Quantitative Methods

Intro to Machine Learning

Dr. Yves-Alexandre de Montjoye

Associate Professor in Dept. of Computing joint with the
Data Science Institute

- Postdoc at Harvard
- PhD from Massachusetts Institute of Technology
- MSc in Applied Mathematics from Université catholique de Louvain, Ecole Centrale Paris, and Katholieke Universiteit Leuven
- BSc in Engineering from Louvain

Head of the Computational Privacy Group
Director of the Algorithmic Society Lab

Teach CO408 (Privacy Engineering) at Imperial College London

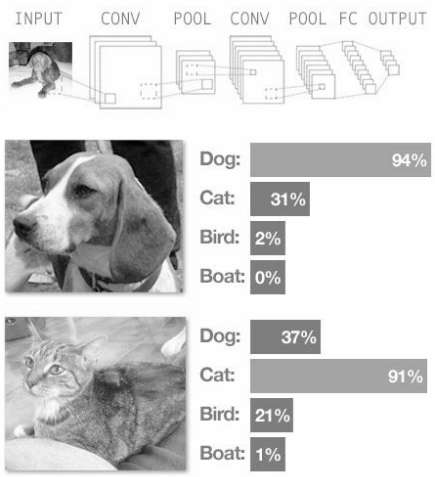
Intro to Machine Learning: Classifiers

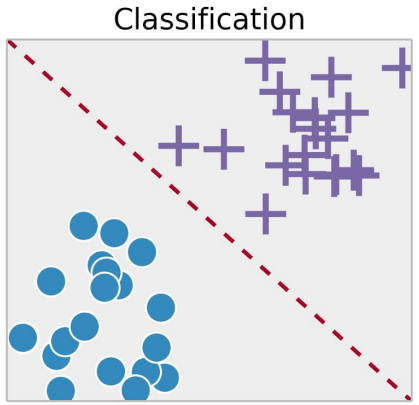
We will here focus on predictive analytics: training a model on labeled data (“where we know the right answer, e.g. dog or cat”) to then guess the answer in another similar* dataset

Examples:

- Predicting whether a picture is a picture of a cat or a dog to prevent spam on social network for cat owners
- Predicting if a mushroom is toxic or not based on a picture
- Predicting if a person is a republican or a democrat based on demographics
- or...

* Similar is a big necessary assumption here, the model learns from the data so if the data is not representative the model will not work well or even completely wrong






NATIONAL
GEOGRAPHIC
CHANNEL

Titanic dataset

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew.

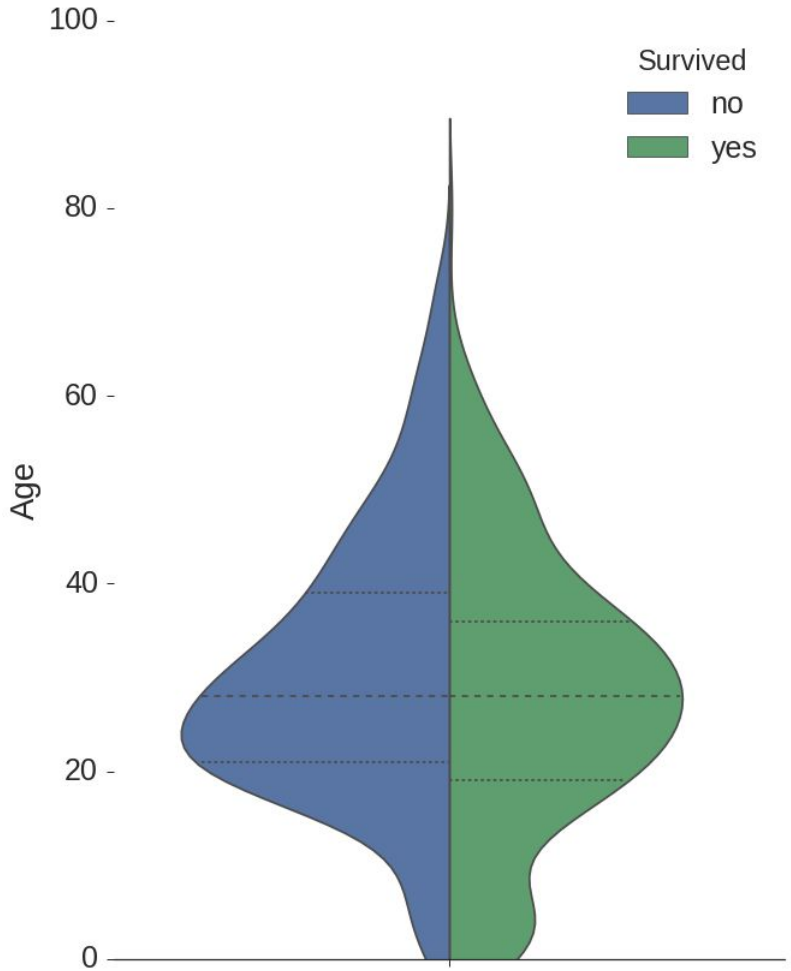
Although there is (as always) some element of luck involved in surviving the sinking of a ship, were some people more likely to survive than others?

