Lead Scoring Case Study Summary

Problem Statement:

X Education is an education company that offers online courses for industry professionals. The company attracts many visitors to its website through various marketing channels. The company faces a problem: its lead conversion rate is very low. Out of 100 leads, only 30 become customers on average.

To solve this problem, X Education wants to identify the most potential leads, also known as 'Hot Leads'. The company has hired you to help them with this task. Your job is to build a model that can assign a lead score to each lead based on various factors, such as their demographics, behavior, preferences, etc. The higher the lead score, the more likely the lead is to convert. The lower the lead score, the less likely the lead is to convert. The company's CEO has set a target of achieving an 80% lead conversion rate with this model.

Solution Summary:

Step1: Reading and Understanding Data:
Read and inspected the data.

Step2: Data Cleaning:

Here's how we prepared the data:

a. First, we dropped variables with unique values.

b. For columns where the value was 'Select' (indicating that the lead didn't pick any option), we replaced those values with nulls.

c. We removed columns with more than 52% null values. However, we kept the *Lead Quality* column (despite 52.9% missing values) because it seemed important. We assumed that missing values meant the employee was unsure, so we imputed them with *Not Sure*.

d. We cleaned and preprocessed the data. Skewed and duplicate variables were removed, and missing numerical values were filled with the median. For categorical variables, new categories were created. Outliers were detected and removed. We also standardized inconsistent labels by converting them all to uppercase.

e. Variables generated by the sales team were excluded to avoid ambiguity in the final results.

Step3: Data Transformation:

Changed the binary variables into ‘0’ and ‘1’

Step4: Dummy Variables Creation:

a. We created dummy variables to represent the categorical data.

b. All repeated and redundant variables were identified and removed.

Step5: Test Train Split:

The next step involved splitting the dataset into training and testing sets, with 70% of the data used for training and 30% for testing.

Step6: Feature Rescaling:

a. We applied Standard Scaling to normalize all the variables.

b. Next, we plotted a heatmap to visualize the correlations between the variables.

Step7: Model Building:

a. We used Recursive Feature Elimination (RFE) to select the top 15 important features.

b. Based on statistical analysis, we iteratively examined P-values to identify the most significant features and removed the insignificant ones.

c. This process helped us finalize the 12 most significant variables, all with acceptable Variance Inflation Factor (VIF) values.

d. For the final model, we determined the optimal probability cutoff by analyzing accuracy, sensitivity, and specificity.

e. We plotted the ROC curve for the features, achieving a strong area under the curve (AUC) of 95%, which validated the model’s performance.

f. We verified that the model correctly predicted outcomes for 80% of cases based on the *converted* column.

g. Precision, recall, accuracy, sensitivity, and specificity were assessed on the training set for the final model.

h. Using the trade-off between precision and recall, we identified an optimal cutoff value of approximately 0.25.

i. Finally, we applied these insights to the test dataset, calculating conversion probabilities using sensitivity and specificity. The results showed an accuracy of 90.78%, sensitivity of 84.12%, and specificity of 94.58%.

Step 8: Conclusion:

- The lead score calculated on the test dataset indicates a conversion rate of 84% with the final predicted model, exceeding the CEO's target of approximately 80%.
- The model's high sensitivity ensures it effectively identifies the most promising leads.
- The features contributing most significantly to the likelihood of lead conversion are:
  i. **Tags_Lost to EINS**
  ii. **Tags_Closed by Horizzon**
  iii. **Tags_Will revert after reading the email**

* * *