Summary

After comprehensively analyzing the provided data from X Education, a multitude of insights have been gleaned regarding customer behavior on the website. This analysis shed light on various aspects, including the frequency of customer visits, their duration on the site, and the diverse channels through which they reach the website.

The model was constructed using the following measures.

1. **Importing and knowing the data**

   - Firstly, the problem statement and business problem were comprehensively understood. Essential libraries were imported.

   - Additionally, the data provided by X Education was imported, and insights were obtained through descriptive statistics, information summaries, shape examination, and the head function.

2. **Data Cleaning:**

   - Handling missing values involved dropping columns with high rates of missing data and eliminating columns/features deemed irrelevant for the analysis.

   - Detecting outliers was conducted to ensure data quality and reliability.

3. **EDA (Exploratory data analysis)**

   - Identify the various classes of specific features using the value_counts() method.
   - Visualize relationships among numerical variables using pair plots.
   - Use box plots to visualize relationships between categorical variables and the dependent variable.
   - Determine the frequency distribution of variables through histograms.
   - Perform both univariate and bivariate analyses.

4. **Data visualization**

   - Analyze and visualize numerical variables through pair plots.
   - Display the frequency distribution of all categorical values using histograms in subplots.
   - Visualize the relationship between categorical variables and the dependent variable using box plots in subplots.
   - Visualize the relationship between categorical variables and the dependent variable using box plots in subplots.

5. **Dummy Variables:**

   - Create dummy variables for categorical features containing more than two classes.
   - Aggregate these dummy variables into a unified dataframe, and

subsequently append it to the original dataframe.

- Eliminate the original variables for which dummy variables have been generated.

6. Scaling

- Used min max scaler from sklearn to scale down the value of those feature
  whose values are not in the form of zeros and one.

7. Train-Test split:

- Initially, divided the data into ‘x’ and ‘y’ where x is independent variable and yis dependent variable.
- Further the x and y were split.
- The split was done at 70% and 30% for train and test data.

8. Model Building:

- Split the data into ‘x’ and ‘Y’ (independent and dependent variable).Again split ‘x’ and ‘y’ into train and test dataset.
- I have chosen 70-30 ratio for splitting; 70 percentage is train part and 30percentage is test part.
- Used min max scaler to scale down the values for those variables whosevalues are not in the form of zero and one.
- Used RFE feature with an output of top 15 selected feature
- Made prediction on train data first and ultimately predicted on test data.
- I repeatedly made models to improve the accuracy and the fifth model was the final model with desired accuracy as there is a multicollinearitybetween the dependent variable and with a high ‘p’ value.

9. ROC and exact cut-off value

- Found region under curve is 0.86. It looks like good model. But need to checkthe in cut off range
- As we found around 0.41 is optimal value where all the three metrics intersectbetween each other (accuracy, sensitivity, and specificity)

10. Model Evaluation and prediction:

- I have used a classic list comprehension method to create a dataframewith variables named actual, predicted, and converted predicted.
- I have found cut-off being 0.4 and decided to cut the prediction value.
- My final model accuracy score on test data is 0.83
- sensitivity (true positive rate) = 0.81
- Specificity (true negative rate) =0.89

11. Conclusion

Found that the variables which are having negative and positive impact on lead conversion.POSITIVE VARIABLES used for lead conversion were

- Total Time Spent on Website
- Lead Origin_LeadAdd Form
- Lead Source_OlarkChat
- Tags_OtherReasons
- Tags_Willrevert after reading the email
- Lead Profile_PotentialLead
- Lead Profile_diploma_dual_n_SomeSchool_lead
- Last Notable Activity_Other_Activity
- Last Notable Activity_SMSSent

12. We should keep above aspect in mind. X Education can improve by seeing their negative and positive variables. By using keeping features in mind X education canachieve their goal.

THANK YOU