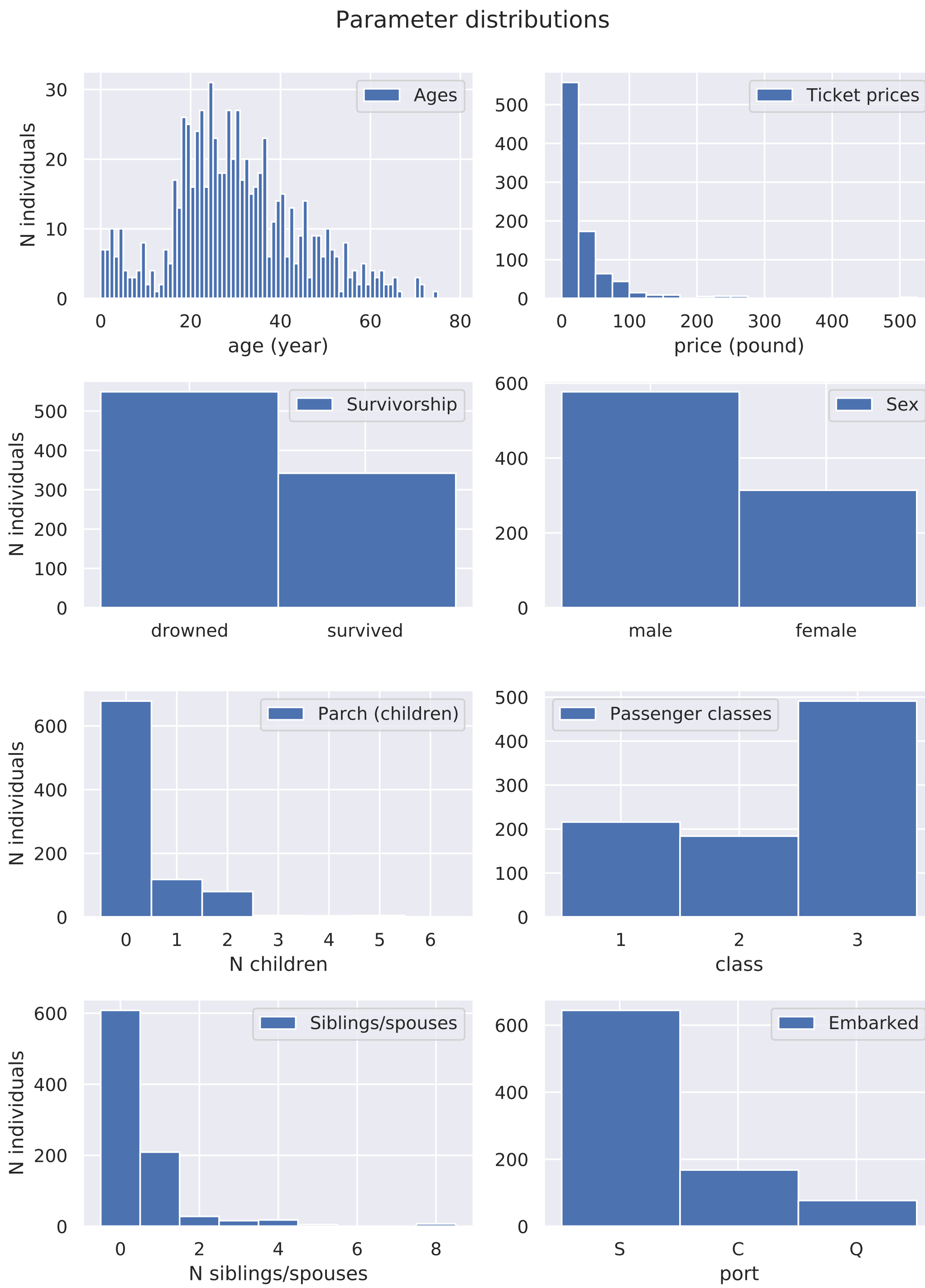
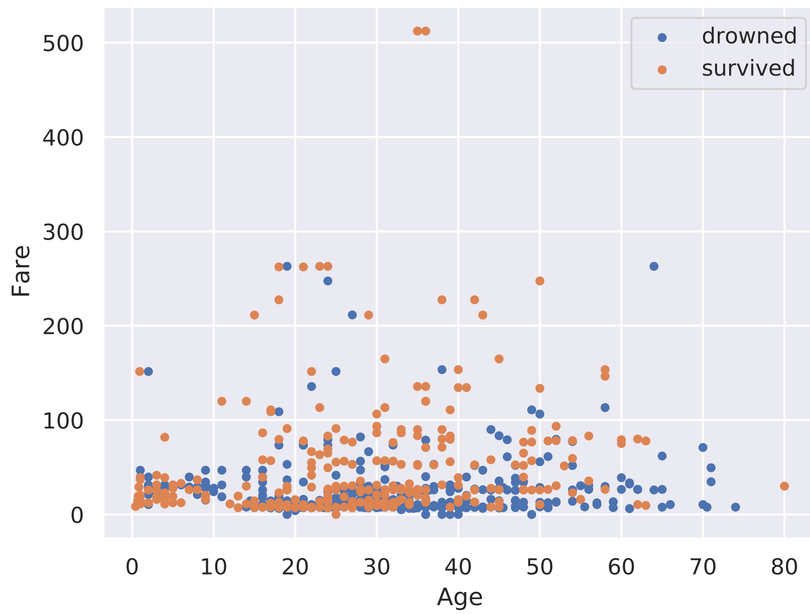
Analysis of Titanic data

