Lead Scoring Case Study - Summary

X Education is facing a challenge with its lead conversion rate, which is currently at 30%. In order to improve this rate and achieve the CEO's target of an 80% conversion rate, the company requires us to develop a lead scoring model. The goal of this model is to assign a score to each lead that reflects their likelihood of converting to a customer, with higher scores indicating a higher chance of conversion. By prioritizing leads with higher scores, the company can focus its resources on the most promising leads and increase its overall conversion rate.

Data Cleaning

- Columns with more than 40% null values were dropped.
- Value counts were checked within categorical columns to decide appropriate action for handling missing values.
- If imputing missing values causes skew, then the column was dropped or a new category (others) was created.
- High frequency values were used to impute missing values in categorical columns.
- Columns with only one unique response from customers were dropped.
- Mode was used to impute missing values in numerical categorical columns.
- Outliers were treated using appropriate methods.
- Invalid data points were fixed or removed.
- Low frequency values in categorical columns were grouped together.
- Binary categorical values were mapped to 0 and 1.

EDA

- Checked for data imbalance and found that only 38.5% of leads converted.
- Performed univariate and bivariate analysis for both categorical and numerical variables.
- Variables like 'Lead Origin', 'Current occupation', 'Lead Source', etc. provided valuable insights on the effect of these variables on the target variable.
- Found that the 'Time spent on the website' had a positive impact on lead conversion.

Data Preparation

- Categorical variables were one-hot encoded to create dummy features
- The dataset was split into a training set and a test set in a 70:30 ratio
- Feature scaling was applied using standardization
- Highly correlated columns were dropped to avoid multicollinearity issues.

Model Building

- RFE (Recursive Feature Elimination) was used to reduce the number of variables from 48 to 15, which helped to make the data frame more manageable.
- A manual feature reduction process was also used to build models by dropping variables with p-values > 0.05.
- A total of 3 models were built before reaching the final Model 4, which was stable with p-values < 0.05 and no sign of multicollinearity with VIF < 5.

- The final model selected was logm4, which had 12 variables. This model was used to make predictions on the train and test set.

Model Evaluation

- A confusion matrix was created and a cut-off point of 0.345 was selected based on accuracy, sensitivity, and specificity plots.
- This cut-off provided accuracy, specificity, and precision scores around 80%, while precision-recall scores were lower around 75%.
- Since the CEO's goal was to increase the conversion rate to 80%, the sensitivity-specificity view was chosen over the precision-recall view.
- Using the cut-off of 0.345, lead scores were assigned to the train data.

Making Predictions on Test Data

- The final model was used to make predictions on the test data after scaling.
- The evaluation metrics for both train and test sets were around 80%, indicating good performance of the model.
- Lead scores were assigned to the test data using the chosen cut off of 0.345.
- The top three features that had the most impact on lead conversion were determined to be:
    - Lead Source_Welingak Website
    - Lead Source_Reference
    - Current_occupation_Working Professional

Recommendations

- Allocate more budget/spend on Welingak Website for advertising and other promotional activities as it has shown to be the top lead source for conversions.
- Offer incentives or discounts to encourage customers to provide references that convert to leads. This can help in increasing the number of high-quality leads.
- Focus on aggressively targeting working professionals as they have a higher conversion rate and are more likely to have a better financial situation to pay higher fees. This can be achieved through tailored messaging and targeted marketing strategies.
- Monitor the performance of other lead sources and optimize communication channels based on lead engagement impact to attract high-quality leads from top-performing sources.
- Consider dropping columns with high null values or impute them with high-frequency values to reduce skewness in data. Additionally, outlier treatment, fixing invalid data, grouping low-frequency values, and mapping binary categorical values can improve the quality of the data.
- Use machine learning techniques like RFE and manual feature reduction to reduce the number of variables and build a stable final model with high predictive power.
- Choose an optimal cut-off for final predictions based on sensitivity-specificity view to achieve the desired conversion rate of 80%.

End of the document