NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

MW 4:30 – 5:50, NFN 203

Slide Set #16
Introduction to Machine Learning

Lecture Outline

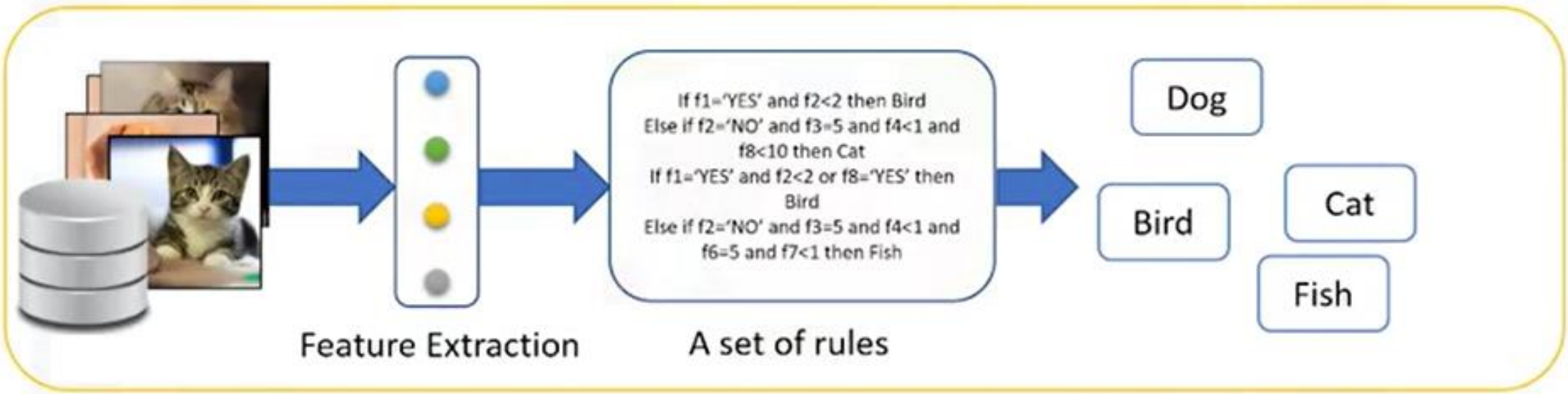
- Machine learning
  - Definition
  - Categories and techniques
- Model Evaluation
- Evaluation Metrics
  - Regression
  - Classification
- Logistic Regression

Machine Learning Defined

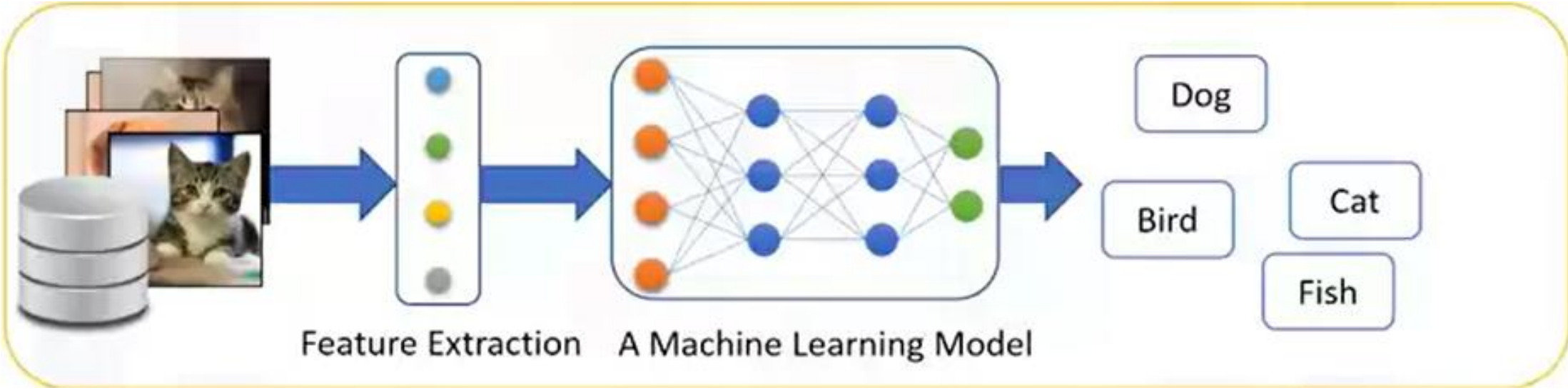
Machine learning is the subfield of computer science that give **“computers the ability to learn without being explicitly programmed.”**

Arthur Samuel
American pioneer in the field of computer gaming and artificial intelligence, coined the term “machine learning” in 1959 while at IBM.

