Classification Problems and Logistic Regression

Classification Problems

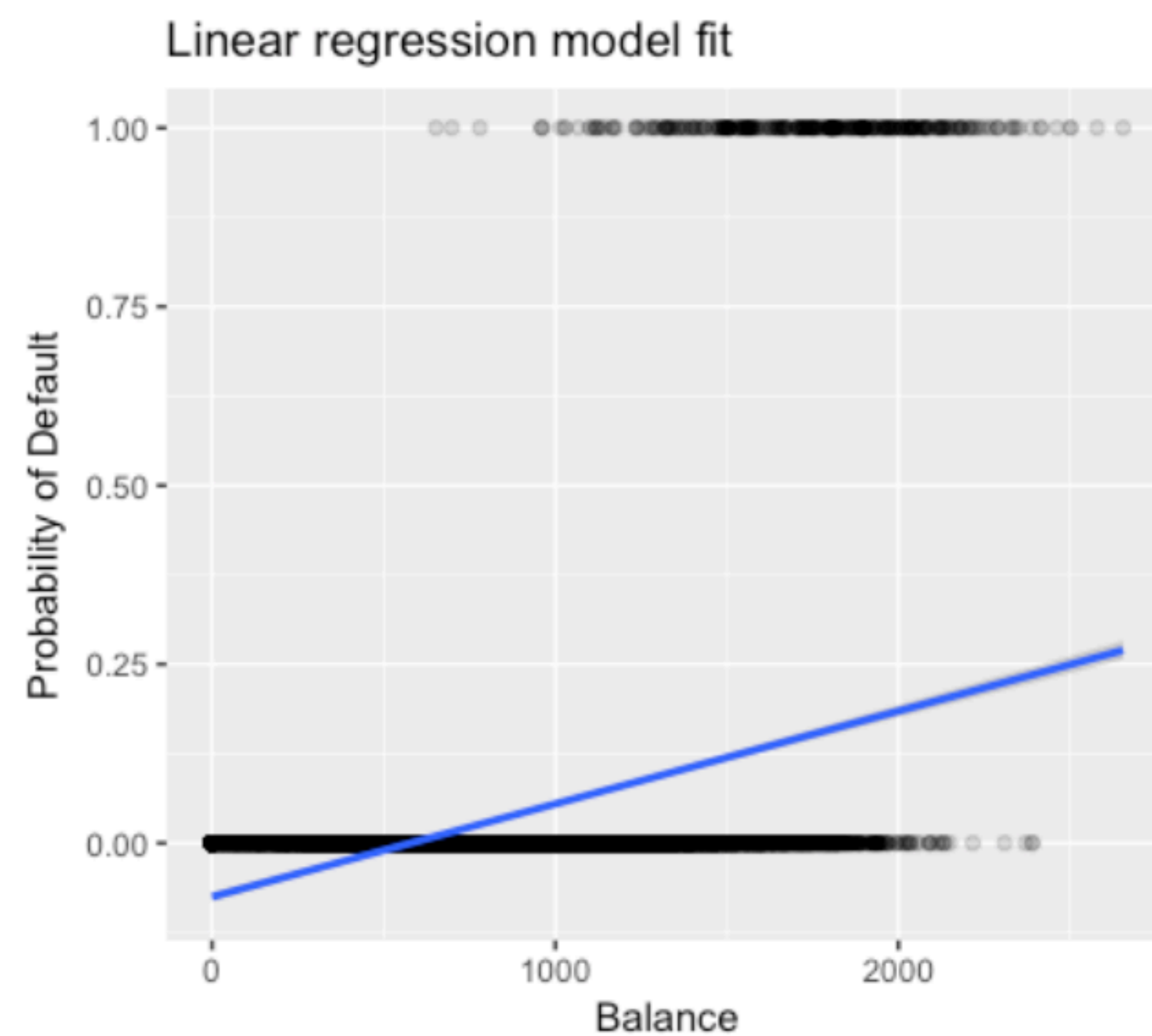
- So far we've only seen regression problems where we try to predict a continuous value.
- The convention for **binary classification** is to have two classes 0 and 1, which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick.
- **Logistic regression** allows us to solve binary classification problems, where we are trying to predict discrete categories.
- Cases where the dependent variable has more than two outcome categories may be analysed with **multinomial logistic regression** (not covered in this course).

