Lead Score Case Study

Purnima NS

Priya Raajendran

Rajesh Iyer

Problem Statement

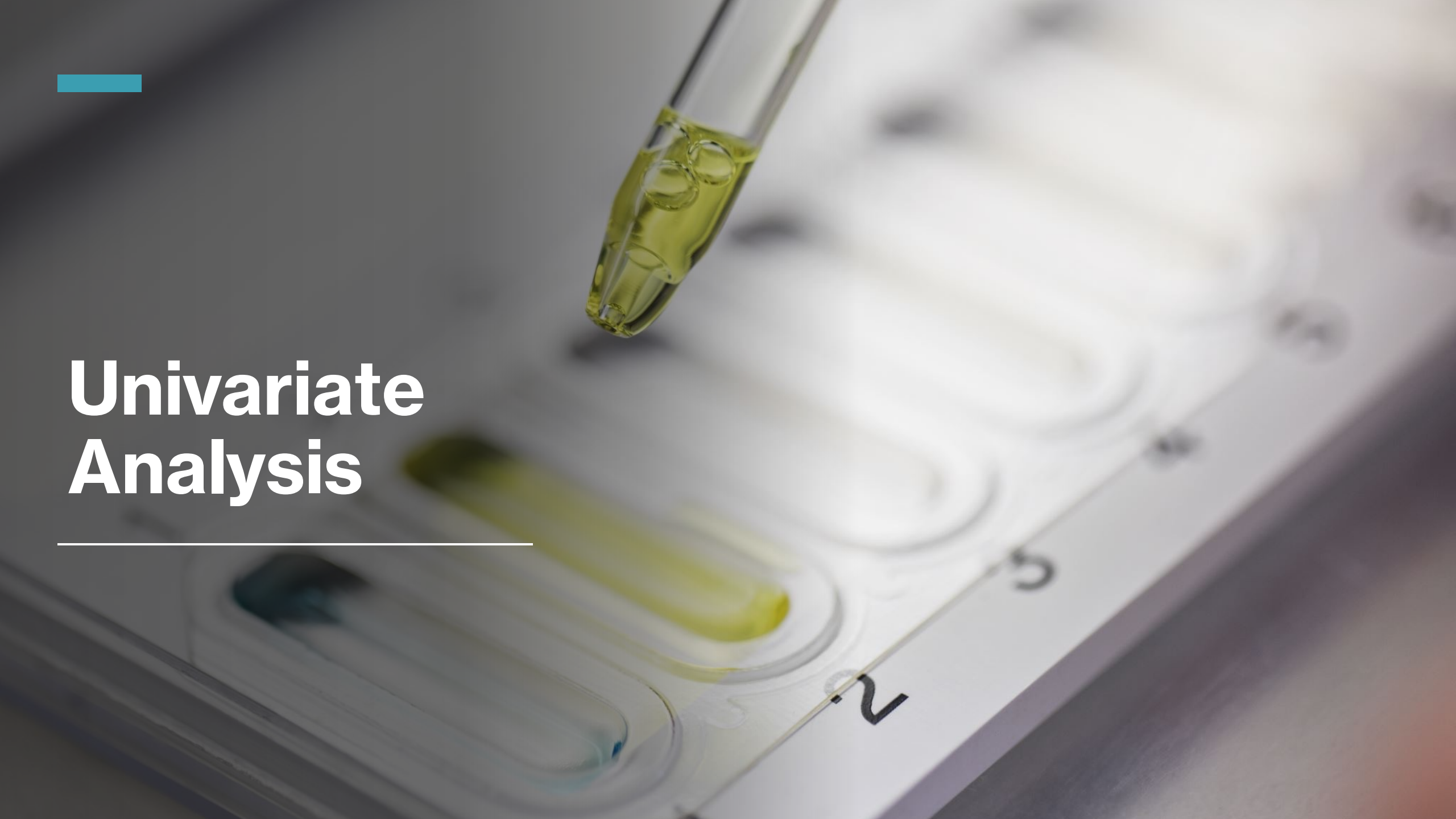
- X Education seeks to improve its low lead conversion rate by identifying 'Hot Leads' – those most likely to convert into paying customers – from a large pool of potential leads. The goal is to build a logistic regression model that assigns a lead score to prioritize these high-potential leads, aiming to increase the conversion rate from 30% to a target of 80%. The dataset provided includes various attributes that will be analyzed to predict lead conversion likelihood.

Process

- The EDA process has the following phases:
  - Reading the data
  - Data Cleaning
  - Exploratory Data Analysis
  - Train-Test Set Split
  - Getting Dummy Variables
  - Applying Logistic Regression Model
  - Verification and Fine-Tuning using VIF
  - Confusion Matrix and ROC Analysis
  - Verifying the Test Set with the Training Model

Data Cleaning

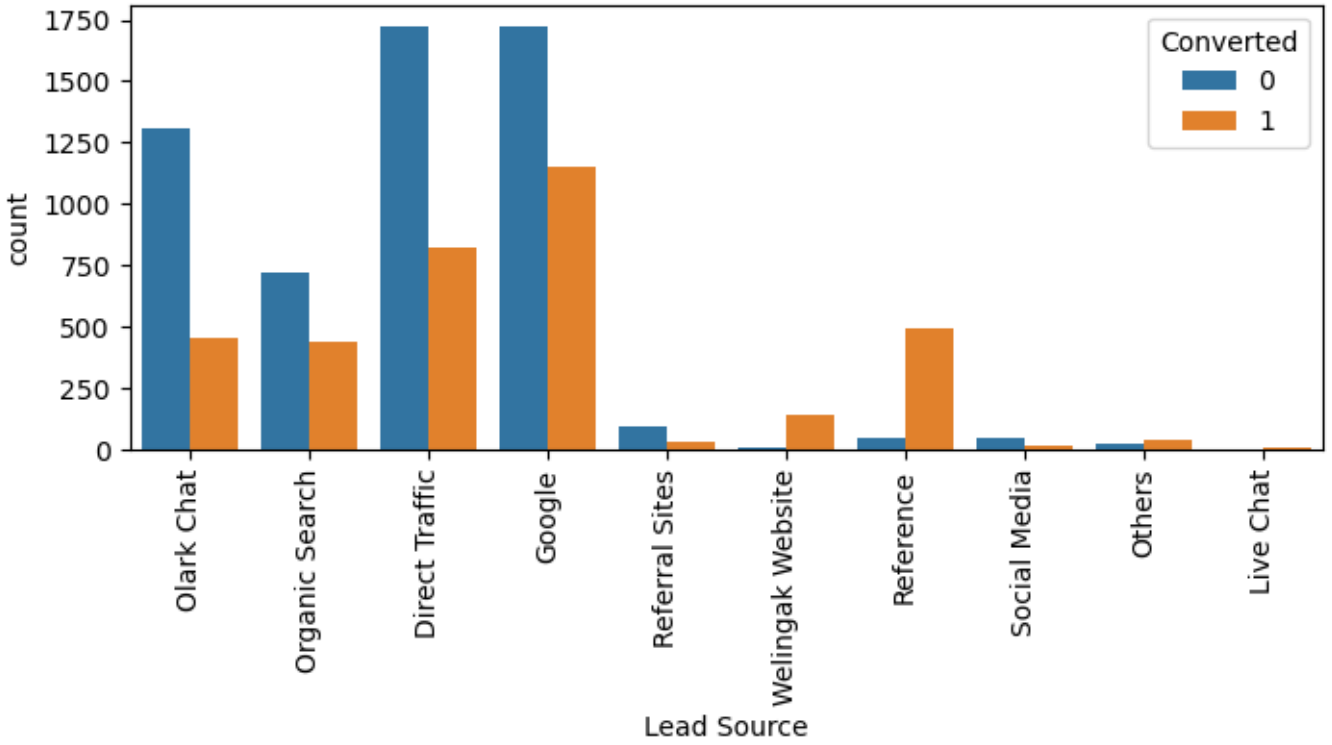
- Following steps were followed:
  - Replace values having “Select” to null and then perform analysis on respective columns
  - Drop all 40% missing value columns
  - Imputed City to ”Mumbai” which has the highest mod compared to all other values.
  - Replaced Current occupation with “Unemployeed” in case the value is null.
  - Dropped Country Column and reduced the labels for the Lead Source.
  - Replaced null and multiple Last Activities and Last Notable activities to Others which are not much relevant or can be clubbed into one.



Univariate Analysis

Analysis of Converted by Lead Source

- The bar chart displays the count of converted and non-converted leads across multiple lead sources.
- **Olark Chat** and **Organic Search** are the most prominent sources in terms of lead generation.
- **Direct Traffic** and **Google** follow closely, with a substantial number of both converted and non-converted leads.
- **Referral Sites** and **Welingak Website** have a comparatively lower number of leads.
- The remaining sources, including **Reference, Social Media, Others,** and **Live Chat**, have minimal lead counts.



