LEAD SCORING CASE STUDY

Problem Statement

- X-Education is an education company that providing online courses to the industrial professionals. If the people that open the website and fill up the form by giving their e-mail and phone number on any given day they are called as Leads.
- And these Leads are also available from the past referrals. Actually, the people that are joining the course are very less i.e., around 30%. So, we have to find the Hot leads (The leads that are highly interested to join the course). For this we have to build a model and find the Lead score for each and every customer. So, from this we can find the people with high score i.e., high Lead score.

The model that is used to find the Lead Score is LOGISTIC REGRESSION MODEL.
The Logistic Regression is a classification model, and this is used to make predictions if the output variable is Categorical variable.

First step is to import the data which is used to analyze.
We have to perform the normal routine check for the data

Info and shape of the dataset: number of columns, row, d-types, memory usage.
There are 37 columns, 9240 rows, dtypes: float64(4), int64(3), object(30)
memory usage: 1.6+ MB
Describe of all columns like count, mean, std, 25%, 50%, 75% percentiles etc.,

Data Cleaning

(i). Changing the rows in some of the columns from the ‘select’ to nan values.

(ii). Finding the missing value percentage and dropping the columns with high percentage. And we drop the rows that are having <= 5 missing values in all columns.

(iii). Checking the number of unique categories in each categorical column. Here we are dropping the columns in which one of the unique categories having very high.

(iV). Imputing the missing values with mean, median, mode for "continuous variables". And for the categorical variables we impute with high unique category.

Finally after data cleaning and performing the outlier treatment we are remained the data with **(7824, 20)**

(V).The percentage of rows after data cleaning process 85% of rows

Data Preparation

(i) There are some categorical columns which are having YES/NO, so we converting these to 1/0.

(ii) For categorical with multiple levels, we are creating dummies.

['Lead Origin', 'Lead Source', 'Last Activity', 'Specialization',

'City', 'I agree to pay the amount through cheque', 'Last Notable Activity']

Finally after concatenating with original data frame and by dropping the original columns after creating the dummies, we are remained with **(7824,90).**

Splitting the data to train and test with train size is 70% and test size is 30%.

Next Feature Scaling is done here we are using Standard Scaler which is used to normalize the independent data.

Building the model

RFE : Here we are running RFE with 20 variables. (Recursive Feature Elimination).

From these variables we have to select the columns that are supported by RFE.

We get 13 that are supported by RFE.

Assessing the model is done, here we will drop some of the columns depending up on the p-value and VIF values. We have to drop one column at once

We have to run the model for each column deletion, and also we have to check for

Accuracy, Sensitivity, Specificity.

- After building a model, here first we take cut-off as 0.5 for making predictions.

  For this cut-off we have the values as:

