Lead Scoring Case Study

Goal: Logistic Regression for X Education Sell Online Course.

Step:

1. Data Cleaning
2. EDA(Checking for outlier)
3. Data Preparation
4. Train and Test Data Split
5. Model Building
6. Conclusion

Data Cleaning

Reading the data from given csv file using pandas and checking the shape and type of data.

After the following step perform for data cleaning

1. Replace ‘Select’ with na value.
2. Checked % of missing value for all column
3. Dropped the missing value where missing % is greater than 40%.
4. Dropped the categorical column where all data align with one category(like no-92.06% and Yes 7.94% avail in column)
5. Replace other missing value with mean(median also same for this data set)

EDA

Check for outlier for all continuous variable. All column don’t have any outlier. No need to perform outlier treatment for any column

Data Preparation

Following step perform in data preparation…

1. Drop the tag because it added by sales team
2. Get dummy value for 'Lead Origin','Lead Source','Last Activity','Specialization','City','A free copy of Mastering The Interview','Last Notable Activity'.
3. Merge the data set with dummy values and create new data set.
4. Drop all columns which value converted into dummy

Train and Test Data Split

Create Train and test data set for new data frame. We splitting data frame ,Train 70% and Test 30%(to verify the moel)

Model Building

Before building the model, We scale the continuous variable(TotalVisits, Total Time Spent on Website,Page Views Per Visit) using StandardScaler.

First Build the modal using RFE(with 25 columns).Then after that building model manually (By checking VIF and p-value). Manual Model build is repeated until we got good VIF and p-value.

Following are the step perform in model building..

- Build the Model using GLM.
- Check the probability value and save in Converted_Prob

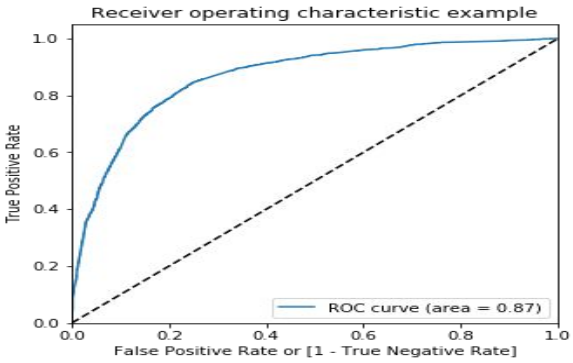
- Create new column predicted where Converted_Prob is greater than .5 then 1 otherwise 0.
- Check the accuracy_score(.80% for our dataset)
- Check VIF and dropped column if more than 5
- Create Confusion Matrix from filter column from model building

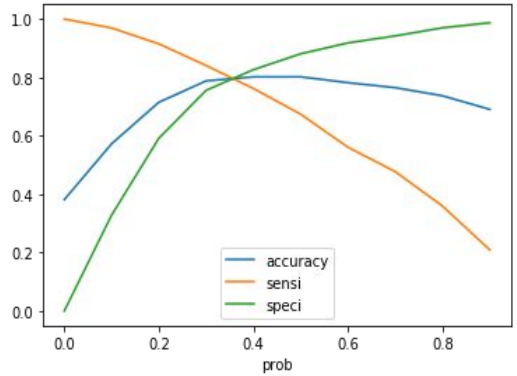
  Confusion Matrix:[[3525,477],[804,1662]]

- Calculate Sensitivity(0.69)
- Calculate specificity(.88)
- Calculate false positive rate(.11) , positive predictive value(.78) and Negative predictive value(.82)

- Plot ROC Curve and look good .
- Plot graph 'prob','accuracy','sensi', 'speci'

Find .3 is best cutoff for accuracy.





After taking cut off .3 following matrix we got.

1-accuracy_score: 79%

2-sensitivity: 84%

3-specificity: 76%

This is score on cut off .3

After training the model, we tested on test data set.And got Accuracy Score(80%).

Please find below Confusion Matrix and final predication.

	Converted	LeadID	Converted_Prob	final_predicted
0	1	4269	0.715870	1
1	1	2376	0.935845	1
2	1	7766	0.327830	0
3	0	9199	0.051512	0
4	1	4359	0.916881	1