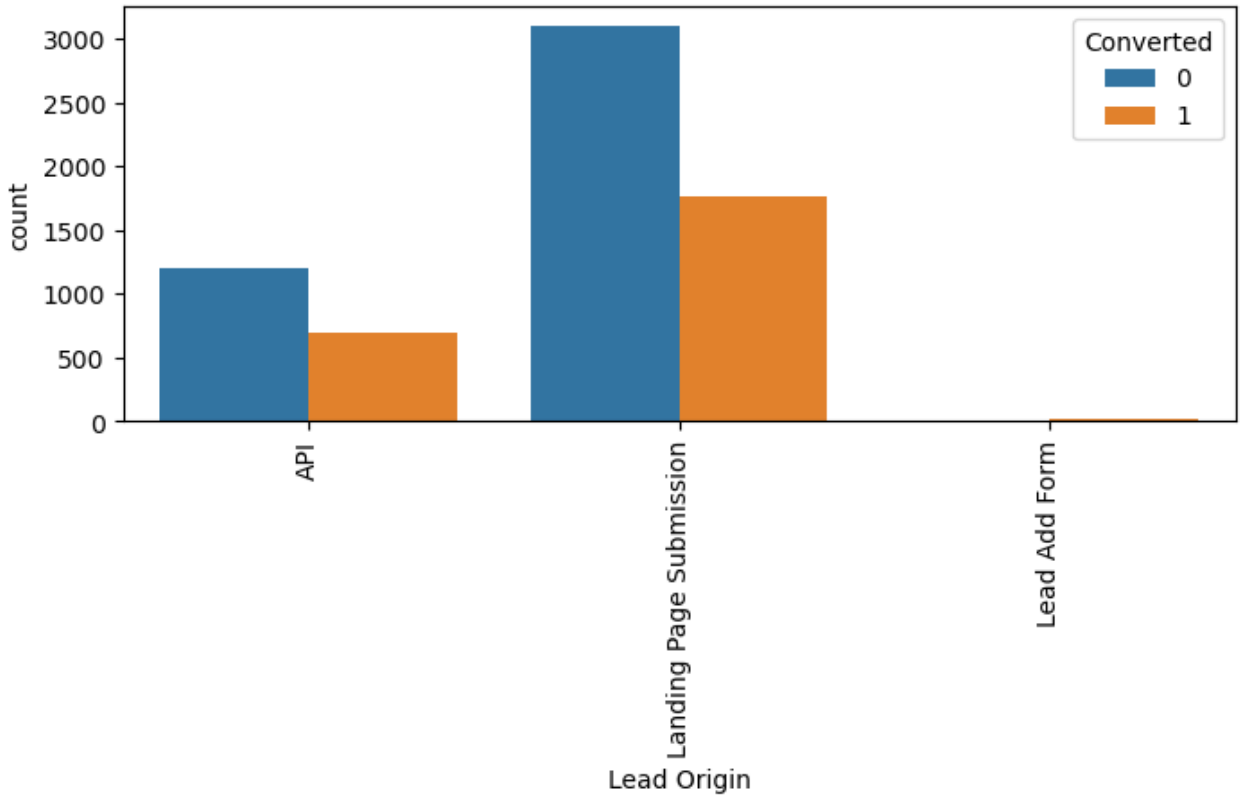
Analysis of Converted by Lead Origin

Key observations:

- **Landing Page Submission** is the lead origin with the highest number of both converted and non-converted leads.
- **API** has a significant number of converted leads, but also a considerable number of non-converted leads.
- **Lead Add Form** has a very small number of leads, with almost all of them being converted.

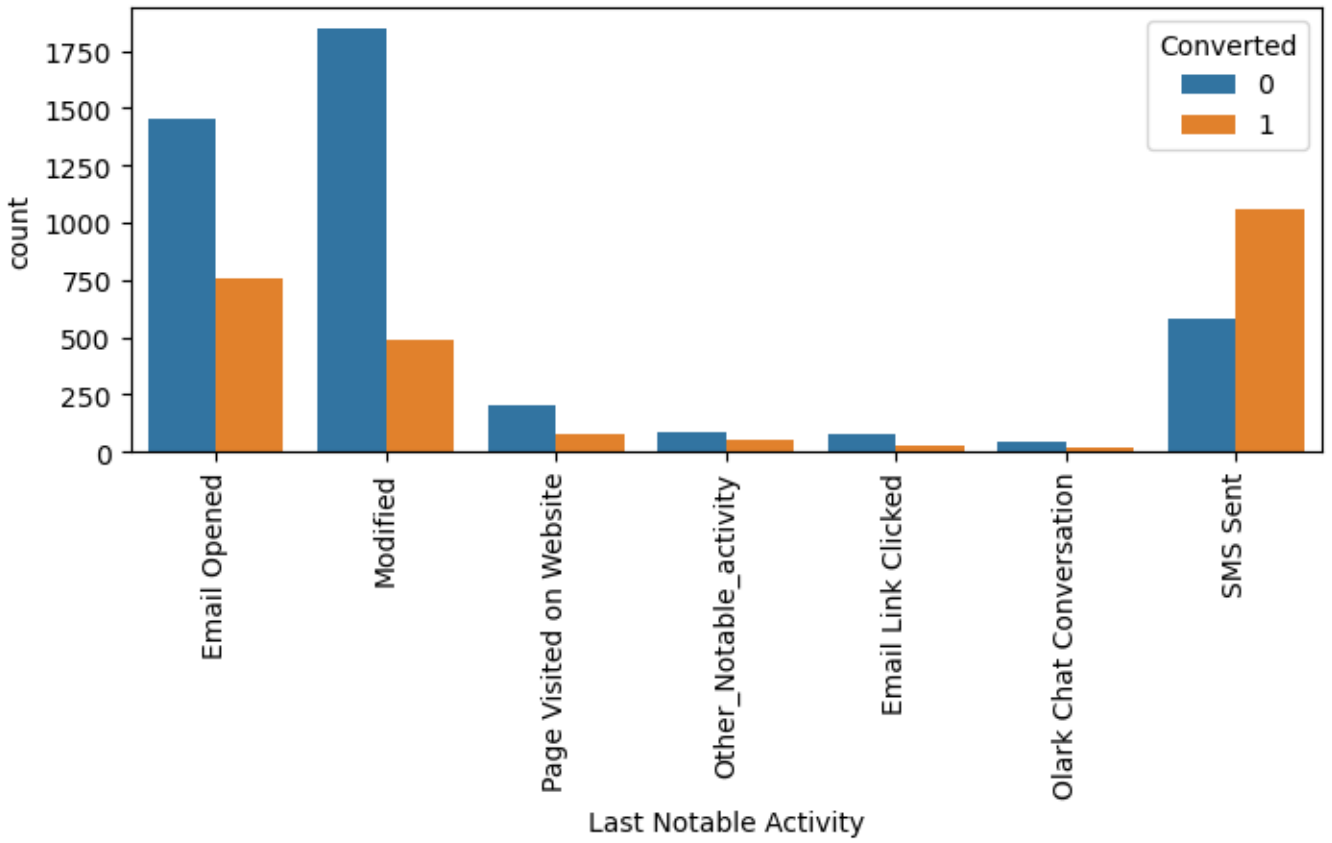
Overall:

- This chart provides insights into which lead origins are generating the most leads and their conversion performance. It can be helpful for optimizing marketing efforts and allocating resources accordingly.