Why not Linear Regression?

- Consider using linear regression for modelling *default* **probabilities** depending on *credit card balance.*

- For balances close to zero we predict a negative probability of defaulting; if we were to predict for very large balances, we would get values bigger than 1.

- These predictions are not sensible, since the true probability of defaulting, regardless of credit card balance, must fall between 0 and 1.



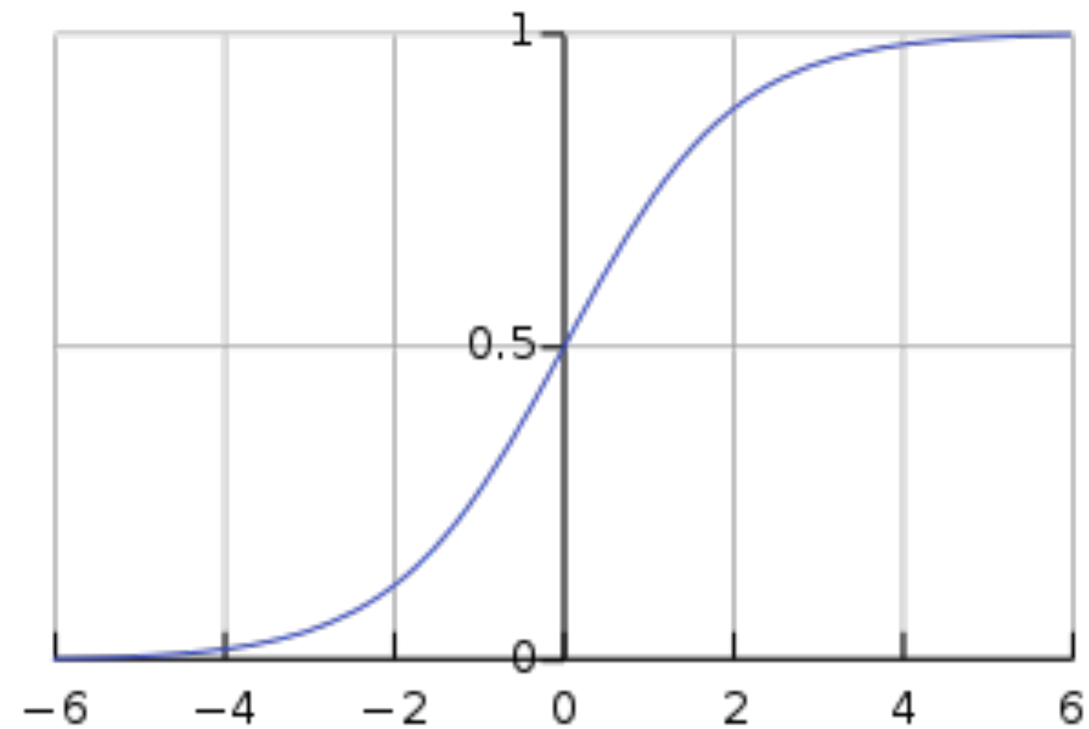
Fälle

Sigmoid Function

- To avoid problem of negative and larger than 1 probabilities, we must model p(X) using a function that gives outputs between 0 and 1 for all values of X.

