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. To prevent losing a lot of data, some of the null values were changed to 'not provided'. Nevertheless, they were later taken out while manufacturing dummies. The elements were altered to "India," "Outside India," and "not provided" because there were a lot of people from India and a small number from elsewhere.

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.

3. Dummy Variables:
   The dummy variables were made, and those that had 'not provided' elements were afterwards taken out. The "MinMaxScaler" was employed for numerical values.

4. Train-Test split:
   The split was done at 70% and 30% for train and test data, respectively.

5. Model Building:

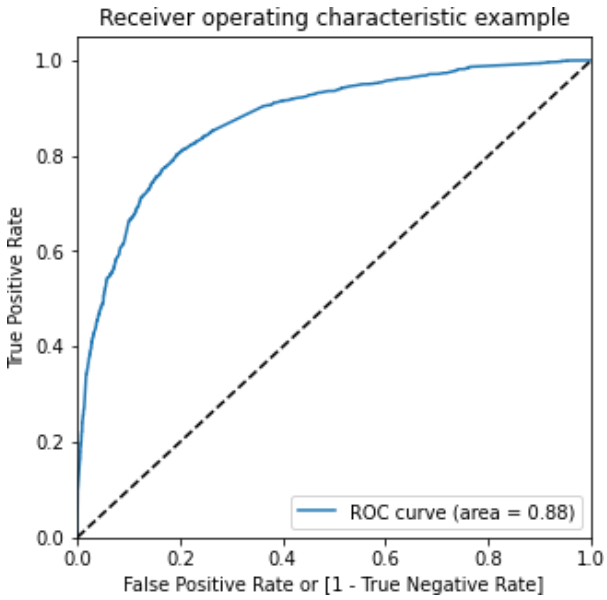
   We have used Recursive Feature Elimination Technique to remove attributes and built a modelon those attributes that remain. RFE uses the model accuracy to identify which attributes (andcombination of attributes) contribute the most to predicting the target attribute.

   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.

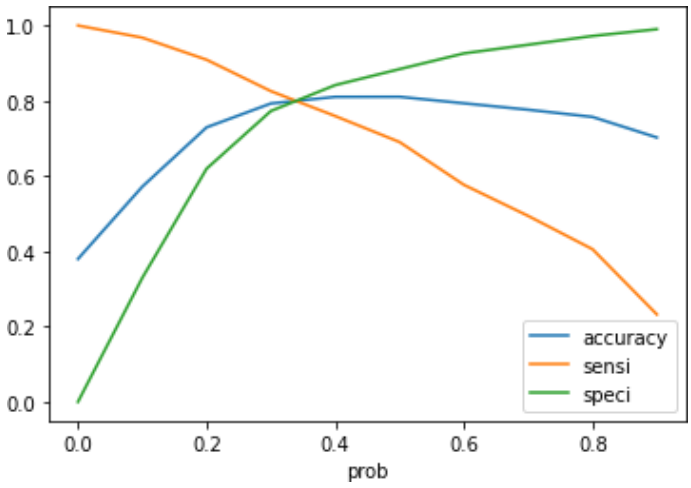


6. Model Evaluation:

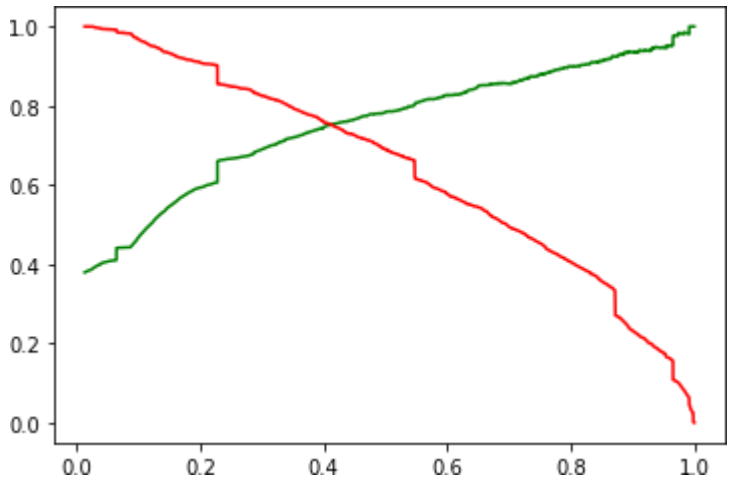
In step 5 we took 0.5 as the cut-off. To confirm that it was the best cut, we calculated theprobabilities with different cut-offs.

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

7. Prediction:
   The prediction was done on the test data frame with an optimum cut-off of 0.35 with accuracy, sensitivity and specificity of 80%.

8. Precision – Recall:
   On the test data frame, a cut-off of 0.41 was discovered with precision and recall averaging approximately 73% and 75%, respectively, using this method.

   It was found that the variables that mattered the most in the potential buyers are (In descending order):
   1. The total time spends on the Website.
   2. a Total number of visits.
   3. When the lead source was:
      a. Google
      b. Direct traffic
      c. Organic search
      d. Welingak website
   4. When the last activity was:
      a. SMS
      b. Olark chat conversation

   5. When the lead origin is Lead add format.
   6. When their current occupation is as a working professional.

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.