Traditional Approach



ML Model Approach



Major Machine Learning Techniques

- Regression/Estimation
  - Prediction of continuous variables
- Classification
  - Predicting a class/category of an observation
- Clustering
  - Finding the structure of data; uncover groupings between features
- Associations
  - Finding items/events that frequently co-occur

Major Machine Learning Techniques

- Anomaly detection
  - Discover abnormal and unusual cases
- Sequence mining
  - Predicting the next *event*; word in a sentence (Markov Model, HMM)
- Dimension Reduction
  - Reducing the size of data (PCA)
- Recommendation systems
  - Recommend items

Supervised vs Unsupervised Learning

- Supervised learning –
  - Definition: **Supervise –** to observe and direct the execution of a task, project or activity.
  - We supervise an ML model by *training* it with labeled data, such as:

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

  - Two types
    - Classification
    - Regression

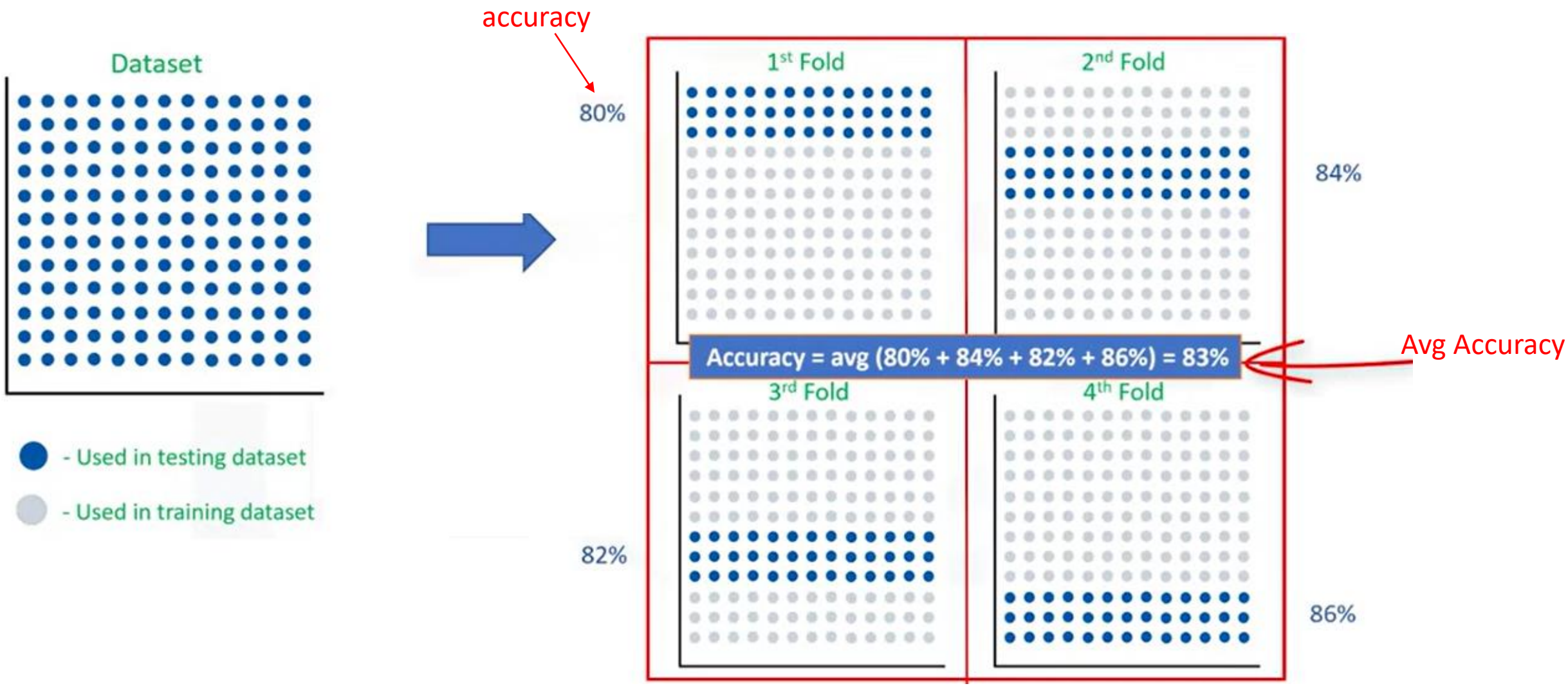
Supervised vs Unsupervised Learning

- Unsupervised learning –
  - Exactly what you might expect based on the definition of supervised learning.
  - Unsupervised models work on their own to discover information about the data that may not be apparent to a human evaluator
