

Lead Scoring Case Study

Optimizing Lead Scoring with Logistic Regression

Submitted by

Tejaswini D V

Problem Statement

- ❑ X Education generates a significant volume of leads through its website and referrals, yet the conversion rate remains at approximately 30%.
- ❑ The sales team interacts with all leads, regardless of their potential, which leads to inefficient use of resources.
- ❑ The company faces challenges in pinpointing high-potential leads, leading to wasted efforts and missed opportunities. Consequently, the conversion process remains inefficient, and valuable prospects are often overlooked.

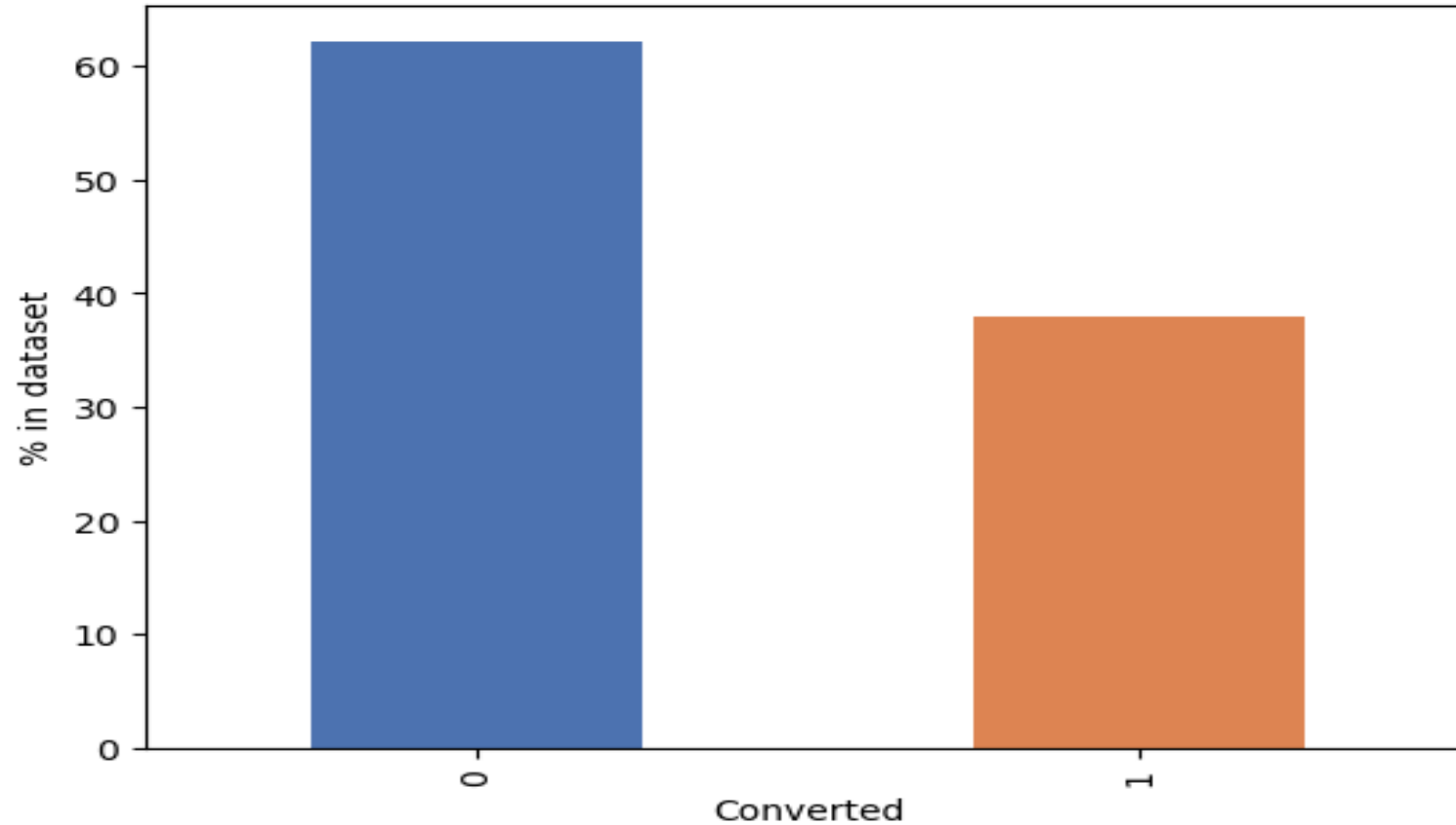
Business Objective

- ❑ The goal is to develop a lead scoring model to prioritize high-potential leads based on their likelihood to convert.
- ❑ By implementing lead scoring, the sales team can prioritize high-potential prospects, enhancing the efficiency of the conversion process.
- ❑ The goal is to raise the conversion rate to 80%, streamlining the sales process, boosting revenue, and enhancing resource allocation for sales and marketing.

Steps Followed

- ❑ Reading Data
- ❑ Cleaning Data
- ❑ Data Visualization
- ❑ Data Preparation
- ❑ Model Building
- ❑ ROC Curve
- ❑ Model Evaluations
- ❑ Prediction on test set
- ❑ Conclusion

