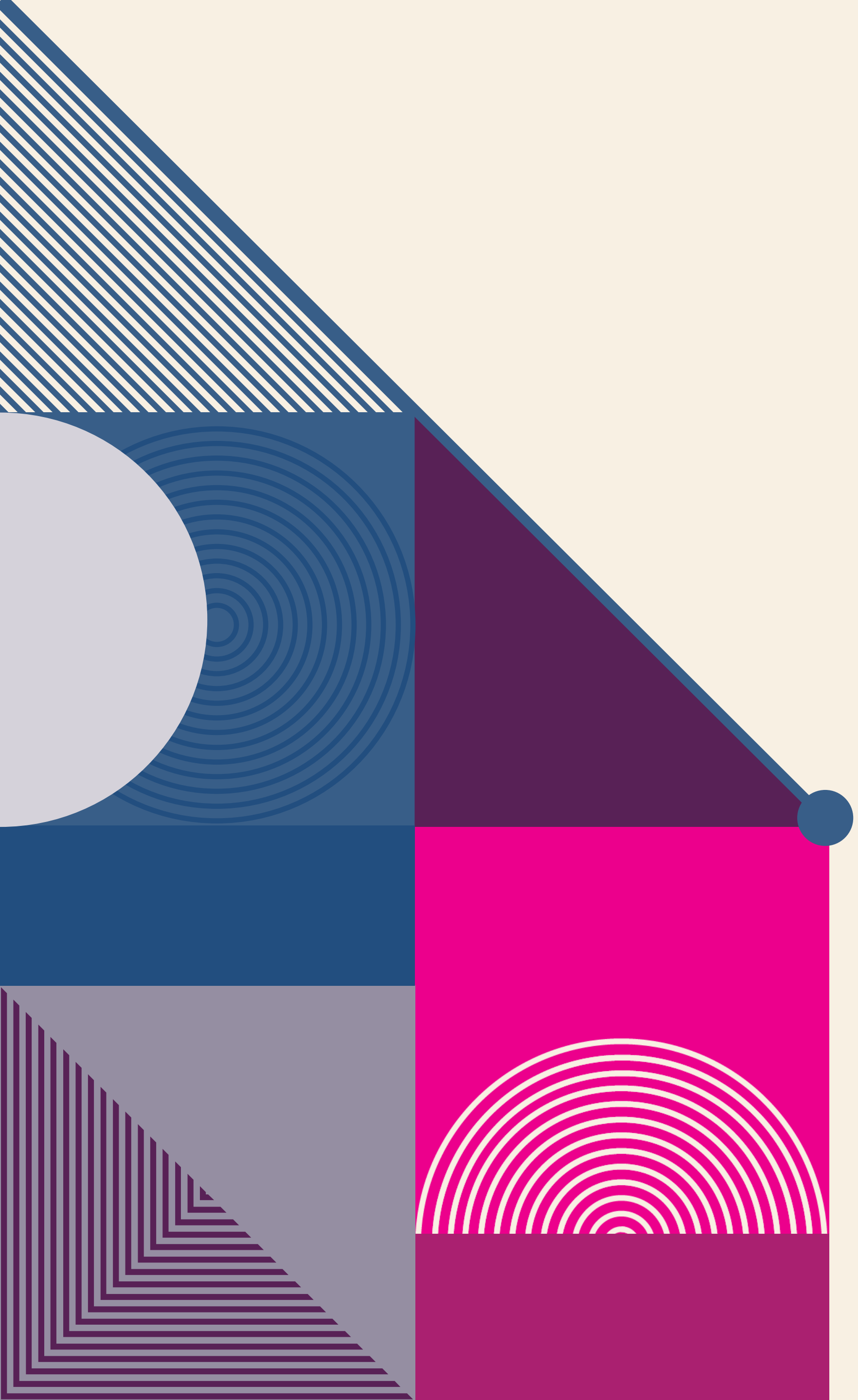
LEAD CONVERSION PREDICTION PROJECT

OBJECTIVE

- **X Education** gets a lot of leads, its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as '**Hot Leads**.
- Our objective is to help select promising leads using logistic regression.