  - An unsupervised algorithm trains on the dataset and draws conclusions about unlabeled data
  - Common types:
    - Dimension reduction
    - Density estimation
    - Clustering

Model Evaluation

- Two types of evaluation approaches
  - Train and test on the same dataset
    - Results in over training
  - Train/Test Split
    - A portion of the data is reserved for testing and the rest is used for training.
    - The test set truly represents “out of sample data”
    - However, particularly with small datasets, biases may occur in the splits that lead to pronounced variations in the model’s performance

Model Evaluation – K-fold Cross Validation



Evaluation Metrics

- How accurate is the classifier model?
  - Predictions vs actual values in a test set
- There are a number of different evaluation metrics for classification models, including:
  - Jaccard index
  - F1-score
  - Log Loss

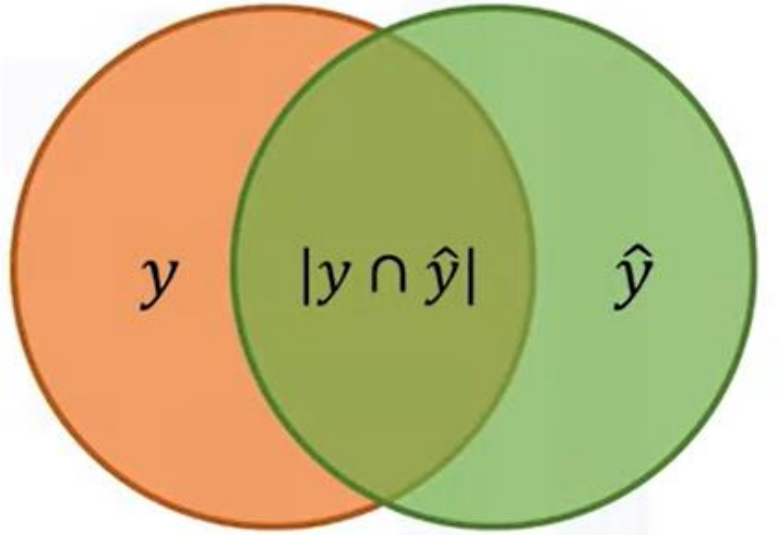
Jaccard Index

- One of the simplest measures of accuracy
- Also known as the “Jaccard similarity coefficient”

y: Actual labels

$\hat{y}$: Predicted labels

$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$



y: [0, 0, 0, 0, 0, 1, 1, 1, ,1 ,1]

$\hat{y}$: [1, 1, 0, 0, 0, 1, 1, 1, ,1 ,1]

