Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 7
Binary Logistic Regression

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Logistic regression

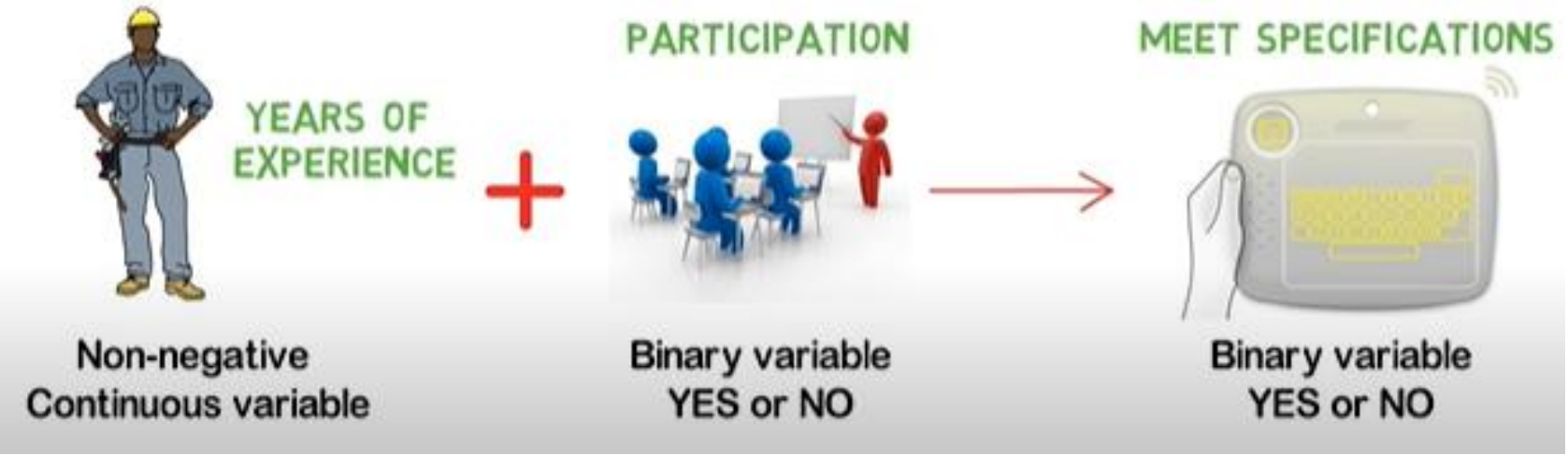
Logistic regression is a multiple regression with an outcome variable that is categorical and predictor variables that are continuous or categorical.

Predictors are continuous or categorical variables

Dependent variable is categorical

Binary Logistic Regression

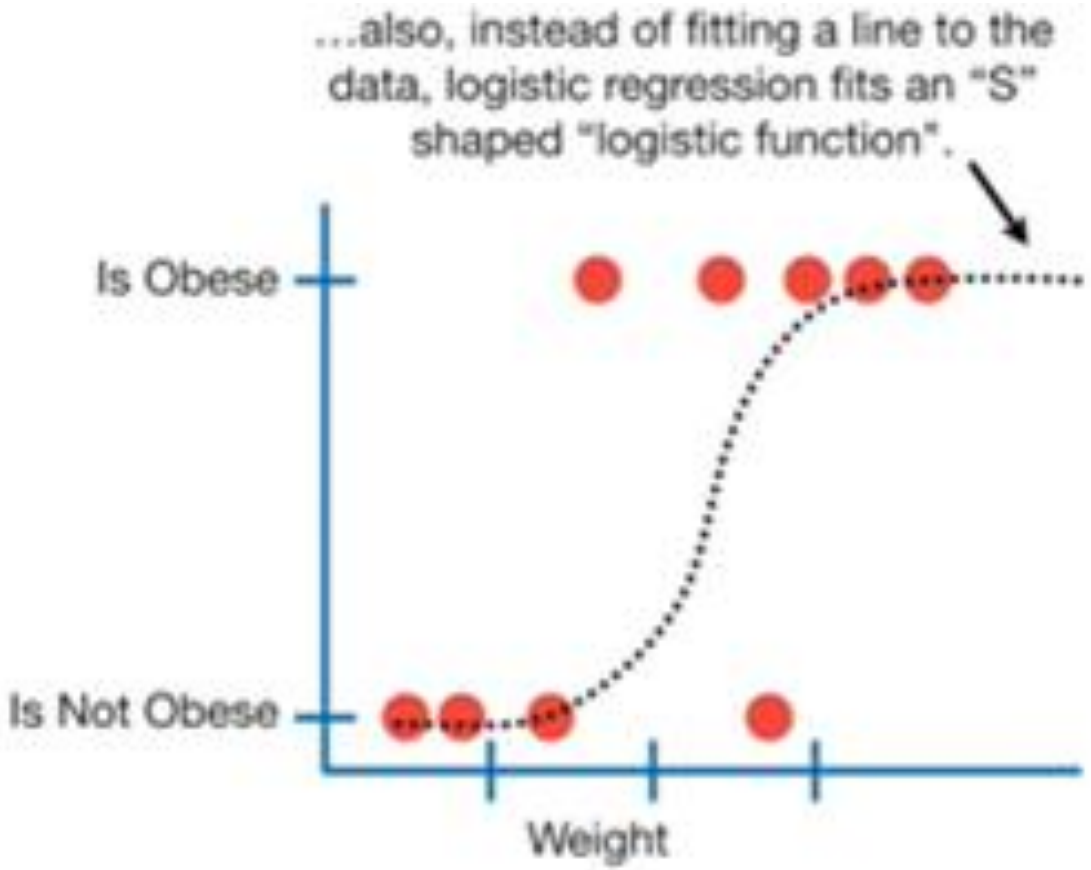
In binary logistic regression the dependent variable is dichotomous. So, we can, for example, predict in which of two categories a person is likely to belong to given certain other information. Permits to evaluate the probability of presence/absence of some characteristic.



