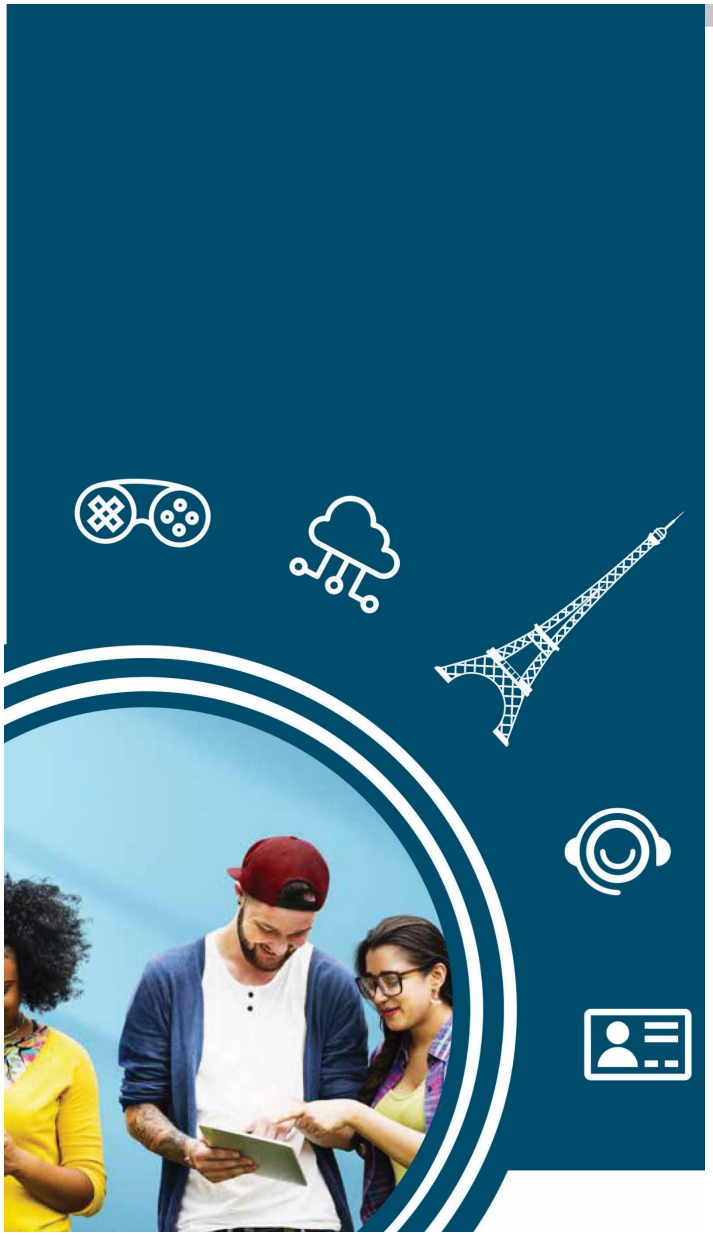


FOUNDATIONS OF STATISTICAL ANALYSIS & MACHINE LEARNING

Amric Trudel
amric.trudel@epita.fr

COURSE PROGRAM

Structure

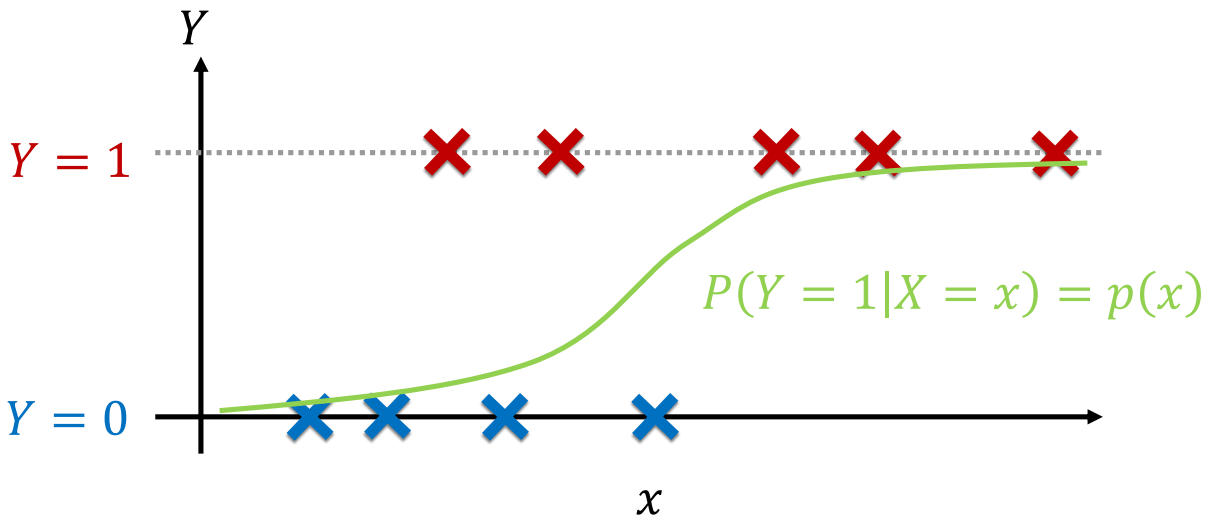
PREPARATION	Data exploration
	Data preprocessing
REGRESSION	Linear regression with one variable
	Multiple and polynomial regression
CLASSIFICATION	Logistic regression
	Classification model assessment
	k-NN, Decision Tree, SVM