$$J(y,\hat{y}) = \frac{8}{10+10-8} = 0.66$$

$J(y,\hat{y}) = 1.0$

Higher Accuracy

F1 - score

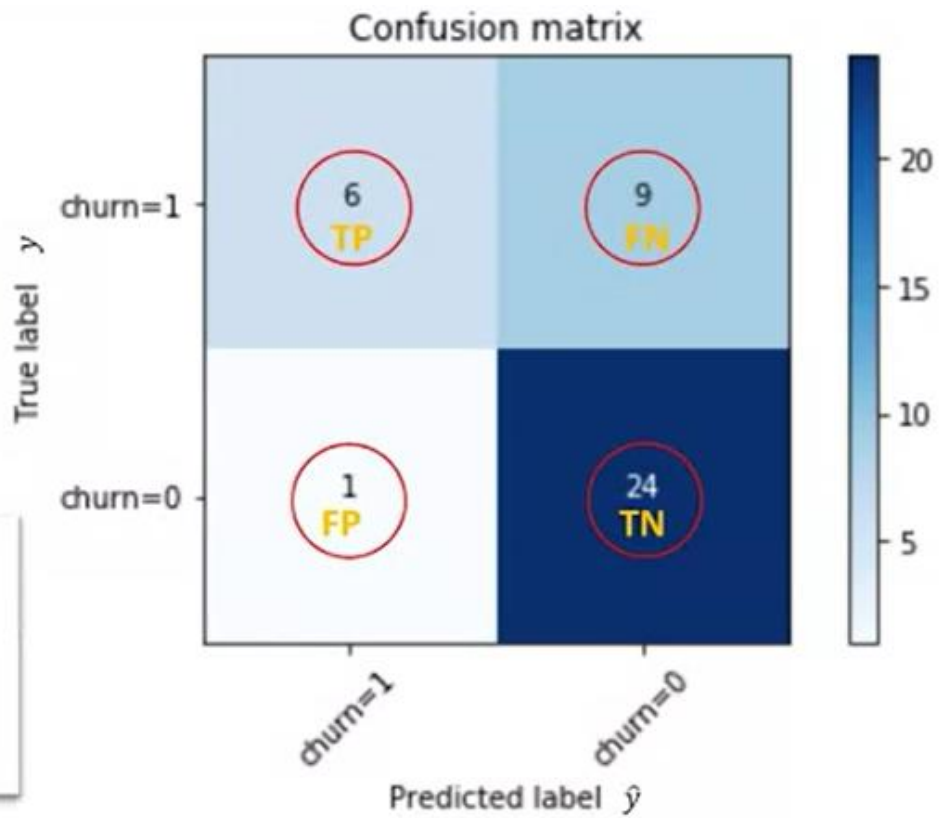
AKA – "Sensitivity"

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-score = 2x (prc x rec) / (prc+rec)

F1-score: 0.00 ... 0.20 .... 0.55 .... 0.83 ... 1.00

Higher Accuracy

	precision	recall	f1-score
Churn = 0	0.73	0.96	0.83
Churn = 1	0.86	0.40	0.55

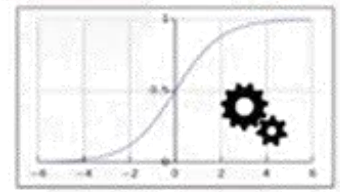


Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.

Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Test

Predicted churn	LogLoss
0.91	0.094
0.13	0.89
0.04	0.04
0.23	0.26
0.43	0.56

$LogLoss = 0.60$

Actual Labels y

$\hat{y}$ Predicted Probability

$$LogLoss = -\frac{1}{n}\sum(y \times \log(\hat{y}) + (1-y) \times \log(1-\hat{y}))$$

Log loss of an individual row

LogLoss: 0.00 ... 0.35 .... 0.60 ... 1.00

Higher Accuracy

Sensitivity and Specificity

Sensitivity, recall, hit rate, or true positive rate (TPR)

$$TPR = TP/P = TP/(TP + FN)$$

Specificity, selectivity or true negative rate (TNR)

$$TNR = TN/N = TN/(TN + FP)$$

Precision or positive predictive vale (PPV)

$$PPV = TP/TP + FP$$

True Positive = A test result that correctly indicates the presence of a condition or characteristic
True Negative = A test result that correctly indicates the absence of a condition or characteristic