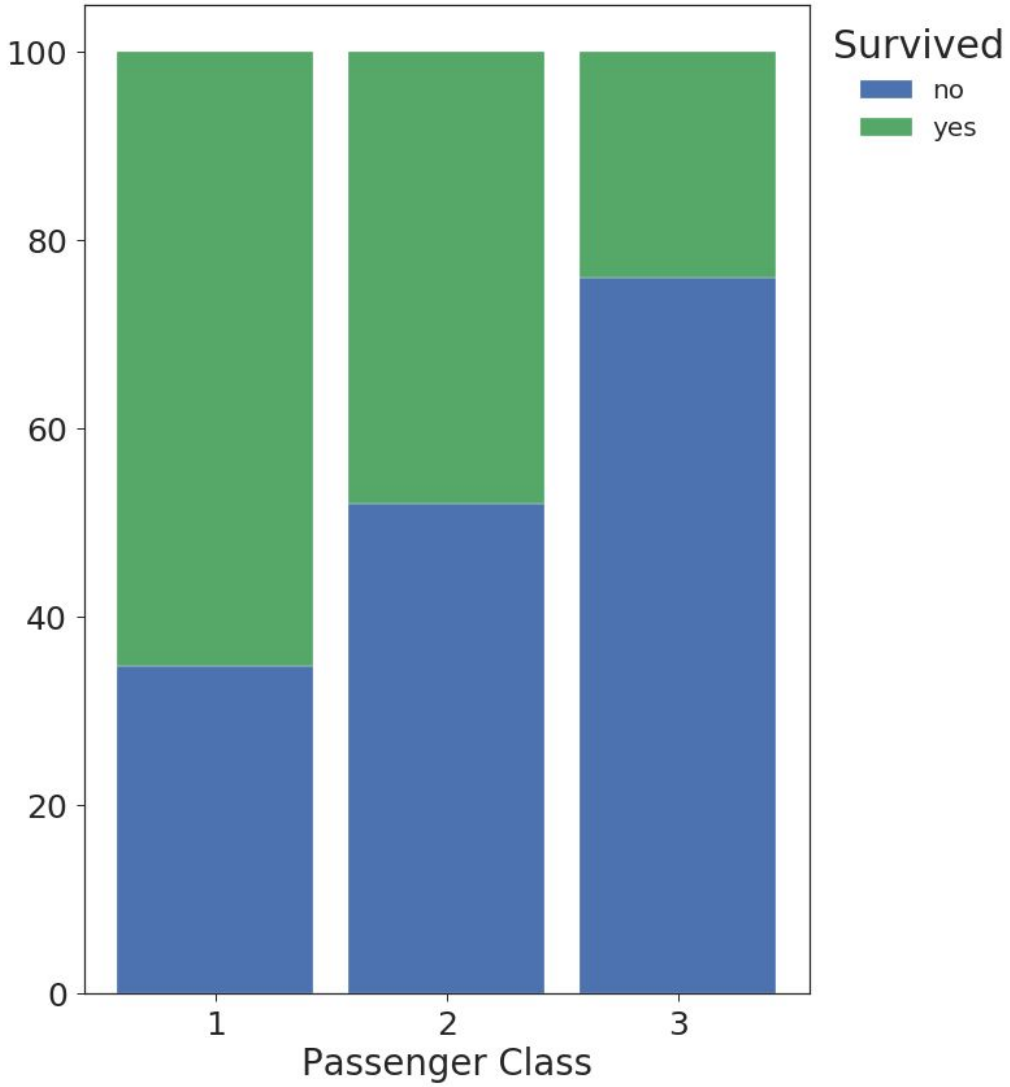
And, if yes, I could have used this to make a prediction and used this prediction to price travel insurance in 1913?

