SUMMARY – LEAD SCORING CASE STUDY

X Education gets a lot of leads, 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 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 appropriate action: if imputation causes skew, then columns were dropped, created new category (others), impute mode, drop columns that don't add any value to the model.

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

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

Exploratory Data Analysis (EDA):

● Data imbalance checked- only 38.5% leads converted.

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

● Time spend on website shows positive impact on lead conversion.

Data Preparation:

● Created dummy features for categorical variables.

● Splitting Train & Test Sets: 70:30 ratio.

● Feature Scaling using Standardization.

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

Model Building:

● Used RFE to reduce variables from 48 to 20. This will make dataframe more manageable.

● Manual Feature Reduction process was used to build models by dropping variables having p – value.

● Total 7 models were built before reaching final Model 8 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.

● logm8 was selected as final model with 13 variables, we used it for making prediction on train and test set.

Model Evaluation:

● Confusion matrix was made and cut off point of 0.346 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.

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

Making Predictions on Test-Data:

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

● Evaluation metrics for train & test are very close to around 80%.

● Lead score was assigned.

● Top 3 features are:

- Lead Source_Welingak Website

- Lead Source_Reference

- Current Occupation_Working Professional

Recommendations:

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

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

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