Source: https://www.youtube.com/watch?v=4UI9qZw7DXg

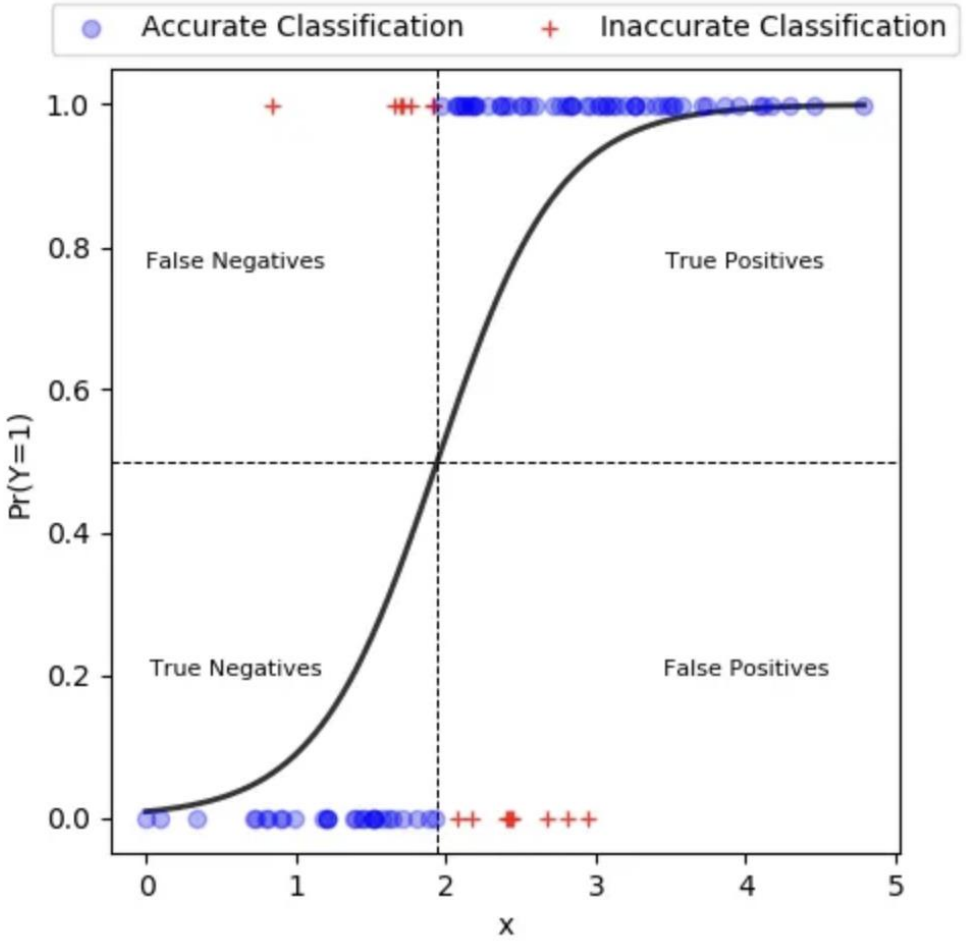
Binary Regression

The curve goes from 0 to 1. The curve tells the probability that a person is obese or not based on the weight. The higher the weight the grater is a probability that the person is obese.



Source: https://www.youtube.com/watch?v=yIYKR4sgzI8

Binary Regression



Probability of Y occurring

In logistic regression, instead of predicting the value of a variable Y from predictor variables we predict the ***probability*** of Y occurring given known values of the predictor variables.

$$p = \frac{1}{1 + e^{-z}}$$

p – probability of Y occurring,
e – base of natural logarithms,
$z = b_1 * X_1 + b_2 * X_2 + \ldots + b_n * X_n + a$,
a – constant,
$X_1, X_2 \ldots X_n$ *– regression coefficients.*

Why can’t we apply linear regression?

We can’t apply linear regression directly to a situation is which the outcome variable is dichotomous because one of the assumptions of linear regression is that the relationship between variables is linear.

In the binary logistic regression, we transform the data using the logarithmic transformation. This has the effect of making the form of the relationship linear whilst leaving the relationship itself as non-linear. In the formula above we express the multiple linear regression equation in logarithmic terms.

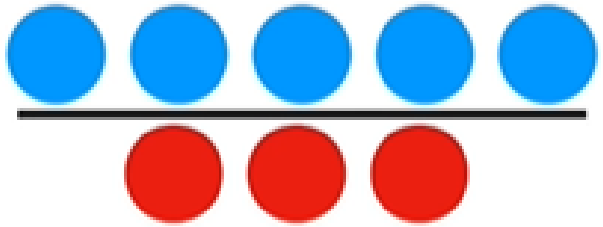
Interpretation of the probability

The resulting value from the equation is a probability value that varies between 0 and 1. It's the probability of occurring an event that corresponds of the highest value of Y. A value close to 0 means that Y is very unlikely to have occurred, and a value close to 1 means that Y is very likely to have occurred. If the probability is less than 0.5, we will conclude that it's unlikely that the event will occur.

Odds and Probabilities

The odds are the ratio of something happening (i.e. my team winning)…

…to something not happening (i.e. my team not winning).



