Summary

A certain analysis has been done by the given data of X education. Understood the data and came to know varieties of insights on all the features. It gave a lot of information about the customers visits the website, how much time they are spending and how they are reaching the website by various means.

The following are the measures used by building the model:

1. **Importing and knowing the data**

   - ✓ First, understood the problem statement and business problem. Imported all the important libraries. Also imported the data provided by X education and knowing the insights of the data by using describe info, shape, and head function.

2. **Data Cleaning:**

   - ✓ Treating the missing values by dropping the columns which have high rate of missing values and dropping those columns/features which aren't important for the analysis
   - ✓ Check the outliers

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

   - ✓ Finding the different classes of certain features by using value_counts()
   - ✓ Visualization among the numerical variables (by using pair plot)
   - ✓ Visualization among the categorical variables with respect to dependent variable (by using box plot)
   - ✓ Find the frequency of distribution among the variables (by using histogram)
   - ✓ Univariate and bivariate analysis

4. **Data visualization**

   - ✓ Analysis and visualizing numerical variable by using pair plot
   - ✓ Frequency distribution of all categorical values by using histogram in sub plot
   - ✓ Visualization of categorical variable vs dependent variable using box plot in sub plot
   - ✓ Visualization of categorical variable vs dependent variable using box plot in sub plot

5. **Dummy Variables:**

   - ✓ Create a dummy variable for those categorical features which is having more than two classes.
   - ✓ Stored all these dummy variables into one data frame. Finally concatenated with original data frame.
   - ✓ Deleted the original variable whose dummy variable is created

6. **Scaling**

   - ✓ Used min max scaler from sklearn to scale down the value of those features

whose values are not in the from of zeros and one.

7. **Train-Test split:**

- ✓ Initially, divided the data into ‘x’ and ‘y’ where x is independent variable and y is 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 30 percentage is test part.
- ✓ Used min max scaler to scale down the values for those variables whose values 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 multicollinearity between 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 check the in cut off range
- ✓ As we found around 0.42 is optimal value where all the three metrics intersect between each other (accuracy, sensitivity, and specificity)

10. **Model Evaluation and prediction:**

- ✓ I have used a classic list comprehension method to create a dataframe with 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 0.78
- ✓ sensitivity (true positive rate) = 0.78
- ✓ Specificity (true negative rate) =0.79
- ✓ Recall=0.78
- ✓ Precision=0.77
- ✓ F1 score is 0.78

Conclusion

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

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was Welingak website and Olark Chat
4. When the last activity was SMS and phone conversation
5. When the lead origin is Lead Add Form.

NEGATIVE VARIABLE which has a negative impact on lead conversation was If the current occupation is student and unemployed.

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 can achieve their goal.