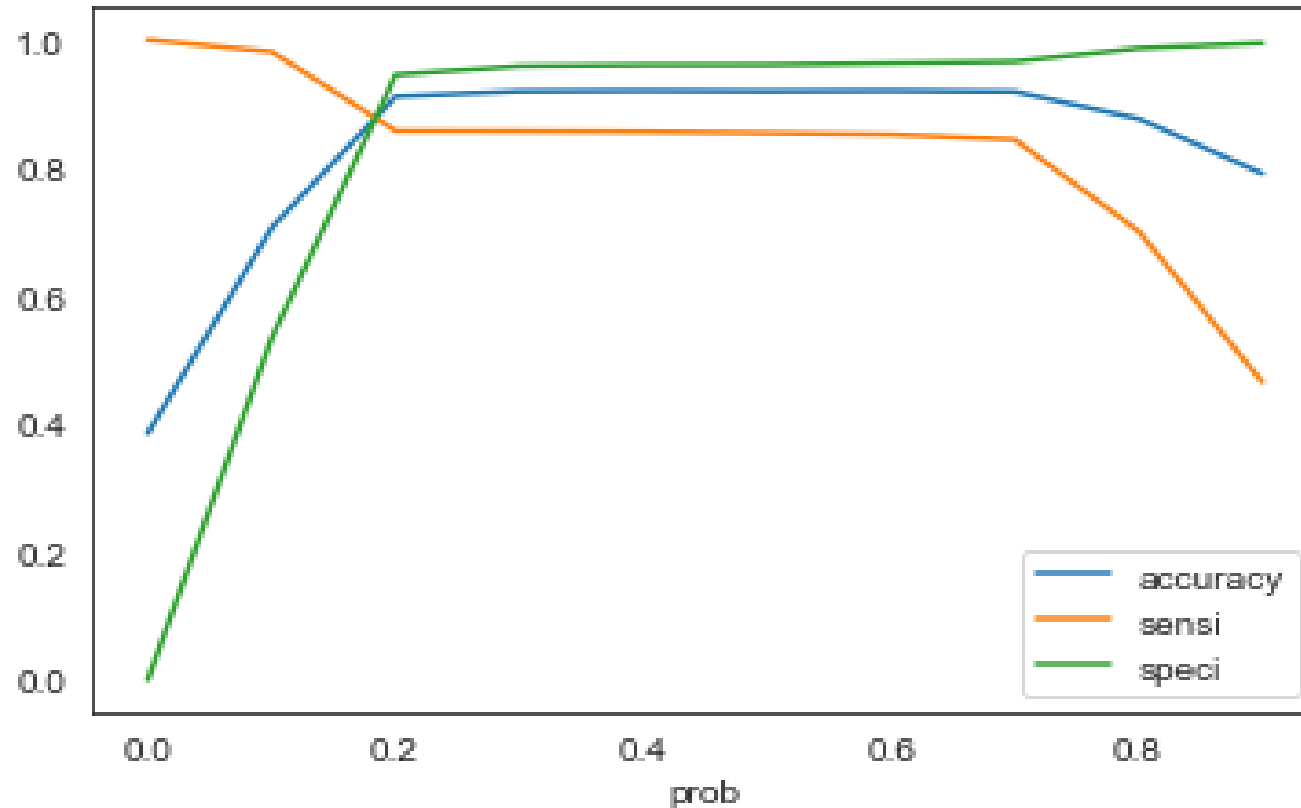
- Correlations between features in the final model are negligible.