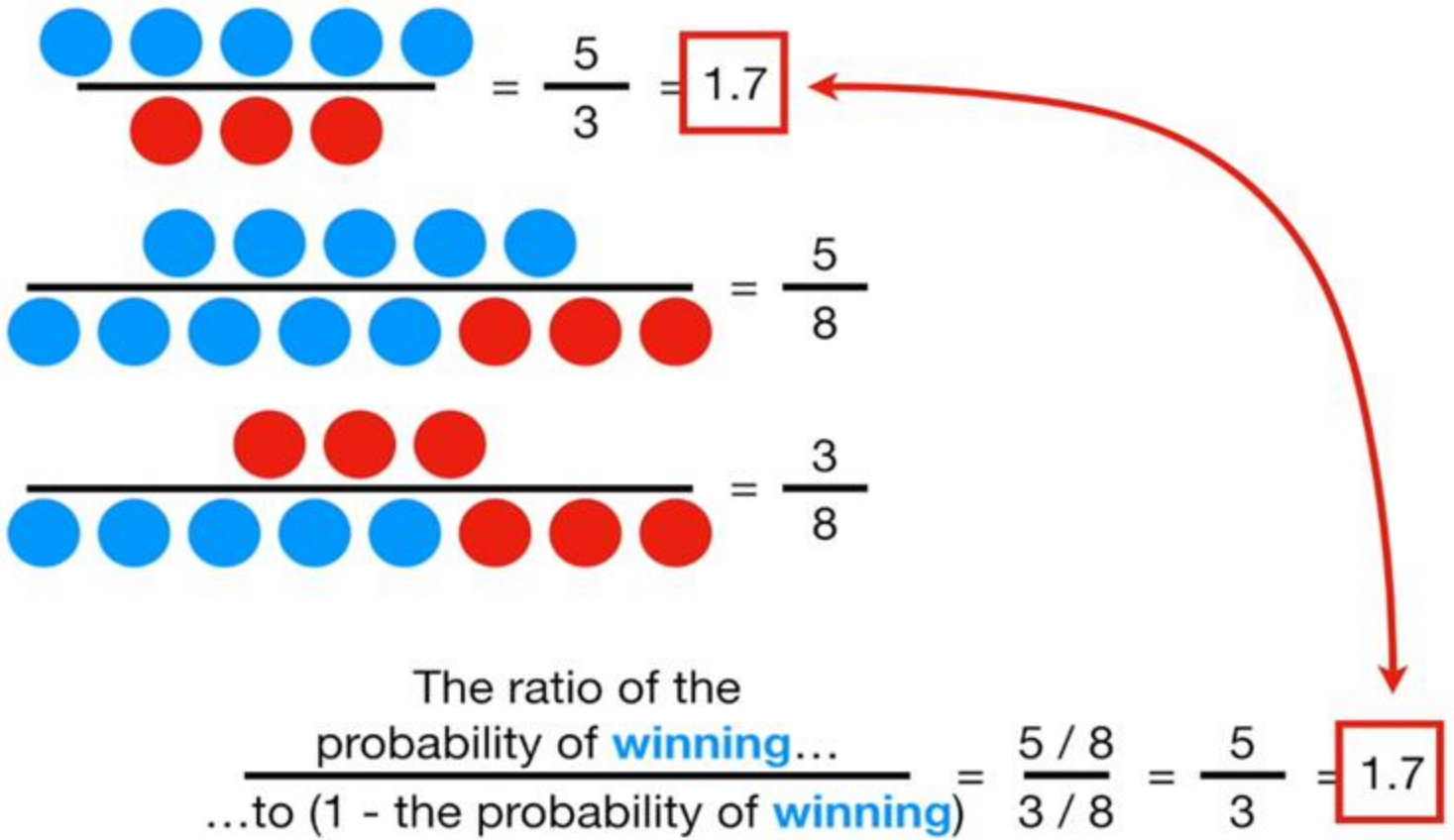
$$odds = \frac{p}{(1 - p)}$$

Probability is the ratio of something happening (i.e. my team winning)…

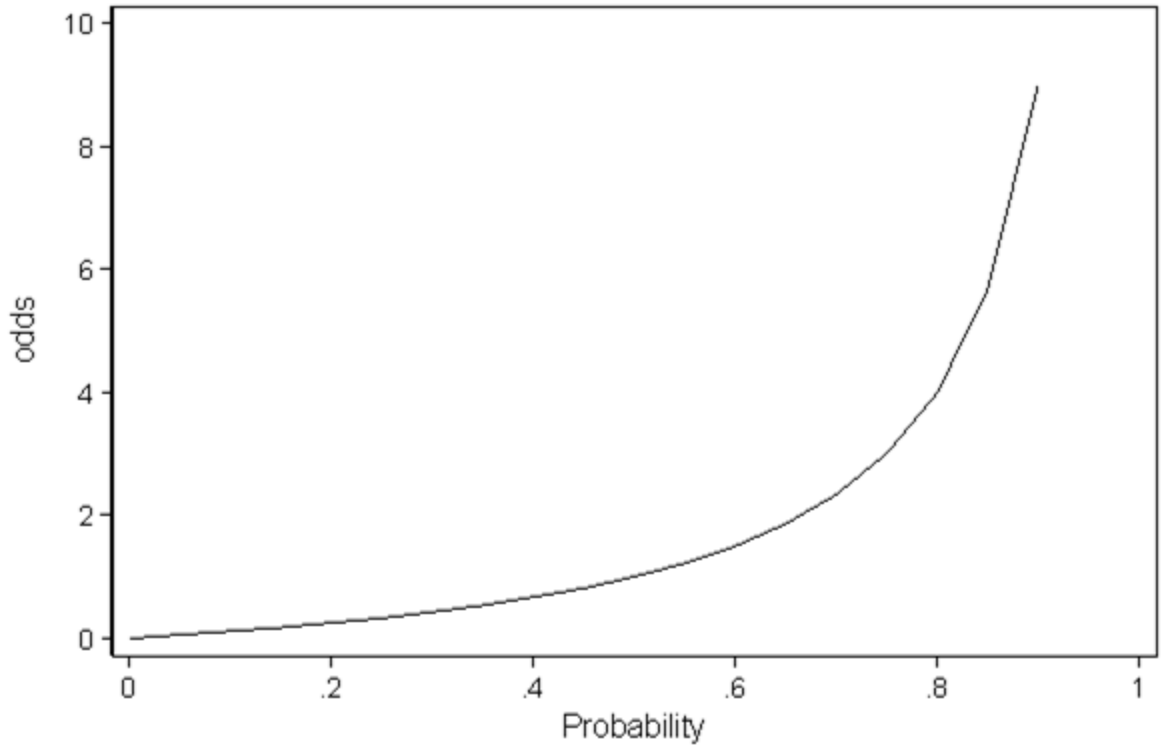
…to *everything* that could happen (i.e. my team winning *and* losing).

Source: https://www.youtube.com/watch?v=ARfXDSkQf1Y

Odds and Probabilities



Odds and Probabilities



$$1{,}5 = \frac{0{,}6}{1 - 0{,}6}$$

Probability	0,01	0,2	0,3	0,5	0,6
Odds	0,0101	0,25	0,42	1	1,5

Odds ratio

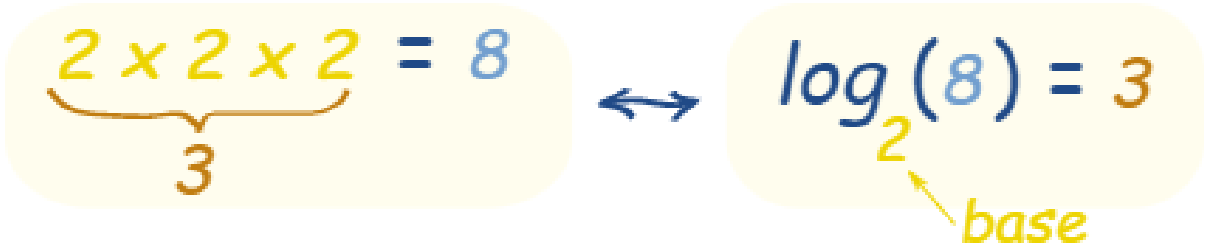
		Has Cancer: Yes	Has Cancer: No
Has the mutated gene	Yes	23	117
	No	6	210

$$\frac{\frac{23}{117}}{\frac{6}{210}} = \frac{0.2}{0.03} = 6.88$$

…and the odds ratio tells us that the odds are 6.88 times greater that someone with the mutated gene will also have cancer.

