

Algorithm Evaluation

UTKARSH GAIKWAD

How to identify which model to select?

```
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVM

dct = {'Logistic':LogisticRegression(),
      'Decision Tree':DecisionTreeClassifier(),
      'Random Forest':RandomForestClassifier(),
      'KNN':KNeighborsClassifier(),
      'SVM':SVM()}
```

Step 1: Create a dictionaries which include base model

Evaluate model

```
from sklearn.metrics import f1_score
from sklearn.model_selection import cross_val_score
tr = []
tr_cv = []
ts = []
for name, model in dct.items():
    model.fit(xtrain,ytrain)
    ypred_tr = model.predict(xtrain)
    ypred_ts = model.predict(xtest)
    f1_tr = f1_score(ytrain,ypred_tr)
    scores_tr = cross_val_score(model,xtrain,ytrain,cv=5,scoring='f1')
    scores_avg = scores_tr.mean()
    f1_ts = f1_score(ytest,ypred_ts)
    tr.append(f1_tr)
    tr_cv.append(scores_avg)
    ts.append(f1_ts)
    print(f'Model Name : {name}')
    print(f'Training F1 Score : {f1_tr}')
    print(f'5 fold Cross Validated F1 Score : {scores_avg}')
    print(f'Testing F1 Score : {f1_ts}')
```

View the results in DataFrame

```
import pandas as pd
eval_dct = {'Model':list(dct.keys()),
            'Training F1':tr,
            'Testing F1':ts,
            'Training F1 CV':tr_cv}
eval_df = pd.DataFrame(eval_dct)
eval_df
```

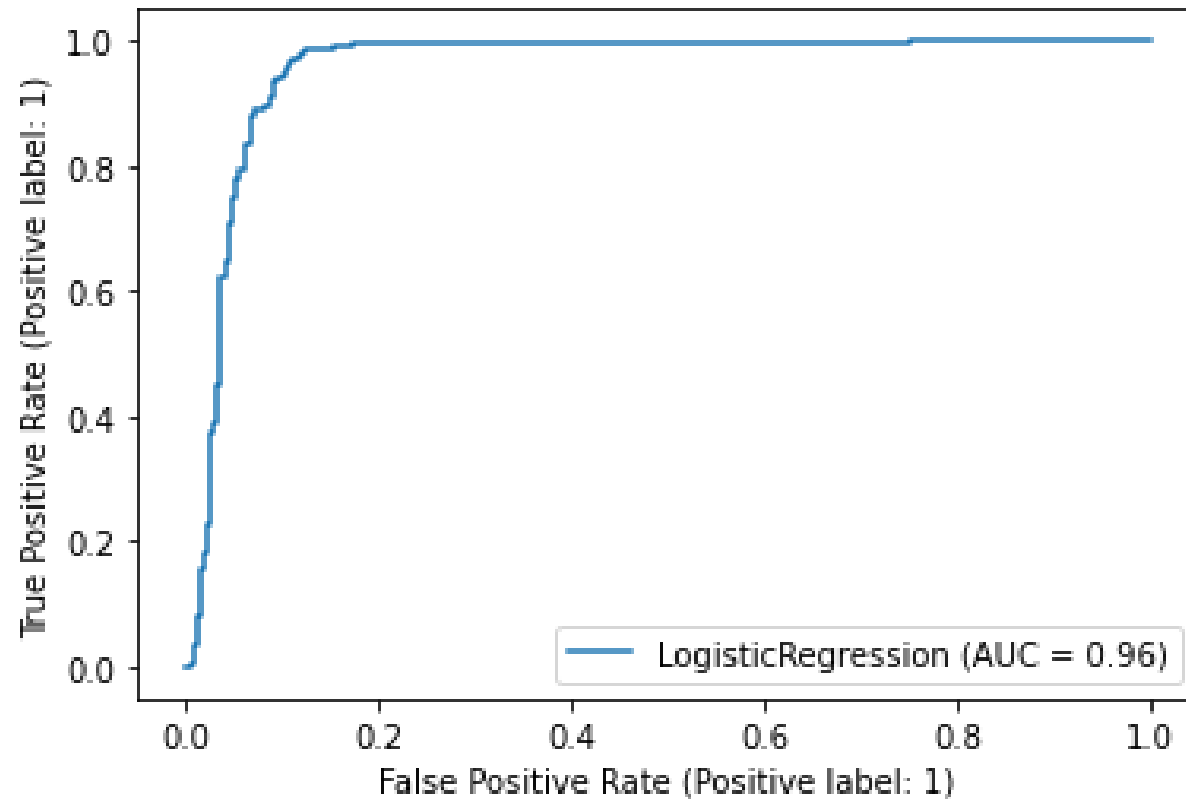
ROC(Receiver Operator Characteristics) Curve

| | Predicted 0 | Predicted 1 |
|--------------------|-----------------------|-----------------------|
| Actual 0 | TN | FP |
| Actual 1 | FN | TP |