ROC Curve



- Area under curve are 95%

Finding Optimal Threshold



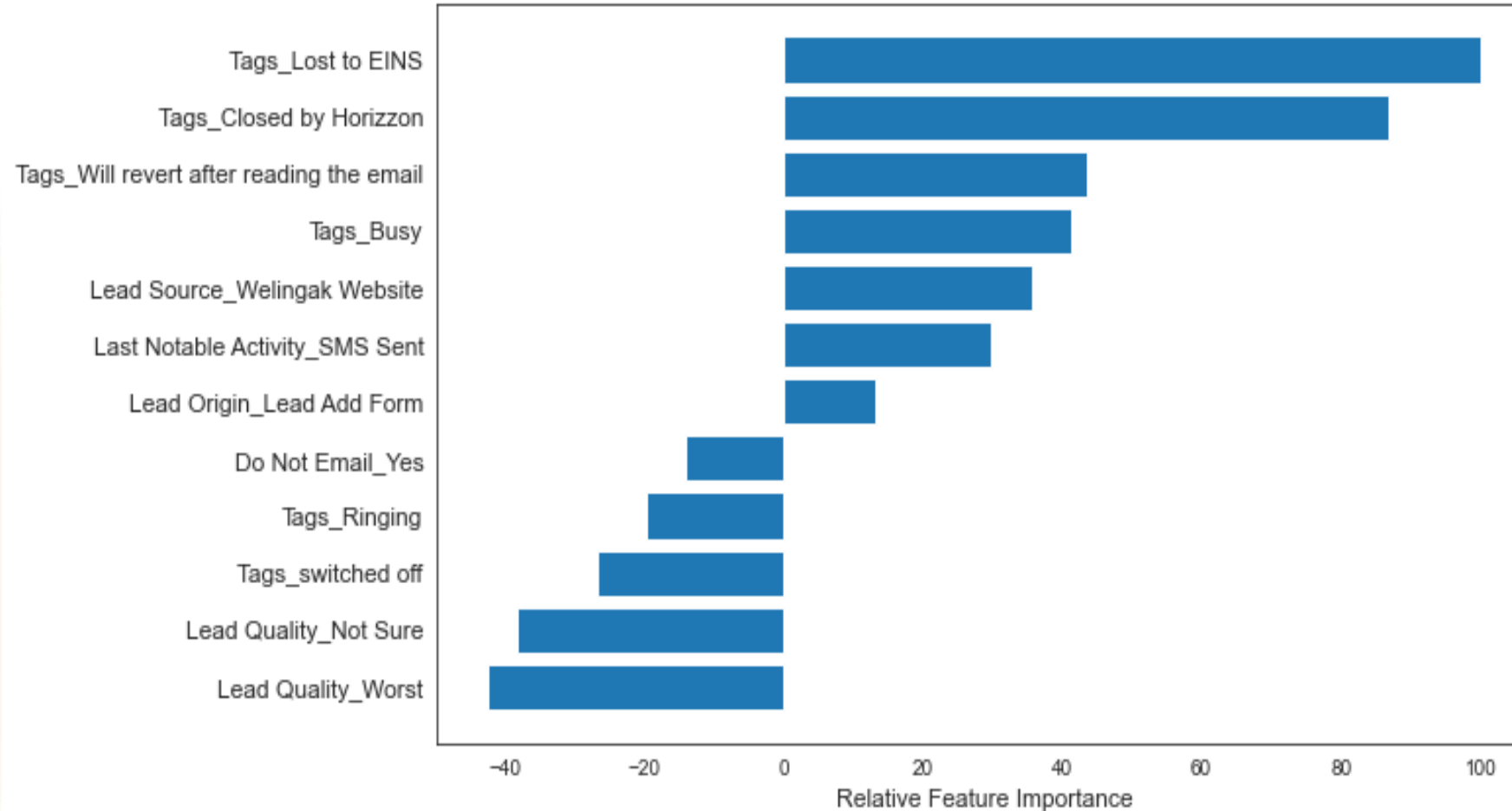
Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values

Optimal cutoff = 0.20

Final Results

Dataset	Accuracy	Sensitivity	Specificity	False Positive Rate	Positive Predictive Value	Negative Predictive Value	AUC
Train	0.9111	0.8573	0.9449	0.0550	0.9070	0.9135	0.9488
Test	0.9078	0.8412	0.9457	0.0542	0.8984	0.9126	0.9388

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:
 1. Tags_Lost to EINS
 2. Tags_Closed by Horizzon
 3. Tags_Will 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.

Situation 1: Company has interns for 2 months. They wish to make lead conversion more aggressive. They want almost all of the potential leads to be converted and hence, want to make phone calls to as much of such people as possible.

Solution:

➤ ***Sensitivity = True Positives / (True Positives + False Negatives)***

➤ Sensitivity can be defined as the number of actual conversions predicted correctly out of total number of actual conversions. As we saw earlier, sensitivity decreases as the threshold increases.

➤ High sensitivity implies that our model will correctly predict almost all leads who are likely to convert. At the same time, it may overestimate and misclassify some of the non-conversions as conversions.

➤ As the company has extra man-power for two months and wants to make the lead conversion more aggressive, it is a good strategy to go for high sensitivity. To achieve high sensitivity, we need to choose a low threshold value.

Situation 2: At times, the company reaches its target for a quarter before the deadline. It wants the sales team to focus on some new work. So during this time, the company's aim is to not make phone calls unless it's extremely necessary.

Solution:

➤ ***Specificity = True Negatives / (True Negatives + False Positives)***

- Specificity can be defined as the number of actual non-conversions predicted correctly out of total number of actual non-conversions. It increases as the threshold increases.
- High specificity implies that our model will correctly predict almost all leads who are not likely to convert. At the same time, it may misclassify some of the conversions as non-conversions.
- As the company has already reached its target for a quarter and doesn't want to make unnecessary phone calls, it is a good strategy to go for high specificity.
- It will ensure that the phone calls are only made to customers who have a very high probability of conversion. To achieve high specificity, we need to choose a high threshold value.

Recommendations

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