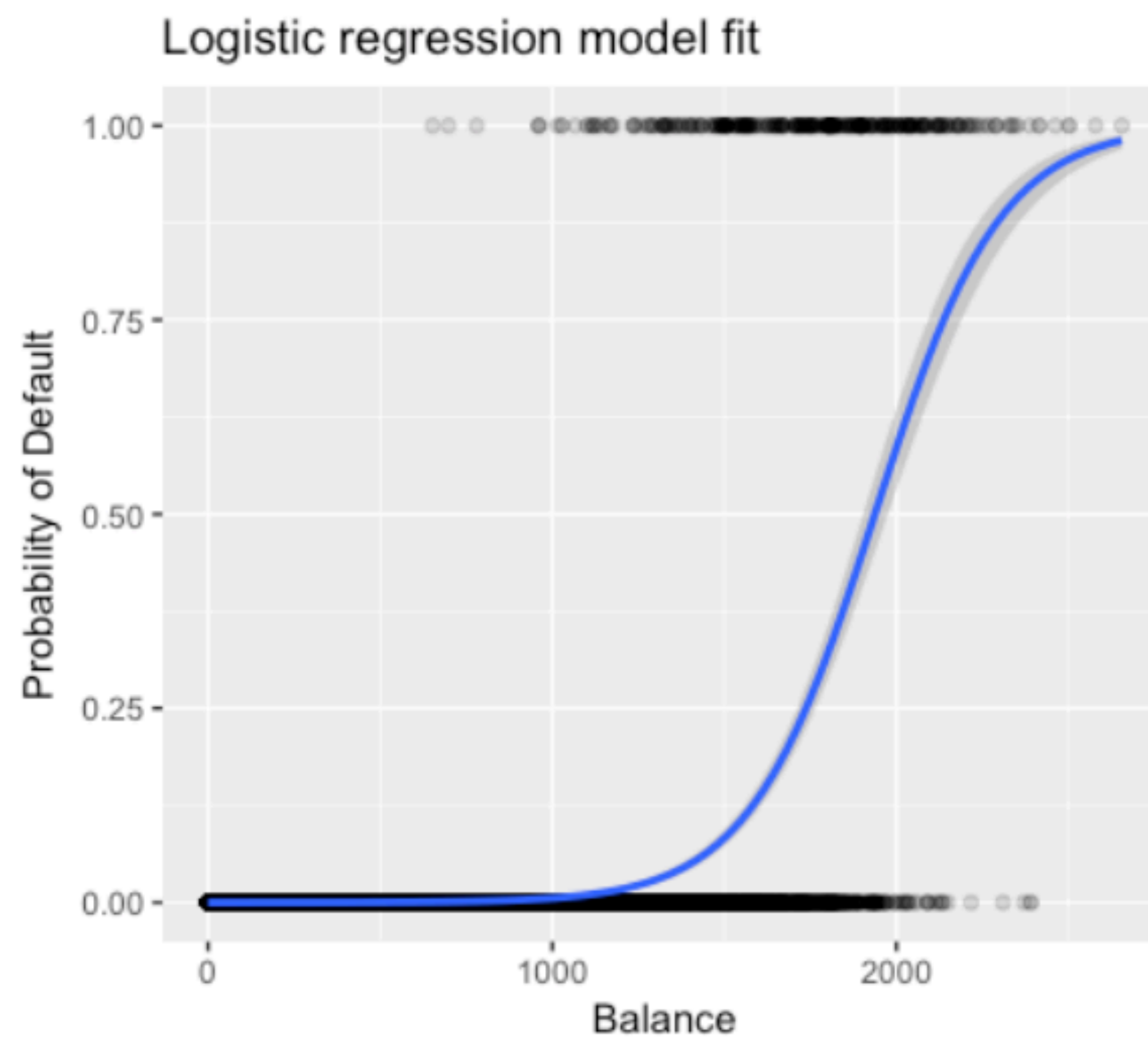
- Many functions meet this description. In logistic regression, we use sigmoid function:

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}.$$



Why not Linear Regression?