Variables:

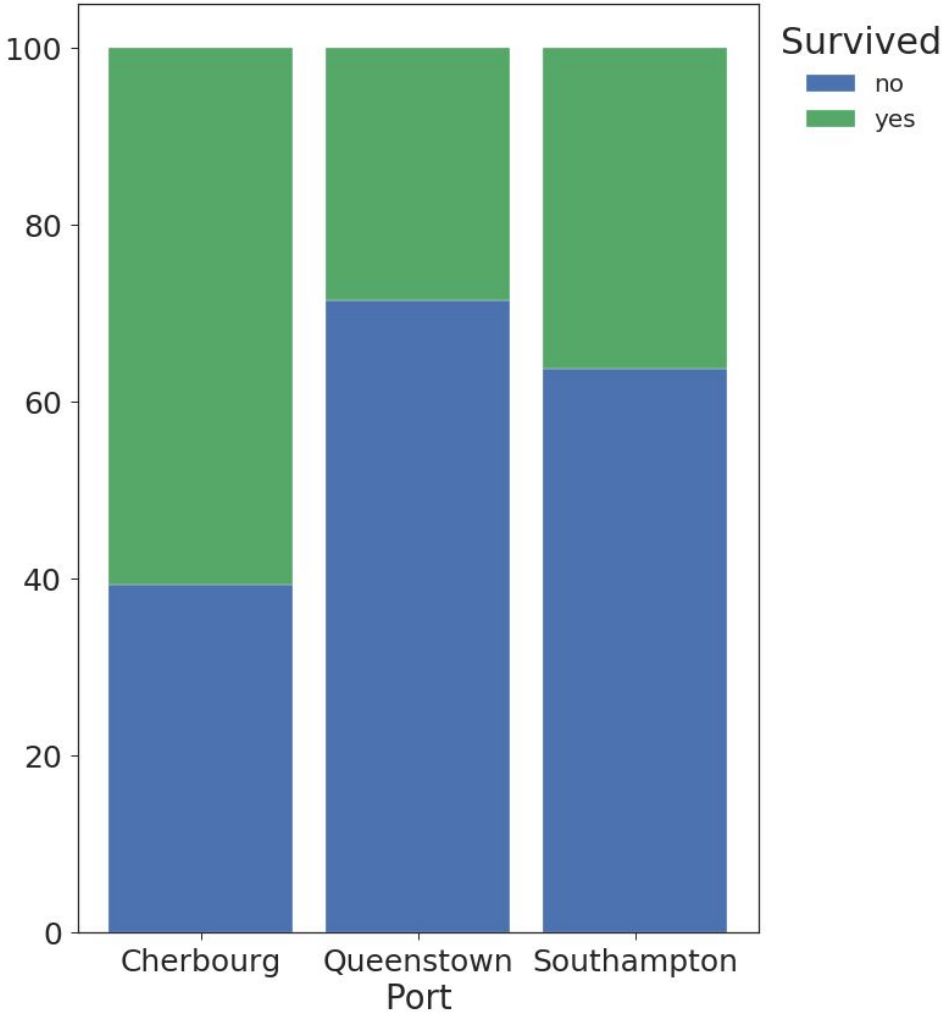
- *Survival*: (0 = No, 1 = Yes)
- *Pclass:* Ticket class (1st, 2nd, 3rd)
- *Sex*: Sex (male/female)
- *age:* Age [years]
- *fare:* Passenger fare in Pre-1970 British Pounds
- *embarked:* Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
- Etc

