Unraveling the Dynamics of Lead Scores: Insights and Recommendations

Lead score analysis involves the intricate evaluation and assignment of scores to potential sales leads, considering diverse criteria and behaviors. The primary objective is to prioritize and pinpoint leads with the highest likelihood of converting into customers. This analytical process empowers sales and marketing teams to channel their efforts effectively, concentrating on leads with superior potential. The result is enhanced efficiency and increased probabilities of successful conversions.

- In our dataset, we encounter a diverse array of variables encompassing various factors and characteristics that warrant thorough analysis.

Variables Description

1. **Prospect ID:** A unique ID with which the customer is identified.
2. **Lead Number:** A lead number assigned to each lead procured.
3. **Lead Origin:** The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
4. **Lead Source:** The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
5. **Do Not Email:** An indicator variable selected by the customer indicating whether or not they want to be emailed about the course.
6. **Do Not Call:** An indicator variable selected by the customer indicating whether or not they want to be called about the course.
7. **Converted:** The target variable indicating whether a lead has been successfully converted or not.
8. **TotalVisits:** The total number of visits made by the customer on the website.
9. **Total Time Spent on Website:** The total time spent by the customer on the website.
10. **Page Views Per Visit:** Average number of pages on the website viewed during the visits.
11. **Last Activity:** Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.
12. **Country:** The country of the customer.
13. **Specialization:** The industry domain in which the customer worked before. Includes the level 'Select Specialization,' which means the customer had not selected this option while filling the form.
14. **How did you hear about X Education:** The source from which the customer heard about X Education.
15. **What is your current occupation:** Indicates whether the customer is a student, unemployed, or employed.
16. **What matters most to you in choosing this course:** An option selected by the customer indicating their main motive behind doing this course.
17. **Search:** Indicates whether the customer had seen the ad in any of the listed items.
18. **Magazine:** Indicates whether the customer had seen the ad in any of the listed items.
19. **Newspaper Article:** Indicates whether the customer had seen the ad in any of the listed items.
20. **X Education Forums:** Indicates whether the customer had seen the ad in any of the listed items.
21. **Digital Advertisement:** Indicates whether the customer had seen the ad in any of the listed items.
22. **Through Recommendations:** Indicates whether the customer had seen the ad in any of the listed items.
23. **Receive More Updates About Our Courses:** Indicates whether the customer chose to receive more updates about the courses.
24. **Tags:** Tags assigned to customers indicating the current status of the lead.

25. **Lead Quality:** Indicates the quality of lead based on the data and intuition of the employee who has been assigned to the lead.
26. **Update me on Supply Chain Content:** Indicates whether the customer wants updates on the Supply Chain Content.
27. **Get updates on DM Content:** Indicates whether the customer wants updates on the DM Content.
28. **Lead Profile:** A lead level assigned to each customer based on their profile.
29. **City:** The city of the customer.
30. **Asymmetrique Activity Index:** An index and score assigned to each customer based on their activity and their profile.
31. **Asymmetrique Profile Index:** An index and score assigned to each customer based on their activity and their profile.
32. **Asymmetrique Activity Score:** An index and score assigned to each customer based on their activity and their profile.
33. **Asymmetrique Profile Score:** An index and score assigned to each customer based on their activity and their profile.
34. **I agree to pay the amount through cheque:** Indicates whether the customer has agreed to pay the amount through cheque or not.
35. **A free copy of Mastering The Interview:** Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.
36. **Last Notable Activity:** The last notable activity performed by the student.

Steps for Lead Score Analysis Project

1. **Define Objectives:**
    - Clearly outline the objectives of your lead score analysis project. What specific outcomes are you aiming for?
2. **Data Gathering:**
    - Collect the dataset containing information on leads, ensuring it covers relevant variables for analysis.
3. **Data Inspection:**
    - Explore the dataset to understand its structure, features, and any initial patterns.
4. **Data Cleaning:**
    - Handle missing values, address outliers, and preprocess the data to ensure its quality.
5. **Exploratory Data Analysis (EDA):**
    - Visualize the data using charts and graphs to gain insights into the distribution and relationships of variables.
6. **Feature Selection:**
    - Identify key features that are likely to impact lead scores and focus on those during analysis.
7. **Define Lead Scoring Model:**
    - Choose a suitable lead scoring model or methodology based on the characteristics of your dataset.
8. **Data Preprocessing:**
    - Prepare the data for modeling, including encoding categorical variables, scaling, and splitting into training and testing sets.
9. **Model Training:**

- Train your lead scoring model using the prepared dataset.

10. **Model Evaluation:**
    - Assess the performance of your model using relevant metrics such as accuracy, precision, recall, or F1 score.
11. **Adjust and Optimize:**
    - Fine-tune your model based on the evaluation results, iterating as needed for better performance.
12. **Interpret Results:**
    - Analyze the results to understand the factors influencing lead scores and draw actionable insights.
13. **Documentation:**
    - Document your entire process, including code, data transformations, and model parameters.
14. **Communication:**
    - Clearly communicate your findings and insights to relevant stakeholders, using visualizations to aid understanding.
15. **Iterate and Improve:**
    - If necessary, iterate on your analysis based on feedback or new information to continuously improve your lead scoring system.
16. **Implementation:**
    - Implement the lead scoring system into your workflow, ensuring seamless integration with sales and marketing processes.
17. **Monitoring:**
    - Regularly monitor and update the lead scoring model to adapt to changes in customer behavior or market conditions.
18. **Feedback Loop:**
    - Establish a feedback loop for continuous improvement based on the performance of the lead scoring system in real-world scenarios.

1) Define Objectives:

Defining objectives is a critical step in any project, including lead score analysis. Objectives provide a clear direction and purpose for the analysis. In the context of a lead score analysis project, the objectives might include:

- **Improve Lead Conversion Rates:**

Increase the efficiency of lead management by identifying and prioritizing leads with a higher likelihood of conversion.

- **Enhance Sales and Marketing Alignment:**

Foster better collaboration between sales and marketing teams by providing a standardized lead scoring system that aligns with business goals.

- **Optimize Resource Allocation:**

Allocate sales and marketing resources more effectively by focusing efforts on leads that are more likely to convert, thereby maximizing the return on investment.