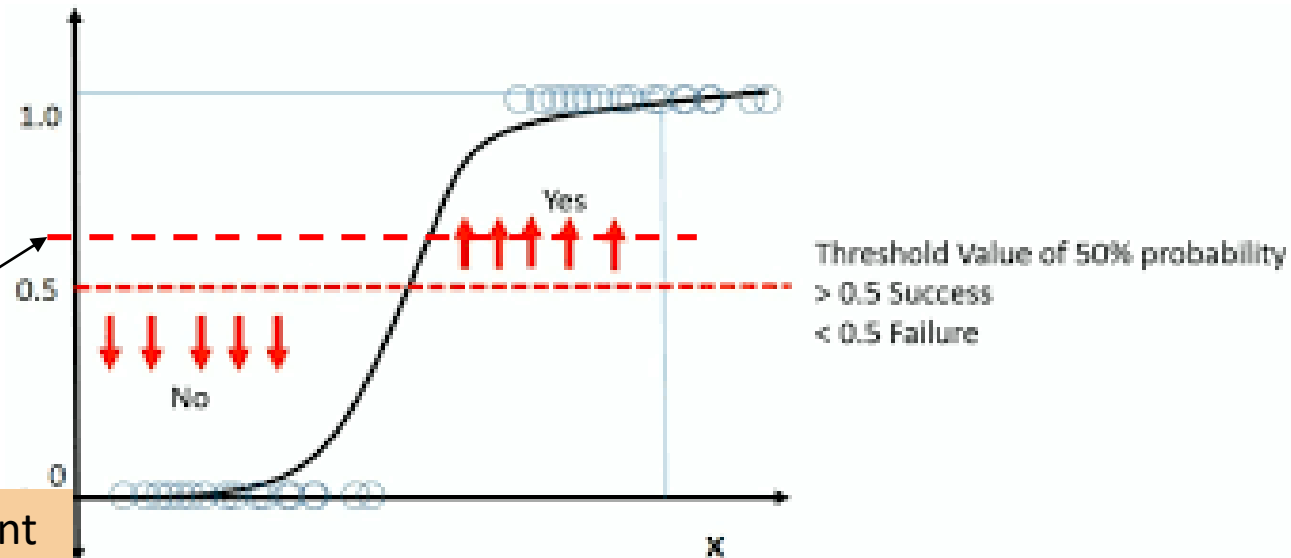
$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}$$

Example ROC AUC Plot



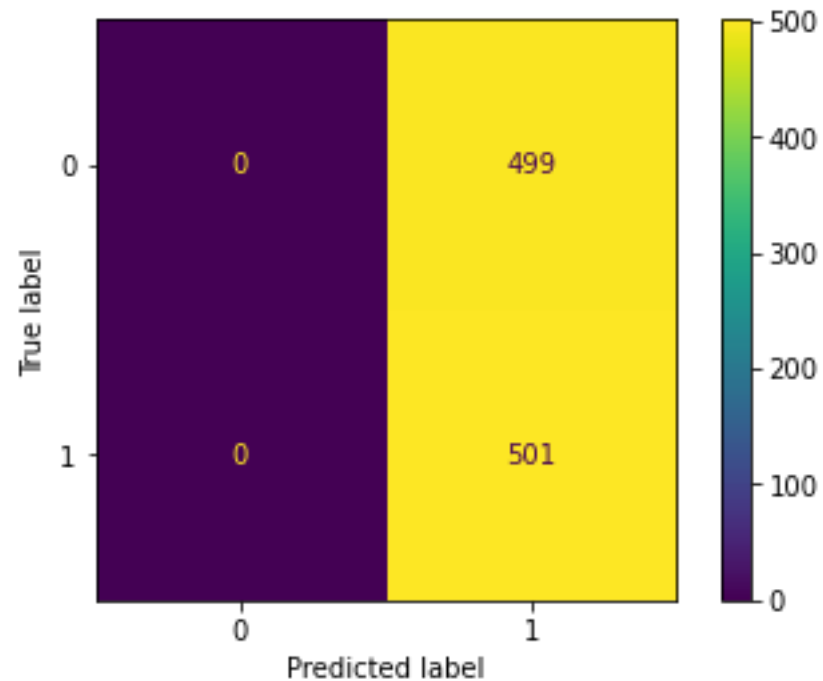
Threshold in Logistic Regression



What if we use different Threshold values?

