upGrad
#LifeKoKaroLift

Lead Scoring: Case Study

Amit Pawar
Rituparna Chakraborty

Problem Statement

- ✓ X Education is an online course selling institution to industry professionals
- ✓ Now, although X Education gets a lot of leads, its lead conversion rate is very poor, 30%
- ✓ To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’

What you need to do?

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires you 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.

Analysis Approach

upGrad

We will use Logistic regression model to find out the lead scores of the users.

For the model building we will use RFE to find out the top 15 features contributing towards

The model building process.

After the model building process, we will find out the optimal cut-off probability to find out whether the lead is converted or not based on the probability predicted by the model and compare it with actual converted to determine accuracy of the model.

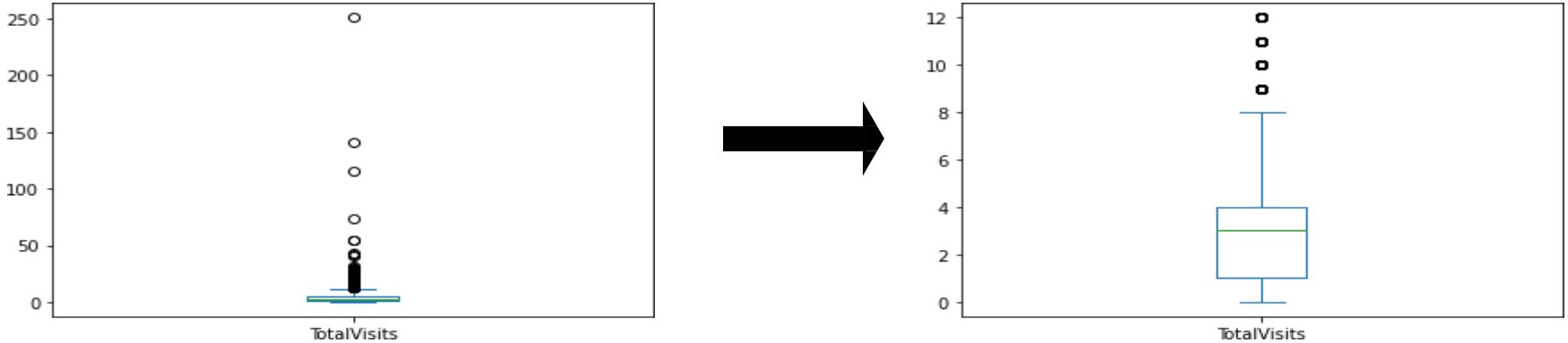
We also need to calculate the Sensitivity/ Specificity of the model and also the accuracy rate of the model.

Data Preparation

upGrad

We have carried out following steps in data preparation.

- ✓ Dropping columns with very high number of missing values.
- ✓ Imputing columns with missing values with their mean, median , mode
- ✓ Some columns found with ‘Select’ as value, so we imputed them with np.nan
- ✓ Outlier Treatment – Some columns had very high outliers . So we removed outliers from such columns.
- ✓ Multicollinearity - Check correlation heatmap to remove highly correlated variables.
- ✓ Train test split
- ✓ Feature scaling to get standardized dataset.
- ✓ Model building and then checking p values and VIF to eliminate more columns.

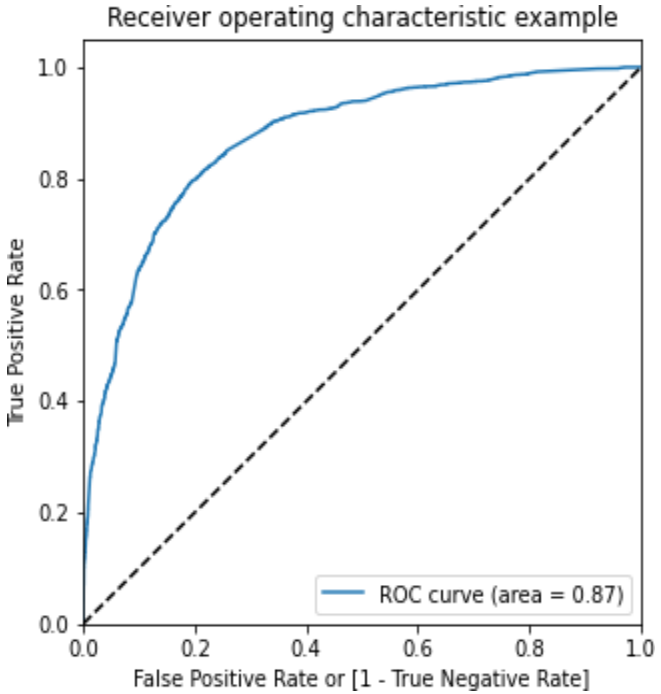


Model Evaluation

upGrad

- After the model Building, the metrics were:
  Accuracy = 80%
  Sensitivity = 67%
  Specificity = 88%

There is room for improvement here.
Lets check ROC curve



- As you can see, ROC curve is more inclined to the top left corner and AUC is 0.87 which is good for our model

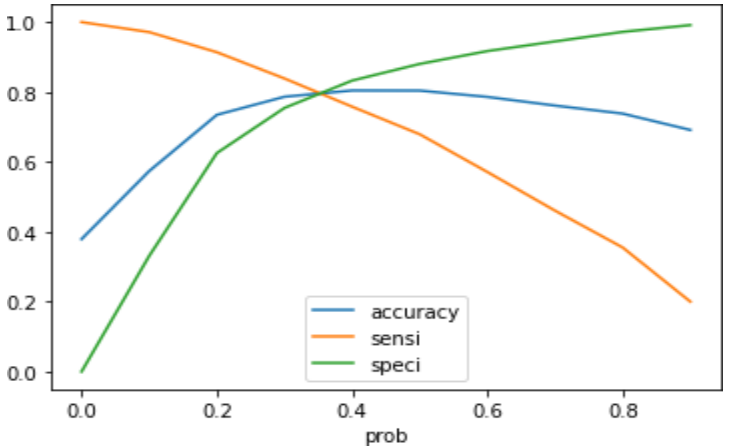
Model Evaluation

upGrad

- Now lets look at Sensitivity and Specificity for different threshold values and its graph to get optimal

threshold value

	prob	accuracy	sensi	speci
0.0	0.0	0.379126	1.000000	0.000000
0.1	0.1	0.573786	0.971404	0.330988
0.2	0.2	0.734628	0.913786	0.625228
0.3	0.3	0.786408	0.838668	0.754496
0.4	0.4	0.804369	0.758003	0.832682
0.5	0.5	0.804045	0.679044	0.880375
0.6	0.6	0.785922	0.571063	0.917123
0.7	0.7	0.761165	0.460521	0.944749
0.8	0.8	0.738188	0.355100	0.972114
0.9	0.9	0.691424	0.200171	0.991400



- From the graph and table, 0.3 looks like a good threshold value. Thus using it on our model, we get

  Accuracy = 78%
  Sensitivity = 83%
  Specificity = 75%

This are good values for our model and we can use the model on our test set.

Evaluation on Test Set and Conclusion

upGrad

- After running the model on Test set we got the following parameters:
  - Accuracy = 77%
  - Sensitivity = 85%
  - Specificity = 73%

- Overall we have achieved a very good model accuracy. We have used Sensitivity/ Specificity

approach to determine model efficiency. Similarly Precision/Recall method can also be used.

Final dataset for the client with Lead scores.

	Student	Lead_Score
0	1523	62
1	5850	11
2	3447	69
3	1011	12
4	7316	52
5	3001	75
6	1495	17
7	5777	19

THANK YOU