CLUSTERING	k-means, hierarchical clustering
DIMENSIONALITY REDUCTION	Principal Components Analysis
ALL NOTIONS	Final assignment

LOGISTIC REGRESSION

Model definition

- Probability of belonging to class 1:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$



CLASSIFICATION MODEL TRAINING

Log loss

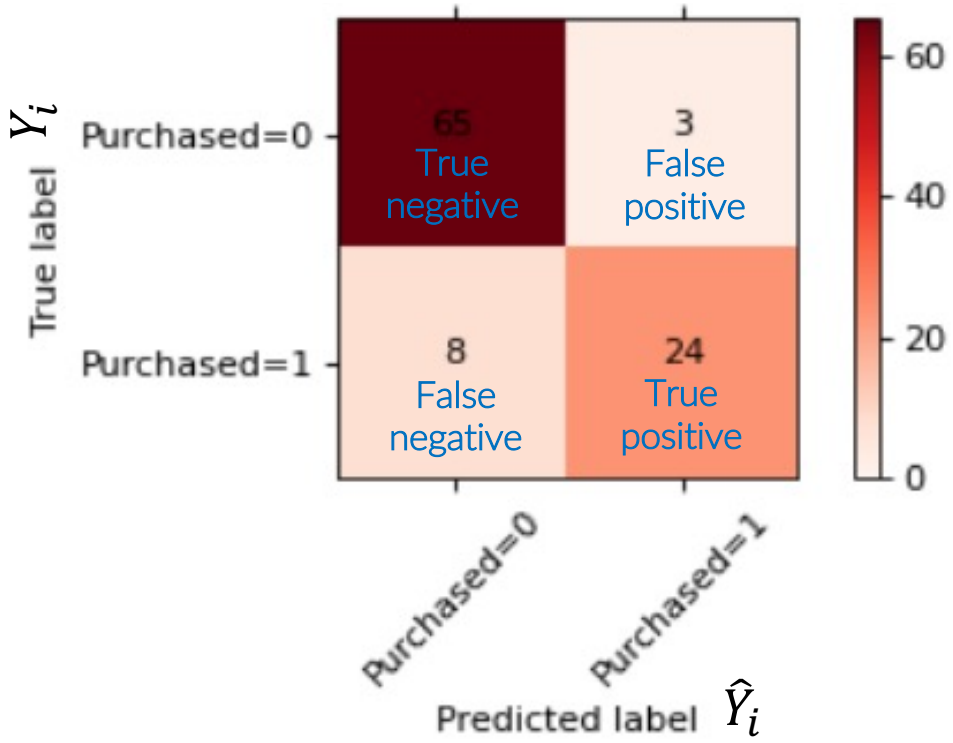
- Loss function of a classifier (higher accuracy when closer to zero):