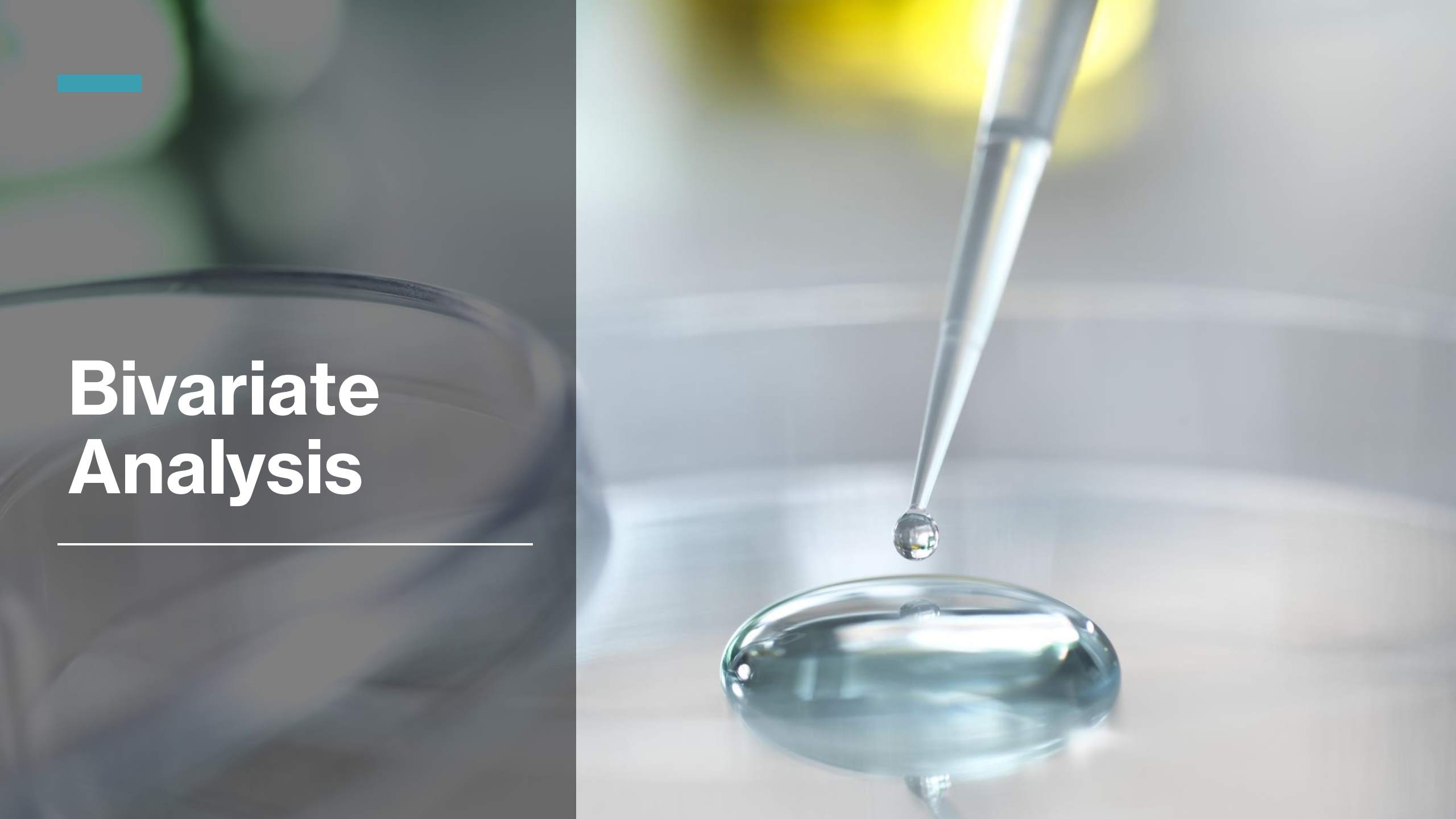


Analysis of Converted by Last Notable Activity

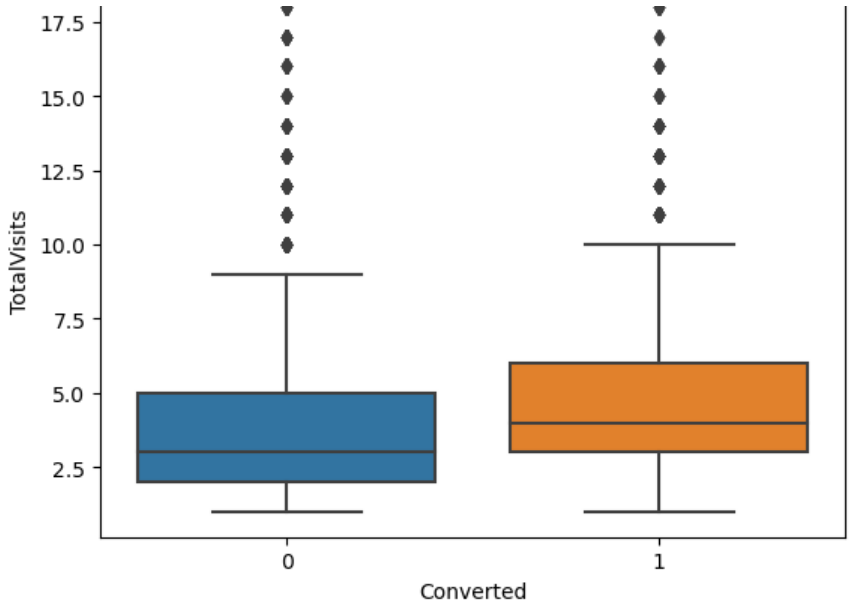
- The bar chart illustrates the number of leads converted and not converted (labeled as 0 and 1, respectively) across various last notable activities.
- Email Opened and Modified are the lead sources with the highest overall number of leads.
- Page Visited on Website and Email Link Clicked have a moderate number of both converted and non-converted leads.
- The remaining sources, including Other Notable Activity, Olark Chat Conversation, and SMS Sent, have fewer leads.



Bivariate Analysis

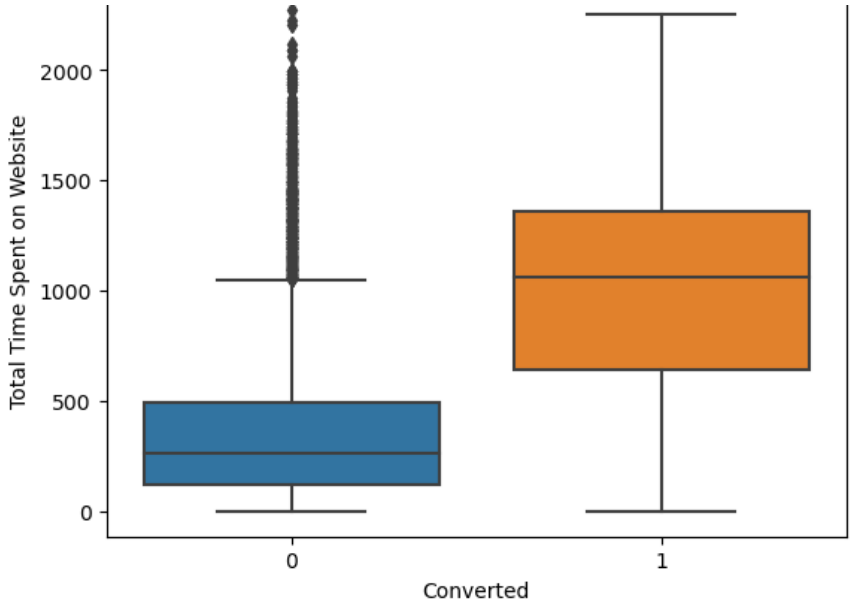


Analysis of Total Visits vs Converted



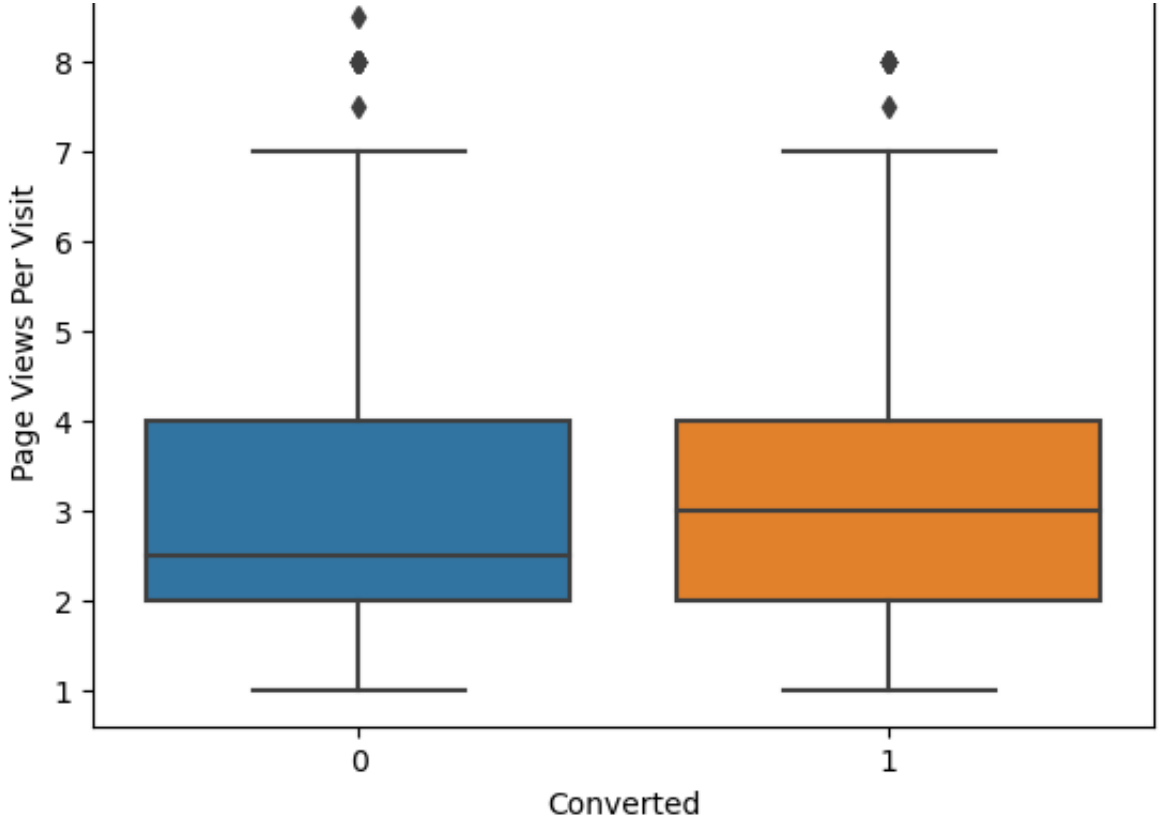
1. The box plot visualizes the distribution of total visits for converted and non-converted leads.
2. The median number of total visits for converted leads is higher than that of non-converted leads.
3. Both groups exhibit a wide range of total visit values, with several outliers present in both categories.
4. The distribution of total visits for both converted and non-converted leads appears to be right-skewed, indicating that a majority of leads have a lower number of total visits.