METHODOLOGY

DataCleaningandmanipulation

ExploratoryDataAnalysis

ModelBuilding

ModelEvaluation

ModelPredictiononTestset

Inferences

Recommendation

DATA CLEANING

Where messes turn into messages

STREAMLINED DATA FOR BETTER INSIGHTS

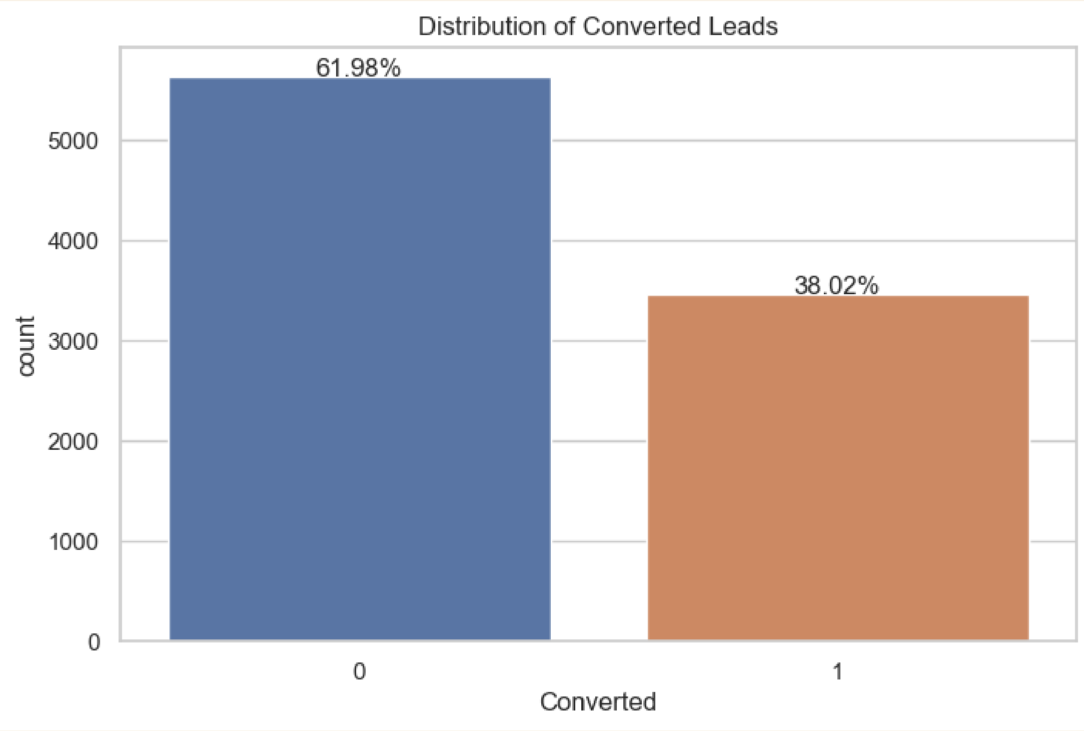
- **Missing Values Addressed:** Columns exceeding 40% missing values were removed to ensure data quality.
- **Unique Identifiers Removed:** Prospect ID and Lead Number served only for record-keeping and were dropped.
- **Reduced Skew:** Skewed data points were eliminated to enhance model predictability.
- **Reshaping the Data:** The data structure might have been reorganized to make it more suitable for analysis
- Started with 9240 rows, 37columns → Ended with 9103 rows, 16 columns