Samuel Knudsen

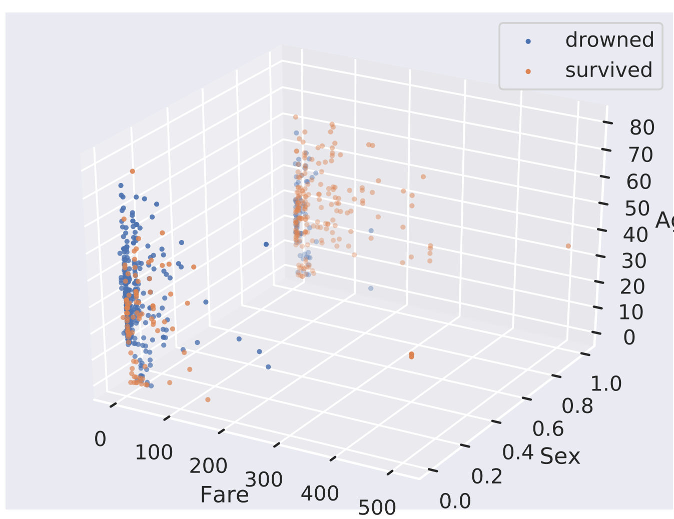
Last update: September 11, 2021



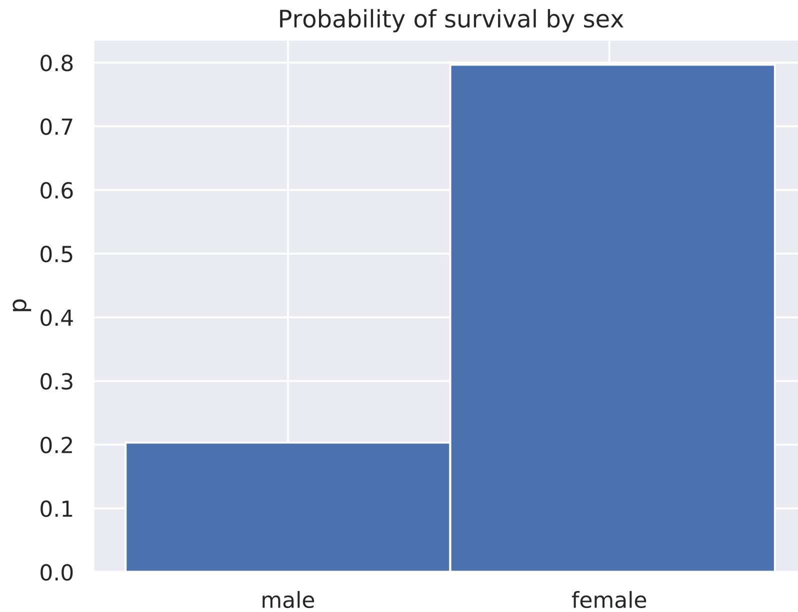
Parameter distributions
Ages
N individuals
age (year)
Ticket prices
price (pound)
Survivorship
drowned
survived
Sex
male
female
Parch (children)
N children
Passenger classes
class
Siblings/spouses
N siblings/spouses
Embarked
S
C
Q
port

