CIS 635 Knowledge Discovery & Data Mining

Predictive modeling: Classification Metrics and Imbalanced Data

Classification Matrices

- Accuray

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

Classification Matrices

- Accuray

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

- Is accuracy a good metric?

Classification Matrices

- Accuray

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

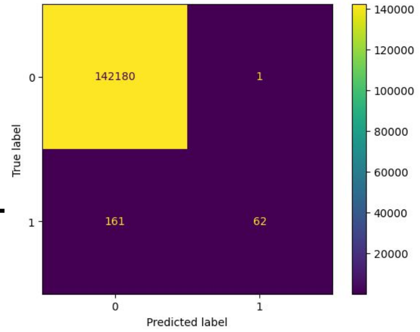
- Is accuracy a good metric?
- Not always

Classification Matrices

- Accuray

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

- Is accuracy a good metric?
- Not always
- Let’s analyze the confusion matrix of our credit card fraud detection notebook
    - Accuray 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

- Accuray

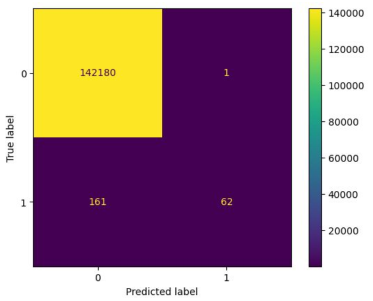
$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

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

Metrics

- Accuray

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

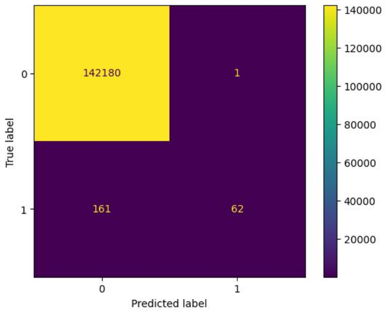


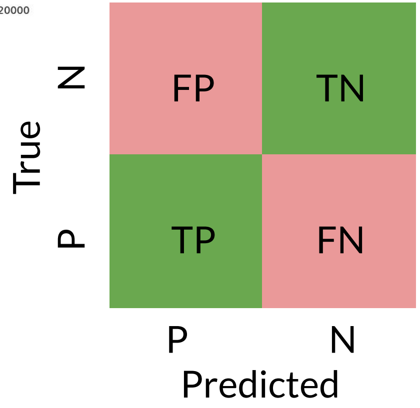

```
[20] # Make predictions using the testing set
     y_pred = clf.predict(X_test)
     # The mean squared error
     print("accuracy: %.5f" % accuracy_score(y_test, y_pred))

     accuracy: 0.99886
```

Metrics

- Accuray

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





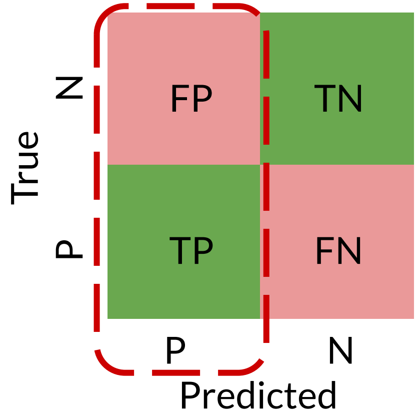
Other important classification metrics

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

Metrics

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

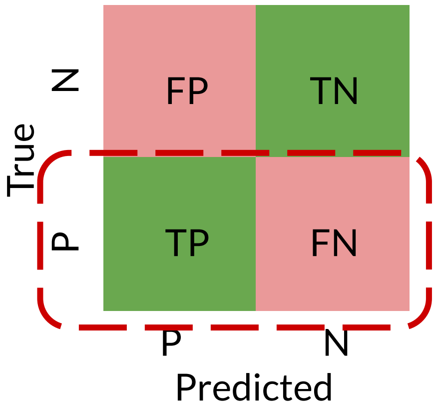
$$\text{Precision} = \frac{TP}{TP + FP}$$



Metrics

- Recall (also called **Sensitivity**)