Source: https://www.youtube.com/watch?v=8nm0G-1uJzA

Logarithm



The number we multiply is called the "base", so we can say:

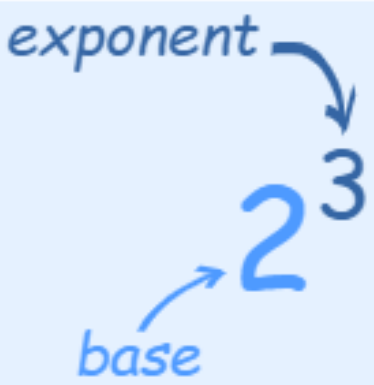
- "the logarithm of 8 with base 2 is 3"
- or "log base 2 of 8 is 3"
- or "the base-2 log of 8 is 3"

$\log_5 25 = 2 \quad \text{as} \quad 5^2 = 25$

$\log_3 81 = 4 \quad \text{as} \quad 3^4 = 81$

Exponent

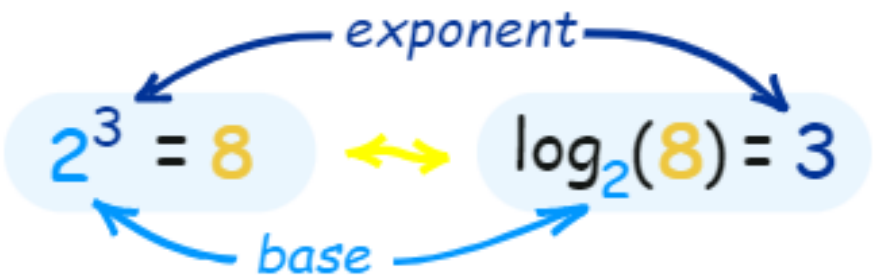
Exponents and Logarithms are related, let's find out how ...



The **exponent** says **how many times** to use the number in a multiplication.

In this example: $2^3 = 2 \times 2 \times 2 = 8$

(2 is used 3 times in a multiplication to get 8)



Natural Logarithm

$$f(x) = \exp(x) = e^x$$

Natural Logarithms: Base "e"

Another base that is often used is e (Euler's Number) which is about 2,71828.

This is called a "natural logarithm". Mathematicians use this one a lot.

On a calculator it is the "ln" button.

It is how many times we need to use "e" in a multiplication, to get our desired number.

Example: $\ln(7{,}389) = \log_e(7{,}389) \approx 2$

Because $2{,}71828^2 \approx 7{,}389$

Source: https://www.mathsisfun.com/algebra/logarithms.html

Exp (B)

Initial number	Exponent	Calculations (in Python)	Result
1	Exp(1)	2.72 ** 1	2.72
2	Exp(2)	2.72 ** 2	7.39
3	Exp(3)	2.72 ** 3	20.1

Exp (B)

Exp(B) is the exponentiation of the B coefficient, which is an odds ratio. Is an indicator of the change in odds resulting from a unit change in the predictor.

Dependent variable: the student has passed the exam or not.

Predictor variable: the student prepared for the exam or not.

The odds of passing the exam are the probability of passing the exam divided by the probability of not passing it.

To calculate the change in odds that results from a unit change in the predictor we first calculate the odds of passing the exam given that there was no preparation for it. Then we calculate the odds of passing the exam given that there was a preparation. Finally, we calculate the proportionate change in these two odds.

Exp (B): interpretation

- The value greater than 1 indicates that as the predictor increases, the odds of the outcome occurring increase.
- The value less than 1 indicates that that as the predictor increases, the odds of the outcome occurring decrease.
- The value equal to 1 indicates that that as the predictor changes, the odds of the outcome occurring will not change.

Interpretation if the B coefficient is positive

Coef (b)	OR=exp(b)	Interpretation
> 0	> 1	Y – candidate gets the job (positive outcome/negative outcome) X_1 – candidate has higher education (yes/no) X_2 – candidate's years of experience (number of years) Candidates with higher education in comparison to the candidates without higher education (reference group) are more likely to get the job (positive income). Exp(b1)=3.1 -> odds of getting a job for candidates with higher education are 3.1 times higher than for candidates without higher education.

Interpretation if the B coefficient is positive

Coef (b)	OR=exp(b)	Interpretation
> 0	> 1	Y – candidate gets the job (positive outcome/negative outcome) X_1 – candidate has higher education (yes/no) X_2 – candidate's years of experience (number of years) Increase in the years of candidate's experience will increase the odds to get a job (positive income). Exp(b2)=1.6 -> one unit increase in the years of candidate's experience will increase by 1.6 or by 60% the odds of a getting a job.

Interpretation if the B coefficient is zero

Coef (b)	OR=exp(b)	Interpretation
= 0	= 1	Candidates with higher education and without higher education have the same odds to get the job (positive income). Changes in the years of candidate's experience will not change the odds to get the job (positive income).

Interpretation if the B coefficient is negative

Coef (b)	OR=exp(b)	Interpretation
< 0	form 0 to 1	Candidates with higher education in comparison to the candidates without higher education (reference group) are less likely to get the job (positive income). Exp(b1)=0.1 -> odds of getting a job for candidates with higher education are 90% lower than for candidates without higher education. Increase in the years of candidate's experience will decrease the odds to get the job (positive income). Exp(b2)=0.6 -> one unit increase in the years of candidate's experience will decrease by 40% the odds of a getting a job.

Which model is the best?

The chosen model will be the one that, when values of the predictor variable are placed in it results in values of Y closest to the observed values.

So, the values of the parameters are estimated using the ***maximum-likelihood estimation***, which selects coefficients that make the observed values most likely to have occurred.

Assessing the model

In logistic regression we use the observed and predicted values to assess the fit of the model. The measure is called **log-likelihood**.

The log-likelihood is based on summing the probabilities associated with the predicted and actual outcomes. It's analogous to the residual sum of squares in multiple regression in the sense that it's an indicator of how much unexplained information there is after the model has been fitted.

So, the larger the value of the log-likelihood, the more unexplained observations there are. The baseline model is the model with no predictors and has only the constant. There are several pseudo-R-squared to assess the goodness of fit.

