Lead Scoring Case Study Summary

Problem Statement

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

To solve this, X Education wants to find out the most potential leads, or 'Hot Leads'. The company has hired you to build a model that assigns a lead score based on demographics, behavior, preferences, etc. The higher the lead score, the more likely the lead is to convert. The CEO has set a target of achieving an 80% lead conversion rate with this model.

Solution Summary

Step 1: Reading and Understanding Data
The dataset was inspected to understand its structure and identify missing values, outliers, and key patterns.

Step 2: Data Cleaning

- Columns with unique values were dropped.
- Entries with the value "Select" were treated as null values.
- Features with over 52% null values were removed, except *Lead Quality*. Its missing values were imputed as "Not Sure," assuming uncertainty.
- Numerical variables with missing values were imputed with medians, and new categories were created for categorical variables with missing values.
- Outliers were figured out and removed, and inconsistent labels were standardized.
- Variables generated by the sales team were excluded to avoid confusion in the final model.

Step 3: Data Transformation
Binary variables were transformed into numerical representations ('0' and '1') for processing.

Step 4: Dummy Variable Creation
Dummy variables were created for categorical features, and redundant variables were removed to avoid multicollinearity.

Step 5: Train-Test Split
The dataset was split into training (70%) and test (30%) sets to validate the model's performance effectively.

Step 6: Feature Rescaling and Correlation Analysis

All features were standardized using Standard Scaling. A heatmap was created to examine correlations between variables, aiding in a robust feature selection process.

Step 7: Model Building

- Recursive Feature Elimination (RFE) is used to select the top 15 features.
- Insignificant features were iteratively removed, resulting in 12 significant features with acceptable Variance Inflation Factors (VIFs).
- The optimal probability cutoff was determined by analyzing accuracy, sensitivity, and specificity.
- The ROC curve demonstrated strong performance with an area under the curve (AUC) of 95%.
- A cutoff value of 0.25 was chosen based on the Precision-Recall trade-off.

Step 8: Model Performance

We applied the insights to the test model and computed the conversion probability using Sensitivity and Specificity metrics. The results showed an accuracy of 91.33%, with a Sensitivity of 84.12% and a Specificity of 95.44%.

Conclusion

The lead score analysis on the test dataset demonstrates a sensitivity of 84%, successfully meeting the CEO's goal of an 80% conversion rate. This indicates that 84% of actual conversions are accurately identified. The model effectively pinpoints the most promising leads.

Features which contribute more towards the probability of a lead getting converted are:

i. Tags_Lost to EINS
ii. Tags_Closed by Horizzon
iii. Tags_Will revert after reading the email