Lead Scoring Case Study

Project By-
Sonu Verma
Siddharth Rahate
Shruti Sinha

INTRODUCTION

X EDUCATION GETS A LOT OF LEADS, ITS LEAD CONVERSION RATE IS VERY POOR APPROXIMATELY 30%. TO MAKE THIS PROCESS MORE EFFICIENT, THE COMPANY WISHES TO IDENTIFY THE MOST POTENTIAL LEADS, ALSO KNOWN AS ‘HOT LEADS’. THIS IS THE MAJOR AIM OF PROJECT TO INCREASE SALES CONVERSION RATHER THEN WASTING TIME CALLING EVERYONE

Business Objective

Preparing a model to identify ‘Hot leads’ for further use of the company which will increase their sales conversion

Methodology

- Import data
- Data cleaning and preparation for further analysis
- Exploratory data analysis
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each lead
- Test the model
- Model Evaluation
- Measure the accuracy of the model

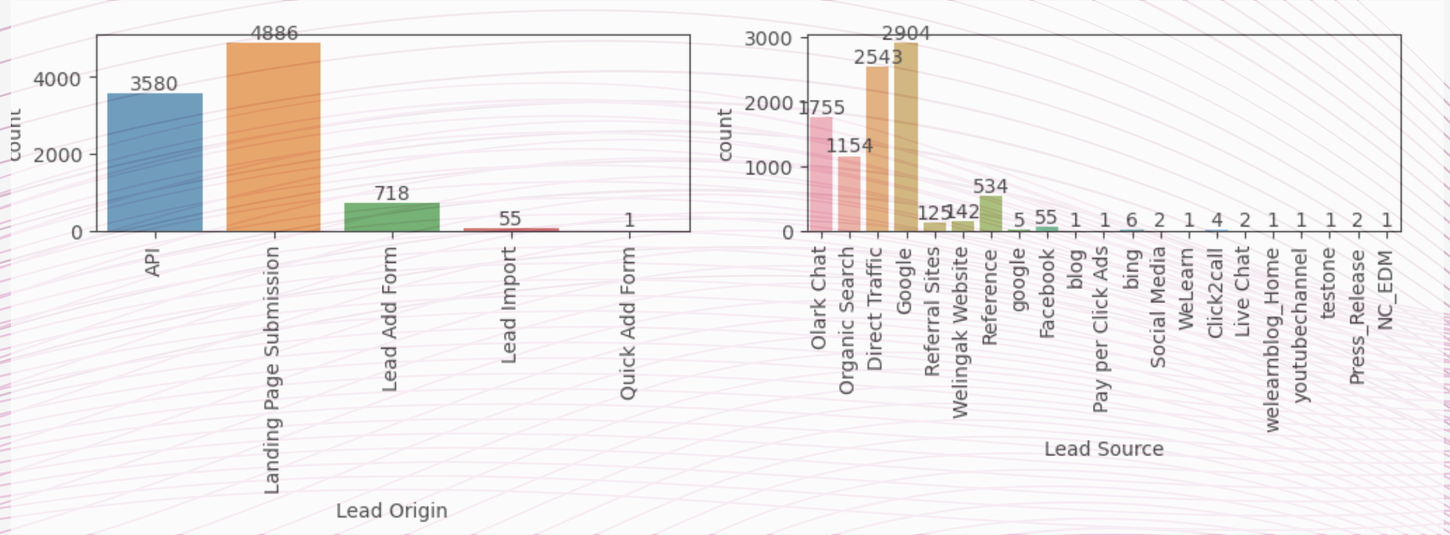
DATA INSIGHTS

- City: The City column exhibits 39.71% missing values.Hence, it's advisable to drop the City column.
- Specialization: With 36.58% missing values, the Specialization column demonstrates an even distribution of selections. In this scenario, creating an additional category labeled 'Others' is more appropriate than imputation or dropping
- Tags: Tags indicate the current status of leads and contain 36.29% missing values, it's recommended to drop the Tags column.
- What matters most to you in choosing a course: This variable, with 29.32% missing values, sees 'better career prospects' selected by 99.95% of customers, indicating significant skewness. Thus, it's unlikely to provide meaningful insights.
- What is your current occupation: Imputing missing values with 'Unemployed', the most prevalent occupation, seems appropriate given X Education's context of selling online courses.
- Country: Around 96% of customers are from India, making it impractical to impute missing values with 'India'. Hence, dropping the Country column is recommended.
- Last Activity: "Email Opened" is the most frequent activity,and only 1.11% of values are missing. Hence, imputing missing values with 'Email Opened' is a reasonable strategy.
- Lead Source: "Google" is the most common source

