Case Study - Leads Scoring

Vu Manh Cuong

Data Understanding

In [3]:
```
# Get the shape of the dataset
shape = leads_data.shape

# Get datatypes of the features
datatypes = leads_data.dtypes

# Count of missing values in each column
missing_values = leads_data.isnull().sum()

# Basic statistical summary for numerical columns
summary = leads_data.describe()

shape, datatypes, missing_values, summary
```

Out[3]:
```
((9240, 37),
 Prospect ID                                    object
 Lead Number                                     int64
 Lead Origin                                    object
 Lead Source                                    object
 Do Not Email                                   object
 Do Not Call                                    object
 Converted                                       int64
 TotalVisits                                   float64
 Total Time Spent on Website                     int64
 Page Views Per Visit                          float64
 Last Activity                                  object
 Country                                        object
 Specialization                                 object
 How did you hear about X Education             object
 What is your current occupation                object
 What matters most to you in choosing a course  object
 Search                                         object
 Magazine                                       object
```

Data Understanding and Preparation

Observations:

- Missing Values: We have missing values in several columns like 'Lead Source', 'Total Visits', 'Page Views Per Visit', 'Last Activity', 'Country', 'Specialization', 'How did you hear about X Education', and so on. Some columns like 'Lead Quality', 'Asymmetrique Activity Index', and 'Asymmetrique Profile Index' have a significant number of missing values.
- Data Types: We have a mix of object (categorical) and numerical data types in the dataset. This will be crucial when we proceed with data preparation and encoding.
- General Statistics: The dataset statistics provide insights into the distribution of numeric columns. For example, the average total time spent on the website is approximately 487 minutes, and the average page views per visit are around 2.36.

Handling Missing Values:

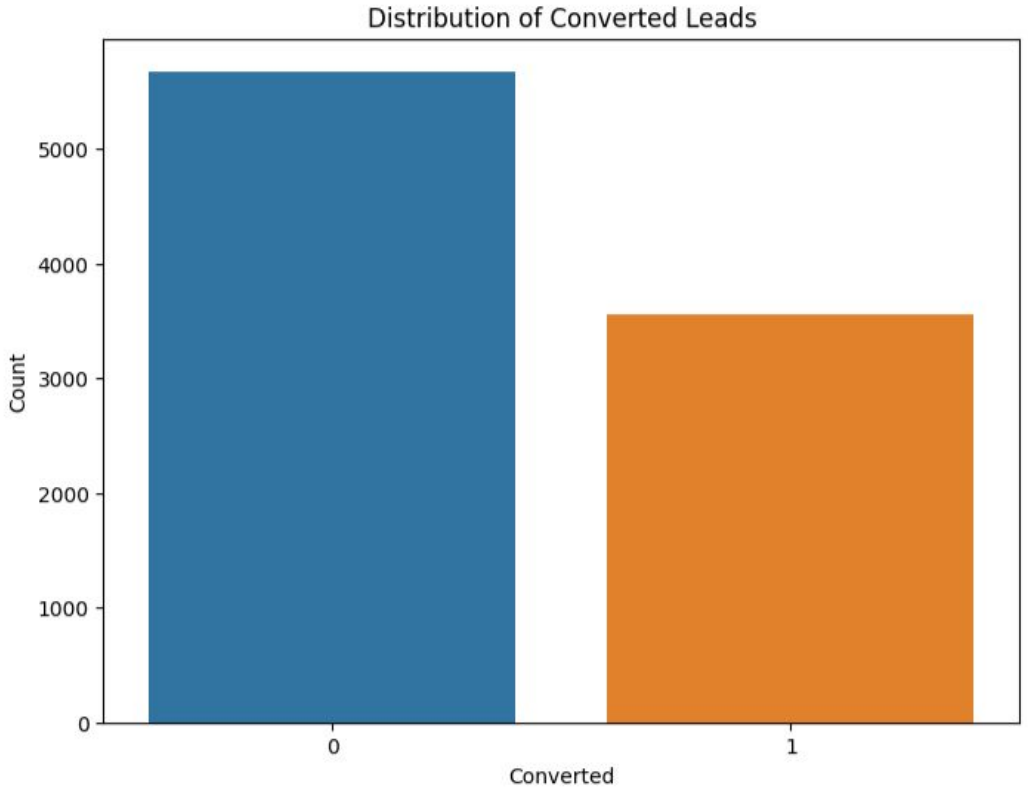
- We will first identify columns with a significant percentage of missing values and decide whether to drop or impute them.
- For categorical columns, we can replace the missing values with the mode or a placeholder like 'Unknown'.
- For numeric columns, we can replace the missing values with the median or mean, depending on the distribution.