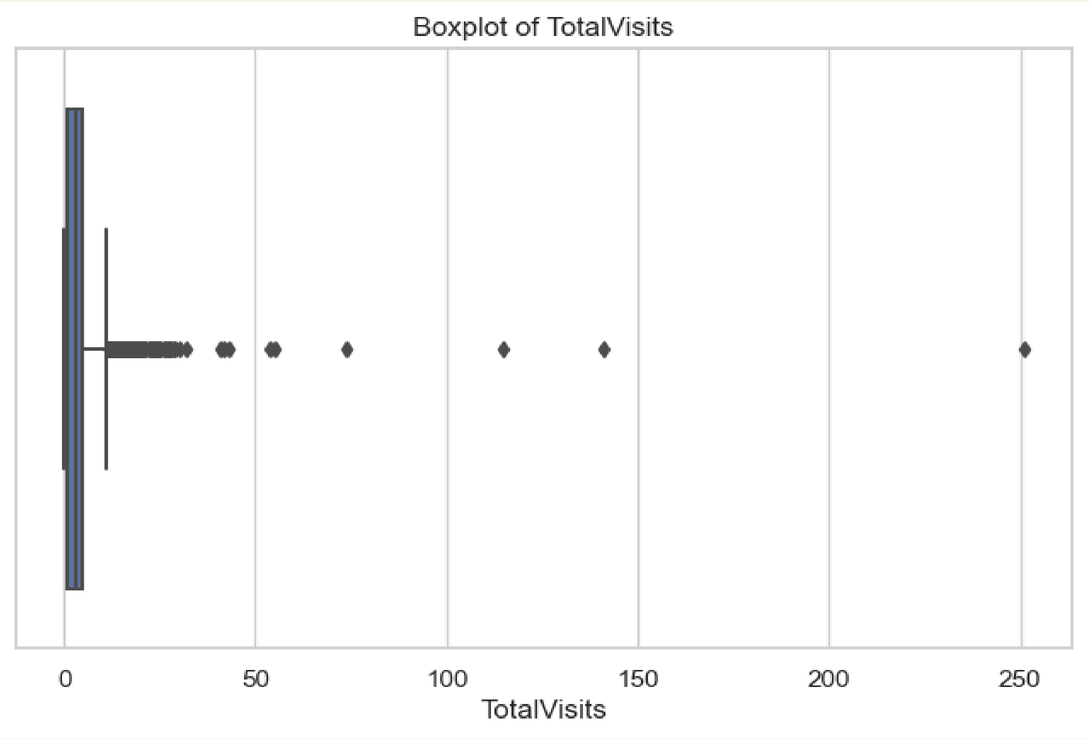
EXPLORATORY DATAANALYSIS

Inspecting, cleaning, transforming, and

understanding data

UNIVARIATE ANALYSIS

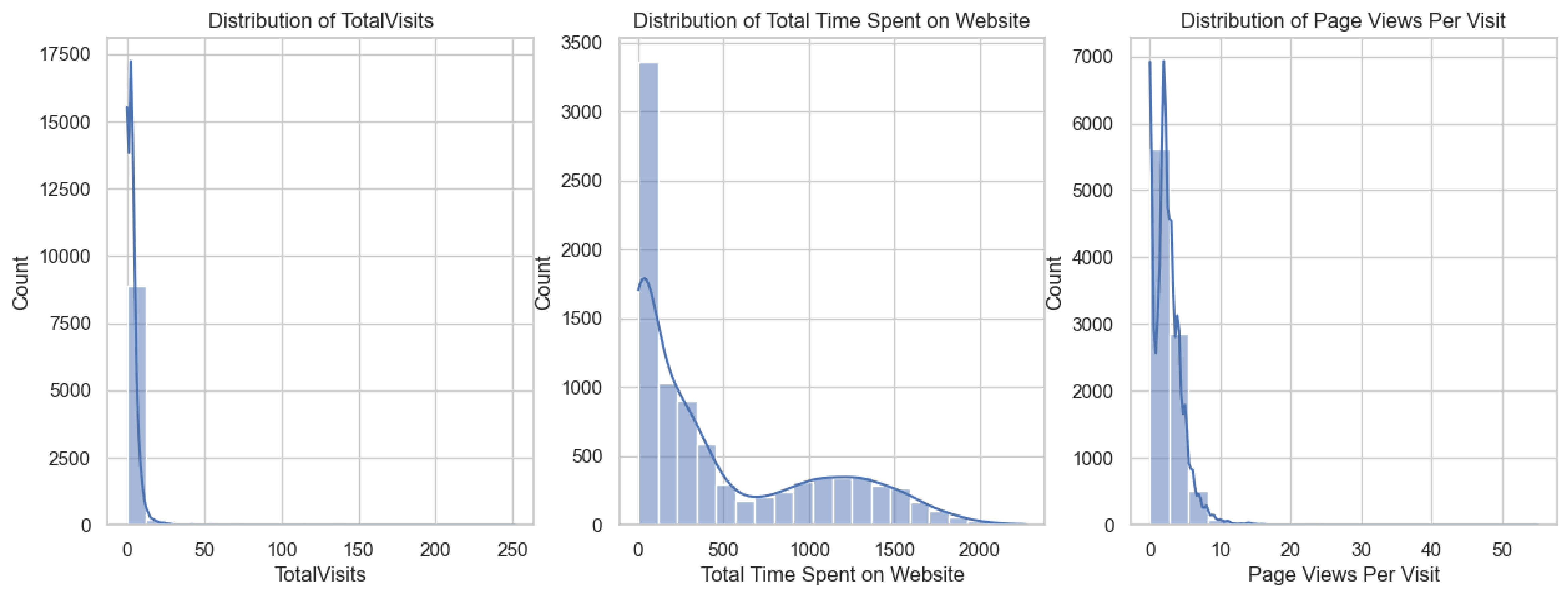




As seen in the graph,most of the leads were not converted

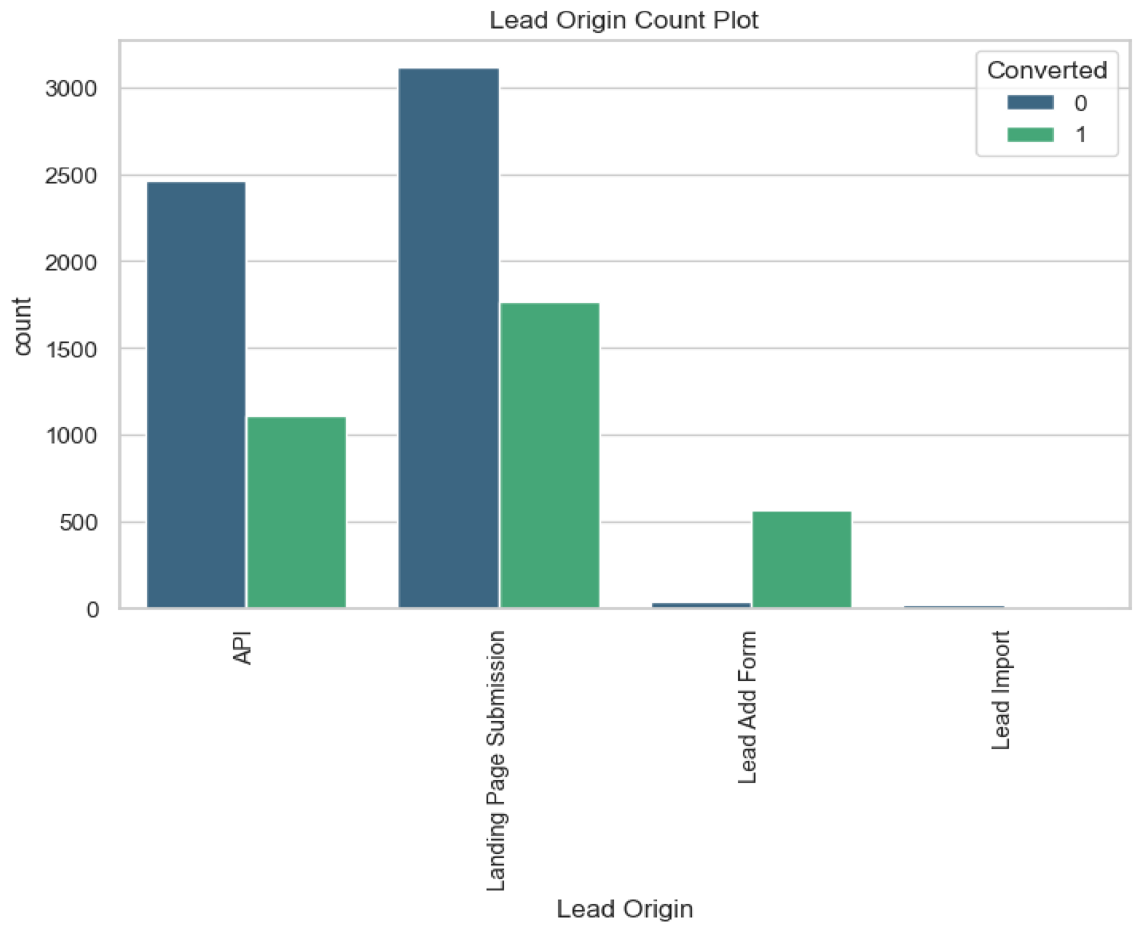
From the boxpot,it is clear that there are outliers that needs to be handled