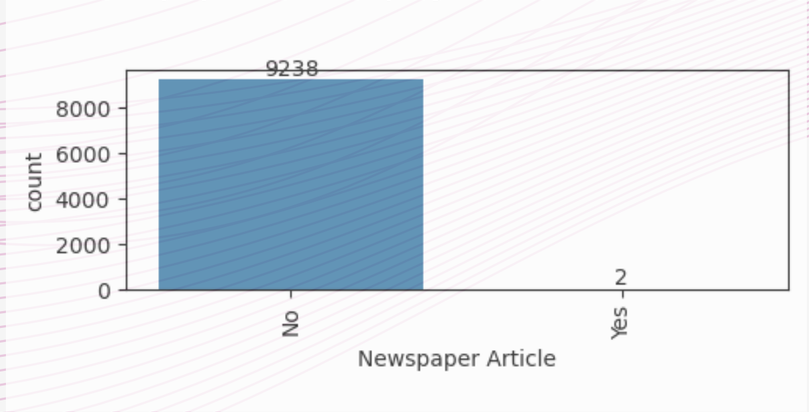
EDA

Univariate

- Lead origin is majorly from landing page submission and source is google

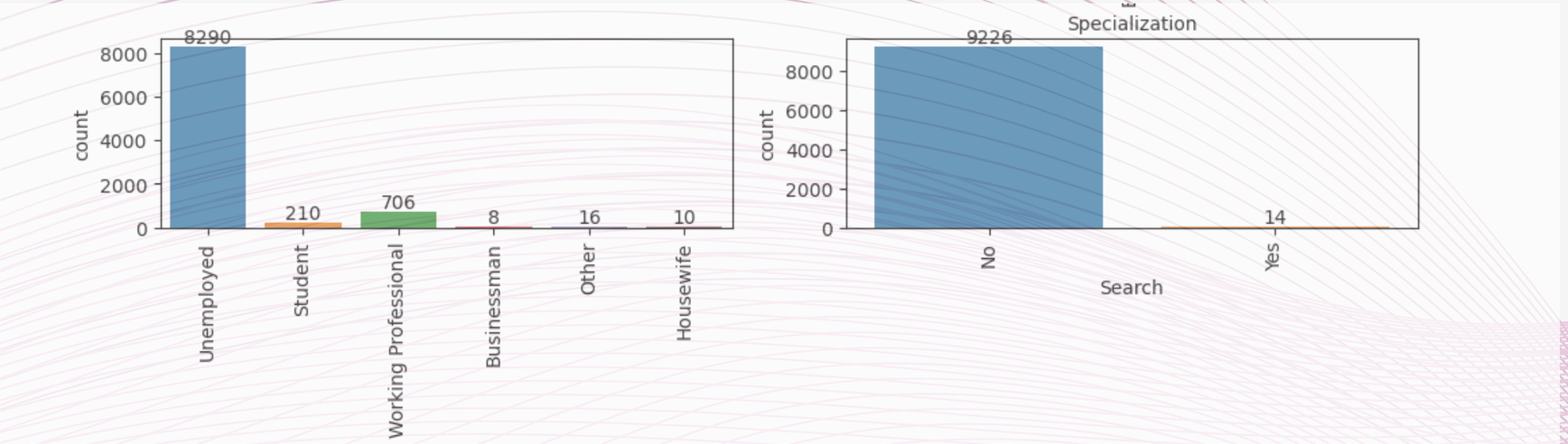


- News paper articles plays a vital role for leads

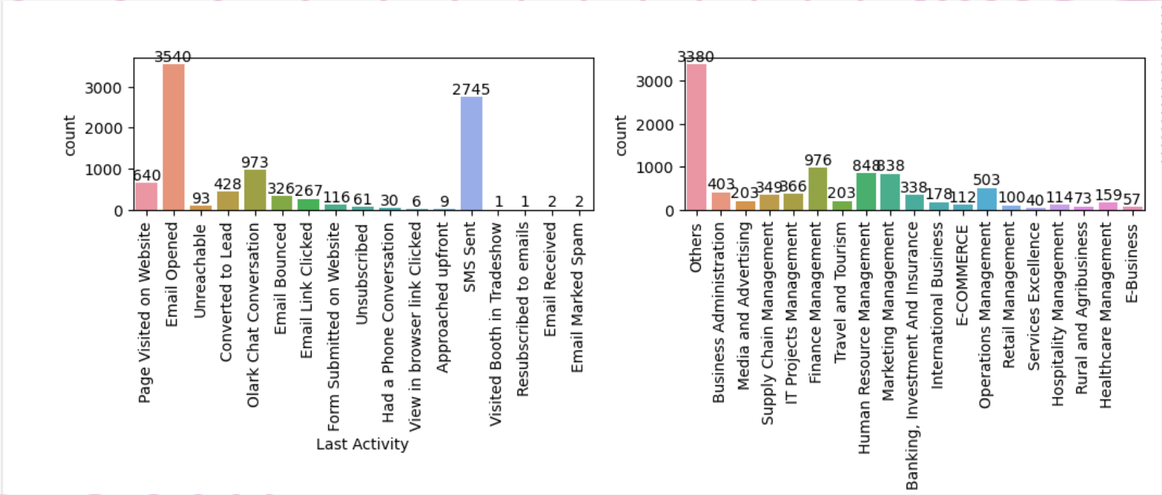


EDA

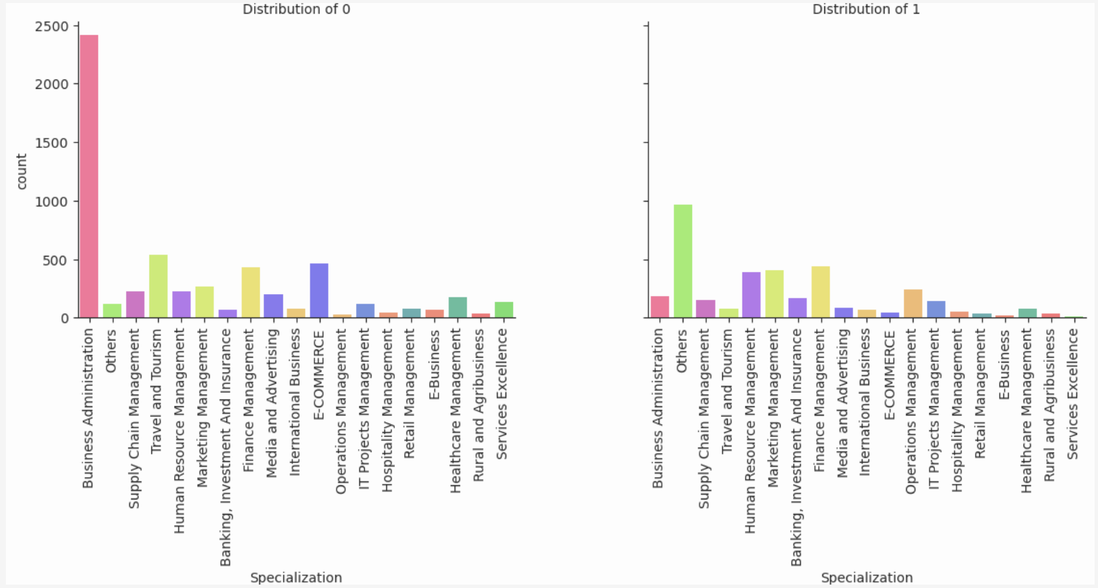
- Unemployed people are most visitors to site