- **Increase Customer Engagement:**

Develop strategies to engage leads more effectively by tailoring communication and marketing efforts based on their lead scores and behaviors.

- **Reduce Time to Conversion:**

Shorten the time it takes for leads to progress through the sales funnel by identifying and addressing bottlenecks or areas of improvement.

- **Improve Overall Customer Acquisition:

Contribute to the growth of the customer base by consistently identifying and converting high-value leads.

- **Enhance Data-Driven Decision-Making:**

Establish a data-driven approach to lead management, enabling continuous improvement based on insights gained from the lead score analysis.

- **Increase Revenue and Return on Investment (ROI):**

Drive revenue growth and improve ROI by focusing on leads that are more likely to generate sales.

- **Implement Predictive Lead Scoring:**

Move towards a more predictive lead scoring model that leverages advanced analytics and machine learning to improve the accuracy of lead predictions.

- **Establish a Scalable and Sustainable System:**

Create a lead scoring system that is scalable and adaptable to changing business conditions, ensuring long-term effectiveness

2) Data Gathering:

- Data gathering is the process of collecting and compiling relevant information or datasets that will be used for analysis, research, or decision-making. In the context of a lead score analysis project, data gathering involves acquiring the necessary information about potential leads to build a comprehensive dataset.

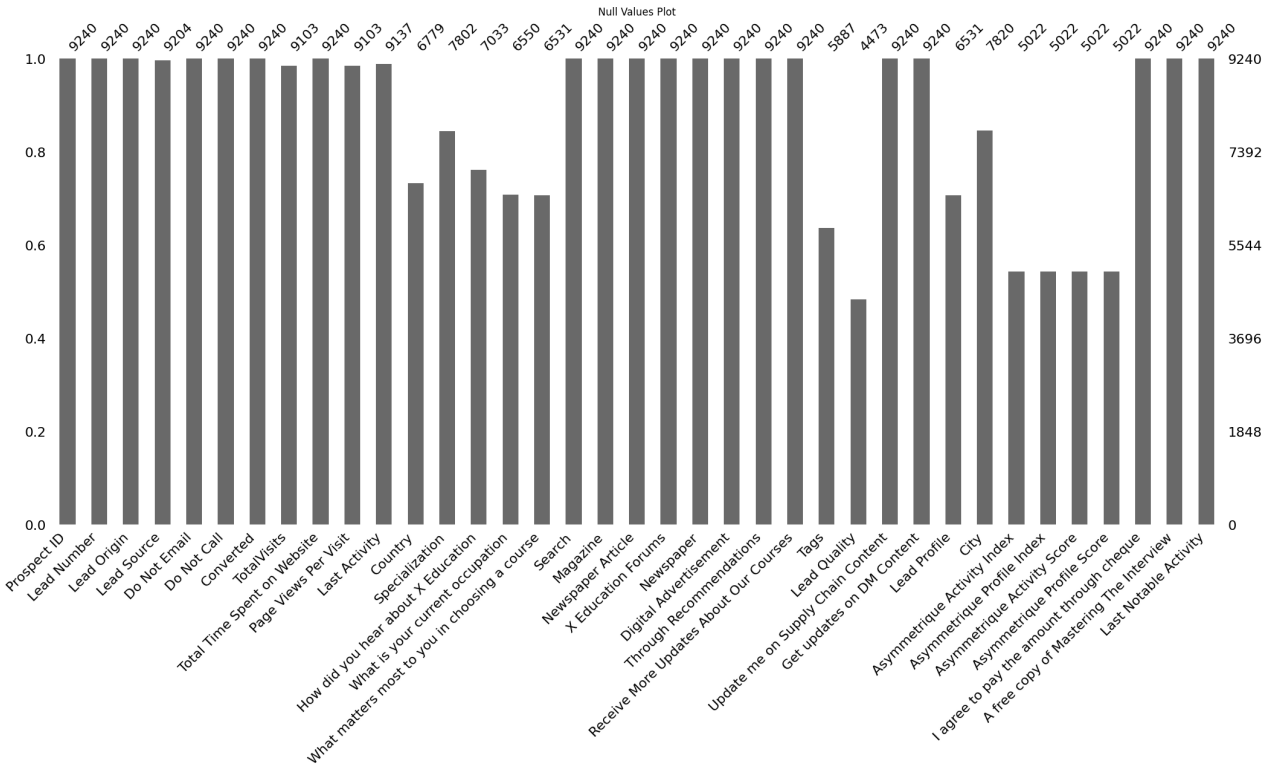
```
import numpy as np
#NumPy is a fundamental package for scientific computing in Python.
It provides support for large, multi-dimensional arrays and matrice
s, along with mathematical functions to operate on these arrays.

import pandas as pd
# Pandas is a powerful data manipulation and analysis library for Py
thon. It provides easy-to-use data structures such as Series and Dat
aFrame, along with a variety of functions for data cleaning, explora
tion, and transformation

import matplotlib.pyplot as plt # use for ploting
import seaborn as sns # use for ploting
import missingno as ms
#missingno is a Python library that provides a visual representation
of missing (null or NaN) values in a dataset. It is particularly use
ful during the data exploration phase to quickly identify patterns o
f missing data.

import warnings
warnings.filterwarnings("ignore") # ignore the warning
```

3) Data Inspection:



The table below shows the percentage of missing values for each variable in the dataset:

Variable	Missing Percentage
Prospect ID	0.00%
Lead Number	0.00%

Variable	Missing Percentage
Lead Origin	0.00%
Lead Source	0.39%
Do Not Email	0.00%
Do Not Call	0.00%
Converted	0.00%
TotalVisits	1.48%
Total Time Spent on Website	0.00%
Page Views Per Visit	1.48%
Last Activity	1.11%
Country	26.63%
Specialization	15.56%
How did you hear about X Education	23.89%
What is your current occupation	29.11%
What matters most to you in choosing a course	29.32%
Search	0.00%
Magazine	0.00%
Newspaper Article	0.00%
X Education Forums	0.00%
Newspaper	0.00%
Digital Advertisement	0.00%
Through Recommendations	0.00%
Receive More Updates About Our Courses	0.00%
Tags	36.29%
Lead Quality	51.59%
Update me on Supply Chain Content	0.00%
Get updates on DM Content	0.00%
Lead Profile	29.32%
City	15.37%
Asymmetrique Activity Index	45.65%
Asymmetrique Profile Index	45.65%
Asymmetrique Activity Score	45.65%
Asymmetrique Profile Score	45.65%
I agree to pay the amount through cheque	0.00%
A free copy of Mastering The Interview	0.00%

- Upon closer examination, we've determined that filling null values with the mode or any specific value might introduce unintended skewness to the data. To address this, we recommend the creation of a new category, such as "Other." This approach enhances transparency and ensures a more accurate representation of missing or unspecified information in our analysis.