  Sensitivity is =70.7%

  Specificity is = 85.2%

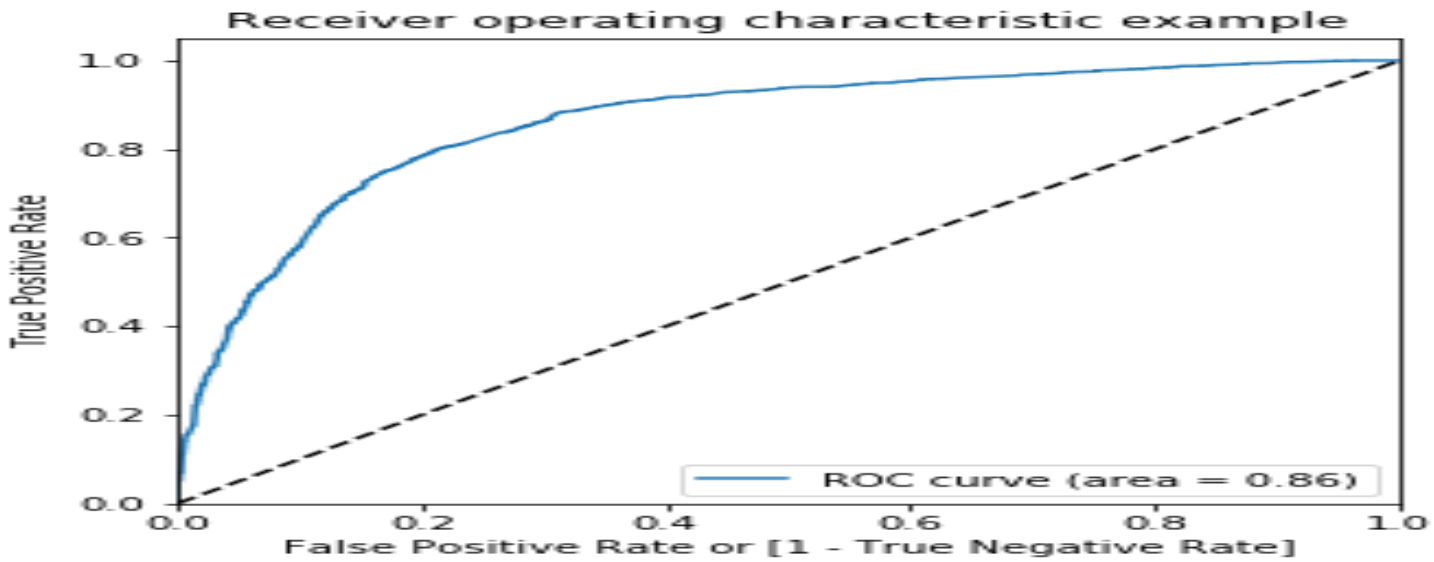
  False positive rate - predicting churn when customer does not have churned is = 14.7%

  positive predictive value is = 78.5%

  Negative predictive value is = 79.2%

ROC Curve

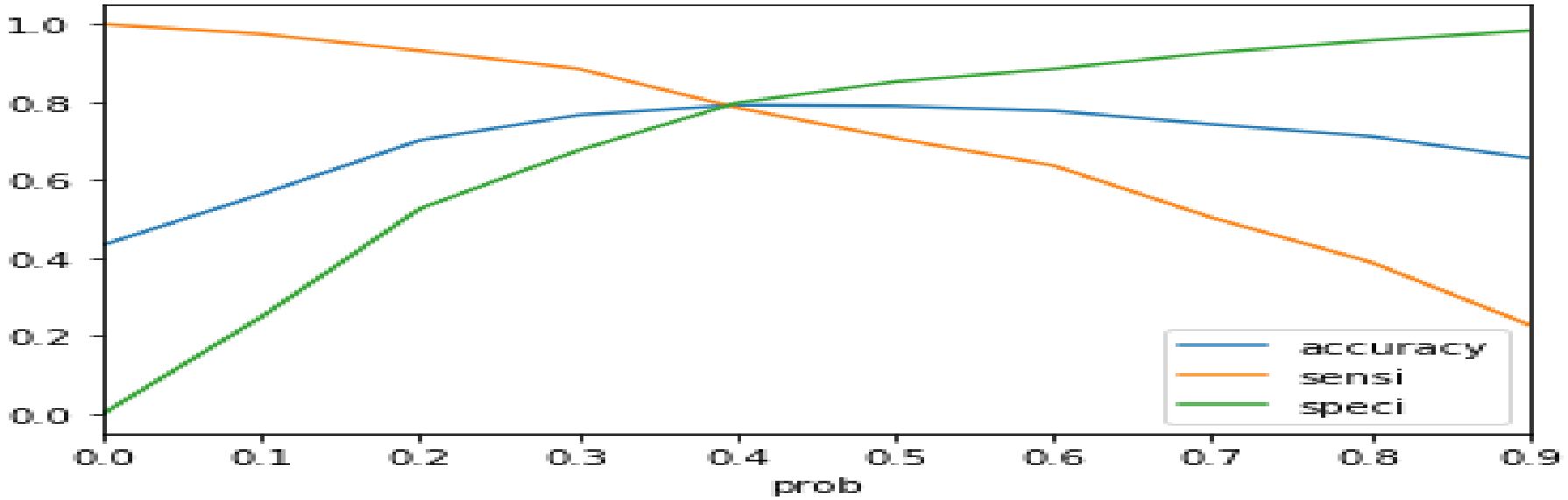
The **ROC curve** is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.