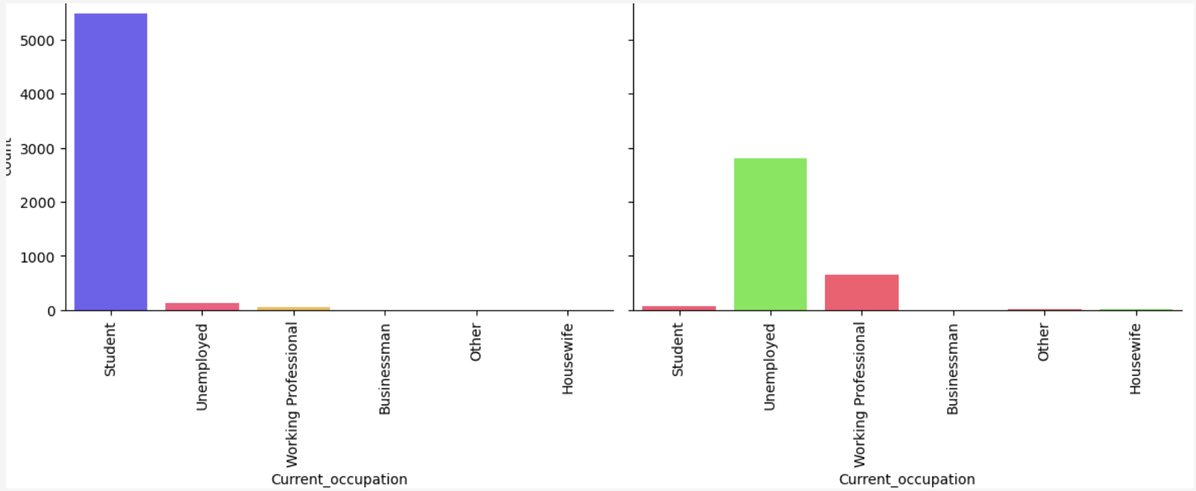


- Last activity of mail opened are majorly in leads



Bivariate Analysis





Bivariate Analysis

- LEAD ORIGIN: APPROXIMATELY 52% OF ALL LEADS ORIGINATED FROM "LANDING PAGE SUBMISSION" WITH A LEAD CONVERSION RATE (LCR) OF 36%. THE "API" IDENTIFIED APPROXIMATELY 39% OF CUSTOMERS WITH A LEAD CONVERSION RATE (LCR) OF 31%.

- CURRENT_OCCUPATION: ABOUT 90% OF THE CUSTOMERS ARE CATEGORIZED AS UNEMPLOYED, WITH A LEAD CONVERSION RATE (LCR) OF 34%. CONVERSELY, WORKING PROFESSIONALS CONSTITUTE ONLY 7.6% OF TOTAL CUSTOMERS, WITH AN IMPRESSIVE LEAD CONVERSION RATE (LCR) OF ALMOST 92%.

- DO NOT EMAIL: 92% OF THE PEOPLE HAVE OPTED NOT TO BE EMAILED ABOUT THE COURSE.

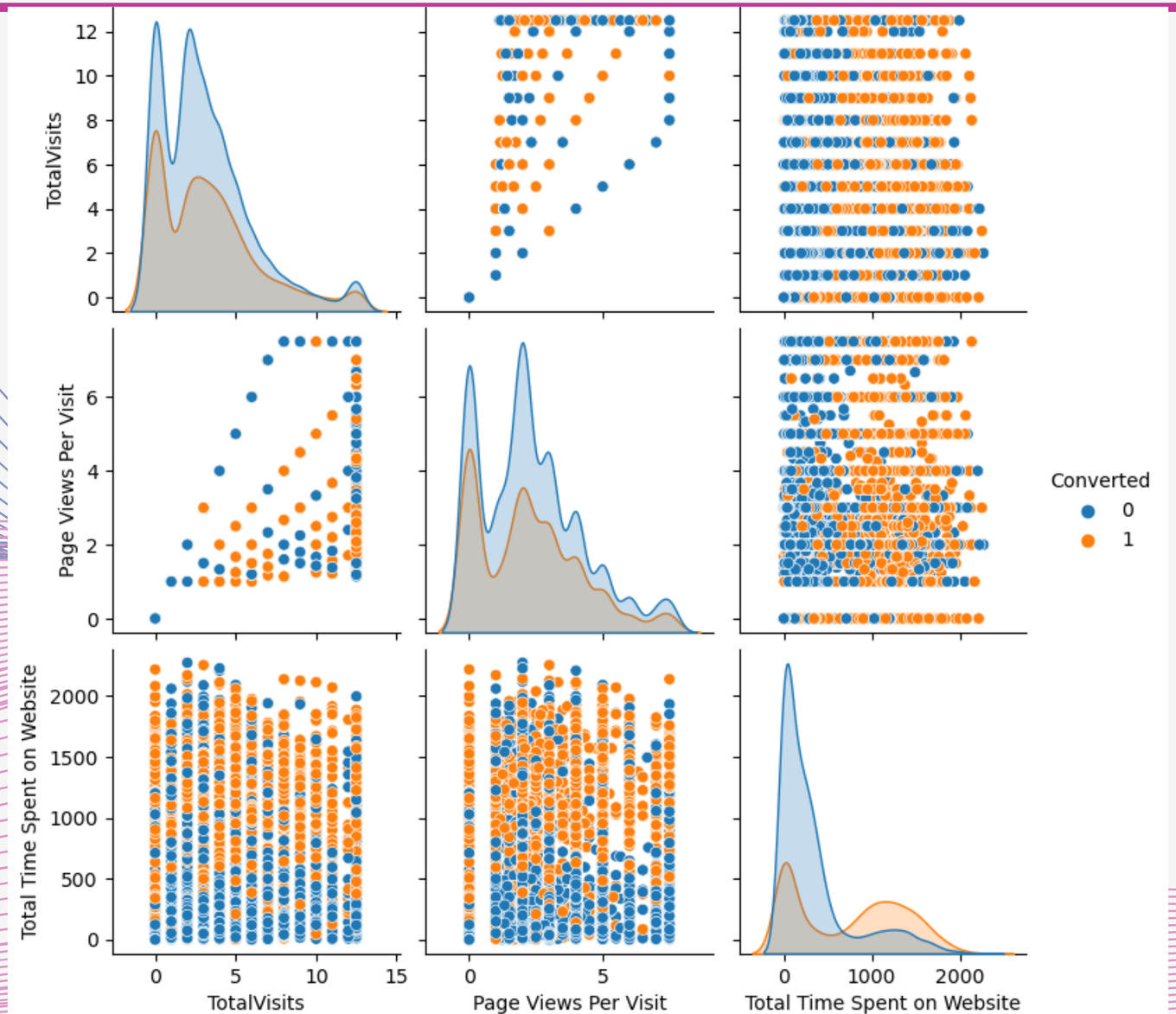
NOTE: WE HAVE ASSUMED LCR AS LEAD CONVERSION RATE.

