Summary of Findings

- **Data Cleaning:**

    a) The dataset was cleaned by handling missing values in columns such as 'Lead Quality', 'City', 'Specialization', and 'Tags'. 'Select' values were replaced with appropriate values or 'Not Sure' where applicable.

- **Data Exploration:**

a) The target variable 'Converted' was analyzed, showing that approximately 37.5% of leads were converted.
b) Univariate and bivariate analysis was performed on various features to understand their distribution and relationships with the target variable.
c) Features like 'Lead Source', 'Lead Add Form', 'Last Activity', 'Specialization', 'What is your current occupation', and 'Tags' showed varying conversion rates.

- **Feature Engineering:**

a) Dummy variables were created for categorical features to convert them into numeric format.
b) Recursive Feature Elimination (RFE) technique was used to select significant features and handle multicollinearity.
c) The optimal features were identified to build the final logistic regression model.

- **Model Building:**

a) A logistic regression model was built using the selected features.
b) The model's performance was evaluated using the log-likelihood, Wald test (p-values), and VIF scores to ensure significance and non-multicollinearity of features.

- **Model Evaluation:**

a) The model achieved an accuracy of approximately 91.5% on the training data and 91.9% on the test data, indicating a good fit.
b) Sensitivity (True Positive Rate) was around 84.8% on both the training and test data, indicating the model's ability to correctly predict positive cases (leads converted).
c) Specificity (True Negative Rate) was around 95.6% on both the training and test data, indicating the model's ability to correctly predict negative cases (leads not converted).
d) The model's precision was around 90%, meaning that when it predicts a lead will convert, it is correct around 90% of the time.
e) The model's recall was around 84.8%, indicating that it identifies around 84.8% of the leads that actually convert.
f) The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score was high, indicating good discriminative power between positive and negative cases.

- **Cutoff Probability Optimization:**

a) The optimal cutoff probability of 0.3 was chosen to balance precision and recall, based on precision-recall tradeoff analysis.

- **Prediction on Test Data:**

a) The model was used to make predictions on the test dataset.
b) Leads were assigned lead scores based on their conversion probabilities, indicating their likelihood of converting.

- **Identifying Hot Leads:**

a) Leads with high conversion probabilities (lead score >= 85) were identified as potential hot leads for contacting.
b) Approximately 623 leads were identified as hot leads.

- **Important Features:**

a) The coefficients of significant features in the final model were provided, indicating their impact on lead conversion.
b) Features with higher coefficients have a more significant influence on lead conversion.

- **Recommendations:**

a) The company should focus on leads with high conversion probabilities (lead score >= 85) as these have a higher chance of conversion.
b) Leads from Google and references, as well as those who have opened emails and received SMS, have higher conversion rates. The company can prioritize these leads in their marketing and sales efforts.
c) Leads with 'Will revert after reading the email' and 'Not Sure' in the 'Tags' column have higher conversion rates, indicating they are potential targets for follow-up.
d) The city 'Mumbai' seems to have a higher conversion rate, so the company can focus more on leads from Mumbai.

- **The company should continue to monitor and refine the model periodically as the business and lead characteristics may change over time.**