- The sigmoid (logistic) function takes in any value and outputs it to be between 0 and 1.
- This means we can take our linear regression solution and fit into the sigmoid function.
- This results in a probability from 0 to 1 of belonging to class 1.



Model Evaluation

- After you train a logistic regression model on some training data, you will evaluate your model's performance on test data.
- To evaluate binary classification models you can use
  - confusion matrix,
  - precision / recall scores,
  - F-score,
  - ROC curve,
  - other.

```
- Confusion
matrix is
output of a
model of one
specific
trashhold
- to evaluate
for all
trashhold
look at ROC
curve
```

Confusion Matrix

- **Confusion matrix** (table of confusion) is a table with two rows and two columns that reports the number of true positives, false positives, true negatives, false negatives

	Predicted: NO	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

- Test for defaulting on credit card payment:
  - YES - will default - positive test - True=1
  - NO - won't default - negative test - False=0

Confusion Matrix - Basic Terminology

	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

- True Positives (TP): Predicted - YES / Actual - Yes
- False Positives (FP): Predicted - Yes / Actual - No
- True Negatives (TN): Predicted - NO / Actual - NO
- False Negatives (FN): Predicted - NO / Actual - Yes

Confusion Matrix - Basic Terminology

- True Positives - *correctly* predicted *positive values*: E.g., if actual class value indicates that the client has defaulted and predicted class tells you the same thing.

- False Positives - *incorrectly* predicted *positive values*: E.g., if the client has not defaulted but predicted class tells that the client will default.

- True Negatives - *correctly* predicted *non-positive values*: E.g., if actual class says the client won't default and predicted class tells the same thing.

- False Negatives - *incorrectly* predicted *non-positive values*: E.g., if actual class value indicates that the client defaulted and predicted class tells you that the client won't default.