- LEAD SOURCE: GOOGLE HAS AN LCR OF 40% OUT OF 31% OF CUSTOMERS, WHILE DIRECT TRAFFIC CONTRIBUTES A 32% LCR WITH 27% OF CUSTOMERS, WHICH IS LOWER THAN GOOGLE. ORGANIC SEARCH ALSO YIELDS A 37.8% LCR, BUT ONLY 12.5% OF CUSTOMERS COME THROUGH THIS LEAD SOURCE. REFERENCE HAS AN LCR OF 91%, BUT THERE ARE ONLY AROUND 6% OF CUSTOMERS THROUGH THIS LEAD SOURCE.

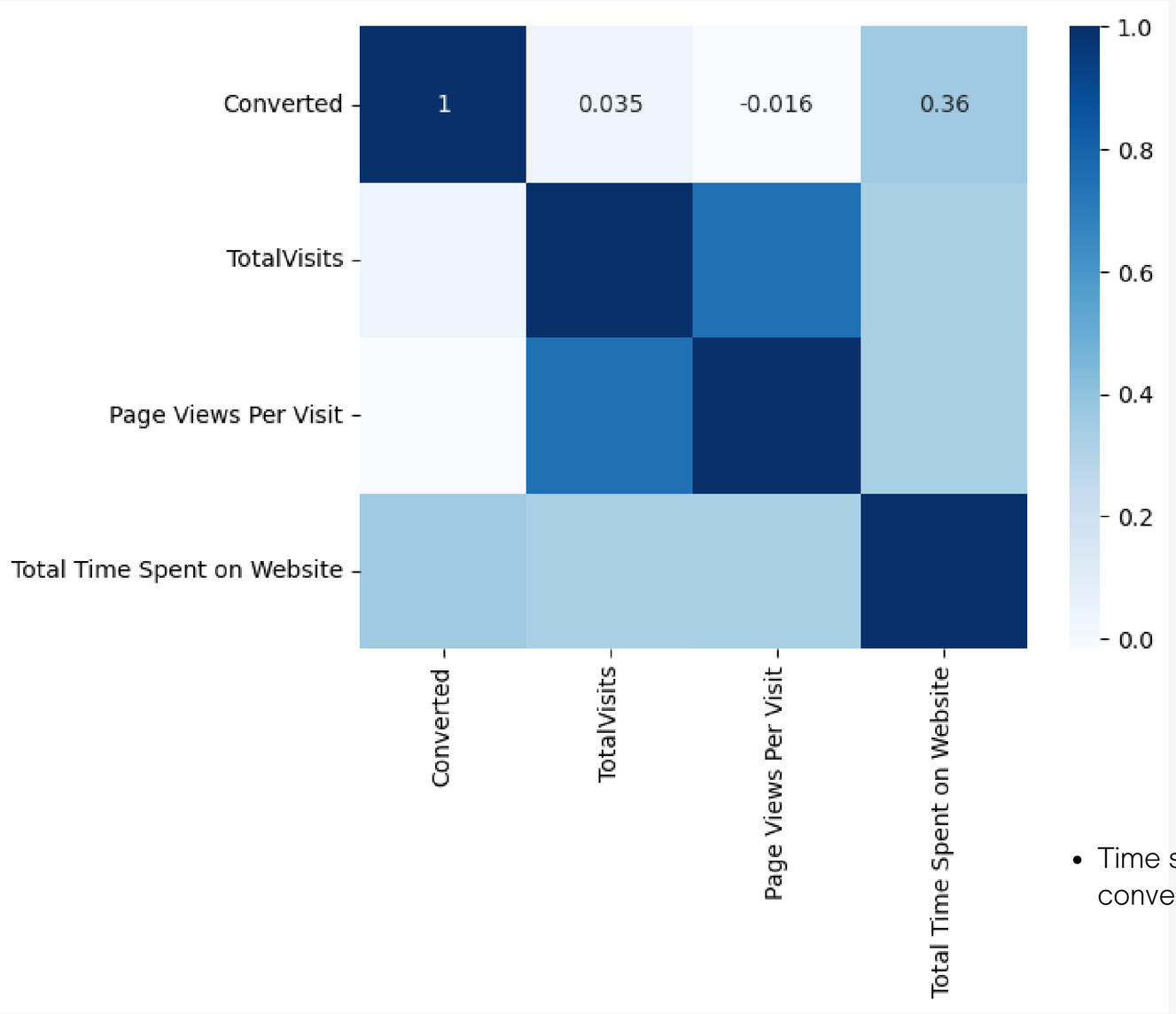
- LAST ACTIVITY: "SMS SENT" BOASTS A HIGH LEAD CONVERSION RATE OF 63%, WITH A 30% CONTRIBUTION FROM THE LAST ACTIVITIES. "EMAIL OPENED" ACTIVITY CONTRIBUTED 38% OF THE LAST ACTIVITIES PERFORMED BY THE CUSTOMERS, WITH A 37% LEAD CONVERSION RATE.