Data Cleaning

- I handled missing values based on their proportion in the columns.
- I replaced the 'Select' level with 'Not Provided' in categorical columns.
- I checked for and ensured there are no duplicate rows.

Exploratory Data Analysis (EDA)

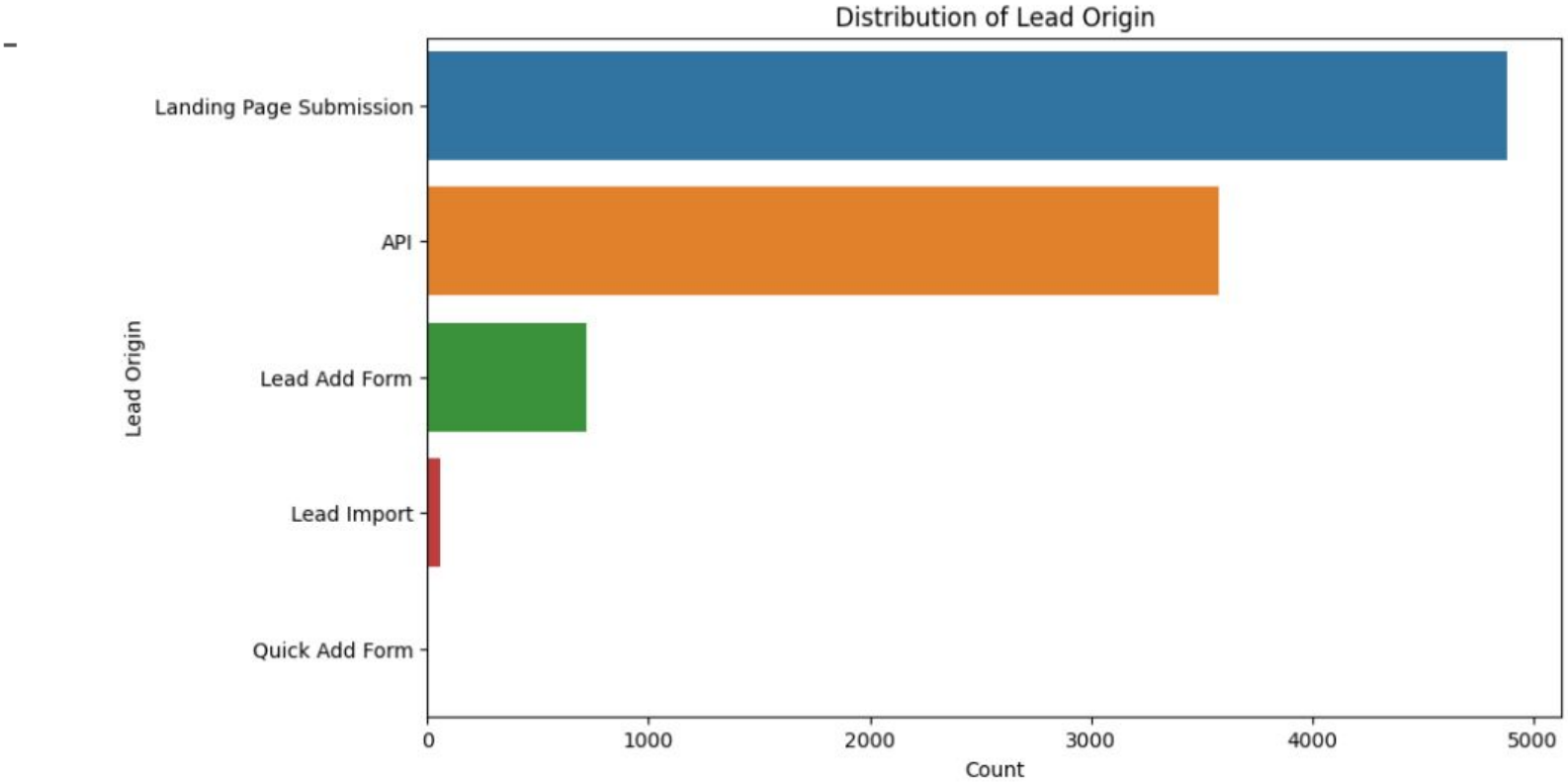
-



Exploratory Data Analysis (EDA)

- The distribution indicates that a significant number of leads were not converted (represented by 0), while a smaller proportion were converted (represented by 1). This aligns with the case-study statement that the lead conversion rate at X Education is around 30%.
- Examining some other key variables:
  - Lead Origin: To understand where the majority of the leads are originating from.
  - Lead Source: To understand the main sources from which potential customers are accessing the courses.
  - Total Time Spent on Website: To gauge engagement of the leads with the website.
  - Last Activity: To understand the most recent activity of the leads.

