



CIS 678 - Machine Learning

Predictive modeling: Classification Metrics and Imbalanced Data



Classification Matrices

- Accuracy

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$



Classification Matrices

- Accuracy

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?



Classification Matrices

- Accuracy

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

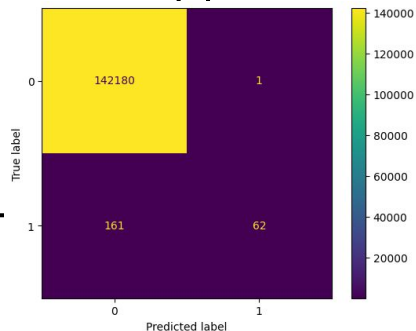
- Is accuracy a good metric?
- Not always

Classification Matrices

- Accuracy

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?
- Not always
- Let's analyze the confusion matrix of our [credit card fraud detection notebook](#)
 - Accuracy metric can be catastrophic



```
[20] # Make predictions using the testing set
y_pred = clf.predict(X_test)
# The mean squared error
print("accuracy: %.5f" % accuracy_score(y_test, y_pred))
```

accuracy: 0.99886



Classification Matrices

- Accuracy

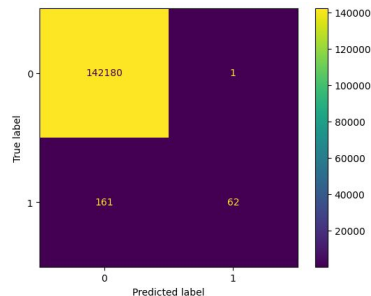
$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?
- Not always
- Let's analyze the confusion matrix of our [credit card fraud detection notebook](#)
 - Accuracy metric can be catastrophic
- What other metrics we may use?

Classification Matrices

- Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$



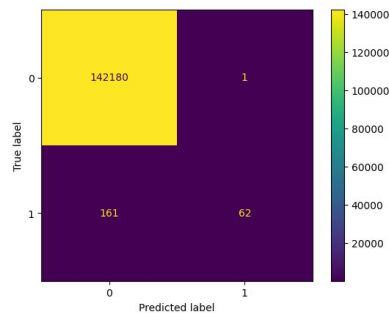
```
[20] # Make predictions using the testing set
y_pred = clf.predict(X_test)
# The mean squared error
print("accuracy: %.5f" % accuracy_score(y_test, y_pred))

accuracy: 0.99886
```

Classification Matrices

- Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$



True \ Predicted	N	P
	N	P
N	TN	FP
P	FN	TP



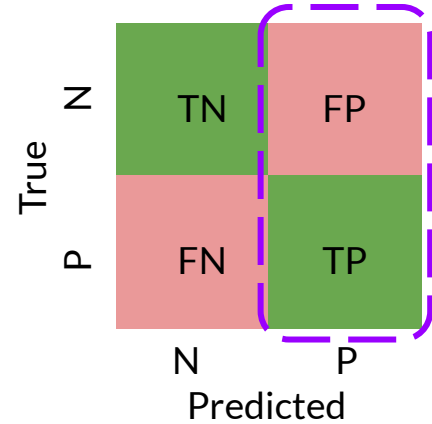
Other important classification metrics

- Precision (also called **Positive Predictive Value**)
- Recall (also called **Sensitivity**)
- F1 Score

Metrics

- Precision (also called **Positive Predictive Value**)

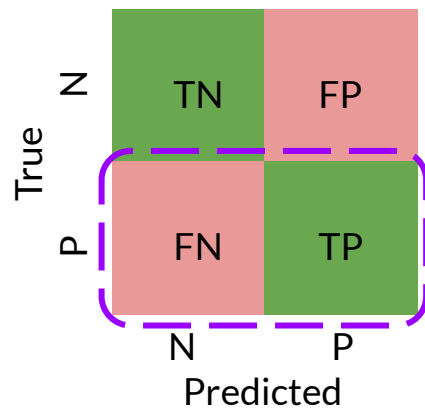
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$



Metrics

- Recall (also called **Sensitivity**)

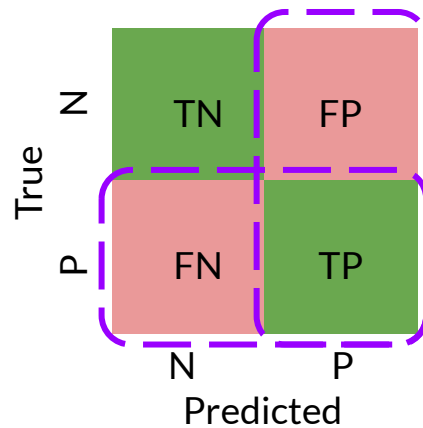
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