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n} Y_i \log\left(\hat{p}(x_i)\right) + (1 - Y_i)\log\left(1 - \hat{p}(x_i)\right)$$

Y_i	$\hat{p}(x_i)$	$\hat{Y}_i$
1	0.85	1
1	0.43	0
0	0.12	0
1	0.62	1
0	0.51	1
0	0.22	0
⋮	⋮	⋮

CLASSIFICATION MODEL ASSESSMENT

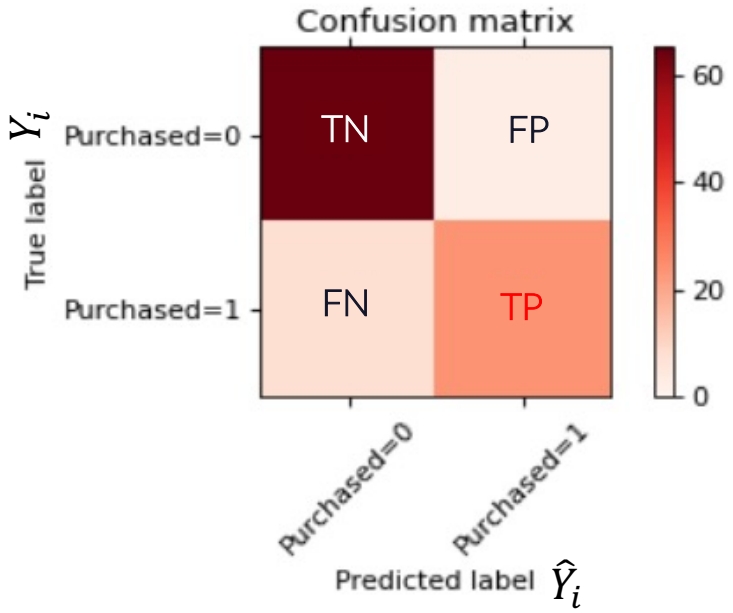
Confusion matrix



Y_i	$\hat{Y}_i$	
1	1	TP
1	0	FN
0	0	TN
1	1	TP
0	1	FP
0	0	TN
⋮	⋮	⋮

CLASSIFICATION MODEL ASSESSMENT

Classification metrics

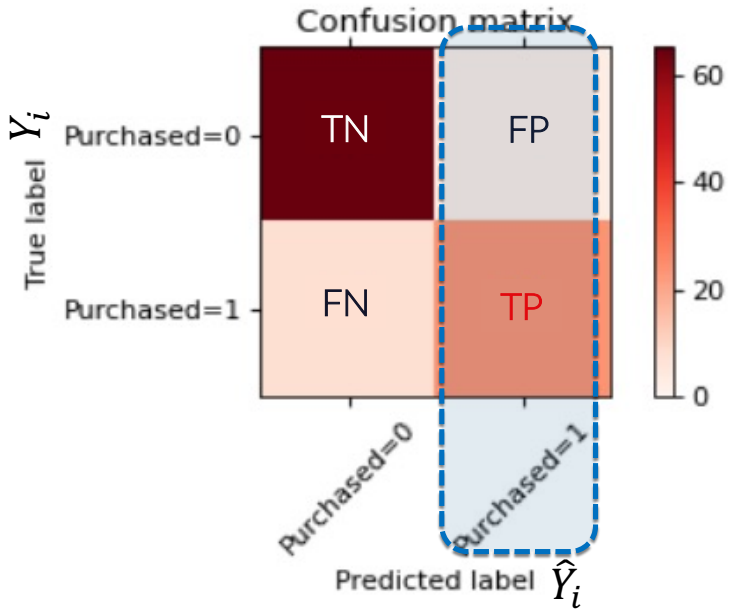


$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

What is the proportion of correct predictions?