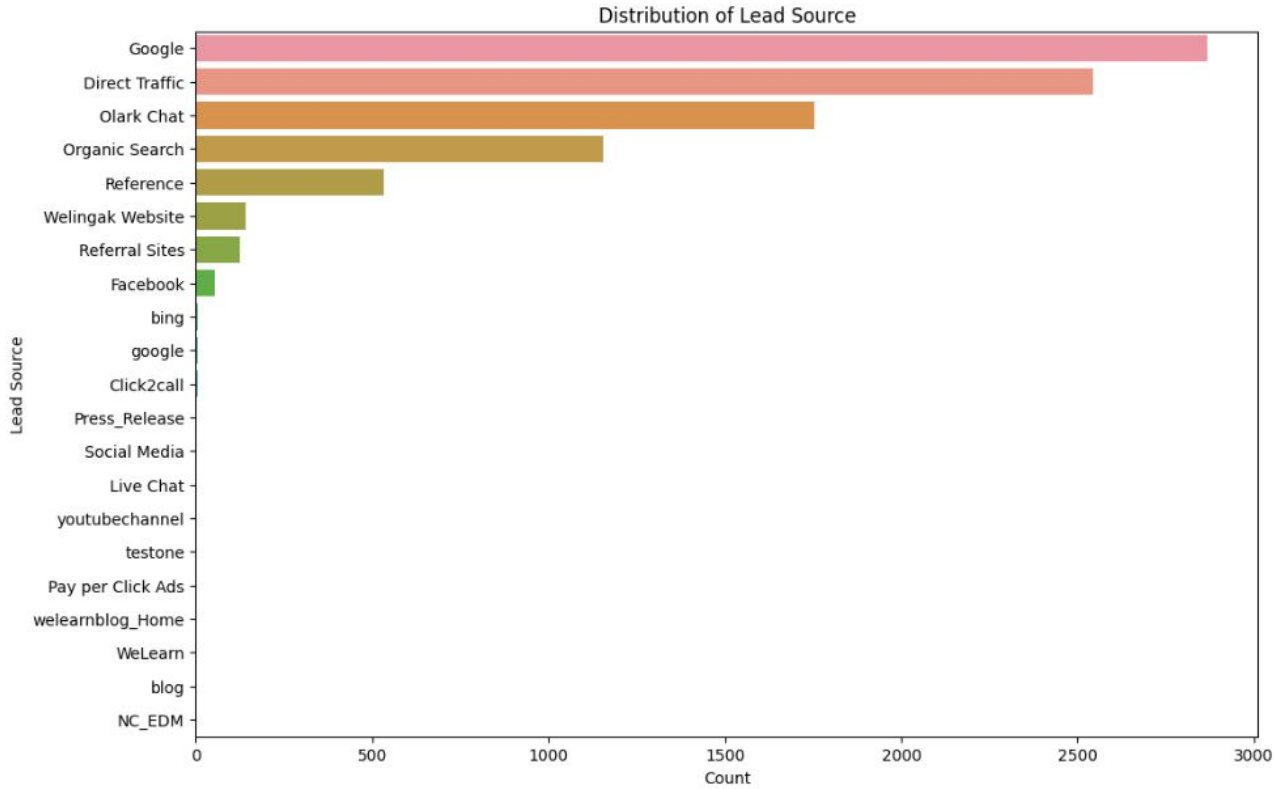
Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

- The majority of the leads seem to come from "Landing Page Submission", followed by "API". The least number of leads come from "Lead Add Form" and "Lead Import".