Confusion Matrix - Basic Terminology

wichtig zu wissen

Confusion Matrix - Accuracy & Error Rate

	Predicted: NO	Predicted: YES	
Actual: NO	50	10	60
Actual: YES	5	100	105
	55	110	

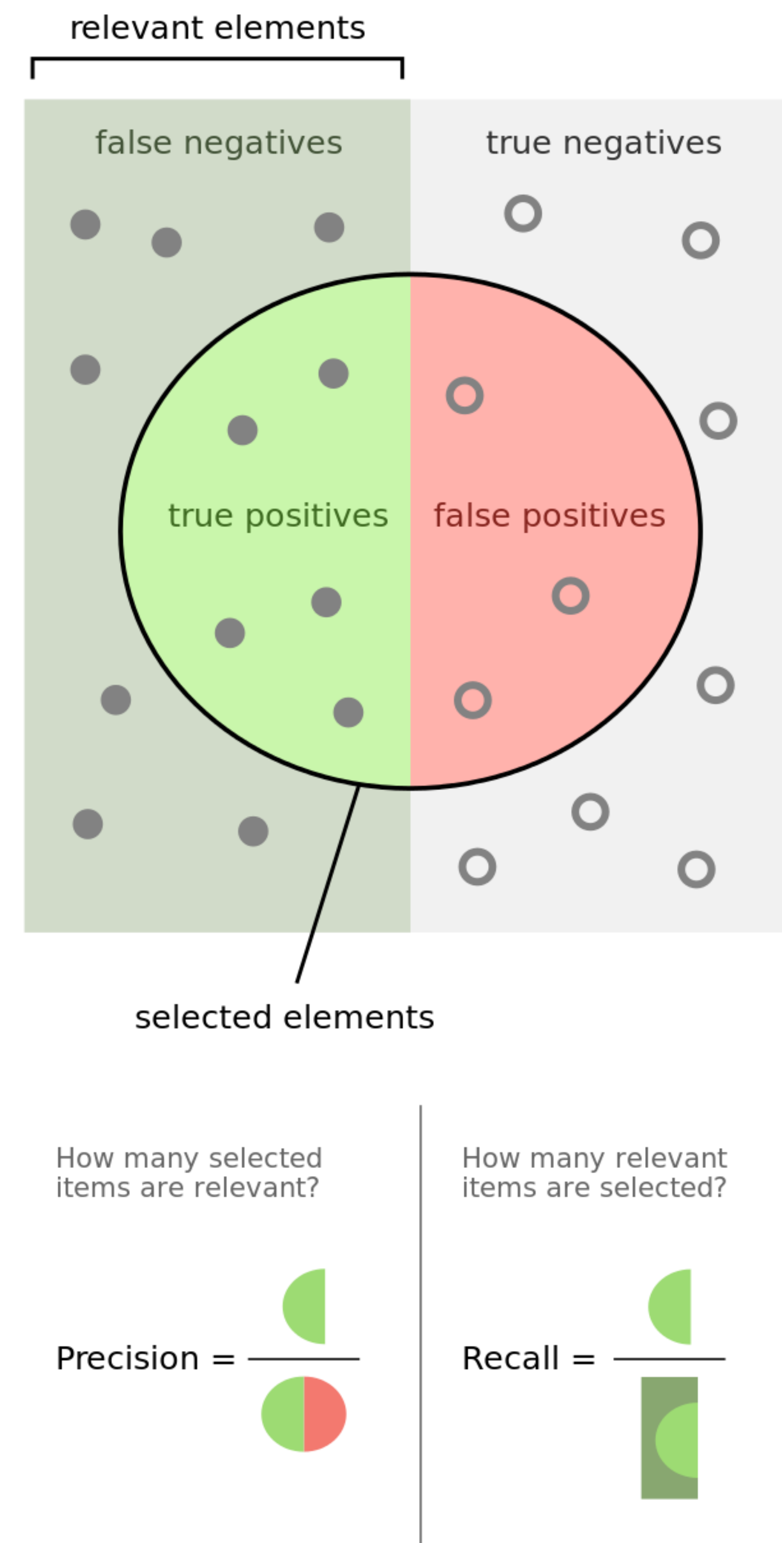
- **Accuracy**
    - How often was classification *correct*?
    - (TP + TN) / total = 150/165 = 0.91
- **Error Rate (Misclassification Rate)**:
    - How often was classification *wrong*?
    - (FP + FN) / total = 15/165 = 0.09

Precision & Recall

- **Precision (Positive Predictive Value)** - ratio of correctly predicted positive observations to the overall positive observations.
  - The question that this metric answer is: Of all clients that labeled as defaulting, how many did actually default?
  - High precision relates to the low false positive rate.
- **Recall (True Positives Rate or Sensitivity**) - ratio of correctly predicted positive observations to the total actual positive observations.
  - The question that this metric answer is: Of all the clients that defaulted, how many did we correctly predict?
  - High recall relates to the low false negative rate.

Precision & Recall

- Precision = TP / (TP + FP)
- Recall = TP / P = TP / (TP + FN)



Source: Wikipedia

F1-Score

- F1-score (F-score) combines precision and recall as the harmonic mean:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- F1-score is between 0 and 1,

- F1-score allows to compare models between each other: in general, the higher is F1-score, the better is the model,

- F1-score does not distinguish between precision and recall, because recall and precision are evenly weighted.

General F-Score

- F1-score is a special case of the general F_β measure

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

- Two other commonly used F measures are
    - F_2 measure: weights recall higher than precision and therefore puts more emphasis on recall than precision,
    - $F_{0.5}$ measure, weights precision higher than recall and therefore puts more emphasis on precision than recall.

Receiver Operating Characteristics Curve

- In binary classification, the class prediction for each instance is often made based on a continuous random variable X, which is a "score" computed for the instance.

- For logistic regression where $\boldsymbol{X}$ is the **estimated probability** and $\boldsymbol{T}$ - **threshold** parameter, then

  - $X>T$ - instance is classified as “positive”,

  - $X \leq T$ - instance is classified as “negative”.