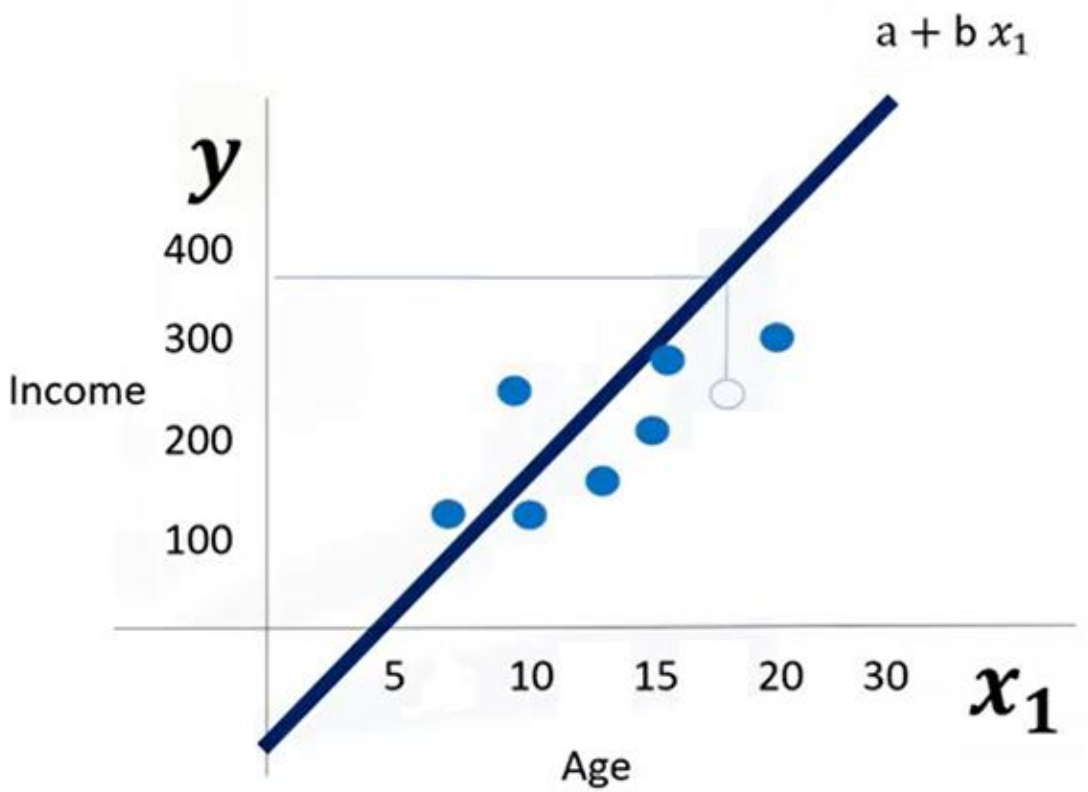
False Positive (Type I error) = A test result that wrongly indicates the presence of a condition or characteristic
False Negative (Type II error) = A test result that wrongly indicates the absence of a condition or characteristic

Classification with Logistic Regression

- Statistical and machine learning technique for classifying observations of a dataset based on values in input fields (“predictors”)
- Analogous to linear regression, but tries to predict a categorical or discrete value instead of a numeric one
  - binary value such as “true/false”, “positive/negative”, etc.. All of which may be coded as 0 or 1
- Predictor variables should be continuous. If categorical, they should be transformed to a continuous value (“dummy” or “indicator” coded)
- Logistic regression returns a probability score between 0 and 1 for which category an observation belongs to
- As with all classifiers, logistic regression use a “decision boundary” (threshold value) to determine which class an observation belongs to

