**Summary Report**

**Lead Conversion Optimization for X Education**

This assignment aimed to address the challenge of improving the lead conversion rate for X Education by building a predictive model that identifies high-potential leads. The goal was to help the sales team focus their efforts on leads that are more likely to convert into paying customers, thereby increasing efficiency and improving conversion rates. The approach focused on data preprocessing, model building, evaluation, and the development of actionable strategies for lead prioritization.

**Data Preprocessing:**

The first step in the process was preparing the dataset for analysis. The dataset contained around 9,000 records with various attributes related to lead behavior, such as **Lead Source**, **Total Time Spent on Website**, **Last Activity**, and the target variable **Converted** (indicating whether the lead converted into a customer).

To ensure that the model would work optimally, the following preprocessing steps were applied:

1. **Handling Missing Data**: Many categorical variables had a 'Select' value, which was treated as a null value. Numerical variables with missing values were imputed using the median, and categorical variables were filled with the most frequent category.
2. **Feature Engineering**: Categorical variables were encoded into dummy variables (one-hot encoding), while numerical variables were standardized to ensure that all features contributed equally to the model.
3. **Data Splitting**: The data was split into a training set (80%) and a test set (20%) to evaluate the model's performance on unseen data.

**Model Building:**

The next step involved selecting a model to predict lead conversion. Given the binary nature of the problem (converted vs. not converted), **logistic regression** was chosen for its simplicity and interpretability. Logistic regression is well-suited for classification tasks and provides coefficients that indicate the relationship between input features and the likelihood of conversion.

I trained the model on the training data and evaluated its performance using metrics like accuracy, precision, recall, and ROC-AUC. The logistic regression model achieved an **accuracy of 80%**, with a **ROC-AUC score of 0.85**, demonstrating that it was effective in distinguishing between leads that were likely to convert and those that were not.

**Model Evaluation & Results:**

The logistic regression model produced lead scores for each lead, representing the probability of conversion. By analyzing the model's coefficients, the top three features that contributed most to the conversion probability were identified:

1. **Total Time Spent on Website**: Leads who spent more time on the website had a higher chance of converting.

2. **Lead Source**: Certain lead sources, such as Google and Referrals, were more strongly associated with conversions.
3. **Last Activity**: Activities such as filling out a form were linked to higher conversion rates.

The importance of categorical variables was also evaluated, with **Lead Source** and **Last Activity** emerging as the most impactful categorical features. These insights helped in formulating strategies for focusing on the most promising leads.

**Strategy Development:**

With the model in place, strategies were devised to optimize lead conversion based on different business scenarios:

1. **Aggressive Conversion Strategy (Intern Phase)**: During the intern phase, the strategy was to focus on leads predicted to have a high chance of conversion. By lowering the threshold for predicting leads as likely to convert (e.g., from 0.5 to 0.4), more leads were identified as high-potential, ensuring that interns focused on leads with the highest predicted conversion probabilities.
2. **Conservative Conversion Strategy (Quarter-End)**: When the company meets its quarterly targets early, the goal is to minimize unnecessary calls. By increasing the threshold for lead conversion (e.g., from 0.5 to 0.7), only the most promising leads were prioritized, reducing wasted efforts and resources.

**Learnings and Recommendations:**

This project reinforced the importance of using data-driven insights to optimize business processes. Key learnings included:

- **Threshold Adjustment**: The ability to adjust thresholds based on business objectives (e.g., aggressive vs. conservative conversion strategies) is crucial in maximizing lead conversion.
- **Feature Importance**: Understanding which features most influence the target variable is essential for developing effective business strategies. In this case, **website engagement** and **lead source** were critical.
- **Actionable Insights**: By leveraging predictive modeling, X Education can allocate resources more effectively, leading to better ROI and higher conversion rates.

Overall, this analysis demonstrated how data science and machine learning can directly contribute to improving business outcomes by enabling more targeted and efficient lead conversion strategies.