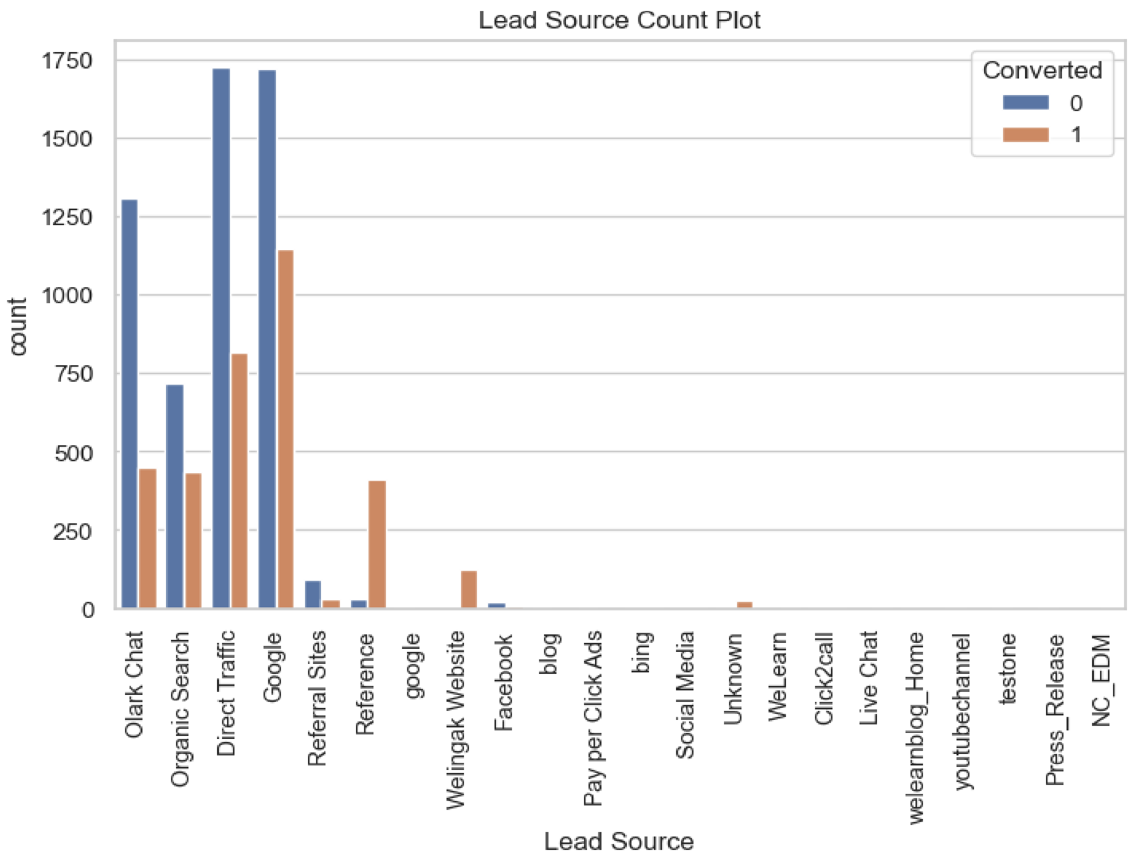
UNIVARIATE ANALYSIS



As seen in the graph the most of the value is 0, in "total visit" and "page view per visit“ ,Most of the prospects had not visited the site,or viewed ,These two column are skewd.

