Summary

X Education gets a lot of leads, but its lead conversion rate is poor at around 30%. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance. CEO's target for lead conversion rate is around 80%.

Data Cleaning:

● Columns with >40% nulls were dropped. Value counts within categorical columns were checked to decide the appropriate action: if imputation causes skew, then the column was dropped, created new category (others), impute high-frequency value was, drop columns that don't add any value.

● Numerical categorical data were imputed with mode and columns with only one unique response from a customer were dropped.

● Other activities like outliers' treatment, fixing invalid data, grouping low-frequency values, and mapping binary categorical values were carried out

EDA:

● Data imbalance checked- only 38.5% of leads converted.

● Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight into the effect on the target variable.

● Time spent on the website shows a positive impact on lead conversion.

Data Preparation:

● Created dummy features (one-hot encoded) for categorical variables

● Splitting Train & Test Sets: 70:30 ratio

● Feature Scaling using Standardization

● Dropped a few columns, they were highly correlated with each other

Model Building:

● Used RFE to reduce variables from 48 to 15. This will make the data frame more manageable.

● A Manual Feature Reduction process was 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). No sign of multicollinearity with VIF < 5.

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

Model Evaluation:

● A Confusion matrix was made and a cut-off point of 0.345 was selected based on the accuracy, sensitivity, and specificity plot. This cut-off gave accuracy, specificity, and precision all around 80%. Whereas the precision-recall view gave fewer performance metrics around 75%.

● As to solving business problem CEO asked to boost the conversion rate to 80%, but metrics dropped when we took a precision-recall view. So, we will choose a sensitivity-specificity view for our optimal cut-off for final predictions

● Lead score was assigned to train data using 0.345 as a cut-off.

Making Predictions on Test Data:

● Making Predictions on Test: Scaling and predicting using the final model.

● Evaluation metrics for train & test are very close to around 80%. ● Lead score was assigned.

● Top 3 features are:

o Lead Source_Welingak Website

o Lead Source_Reference

o Current_occupation_Working Professional

Recommendations:

● More budget/spend can be done on the Welingak Website in terms of advertising, etc.

● Incentives/discounts for providing references that convert to lead, encourage to provide more references.

● Working professionals to be aggressively targeted as they have a high conversion rate and will have a better financial situation to pay higher fees too.