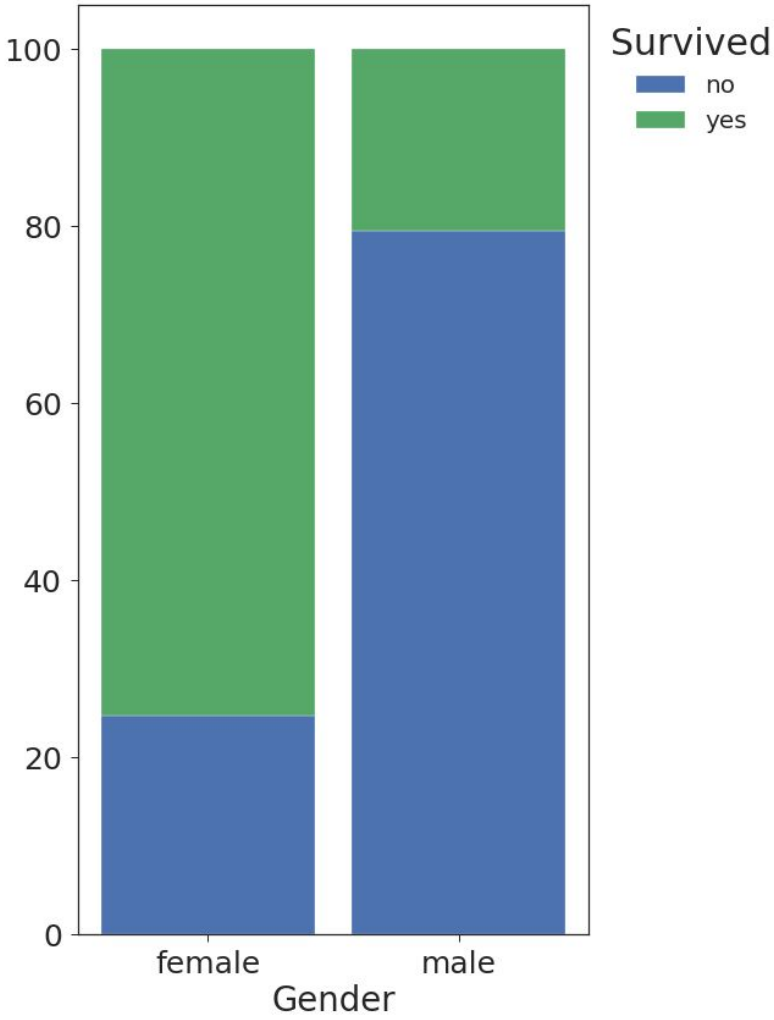
Which feature might help?





Which feature might help?





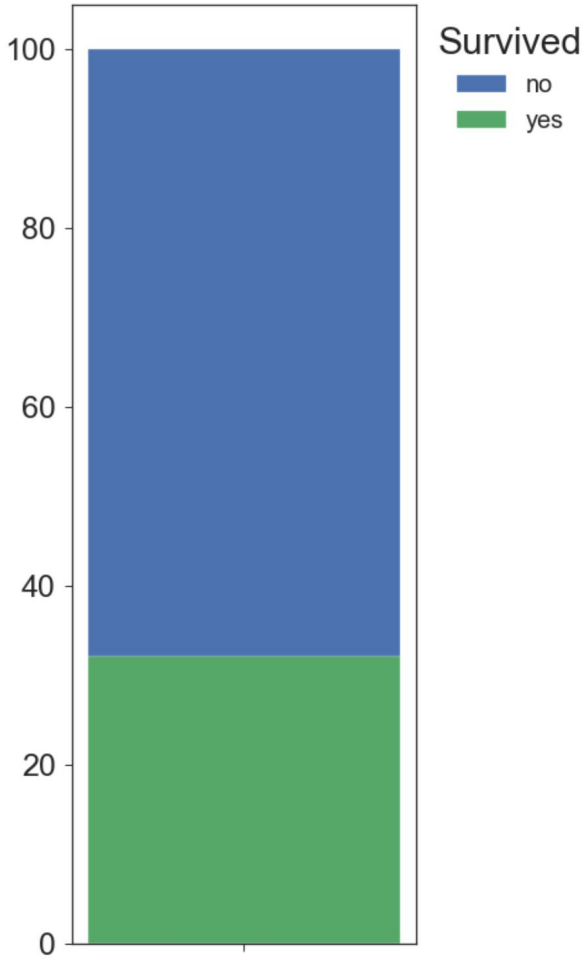
Building model to explain a **categorical** variable