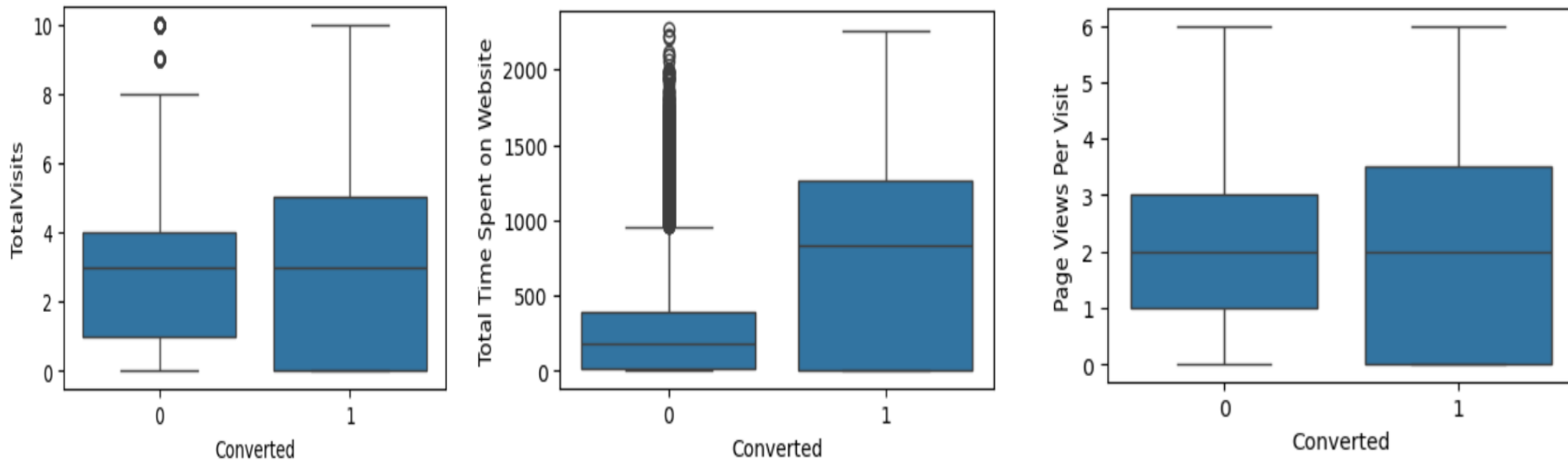
Target variable



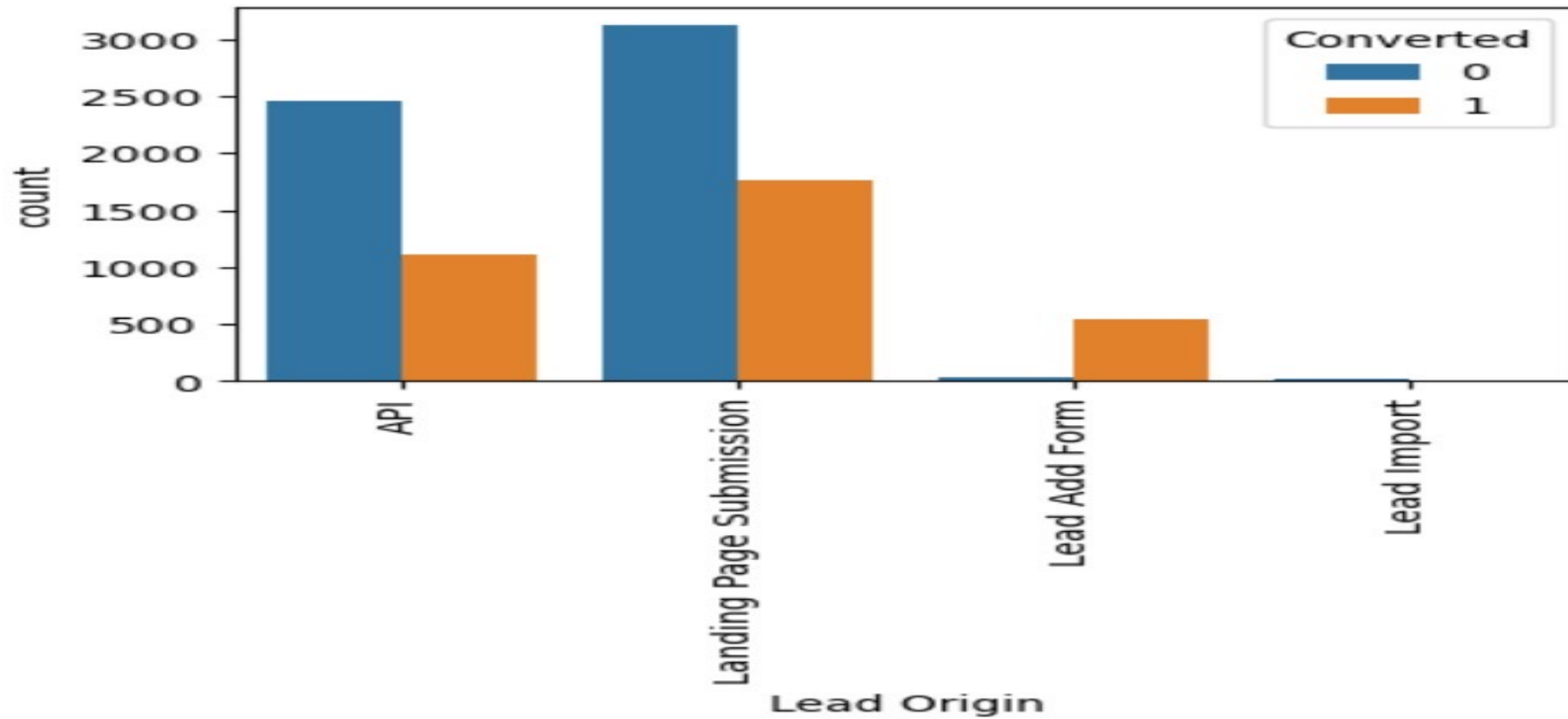
The data is imbalanced, with 38% of leads converting and 62% not converting, showing a disproportionate distribution between the two classes.