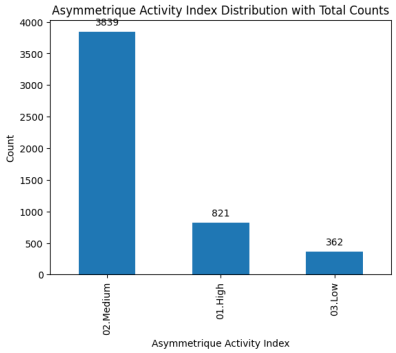
4) Data Cleaning:

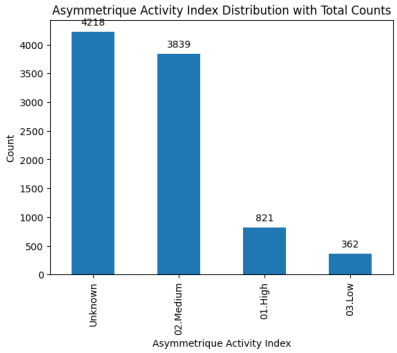
- For columns with null values exceeding 25%, the approach adopted is to replace the missing values with the label "**unknown**"

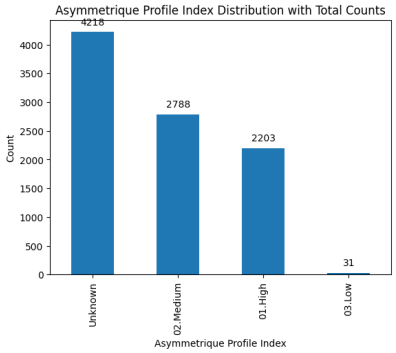
The table below shows more than 25% percentage of null values for the specified columns:

Column	Null Percentage
Lead Quality	51.59%
Asymmetrique Activity Index	45.65%
Asymmetrique Profile Index	45.65%
Tags	36.29%
Lead Profile	29.32%
Country	26.63%

/table>



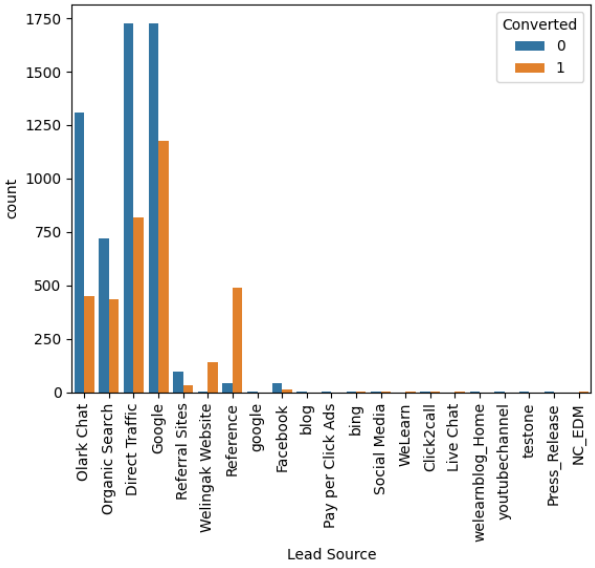


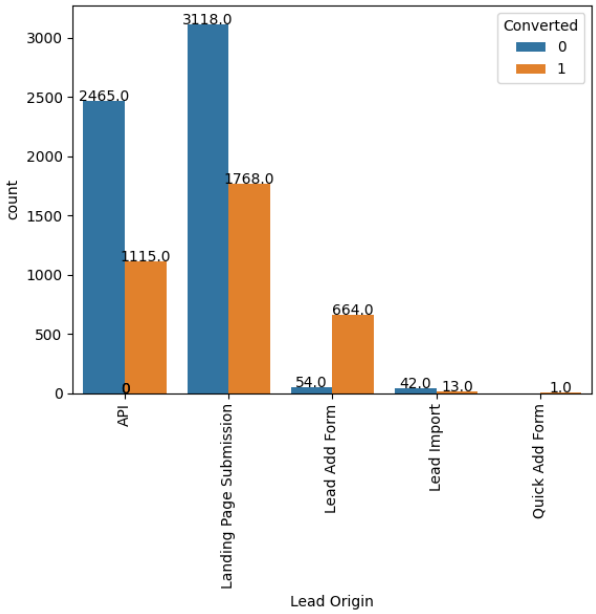


For columns with less than 25% null values,We handel with

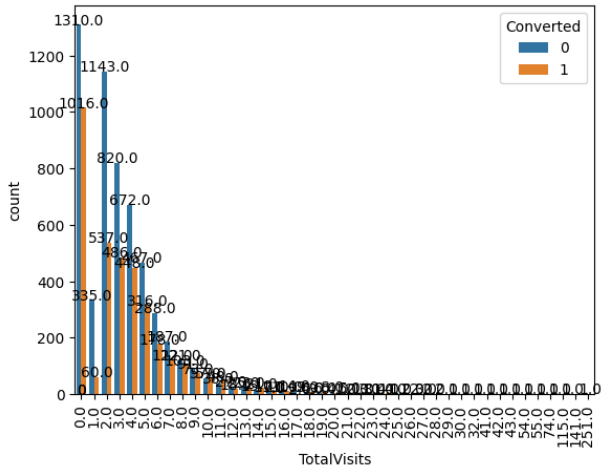
- **TotalVisits:** 0
- **Asymmetrique Profile Score:** 0
- **Asymmetrique Activity Score:** 0
- **What matters most to you in choosing a course:** Better Career Prospects
- **Current Occupation:** Unemployed
- **How did you hear about X Education:** Select
- **Specialization:** Select
- **City:** Select
- **Page Views Per Visit:** 0
- **Last Activity:** Unreachable
- **Lead Source:** `df['Lead Source'].mode()[0]`