Analysis of Total Time Spent on the Website vs Converted

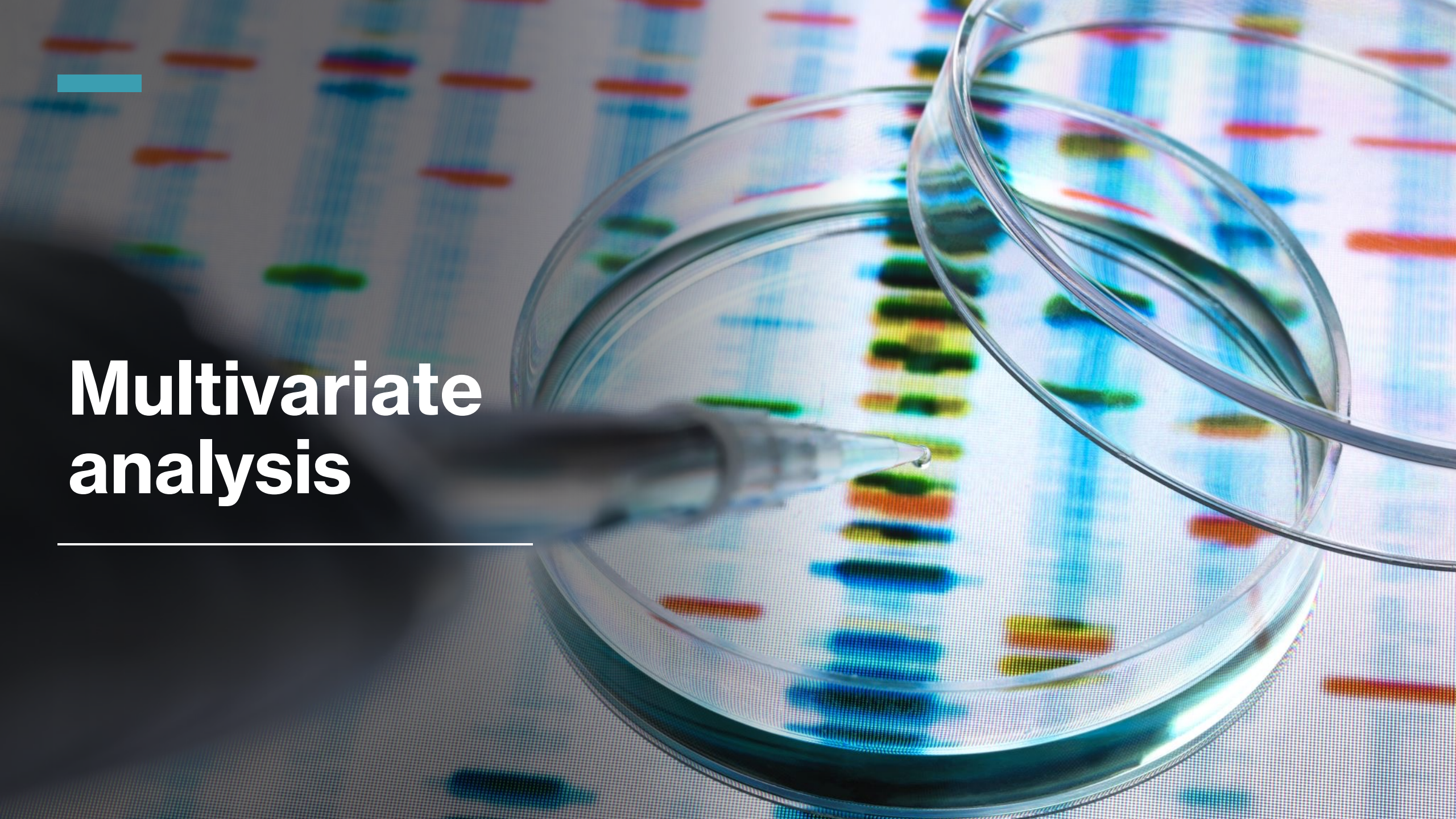


1. The box plot visualizes the distribution of total time spent on the website for converted and non-converted leads.
2. The median time spent on the website for converted leads is significantly higher than that of non-converted leads.
3. Both groups exhibit a wide range of time spent values, with several outliers present in both categories.
4. The distribution of time spent on the website for both converted and non-converted leads appears to be right-skewed, indicating that a majority of leads spend a lower amount of time on the website.

Analysis of Page Views Per Visit vs Converted



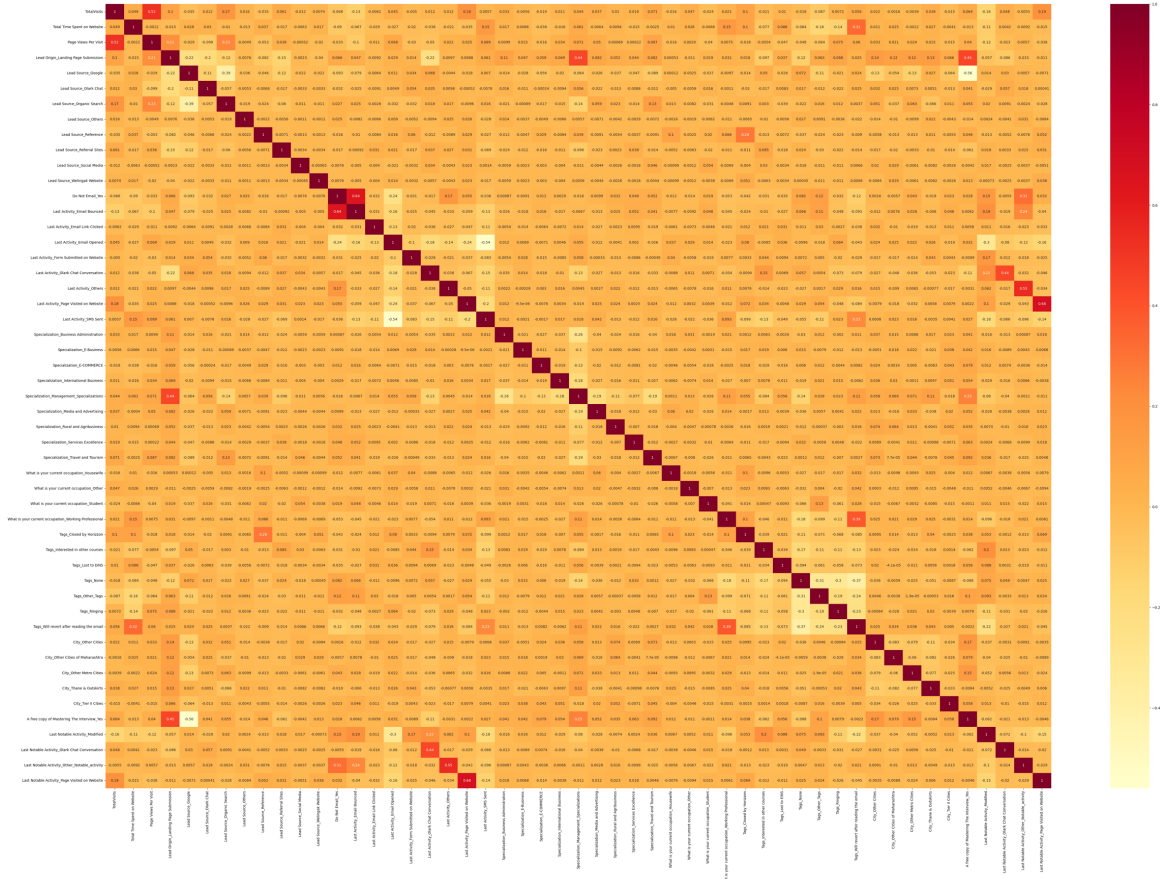
1. The box plot visualizes the distribution of total time spent on the website for converted and non-converted leads.
2. The median time spent on the website for converted leads is significantly higher than that of non-converted leads.
3. Both groups exhibit a wide range of time spent values, with several outliers present in both categories.
4. The distribution of time spent on the website for both converted and non-converted leads appears to be right-skewed, indicating that a majority of leads spend a lower amount of time on the website.



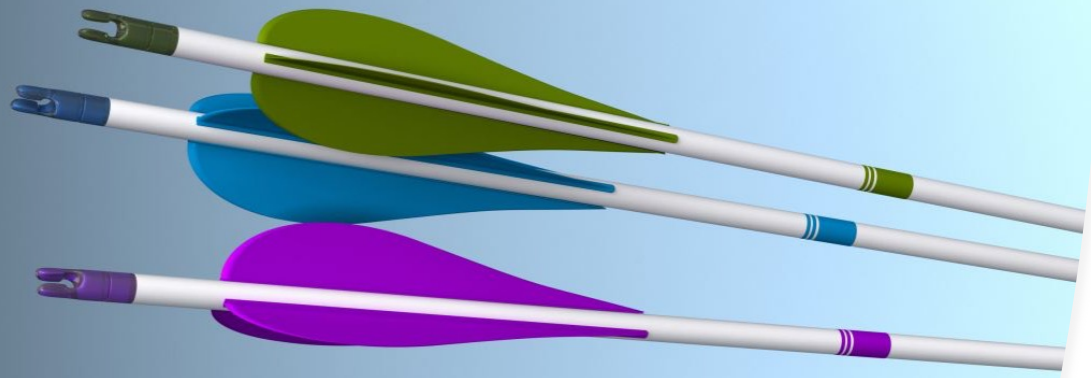
Multivariate analysis

Analysis of Heat Maps for all the Numerical Values in the dataset

1. **Lead Origin and Conversion:** It's highly likely that certain lead origins (e.g., Google, Organic Search) have higher conversion rates compared to others (e.g., Olark Chat, Referral Sites). This could indicate a positive correlation between specific lead sources and conversions.
2. **Last Activity and Conversion:** The last recorded activity before conversion might influence the outcome. For example, a positive correlation between "Email Opened" or "Form Submitted" and conversion could be expected, while a negative correlation with "Olark Chat Conversation" might suggest lower conversion rates from this channel.
3. **Total Visits and Conversion:** Generally, a higher number of total visits is associated with increased conversion probability (positive correlation). However, there might be exceptions (e.g., excessive visits leading to bounce rates).
4. **Total Time Spent on Website and Conversion:** Similar to total visits, more time spent on the website often correlates positively with conversions. Yet, there could be scenarios where longer visit durations indicate indecision or disinterest.
5. **Tags and Conversion:** Certain tags (e.g., "Interested in other courses") might negatively correlate with conversion, while tags like "Will revert after reading the email" could indicate a positive potential.