CLASSIFICATION MODEL ASSESSMENT

Classification metrics

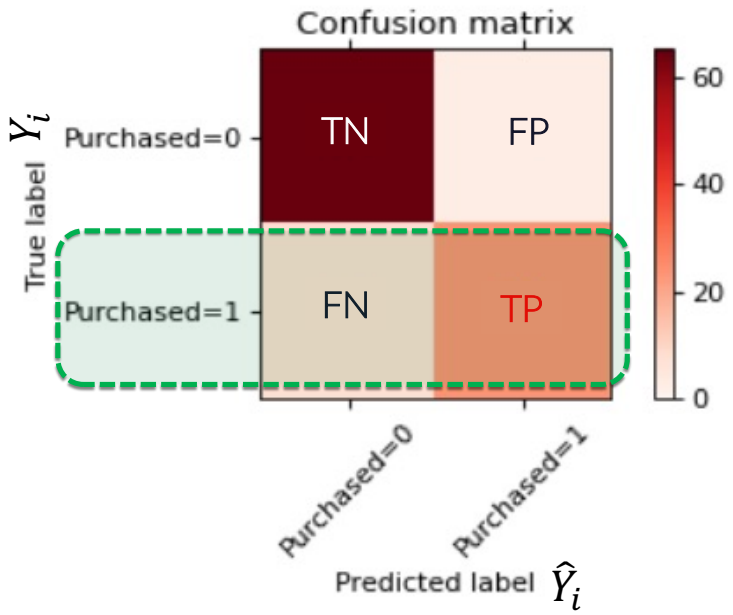


$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

How many selected items are relevant?

CLASSIFICATION MODEL ASSESSMENT

Classification metrics

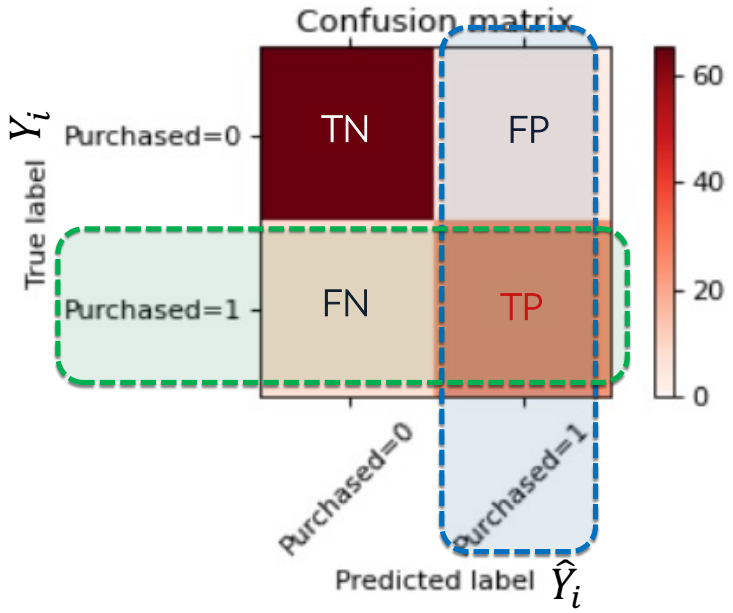


$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

How many relevant items are selected?

CLASSIFICATION MODEL ASSESSMENT

Classification metrics



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

How many selected items are relevant?