Log-likelihood

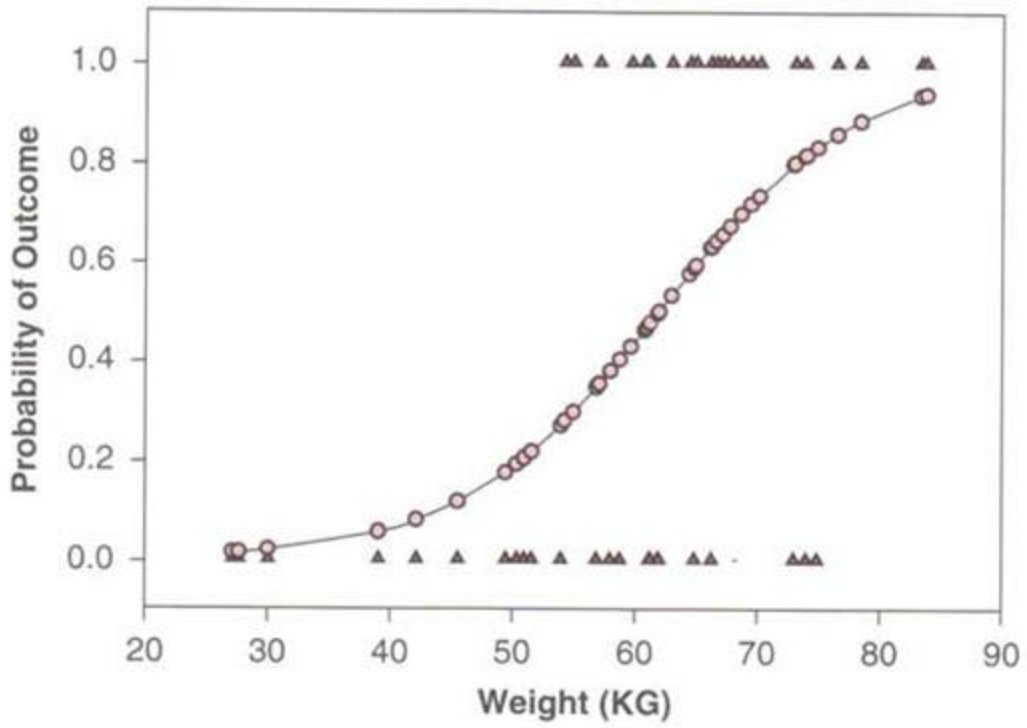
$$\text{log-likelihood} = \sum_{i=1}^{N} \{Y_i \ln(P(Y_i)) + (1 - Y_i) \ln[1 - P(Y_i)]\}$$

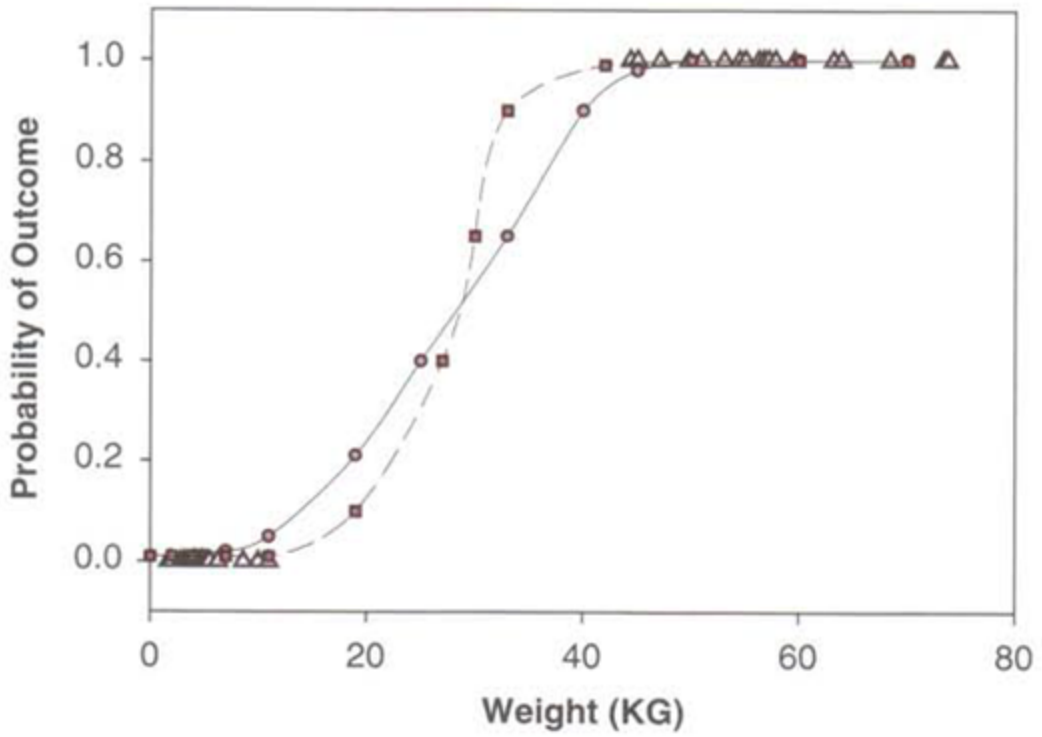
log(0.55) + log(0.55) + log(0.55) + log(0.55) +
log(0.55) + log(1 - 0.55) + log(1 - 0.55) +
log(1 - 0.55) + log(1 - 0.55)

Incomplete information

Do you smoke?	Do you eat tomatoes?	Do you have Cancer?
Yes	No	Yes
Yes	Yes	Yes
No	No	Yes
No	Yes	??????