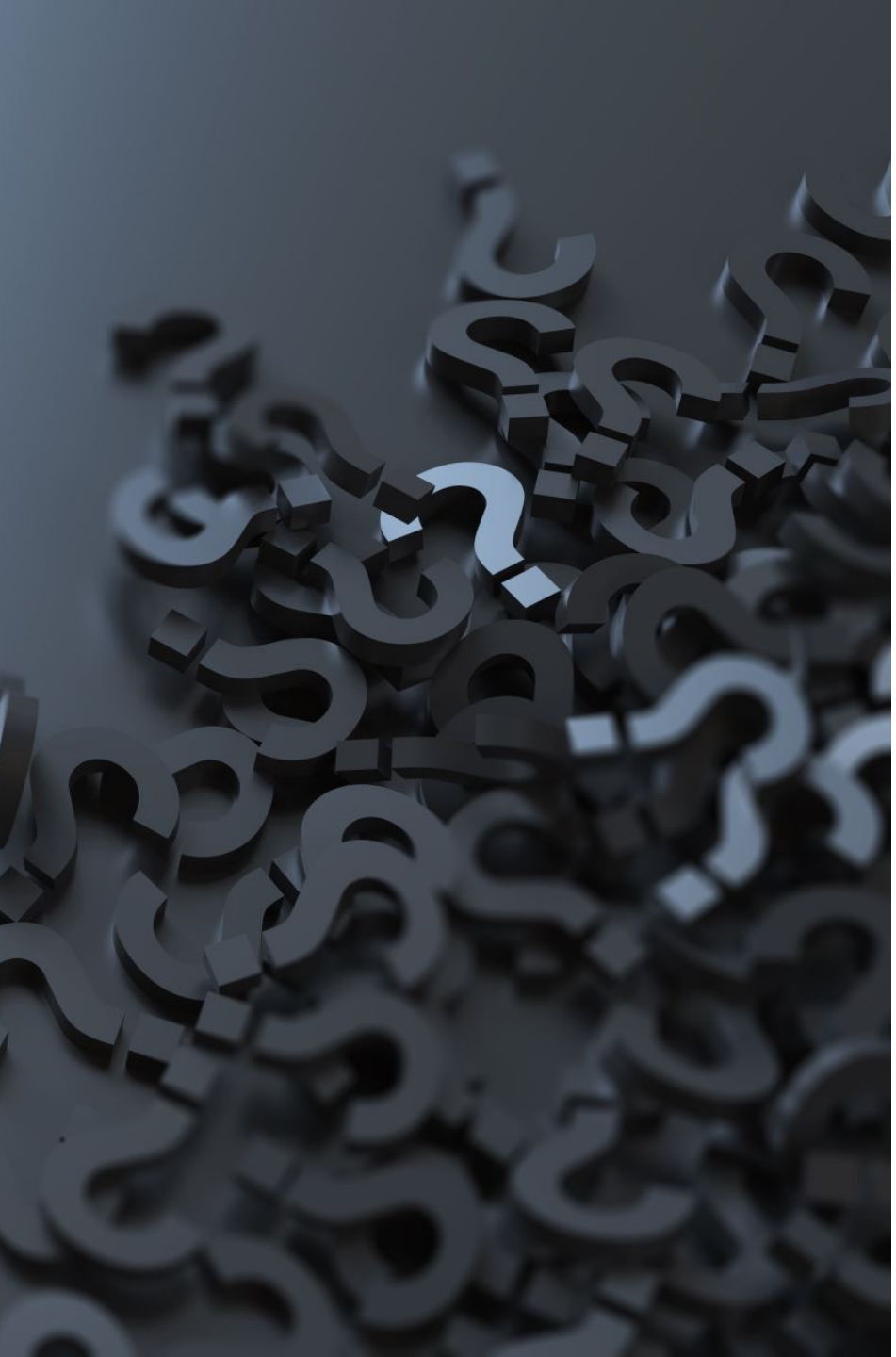
Define the target variables



The target variable: Converted

- The target variable **Converted** indicates whether a lead was successfully turned into a paying customer, with a value of 1 representing conversion and 0 representing non-conversion.
- In the provided dataset, **Converted** serves as the outcome that the model will predict, helping X Education identify the leads most likely to result in sales.
- The goal is to maximize the **Converted** rate by focusing efforts on leads with the highest potential, as determined by the model's lead scoring system.

Create dummy variables for the independent variables



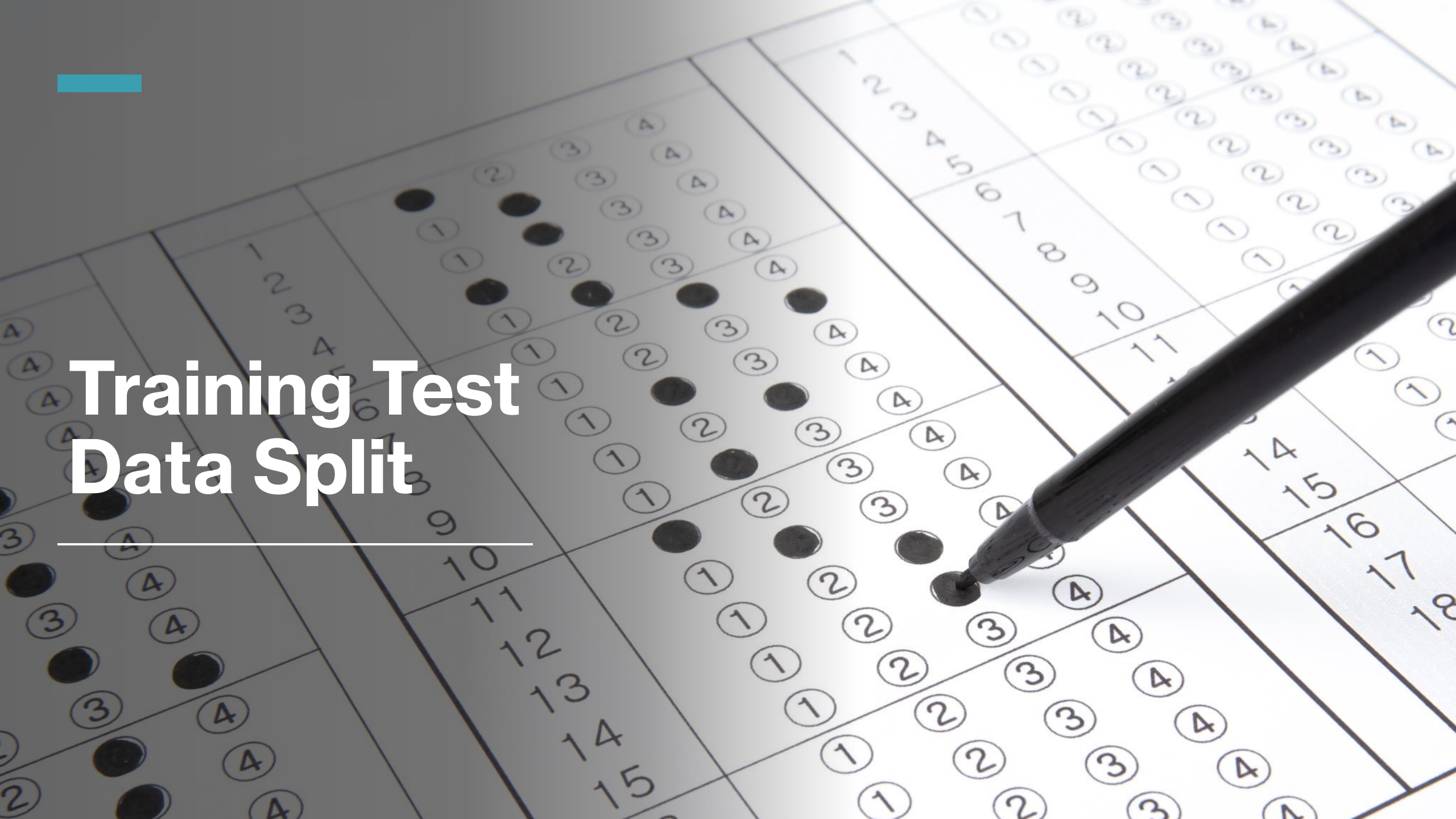
The dummy variables

Objective: To prepare the dataset for modelling, categorical variables are converted into dummy variables, enabling effective use in machine learning models.