Data Visualization

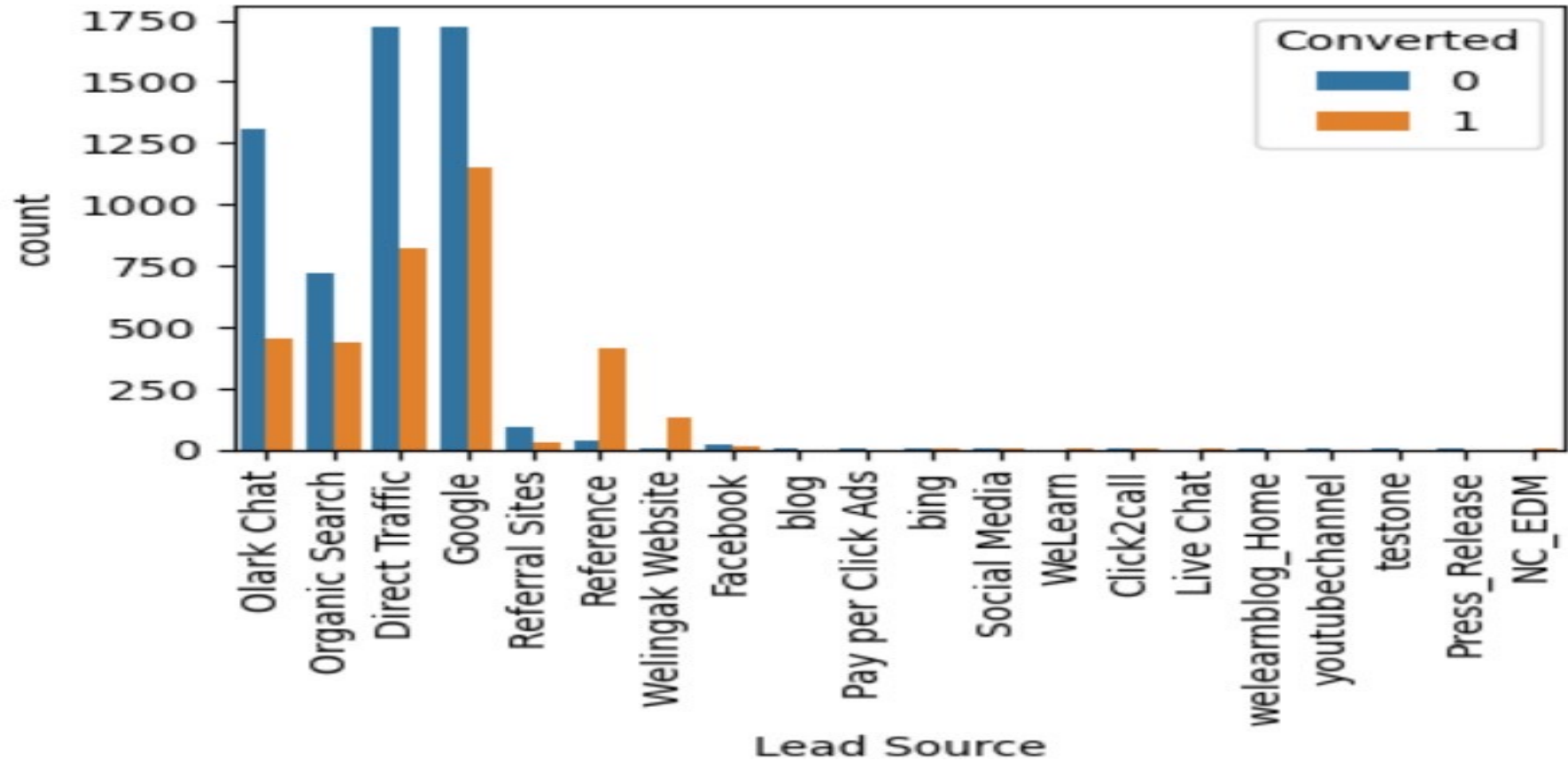
Identifying important features



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

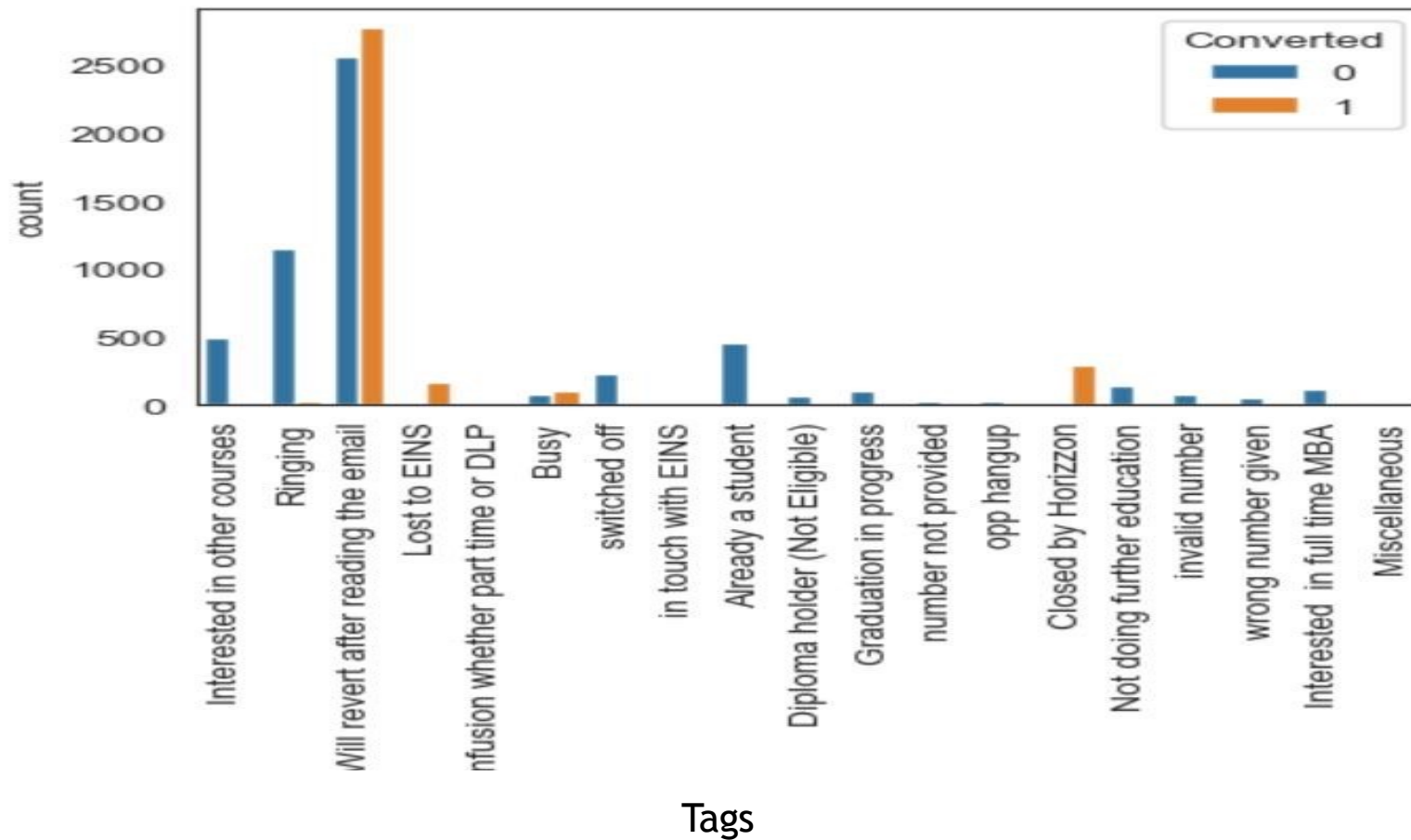


The 'API' and 'Landing page submission' channels generate the highest number of leads but have lower conversion rates, whereas the 'Lead add form' produces fewer leads but achieves a higher conversion rate.



* High conversion rate for lead sources '**Reference**' and '**Welingak Website**'.

* Most of the leads are generated through '**Direct Traffic**' and '**Google**'.



Tags like 'Will revert after reading the email', 'Busy', and 'Closed by Horizzon' show high conversion rates, indicating engaged leads. Strategic follow-ups can maximize their conversion potential.