Linear Regression - reminder

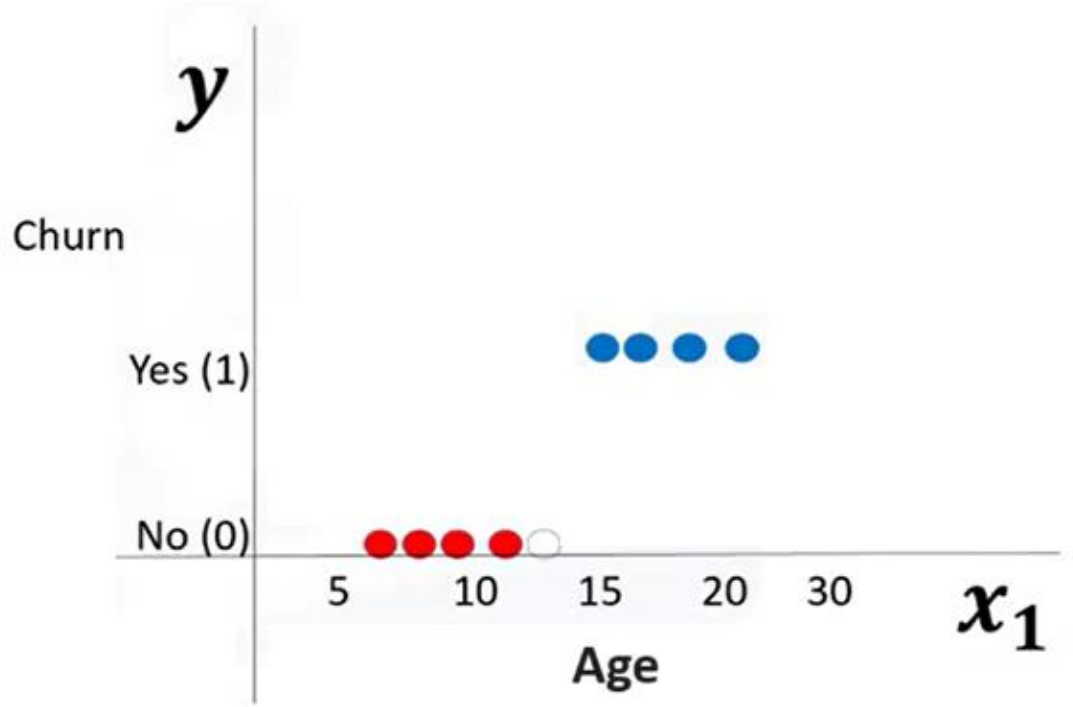
	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Linear Regression for Classification?

"churn" = customer loss

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Linear Regression for Classification?

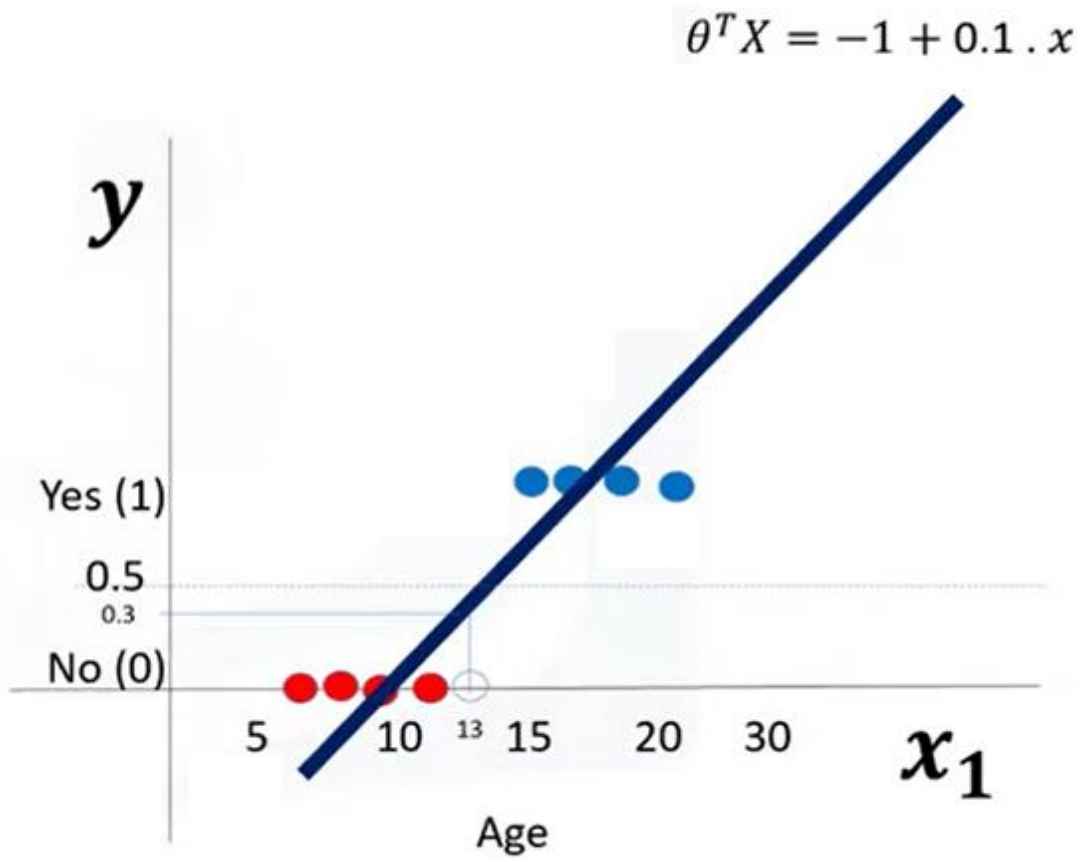
$$\theta^T X = \theta_0 + \theta_1 x_1$$

