# PREDICTING PASSENGER SURVIVAL

# -

# TITANIC DATASET

Bilal Ullah

# USE CASE

- I have been approached by the National Transportation Safety Board (NTSB) to investigate data of crashes of large vehicles
- The aim for the NTSB is to investigate whether or not the survival of a passenger can be predicted and if so what does this depend on
- In particular I have been tasked with the titanic dataset and have to predict the survival of passengers

# PREVIEW OF DATA

Size of data:

- 891 rows
- 69 features

All Numeric data

Passenger profile

Ticket location

Cabin type

| | Sex | Age | SibSp | Parch | Fare | Title_Master | Title_Miss | Title_Mr | Title_Mrs |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 22.0 | 1 | 0 | 7.2500 | 0 | 0 | 1 | 0 |
| 1 | 0 | 38.0 | 1 | 0 | 71.2833 | 0 | 0 | 0 | 1 |
| 2 | 0 | 26.0 | 0 | 0 | 7.9250 | 0 | 1 | 0 | 0 |
| 3 | 0 | 35.0 | 1 | 0 | 53.1000 | 0 | 0 | 0 | 1 |

# FEATURE SELECTION

- 69 features to select from
- Features needed to be reduced significantly to optimize time and model results
- As there are so many features it is not efficient to manually pick features

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.000000 | 0.100952 | -0.114631 | -0.245489 | -0.182333 | 0.159934 | -0.691548 | 0.867334 | -0.552686 | 0.089228 | -0.007483 | -0.082853 | -0.074115 | [illegible] | [illegible] | [illegible] |
| 0.100952 | 1.000000 | -0.267801 | -0.184627 | 0.125602 | -0.391855 | -0.303490 | 0.205098 | 0.203422 | 0.188543 | 0.068702 | 0.052410 | -0.107742 | 0.021889 | 0.144459 | 0.106245 |
| -0.114631 | -0.267801 | 1.000000 | 0.414838 | 0.159651 | 0.349559 | 0.084945 | -0.250489 | 0.059941 | -0.024712 | -0.008384 | -0.059528 | -0.026354 | 0.068734 | -0.046266 | -0.034538 |
| -0.245489 | -0.184627 | 0.414838 | 1.000000 | 0.216225 | 0.267344 | 0.102514 | -0.333905 | 0.221318 | -0.048211 | -0.035583 | -0.011069 | -0.081228 | 0.060814 | -0.040325 | 0.056498 |
| -0.182333 | 0.125602 | 0.159651 | 0.216225 | 1.000000 | 0.010908 | 0.120829 | -0.183766 | 0.105665 | 0.010357 | 0.015044 | 0.269335 | -0.117216 | -0.162184 | 0.019549 | 0.386297 |
| 0.159934 | -0.391855 | 0.349559 | 0.267344 | 0.010908 | 1.000000 | -0.110602 | -0.254903 | -0.088394 | -0.031131 | -0.016287 | -0.035225 | 0.010478 | 0.024264 | 0.013759 | -0.026914 |
| -0.691548 | -0.303490 | 0.084945 | 0.102514 | 0.120829 | -0.110602 | 1.000000 | -0.599803 | -0.207996 | -0.073253 | -0.038324 | 0.037613 | 0.168720 | -0.139126 | -0.066756 | 0.065664 |
| 0.867334 | 0.205098 | -0.250489 | -0.333905 | -0.183766 | -0.254903 | -0.599803 | 1.000000 | -0.479363 | -0.168826 | -0.088324 | -0.072567 | -0.078338 | 0.112870 | 0.040591 | -0.114673 |
| -0.552686 | 0.203422 | 0.059941 | 0.221318 | 0.105665 | -0.088394 | -0.207996 | -0.479363 | 1.000000 | -0.058544 | -0.030628 | 0.066101 | -0.091121 | -0.000565 | -0.053352 | 0.061767 |
| 0.089228 | 0.188543 | -0.024712 | -0.048211 | 0.010357 | -0.031131 | -0.073253 | -0.168826 | -0.058544 | 1.000000 | -0.010787 | -0.008034 | 0.012618 | -0.000902 | 0.043217 | 0.073177 |
| -0.007483 | 0.068702 | -0.008384 | -0.035583 | 0.015044 | -0.016287 | -0.038324 | -0.088324 | -0.030628 | -0.010787 | 1.000000 | 0.079020 | -0.023105 | -0.054685 | 0.223735 | 0.049486 |
| -0.082853 | 0.052410 | -0.059528 | -0.011069 | 0.269335 | -0.035225 | 0.037613 | -0.072567 | 0.066101 | -0.008034 | 0.079020 | 1.000000 | -0.148258 | -0.782742 | 0.093040 | 0.168642 |
| -0.074115 | -0.107742 | -0.026354 | -0.081228 | -0.117216 | 0.010478 | 0.168720 | -0.078338 | -0.091121 | 0.012618 | -0.023105 | -0.148258 | 1.000000 | -0.499421 | -0.040246 | -0.072579 |
| 0.119224 | 0.021889 | 0.068734 | 0.060814 | -0.162184 | 0.024264 | -0.139126 | 0.112870 | -0.000565 | -0.000902 | -0.054685 | -0.782742 | -0.499421 | 1.000000 | -0.056180 | -0.102063 |
| 0.078271 | 0.144459 | -0.046266 | -0.040325 | 0.019549 | 0.013759 | -0.066756 | 0.040591 | -0.053352 | 0.043217 | 0.223735 | 0.093040 | -0.040246 | -0.056180 | 1.000000 | -0.030880 |
| -0.109689 | 0.106245 | -0.034538 | 0.056498 | 0.386297 | -0.026914 | 0.065664 | -0.114673 | 0.061767 | 0.073177 | 0.049486 | 0.168642 | -0.072579 | -0.102063 | -0.030880 | 1.000000 |
| -0.058649 | 0.151943 | 0.029251 | 0.030736 | 0.364318 | -0.035937 | 0.009098 | -0.047873 | 0.072174 | 0.025924 | -0.020005 | 0.113952 | -0.049776 | -0.068502 | -0.034846 | -0.062841 |
| -0.079248 | 0.145891 | -0.017575 | -0.019125 | 0.098878 | -0.042519 | 0.017400 | -0.049952 | 0.073034 | 0.014080 | -0.014733 | 0.102977 | -0.060318 | -0.052254 | -0.025663 | -0.046280 |