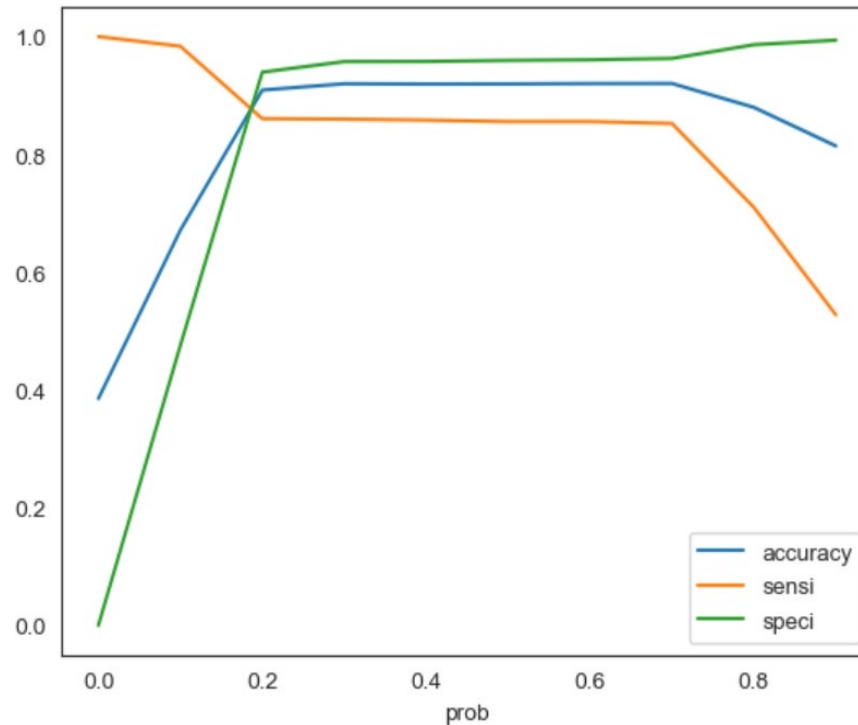
Model Evaluation

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:           -1611.4
Date:                   Mon, 16 Dec 2024   Deviance:                3222.8
Time:                   11:02:30    Pearson chi2:             2.61e+04
No. Iterations:         8           Pseudo R-squ. (CS):       0.5620
Covariance Type:        nonrobust
=====
```