The ROC curve (area=0.86)

Optimal Cut-Off

We have to create columns with different probability cutoffs. And for this we have to find the accuracy, sensitivity and specificity for various probability cutoffs. The cut-off value from graph is **0.4**

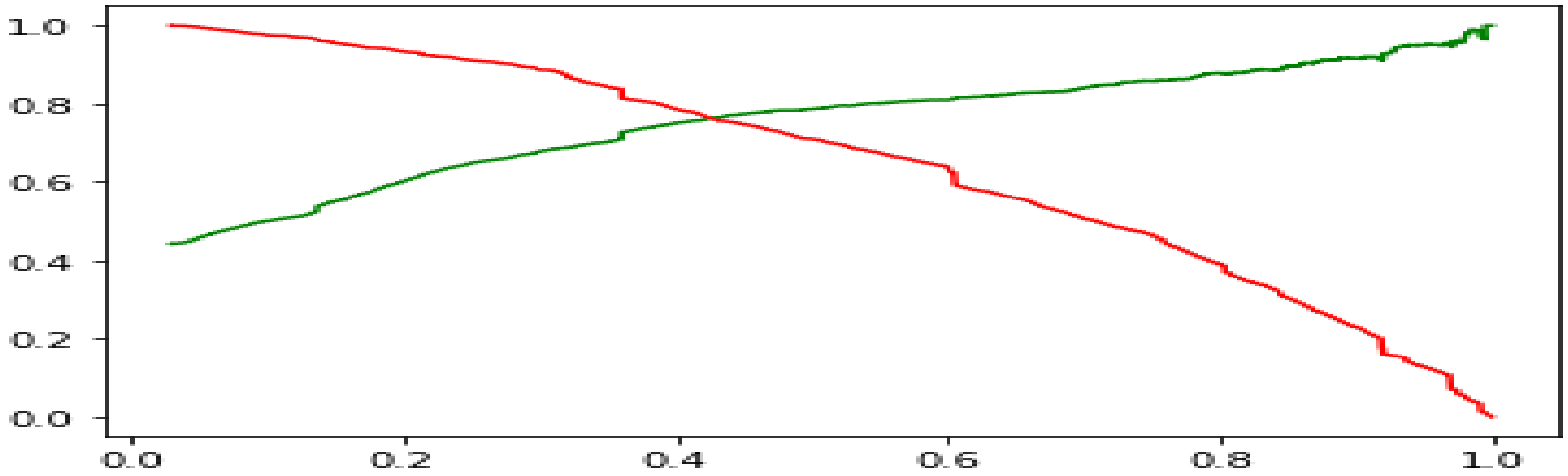


Precision and Recall Trade-off

The below curve is the precision-Recall curve.

From the curve the cut-off is around 0.43.

So, finally the predictions is made by using the cut-off as 0.43



For the final model with cut-off as 0.43 the values are as:

1. Accuracy=79.2%
2. Sensitivity=78.5%
3. Specificity 79.8%
4. false positive rate - predicting churn when customer does not have churned =20.1%
5. Positive predictive value=74.8%
6. Negative predictive value=82.9%
7. Precision= 78.5%
8. Recall=70.7%

Likewise, we have to make the predictions on the test data set. As the cut-off is also known we have to make the predictions by using it.

The Accuracy = 77.3%

Sensitivity= 73.4%

Specificity = 80.5%

The classification report is as follows.

	precision	recall	f1-score	support
0	0.79	0.81	0.80	1310
1	0.75	0.73	0.74	1038
accuracy			0.77	2348
macro avg	0.77	0.77	0.77	2348
weighted avg	0.77	0.77	0.77	2348

- Finally, we have to make the predictions for each customer i.e., the original data frame.

  To make predictions first we have to treat missing values and outliers for the original data frame.

  Also performs the dummies creation and scaling.

  We have to perform the final steps for the prediction of the Hot leads for the original data with 0.43% cut-off.

  The final model for the total data frame after finding the prediction has values for as

  Accuracy= 77.3%

  Sensitivity= 75.7%

  Specificity= 78.3%