Summary

This analysis is carried out for X Education in an effort to attract more business professionals to their courses. We learned a lot from the fundamental data on how potential customers use the site, how long they stay there, how they got there, and the conversion rate.

The following are the steps used:

1. Cleaning data:

   With the exception of a few null values, the data was mostly clean. The choice chosen had to be replaced with a null value because it provided little useful information. The "Select" value was replaced with NAN. Calculation of missing values for each column was done and dropped the columns with a high percentage of missing values. Checked the unique category for columns. If the columns are highly skewed with one category, such columns were also dropped. Combined different categories of the columns with fewer percentage values into another category and named it "Others". Input the no of columns with the least missing values percentage. Finally checked for the number of rows kept after performing all the above steps.

2. EDA:
   To quickly assess the state of our data, an EDA was performed. It was discovered that several of the categorical variables' components were unnecessary. The numerical results seem accurate, and no outliers were discovered.
   - We also performed univariate analysis on a categorical column to see which columns make more sense and removed those columns whose variance is nearly zero.
   - We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
   - We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
   - We performed bivariate analysis on numerical columns with Converted columns to see how the leads are related to these columns.
   - We have used the IQR method to treat the outliers in the data set.

3. Dummy Variables:
   - At this stage, most of our data were clean and had no outliers. We know that logistic regression takes the ut parameters as numerical values. Hence, we converted all the categorical columns to numerical ones.
   - Columns which have only two levels "Yes" and "No" were converted to numerical using binary mapping.
   - Columns which have more than two levels were converted to dummies using pd.get_dummies function.
   - Now, the data contained only numerical columns and dummy variables. Before proceeding with model building, we rescaled all numerical columns by using the standard Scaler method.

4. Model Building:

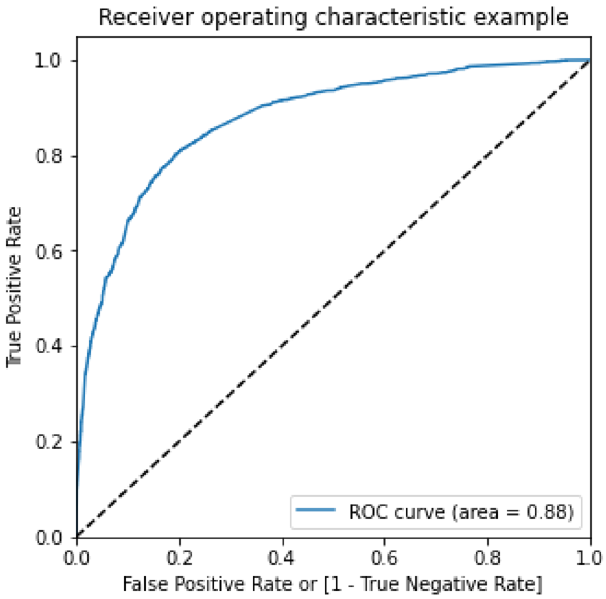
We have used Recursive Feature Elimination Technique to remove attributes and built a modelon those attributes that remain. RFE uses the model accuracy to identify which attributes (andcombination of attributes) contribute the most to predicting the target attribute.

First, the top 15 pertinent factors were determined by RFE. Later, based on the VIF values and p-value, the remaining variables were manually deleted (the variables with VIF <5 and p-value <0.05 were retained).

In this step, we made the model stable by using the stats library, where we checked the p-values tobe less than 0.05 and VIF values to be under 5. Variance inflation factor(VIF) is used to treat the multi-collinearity.

Once the stable model was created, we predicted probabilities on the train set and created anew column predicted with 1 if the probability is greater than .5 else 0.

We calculated the confusion matrix on this predicted column to the actually converted column.We also calculated the metrics sensitivity, specificity, precision, recall and accuracy. We also plotted the roc curve to find the area under the curve.

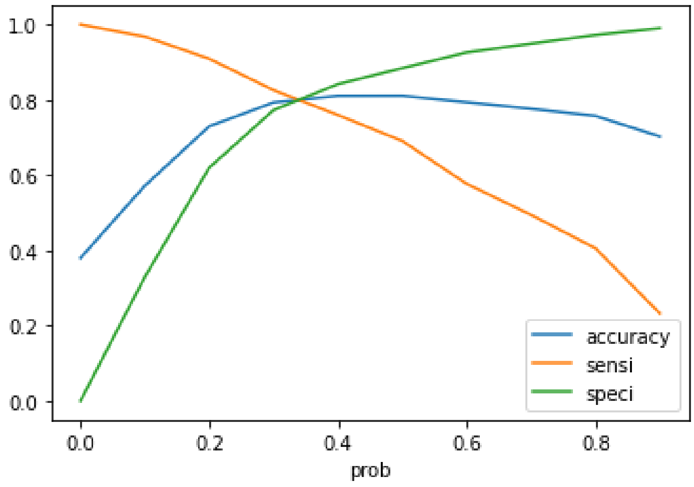


5. Model Evaluation:

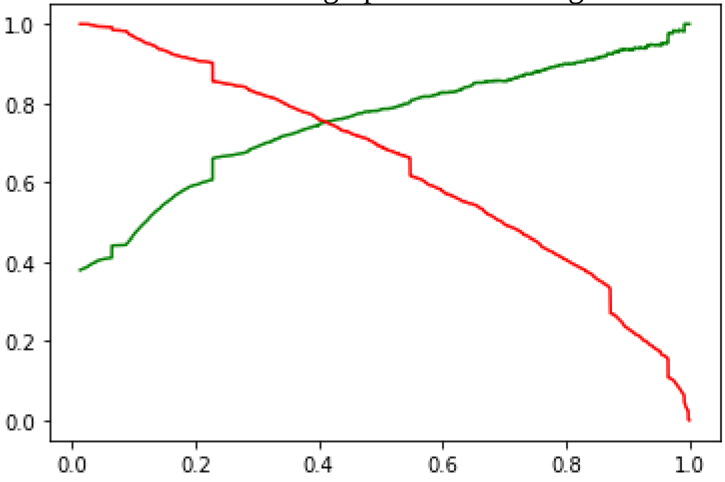
In step 5 we took 0.5 as the cut-off. To confirm that it was the best cut, we calculated theprobabilities with different cut-offs.

With probabilities from 0.0 to 0.9, we calculated the 3 metrics -accuracy, sensitivity andspecificity.

To make predictions on the training dataset, an optimum cut-off of 0.35 was found from theintersection of sensitivity, specificity and accuracy as shown in below figure:



To make predictions on the test dataset, the optimum cut-off was considered as obtained from the Precisionrecall graph of the training dataset as shown below figure:



We can observe that 0.4 is the trade-off between Precision and Recall. Thus, we can safely choose toconsider any Prospect Lead with a Conversion Probability higher than 40 % to be a positive Lead

6. Prediction:

   After finalizing the optimum cut-off and calculating the metrics on the train set, we predicted the data on the test data set.

   Below are the observations:

   Train Data:
   Accuracy: 80.5%
   Sensitivity: 79.4%

Specificity: 81.1%

Test Data:
Accuracy: 80.88%
Sensitivity: 76.1%
Specificity: 83.77%

7. Conclusion

The model seems to predict the conversion rate very well and we should be able to give the CEO confidence in making good calls based on this model.

- We must majorly focus on working professionals.
- It's always good to focus on customers, who have spent significant time on our website.
- More time spent on the website was when the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
    e.
- They must majorly focus on leads whose last activity is an SMS sent or Email opened.
-

Keeping these in mind, X Education can prosper since they have a very high possibility of persuading nearly all prospective customers to purchase their courses.