BIVARIATE ANALYSIS

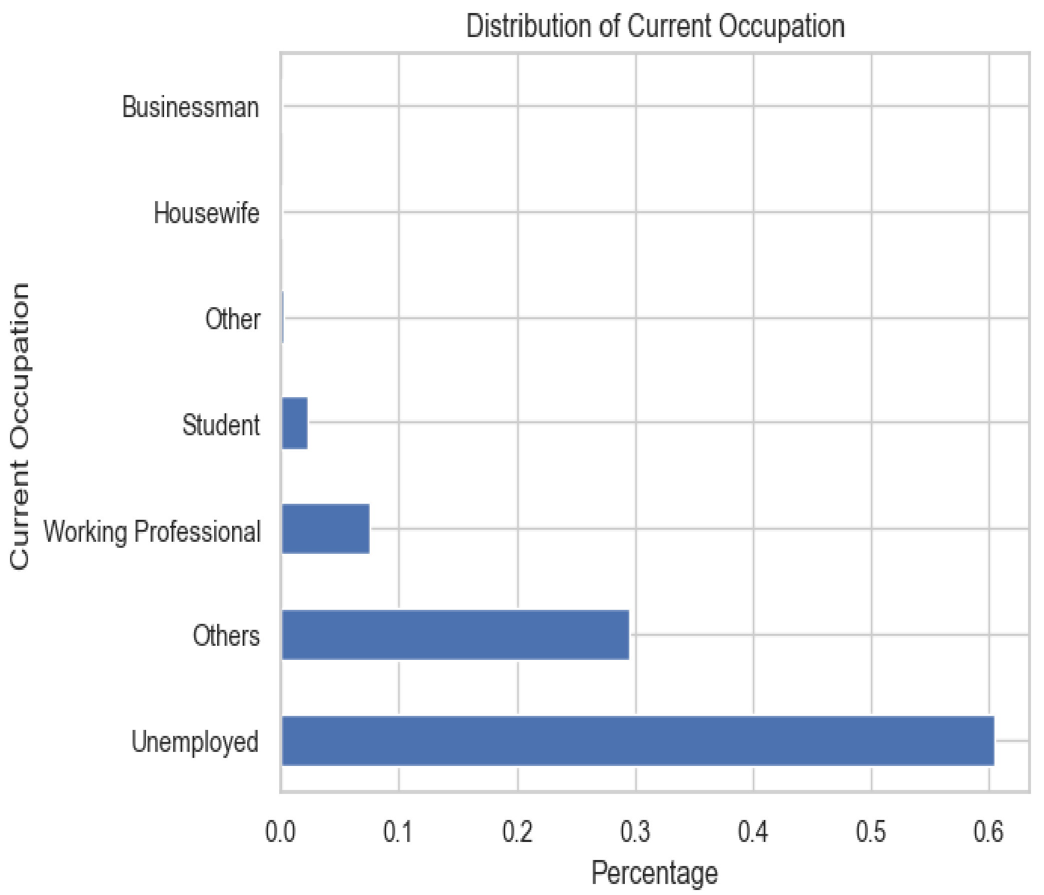


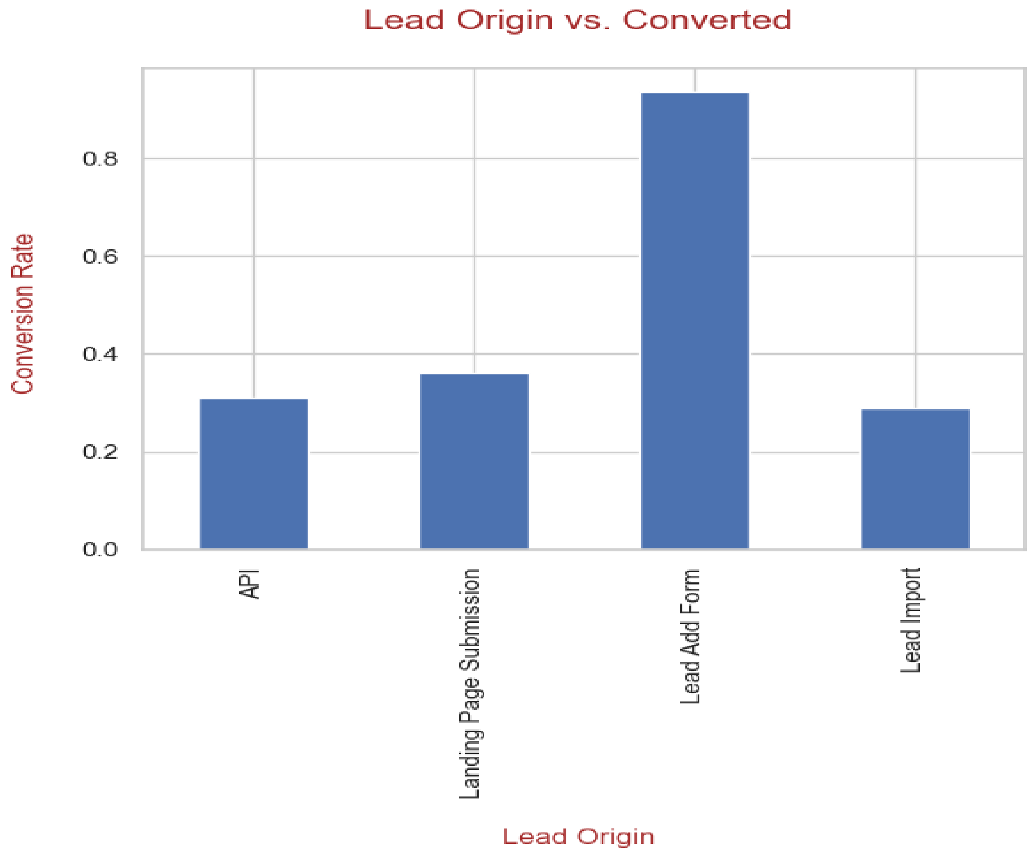


## 'Google' ,Direct Traffic' is the best lead source ,followed by `Olark Chat` and `Organic Search`, And percentage conversion of leads is less in Olark Chat

# Most of the leads originated from `Landing Page Submission` followed by `API`

BIVARIATE ANALYSIS





# Most of the lead were "unemployed".Might be the reson that they want to upskill.

# Most of the leads originated from `Landing Page Submission` followed by `API`

MODEL BUILDING

1. Data Splitting:

•We strategically divided the data into 70% for training and 30% for testing our model. This ensures the model learns from a substantial dataset and generalizes well on unseen data.

2. Feature Scaling:

•To ensure all features contribute equally, we applied Standard scaling to numerical features. This prevents features with larger ranges from dominating the model.

3. Identifying Key Features:

•To focus on the most impactful factors, we employed Recursive Feature Elimination (RFE). This technique iteratively removes the least important feature until we reached the optimal set of 15 features.(Coarse Tuning)

MODEL BUILDING

4. Feature Refinement:

•We took a two-pronged approach for further refinement:

- •**P-value analysis:** Eliminated features with low statistical significance, ensuring the chosen features genuinely influence conversion rates.
- •**Variance Inflation Factor (VIF):** To features with high multicollinearity, which can mislead the model.(Our model had none)