The variable of interest is categorical (in fact even binary, the person *survived* or *died*) so we will use the following formula to model the **probability of someone surviving the sinking of the titanic** using e.g. *fare* as the explanatory variable:

$$P_{surviving} = \frac{1}{(1 + \exp(-(a + b * fare)))}$$

We then again try to find the *a* and *b* that would give us the best model

To transform a categorical variable into a continuous one, a logistic regression uses a **logit function.**



Your first classifier: Logistic regression

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           Survived_bin   No. Observations:                  712
Model:                            GLM   Df Residuals:                      710
Model Family:                Binomial   Df Model:                            1
Link Function:                  logit   Scale:                           1.0
Method:                          IRLS   Log-Likelihood:                -449.58
Date:                Sat, 09 Jun 2018   Deviance:                       899.16
Time:                        15:49:38   Pearson chi2:                     756.
No. Iterations:                     5
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.8945      0.107     -8.330      0.000      -1.105      -0.684
Fare           0.0157      0.002      6.323      0.000       0.011       0.021
==============================================================================
```

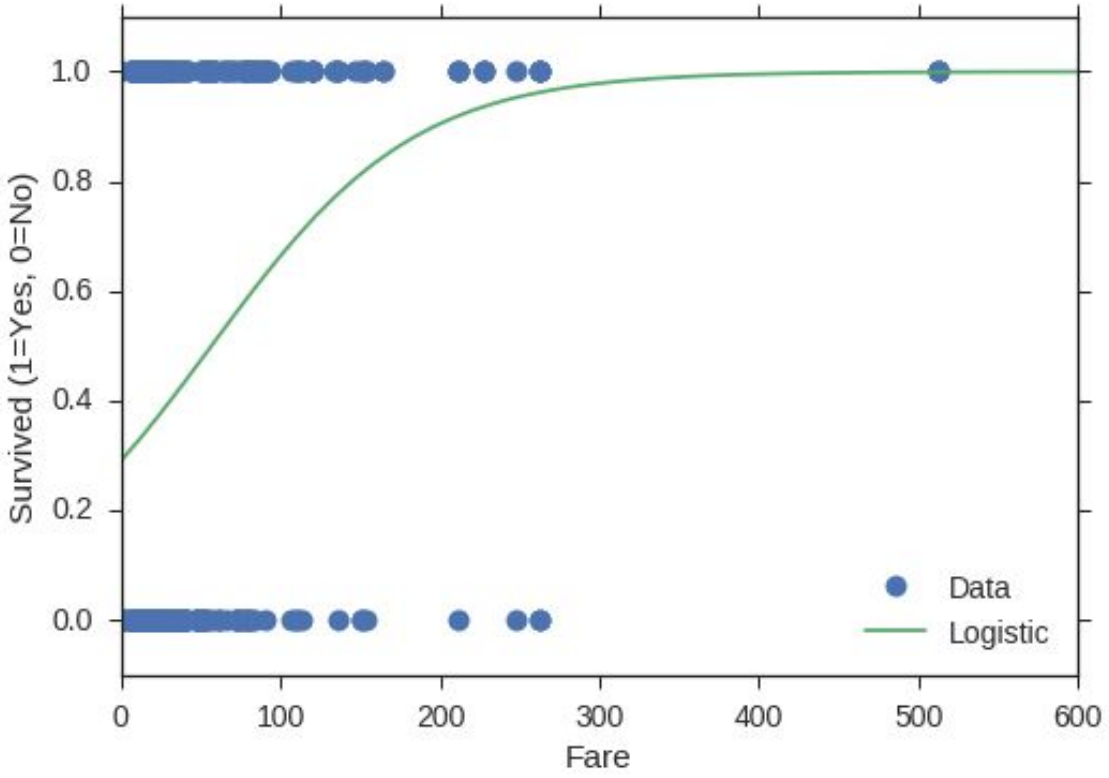
Caution: interpreting the coefficient

The model we fit is now:

$$P_{surviving} = \frac{1}{(1 + \exp(-(-0.8945 + 0.0157 * fare)))}$$

So the coefficient of the variable *fare* is **inside the exponent** instead of directly related to the variable we try to explain.

We can therefore not interpret it anymore as “the increase in Y resulting from an increase in X”

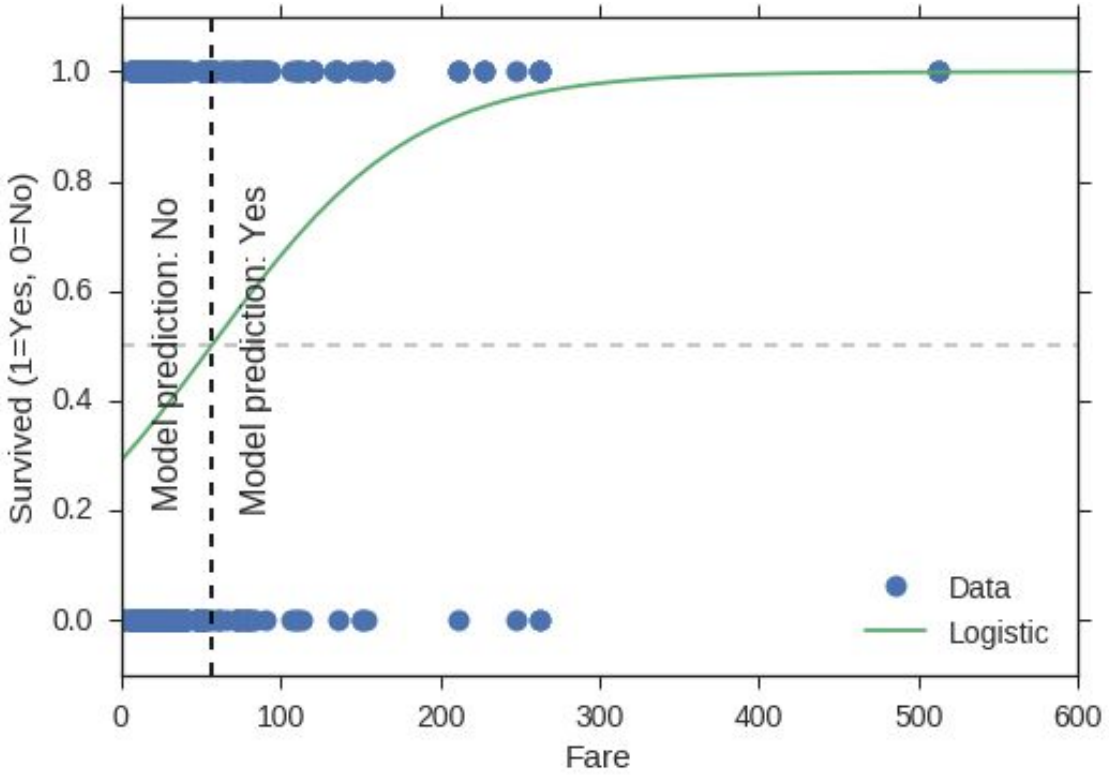


But how good it our model?

We can’t compute an R^2 anymore as the results are binary: survived or died

We instead use our model to make a prediction e.g. assuming that if the probability is >0.5, the model predicts that the person survived.

We can then look at **how often our model make the correct prediction**