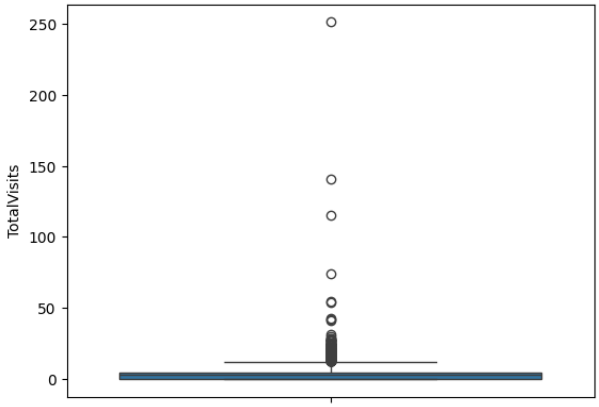
5) Exploratory Data Analysis (EDA)

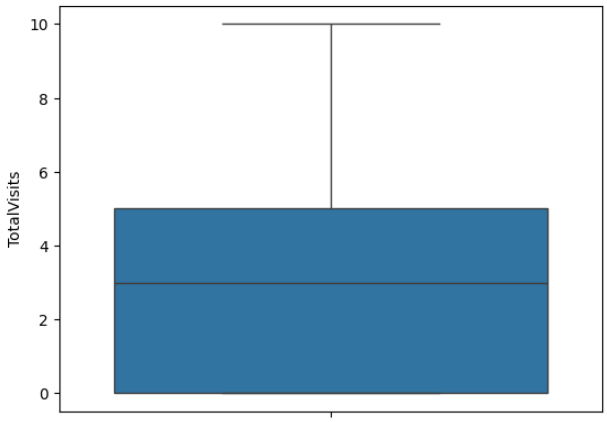


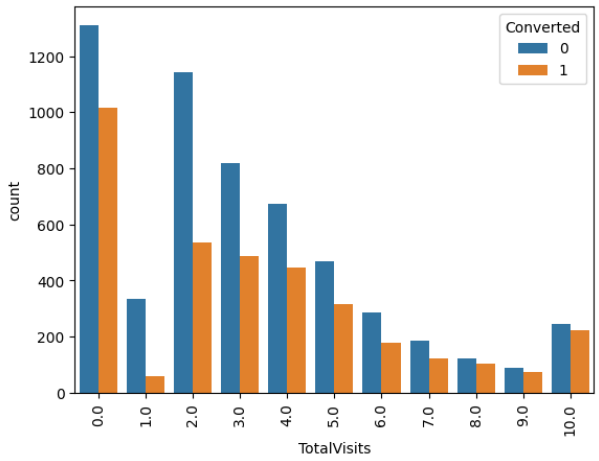


- **Lead Score:** It is evident that a substantial majority of leads originate from Google and direct traffic.
- **Lead Origin:** The analysis reveals a significant disparity in lead conversion rates based on the chosen method. Notably, individuals who utilized the "lead add form" exhibited an impressive conversion rate exceeding 92%, while those who opted for landing page submission demonstrated a still commendable conversion rate surpassing 36%.

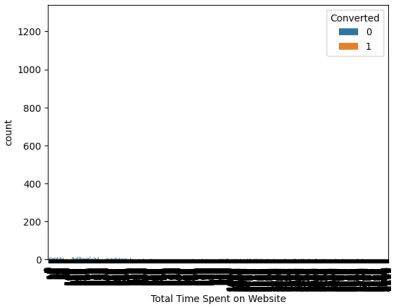


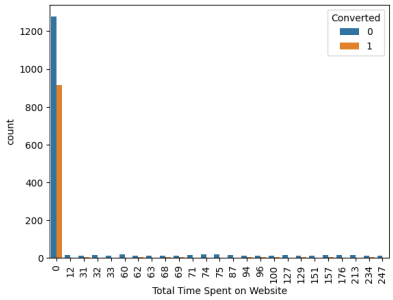


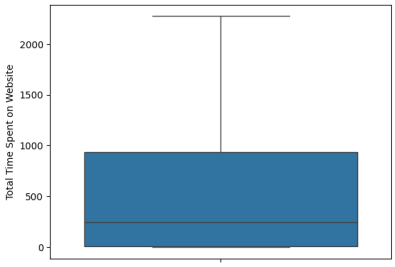




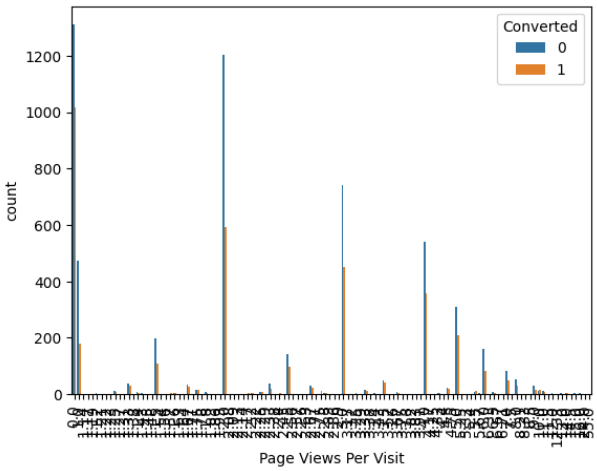
An intriguing pattern emerges as individuals visiting the website for the first time show a high likelihood of course enrollment. However, as the frequency of visits gradually increases, the lead conversion rate experiences a continuous decline. Beyond 30 visits to the website, the chance of conversion diminishes significantly, approaching nearly zero.