Complete separation





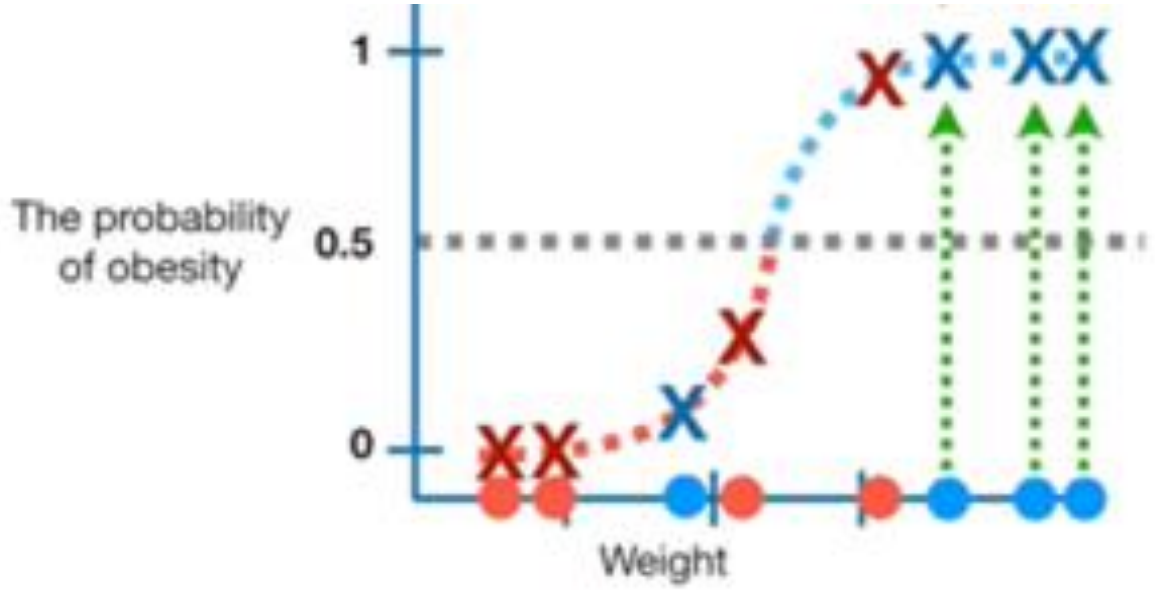
Confusion matrix

Actual		Predicted: Positive	Predicted: Negative
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

P and N - number of observations of positive and negative class.
P = TP + FN, N = TN + FP

Accuracy:	$\frac{TP+TN}{P+N}$
Error rate:	1-accuracy$=\frac{FP+FN}{P+N}$

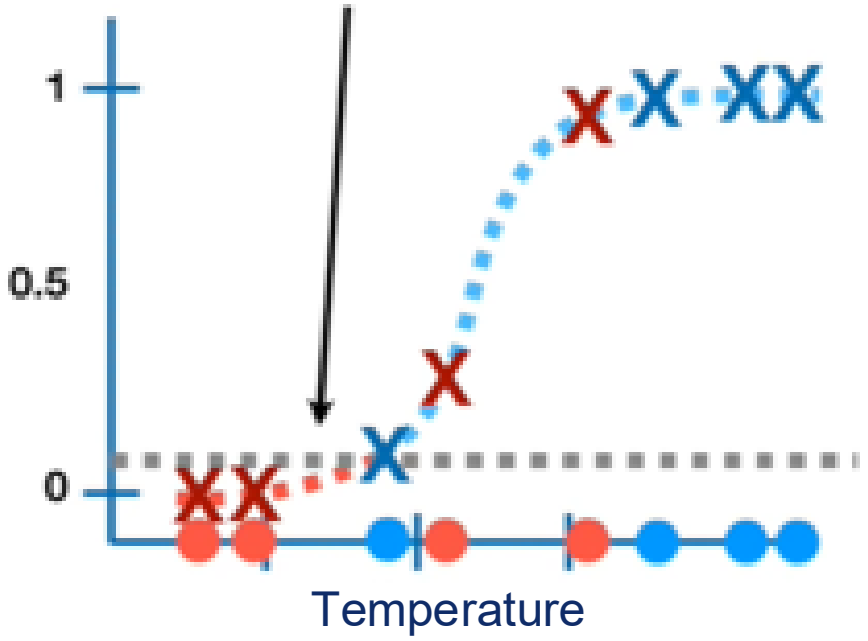
Confusion matrix with threshold 0.5



		Actual: Is Obese	Actual: Is Not Obese
Predicted	Is Obese	3	1
Predicted	Is Not Obese	1	3

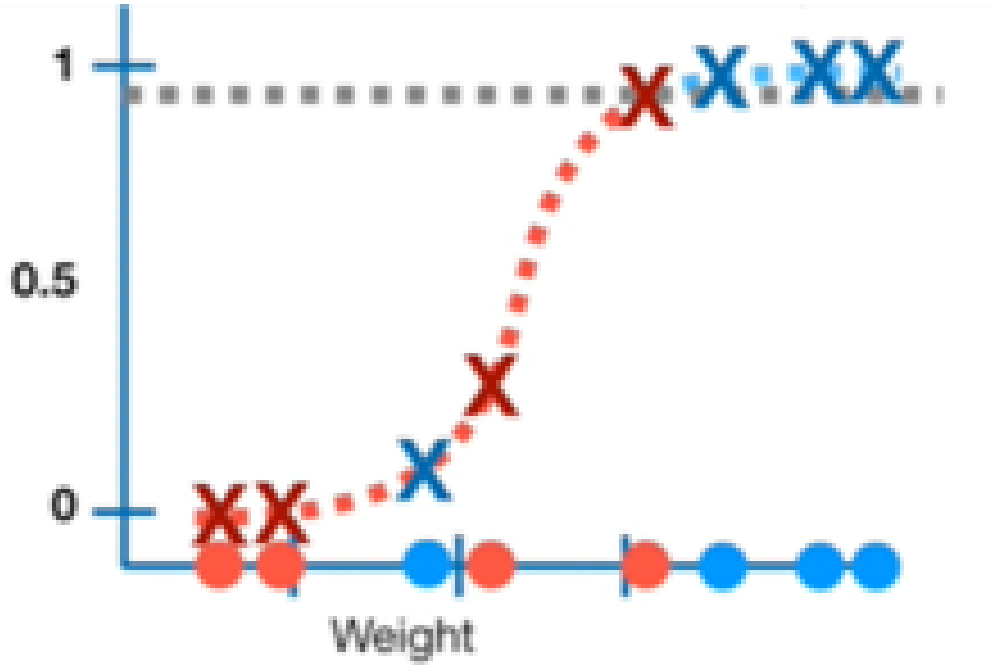
Once the **Confusion Matrix** is filled in, we can calculate **Sensitivity** and **Specificity** to evaluate this Logistic Regression when **0.5** is the threshold for **obesity**.