```
Accuracy:          0.666
Precision:         0.723
Recall:            0.281
```

Survived ~ Fare + Sex

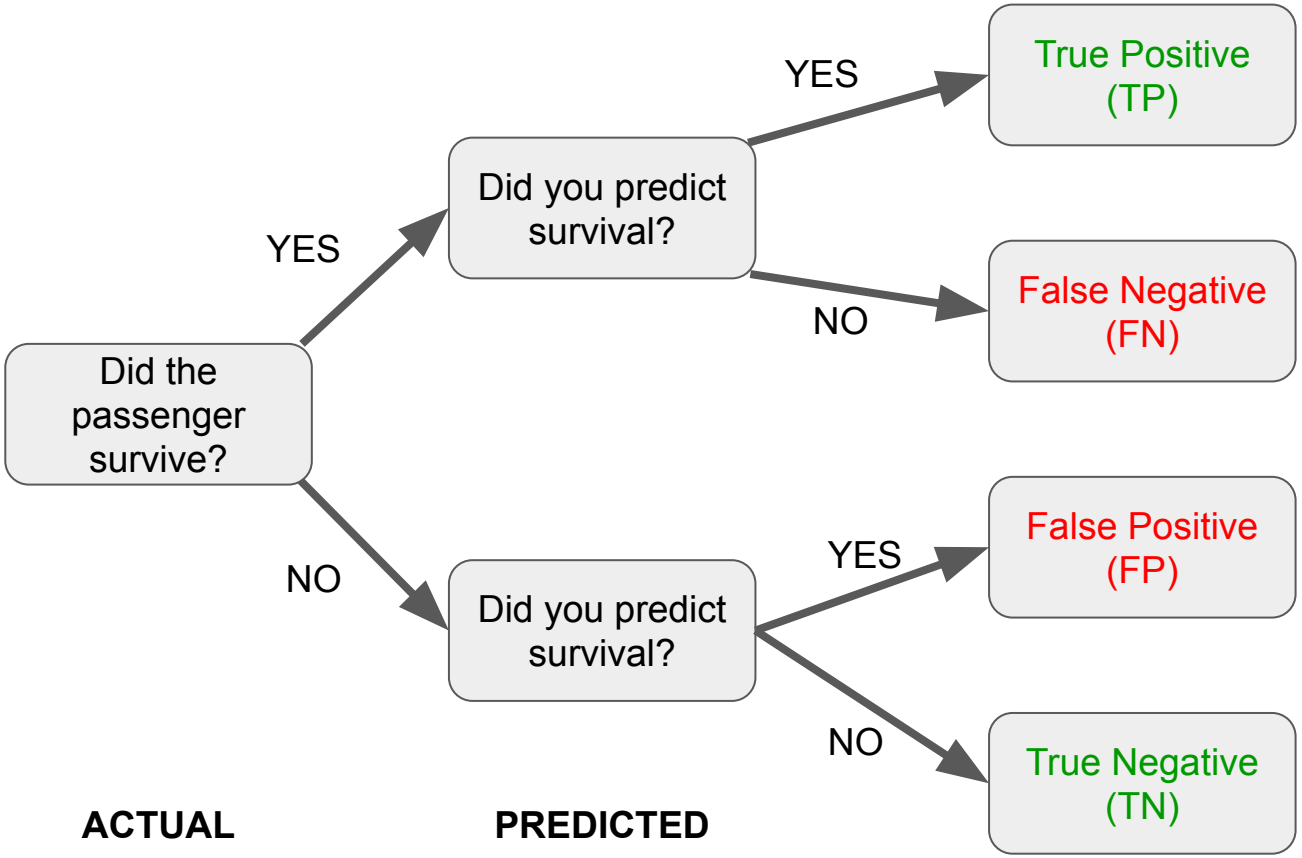
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           Survived_bin   No. Observations:                  712
Model:                            GLM   Df Residuals:                      709
Model Family:                Binomial   Df Model:                            2
Link Function:                  logit   Scale:                           1.0
Method:                          IRLS   Log-Likelihood:                -359.02
Date:                Sat, 09 Jun 2018   Deviance:                       718.03
Time:                        15:45:37   Pearson chi2:                     696.
No. Iterations:                     5
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.6590      0.167      3.935      0.000       0.331       0.987
Sex[T.male]   -2.3711      0.189    -12.524      0.000      -2.742      -2.000
Fare           0.0121      0.003      4.595      0.000       0.007       0.017
==============================================================================
Confusion Matrix (total:712)        Accuracy:                0.777
  TP: 198 | FN: 90                  Precision:               0.742
  FP: 69 | TN: 355                  Recall:                  0.688
```