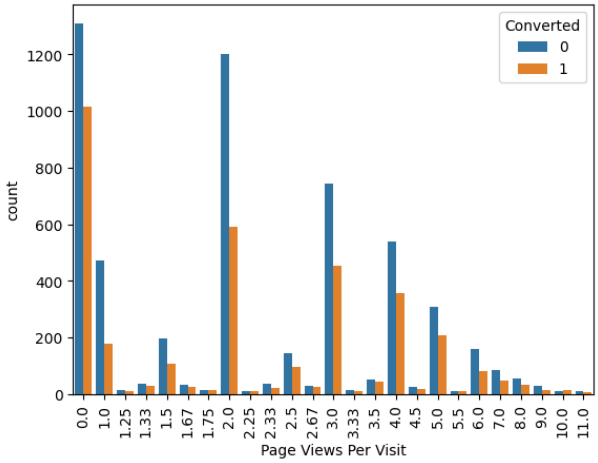




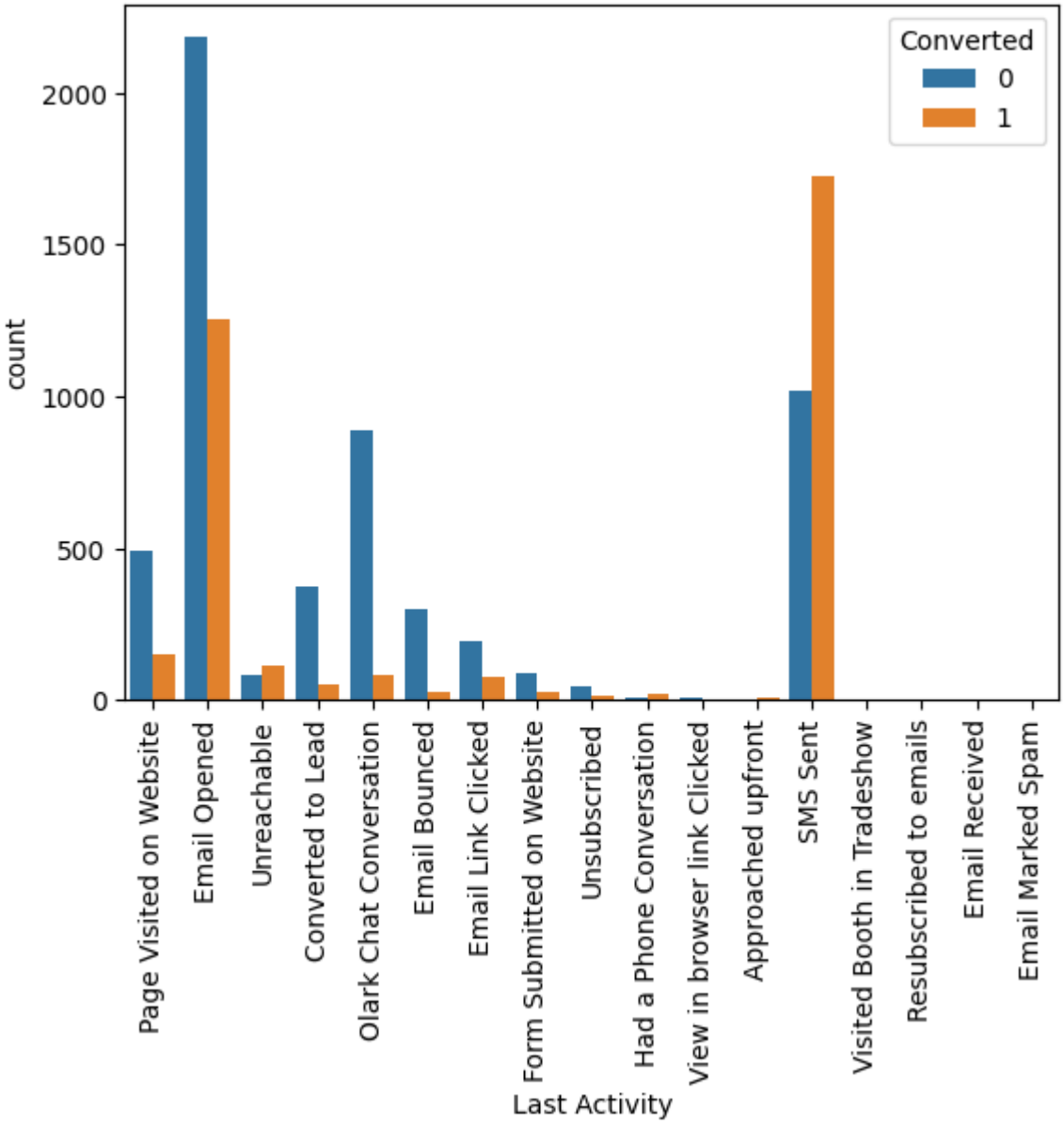


In the second image, our focus shifts to extracting the top 25 values. Intriguingly, individuals who spend zero time on the website exhibit a notably high chance of 'conversion,' whereas those who invest time on the website experience lower conversion rates. This challenges the notion that conversion is contingent on the duration of time spent on the website, as suggested by the data in the specified column.





An intriguing pattern unfolds, indicating that individuals with fewer page views per visit demonstrate a notably higher chance of conversion. However, this probability gradually diminishes as the number of visits increases, revealing a inverse relationship between page views per visit and the likelihood of conversion."

