Summary

The problem statement and database of X education itself were highly insightful to navigate the best possible solution/ Model of the problem as the company CEO wants to increase the conversion rate extensively from 30% to 80%.

Strategy followed:

1. **Data Cleaning:**
    - We dropped columns with more than 45% missing data
    - Other categorical columns where missing data is less than 45%, we imputed with another category as 'Missing' so can be used in analysis
    - There were few columns where 'Select' value was then instead of specific category, as this no meaning of that column and considered as NULL value only and treated as Missing/Null only.
    - For Numerical columns we used median of data to impute missing values.

2. **Outliers Treatment:**
    - Few numerical columns had outliers we used them treated them with proper investigation and removed those rows with suited quantiles.

3. **EDA:**
    - We performed EDA also on many features and treated those columns accordingly like there were some features where majority of the rows with only 1 value and very less/No rows with other values, as these columns also not usable in analysis we dropped them too with the help of EDA.

4. **Model Building:**
    - For Model building we used 'sklearn' and 'statsmodels' both modules
    - We divided data in 70% training set and 30% in test set
    - Created dummy variables for all categorical columns
    - We used MinMax Scaler for scaling of numerical columns
    - We started with automatic feature selection RFE with 15 features
    - And with those selected features we started building model with statsmodel
    - And based on p-value and VIFs checks we finalized one model which had 11 features with all having less than 0.05 p-value and VIFs less than 5
    - After this we evaluated model with both Accuracy, Sensitivity & Specificity which was ~91% on both training and test sets
    - With this trade-method optimum cut-off arrived is 0.3
    - In Precision-Recall trade-off also we achieved very good values of Precision, Recall and F1-Score and all values are ~90%, and optimum cut-off arrive 0.4.
    - ROC Curve value also we got 0.97 which indicates very good predictive value.

5. Conclusion:

- So, after an entire analysis we concluded that following 11 features affecting the most for lead conversion. (Listed in Descending Order of their Co-efficient)
    - Tags_Closed by Horizzon
    - Tags_Lost to EINS
    - Lead Source_Welingak Website
    - Tags_Will revert after reading the email
    - Total Time Spent on Website
    - Last Activity_sent SMS
    - Last Notable Activity_Modified
    - Page Views Per Visit
    - Tags_Ringing
    - Lead Profile_Student of SomeSchool
    - Tags_switched off

- With this Model we can predict Conversion Rate in a better way and X Education company can utilize this and identify potential buyers and can arrange call back and convince them to buy their course.
- And this way X Education company can improve their Lead Conversion Rate and achieve provided target of 80% Conversion Rate through this model.