Metrics

- F1 Score

$$\text{F1 Score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$





Data Imbalance Problem

- Demonstration through a practical example
 - [CC fraud detection](#)

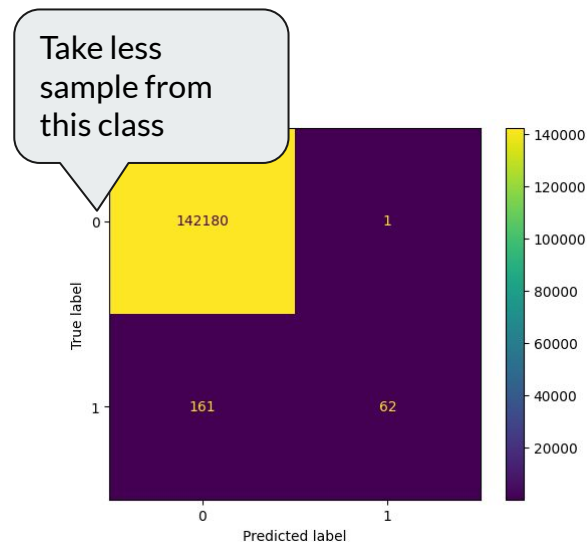


Data Imbalance Problem

- How to deal with Data Imbalance Problems
 - Through Sampling Bias

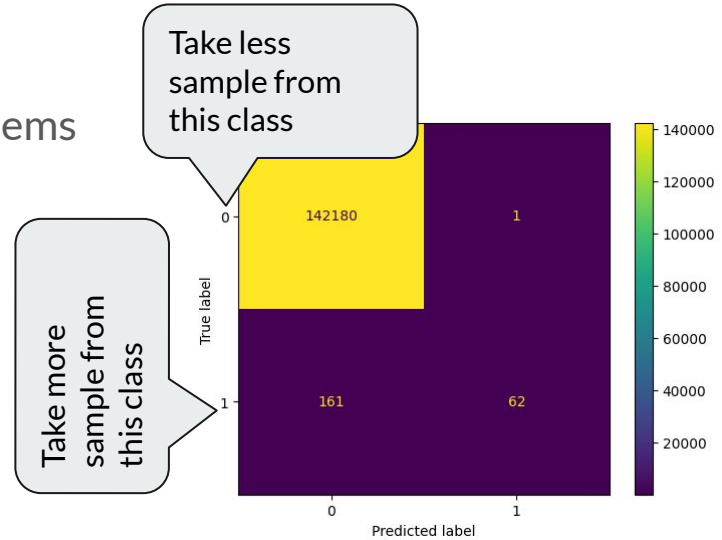
Data Imbalance Problem

- How to deal with Data Imbalance Problems
 - Through Sampling Bias
 - Undersampling
 - Oversampling



Data Imbalance Problem

- How to deal with Data Imbalance Problems
 - Through Sampling Bias
 - Undersampling
 - Oversampling



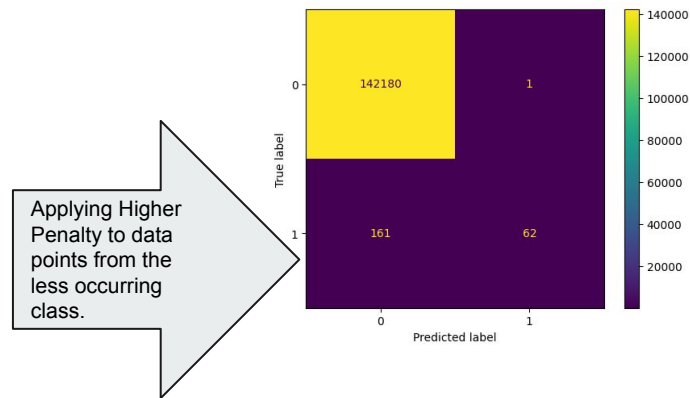


Data Imbalance Problem

- How to deal with Data Imbalance Problems
 - Through Sampling Bias
 - Undersampling
 - Oversampling
 - Redefining model (loss function for an example)

Data Imbalance Problem

- How to deal with Data Imbalance Problems
 - Through Sampling Bias
 - Undersampling
 - Oversampling
 - Redefining model (loss function for an example)





QA