Demo on Models

Rabi Kumar Singh

Use case-1 – Airlines Passenger Satisfaction

Problem:- Need to find the factors which are highly correlated to a satisfied (or dissatisfied) passenger.

Overview of dataset-

- Shape- (103904, 25)
- No Null Values
- Data type
- Models are used

1.Logistic Regression

acc_train -	87.475695
Test Accuracy	87.56395
Roc Score	87.107435
COrrect	27214
Incorrect	3865
Confusion	[[15912, 1647], [2218, 11302]]

2. Random Forest

acc_train	100.0
Test Accuracy	95.990862
Roc Score	95.723778
COrrect	29833
Incorrect	1246
Confusion	[[17169, 390], [856, 12664]]

3. XGboost

acc_train	94.296352
Test Accuracy	94.102127
Roc Score	93.797066
COrrect	29246
Incorrect	1833
Confusion	[[16882, 677], [1156, 12364]]

Tuned Random Forest

acc_train	99.972419
Test Accuracy	96.135654
Roc Score	95.872334
COrrect	29878
Incorrect	1201
Confusion	[[17190, 369], [832, 12688]]

Use-case-2: Gender Classification

Problem:- Classifying gender based on personal preferences. The way males and females are treated differently since birth moulds their behaviour and personal preferences into what society expects for their gender.

Overview of dataset-

- Shape- (569, 5).
- No Null Values.
- Data type.
- Label Encoding and one hot encoding.
- Models are used.

1.Logistic Regression

acc_train	69.565217
Test Accuracy	60.0
Roc Score	61.616162
COrrect	12
Incorrect	8
Confusion	[[5, 6], [2, 7]]

2.Random Forest

acc_train	93.478261
Test Accuracy	75.0
Roc Score	76.262626
COrrect	15
Incorrect	5
Confusion	[[7, 4], [1, 8]]

3.XGboost

acc_train	93.478261
Test Accuracy	60.0
Roc Score	60.606061
COrrect	12
Incorrect	8
Confusion	[[6, 5], [3, 6]]

4.Tuned Random Forest

acc_train	91.304348
Test Accuracy	80.0
Roc Score	81.818182
COrrect	16
Incorrect	4
Confusion	[[7, 4], [0, 9]]

Use-case-3: Water Potability Prediction

Problem:- We need to find the factor which affect the potability of drinking water. 1 mean potable and 0 means not potable.

Overview of dataset-

- Shape- (3276, 10).
- Null Values are replaced by mean.
- Data type.
- Correlation Matrix & VIF.
- Models are used.

1.Logistic Regression

acc_train	60.561661
Test Accuracy	62.271062
Roc Score	50.0
COrrect	510
Incorrect	309
Confusion	[[510, 0], [309, 0]]

2.Random Forest

acc_train	100.0
Test Accuracy	80.09768
Roc Score	76.239292
COrrect	656
Incorrect	163
Confusion	[[469, 41], [122, 187]]

3.XGboost

acc_train	85.673586
Test Accuracy	77.899878
Roc Score	72.816486
COrrect	638
Incorrect	181
Confusion	[[477, 33], [148, 161]]

4.Tuned Random Forest

acc_train	100.0
Test Accuracy	78.510379
Roc Score	75.666286
COrrect	643
Incorrect	176
Confusion	[[445, 65], [111, 198]]

Use-case:-4 Heart Disease classification

Problem:- Predicting probability of heart disease in patients.

Overview of dataset:-

- Shape- (4238, 16).
- No Null Values.
- Data type.
- Correlation Matrix & VIF.
- Models are used.

1.Logistic Regression

acc_train	98.415374
Test Accuracy	98.899371
Roc Score	77.697456
COrrect	1258
Incorrect	14
Confusion	[[1243, 2], [12, 15]],

2.Random Forest

acc_train	100.0
Test Accuracy	98.899371
Roc Score	77.697456
COrrect	1258
Incorrect	14
Confusion	[[1243, 2], [12, 15]],

3.XGboost

acc_train	99.527984
Test Accuracy	98.034591
Roc Score	75.443998
COrrect	1247
Incorrect	25
Confusion	[[1233, 12], [13, 14]]

4.Tuned Random Forest

acc_train	99.730276
Test Accuracy	98.820755
Roc Score	77.657296
COrrect	1257
Incorrect	15
Confusion	[[1242, 3], [12, 15]]

Use-case-5: Fetal Health Classification

Problem:- Classify the health of a fetus as Normal, Suspect or Pathological using CTG data

Overview of dataset-

- Shape- (2126, 22).
- No Null Values.
- Data type.
- Correlation Matrix & VIF.
- Models are used.

1.Logistic Regression

acc_train	90.915751
Test Accuracy	89.059829
Roc Score	70.258266
COrrect	521
Incorrect	64
Confusion	[[481, 10], [54, 40]]

2.Random Forest

acc_train	99.85348
Test Accuracy	94.017094
Roc Score	83.96347
COrrect	550
Incorrect	35
Confusion	[[485, 6], [29, 65]]

3.XGboost

acc_train	99.047619
Test Accuracy	93.675214
Roc Score	83.759804
COrrect	548
Incorrect	37
Confusion	[[483, 8], [29, 65]]

4.Tuned Random Forest

acc_train	99.85348
Test Accuracy	93.675214
Roc Score	83.759804
COrrect	548
Incorrect	37
Confusion	[[483, 8], [29, 65]]

Use case-6: Delhi fatality classification

Problem:- Classify the accident type whether it is fatal or simple on the basis different factors.

Overview of dataset-

- Shape- (75748, 7).
- No Null Values.
- Data type..
- Models are used.

1.Logistic Regression

acc_train	75.293439
Test Accuracy	75.213253
Roc Score	50.0
COrrect	16753
Incorrect	5521
Confusion	[[0, 5521], [0, 16753]]

2.Random Forest

acc_train	75.301135
Test Accuracy	75.208764
Roc Score	49.997015
COrrect	16752
Incorrect	5522
Confusion	[[0, 5521], [1, 16752]]

3.XGboost

acc_train	75.301135
Test Accuracy	75.208764
Roc Score	49.997015
COrrect	16752
Incorrect	5522
Confusion	[[0, 5521], [1, 16752]]

4.Tuned Random Forest

acc_train	75.301135
Test Accuracy	75.208764
Roc Score	49.997015
COrrect	16752
Incorrect	5522
Confusion	[[0, 5521], [1, 16752]]

Use-Case-7: Breast Cancer

Problem:- Find the factors which causes cancer.

Overview of dataset-

- Shape- (569, 32).
- No Null Values.
- Data type.
- Correlation Matrix and Multi-collinearity calculated.
- Models are used.

1.Logistic Regression

acc_train	38.442211
Test Accuracy	34.502924
Roc Score	50.0
COrrect	59
Incorrect	112
Confusion	[[0, 112], [0, 59]]

2.Tuned Logistic Regression

Accuracy Score:- 0.9473684210526315
F1 Score:- 0.9279999999999999
Average Precision Score:- 0.869741122194289
Log Loss:- 1.817867744641157
Precision Score:- 0.8787878787878788
Recall Score:- 0.9830508474576272
ROC-AUC Score:- 0.9558111380145279

Note- Only Logistic was performing well; rest of the model was overffing.