$p_1 = [13]$ → $\theta^T X = -1 + 0.1 \,.\, x_1$

= -1 + 0.1 x 13

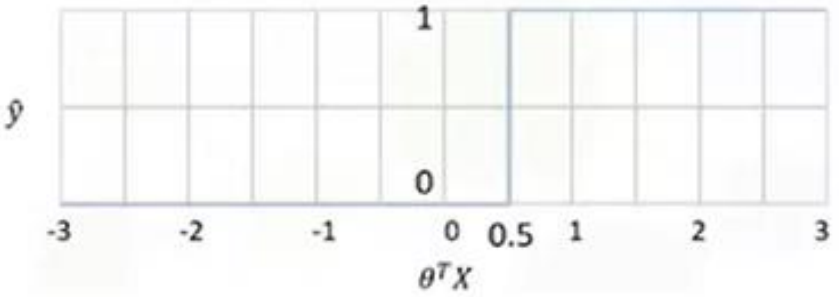
= 0.3

$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$



Linear Regression for Classification?

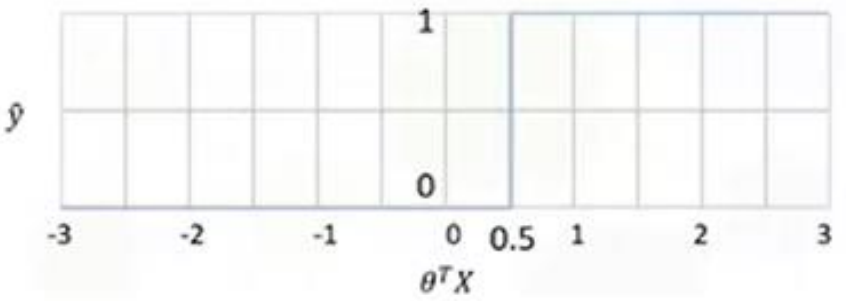
$$\theta^T X = \theta_0 + \theta_1 x_1 + \cdots$$



$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

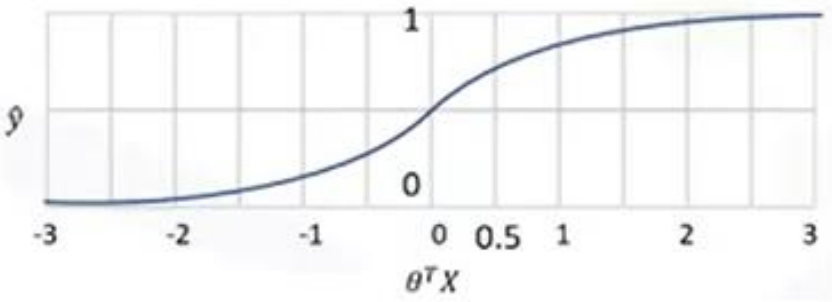
The Sigmoid Function for Logistic Regression

$$\theta^T X = \theta_0 + \theta_1 x_1 + \cdots$$



$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \cdots)$$



$$\hat{y} = \sigma(\theta^T X)$$

The Sigmoid (AKA Logistic) Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$\sigma(\theta^T X) = 1$

$\sigma(\theta^T X) = 0$

$[0, 1]$