- Adjusting the threshold changes the number of true positives (TP):

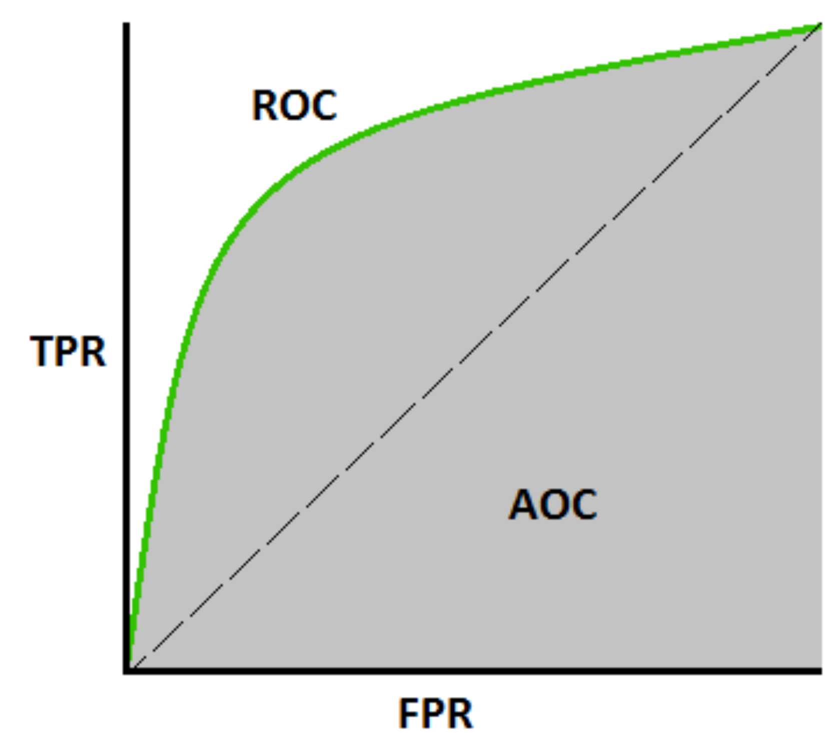
  - if $T=0$, then TP=P,

  - if $T=1$, then TP=0.

jedes mal
wenn man den
trashhodl
ändert ändert
sicvh auch TP
und TN usw

Receiver Operating Characteristics Curve

- **Receiver Operating Characteristics (ROC) curve** is a performance measurement for classification problem at various **thresholds** settings.
  - It tells how much model is capable of distinguishing between classes.
  - It is plotted with **True Positives Rate** (TPR or recall or sensitivity) at y-axis against the **False Positives Rate** (FPR or false alarm rate) at x-axis, where
    - TPR = TP/P = TP / (TP + FN),
    - FPR = FP/N = FP / (FP + TN),
    - The higher is the **Area Under the Curve** (**AUC**), the better is the model at distinguishing between the classes.
    - Diagonal dashed blue line approximates the line of the classification that is based on a random guess.