- SPECIALIZATION: MARKETING MANAGEMENT, HR MANAGEMENT, AND FINANCE MANAGEMENT SHOW SIGNIFICANT CONTRIBUTIONS.

BIVARIATE ANALYSIS NUMERICAL VARIABLES



CORRELATION



- Time spent on website have higher chances of conversion

TEST-TRAIN SPLIT

We'll assign predictor variables to X and target variables to y. Then, we'll split the data into training and testing sets. in 70:30.

Feature Scaling

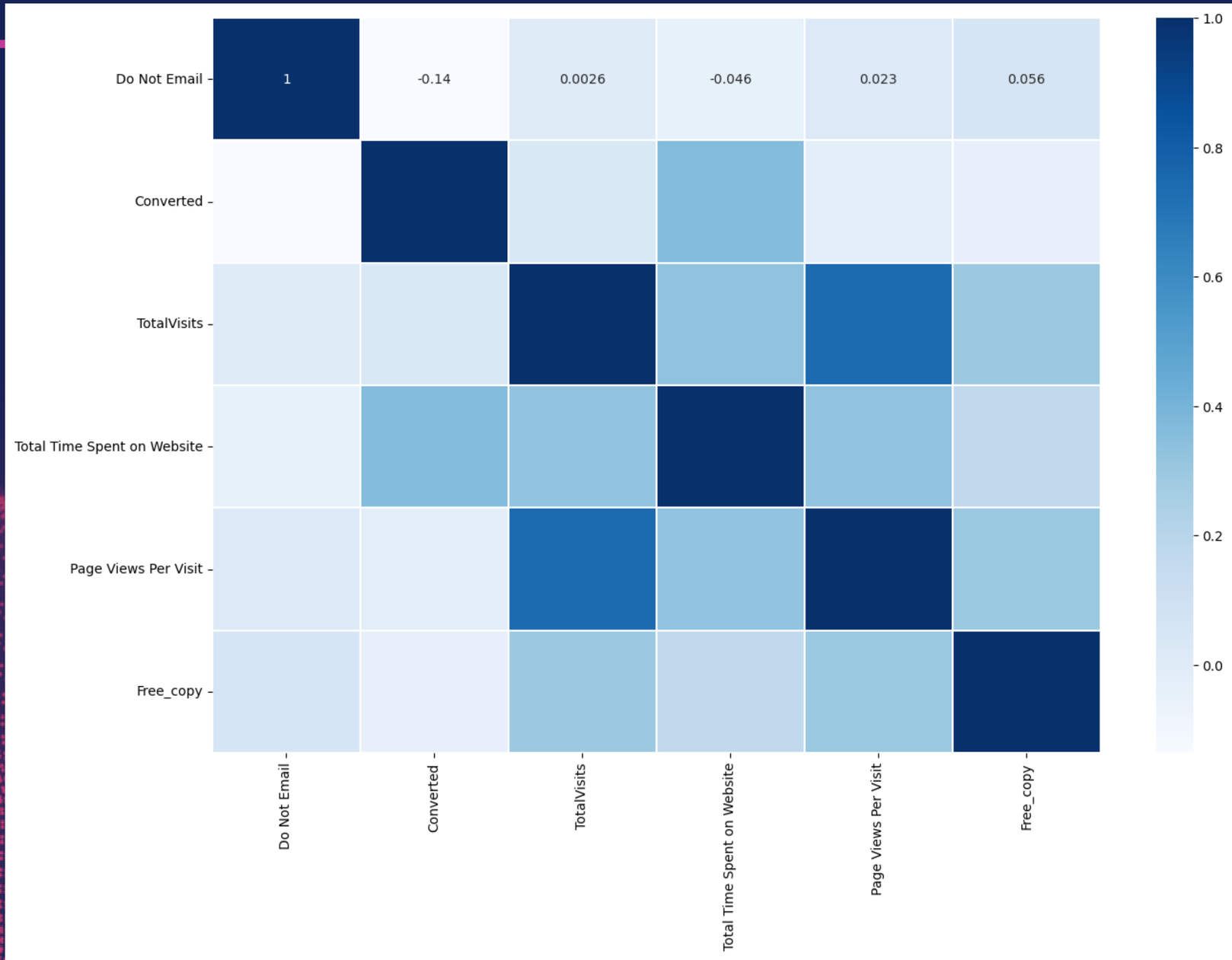
We'll use the StandardScaler for scaling the features. First, we'll fetch the columns with int64 and float64 data types from the data frame for scaling. Then, we'll display the X-train data frame after standard scaling.

We'll check the Lead Conversion Rate (LCR) where "Converted" is our target variable. We'll denote Lead Conversion Rate with 'LCR' as its short form.
We are having 38.53% conversion for that

FEATURECORRELATION

Correlation matrix to understand the relationships between different features in the dataset.

Analyzing variables that might be highly correlated with each other from the same class, we can focus on specific sections of the correlation matrix. We'll identify clusters of variables that have high correlation coefficients and might indicate multicollinearity. Then, we can further investigate these variables to determine if they should be addressed.



MODEL BUILDING

We will be using Recursive Feature Elimination (RFE) for feature selection. RFE is a method that recursively removes features from the model and selects the optimal subset of features based on their importance in predicting the target variable.

