Lead Scoring Case Study Summary

Problem Statement:

- **A/B testing:** Experiment with different scoring systems and weights to find the optimal combination.
- **Continuous monitoring:** Track the performance of the model and make adjustments as needed.
- **Leverage machine learning:** Consider using machine learning algorithms to automatically identify patterns in lead data and predict conversion likelihood.

Achieving the 80% Target Conversion Rate

- **Focus on high-scoring leads:** Prioritize leads with the highest scores for nurturing and follow-up.
- **Optimize lead nurturing campaigns:** Tailor content and messaging to the specific needs and interests of high-scoring leads.
- **Measure and analyze results:** Track conversion rates for different lead score segments to identify areas for improvement.

By implementing a well-designed lead scoring model, X Education can significantly improve its lead conversion rates and achieve its target of 80%.

Solution Summary:

Step1: **Reading and Understanding Data**. Read and analyze the data.

Step2: **Data Cleaning**:

1. **Variable Removal:** You identified and removed variables with excessive missing data, which can introduce bias and noise into your analysis.
2. **Missing Value Imputation:** You used median imputation for numerical variables, which is a common and effective approach when the data is skewed. Creating new classification variables for categorical variables is also a good practice to avoid introducing bias.
3. **Outlier Identification and Removal:** Outliers can significantly impact your model's performance. By identifying and removing them, you've ensured that your analysis is based on representative data.

Key considerations for future analysis:

- **Impact of outlier removal:** Evaluate the impact of outlier removal on your dataset's distribution and potential biases.
- **Alternative imputation methods:** Explore other imputation techniques, such as K-nearest neighbors or regression, if appropriate for your data.
- **Feature engineering:** Consider creating new features from existing variables to capture more meaningful relationships.

Step3: **Data Analysis**

Dropping variables with only one unique value is a crucial step in data exploration. These variables, often referred to as "constant variables," provide no variance and therefore offer no predictive power in your analysis.

By removing these variables, you've simplified your dataset and improved its efficiency for modeling.

Here are some additional considerations for your exploratory data analysis:

- **Variable distributions:** Examine the distribution of your remaining variables to identify any skewness or outliers.
- **Correlation analysis:** Assess the relationships between variables to identify potential multicollinearity issues.
- **Data visualization:** Use visualizations like histograms, box plots, and scatter plots to gain insights into the data's characteristics.
- **Missing value patterns:** Investigate any remaining missing values to determine if they follow specific patterns or are randomly distributed.

Step4: **Creating Dummy Variables**

We went on with creating dummy data for the categorical variables.

Step5: **Test Train Split**:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling

Min-Max scaling is an effective technique for scaling numerical variables to a specific range (usually 0 to 1). This ensures that all features contribute equally to the model, preventing features with larger magnitudes from dominating the analysis.

Using statsmodels to create your initial model is a wise decision. This statistical modeling library provides valuable insights into the significance of each parameter, their coefficients, and other relevant metrics.

Here are some additional considerations for your model building:

- **Feature selection:** If you have a large number of features, consider using feature selection techniques (e.g., correlation analysis, recursive feature elimination) to identify the most relevant variables.
- **Model evaluation:** Use appropriate metrics (e.g., R-squared, adjusted R-squared, RMSE) to evaluate your model's performance.
- **Regularization:** Explore regularization techniques (e.g., L1, L2) to prevent overfitting and improve model generalization.

- **Cross-validation:** Employ cross-validation techniques (e.g., k-fold cross-validation) to assess your model's performance on unseen data.

Step7: **Feature selection using RFE**:

Using the Recursive Feature Elimination we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.
Finally, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good.

We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

Step8: Plotting the ROC Curve

We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 89% which further solidified the of the model.

Step9: Finding the Optimal Cutoff Point

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.37

Based on the new value we could observe that close to 80% values were rightly predicted by the

model. We could also observe the new values of the 'accuracy=81%, 'sensitivity=79.8%',

'specificity=81.9%'.

Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 80%

Step10: Computing the Precision and Recall metrics

We also found out the Precision and Recall metrics values came out to be 79% and 70.5% respectively on the train data set.

Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.42

Step11: **Making Predictions on Test Set**

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 80.8%; Sensitivity=78.5%; Specificity= 82.2%.