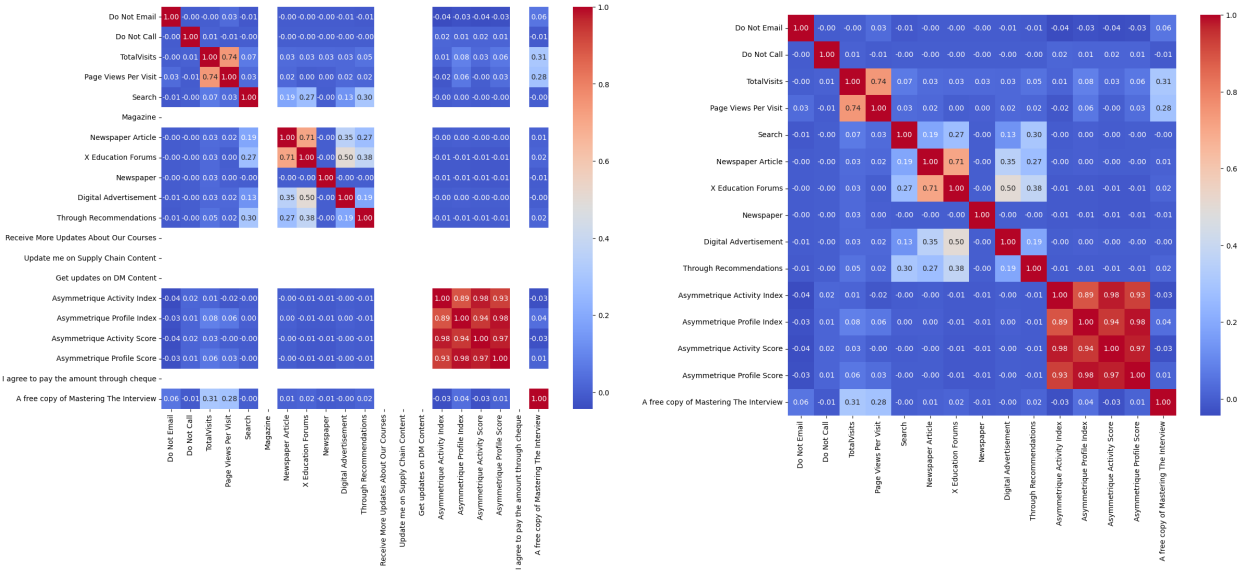


Individuals who actively send messages demonstrate a notably higher chance of conversion to a lead. Additionally, those who open emails exhibit an approximately 40% conversion rate, emphasizing the positive correlation between engagement through messaging and a heightened likelihood of successful conversions.

- To facilitate analysis, categorical columns with only two categories, such as 'Yes' and 'No', have been mapped to numerical values. 'Yes' has been encoded as 1, while 'No' has been encoded as 0 across the specified columns.

```
df.replace({'Yes': 1, 'No': 0}, inplace=True)
```

- except out target coloum "Converted" we have some work with the coloum right now then at the end of our analysis we can again change change it to neumerical




```
df.drop(columns=zero_columns,axis=1,inplace=True)
```

- **Columns that exclusively contain zero values (denoted as 'NO') contribute no meaningful information and can be safely removed from the dataset. This not only simplifies the data but also has the potential to enhance model performance**

After the multivariate analysis, the target column is transformed back into numeric form

```
df['Converted'].replace({'Yes': 1, 'No': 0}, inplace=True)
```

```
# Identifying categorical columns
categorical_columns = df.select_dtypes(include=['object']).columns

# Displaying the list of categorical columns
categorical_columns

# Creating dummy variables for categorical columns and dropping the
first one
dummy_data = pd.get_dummies(df[categorical_columns], drop_first=Tru
e)

# Converting boolean values to integers (1 and 0)
dummy_data = dummy_data.astype(int)

# Concatenating the dummy_data with the original DataFrame
df = pd.concat([df, dummy_data], axis=1)

# Converting boolean values to integers (1 and 0)

# Displaying the resulting DataFrame
df.head()
```

Now that our data has been successfully cleaned and converted into a numeric format, we are ready to proceed with splitting the dataset for further prediction tasks.

Reasons for Splitting Data in Machine Learning

1. **Model Evaluation:** The primary goal of splitting the data is to assess how well your machine learning model performs on unseen data. By training the model on one subset (training set) and testing it on another, you can get an estimate of its performance on data it has never seen before.
2. **Preventing Overfitting:** If a model is trained on the entire dataset, it might memorize the data and perform well on it but fail to generalize to new, unseen data. Splitting the data helps to identify if the model is overfitting (performing well on the training data but poorly on new data) by evaluating its performance on a separate test set.
3. **Hyperparameter Tuning:** When fine-tuning or selecting the best hyperparameters for a model, it is essential to have a separate validation set. This set is used to evaluate different configurations and choose the one that performs well on both the training and validation data.
4. **Performance Metrics:** Splitting the data allows you to calculate various performance metrics, such as accuracy, precision, recall, F1 score, etc., on the test set. These metrics provide insights into how well the model generalizes and whether it meets the desired criteria.
5. **Avoiding Data Leakage:** Mixing training and testing data may lead to data leakage, where information from the test set influences the model during training. This can result in overly optimistic performance estimates. Separating the sets helps prevent such issues.

Typical Split:

- **Training Set:** Used to train the machine learning model.
- **Validation Set:** Used to fine-tune hyperparameters and evaluate model performance during training.
- **Test Set:** Reserved for the final evaluation of the model's generalization to unseen data.

The specific ratio of the split (e.g., 70% training, 15% validation, 15% test) depends on the size of the dataset and the goals of the machine learning project. Cross-validation techniques may also be

