Summary

X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. **A model is required to be built wherein a lead score is assigned** to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead **conversion rate to be around 80%.**

Data Cleaning:

- Columns with >40% nulls were dropped. Value counts within categorical columns were checked to decide appropriate action: if there is null value we have imputed with mode value.
- Imputing numerical columns with median
- Other activities like checking for unique values, Checking the value counts in each column for the data distribution and data imbalance.

EDA:

● Checked the co-relation between the numeric columns using Heatmap and Pairplot.
● Checked outliers and done outlier treatment by capping them.
● Done Categorical analysis using countplot.

Data Preparation:

● Created dummy features for categorical variables.
● Splitting Train & Test Sets with 70:30 ratio.
● Feature Scaling using Standardization

Model Building:

● Used RFE to reduce variables from 30 to 18. This will make dataframe more manageable.
● Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
● Total 12 models were built before reaching final Model 13 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
● Model 13 was selected as final model with 18 variables, we used it for making prediction on train and test set.

Model Evaluation:

● Confusion matrix was made and cut off point of 0.299 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all above 80%. Whereas there was a slight drop in precision recall view and gave less performance metrics when cut off point of 4.2 was selected.
● Lead score was assigned to train data using 0.299 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 80%.
● Lead score was assigned.
    ● Top 3 features are:
        Tags_Closed by Horizzon
        Tags_Lost to EINS
        Last Notable Activity_Had a Phone Conversation

Recommendations:

- The company should make calls to the leads coming from the `Tags_Closed by Horizzon` , `Tags_Lost to EINS` `Last Notable Activity_Had a Phone Conversation`, `Lead Origin_Lead Add Form `, `Last Notable Activity_Email Bounced ` and `Last Notable Activity_SMS Sent` as these are more likely to get converted.
- The company should make calls to the leads who are the `working professionals` as they are more likely to get converted.
- The company should make calls to the leads coming from `Lead Source_Welingak Website` and who spent `more time on the websites` as these are more likely to get converted.