Summary Report: Lead Score Case Study

Objective:

The objective of this assignment was to develop a predictive model to identify potential leads who are more likely to convert into customers for a training institute, X Education. The model aimed to help the sales team prioritize leads, especially during the enrolling of students, and make the lead conversion process more efficient.

Data Understanding and Preprocessing:

The dataset consisted of various features related to the leads, including their activities, demographics, and preferences. The initial step involved understanding the data and performing necessary preprocessing.

1. **Data Cleaning**:
    - **Dropping Irrelevant Columns**: We started by dropping ID columns like Prospect ID and Lead Number as they don't contribute to the predictive power of the model. Columns with uniform data were also removed as they offered no variance.
    - **Handling Missing Values**: Features with a high percentage of missing values (greater than 40%) were dropped, while others were imputed with the mode or removed depending on their business significance.
    - **Categorical Columns**: Categorical features were identified, and necessary transformations were applied. Features like Lead Source, Last Activity, and What is your current occupation were converted to dummy variables for inclusion in the logistic regression model.
2. **Exploratory Data Analysis (EDA)**:
    - **Correlation Analysis**: We explored correlations between numerical features and the target variable (Converted). A heatmap was generated to visualize these correlations, highlighting that features like TotalVisits, Total Time Spent on Website, and Do Not Email had significant relationships with lead conversion.
    - **Box Plots and Outlier Detection**: Box plots were used to detect outliers in the data. While some values were outside the typical range, they were considered valid and not treated as outliers.
    - **Bar Charts for Categorical Features**: Bar charts were plotted for categorical features to understand their distribution and relation to the target variable. Insights included that leads from Lead Add Form and those with India as the country had higher etc have conversion rates.

Model Building:

We used a logistic regression model for this binary classification problem.

1. **Feature Selection**:
    - **Recursive Feature Elimination (RFE)**: Given the large number of features, RFE was employed to select the most significant ones. This process helped in reducing multicollinearity and retaining only the most predictive features.

- **Handling Multicollinearity**: Variance Inflation Factor (VIF) was calculated for all features, and those with high VIF values were removed to avoid multicollinearity issues.

2. **Model Evaluation**:
   - **Training Model**: The model was first trained on the training dataset, achieving an accuracy of approximately 81%. The sensitivity was around 70%, specificity 89%, and the area under the ROC curve was 0.89.
   - **Cutoff Selection**: Different probability cutoffs were explored to balance sensitivity and specificity. A cutoff of 0.4 was found to be optimal, improving the sensitivity to 77% and specificity to 84%.
   - **Test Model**: The model was then evaluated on the test dataset, where it maintained similar performance metrics, indicating good generalizability.

Business Implications and Learnings:

The logistic regression model revealed several key features that strongly influence lead conversion:

- **TotalVisits (6.5446)**: This feature had the highest positive coefficient, indicating that the more a lead visits the website, the higher their likelihood of conversion.
- **Lead Source - Welingak Website (6.1228)**: Leads coming from the Welingak website showed a high propensity to convert, making it a crucial source.
- **Total Time Spent on Website (4.6192)**: The amount of time spent on the website was also a significant predictor of conversion, reflecting the lead's engagement level.
- **Lead Source - Reference (4.0839)**: Referrals had a strong positive impact, suggesting that word-of-mouth remains a powerful conversion tool.
- **Occupation - Working Professional (2.1704)**: Working professionals were more likely to convert, indicating a targeted demographic.
- **Negative Influences:** Several activities and behaviors negatively impacted conversion. For instance, Last Notable Activity - Email Link Clicked (-1.9029) and Do Not Email (-1.6363) had high negative coefficients, suggesting that such leads were less likely to convert. Activities like Olark Chat Conversation (-1.5682) and Email Opened (-1.4323) also showed negative correlations, indicating potential disengagement.
- **Adjust the Model's Threshold:** Lower/increase the cutoff probability threshold (cutoff_prob variable in notebook/code) threshold of the logistic regression model from the 0.4 to a lower value of 0.3 or 0.2 or increase to 05, 0.6, 0.7 depending on how aggressive the campaign should be. Lowering the threshold will classify more leads as "high potential," ensuring that a greater number of leads are flagged for follow-up. Increasing threshold value will reduce the number of conversions but leads will have higher potentials of concersion.
- **Follow the Lead_Score:** The lead score (100 to 0) column in the model output is the indicator of how much potential is there for conversion. Higher value indicates higher probability of conversion. Sales people can start from highest lead score and work their way down.

This model provides X Education with a clear roadmap to focus their efforts on leads with the highest potential for conversion, while understanding the behaviours that may require intervention to turn around. It offers a strategic advantage by highlighting where to allocate resources effectively, ensuring that the sales team targets leads who are most likely to bring value.