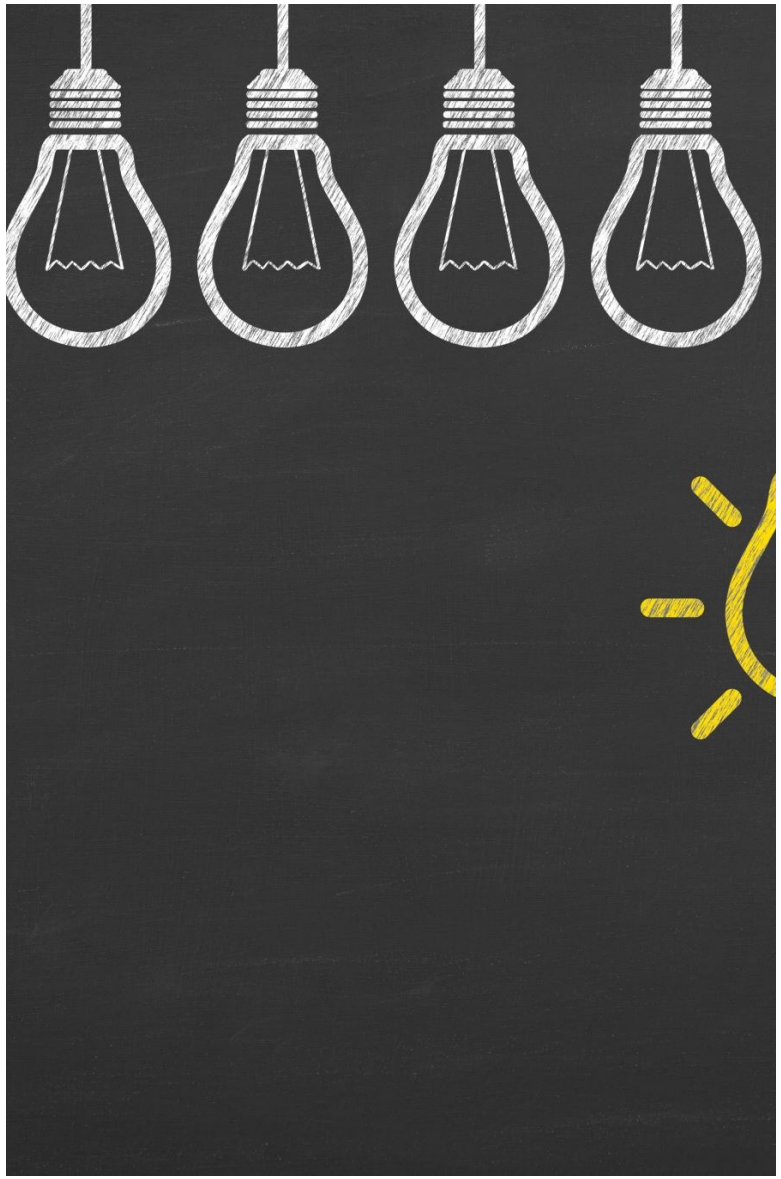
Key Categorical Columns:

1. **Lead Source & Origin:** Captures where the leads come from (e.g., Google, Direct Traffic, API). Creating dummies will help identify which channels are most effective in driving conversions.
2. **Last Interaction & Notable Activity:** Includes recent actions like Email Opened or SMS Sent. Dummy variables will highlight which activities most strongly influence conversion likelihood.
3. **Demographics:** Covers Specialization, Occupation, and City, providing insights into how personal and geographic factors impact lead conversion. Dummies here will reveal the most promising customer segments.
4. **Engagement Preferences:** Tracks whether leads have opted out of emails or requested a free copy of resources. These binary categories will directly translate into dummy variables, aiding in understanding lead engagement.
5. **Lead Status & Tags:** Labels like Interested in other courses or Will revert after reading the email provide context on lead readiness. Converting these into dummies will help in prioritizing sales efforts based on lead status.

Conclusion: By creating dummy variables for these key categories, we can enhance our model's ability to predict and prioritize high-potential leads, ultimately improving X Education’s conversion rate.



Training Test Data Split



Training-Test Data Split

Purpose:

- To assess model performance and ensure generalization to unseen data.

Technique:

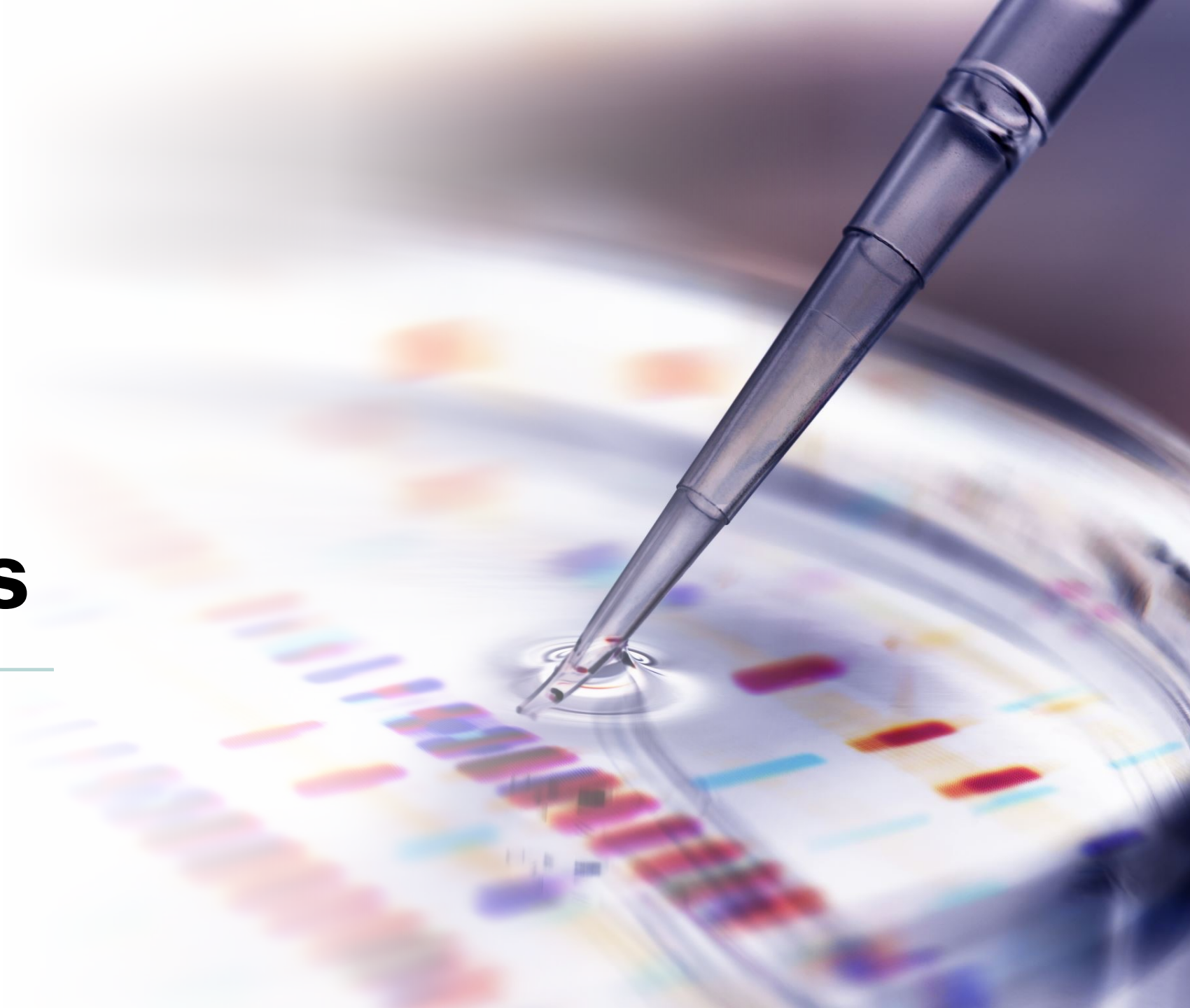
Training Set:

- **Feature Selection**: Applied Recursive Feature Elimination (RFE) with Logistic Regression to select the top 15 features.
- **Model Training**: Used selected features to build a Generalized Linear Model (GLM) with a binomial family for prediction.

Test Set:

- **Evaluation**: Reserved for testing model performance on new data after training.

