Lead Scoring Case Study Summary

Problem Statement:

X Education, an online education company for industry professionals, attracts many potential clients to their website daily, where they explore various courses. While some of these leads eventually convert into customers, the conversion rate is typically around 30%.

To enhance efficiency, the company aims to identify the most promising leads, referred to as 'Hot Leads.' By targeting this specific group, they expect the conversion rate to improve, as the sales team will concentrate their efforts on high-potential prospects rather than reaching out to all leads.

The company needs us to develop a model that assigns a lead score to each prospect. This scoring system should ensure that leads with higher scores are more likely to convert, while those with lower scores are less likely to convert. The CEO has indicated that the target conversion rate should be approximately 80%.

Solution Summary

Step 1: Importing and analyzing data.

- Import the dataset and review the data types.
- Calculate the percentage of missing values for each column.
- Analyze the range and distribution of data in numerical columns.
- Count the unique categories in each column.

Step 2: Data Cleaning

- Replace 'Select' values with nulls.
- Remove highly skewed and unnecessary columns.
- Eliminate columns with 30% or more missing values.
- Impute missing values in columns with low null percentages: use the median for numerical columns with outliers, and the mode for categorical columns.
- Clean categorical columns by removing duplicates and consolidating low-frequency categories.
- Address outliers by capping their values.

Step 3: Data Preparation

- Create dummies for the categorical columns.
- Split the data into training and testing sets.
- Apply MinMaxScaler to scale numerical columns in the training set

Step 4: Exploratory Data Analyst

- Conduct EDA on categorical variables to examine category-wise counts in relation to conversion types.
- Create boxplots for numerical attributes relative to conversion status.
- Observe that customers who converted typically spent more time on the website.

Higher conversion rates were observed for:

- Lead origin as Lead Add form
- Lead source as Reference
- Last activity as SMS sent
- Current occupation as Working professional

Step 5: Model Building

- Initialize Logistic Regression model.
- Perform coarse feature selection using Recursive Feature Elimination to identify the top 15 features.
- Refine feature selection by removing features with high p-values and high Variance Inflation Factor (VIF).
- The final model included 11 features, all with p-values < 0.05 and VIF < 5.

Step 6: Model Performance Evaluation

- Make predictions with the final model on the training set.
- Plot the ROC curve, which demonstrated a strong performance with an AUC of 0.87.
- Use the confusion matrix to calculate Accuracy, Sensitivity (Recall), and Specificity at various probability thresholds.
- Sensitivity or Recall is the Total number of Leads correctly predicted as Converted out of the Total number of actual Converted Leads.
- Recall = $\frac{\text{True Positive (TP)}}{\text{Total Actual Positives (TP + FN)}}$
- Precision is the Total number of leads correctly predicted as Converted out of the total number of leads which are predicted as Converted.
- Precision = $\frac{\text{True positives (TP)}}{\text{Total Predicted positives (TP+FP)}}$
- Given that a slightly lower precision might involve reaching some non-potential leads, we balanced this by targeting higher recall to identify more potential leads.
- Set a cut-off probability of 0.35 to achieve higher recall and reasonable precision, resulting in a target lead conversion rate of approximately 79%.

Step 7: Evaluation of model on the test set

- Transform the numerical columns in the test set using the MinMaxScaler.
- Select columns from the test set that were used in the final model.
- Generate predictions on the test set and evaluate performance metrics.

	Accuracy	Precision	Recall	f1-score
Train	0.790816	0.70247	0.789855	0.743604
Test	0.797619	0.711823	0.805014	0.755556

Step 8: Forming the Scores dataframe

- Join the initial dataframe and predictions dataframe on 'ID' column which is the index of the rows from initial dataframe.
- Merge the original dataframe with the predictions dataframe using the 'ID' column, which serves as the index in the initial dataframe.
- Calculate the Score as \(\text{Probability of Conversion} \times 100 \).
- The final scores dataframe includes columns for Prospect ID, Lead Number, Actual Converted status, Predicted Converted status, and the Score.

	Prospect ID	Lead Number	Converted	final_predicted	Score
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	0	0	18.59
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	0	1	38.12
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	1	1	73.17

Submitted By:
Adwait Mohgaonkar
Ankush Sharma
Mohanpriya Pichandi