Summary

Done By -

- ***Shweta Alukuru Trikutam***
- ***Trupti Nidhalkar***
- ***Tiana Viola Pinto***

X Education faces a challenge with a low lead conversion rate of approximately 30%. To meet the CEO's target of achieving an 80% conversion rate, we developed a lead scoring model. This model assigns a score to each lead, prioritizing those with higher scores, indicating a greater likelihood of conversion.

Data Cleaning

- Initially, columns with more than 40% null values were dropped.
- Categorical columns were examined to determine appropriate actions, such as imputation, dropping columns, creating new categories, or using high-frequency values.
- Numerical categorical data were imputed using the mode, and columns with only one unique response were dropped.
- Other activities included treating outliers, fixing invalid data, grouping low-frequency values, and mapping binary categorical values.

EDA (Exploratory Data Analysis)

- We checked for data imbalance, noting that only 38.5% of leads converted.
- Univariate and bivariate analyses were conducted on categorical and numerical variables.
- Insights were gained into variables such as 'Lead Origin,' 'Current Occupation,' and 'Lead Source,' which showed significant effects on the target variable.
- The time spent on the website positively correlated with lead conversion.

Data Preparation

- We created dummy features (one-hot encoded) for categorical variables and split the data into training and testing sets using a 70:30 ratio.
- Feature scaling was performed using standardization.
- Highly correlated columns were dropped to improve model performance.

Model Building

- We used Recursive Feature Elimination (RFE) to reduce the number of variables from 48 to 15.
- Additionally, a manual feature reduction process was employed, dropping variables with a p-value > 0.05.
- Three models were developed before finalizing Model 4, which demonstrated stability with all p-values < 0.05 and no multicollinearity (VIF < 5).
- Model 4, consisting of 12 variables, was selected for predicting outcomes on both the train and test sets.

Model Evaluation

- We created a confusion matrix and selected a cutoff point of 0.345 based on accuracy, sensitivity, and specificity plots.
- This cutoff achieved around 80% accuracy, specificity, and precision.
- Despite lower performance metrics in the precision-recall view, this cutoff aligns with the CEO's objective of boosting the conversion rate to 80%.

Making Predictions on Test Data

- Using the finalized model, we scaled and predicted outcomes on the test data.
- Evaluation metrics on both train and test sets approached 80%.
- Lead scores were assigned based on the 0.345 cutoff.

Top 3 Features

The top three features influencing lead conversion were identified as:

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional

Recommendations

To enhance lead conversion rates:

- Allocate more budget towards advertising on the Welingak Website.
- Offer incentives or discounts for references that convert into leads, encouraging more referrals.
- Implement targeted strategies to engage working professionals, leveraging their higher conversion rates and better financial capability to pay higher fees.