Note that all p-value are significant (<0.05 with CI not containing 0)

Finally we can see that, compared to our previous model, accuracy increased

Beyond accuracy: Evaluating how often and how is my model wrong

There are four possible outcomes of a prediction I (well, my model) made:



My model predicted Positive and was right

My model predicted Negative and was wrong

My model predicted Positive and was wrong

My model predicted Negative and was right

This is called a confusion matrix

This is usually summarised in a matrix:

	Prediction: Yes - Survived	Prediction: No - Died
Actual: Yes - Survived	True Positives (TP)	False Negatives (FN)
Actual: No - Died	False Positives (FP)	True Negatives (TN)

Clearly, we would like to get as many True Negatives and True Positives as possible (my model was **right**) and as few False Positives and False Negatives as possible (my model was **wrong**).

From this I can compute accuracy which is $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

Survived ~ Fare + Age + Pclass + Sex

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           Survived_bin   No. Observations:                  712
Model:                            GLM   Df Residuals:                      706
Model Family:                Binomial   Df Model:                            5
Link Function:                  logit   Scale:                           1.0
Method:                          IRLS   Log-Likelihood:                -323.32
Date:                Sat, 09 Jun 2018   Deviance:                       646.64
Time:                        15:51:48   Pearson chi2:                    766.
No. Iterations:                     5
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.7150      0.464      7.998      0.000       2.805       4.625
Pclass[T.2]   -1.2682      0.313     -4.055      0.000      -1.881      -0.655
Pclass[T.3]   -2.5336      0.328     -7.728      0.000      -3.176      -1.891
Sex[T.male]   -2.5096      0.208    -12.041      0.000      -2.918      -2.101
Fare           0.0005      0.002      0.230      0.818      -0.004       0.005
Age           -0.0369      0.008     -4.769      0.000      -0.052      -0.022
==============================================================================
Confusion Matrix (total:712)        Accuracy:                0.792
  TP: 207 | FN: 81                  Precision:               0.755
  FP: 67 | TN: 357                  Recall:                  0.719
```

Pclass can take 3 values, T.2 and T.3 are the effect of being in 2nd and 3rd class as opposed to being in 1st which is the default value

Note that while fare was significant in the previous model, it is not significant anymore. Why?

Finally, note that we have again improved accuracy

Survived ~ Embarked + Age + Pclass + Sex

Generalized Linear Model Regression Results

Dep. Variable:	Survived_bin	No. Observations:	712
Model:	GLM	Df Residuals:	705
Model Family:	Binomial	Df Model:	6
Link Function:	logit	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-321.34
Date:	Sat, 09 Jun 2018	Deviance:	642.68
Time:	15:54:02	Pearson chi2:	754.
No. Iterations:	5		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	4.0368	0.431	9.371	0.000	3.193	4.881
Pclass[T.2]	-1.1446	0.291	-3.938	0.000	-1.714	-0.575
Pclass[T.3]	-2.4096	0.291	-8.275	0.000	-2.980	-1.839
Embarked[T.Q]	-0.8142	0.568	-1.434	0.152	-1.927	0.299
Embarked[T.S]	-0.4937	0.267	-1.850	0.064	-1.017	0.029
Sex[T.male]	-2.5158	0.209	-12.020	0.000	-2.926	-2.106
Age	-0.0361	0.008	-4.677	0.000	-0.051	-0.021

Confusion Matrix (total:712)

TP: 209 | FN: 79	
FP: 65 | TN: 359	

Accuracy:	0.798
Precision:	0.763
Recall:	0.726

Removing Fare and adding the port people embarked at

Not embarking at Cherbourg decreases the odds of surviving

Overall this looks like a good model
1) Accuracy is good with low FP and FN
2) It confirms what we saw visually: 2nd and 3rd class, male, older people and people who did not embark in Cherbourg are less likely to have survived the sinking
3) all variables are significant

Let’s compete:
https://ml101-cpg.doc.ic.ac.uk/

Competition: best accuracy

For this exercise you will need to estimate a logistic regression on a crime dataset.

The variable of interest is whether larcenies are high or low in a region

The variables you can use are listed on the app:
-- **agePct12t29**: percentage of population that is 12-29 in age (numeric - decimal)
-- **agePct16t24**: percentage of population that is 16-24 in age (numeric - decimal)
-- **agePct65up**: percentage of population that is 65 and over in age (numeric - decimal)
-- **numbUrban**: number of people living in areas classified as urban (numeric - expected to be integer)
-- **pctUrban**: percentage of people living in areas classified as urban (numeric - decimal)
-- **medIncome**: median household income (numeric - may be integer)