Confusion matrix with threshold 0.1



		Actual: Infected	Actual: Not Infected
Predicted	Infected	4	2
Predicted	Not Infected	0	2

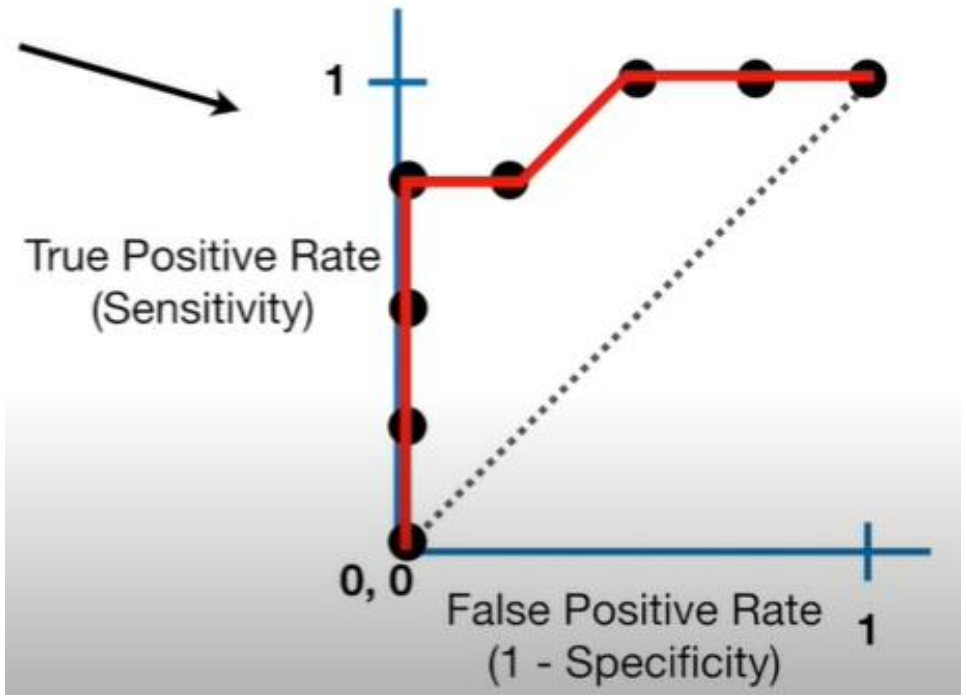
Confusion matrix with threshold 0.9



		Actual: Is Obese	Actual: Is Not Obese
Predicted	Is Obese	3	0
Predicted	Is Not Obese	1	4

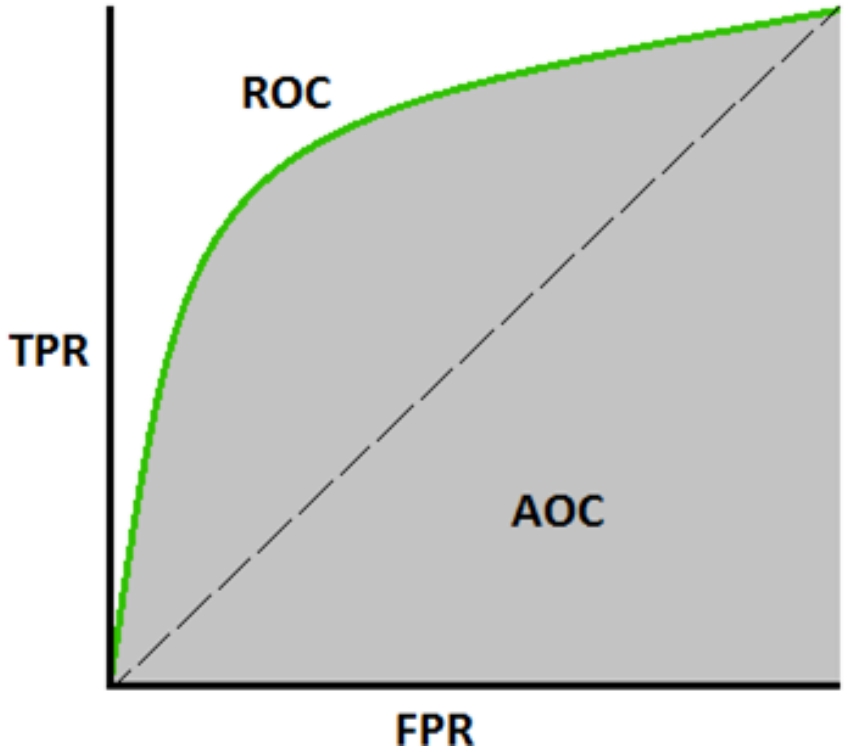
Confusion matrix with different thresholds

The threshold could be set to anything between 0 and 1. If we made one confusion matrix for each threshold that matters, it would result in a large number of different confusion matrices. ROC graph helps to summarize all of the information.



ROC-curve

Receiver Operator Characteristic Curve – a graph that permits to evaluate the quality of the binary classification. It's based on two values – True Positive Rate and False Positive Rate.

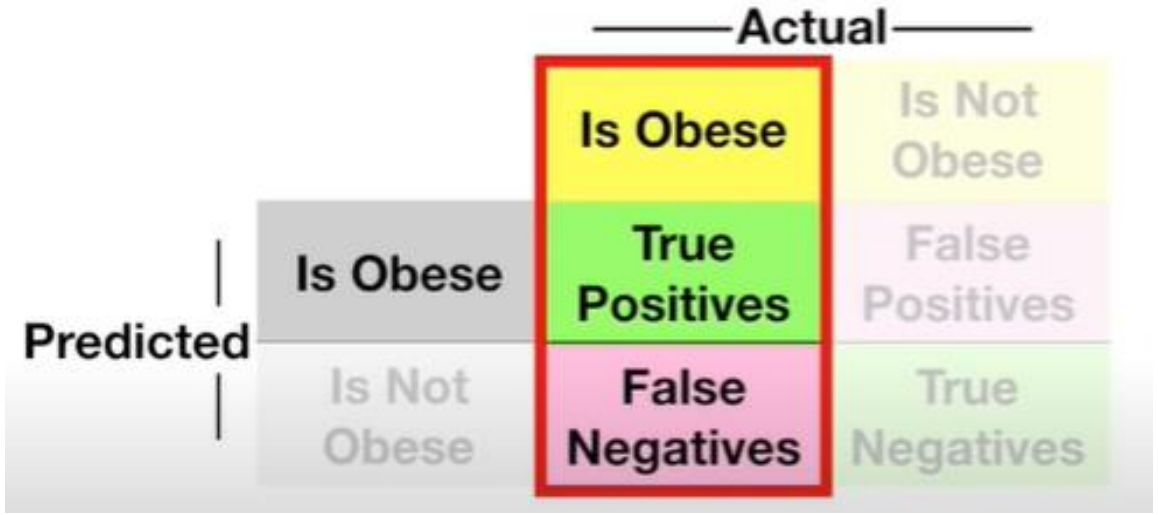


http://www.navan.name/roc/

True Positive Rate

True Positive Rate shows the proportion of cases with positive outcome (value 1), which were correctly classified.