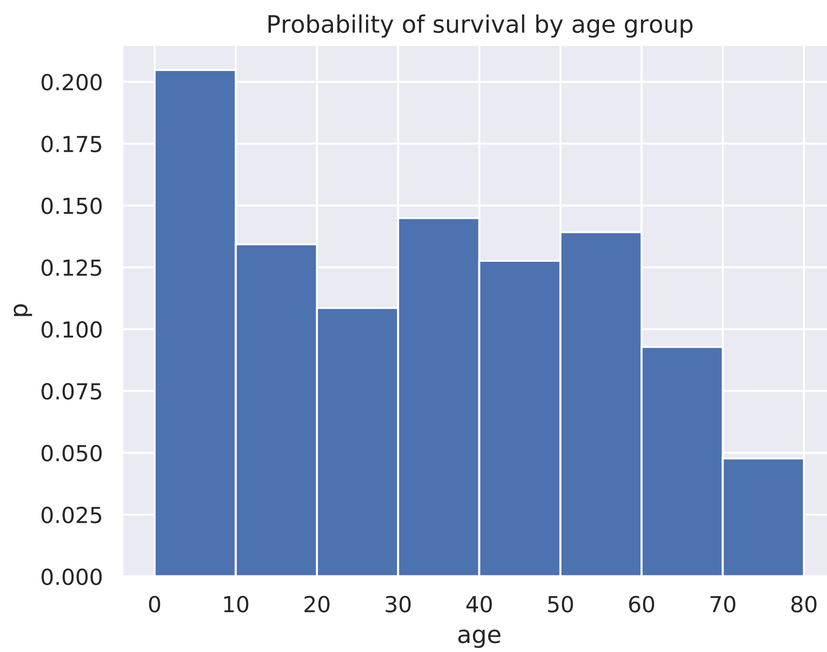


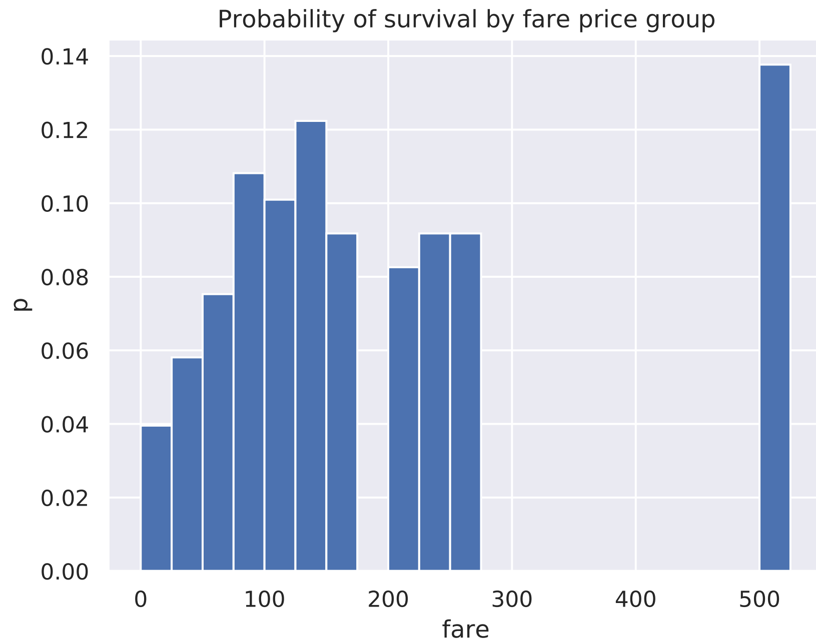
drowned
survived
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80