```
=====
                                coef    std err          z      P>|z|      [0.025    0.975]
-----
const                        -2.1838      0.216    -10.106     0.000     -2.607    -1.760
Lead Origin_Lead Add Form      1.0856      0.361      3.006     0.003      0.378     1.794
Lead Source_Welingak Website   3.3105      0.811      4.080     0.000      1.720     4.901
What is your current occupation_Working Professional  1.2699      0.282      4.506     0.000      0.718     1.822
Tags_Busy                      3.8972      0.331     11.781     0.000      3.249     4.546
Tags_Closed by Horizzon       8.0859      0.763     10.596     0.000      6.590     9.582
Tags_Lost to EINS             9.2585      0.753     12.294     0.000      7.782    10.735
Tags_Ringing                  -1.6717      0.336     -4.979     0.000     -2.330    -1.014
Tags_Will revert after reading the email  4.0114      0.230     17.477     0.000      3.562     4.461
Tags_switched off             -2.3371      0.584     -4.003     0.000     -3.481    -1.193
Lead Quality_Not Sure         -3.3829      0.128    -26.455     0.000     -3.633    -3.132
Lead Quality_Worst            -3.8266      0.842     -4.547     0.000     -5.476    -2.177
Last Notable Activitv SMS Sent  2.7506      0.119     23.137     0.000      2.518     2.984
=====
```

Finding Optimal Threshold



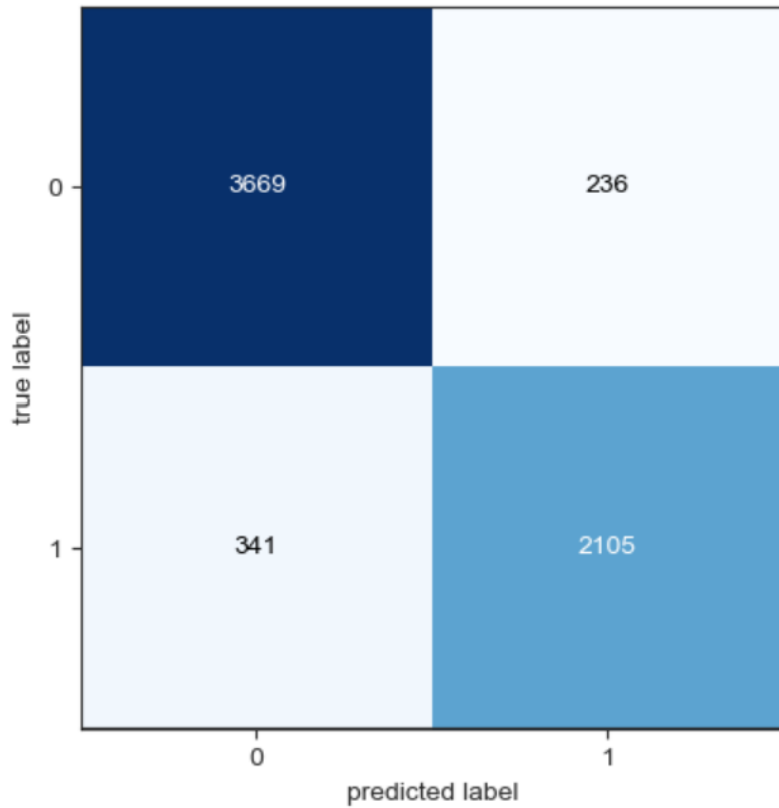
optimal cutoff = 0.20

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

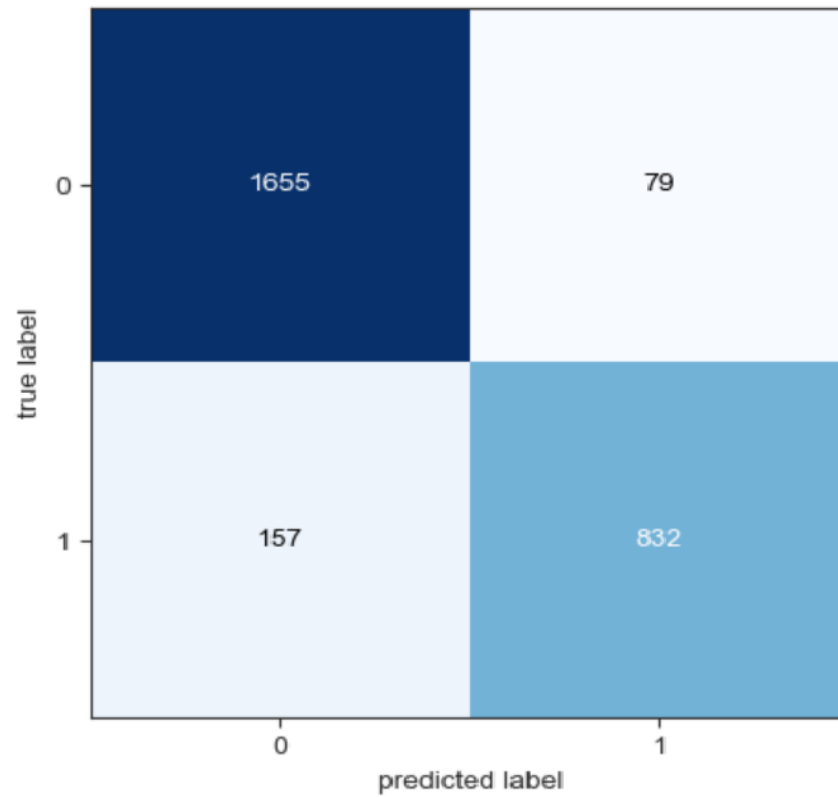
FINAL RESULTS

Data	Train Set	Test Set
Accuracy	0.9091	0.9133
Sensitivity	0.8605	0.8412
Specificity	0.9395	0.9544
False Positive Rate	0.0604	0.0455
