Summary

X Education gets a lot of leads, its lead conversion rate is very poor at around 30%. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have higher conversion chance. CEO's target for lead conversion rate is around 80%.

Data Cleaning:

- **Columns with more than 35% null values were dropped**. The value counts within categorical columns were examined to determine the appropriate action: if imputation caused skew, the column was either dropped, a new category ("others") was created, the high-frequency value was imputed, or columns that didn't add value were removed.
- **Numerical categorical data were imputed with the mode**, and columns with only one unique response from customers were dropped.
- **Other activities included**: treating outliers, fixing invalid data, grouping low-frequency values, and mapping binary categorical values.

EDA:

- **Checked for data imbalance**: Only 38.0% of leads converted.
- **Conducted univariate and bivariate analysis**: Analyzed categorical and numerical variables. Variables such as 'Lead Origin', 'Current Occupation', and 'Lead Source' provided valuable insights into their effect on the target variable.
- **Observed website engagement**: Time spent on the website showed a positive impact on lead conversion.

Data Preparation:

- **Generated dummy features** for categorical variables.
- **Split the data into training and test sets** using a 70:30 ratio.
- **Applied feature scaling** through standardization.
- **Removed highly correlated columns** to avoid redundancy.

Model Building:

- **Applied Recursive Feature Elimination (RFE)** to reduce the number of variables from 31 to 14, making the DataFrame more manageable.
- **Performed manual feature reduction** by dropping variables with p-values greater than 0.05 to build the models.
- **Developed two models**: The final model (Model 2) was stable with p-values less than 0.05 and showed no signs of multicollinearity (VIF < 5).
- **Selected logm2 as the final model** with 14 variables, and used it to make predictions on both the training and test sets.

Model Evaluation:

- **Constructed a confusion matrix** and selected a cut-off point of 0.36 based on the accuracy, sensitivity, and specificity plot. This cut-off provided accuracy, specificity, and precision all around 80%, while the precision-recall view showed lower performance metrics around 74%.
- **To address the business problem**, the CEO requested boosting the conversion rate to 80%. However, metrics dropped when using the precision-recall view. Therefore, we chose the sensitivity-specificity view for our optimal cut-off for final predictions.
- **Assigned lead scores** to the training data using a cut-off of 0.36

Making Predictions on Test Data:

- **Making Predictions on Test Data**: Applied scaling and used the final model for predictions.
- **Evaluation Metrics**: The metrics for both training and test sets are close to 80%.
- **Assigned Lead Scores**: Lead scores were assigned based on the model.
- **Top 3 Features**:

- Occupation_Working Professional: 3.7478
- Lead Origin_Lead Add Form: 3.6829
- Occupation_Businessman: 2.1988

Recommendations:

Based on the results of the logistic regression model, here are some recommendations for the business:

1. **Prioritize High-Impact Features**:
   - **Occupation_Working Professional**: With the highest coefficient (3.7478), leads identified as working professionals should be prioritized as they have the highest likelihood of conversion.
   - **Lead Origin_Lead Add Form**: This feature also has a strong positive impact (3.6829). Focus on leads originating from the Lead Add Form.
   - **Occupation_Businessman**: Another significant feature (2.1988). Leads identified as businessmen should be given priority.
2. **Enhance Website Engagement**:
   - **Total Time Spent on Website**: This variable shows a positive impact on lead conversion (1.1180). Encourage potential leads to spend more time on the website through engaging content and interactive features.
3. **Utilize Effective Communication Channels**:
   - **Last Activity_SMS Sent**: This has a very high positive impact (2.0291). Continue sending SMS messages to leads as it significantly boosts conversion rates.
   - **Last Activity_Email Opened**: Also shows a strong positive impact (1.0177). Ensure that email campaigns are effective and track email opens to identify engaged leads.
4. **Leverage Multiple Lead Sources**:
   - **Lead Source_Olark Chat**: This source has a significant positive impact (1.2122). Utilize the Olark Chat feature on the website to engage with potential leads.
   - **Lead Source_Welingak Website**: Although less impactful than others, it still shows a positive effect (1.7369). Continue to optimize this source for lead generation.
5. **Address Negative Indicators**:
   - **Last Activity_Email Bounced**: This has a negative impact (-1.0726). Minimize email bounces by maintaining a clean email list and verifying email addresses.
6. **Focus on Diverse Occupations**:
   - **Occupation_Other, Occupation_Student, and Occupation_Unemployed**: These categories also show positive impacts. Tailor marketing strategies to effectively engage with these groups.
7. **Monitor and Adjust Strategies**:
   - Regularly review the model’s performance and adjust strategies based on new data and changing market conditions. This will help in maintaining high conversion rates and adapting to new challenges.

By implementing these recommendations, the business can enhance its lead conversion rates and achieve its sales targets more effectively