Assume we have 1000 Samples
499 have labels 0
501 have labels 1

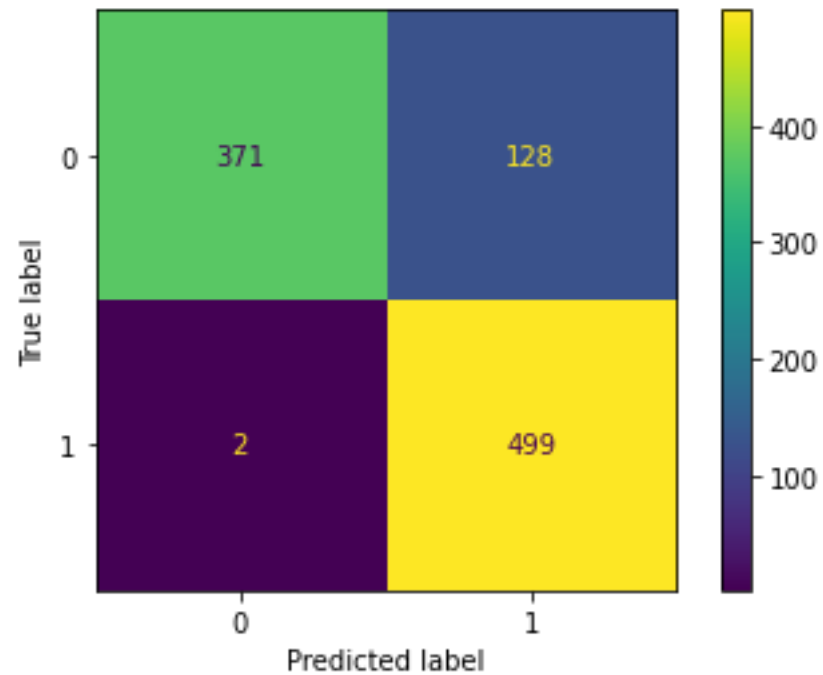
Threshold = 0, 1000 Samples



$$\text{False Positive Rate (FPR)} = \frac{499}{0 + 499} = 1$$

$$\text{True Positive Rate (TPR)} = \frac{501}{0 + 501} = 1$$

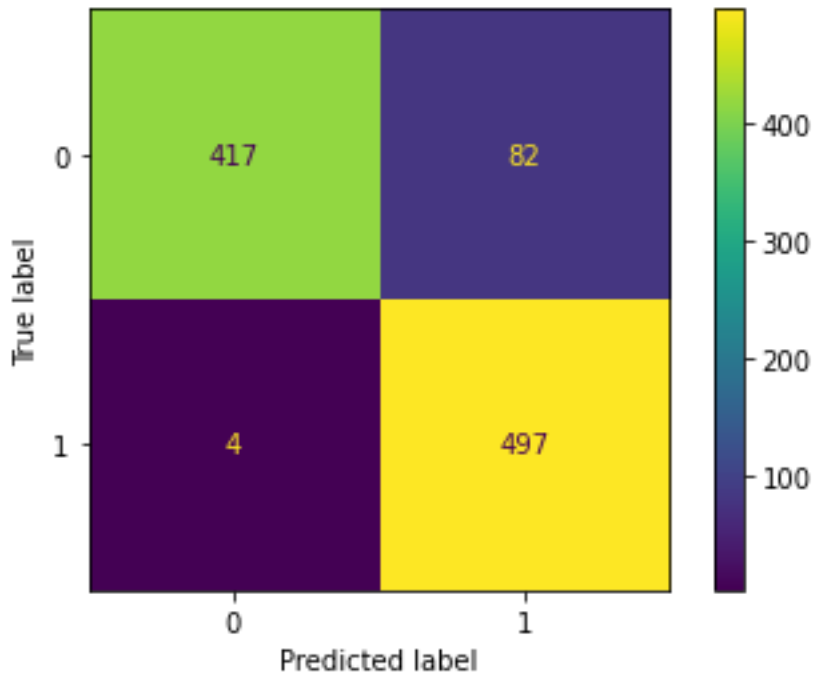
Threshold = 0.1



$$\text{False Positive Rate (FPR)} = \frac{128}{128 + 371} = 0.2565$$

