Lead Score Case Study

Predicting the Hot leads with ML algorithm

Problem Statement

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. They need to increase their conversion rate from 30% to 80%.

What we need to do?

●X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

● The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

How score variable can help the sales team to optimise the process?

Score variable is nothing different than the probability of the Converted variable. As converted variable has only two values we can easily say it is a logistic regression problem. By considering lead scores, sales team can easily identify the hot leads that has the higher probability to be converted. In turn Company can achieve its target in time.

Data cleaning and preparation

- Removed the columns having high percentage of missing values.
- Imputed missing values with mean, median and mode
- Dropped the columns having very high weight of one category than the other
- Dropped the highly skewed columns
- Grouped the category variables in single category having highly skewed variables with many variables.
- Dropped the category variables having highly correlated with each other.
- We also scaled our numerical features but after train test split.

Model Building

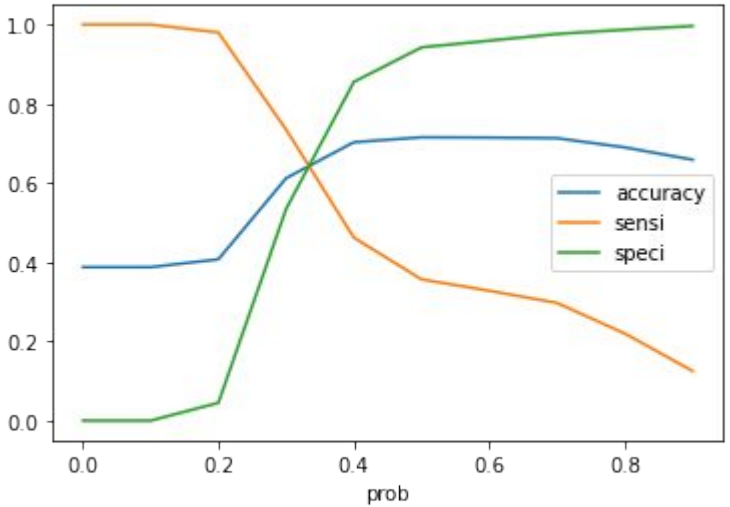
- We performed feature selection using RFE, p-value and VIF and got the final data frame.
- We first consider optimal cut off point as 0.5 for our converted probabilities. And got the accuracy of the model around 70%.
- But according to the problem statement we have to consider the performance of recall function and by using default cut off of 0.5 we only recall 35% of the data points.

		Predicted	Class
		No	Yes
Actual	No	3730: TN	228: FP
Class	Yes	1611: FN	893: TP

Fig. - Confusion Matrix (0.5 Cutoff)

Optimal Cut Off Point

- We use accuracy , sensitivity and specificity curve to find our optimal cut off point, which turns out to be 0.3.
- But using 0.3 we only recall 73% of the data points so we need to go down further and selected 0.272 as optimal cut off point.



Confusion Matrix

True Negatives(TP): if actual class says this lead did not convert and predicted class tells you the same thing.

False Positives (FP) – if actual class says this Lead did not convert but predicted class tells you that this lead will convert.

True Positives (TP) - if actual class value indicates that this lead converted and predicted class tells you the same thing.

False Negatives (FN) – if actual class value indicates that this lead converted and predicted class tells you that lead will not convert.

		Predicted	Class
		No	Yes
Actual	No	1674: TN	2284: FP
Class	Yes	494: FN	2010: TP

Fig. - Confusion Matrix (0.272 Cutoff)

Evaluation Matrices

Accuracy - accuracy is a great measure but only when we have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of our model. For our model, we have got 0.57 which means our model is approx. 57% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all leads that labeled as converted, how many actually converted? High precision relates to the low false positive rate. We have got 0.468 precision.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly converted, how many did we label? We have got recall of 0.80 which is what we need for this model.

Recall = TP/TP+FN

Precision and recall tradeoff

- you can't have both precision and recall high. If you increase precision, it will reduce recall, and vice versa. This is called the precision/recall tradeoff.
- Our model which detects a Lead will be Converted or not. The aim of the model is high recall {TP/(TP+FN)}. If model detects that lead is not converted so, lead must not be converted. So, model should reduce the false negative, which will increase the recall.
- Also we can see that cut off point is around 0.3 which is exactly same as accuracy, sensitivity and specificity curve.

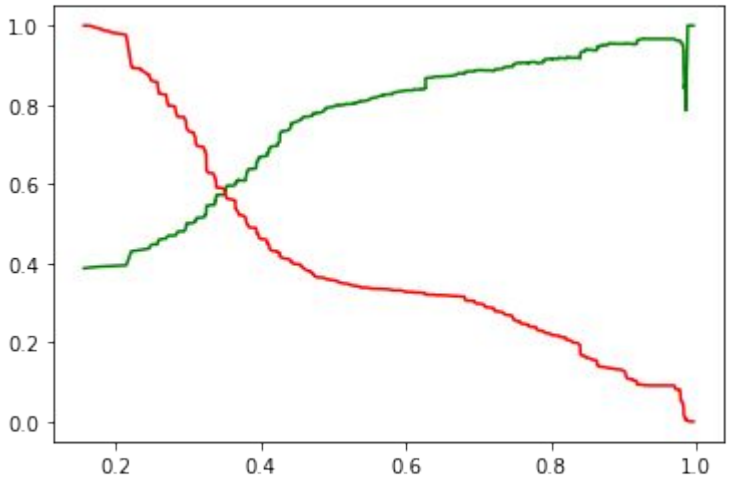


Figure: Precision Recall Tradeoff

F-1 Score

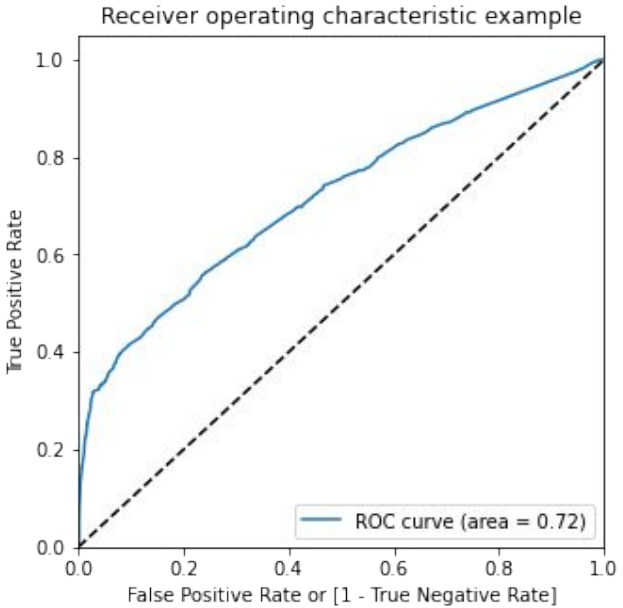
F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall. In our case, F1 score is 0.591.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

ROC Curve and AUC

An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Predictions on the Test set

Now as we trained our model on the train data set its time to test our model on the test data set and calculated evaluation matrices as following:

Precision: 0.468

Recall: 0.802

As you can see we achieved our target to get the conversion rate of 80%.

		Predicted	Class
		No	Yes
Actual	No	760: TN	954: FP
Class	Yes	210: FN	846: TP

Fig. - Confusion Matrix (0.272 Cutoff)