Summary

To solve the given problem of X Education Company, we performed Logistic Regression on the given dataset. We followed below steps to generate score against each lead.

Step1: Importing the Data
Imported data and libraries.

Step 2: Inspecting the Data Frame
Initially we had 9240 rows and 37 columns in the dataset. Out of 37 variables 30-object,3-integer and 4-float datatypes. We also have few null values columns.

Step 3: Data Cleaning and Preparation

The columns with more than 30% of null values are dropped and others are treated by dropping some of the unnecessary columns and by removing null values rows alone from the columns for important columns as they will contribute more for our analysis.

Categorical columns 'Lead Profile', 'How did you hear about X Education' and 'Specialization' have level 'Select' . columns 'Lead Profile', 'How did you hear about X Education' columns dropped because of high number of 'Select' labels. For 'Specialization' feature, only rows with 'Select' are removed from the data set. After cleaning the data, we had 12 columns with 6373 rows.

Step 4: Dummy Variable creation
For all the categorical columns, dummy variable created. For column 'Specialization', dummy column with 'Specialization_Select' is dropped.
Outliers are checked using boxplot not treated at this stage as it will get normalized in feature scaling.

Step 5: Test-Train Split
Prepared X and y variables by considering 'Converted' column as target variable and all other variables as independent variables. Splitting the data into train and test data with 70% and 30% respectively. Considered random state as 100.

Step 6: Feature Scaling
We have taken minmax scaling for normalizing the numeric values.

Step 7: Checking Correlation
Multicollinearity check is done using corr() function, it will get eliminated using RFE.

Step 8: Model Building
Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).A logistic regression model was built using the function GLM() under statsmodel library. Hence, some of these variables were removed first based on an automated approach, i.e. RFE and then a manual approach based on the VIFs and p-values are used for further process.

	coef
const	0.2040
TotalVisits	11.1489
Total Time Spent on Website	4.4223
Lead Origin_Lead Add Form	4.2051
Lead Source_Olark Chat	1.4526
Lead Source_Welingak Website	2.1526
Do Not Email_Yes	-1.5037
Last Activity_Had a Phone Conversation	2.7552
Last Activity_SMS Sent	1.1856
What is your current occupation_Student	-2.3578
What is your current occupation_Unemployed	-2.5445
Last Notable Activity_Unreachable	2.7846

Fig:1 Final model features with coefficients

Step 9: Model Evaluation

A confusion matrix was made from the model. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity.

Step 8: Plotting the ROC Curve

When we plotted the true positive rate against the false positive rate, and got a graph which showed the trade-off between them and this curve is known as the ROC curve. The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model.

