PATTERN RECOGNITION AND MACHINE LEARNING

LAB ASSIGNMENT = 1

Akshaykumar Kanani(B19EE008)

PROBLEM STATEMENT:

A csv file has been provided to you at this link. It contains three columns. First column is the actual labels for a binary classification problem. Second, and third column are predicted probabilities from two classifiers. You will be converting these probabilities values in the final label based on the threshold value. Helping code-script is in Notebook. You are supposed to complete the functions computing the different evaluation metrics as described in the Colab Notebook at this link. You may download the Notebook and may start working on it on your local device or on Colab Notebook. The Notebook is provided to you for a quick start. You will define the functions for the following tasks

i.) To calculate accuracy.

ii.) To calculate precision and recall. iii.) To calculate F1 score.

Additionally you are required to change the threshold value (0.5, 0.4, 0.6 etc.) and compare, contrast the difference in metrics for both the models.

COMPARISON BASED ON THRESHOLD VALUES

THRESHOLD = 0.4

TP-True positive, TN-True negative, FP- False positive, FN-False negative

- TP for Logistic Reg : 6535
- TN for Logistic Reg : 2862
- FP for Logistic Reg : 5017
- FN for Logistic Reg : 1344

Logistic Reg.	Predicted class		
Actual class		Predicted as class 0	Predicted as class 1
	Class 0	2862	5017
	Class 1	1344	6535

- TP for Random Forest : 7411
- TN for Random Forest : 1930
- FP for Random Forest : 5949
- FN for Random Forest : 468

Random Forest	Predicted class		
Actual class		Predicted as class 0	Predicted as class 1
	Class 0	1930	5949
	Class 1	468	7411

ACCURACY-

Accuracy in both cases class wise and of whole sample is same

- Accuracy for Logistic Regression : 59.633202183018156
- Accuracy for Random Forest : 59.27782713542328

PRECISION-

- Precision for Logistic Regression : 56.570290858725755
- Precision for Random Forest : 55.471556886227546

RECALL-

- Recall for Logistic Regression : 82.94199771544612
- Recall for Random Forest : 94.06015991877142

F1 SCORE-

- F1 score for Logistic Regression : 0.6726365086717101
- F1 score for Random Forest : 0.6978671312208673

THRESHOLD = 0.5

TP-True positive, TN-True negative, FP- False positive, FN-False negative

- TP for Logistic Reg : 4279
- TN for Logistic Reg : 5425
- FP for Logistic Reg : 2454
- FN for Logistic Reg : 3600

Logistic Reg.	Predicted class

Actual class		Predicted as class 0	Predicted as class 1
	Class 0	5425	2454
	Class 1	3600	4279

- TP for Random Forest : 5047
- TN for Random Forest : 5519
- FP for Random Forest : 2360
- FN for Random Forest : 2832

Random Forest	Predicted class		
Actual class		Predicted as class 0	Predicted as class 1
	Class 0	5519	2360
	Class 1	2832	5047

ACCURACY-

Accuracy in both cases class wise and of whole sample is same

- Accuracy for Logistic Regression : 61.58141896179718
- Accuracy for Random Forest : 67.05165630156111

PRECISION-

- Precision for Logistic Regression : 63.55265112134264
- Precision for Random Forest : 68.1382476036182

RECALL-

- Recall for Logistic Regression : 54.30892245208783
- Recall for Random Forest : 64.05635232897576

F1 SCORE-

- F1 score for Logistic Regression : 0.5856830002737475
- F1 score for Random Forest : 0.6603427973308911

THRESHOLD = 0.6

TP-True positive, TN-True negative, FP- False positive, FN-False negative

- TP for Logistic Reg : 2406
- TN for Logistic Reg : 6858
- FP for Logistic Reg : 1021
- FN for Logistic Reg : 5473

Logistic Reg.	Predicted class		
Actual class		Predicted as class 0	Predicted as class 1
	Class 0	6858	1021

	Class 1	5473	2406

- TP for Random Forest : 2239
- TN for Random Forest : 7417
- FP for Random Forest : 462
- FN for Random Forest : 5640

Random Forest	Predicted class		
Actual class		Predicted as class 0	Predicted as class 1
	Class 0	7417	462
	Class 1	5640	2239

ACCURACY-

Accuracy in both cases class wise and of whole sample is same

- Accuracy for Logistic Regression : 58.789186444980324
- Accuracy for Random Forest : 61.276811778144435

PRECISION-

- Precision for Logistic Regression : 70.20717829004961
- Precision for Random Forest : 82.89522399111439

RECALL-

- Recall for Logistic Regression : 30.536870161187966
- Recall for Random Forest : 28.417311841604263

F1 SCORE-

- F1 score for Logistic Regression : 0.42561471784892974
- F1 score for Random Forest : 0.42325141776937614

All required graphs are on code file.
Here is the conclusion from those graphs:-

COMPARISON TABLE-

thres hold value	logistic regression				random forest			
	accuracy	precision	recall	f1-score	accuracy	precision	recall	f1-score
0.4	59.6332021830181 56	56.570290858725755	82.941997715 44612	0.67263650867171 01	59.27782713542328	55.47155688622 7546	94.06015991877 142	0.69786713122086 73
0.5	61.5814189617971 8	63.552651121 34264	54.308922452 08783	0.58568300027374 75	67.0516563015611 1	68.13824760361 82	64.05635232897 576	0.66034279733089 1
0.6	58.7891864449803 24	70.207178290 04961	30.536870161 187966	0.42561471784892974	61.27681177814 4435	82.89522399111 439	28.41731184160 4263	0.42325141776937 614

CONCLUSION-

NOTE: all graphs are done on code file. By observing them:

1) By increasing threshold value accuracy is first increase and then decrease.
2) By increasing threshold value precision value is increase.
3) By increasing threshold value recall value is decrease .
4) By increasing threshold value f1-score is first increase and then decrease.

This conclusion is for both logistic regression and random forest.

About ROC and AUC:

Graphs are on codefile. From that We get

ROC-AUC of Logistic Regression: 0.665744

ROC-AUC of Random Rainforest: 0.738295

So we can clearly see that Random forest is better approach then logistic regression as its AUC is greater than logistic regression.