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

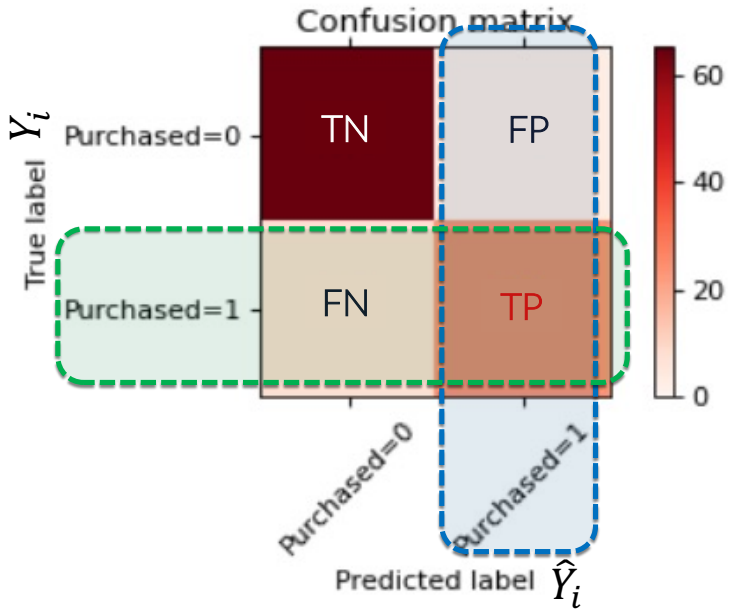
How many relevant items are selected?

$$\text{F1}-\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Combination of precision and recall

CLASSIFICATION MODEL ASSESSMENT

Classification metrics



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

What is the proportion of correct predictions?

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

How many selected items are relevant?

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

How many relevant items are selected?

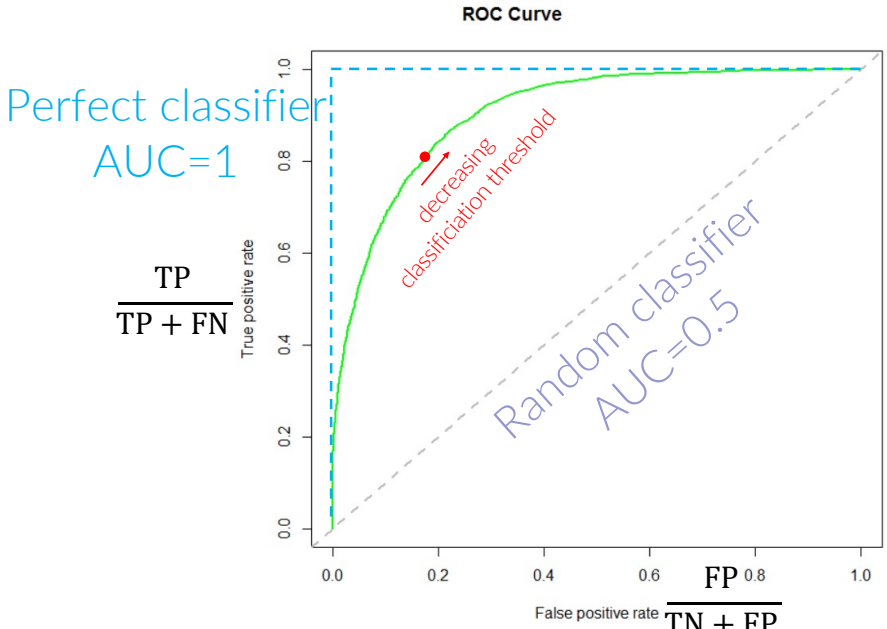
$$\text{F1}-\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

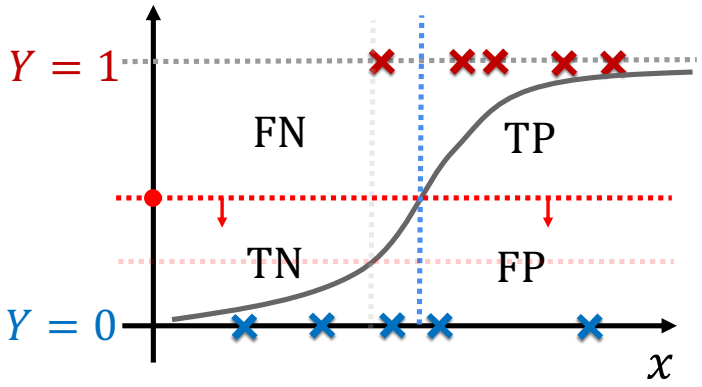
Combination of precision and recall

CLASSIFICATION MODEL ASSESSMENT

Receiver Operating Characteristic (ROC)

- ROC curve: plot of true positive rate vs. false positive rate.
- Area Under the Curve (AUC): metric that represents the quality and performance of a classifier.