Exploratory Data Analysis (EDA)



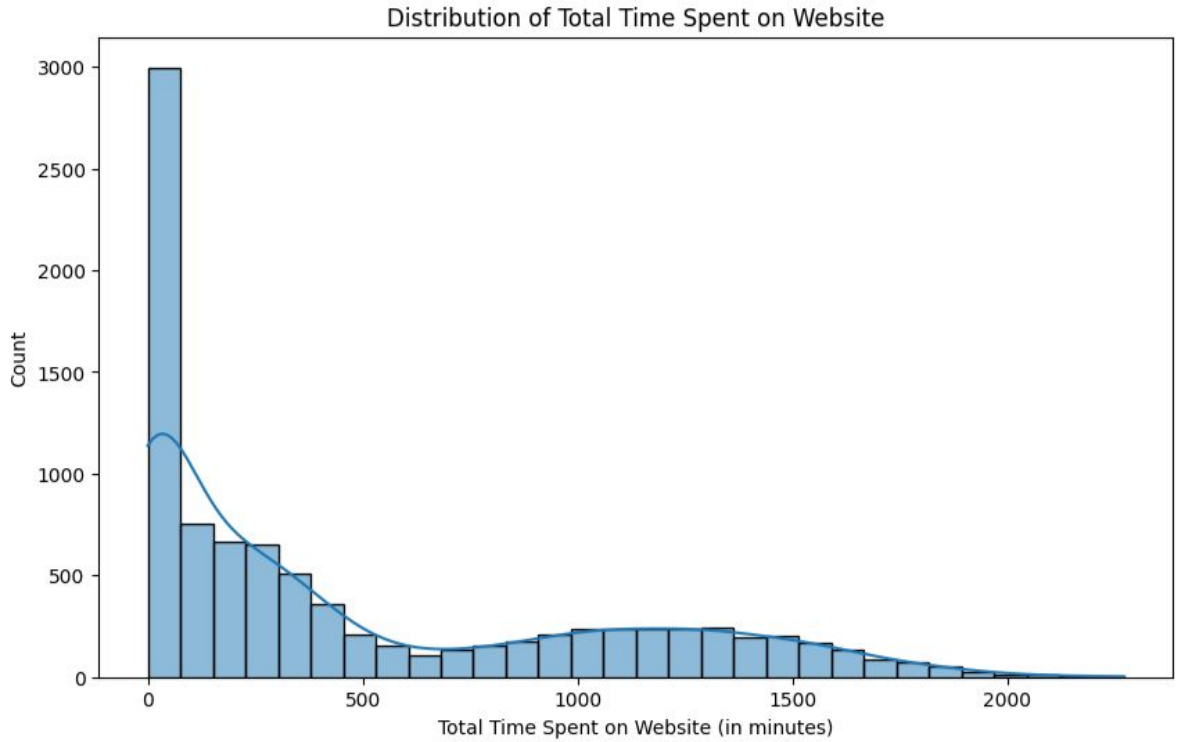
Exploratory Data Analysis (EDA)

From the distribution of the 'Lead Source' variable, we can observe:

- A significant number of leads come from "Google", indicating that search engine marketing or organic search plays a crucial role.
- The next major sources are "Direct Traffic" and "Olark Chat", followed by "Organic Search".
- Several other sources contribute fewer leads.

Exploratory Data Analysis (EDA)