Model 1

TotalVisits column will be removed from model due to high p-value of 0.025, which is above the accepted threshold of 0.05 for statistical significance.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6462
Model Family:	Binomial	Df Model:	5
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3710.6
Date:	Fri, 12 Apr 2024	Deviance:	7421.2
Time:	19:07:23	Pearson chi2:	6.51e+03
No. Iterations:	5	Pseudo R-squ. (CS):	0.1664
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5538	0.029	-19.325	0.000	-0.610	-0.498
Do Not Email	-0.3448	0.036	-9.495	0.000	-0.416	-0.274
TotalVisits	0.0953	0.042	2.245	0.025	0.012	0.179
Total Time Spent on Website	0.9363	0.033	28.802	0.000	0.873	1.000
Page Views Per Visit	-0.4153	0.045	-9.287	0.000	-0.503	-0.328
Free_copy	-0.1773	0.031	-5.789	0.000	-0.237	-0.117

MODEL 2

No Need To remove any columns.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6463
Model Family:	Binomial	Df Model:	4
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3713.1
Date:	Fri, 12 Apr 2024	Deviance:	7426.2
Time:	22:09:50	Pearson chi2:	6.51e+03
No. Iterations:	5	Pseudo R-squ. (CS):	0.1657
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5534	0.029	-19.319	0.000	-0.610	-0.497
Do Not Email	-0.3458	0.036	-9.519	0.000	-0.417	-0.275
Total Time Spent on Website	0.9429	0.032	29.095	0.000	0.879	1.006
Page Views Per Visit	-0.3488	0.033	-10.486	0.000	-0.414	-0.284
Free_copy	-0.1700	0.030	-5.585	0.000	-0.230	-0.110

Model 3

No Need To remove any columns.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6463
Model Family:	Binomial	Df Model:	4
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3713.1
Date:	Fri, 12 Apr 2024	Deviance:	7426.2
Time:	22:10:33	Pearson chi2:	6.51e+03
No. Iterations:	5	Pseudo R-squ. (CS):	0.1657
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5534	0.029	-19.319	0.000	-0.610	-0.497
Do Not Email	-0.3458	0.036	-9.519	0.000	-0.417	-0.275
Total Time Spent on Website	0.9429	0.032	29.095	0.000	0.879	1.006
Page Views Per Visit	-0.3488	0.033	-10.486	0.000	-0.414	-0.284
Free_copy	-0.1700	0.030	-5.585	0.000	-0.230	-0.110

MODEL 4

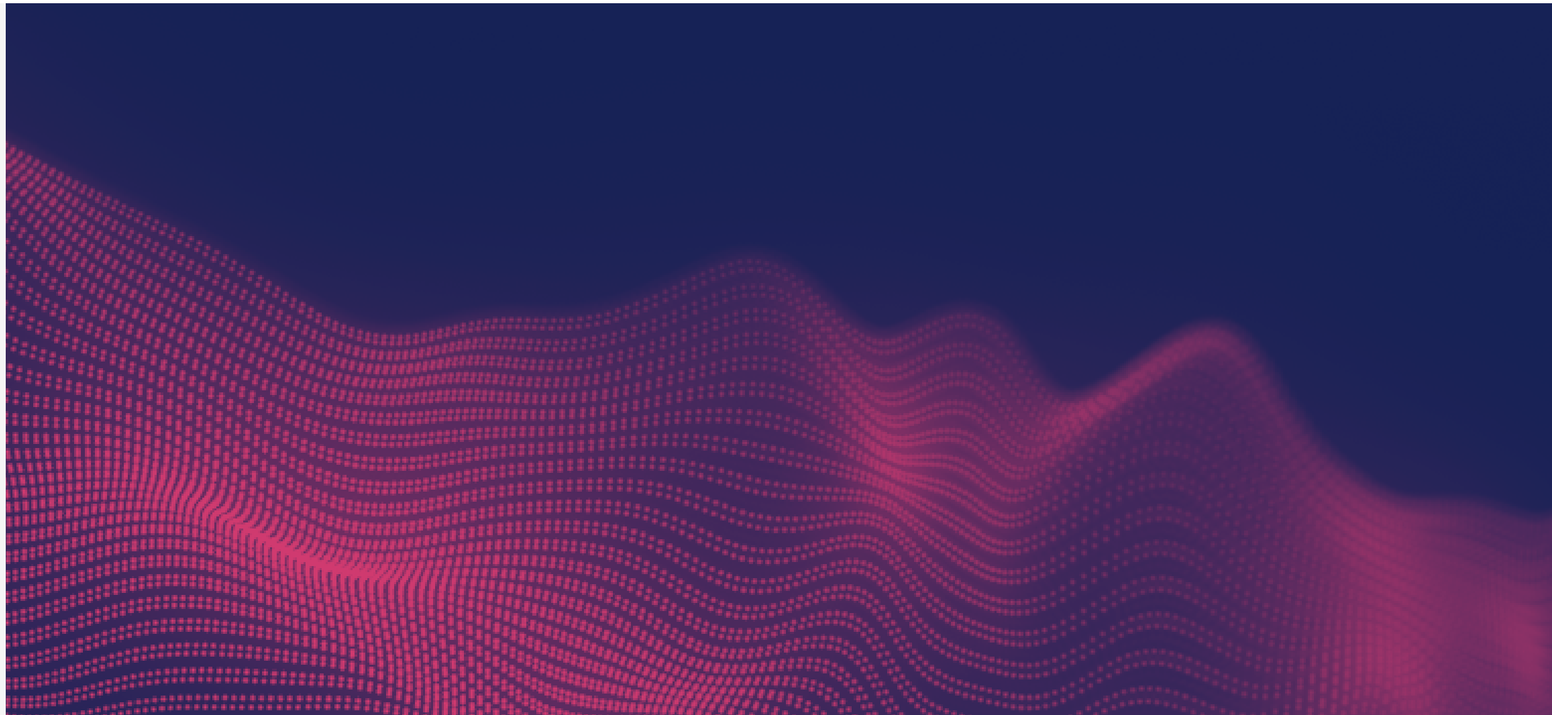
No Need To remove any columns.

