Summary Report: Lead Scoring Analysis for X Education

Project Overview:

The primary goal of this project is to improve the lead conversion rate at X Education by building a logistic regression model that identifies the most promising leads, also known as "Hot Leads." By assigning a lead score to each potential customer, the sales team can prioritize their efforts toward leads that are more likely to convert into paying customers. The objective is to help the sales team focus on high-conversion potential leads and thus increase the lead conversion rate from its current rate of 30% to a targeted 80%.

Data Understanding and Preprocessing:

The dataset contains approximately 9000 records, with each row representing a lead along with features such as Lead Source, Total Time Spent on Website, Last Activity, and Converted (the target variable).

Several data preprocessing steps were conducted to clean the dataset:

- **Missing Values**: Missing values in the dataset, such as 'Select' in categorical variables, were replaced with NaN. Missing values were handled by dropping rows in some cases and imputing using mode (for categorical variables) or median (for numerical variables).
- **Duplicates**: Duplicate rows were removed to ensure data quality.
- **Dummy Variables**: Categorical variables were converted into dummy variables for model compatibility.
- **Feature Scaling**: Numerical variables were scaled using `StandardScaler` to ensure that all features are on the same scale, which is essential for logistic regression.

Model Building:

The logistic regression model was chosen because of its interpretability and effectiveness in binary classification tasks. It assigns probabilities to each lead, indicating the likelihood of conversion. Hyperparameter tuning was performed using `GridSearchCV` to optimize model parameters like the regularization strength (`C`) and solver type. This ensured that the model was both accurate and generalizable.

Key steps in model building included:

- **Data Splitting**: The data was split into training (70%) and testing (30%) sets.
- **Logistic Regression Model**: The model was trained on the training set and evaluated on the test set to assess its performance.
- **Hyperparameter Tuning**: `GridSearchCV` was used to find the best combination of hyperparameters for optimal performance.

Model Performance:

The logistic regression model's performance was evaluated using several key metrics:

- **Accuracy**: 92% – The model correctly predicted whether a lead would convert or not in 92% of the cases.
- **Precision**: 94% – Of the leads predicted to convert, 94% were actual conversions.
- **Recall**: 95% – Of the actual converted leads, the model was able to correctly identify 95%.
- **F1-Score**: 95% – The F1-score, which balances precision and recall, was 95%.
- **AUC-ROC Score**: 0.95 – This indicates that the model has a high ability to distinguish between converting and non-converting leads.

A confusion matrix and ROC curve were generated to provide further insights into the model's classification abilities. The high precision and recall values indicate that the model successfully identifies high-potential leads with minimal false positives and negatives.

Feature Importance:

One of the strengths of logistic regression is its interpretability. The most important features influencing lead conversion were:

- **Total Time Spent on Website**: Leads who spent more time on the website were more likely to convert.
- **Lead Source (Google)**: Leads generated from Google showed a higher probability of conversion.
- **Last Activity (Email Opened)**: Leads who opened emails were highly likely to convert.

These insights help the sales and marketing teams to focus their efforts on optimizing website engagement and targeting high-quality leads from effective sources like Google.

Business Strategy Recommendations:

The model provides a clear strategy for the sales team to focus on high-potential leads, thus improving efficiency:

1. **Aggressive Conversion Strategy** (During periods with more resources, such as interns): Focus on leads with high predicted conversion probabilities. Prioritize contacting leads from high-converting sources (e.g., Google) and with high engagement metrics (e.g., high website time spent or email activity).
2. **Conservative Strategy** (During low activity periods or after reaching sales targets): The sales team should focus only on leads with the highest confidence scores for conversion. This minimizes unnecessary calls and optimizes resource allocation.

Conclusion:

The logistic regression model built in this project successfully identifies high-conversion potential leads, providing actionable insights for the sales and marketing teams at X Education. By implementing this model, X Education can prioritize resources, improve sales efficiency, and ultimately aim to increase their lead conversion rate to 80%. The analysis also highlights important features like website engagement and lead source, which can be used to guide future marketing efforts