In this case study, we have finally been able to solve the problem of enhancing the lower lead conversion rate by implementing a logistic regression machine learning model. Let us summarize the steps involved in the project:

1. Data Collection: Based on the problem statement, we collected the right data and imported it into our notebook. This includes various types of data such as integer, float, and object. We have checked the data dictionary thoroughly to understand what the data attributes are and what they mean, also how significant are they in our problem and what kind of imputations are needed to further enhance the usability of the data.

2. Data Cleaning: The first step here is to remove unwanted values. We have recognized that a lot of columns have the value 'Select' in them, and they are replaced by Null values to give us a true picture of the Data. Further, we have dropped columns with more than 30% of Null values and rows with less than 3% of them. Additionally, we also dropped columns that denoted ID numbers and where more than 90% of the data pertained to a single value.

3. Exploratory Data Analysis: In this step, we compared all the categorical and numerical variables with the target variable and found some insights. Outliers were also observed in the Numerical columns and were treated accordingly.

4. Model Building: Model Building: We encoded categorical variables using one hot encoding and standardized numerical features. One hot encoding was chosen for its effectiveness in converting categorical data into a format suitable for modeling. The dataset was split into training and testing sets with an 80:20 ratio to assess the model's generalizability. Feature selection was critical, and we used Recursive Feature Elimination (RFE) to select the most relevant features, enhancing the model's predictive power. Logistic Regression was chosen as the base model. To improve the model's performance, we created multiple models based on improving p-values, fine-tuning them for optimal performance. The model evaluation included the analysis of a confusion matrix, which revealed a precision of around 75%, indicating the model's ability to accurately predict positive cases. This iterative approach allowed us to achieve an accuracy of around 80%, demonstrating the effectiveness of our model in predicting lead conversions. The model's performance was further validated through rigorous testing on unseen data, confirming its reliability in real-world scenarios. Overall, our methodical approach and attention to detail have resulted in a robust predictive model that can provide valuable insights for improving lead conversion rates and optimizing resource allocation strategies.

5. Conclusion: Overall, our approach emphasized precision over recall, aiming to minimize resources wasted on low-quality leads while maximizing the identification of high-quality leads. Our most influential variables include ' Lead Origin_Lead Add Form', 'Total Time Spent on Website', 'What is your current occupation_Unemployed'.