CLASSIFICATION MODEL ASSESSMENT

Python implementation

- Computing the confusion matrix:

```
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)
```

- Computing the accuracy, the precision, the recall:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score
accuracy_score(y_test, y_pred)
precision_score(y_test, y_pred)
recall_score(y_test, y_pred)
f1_score(y_test, y_pred)
```

CLASSIFICATION MODEL ASSESSMENT

Python implementation

- Plotting the ROC curve:

```
from sklearn.metrics import roc_curve
y_score = y_proba[:,1]
FP_rate, TP_rate, threshold = roc_curve(y_test, y_score)
plt.plot(FP_rate, TP_rate)
```

- Computing the AUC:

```
from sklearn.metrics import roc_auc_score
roc_auc_score(y_test, y_score)
```

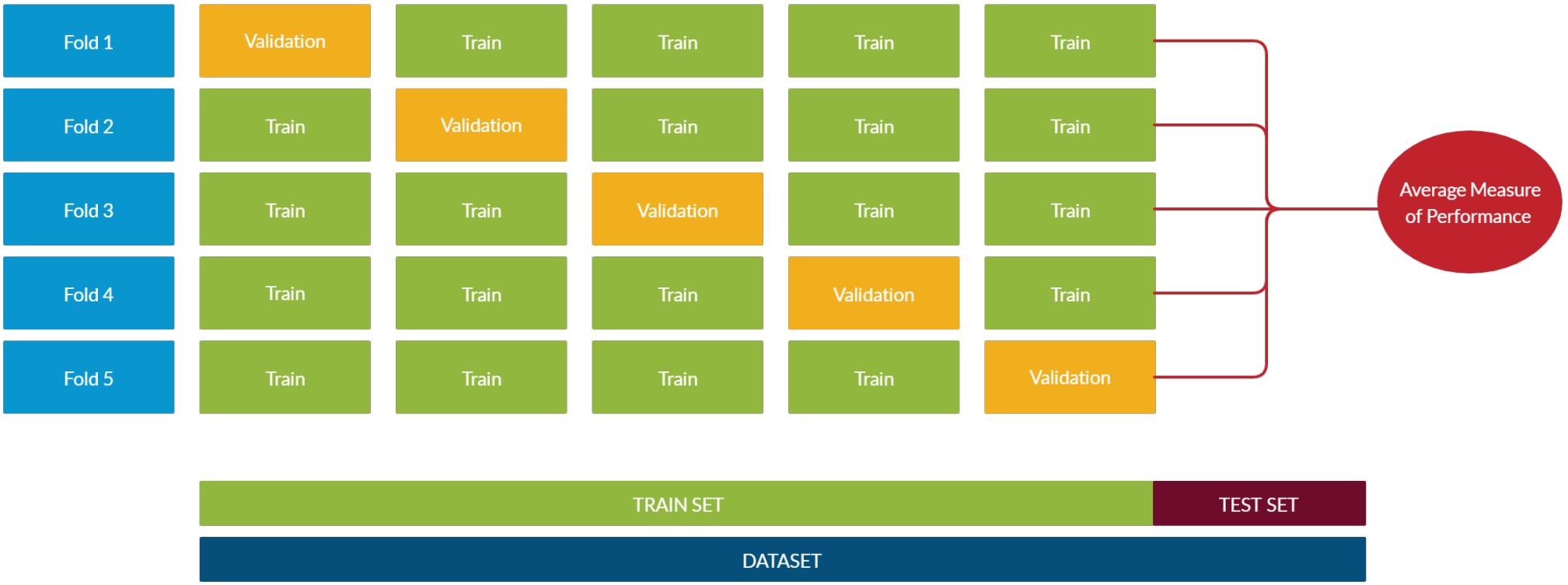
MODEL ASSESSMENT TECHNIQUE

K-fold Cross Validation

- We saw in the lecture on multiple linear regression that we should use a (hold out) **validation set** to evaluate model alternatives in order to choose the best one.
- The issue with using a hold out validation set of about 25% is that it ''wastes'' about 25% of the data.
- There is a more sophisticated way: **K-fold cross-validation**

MODEL ASSESSMENT TECHNIQUE

K-fold Cross Validation



MODEL ASSESSMENT TECHNIQUE

K-fold Cross Validation

- Fitting and scoring a model with cross-validation:

