Summary

X Education gets a lot of leads, but its lead conversion rate is very 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. Imputing the columns with less than 40 % null values with suitable values.

● Checked for duplicates. Treated columns having 'Select' value with NaN

● Other activities like outliers' treatment, fixing invalid data

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 effect on the target variable.

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

Data Preparation:

● Converting 'Yes' / 'No' column values to 1 and 0

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

● Dropped a few columns, as they were insignificant for model building

Model Building:

● Splitting Train & Test Sets: 70:30 ratio

● Feature Scaling using Standardization

● Used RFE to reduce variables from 90 to 15. This will make data frame more manageable.

● Manual Feature Reduction process was used to build models by dropping variables with p-value> 0.05.

● A total of 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 the final model with 17 variables, we used it for predicting train and test sets.

Model Evaluation:

● Confusion matrix was made and a cut-off point of 0.34 was selected based on 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, an average of 76%.

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

● Plotted the ROC curve with an area of 0.89.

● Lead score was assigned to train data using 0.34 as cut-off.

Making Predictions on Test Data:

● 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:

- What is your current occupation_Working Professional
- Last Activity_Had a Phone Conversation
- Lead Origin_Lead Add Form

Recommendations:

● More budget/spending 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.