Here we use 80% traing and 20% test

```
# Split the data into training and testing sets (adjust test_size as
needed)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=
0.2, random_state=42)

# Display the shapes of the resulting sets
print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

- X_train shape: (7392, 177)
- X_test shape: (1848, 177)
- y_train shape: (7392,)
- y_test shape: (1848,)

before going with model creation we have to scale the data for better prediction

Importance of Scaling Data in Machine Learning

Scaling data is a crucial preprocessing step in machine learning to ensure that all features contribute equally to the model training process. It aids in achieving better prediction performance and is particularly important for certain algorithms sensitive to the scale of input features.

Reasons for Scaling:

1. **Equal Contribution:** Features with larger scales might dominate the learning process, causing the model to give more weight to those features. Scaling ensures that all features contribute more equally to the learning process.
2. **Convergence Speed:** Many machine learning algorithms converge faster when the features are on a similar scale. This is especially true for optimization algorithms like gradient descent.
3. **Regularization:** Some regularization techniques, like L1 and L2 regularization, are sensitive to the scale of the features. Scaling can help ensure that regularization has a similar impact on all features.
4. **Distance-based Algorithms:** Algorithms that rely on distances between data points, such as k-nearest neighbors or support vector machines, can be affected by the scale of the features. Scaling helps prevent one feature from dominating the distance computation.

Common Scaling Methods:

- **Min-Max Scaling:** Scales the data to a specific range, often between 0 to 1 or -1 to 1.
- **Standard Scaling (Z-score normalization):** Scales the data to have a mean of 0 and a standard deviation of 1.

- **Robust Scaling:** Scales the data based on the interquartile range, making it more robust to outliers.

The choice of scaling method depends on the characteristics of your data and the requirements of your machine learning algorithm. It's often a good practice to scale your features before training a

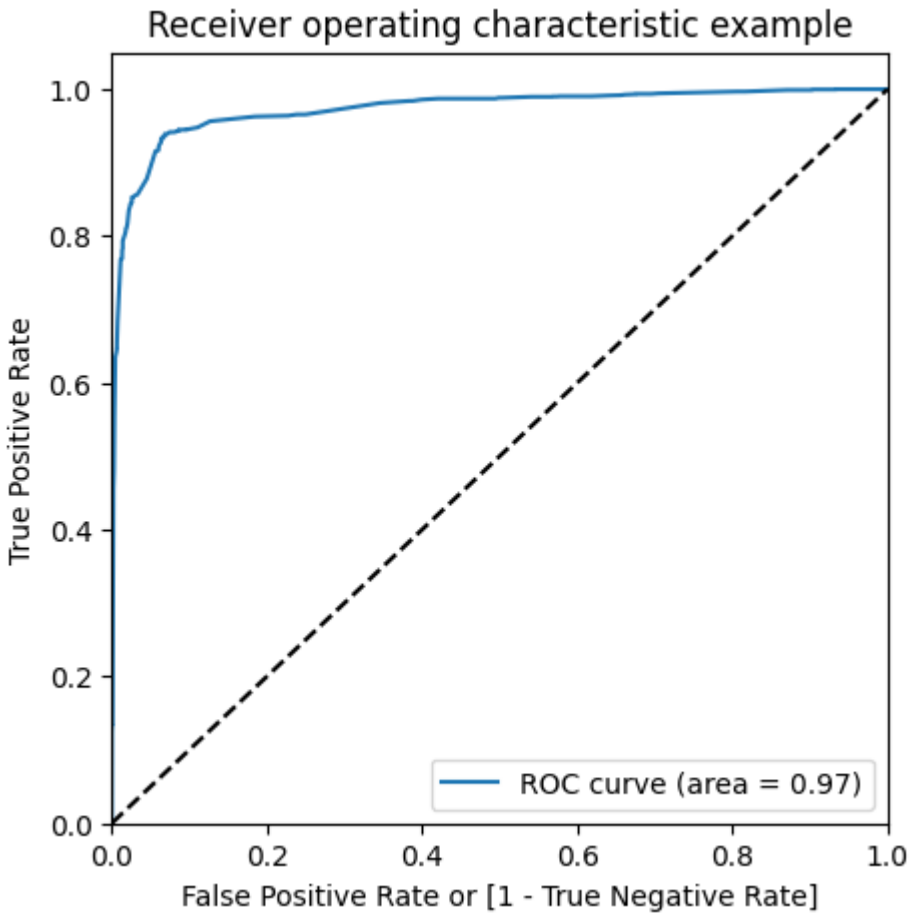
After scaling the features, performing a thorough P-value analysis, and calculating the Variance Inflation Factor (VIF), a decision has been made to drop the 'Tags_Unknown' column. This decision is informed by several considerations. The P-value analysis indicates that 'Tags_Unknown' does not significantly contribute to the model, as its P-value exceeds the conventional significance level. Additionally, the VIF calculation suggests a potential issue of multicollinearity, as 'Tags_Unknown' is exhibiting a high correlation with other variables in the dataset.

- By dropping the 'Tags_Unknown' column, we aim to enhance the model's interpretability, reduce redundancy, and mitigate any adverse effects of multicollinearity on the accuracy and stability of our predictive model.

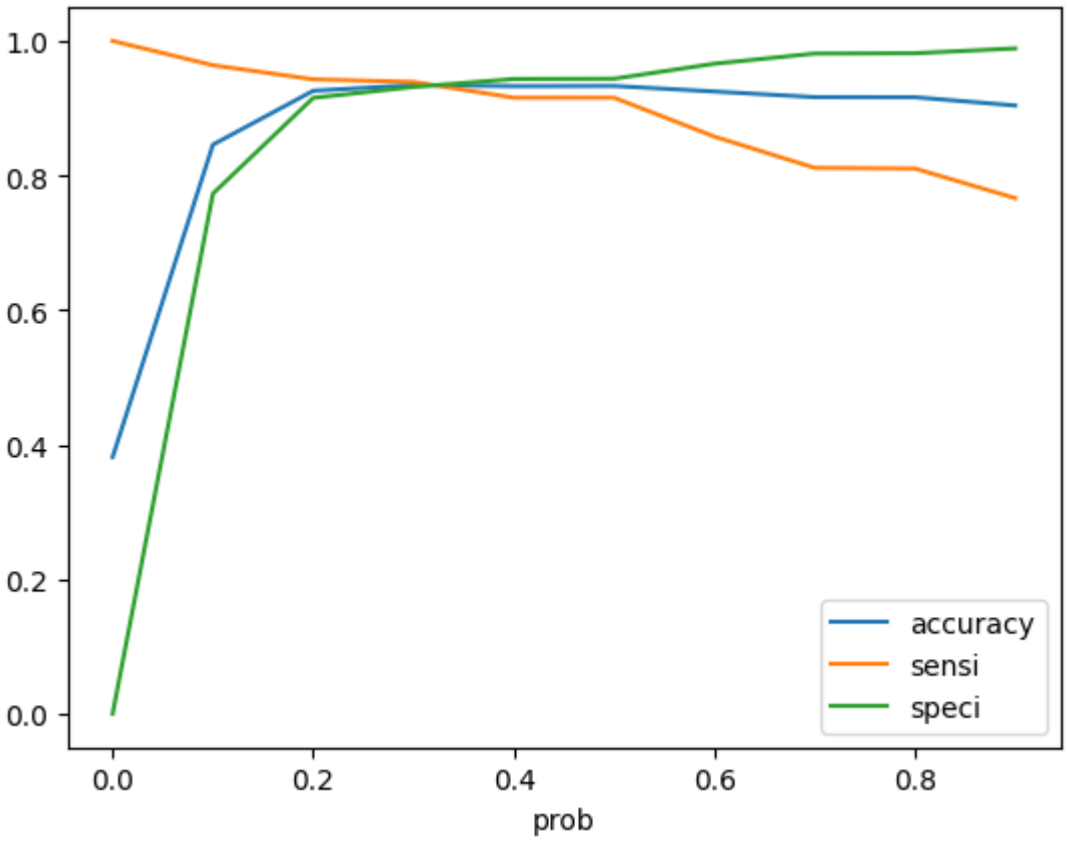
```
X_train[col1].drop('Tags_Unknown', axis=1,inplace=True)
```

- After that, we drew the ROC curve to better understand the model's performance in terms of sensitivity and specificity.

