Summary

X Education receives a large number of leads; yet, its lead conversion rate is just about 30%. The organization requires us to create a model in which we assign a lead score to each lead so that clients with a higher lead score have a greater conversion rate. The CEO's aim for lead conversion rate is approximately 80%.

Data Cleaning:

● Columns containing more than 40% nulls were dropped. Value counts within categorical columns were evaluated to determine the appropriate action: if imputation produces skew, the column was discarded; otherwise, a new category (others) was created; high frequency values were imputed; and columns that added no value were dropped.

● Numerical categorical data were imputed using mode, and columns containing just one unique customer response were removed.

● Other exercises included dealing with outliers, incorrect data, grouping low frequency values, and mapping binary categorical values.

EDA:

● Data imbalance: only 38.5% of leads converted.

● Conducted univariate and bivariate analyses on categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', and other terms provide useful information about the effect on the target variable.

● Spending time on the website improves lead conversion rates.

Data Preparation:

● Generated dummy features (one-hot encoded) for categorical variables. ● Splitting Train and Test Sets: 70:30 ratio.

● Using standardization to scale features

● Removed a few highly connected columns.

Model Building:

● Utilized RFE to reduce variables from 48 to 15. To make dataframes more comprehensible, a manual feature reduction procedure was employed to remove variables with p-values greater than 0.05.

● Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.

● logm4 was selected as final model with 12 variables, we used it for making prediction on train and test set.

Model Evaluation:

● A confusion matrix was created and a cut off value of 0.345 was chosen based on accuracy, sensitivity, and specificity plots. This cutoff resulted in accuracy, specificity, and precision of roughly 80%. The precise recall view provided reduced performance metrics, roughly 75%.

● Despite the CEO's request to increase conversion rate to 80%, metrics decreased when precision-recall was considered. As a result, we will use the sensitivity-specificity view to determine our ideal cut-off for final forecasts.

● The lead score was awarded to the training data using a cut off of 0.345.

Making Predictions on Test Data:

● Scaling and forecasting test results using the final model.

● Evaluation metrics for training and testing are close to 80%.

● A lead score was assigned.

● Top three features are:

- Lead Sources: Welingak Website
- Lead Source_Reference.
- Current occupation_Working Professional

Recommendation

● Increase advertising budget for Welingak website.

● Provide referral incentives/discounts to encourage additional leads.

● Target working professionals with a high conversion rate and the ability to pay higher charges.