Summary

X Education gets a lot of leads there are a lot of columns which have high number of missing values. Clearly, these columns are not useful. Since, there are 9000 datapoints in our data frame, let's eliminate the columns having greater than 3000 missing values as they are of no use to us. CEO's target for lead conversion rate is around 80%.

- There are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that column which is why it shows 'Select'. To get some useful data we must make compulsory selection. Likewise, Customer occupation, Specialization, etc.
- Columns with >40% nulls were dropped. Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
- Numerical categorical data were imputed with mode 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.
- Data imbalance checked- only 38.5% leads converted.
- There are a lot of leads in the initial stage but only a few of them are converted into customers who can pay for the course offered to them. The most numbers of leads are from INDIA and in terms of city highest number are from Mumbai
- Created dummy features (one-hot encoded) for categorical variables
- Feature Scaling using Standardization
- Dropped few columns, they were highly correlated with each other
- Used RFE to reduce variables from 48 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.
- 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.
- Confusion matrix was made and cut off point of 0.345 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%.
- Most of leads current occupation is Unemployed, which means gave more focus on unemployed leads and to convert the leads in customers employment prospects should be presented with importance to them.

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.
