Summary

This analysis has been done for X Education to find ways to entice more business professionals to enroll in their courses. The dataset provided gives us a wealth of details about how potential customers approach the site, how long they stay there, how they leave, as well as the conversion rate.

The steps that have been followed are:

- **Data Cleaning**:

  - Columns containing over 40% null column values were removed. Value counts within categorical columns were reviewed to determine the best course of action: eliminate the column, create a new category (others), impute high frequency values, and drop columns that don't contribute any value if imputation creates skew.

  - Converted all values having 'Select' in the data with NaN.

  - Median was imputed to numerical categorical data, and columns with just one unique value were removed.

  - Other tasks including handling outliers, correcting inaccurate data, grouping low frequency values, and mapping binary category values have been performed.

- **Exploratory Data Analysis**:

  - Found that most of the variables are categorical. Several categorical variables were found to be irrelevant. Numerical variables had some outliers.

  - Univariate analysis was performed on both numerical and categorical variables.

  - Bivariate Analysis was done on Target Variable.

- **Data preparation:**
    - Converted Binary variables into 0 & 1
    - Created Dummy Variables for categorical variables.
    - Data was split in 70:30 ratio of train-test set.
    - Feature Scaling was done.
- **Model Building:**
    - Used RFE to number of variables from 57 to 15.
    - Manual Feature Reduction process was used to build models by dropping variables with p - value > 0.05. Total 2 models were built which were stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
    - Model 2 was selected as final model with 14 variables, used it for making prediction on train and test set.
- **Model Evaluation:**
    - Confusion matrix was made and cut off point of 0.325 was selected based on accuracy, sensitivity and specificity plot. This ROC cut off gave accuracy, specificity and sensitivity all around 98%.
- **Prediction:**
    - Prediction was done on the test data frame and with an optimum cut off as 0.325 with accuracy, sensitivity and specificity of 98%.
    - It was found that the variables that mattered the most in the potential buyers are (In descending order):

- Tags_Lost to EINS
- Tags_Closed by Horizzon
- Lead Source_Welingak Website
- Tags_Will revert after reading the email
- Last Activity_SMS Sent
- Total Time Spent on Website
- Last Notable Activity_Modified
- Tags_Busy
- Last Activity_Email Bounced
- Last Notable Activity_Olark Chat Conversation
- What is your current occupation_Unknown
- Tags_Interested in other courses
- Tags_Other Tags
- Tags_Ringing