We thoroughly evaluated the model and found it to be acceptable based on the criteria you've mentioned:
VIF Values: All variables have VIF values less than 5, indicating that multicollinearity is not a significant issue. P-values: The p-values for all variables are less than 0.05, suggesting that they are statistically significant in predicting the target variable. Overall Assessment: Based on the VIF values and p-values, the model appears to be well-constructed and suitable for further evaluation. Considering these factors, you've decided to finalize "Model 4" for model evaluation.
It sounds like you've done a comprehensive job evaluating your model and ensuring its suitability for further analysis. This decision sets a solid foundation for the next steps in your modeling process.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6463
Model Family:	Binomial	Df Model:	4
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3713.1
Date:	Fri, 12 Apr 2024	Deviance:	7426.2
Time:	22:11:37	Pearson chi2:	6.51e+03
No. Iterations:	5	Pseudo R-squ. (CS):	0.1657
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5534	0.029	-19.319	0.000	-0.610	-0.497
Do Not Email	-0.3458	0.036	-9.519	0.000	-0.417	-0.275
Total Time Spent on Website	0.9429	0.032	29.095	0.000	0.879	1.006
Page Views Per Visit	-0.3488	0.033	-10.486	0.000	-0.414	-0.284
Free_copy	-0.1700	0.030	-5.585	0.000	-0.230	-0.110



Model Evaluation

Breakdown of the model evaluation metrics you can consider:

1. Confusion Matrix: A table that describes the performance of a classification model, showing the counts of true positives, true negatives, false positives, and false negatives.
2. Accuracy: The proportion of correctly classified instances out of the total instances. It is calculated as (TP + TN) / (TP + TN + FP + FN).
3. Sensitivity and Specificity:
   - Sensitivity (True Positive Rate or Recall): The proportion of actual positive cases that were correctly identified by the model. It is calculated as TP / (TP + FN).
   - Specificity (True Negative Rate): The proportion of actual negative cases that were correctly identified by the model. It is calculated as TN / (TN + FP).
4. Threshold Determination using ROC & Finding Optimal Cutoff Point: Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. You can use the ROC curve to find the optimal cutoff point that balances sensitivity and specificity, usually by maximizing the area under the curve (AUC).
5. Precision and Recall:
   - Precision (Positive Predictive Value): The proportion of predicted positive instances that were correctly classified. It is calculated as TP / (TP + FP).
   - Recall (Sensitivity): The proportion of actual positive instances that were correctly classified. It is calculated as TP / (TP + FN).

Model Performance

Metrics beyond accuracy, such as sensitivity and specificity, provide a more comprehensive evaluation of a classification model's performance.

Sensitivity (True Positive Rate):

- Sensitivity measures the proportion of actual positives that are correctly identified by the model.
- It helps assess the model's ability to detect positive instances correctly.
- Mathematically, sensitivity is calculated as the number of true positives (TP) divided by the sum of true positives and false negatives (TP + FN).
- High sensitivity indicates that the model is effective at identifying positive cases.

Specificity (True Negative Rate):

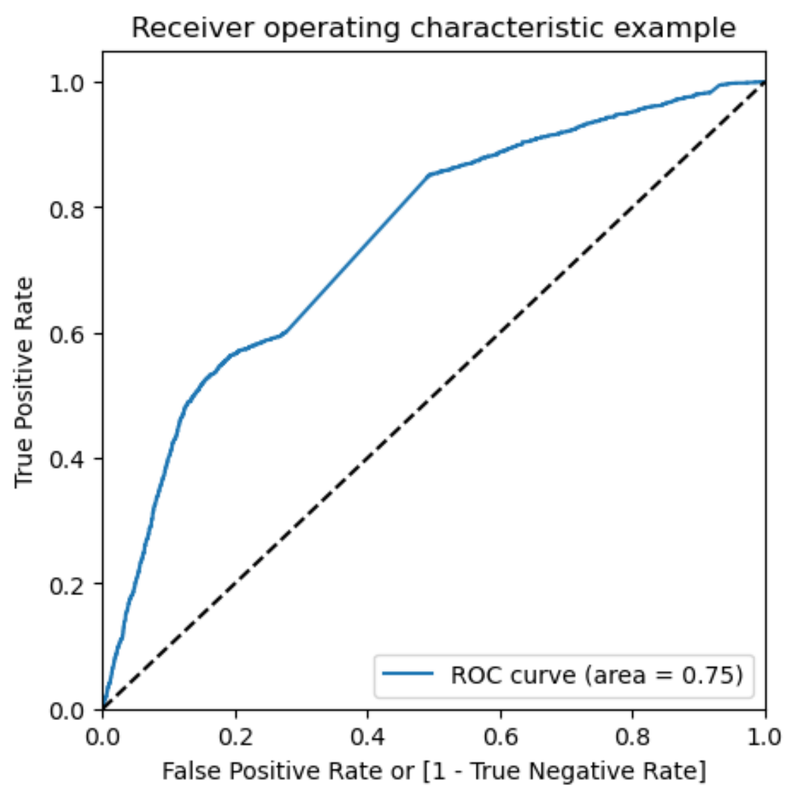
- Specificity measures the proportion of actual negatives that are correctly identified by the model.
- It helps assess the model's ability to avoid false alarms by correctly identifying negative instances.
- Mathematically, specificity is calculated as the number of true negatives (TN) divided by the sum of true negatives and false positives (TN + FP).
- High specificity indicates that the model is effective at avoiding false positive predictions.