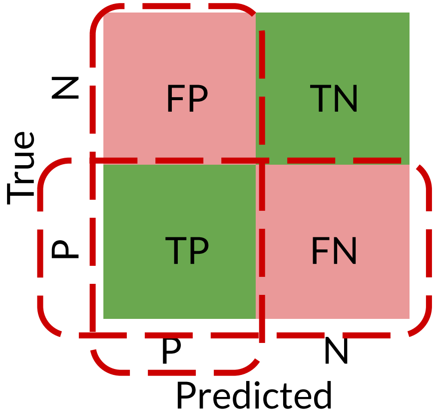
$$\text{Recall} = \frac{TP}{TP + FN}$$



Metrics

- F1 Score

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



Data Imbalance Problem

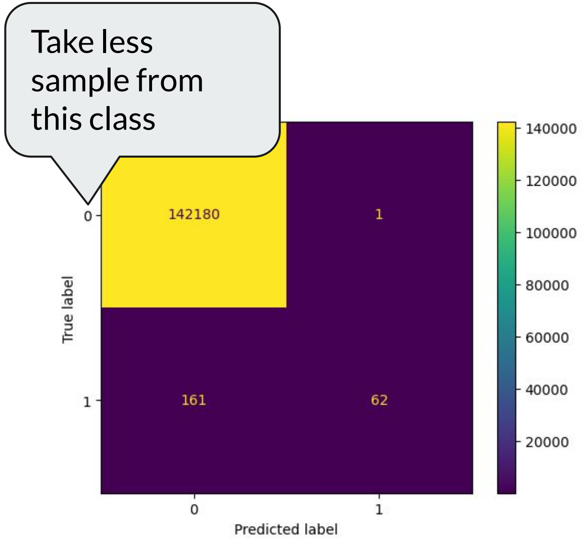
- Demonstration through a practical example
  - CC fraud detection

Data Imbalance Problem

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

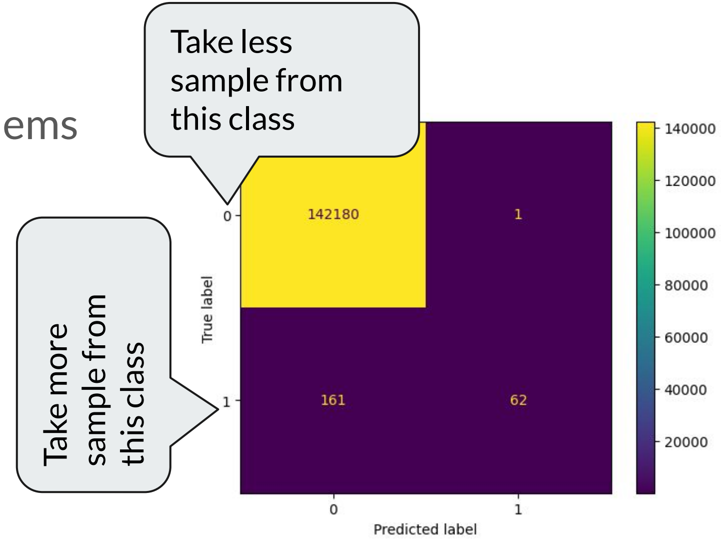
Data Imbalance Problem

- 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



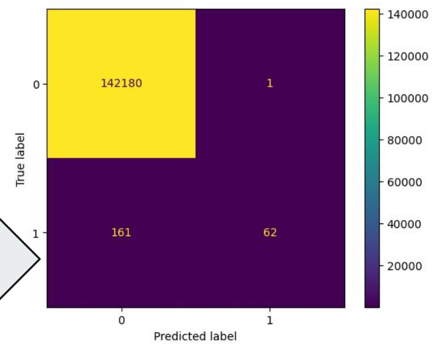
Data Imbalance Problem

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

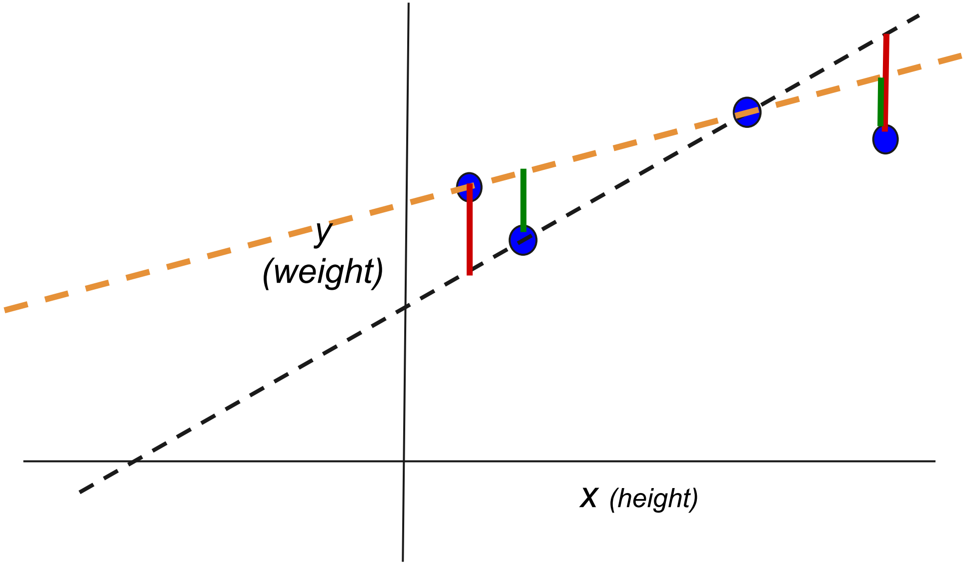
Data Imbalance Problem

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

Applying Higher Penalty to data points from the less occurring class.



Fitting a linear function/model



Model

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\Theta = \{\beta_0, \beta_1\}$$

Fitting Error

$$\epsilon = |\hat{y} - y|$$

Optimization function

$$E_\Theta = \frac{1}{2}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$\Theta^* = \mathrm{argmin}_\Theta\, E\{(x_i, y_i)\}_{i=1,\cdots,N}$$

QA