Summary Report

The goal was to build a predictive model to help X Education improve it's lead conversion process by identifying "Hot Leads".

Understanding the problem

X Education had a typical lead conversion rate of around 30%, and they aimed to increase this rate by focusing on the most promising leads. The company wanted to assign a lead score to each lead, where higher scores indicated a higher likelihood of conversion. The objective was to predict which leads had the highest chance of converting, enabling the sales team to focus their efforts more effectively.

Data Preprocessing

The first step was to clean and preprocess the data. This included handling missing values, handling categorical variables, and scaling continuous features. Key variables such as 'Lead Origin,' 'Total Visits,' 'Total Time Spent on the Website,' 'Lead Quality,' and other relevant factors were included in the analysis. We also identified the presence of a 'Select' level in categorical variables, which was treated as a null value and imputed for further analysis.

The target variable, "Converted," indicated whether a lead was converted (1) or not (0), which was used to train the logistic regression model. Additionally, we checked for class imbalance, as the dataset contained more non-converted leads (0) than converted leads (1)

Model Building

For the modeling process, we chose logistic regression which is suitable for binary classification tasks. We used Recursive Feature Elimination (RFE) to select the most important features and the selected features were then used to train the model.

We split the dataset into a training and testing set, ensuring the model was trained on 80% of the data and evaluated on the remaining 20%. We tuned the model by adjusting the decision threshold to optimize the conversion rate. The default threshold of 0.5 has been adjusted after calculating the Optimal Cutoff Point as 0.4. The Lead Score has been calculated to target the potential leads. The sales team can adjust the threshold to target a larger pool of potential leads during aggressive lead conversion phases and reduce outreach to low-conversion leads when goals are already met.

Model Evaluation

The model's performance was evaluated using metrics such as accuracy, precision, recall, F1-score, specificity, and false positive rate. These metrics provided insights into how well the model was identifying true positives (conversions) and avoiding false positives. The results showed that the model had a balanced performance, with a precision of around 77% and recall of 81%, meaning the model was good at identifying actual conversions while maintaining a low false positive rate.

Business Insights and Recommendations:

Based on the model's output, several actionable recommendations were made:

- **Aggressive Conversion Strategy:** During phases like the intern hiring season, lower the conversion threshold to target more leads with a reasonable chance of converting.
- **Target High-Priority Leads:** Focus on leads with higher predicted scores, especially those with top-tier features, to increase conversion rates.
- **Minimize Effort During Surplus Quarters:** Increase the threshold to reduce unnecessary outreach when the company has already met its targets.