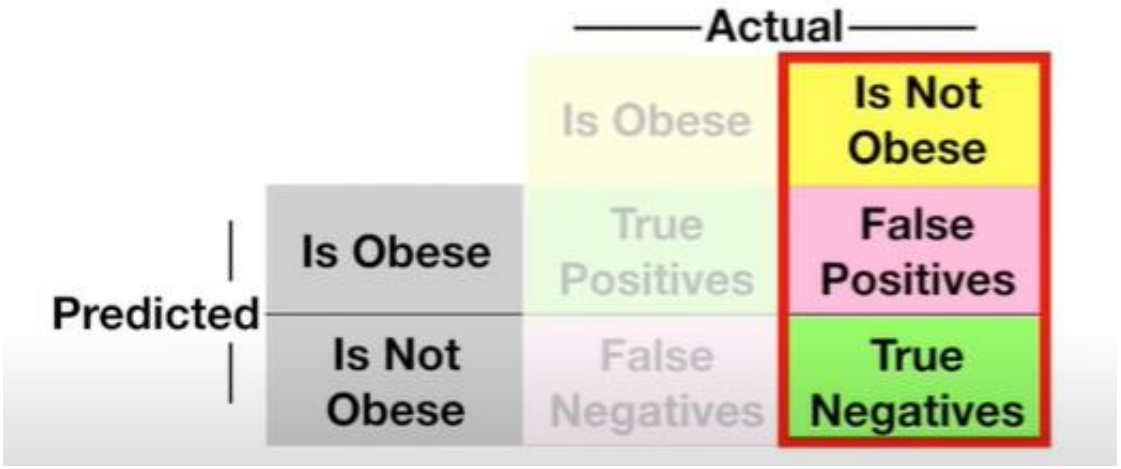
$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$



False Positive Rate

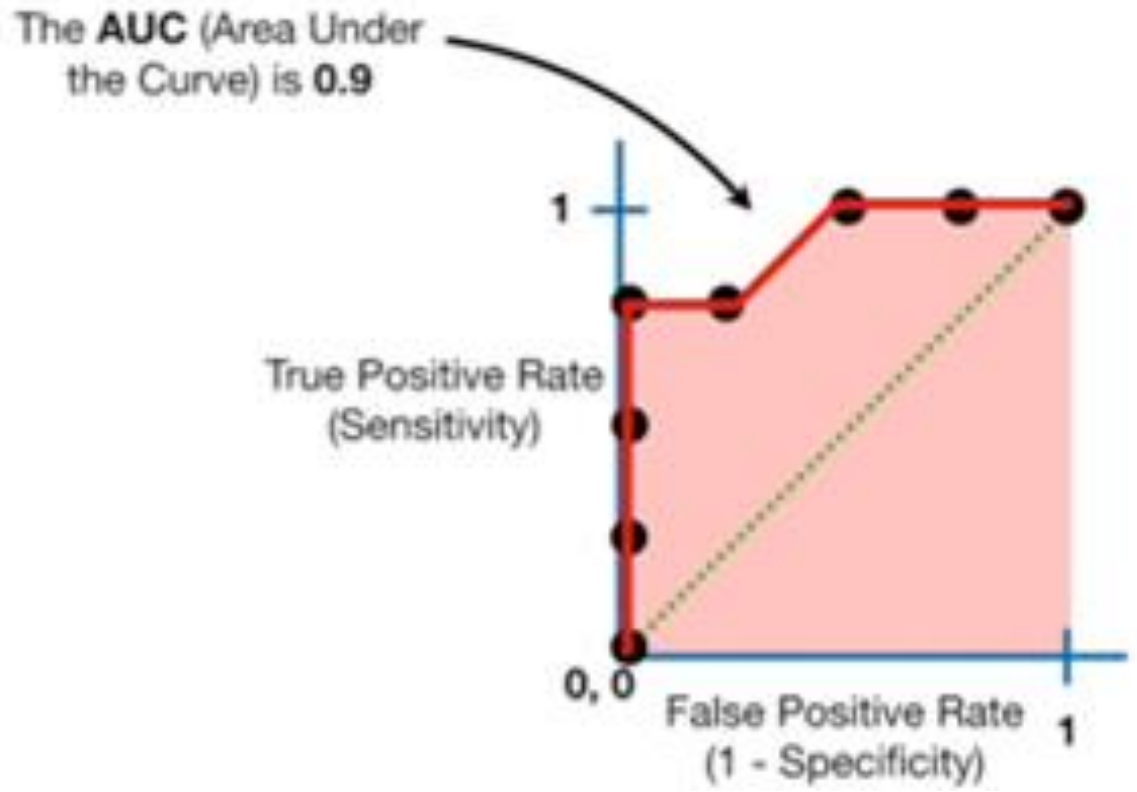
False Positive Rate shows the proportion of cases with negative income (value 0), which were incorrectly classified.

$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$



Area Under the Curve (AUC)

Area under the curve helps to compare the quality of classifications and decide which categorization method is better.



Statistical significance of the model

Log-Likelihood:	-316.46
LL-Null:	-482.26
LLR p-value:	5.069e-66

The model is significantly better (Prob <0.05) than the baseline model with no predictors. It means that by adding the predictors we managed to significantly decrease the log-likelihood (amount of unexplained information).

Pseudo R-squared

$$R^2_{CS} = 1 - e^{\left[-\frac{2}{n}(\mathrm{LL(New)} - \mathrm{LL(Baseline)})\right]}$$

$$R^2_{N} = \frac{R^2_{CS}}{1 - e^{\left[\frac{2(\mathrm{LL(Baseline)})}{n}\right]}}$$

Pseudo R-squared represents the percent of the variation of the dependent variable explained by the model.

Performance Measurement from scikit-learn

Precision measures the accuracy of positive predictions. Also called the precision of the classifier.

$$\text{precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sencitivity) measures the ratio of positive instances that are correctly detected by the classifier. Increasing precision reduced recall and vice versa

$$\text{recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1 score is the harmonic mean of precision and recall. Regular mean gives equal weight to all values. Harmonic mean gives more weight to low values.

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

Useful links

- http://www.dataschool.io/roc-curves-and-auc-explained/
- https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/
- https://realpython.com/logistic-regression-python/
- https://mlu-explain.github.io/logistic-regression/

Thank you for your attention!