ROC-Space

A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

TPR = 0.63

FPR = 0.28

B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

TPR = 0.77

FPR = 0.77

C

TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

TPR = 0.24

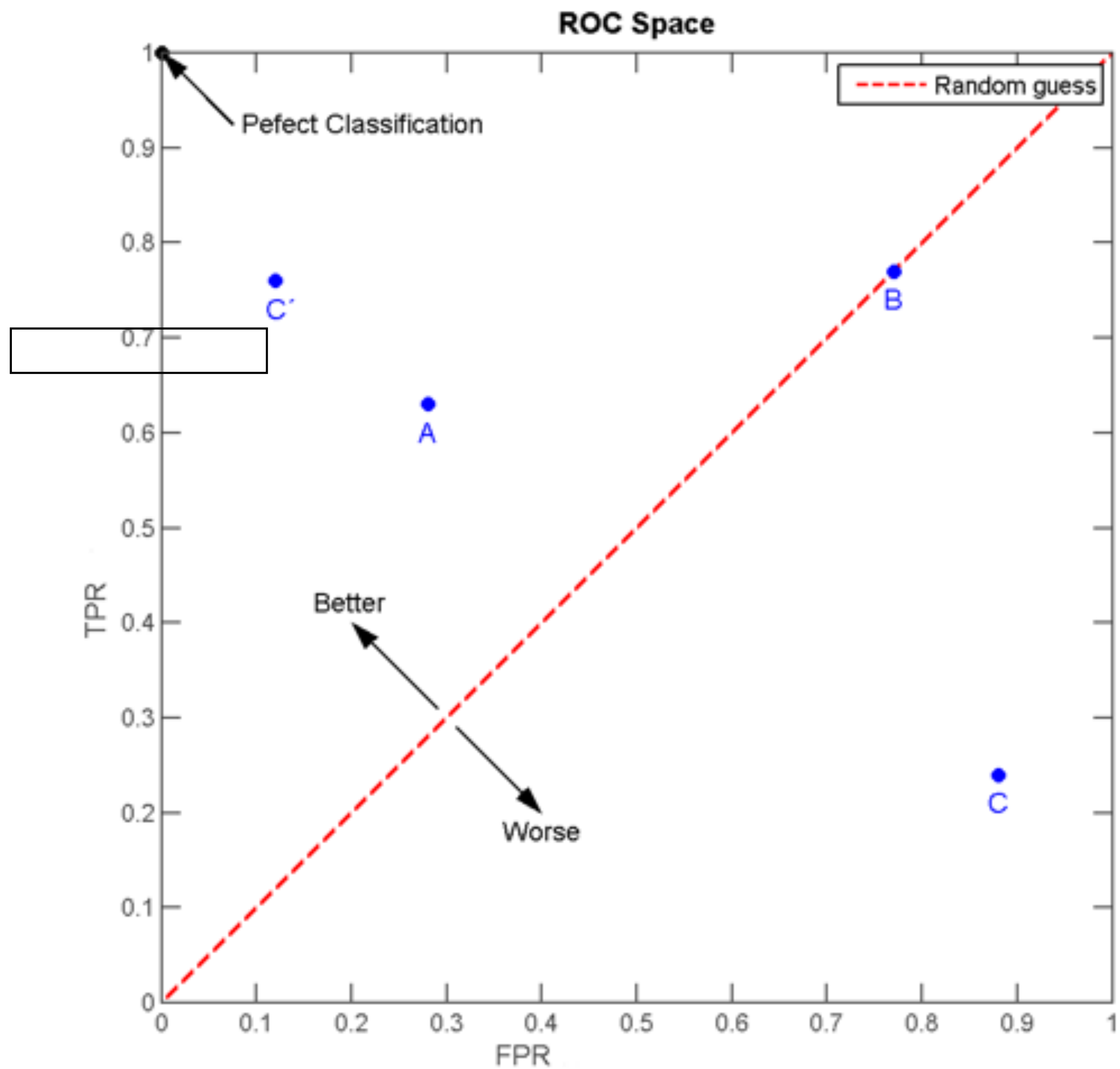
FPR = 0.88

C′

TP=76	FP=12	88
FN=24	TN=88	112
100	100	200

TPR = 0.76

FPR = 0.12



Source: Wikipedia

ROC-Curve and AUC

- ROC-curve is one of the most important evaluation metrics for checking any classification model's performance
  - ROC-curves of different models can be compared directly in general or for different thresholds.
  - The area under the curve (AUC) can be used as a summary of the model skill.
    - AUC=1.0 signifies perfect classification accuracy,
    - AUC=0.5 is the accuracy of making classification decisions via coin toss (or rather a continuous coin that outputs values in [0,1])