Fine Tuning and Analysis



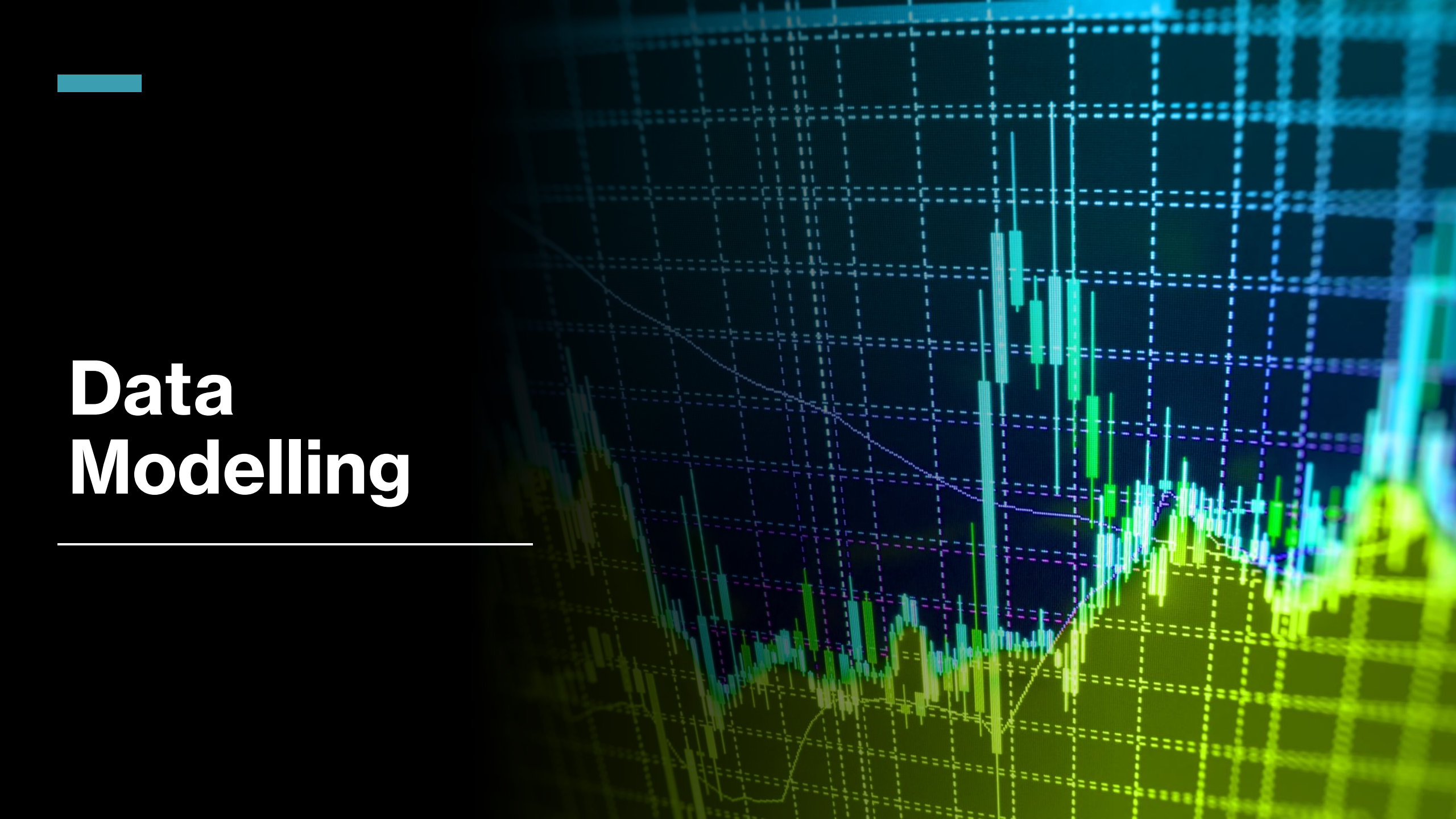
Fine Tuning and Analysis

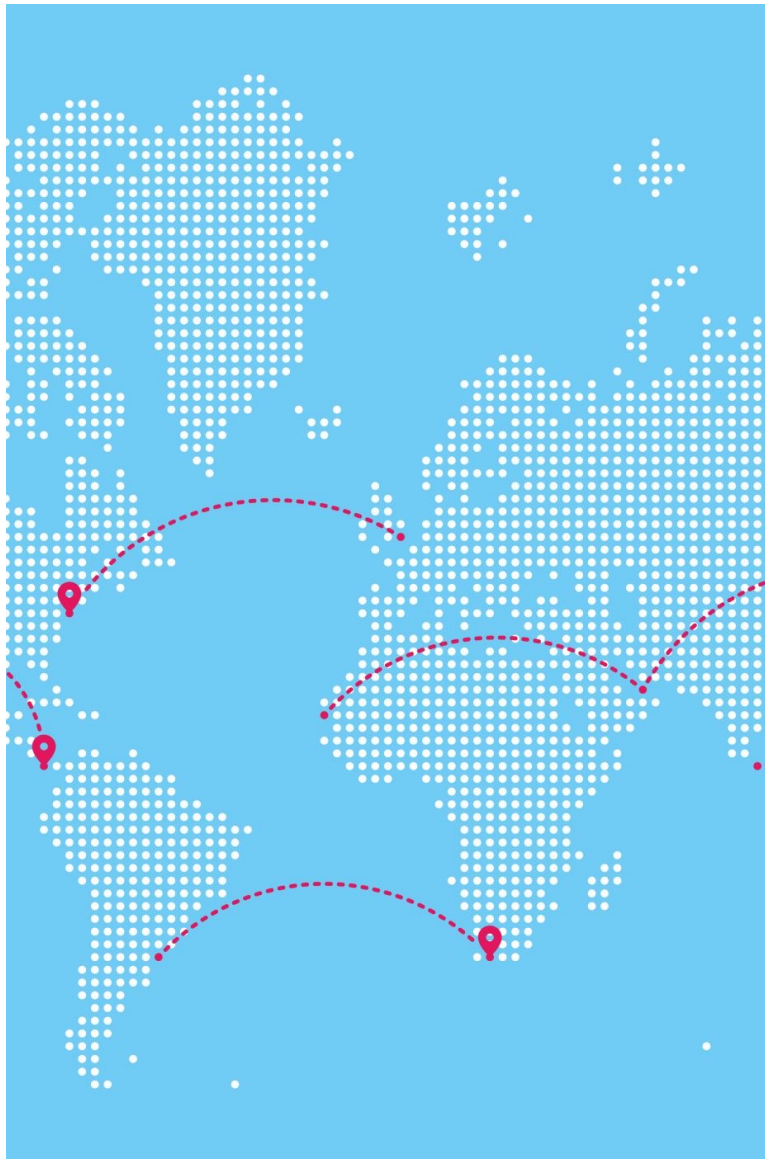
1. **Model Fine-Tuning:**
   - **Objective:** To optimize the model's performance by adjusting hyperparameters.
   - **Hyperparameters Considered:**
     - Max Depth
     - Min Samples Split
     - Min Samples Leaf
   - **Method:** Grid Search/Random Search to find the best combination of hyperparameters.
2. **VIF (Variance Inflation Factor) Analysis:**
   - **Objective:** To check for multicollinearity among the independent variables.
   - **Key Insights:**
     - A VIF value below 5 indicates no significant multicollinearity.
     - Variables with high VIF were removed/adjusted to improve model stability.

Fine Tuning and Analysis

1. **Confusion Matrix Analysis:**
   - **Objective:** To evaluate the model's classification performance.
   - **Metrics Derived:**
     - True Positives (TP)
     - False Positives (FP)
     - True Negatives (TN)
     - False Negatives (FN)
   - **Interpretation:** Understanding the trade-off between accuracy and error rates.

Data Modelling





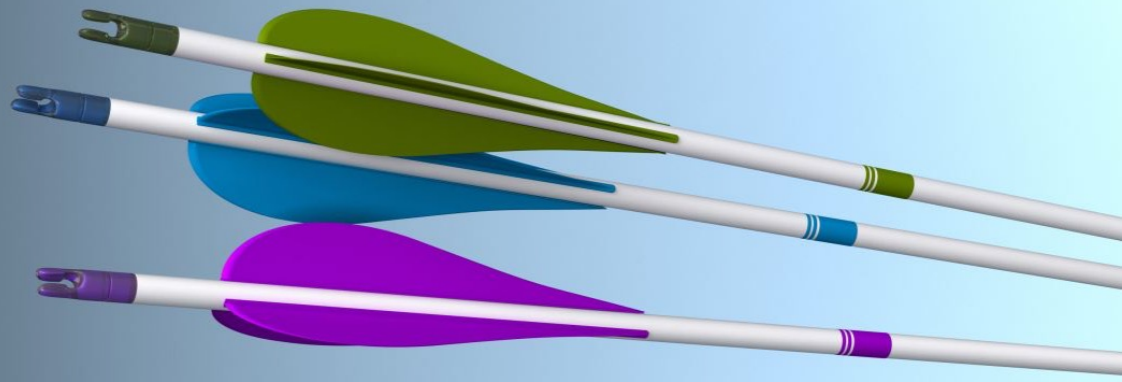
Data Modelling

1. **Model Overview:**
   - **Model Type:** Logistic Regression
   - **Objective:** Predicting the likelihood of conversion based on lead features.
2. **Feature Selection:**
   - **Method:** Recursive Feature Elimination (RFE)
3. **Selected Features:**
   - Total Time Spent on Website
   - Lead Source_Others
   - Do Not Email
   - Last Activity_SMS Sent
   - Specialization
   - What is your current occupation_Working Professional
   - Tags_Closed by Horizon
   - Tags_Interested in other courses
   - Tags_Lost to EINS
   - Tags_None
   - Tags_Other_Tags
   - Tags_Ringing
   - Tags_Will revert after reading the email
   - Last Notable Activity_Modified
   - Last Notable Activity_Olark Chat Conversation

Accuracy, Precision and Other Analysis

Accuracy and Confusion Matrix

1. **Overall Accuracy:**
    1. Training Accuracy: 93.05%
    2. Test Accuracy: 91.87%
2. **Confusion Matrix (Test Data):**
    1. True Positives (TP): 657
    2. True Negatives (TN): 1173
    3. False Positives (FP): 85
    4. False Negatives (FN): 77
3. **Performance Metrics Derived from Confusion Matrix:**
    1. Precision: 0.87
    2. Recall (Sensitivity): 0.90
    3. Specificity: 0.93

