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 identify 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 detected and removed, and inconsistent labels were standardized.
- Variables generated by the sales team were excluded to avoid ambiguity 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 scaled using Standard Scaling. A heatmap was generated to analyze correlations among variables and ensure a robust feature selection process.

Step 7: Model Building

- Recursive Feature Elimination (RFE) was 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 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 91.33%; Sensitivity= 84.12%; Specificity= 95.44%.

Conclusion

The lead score calculated in the test set of data shows a sensitivity of 84%, which clearly meets the CEO's target of an 80% conversion rate.

- A sensitivity of 84% means we are correctly identifying 84% of actual conversions.
- The model helps select 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