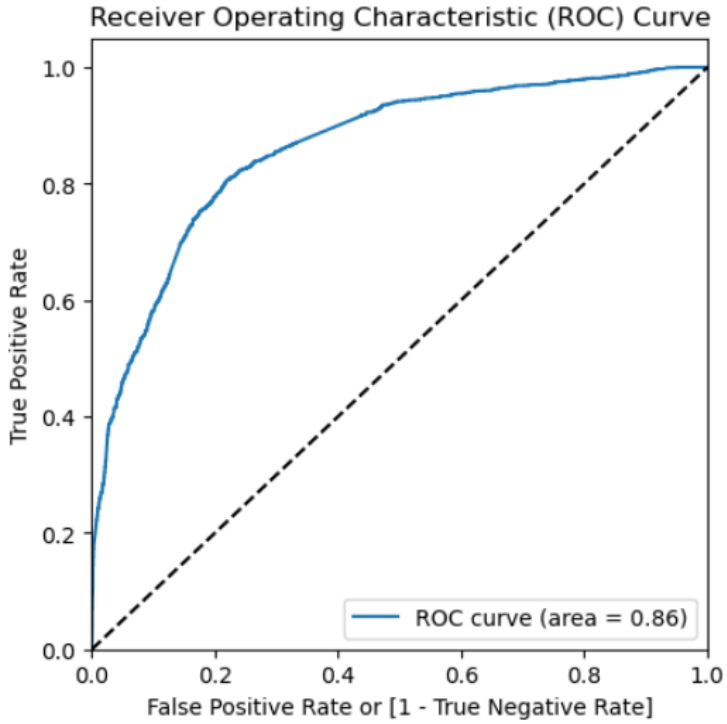

Fig2: ROC Curve

Step 9: Finding Optimal Cutoff Point

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity**.** The optimal cut-off for the model was around 0.42 and we chose this value to be our threshold and got decent values of all the three metrics – Accuracy (79.08%), Sensitivity (79.34%), and Specificity (78.85%).

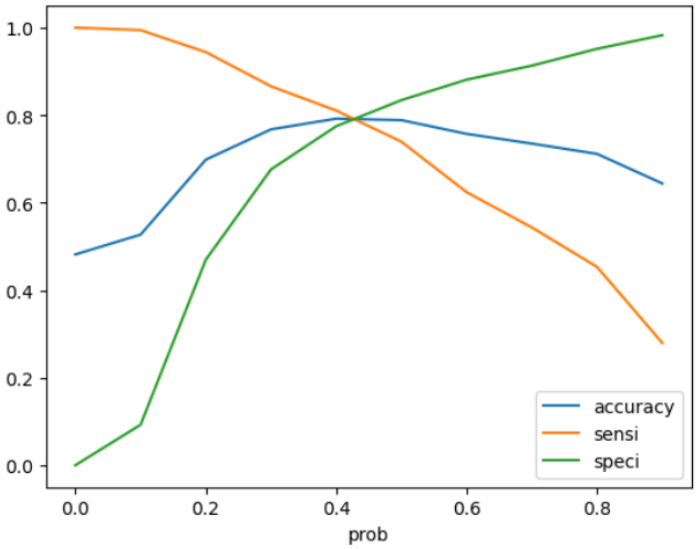

Fig3: Optimal Cutoff

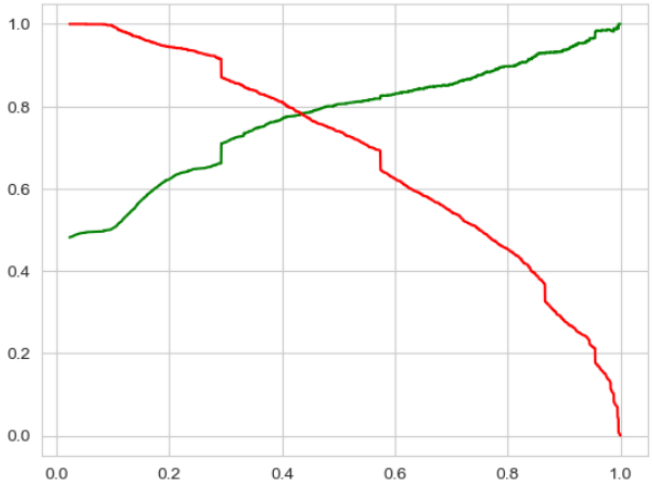

Fig4: Tradeoff curve between Precision and Recall

Step 10: Making predictions on the test set

We also have checked precision and recall which was another pair of industry-relevant metric used to evaluate the performance of a logistic regression module. We have chosen a cut-off point of 0.42, made predictions on the test dataset and have obtained a descent values for metrics such as

1. Accuracy: 78.45%
2. Sensitivity (Recall): 77.94%
3. Specificity: 78.92%
4. Precision:77.27%

Step 11: Lead Score Assigning

We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads. There are 932 leads who have high chance of getting converted. So company can contact them for further information.

Conclusion:

Learning gathered below:

- Test set is having accuracy, recall/sensitivity in an acceptable range.
- In business terms, our model is having stability and accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.
- Top features for good conversion rate:
    1) **TotalVisits**
    2) **Total Time Spent on Website**
    3) **Lead Origin_Lead Add Form**