$$\text{True Positive Rate (TPR)} = \frac{499}{2 + 499} = 0.9960$$

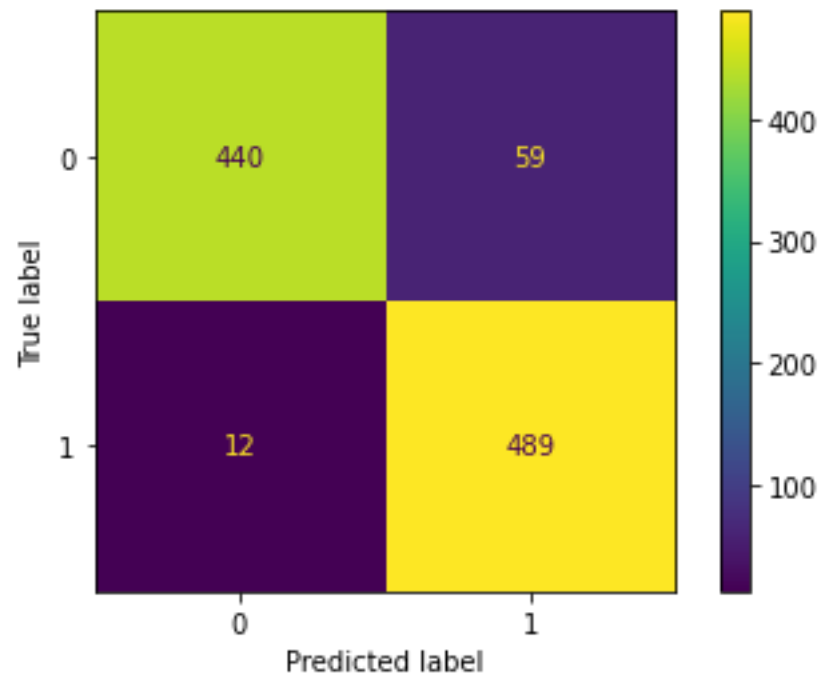
Threshold = 0.3



$$\text{False Positive Rate (FPR)} = \frac{82}{82 + 417} = 0.1643$$

$$\text{True Positive Rate (FPR)} = \frac{497}{497 + 4} = 0.9920$$

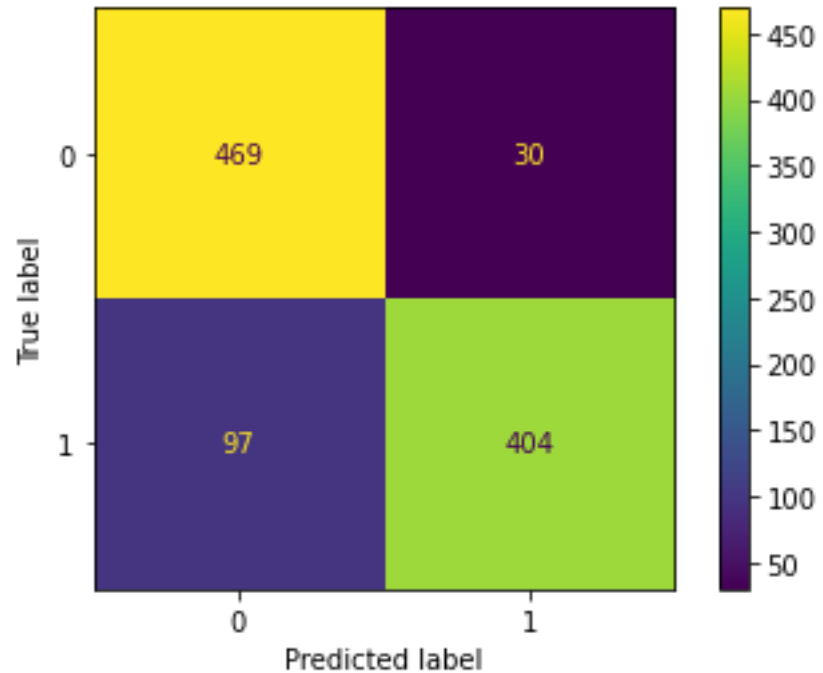
Threshold = 0.5



$$\text{False Positive Rate (FPR)} = \frac{59}{59 + 440} = 0.1182$$