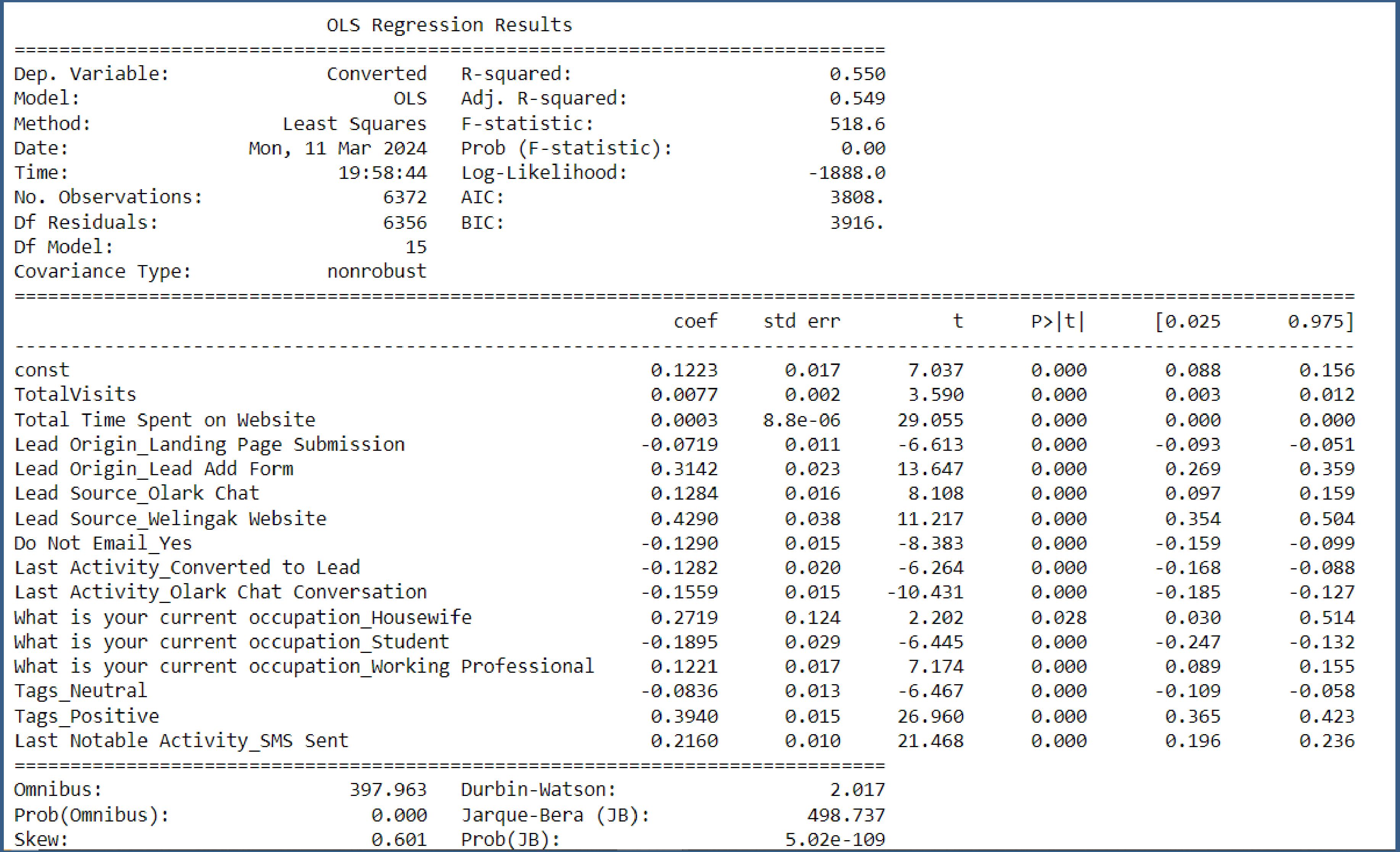
5. Optimized Feature Set:

•Through this rigorous process, we arrived at a concise set of features most relevant to predicting lead conversion.

6. Lead Scoring:

•To simplify lead prioritization, we created a "lead score." This score multiplies the conversion probability by 100, resulting in a value between 0 and 100. Higher scores indicate leads with a greater chance of conversion, allowing you to effectively target your sales efforts.