Predictions at Threshold 0.5 Probability:

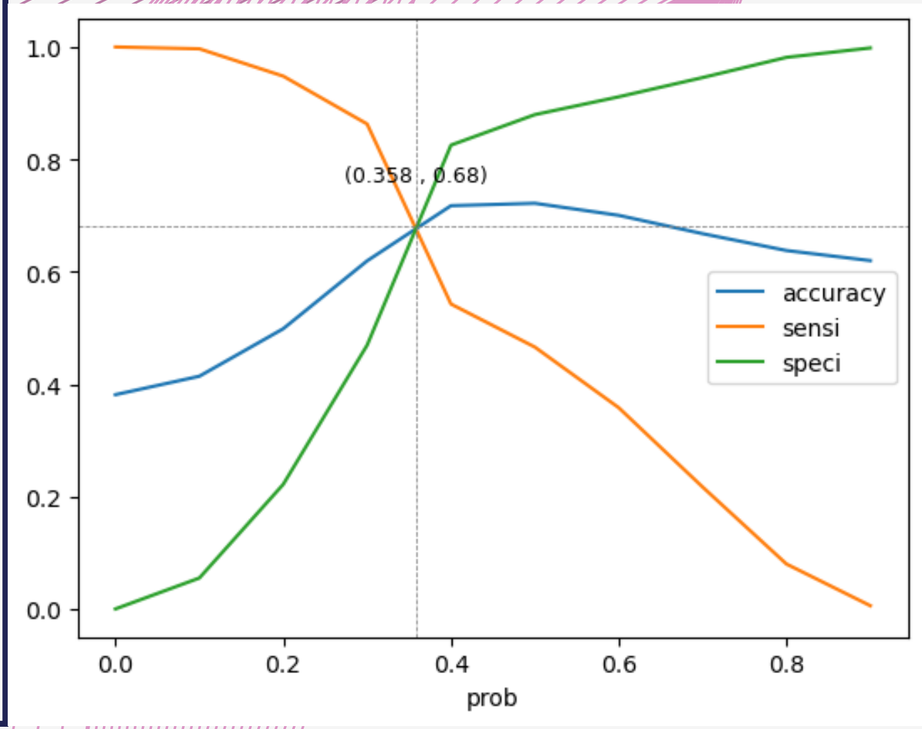
- When we use a threshold of 0.5 probability, it means that the model classifies instances as positive if their predicted probability of belonging to the positive class is 0.5 or higher.
- Instances with predicted probabilities below 0.5 are classified as negative.
- This threshold is commonly used as a default, but it may not always be the optimal threshold for a specific problem.
- Analyzing sensitivity, specificity, and other metrics across different threshold values can help identify the optimal threshold for a given problem.

ROC Model Evaluation

ROC curve is 0.75 out of 1 which indicates a good predictive model



PLOT ACCURACY SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITIES.



Test Model Evaluation

```
*******************************************************

Confusion Matrix
[[1272  405]
 [ 461  634]]

*******************************************************

True Negative                   :  1272
True Positive                   :  634
False Negative                  :  461
False Positve                   :  405
Model Accuracy                  :  0.6876
Model Sensitivity               :  0.579
Model Specificity               :  0.7585
Model Precision                 :  0.6102
Model Recall                    :  0.579
Model True Positive Rate (TPR)  :  0.579
Model False Positive Rate (FPR) :  0.2415


*******************************************************
```

THE EVALUATION METRICS FOR THE TEST SET ARE CONSISTENT WITH THOSE OF THE TRAINING SET, INDICATING THAT OUR FINAL MODEL, LOGM4, PERFORMS CONSISTENTLY ACROSS BOTH DATASETS:

ACCURACY: APPROXIMATELY 68.76%
SENSITIVITY: APPROXIMATE 57.90.0 0% (APPROXIMATELY 58%)
SPECIFICITY: APPROXIMATELY 75.85% (APPROXIMATELY 76%)
THESE METRICS DEMONSTRATE THAT OUR MODEL MAINTAINS ITS PERFORMANCE ACROSS DIFFERENT DATASETS, SUGGESTING ITS RELIABILITY AND EFFECTIVENESS IN PREDICTING CONVERSIONS.

Model parameters

- The final Logistic Regression Model has 12 features

The top 3 features that contribute positively to predicting hot leads in the model are:

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional
-

NOTE: The Optimal cutoff probability point is 0.365. A converted probability greater than 0.345 will be predicted as Converted lead (Hot lead) & probability smaller than 0.345 will be predicted as not Converted lead (Cold lead).

Conclusion

Train - Test

Train Data Set:

- Accuracy: 71.10%
- Sensitivity: 46.59%
- Specificity: 87.95%

Test Data Set:

- Accuracy: 68.76%
- Sensitivity: 57.90% ≈ 58%
- Specificity: 75.85%

NOTE: The evaluation metrics demonstrate consistency between the performance of the model on both the train and test datasets.

- In the train set, the model achieved a sensitivity of 46.59%, while in the test set, it reached 57.90%, using a cut-off value of 11.31.
- Sensitivity reflects the proportion of correctly identified converting leads out of all potential converting leads.
- The CEO's target sensitivity of approximately 75% remains unmet by the model.
- Additionally, the model attained an accuracy of 71.10%, aligning with the study's objectives.

RECOMMENDATION

- Emphasize features with positive coefficients to tailor marketing strategies effectively.
- Implement targeted campaigns to attract high-quality leads from the most successful lead sources.
- Craft personalized messages aimed at engaging working professionals effectively.
- Optimize communication channels based on their impact on lead engagement.
- Allocate more budget to advertising on the Welingak Website for increased visibility.
- Offer incentives or discounts for successful lead referrals to encourage more referrals.
- Aggressively target working professionals due to their high conversion rates and potentially better financial situations for higher fees.
- Identify specialized offering
- Improve landing page for more conversion