$$\text{True Positive Rate (TPR)} = \frac{489}{12 + 489} = 0.9760$$

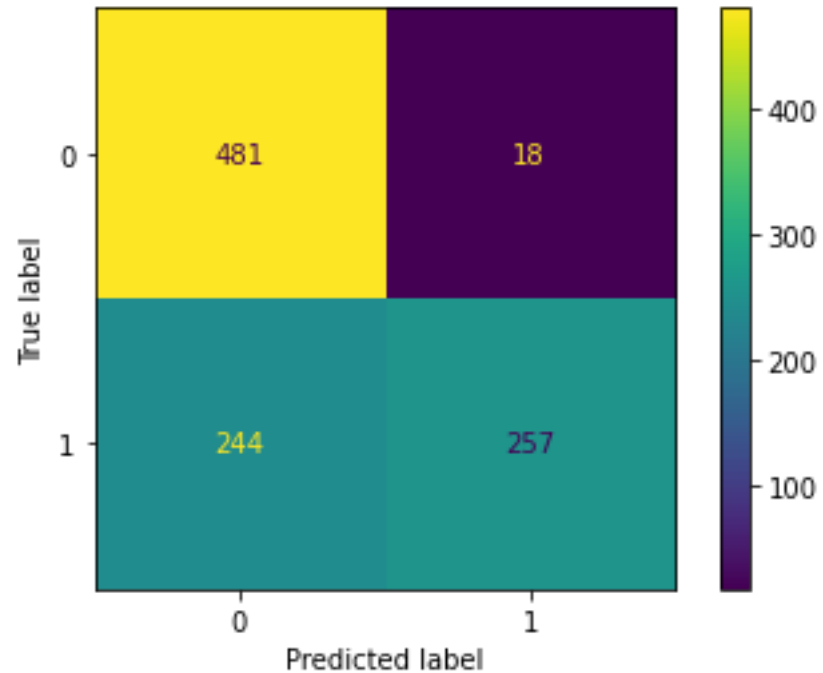
Threshold = 0.8



$$\text{False Positive Rate (FPR)} = \frac{30}{30 + 469} = 0.0601$$

$$\text{True Positive Rate (TPR)} = \frac{404}{404 + 97} = 0.8064$$

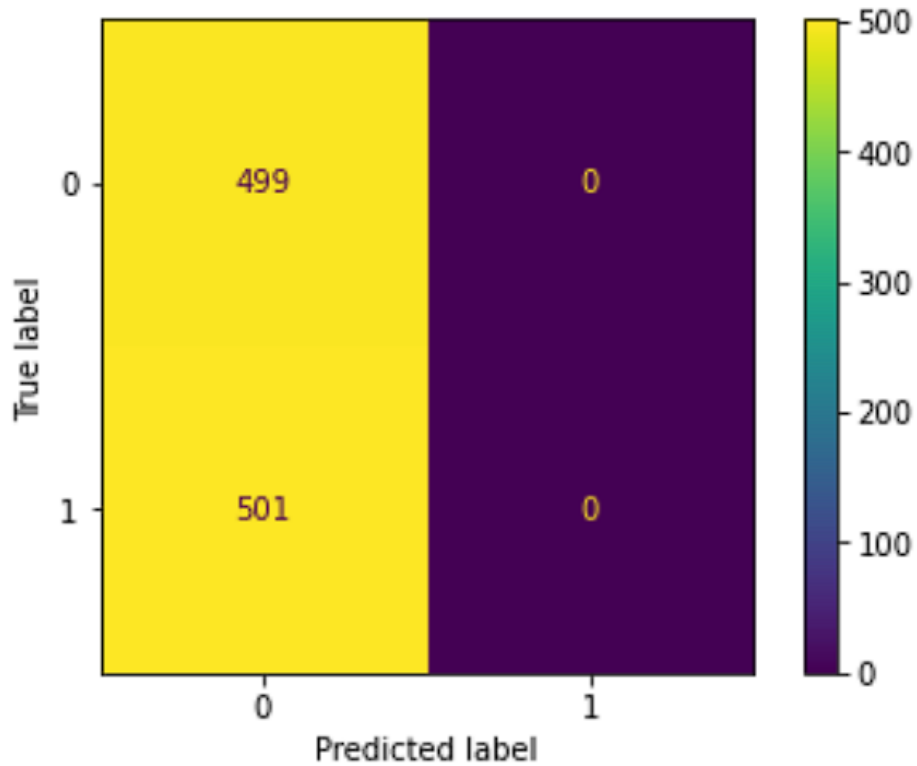
Threshold = 0.9



$$\text{False Positive Rate (FPR)} = \frac{18}{18 + 481} = 0.0361$$

