Summary

X Education, which sells online courses to industry professionals, gets a lot of leads, although its lead conversion rate is poor around 30%. The company requires us to build a model where 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%. To build the model we employed the following approach.

Data Preparation:

- Columns with yes/No converted to its binary equivalent, for yes 1 and for no 0
- Converted values with "SELECT" in all the categorical columns to NaNs. Since user didn't select any option from the list, therefore SELECT is good as NaN.
- Handled missing values by dropping all the columns having more than 70% of missing value, later dropped columns having more than 40% missing values.
- Imputed the column with Nulls values with the most frequent values. For example for City column most of the value belong to Mumbai, hence imputed NaNs in that column to Mumbai
- For columns with around 1% or less number of NaNs, we simply dropped the rows.

EDA:

- Checked the Data Imbalance for target value (lead converted). Around 38% of leads converted.
- Performed Univariate and Binvariate analysis for categorical and numerical variable. Columns such as "Lead Origin", "Lead Source", "Total time spent on Website" provided much better insights into the data and behaviour of leads. "Tags" also played the crucial role in predicting the lead action or kind (hot or cold).

Dummy Variable Creation:

- Created dummy variable for categorical variable such as "Lead Origin", "Lead Source", "Tags", "Lead Quality" etc.
- Dropped the unnecessary and redundant columns after creating dummy variable

Test Train Split:

- Divided the dataframe into X dataset (without target variable and Prospect ID column) and y dataset (only contain target variable)
- Split X and y dataset into X train, X test, y train and y test dataset for model building.

Feature Scaling:

- Scaled the feature using the standard scaler.

Correlation Heatmap Analysis:

- Performed the correlation analysis between variables/features and dropped all the columns with high correlation values.

Model Building:

- Added the constant to the training dataset and used the RFE approach to select the top 15 feature.
- With manual elimination dropped the feature with high p-values. Checked the accuracy of the model.

Model Evaluation:

- In our case, sensitivity holds the utmost importance, therefore we tried to keep the sensitivity as high as possible.
- To find the optimal cutoff value, we plotted the ROC curve and line graph for accuracy, sensitivity, and specificity. And finalized the cutoff value of 0.27.
- Finalized the model with the optimal cutoff values of 0.27 with around 93% sensitivity of test dataset.
- Assigned the lead score to each lead.
- Found the following top features:
    1. Tags_Lost to EINS with importance of 9.4
    2. Tags_Closed by Horizzon with importance of 8.5
    3. Tags_Will revert after reading the email with importance of 3.7

Recommendations:

- Focus on the lead having high potential based on the lead score.
- Leads who had last activity as “SMS sent” can be focused as they have higher conversion rate
- In order to enhance the overall rate at which leads are converted, efforts should be directed towards improving the conversion rates of 'API' and 'Landing Page Submission' Lead Origins, as well as increasing the quantity of leads originating from the 'Lead Add Form' channel.
- Focus is required on leads tagged with "Will revert after reading the email" as they have high count and conversion rate is also good. There are chances that customer forgets to read the email, therefore can be approached or give a reminder to increase the conversion rate.
- Website could be made more attractive, with fun-quality content and interactive component might to increase the time spent by users on website.
- Since, lead count of Google, Direct Traffic, Olark Chat, and Organic search is high. Hence more focus is required in regards of conversion rate. Also Reference and Welingak Website have pretty good conversion, more attention is required to increase its count.