Summary Report

Lead Conversion Prediction and Strategic Analysis

This project was a comprehensive exercise in applying machine learning to solve a real-world business problem: predicting lead conversion for X Education. The assignment required me to build a logistic regression model, derive a lead scoring system, and use the model's insights to provide actionable business strategies. The entire process, from data handling to final recommendations, was critical.

The initial step was to understand and clean the provided data from Leads.csv and the data dictionary. A key challenge was the high number of missing values and the presence of "Select" as a placeholder for unknown values. I decided to drop columns with a high percentage of missing data (greater than 30%) and impute the remaining missing values with the mode or median. This pre-processing step was crucial for building a reliable model.

Following data cleaning, I prepared the data for modelling by converting binary categorical variables to 0/1 and using one-hot encoding for multi-level categorical features. This allowed the logistic regression model to effectively interpret all the variables. The dataset was then split into training and testing sets to ensure the model's performance could be evaluated on unseen data.

The logistic regression model was trained, and its performance was evaluated using standard metrics. The model performed well, achieving an accuracy of approximately 80% and a high ROC-AUC score of 0.85. The ROC-AUC score is particularly important as it measures the model's ability to discriminate between the two classes (converted vs. not converted) across all possible classification thresholds.

One of the most valuable outcomes of the logistic regression model was the ability to interpret the importance of each feature. By examining the model's coefficients, I could identify the most significant predictors of conversion. The time spent on the website and the last notable activity being an SMS sent were the most influential factors. This provided clear, data-driven insights for the sales team.

The model's predicted probabilities were then transformed into a lead score between 0 and 100. This score is a simple and intuitive metric for the sales team to prioritize their efforts. The assignment also required addressing two specific business problems: a period of aggressive sales and a period of conservative sales. I leveraged the concept of a classification threshold to provide a solution. By lowering the threshold, the company can be more aggressive, prioritizing recall and ensuring no potential lead is missed. Conversely, by raising the threshold, the company can be more conservative, prioritizing precision and minimizing wasted effort on unlikely leads.

In conclusion, this project demonstrated the power of a complete data science workflow, from raw data to actionable business recommendations. The lessons learned highlight the importance of meticulous data cleaning, the interpretability of a logistic regression model, and the ability to translate model output into tangible business strategies tailored to different objectives. The final model provides a robust framework for X Education to optimize its lead management and sales processes.