-

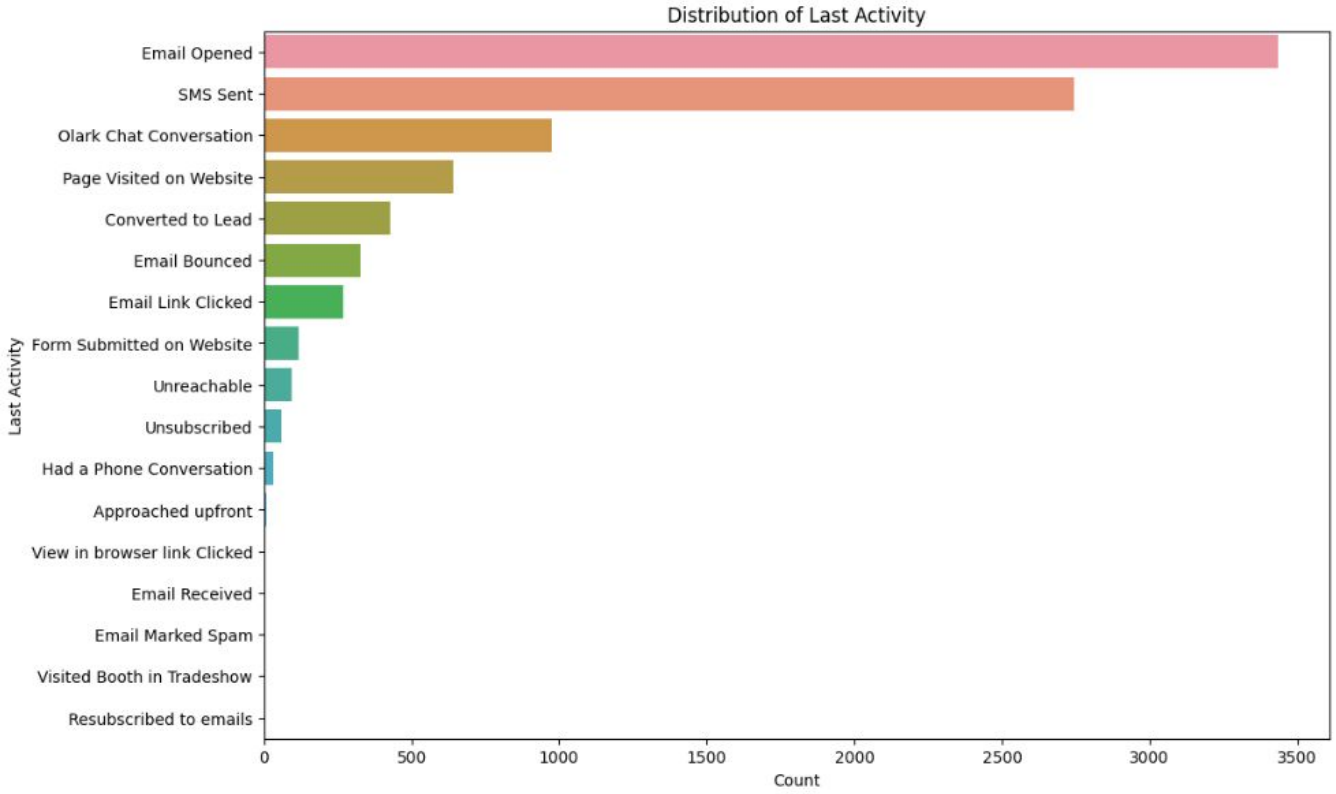


Exploratory Data Analysis (EDA)

The distribution of 'Total Time Spent on Website' shows:

- A significant number of leads spend very little time on the website, which might indicate that they just land on the website and leave without much interaction.
- There's a smooth increase in the number of leads as the time spent increases, reaching a peak around 500-600 minutes.
- After this peak, the number of leads decreases as the time spent continues to increase, indicating fewer leads spend an extremely long time on the website.

Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

From the distribution of the 'Last Activity' variable:

- The most common last activity is "Email Opened", indicating that many leads were last engaged through an email campaign or communication.
- The next most frequent activity is "SMS Sent", suggesting that SMS campaigns also play a significant role in engagement.
- "Olark Chat Conversation" is another notable activity, implying that chat support or inquiries also form a part of the lead's journey.
- Other activities have comparatively fewer counts.

Handle Missing Value

In [9]:
```
# Calculate the percentage of missing values for each column
missing_percentage = (leads_data.isnull().sum() / len(leads_data)) * 100

# Filter out columns with missing values for further inspection
missing_columns = missing_percentage[missing_percentage > 0].sort_values(ascending=False)

missing_columns
```

Out[9]:

Lead Quality	51.590909
Asymmetrique Profile Score	45.649351
Asymmetrique Activity Score	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Index	45.649351
Tags	36.287879
Lead Profile	29.318182
What matters most to you in choosing a course	29.318182
What is your current occupation	29.112554
Country	26.634199
How did you hear about X Education	23.885281
Specialization	15.562771
City	15.367965
TotalVisits	1.482684
Page Views Per Visit	1.482684
Last Activity	1.114719
Lead Source	0.389610

dtype: float64

Handle Missing Value

High Percentage of Missing Values (> 40%):

- 'Lead Quality', 'Asymmetrique Profile Score', 'Asymmetrique Activity Score', 'Asymmetrique Profile Index', and 'Asymmetrique Activity Index' have a significant amount of missing data. Since imputing such a large number of missing values might introduce bias, it might be best to drop these columns.

Medium Percentage of Missing Values (15%-40%):

- For 'Tags', 'Lead Profile', 'What matters most to you in choosing a course', 'What is your current occupation', 'Country', 'How did you hear about X Education', 'Specialization', and 'City', we can impute the missing values with a placeholder value, e.g., 'Unknown' or 'Not Provided'. This is especially true for categorical columns.