```
draw_roc(y_train_pred_final.Converted, y_train_pred_final.Converted_
prob)
```

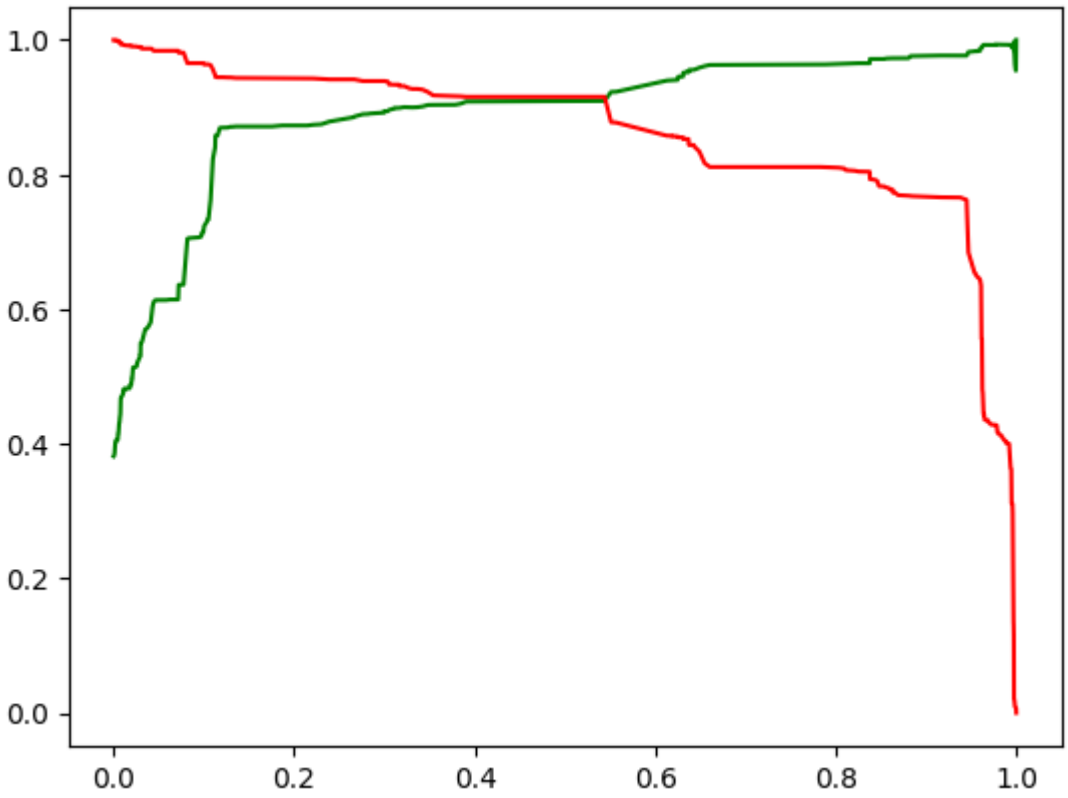


Finding Optimal Cutoff Point



Precision and Recall

```
p, r, thresholds = precision_recall_curve(y_train_pred_final.Convert
ed, y_train_pred_final.Converted_prob)
# plotting a trade-off curve between precision and recall
plt.plot(thresholds, p[:-1], "g-")
plt.plot(thresholds, r[:-1], "r-")
plt.show()
```



Making predictions on the test set

Observations: After running the model on the Test Data , we obtain:

- Accuracy : 91.9 %
- Sensitivity : 91.9 %
- Specificity : 93.6 %

Final Results

Train Data:

- **Accuracy:** 93.3%
- **Sensitivity:** 92.6%
- **Specificity:** 93.7%

Test Data:

- **Accuracy:** 92.9%
- **Sensitivity:** 91.9%
- **Specificity:** 93.6%

Finding out the leads which should be contacted:

Business Recommendations

Tags Recommendations:

- Focus on leads with the tag "Lost to EINS" as it has the highest positive coefficient. These leads are more likely to convert.
- Consider engaging with leads tagged as "Closed by Horizzon" and "Will revert after reading the email," as they also have significant positive coefficients.
- Investigate and address leads with the "Unknown" tag. It has a positive coefficient but lower than the top tags.
- Monitor and engage with leads tagged as "Ringing," "invalid number," and "switched off," as they have negative coefficients. There might be issues that need attention.

Lead Source:

- Pay attention to leads from the "Welingak Website" as it has a positive coefficient, indicating a higher likelihood of conversion.

Last Notable Activity:

- Leads with the last notable activity being "SMS Sent" are more likely to convert, according to the positive coefficient.
- Consider addressing and understanding leads marked as "Unsubscribed."

Other Features:

- Leads with higher "Asymmetrique Activity Score" are more likely to convert.
- Engage with leads involved in "Olark Chat Conversation" as it has a negative coefficient but might present opportunities for conversion.
- Consider the impact of "Do Not Email" on lead conversion. The negative coefficient suggests caution with this feature.
- Evaluate the effect of leads that were "Converted to Lead" in the last activity.
- Pay attention to the "Worst" category in "Lead Quality," as it has a negative coefficient.
- Leads with "invalid number" and "switched off" tags may need verification and correction of contact details.
- Address leads with "Unknown" lead profiles, as it has a negative coefficient.

These recommendations are based on the coefficients or importance scores from your model. It's crucial to interpret these results in the context of your business and make informed decisions. Additionally, consider conducting further analysis and A/B testing to validate the impact of these recommendations on actual lead conversions.