MODEL EVALUATION



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Converted   R-squared:                       0.550
Model:                            OLS   Adj. R-squared:                  0.549
Method:                 Least Squares   F-statistic:                     518.6
Date:                Mon, 11 Mar 2024   Prob (F-statistic):               0.00
Time:                        19:58:44   Log-Likelihood:                -1888.0
No. Observations:                6372   AIC:                             3808.
Df Residuals:                    6356   BIC:                             3916.
Df Model:                          15
Covariance Type:            nonrobust
========================================================================================================================
                                                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                    0.1223      0.017      7.037      0.000       0.088       0.156
TotalVisits                                              0.0077      0.002      3.590      0.000       0.003       0.012
Total Time Spent on Website                              0.0003    8.8e-06     29.055      0.000       0.000       0.000
Lead Origin_Landing Page Submission                     -0.0719      0.011     -6.613      0.000      -0.093      -0.051
Lead Origin_Lead Add Form                                0.3142      0.023     13.647      0.000       0.269       0.359
Lead Source_Olark Chat                                   0.1284      0.016      8.108      0.000       0.097       0.159
Lead Source_Welingak Website                             0.4290      0.038     11.217      0.000       0.354       0.504
Do Not Email_Yes                                        -0.1290      0.015     -8.383      0.000      -0.159      -0.099
Last Activity_Converted to Lead                         -0.1282      0.020     -6.264      0.000      -0.168      -0.088
Last Activity_Olark Chat Conversation                   -0.1559      0.015    -10.431      0.000      -0.185      -0.127
What is your current occupation_Housewife                0.2719      0.124      2.202      0.028       0.030       0.514
What is your current occupation_Student                 -0.1895      0.029     -6.445      0.000      -0.247      -0.132
What is your current occupation_Working Professional     0.1221      0.017      7.174      0.000       0.089       0.155
Tags_Neutral                                            -0.0836      0.013     -6.467      0.000      -0.109      -0.058
Tags_Positive                                            0.3940      0.015     26.960      0.000       0.365       0.423
Last Notable Activity_SMS Sent                           0.2160      0.010     21.468      0.000       0.196       0.236
==============================================================================
Omnibus:                      397.963   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              498.737
Skew:                           0.601   Prob(JB):                    5.02e-109