Positive Predictive Value	0.8991	0.9132
Negative Predictive Value	0.9149	0.9133
AUC	0.9461	0.9372

CONFUSION MATRIX

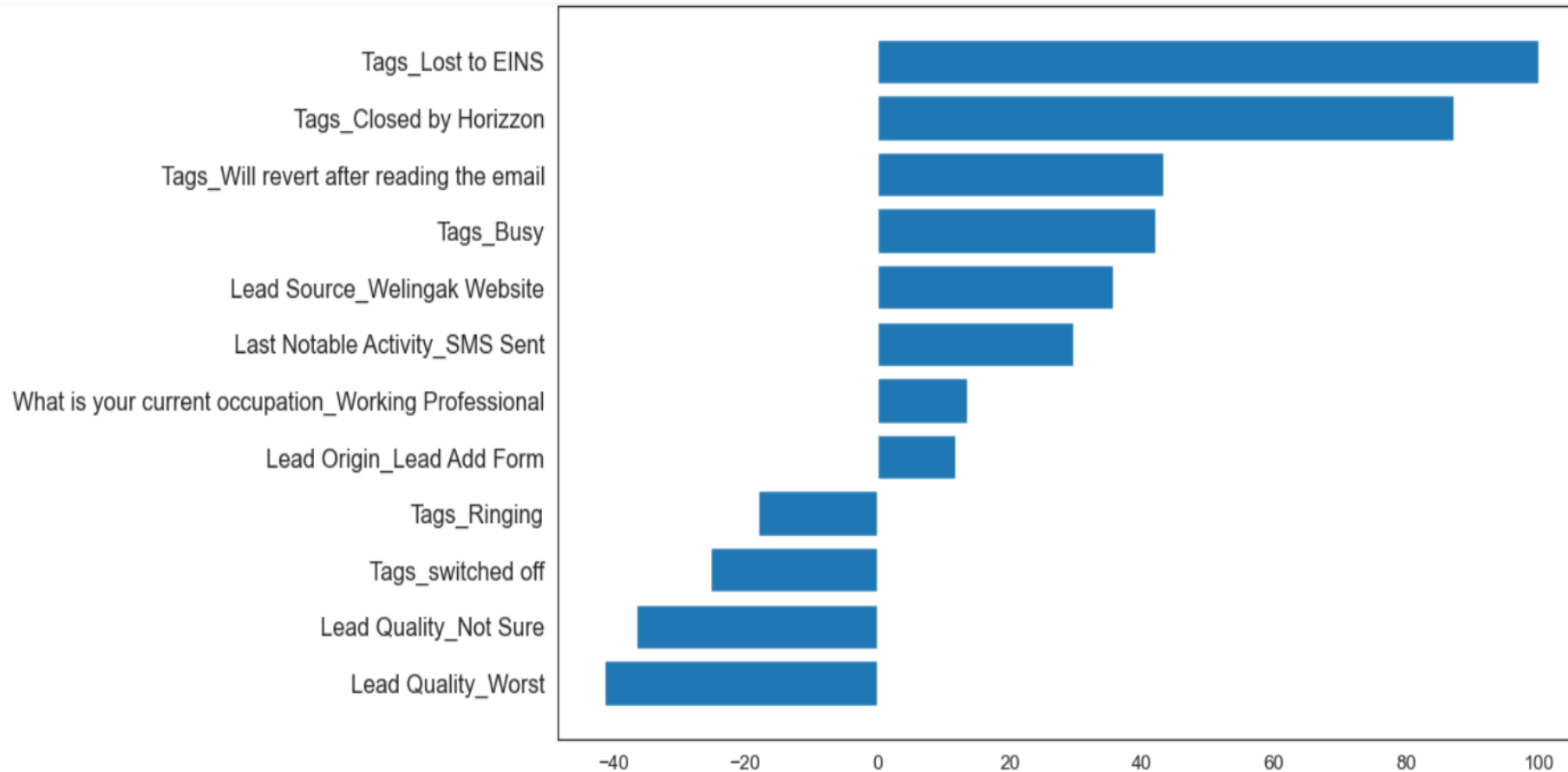


For Train Set



For Test Set

Relative Importance Of Features



RECOMMENDATIONS

- ❑ Leads with a high score can be categorized as 'hot' leads and should be given priority by the sales team, as they have a strong potential for conversion.
- ❑ Leads classified as 'Might Be' or 'Worst' based on previous interactions can be given lower priority, as their likelihood of conversion is low.
- ❑ Leads who have shared their contact numbers through the website or email but are marked as 'Busy' or 'Ringing' (i.e., not answering calls) should be considered as lower priority and can be ignored as well, as they are less likely to engage or convert as customers.

THANK YOU