Low Percentage of Missing Values (< 15%):

- For 'Total Visits', 'Page Views Per Visit', 'Last Activity', and 'Lead Source', we can either drop the rows with missing values (since the percentage is small) or impute them using appropriate strategies like median for numerical columns and mode for categorical columns.

Handle Missing Value

High Percentage of Missing Values (> 40%):

- 'Lead Quality', 'Asymmetrique Profile Score', 'Asymmetrique Activity Score', 'Asymmetrique Profile Index', and 'Asymmetrique Activity Index' have a significant amount of missing data. Since imputing such a large number of missing values might introduce bias, it might be best to drop these columns.

Medium Percentage of Missing Values (15%-40%):

- For 'Tags', 'Lead Profile', 'What matters most to you in choosing a course', 'What is your current occupation', 'Country', 'How did you hear about X Education', 'Specialization', and 'City', we can impute the missing values with a placeholder value, e.g., 'Unknown' or 'Not Provided'. This is especially true for categorical columns.

Low Percentage of Missing Values (< 15%):

- For 'Total Visits', 'Page Views Per Visit', 'Last Activity', and 'Lead Source', we can either drop the rows with missing values (since the percentage is small) or impute them using appropriate strategies like median for numerical columns and mode for categorical columns.

Cleaning Data

To summarize the data cleaning process:

- I handled missing values based on their proportion in the columns.
- I replaced the 'Select' level with 'Not Provided' in categorical columns.
- I checked for and ensured there are no duplicate rows.

Feature engineering

To summarize Feature engineering process:

- Feature Selection: Based on domain knowledge, we handpicked relevant features from the original dataset that were most likely to influence the lead conversion. This reduced the number of features and helped focus on the most informative ones.
- Handling Missing Values: We identified and managed missing values in the dataset. For numerical features like 'TotalVisits' and 'Page Views Per Visit', the median was used for imputation, while for categorical features, missing values were often treated as a separate category or were imputed with the most frequent value.
- Encoding Categorical Variables: We transformed categorical variables into a format that could be provided to machine learning algorithms to better understand. This was done using one-hot encoding, which creates binary columns for each category and indicates the presence of the categories with 1 or 0.
- Scaling the Features: Feature scaling was performed to standardize the range of independent variables or features of the data. This ensures that each feature contributes equally to the model's performance.
- Feature Importance: After training models like Random Forest and GBM, we gained insights into the importance of each feature in predicting the target variable. This can be further used to refine the feature set if necessary.

Modeling

To summarize modeling process:

- Logistic Regression as Baseline: We began with a logistic regression model as our baseline. This model provided an initial understanding of the data's behavior and set a performance benchmark for subsequent models.
- Random Forest Model: An ensemble learning method, Random Forest, was employed to capture complex relationships in the data. This model typically offers higher accuracy by combining multiple decision trees.
- Gradient Boosting Machine (GBM): We explored the GBM model, another ensemble method known for its effectiveness in predictive modeling. It builds trees sequentially, where each tree corrects the errors of its predecessor.
- Hyperparameter Tuning: To extract the maximum performance from the GBM model, we performed hyperparameter tuning using a grid search method. This helped in finding the best combination of parameters that optimize the model's performance on our dataset.
- Performance Evaluation: For each model, we calculated key metrics like accuracy, ROC-AUC score, precision, recall, and F1-score on the test set to evaluate and compare their performances. This ensured that we selected the model that best meets the business objective.

Summary

To summarize entire process:

- Significant Features Identified: Through feature selection and domain knowledge, we identified key features that significantly influence lead conversion. These features can be prioritized in future marketing strategies.
- Model Performance: Among the models explored, the Gradient Boosting Machine (GBM) demonstrated the best performance. The tuned GBM achieved an accuracy of approximately 84.56% and an ROC-AUC score of approximately 83.26%.
- Lead Scoring: We calculated lead scores based on the predicted conversion probabilities. These scores, ranging from 0 to 100, allow the sales team to prioritize leads and tailor their approach accordingly.
- Potential for Improvement: While the current models provide robust predictions, there's always potential for further improvement. Advanced feature engineering, exploration of more algorithms, and deeper hyperparameter tuning can enhance performance.
- Business Impact: By accurately predicting the conversion probability of leads, the sales team can focus their efforts on the most promising leads, potentially increasing the conversion rate and maximizing return on investment in marketing efforts.

Thank you!