```

MODEL EVALUATION

R-squared and Adjusted R-squared: The R-squared and Adjusted R-squared values are 0.550 and 0.549, respectively. These values indicate that the model explains around 55% of the variance in the dependent variable.

F-statistic and Prob (F-statistic): The F-statistic is 518.6, and the corresponding probability (Prob F-statistic) is very close to zero. This suggests that the overall model is statistically significant.

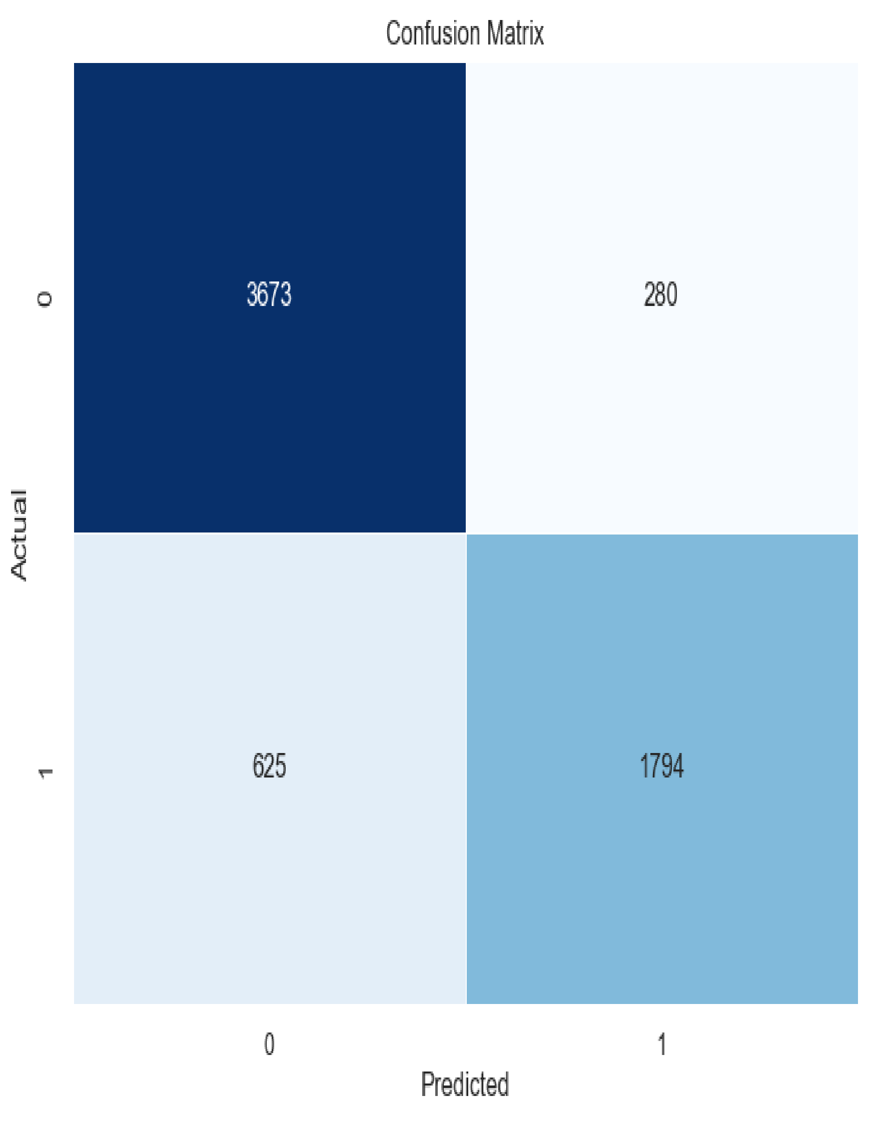
Coefficients:
Intercept (const): The intercept is 0.1223, indicating the estimated conversion rate when all other variables are zero.

TotalVisits: An increase in the total number of visits is associated with a positive change in the probability of conversion.

Total Time Spent on Website: More time spent on the website is associated with a positive change in the probability of conversion.

Lead Origin, Lead Source, Do Not Email, Last Activity, Current Occupation, Tags, and Last Notable Activity: Specific categories within these variables have significant coefficients, affecting the conversion probability.
P-values: The p-values for most of the coefficients are very low (close to zero), suggesting that these features are statistically significant in predicting the conversion outcome

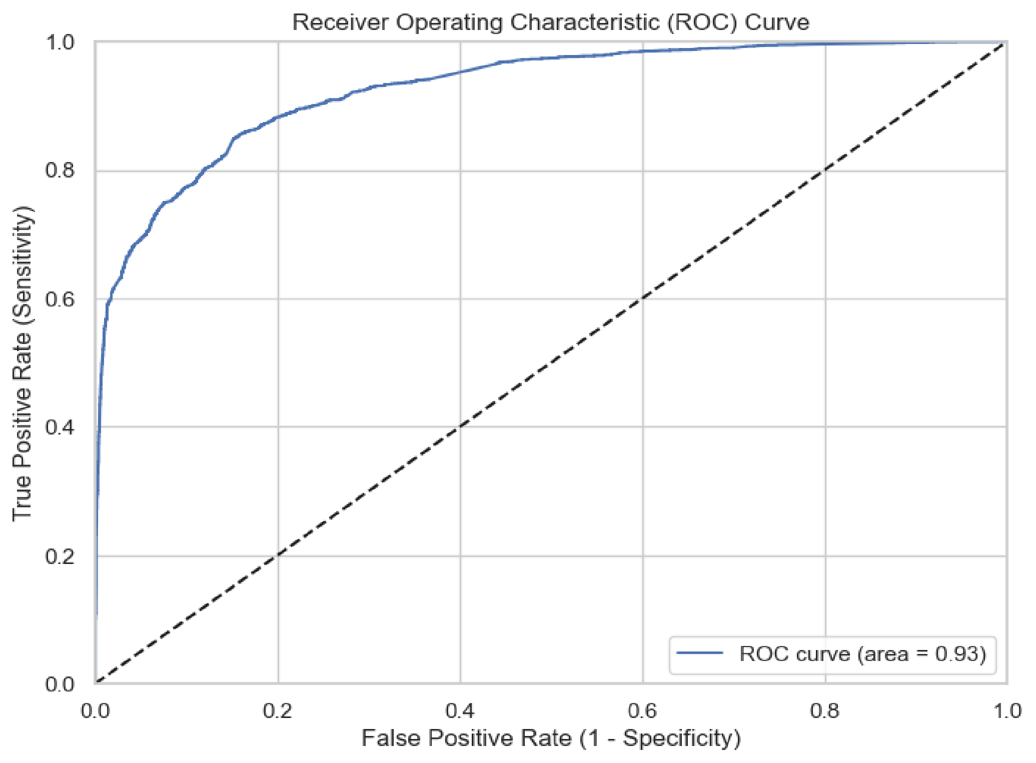
MODEL EVALUATION- TRAIN

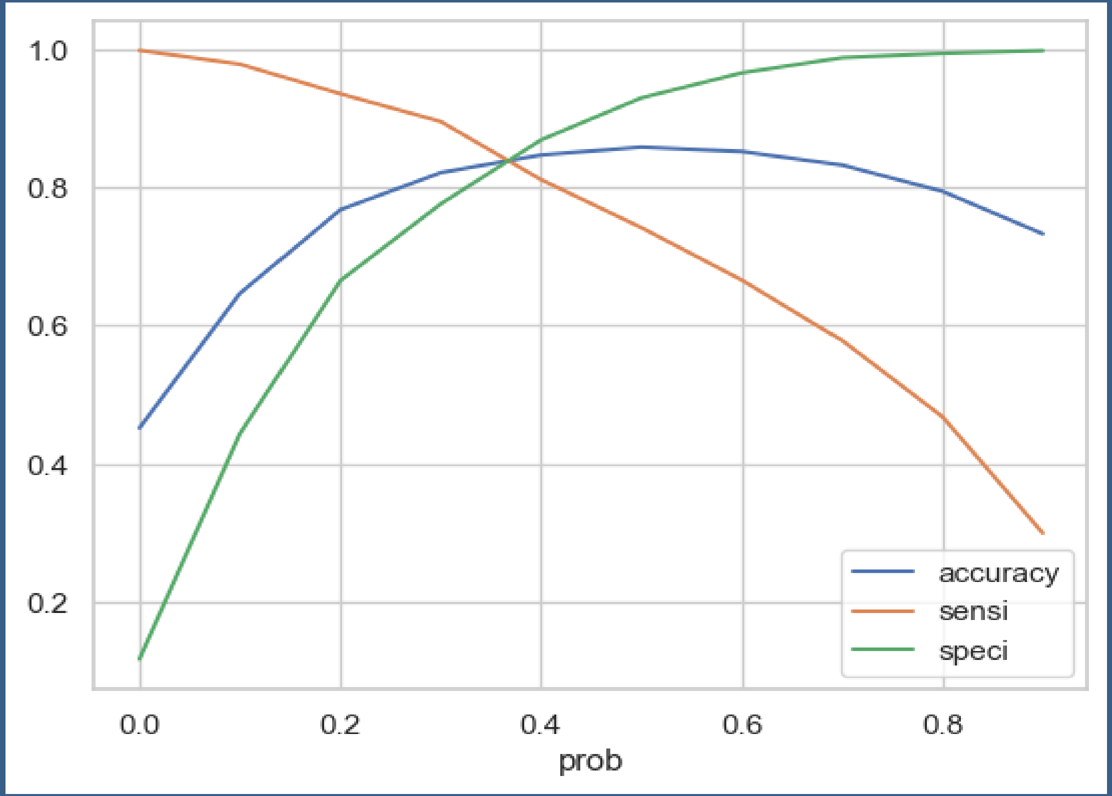


The performance metrics for train model

- Accuracy: 85.80%
- Sensitivity/Recall/True Positive Rate: 74.16%
- Specificity/True Negative Rate: 92.92%
- Precision/Positive Predictive Value: 86.50%
- Negative Predictive Value: 85.46%
- False Positive Rate: 7.08%
- False Negative Rate: 25.84%
- F1 Score: 79.86%

MODEL EVALUATION - ROC/CUTOFF





- A ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.93 is indicative of a good model performance.
- THe ROC AUC score measures the ability of the model to distinguish between positive and negative instances.

As you can see that around 0.38, you get the optimal values of the three metrics. So let's choose 0.38 as ou cutoff now.

MODEL EVALUATION- TEST

The performance metrics for TEST model

•**Accuracy:** The proportion of correctly classified instances is 85.02%, indicating the overall correctness of the model.

•**Sensitivity/Recall/True Positive Rate:** The model can identify approximately 86.47% of the actual positive instances, suggesting good performance in capturing leads that actually converted.

•**Specificity/True Negative Rate:** The model has a specificity of around 84.13%, implying it is effective in correctly identifying non-converted leads.

•**Precision/Positive Predictive Value:** Out of the instances predicted as positive, around 77.07% are actually positive. This measures the accuracy of the positive predictions.

•**Negative Predictive Value:** Among instances predicted as negative, approximately 90.97% are genuinely negative. This reflects the accuracy of negative predictions.

THANK YOU

DIGANTH RAI

UPGRAD-BATCH ID **5831**

dIganthrai048@gmail.com