```
from sklearn.model_selection import cross_val_score
accuracies = cross_val_score(classifier, X, y, cv=5,
                             scoring='accuracy')
accuracies.mean()
```

MODEL ASSESSMENT TECHNIQUE

K-fold Cross Validation

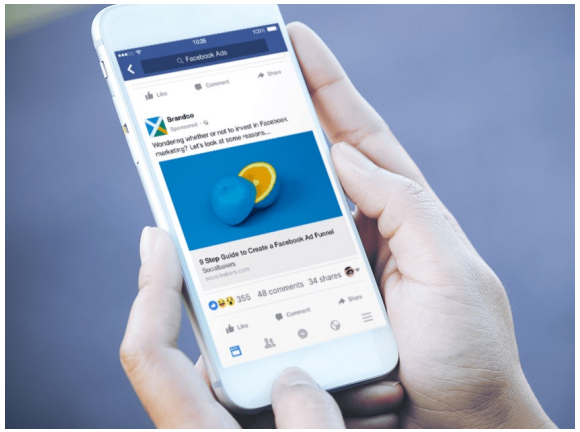
Procedure:

- Choose a metric you want to optimize (accuracy, precision, recall, f1)
- Perform k-fold cross validation on each candidate model
- Average the scores obtained among all folds for each model
- Pick the best model
- Retrain that model (with the same hyperparameters) on the entire training dataset
- Measure the generalization error with the test set (that you should still keep aside)

CLASSIFICATION ASSESSMENT

Example of implementation

- Data set: user profiles and sales information
- Objectives:
  - Train a logistic regression to predict purchase based on profile information
  - Assess the performance



	User ID	Gender	Age	EstimatedSalary	Purchased
1	15624510	Male	19	19000	no
2	15810944	Male	35	20000	no
3	15668575	Female	26	43000	no
4	15603246	Female	27	57000	no
5	15804002	Male	19	76000	no
6	15728773	Male	27	58000	no
7	15598044	Female	27	84000	no
8	15694829	Female	32	150000	yes
9	15600575	Male	25	33000	no
10	15727311	Female	35	65000	no
11	15570769	Female	26	80000	no
12	15606274	Female	26	52000	no
13	15746139	Male	20	86000	no

CLASSIFICATION ASSESSMENT

Student practice

- Data set: breast cancer diagnosis based on tumor characteristics
- Objectives:
  - Work again on previous subject
  - Assess and compare the models

mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	tumor type
14.69	13.98	98.22	656.1	0.10310	0.18360	0.14500	0.06300	0.2086	0.07406	benign
13.17	18.66	85.98	534.6	0.11580	0.12310	0.12260	0.07340	0.2128	0.06777	malignant
12.95	16.02	83.14	513.7	0.10050	0.07943	0.06155	0.03370	0.1730	0.06470	benign
18.31	18.58	118.60	1041.0	0.08588	0.08468	0.08169	0.05814	0.1621	0.05425	malignant
15.13	29.81	96.71	719.5	0.08320	0.04605	0.04686	0.02739	0.1852	0.05294	malignant
16.16	21.54	106.20	809.8	0.10080	0.12840	0.10430	0.05613	0.2160	0.05891	malignant
19.19	15.94	126.30	1157.0	0.08694	0.11850	0.11930	0.09667	0.1741	0.05176	malignant
18.08	21.84	117.40	1024.0	0.07371	0.08642	0.11030	0.05778	0.1770	0.05340	malignant