UPGRAD ASSIGNMENT

Lead Scoring Case Study Overview

Problem Statement:

X Education offers online courses and gathers leads from various channels. Metadata for each lead is recorded, and a team works to nurture promising leads into confirmed opportunities.

Proposed Solution:

To enhance conversion rates and optimize time usage, the focus should be on "hot leads," which have a higher likelihood of conversion. A logistic regression model will be used to determine these leads by assigning a score based on their metadata.

Data Analysis:

1. Columns with a high percentage of missing data are initially considered as having missing values.
2. Categorical columns with less than 5% missing values will be imputed using the mode.
3. Quantitative columns with minimal missing values will be imputed using the median, as there is no significant difference between median and mean.
4. Categorical columns with more than 70% missing values will be excluded.
5. Other missing values will remain unfilled to avoid distorting the data.

Data Preparation:

1. Outliers, identified through boxplots and descriptive statistics, will not be removed to ensure all leads are scored.
2. Key categorical variables critical for lead conversion will be identified through bivariate analysis.
3. Categorical data will be converted to numerical data using:
   - Dummy Variables for low/moderate-level categories.
   - Label Encoding for high-level categories to manage dataframe size.
4. Columns with no variance will be removed.
5. Correlations between variables will be examined using a heatmap, and VIF will be employed during model construction.

Model Development:

1. RFE and PCA techniques will be used to identify the most effective model.
2. The dataset will be split into training and testing sets.
3. Numerical data will be standardized using a standard scaler.
4. Functions for repetitive tasks include:
   - Createmodel: Outputs model summary, VIF, and returns the model.
   - Confscores: Provides accuracy, sensitivity, and specificity from the confusion matrix.

Author: Keshav Atchutuni, Pranav Keshav & Khushi Patel (December 2023 Batch)

  - Calctrainseult: Produces confusion metrics and scores based on the cutoff.

5. RFE will identify the top 20 variables, with model tuning focusing on high p-values and high VIFs.
6. ROC and AUC metrics will validate model performance.
7. The optimal cutoff value will be identified by plotting accuracy, sensitivity, specificity, recall, and precision.

Prediction Making:

1. Model6 and the optimal cutoff will be used for test dataset predictions.
2. PCA will be applied to assess if it provides a better model by addressing multicollinearity, though it may result in lower scores and complexity in identifying original variables.

Model Selection and Lead Scoring:

1. The final model, developed using RFE, will be employed for predictions.
2. Lead scores will be assigned based on predicted probabilities (Lead Score = Predicted Probability * 100).
3. A dataframe will be created to plot conversion versus cutoff.

Conclusions:

1. Key features impacting decisions:
    - Tags
    - Lead Quality
    - Asymmetries Profile Index
2. Major categories influencing decisions:
    - Lead Origin: Landing Page Submission
    - Lead Origin: Lead Add Form
    - Lead Source: Olark Chat

Key Learnings:

1. Conducting Exploratory Data Analysis (EDA) is essential for constructing an accurate model, as it helps in proper data treatment.
2. Data cleaning, including missing value imputation, scaling, and outlier handling, is critical to maintain data integrity.
3. RFE is effective for identifying key features, whereas PCA is useful for reducing dimensionality.
4. Developing modular code with functions for repetitive tasks enhances reusability.
5. Balancing sensitivity and specificity is crucial for determining the optimal cutoff.
6. Confusion metrics provide valuable insights into model performance, facilitating the calculation of accuracy, sensitivity, and specificity.