Name – Samit Kumar Das

Course – DS C65

Date – 20/08/2024

Summary Report: Lead Scoring Analysis for X Education

Overview:

The primary objective of this assignment was to develop a predictive model to determine whether a lead is likely to convert into a paying customer for X Education. The dataset provided included 36 predictor variables, and the target variable was Converted, which indicated whether a lead had been converted or not.

Approach:

The analysis began with an exploration of the dataset, focusing on understanding the structure and identifying key features that might influence lead conversion. The dataset initially contained 9,240 entries and 37 columns, including various demographic and behavioral attributes of the leads.

The first step involved **data cleaning**: handling missing values, removing unnecessary columns, and encoding categorical variables. Columns with a high percentage of missing values and those irrelevant to the analysis, such as specific lead IDs and obscure categories, were dropped. This step reduced the dataset to 4,327 entries with 12 columns.

Next, I used **Recursive Feature Elimination (RFE)** with logistic regression to select the top features contributing most to lead conversion. This process helped in reducing the dimensionality of the dataset, ensuring that only the most predictive variables were retained for model training. The final model included key variables like Lead Source, Total Time Spent on Website, Lead Quality, and Tags.

The **logistic regression model** was then built using the selected features, and its performance was evaluated using metrics like accuracy, precision, recall, and AUC-ROC. The model provided insights into which variables positively or negatively influenced the likelihood of lead conversion. For instance, leads with high Total Time Spent on Website were more likely to convert, highlighting the importance of user engagement.

Learnings:

This assignment reinforced several key data science principles:

1. **The Importance of Data Cleaning**: Effective data cleaning is crucial to ensure that the analysis is based on accurate and relevant data. Removing irrelevant columns and handling missing values allowed for a more focused and effective model.
2. **Feature Selection**: The use of RFE highlighted the importance of selecting the most predictive features to improve model performance and interpretability. This step helped in simplifying the model while retaining its predictive power.
3. **Model Interpretation**: The logistic regression model provided interpretable coefficients, offering insights into how different features impacted lead conversion. This interpretability is critical for translating data findings into actionable business strategies.

Overall, this assignment provided valuable experience in the end-to-end process of data analysis, from data preparation and feature selection to model building and result interpretation. The

learnings from this project will be instrumental in future data science tasks, particularly in the context of predictive modeling and business analytics.