Summary

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:

- Treated placeholder terms as null values, as they provide the same type of information.
- Dropped those columns which have more than 40% null values.
- Replaced all null values with 'Not Specified' for the columns having null values between 5% and 40%.
- Dropped null values for null values less than 5%.
- If a column stands out to be not providing information, we will drop the column. Numerical categorical data will be imputed with mode and columns with only one unique response from customerwere dropped.
- Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

EDA:

- Data imbalance checked- only 38.5% leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables. 'Tags', 'Current Occupation', 'Last Known Activity', etc. provide valuable insight on effect on target variable.
- Time spend on website shows 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

Model Building:

- Used RFE to reduce variables from 48 to 15. This will make dataframe more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 5 models were built before reaching final Model 6 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
- lm6 was selected as final model with 10 variables, we used it for making prediction on train and test set.

Model Evaluation:

- Confusion matrix was made and cut off point of 0.35 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 92%.
- 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 classification.
- Lead score was assigned to train data using 0.35 as cut off.

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 92%.
- Lead score was assigned.
- Top 4 features are:
    - Tags_Closed by Horizzon
    - Tags_Lost to EINS
    - Tags_Will revert after reading the email
    - What is your current occupation_Not Specified