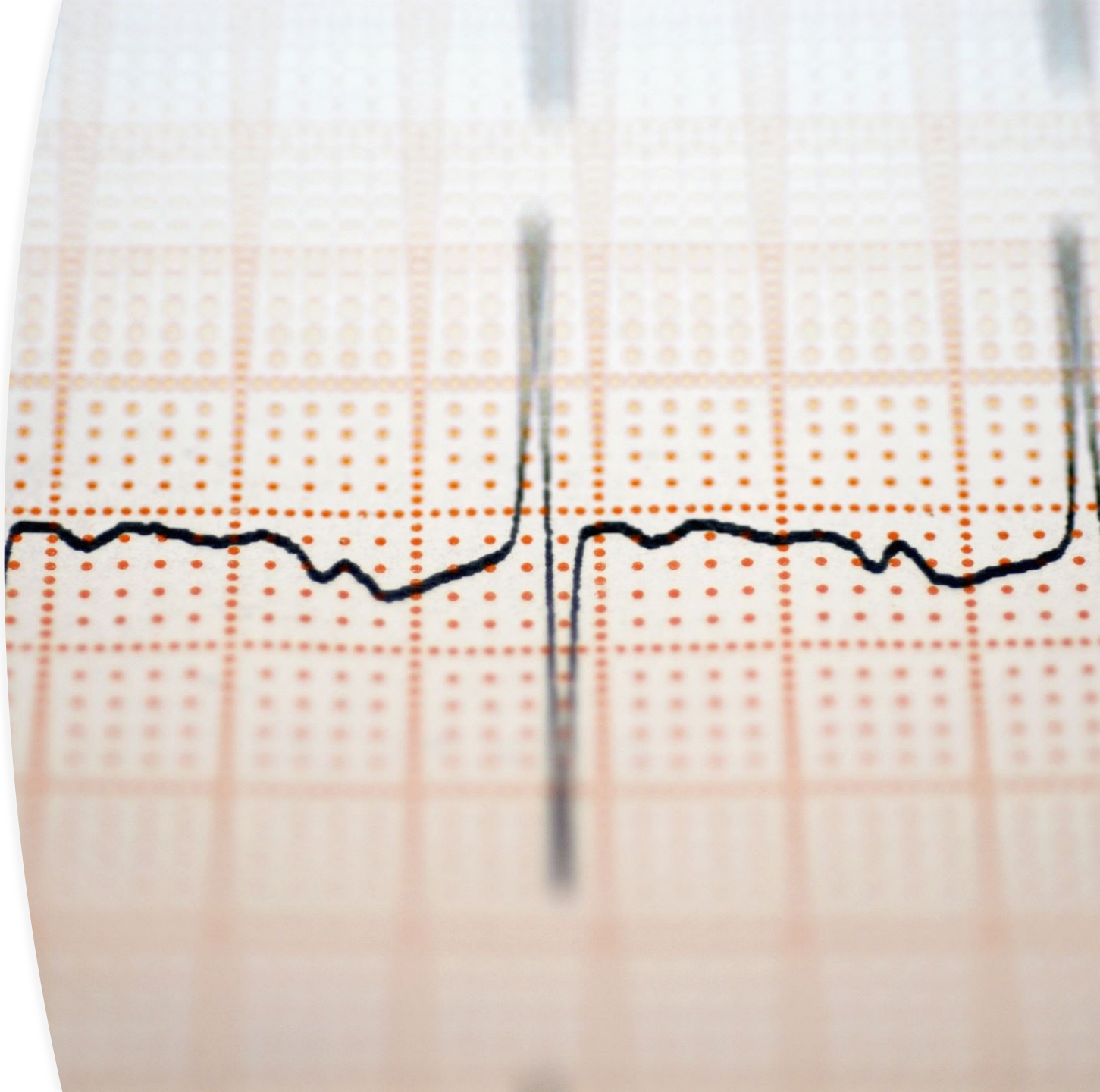


Precision, Recall and F1-Score

- **Precision:**
  - **Definition:** Proportion of positive identifications that were actually correct.
  - **Value:** 0.87
- **Recall (Sensitivity):**
  - **Definition:** Proportion of actual positives that were correctly identified.
  - **Value:** 0.90
- **F1-Score:**
  - **Definition:** Harmonic mean of Precision and Recall.
  - **Value:** 0.88
- **Calculation Summary:**
  - **Precision Formula:** TP / (TP + FP)
  - **Recall Formula:** TP / (TP + FN)

Precision and Recall Analysis

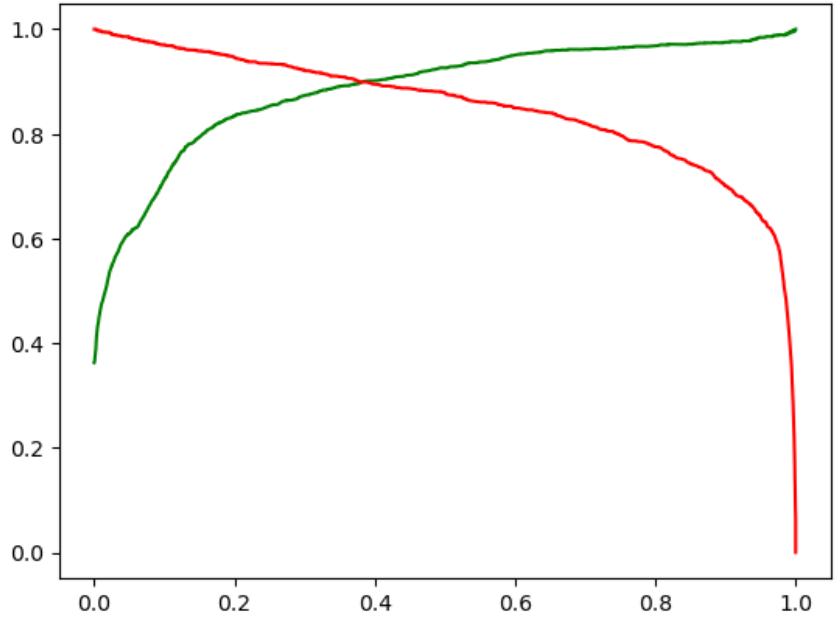
1. **Cutoff Analysis:**
   - **Optimal Cutoff Probability:** 0.3
   - **At Cutoff of 0.3:**
     - **Precision:** 0.87
     - **Recall:** 0.90
2. **Precision-Recall Curve:**
   - **Purpose:** Evaluate the trade-off between Precision and Recall.
   - **Precision and Recall Plot:** (Include the Precision-Recall curve plot here)
3. **Interpretation:**
   - **Precision:** Indicates the proportion of positive predictions that are actually correct.
   - **Recall:** Indicates the proportion of actual positives that are correctly identified.

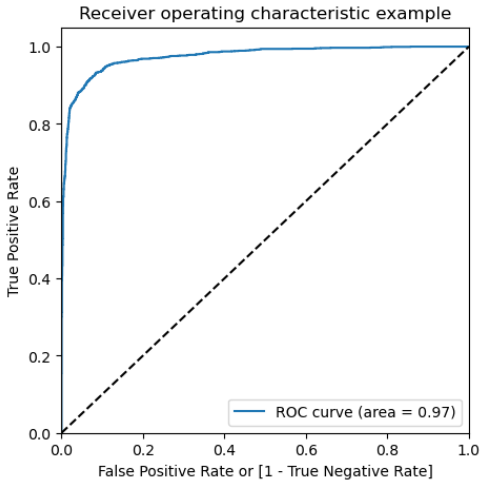


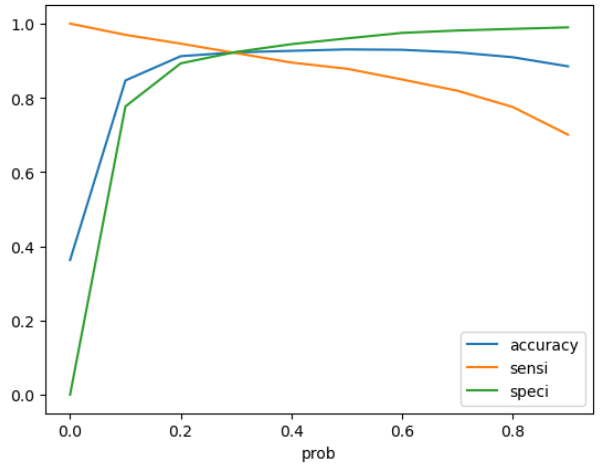
Summary of Model Performance

1. **Overall Accuracy:**
   - Training Accuracy: 93.05%
   - Test Accuracy: 91.87%
2. **Confusion Matrix Highlights:**
   - True Positives (TP): 657
   - True Negatives (TN): 1173
   - False Positives (FP): 85
   - False Negatives (FN): 77
3. **Key Metrics:**
   - Precision: 0.87
   - Recall: 0.90
   - F1-Score: 0.88
   - AUC: 0.97
4. **Optimal Cutoff:**
   - Probability: 0.3

Analysis: Charts







Recommendations

- **Recommendations:**
  - The company can focus on the below categories in order to get the highest leads,
  - Tags Closed by Horizon
  - Tags Lost to EINS
  - Lead Source Welingak Website
  - Tags Will revert after reading the email
  - Lead Source Others
  - Lead Source Reference
  - What is your current occupation Working Professional
  - Total Time Spent on Website
- **The company need not focus on the below categories:**
  - Lead Origin Landing Page Submission
  - Tags Interested in other courses
  - Last Notable Activity Olark Chat Conversation
  - Last Notable Activity Modified
  - Tags Ringing
  - Tags Other tags

Conclusion

- **Key Findings**:
  1. **Feature Selection**: Implemented RFE to identify the top 15 features contributing to lead conversion.
  2. **Model Performance**: The logistic regression model demonstrated high overall accuracy (93.05%) and well-calibrated probabilities.
  3. **Multicollinearity Check**: VIF values indicated no significant multicollinearity issues among selected features.
  4. **Confusion Matrix Analysis**: The model showed strong performance with a sensitivity of 92.06% and specificity of 92.37% at the chosen cutoff of 0.3.
  5. **ROC and AUC**: The ROC curve with an AUC of 0.97 confirmed excellent model discriminative ability.
  6. **Precision and Recall**: At the optimal cutoff, precision and recall were effectively balanced, ensuring reliable predictions.
- **Next Steps**:
  - **Re-evaluate Objectives:** Review and reassess objectives to align with the enhanced capabilities of the model.
  - **Exploit New Opportunities:** Identify and pursue additional opportunities or use cases that could benefit from the model's early success.
  - **Innovative Applications:** Explore innovative applications or extensions of the model to maximize its impact and value.


Thank You