$$\text{True Positive Rate (TPR)} = \frac{257}{257 + 244} = 0.5129$$

Threshold = 1.0



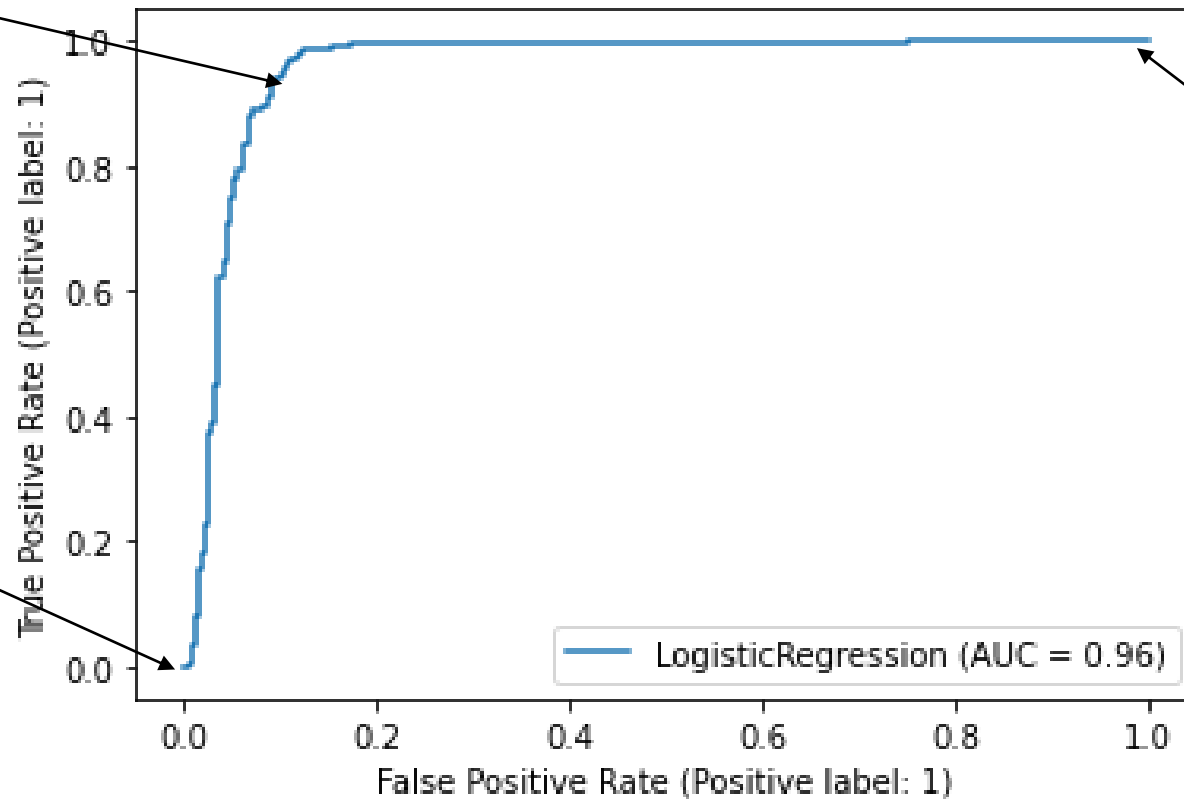
$$\text{False Positive Rate (FPR)} = \frac{0}{0 + 499} = 0$$

$$\text{True Positive Rate (TPR)} = \frac{0}{0 + 501} = 0$$

Example ROC AUC Plot

Threshold=0.5
FPR=0.11, TPR=0.97

Threshold=1
FPR=0, TPR=0



Threshold=0
FPR = 1, TPR=1

Thank you

UTKARSH GAIKWAD

PING ME ON SKYPE FOR ANY QUERIES