SUMMARY

X Education receives a lot of leads, but only about 30% of those leads really become customers. The business wants us to create a model in which we score each lead individually so that leads with higher scores have a higher likelihood of converting. The CEO aims to convert leads at a rate of about 80%.

Data Cleaning:

- Over 40% null column values were removed. Value counts within categorical columns were reviewed to determine the best course of action: eliminate the column, create a new category (others), impute high frequency values, and drop columns that don't contribute any value if imputation creates skew.
- Columns and mode were used to impute numerical categorical data from a single client response were abandoned.
- Additional tasks like handling outliers, correcting flawed data, grouping low frequency values, and binary mapping
- The use of categorical values was done.

EDA:

- Checked for data imbalance, only 38.5% of leads were converted.
- performed categorical and numerical variables' univariate and bivariate analyses. 'Lead Origin', 'Current occupation', 'Lead Source', etc. give important information about the impact on the target variable.
- Time spent on a website has a favorable effect on converting visitors to leads.

Data Prep:

- For categorical variables, dummy features (one-hot encoded) were created.
- 70:30 split between train and test sets
- Feature Using standardization to scale
- When a few columns were dropped, they were very correlated with one another.

Model Building:

- We used Recursive Feature Elimination (RFE) to reduce the number of variables from 48 to 15 in order to make the dataset more manageable.
- We also employed a manual feature reduction process by dropping variables with a p-value greater than 0.05.
- We initially constructed three models, but ultimately settled on Model 4, which exhibited stability with p-values below 0.05. There were no indications of multicollinearity, as indicated by Variance Inflation Factors (VIF) less than 5.
- The final model, logm4, consisted of 12 variables, and we used it for making predictions on both the training and test sets.

SUMMARY

Model Evaluation:

- To assess the model's performance, we created a confusion matrix and selected a cut-off point of 0.345 based on accuracy, sensitivity, and specificity plots. This cut-off point yielded accuracy, specificity, and precision all around 80%. However, precision-recall metrics showed lower performance, around 75%.
- Our business objective was to increase the conversion rate to 80% as per the CEO's request. However, when we considered the precision-recall view, the metrics dropped. Therefore, we decided to use the sensitivity-specificity view to determine the optimal cut-off for final predictions.
- We assigned lead scores to the training data using a cut-off of 0.345.

Making Predictions on Test Data:

- We applied the final model to make predictions on the test data, scaling it appropriately.
- The evaluation metrics for both the training and test data were very close to 80%.
- The top three features that had a significant impact on lead conversion were:

a) Lead Source_Welingak Website
b) Lead Source_Reference
c) Current Occupation_Working Professional

Recommendations:

- We suggest allocating a higher budget for advertising and promotion of the Welingak Website, as it has proven to be an effective lead source.
- Providing incentives or discounts for individuals who refer potential leads that ultimately convert into leads can encourage more references.
- Targeting working professionals more aggressively is advisable, as they exhibit a high conversion rate and typically have better financial capacity to afford higher fees.