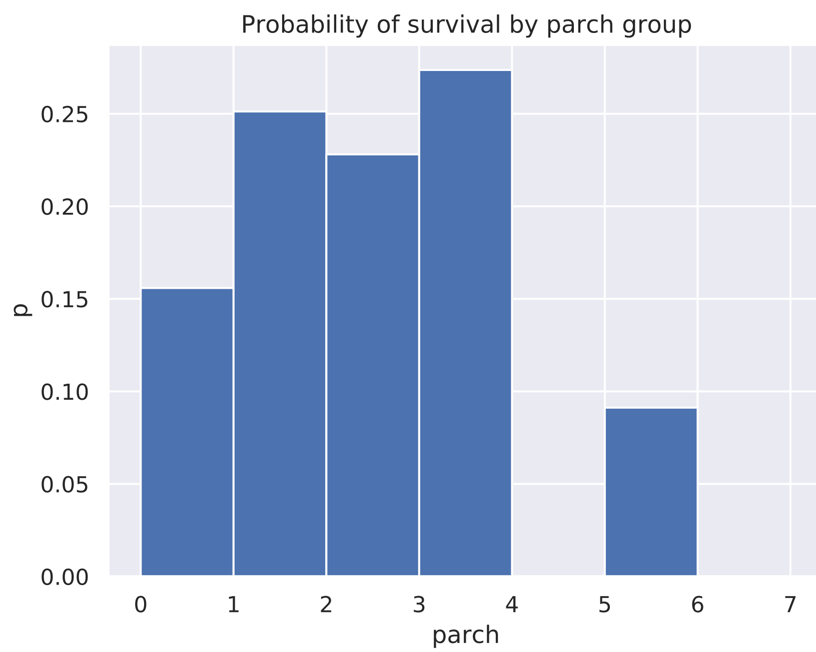


drowned
survived
Age
Sex
Fare
0
10
20
30
40
50
60
70
80
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500

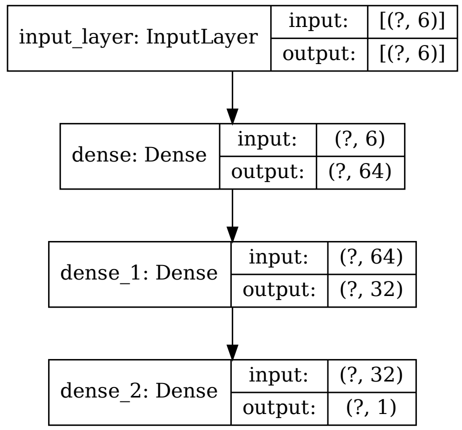

Probability of survival by sex
p
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
male
female


Probability of survival by age group
p
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
0
10
20
30
40
50
60
70
80
age


Probability of survival by fare price group
p
fare
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
100
200
300
400
500


Probability of survival by parch group
p
parch
0.00
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
6
7

Training a neural network to predict survival



Bayesian probability analysis

Let us look at which piece of data about an individual gives us most information about the likelihood of survival, given our dataset D.

For some choice of $x = 0, 1$, $y = 0, 1$, we have by Bayes rule:

$$\text{Prob}(\text{survived} = x|\text{sex} = y, D) \propto \text{Prob}(\text{sex} = y|\text{survived} = x, D) \cdot \text{Prob}(\text{survived} = x|D) \tag{1}$$

We can calculate the prior (the probability estimate of survival prior to knowing the sex of the given person)

$$\text{Prob}(\text{survived} = x|D) = \frac{N_{\text{survived}=x}}{N_{\text{total passengers}}}, \tag{2}$$

and similarly the likelihood

$$\text{Prob}(\text{sex} = y|\text{survived} = x, D) = \frac{N_{\text{survived}=x,\text{sex}=y}}{N_{\text{total survivors}}} \tag{3}$$

with the normalization condition

$$\text{Prob}(\text{survived} = 0|\text{sex} = y, D) + \text{Prob}(\text{survived} = 1|\text{sex} = y, D) = 1. \tag{4}$$

$$\begin{aligned} p(A, B|D) &= p(A|B, D)p(B|D) \\ p(B, A|D) &= p(B|A, D)p(A|D) \end{aligned} \tag{5}$$

and

$$p(A, B|D) = p(B, A|D) \tag{6}$$

meaning

$$p(A|B, D) = \frac{p(B|A, D)p(A|D)}{p(B|D)} \tag{7}$$