# FEATURE SELECTION

## ANOVA

- We want to determine whether a set of means are all equal.
- To evaluate this with an F-test, we need to use the proper variances in the ratio.

Most Significant:

- Ticket_XXX
- Cabin_T

$$F = \frac{\text{between-groups variance}}{\text{within-group variance}}$$

## Corr plot

- Features selected from the correlation included any correlation that was outside the ranges of -0.02 - 0.02

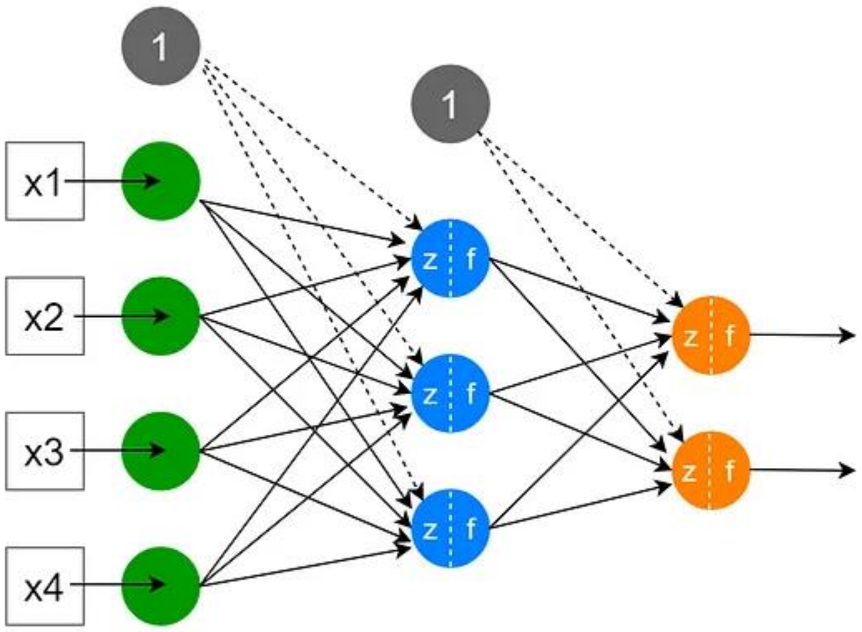
# STANDARDIZATION

- This was carried out on the continuous variables remaining
- Scaling data to fit a standard normal distribution. A standard normal distribution is defined as a distribution with a mean of 0 and a standard deviation of 1.
- This was carried out on these features:
- 'Age', 'Fare', 'Parch', 'SibSp', 'FamilySize'

# SHALLOW ANN

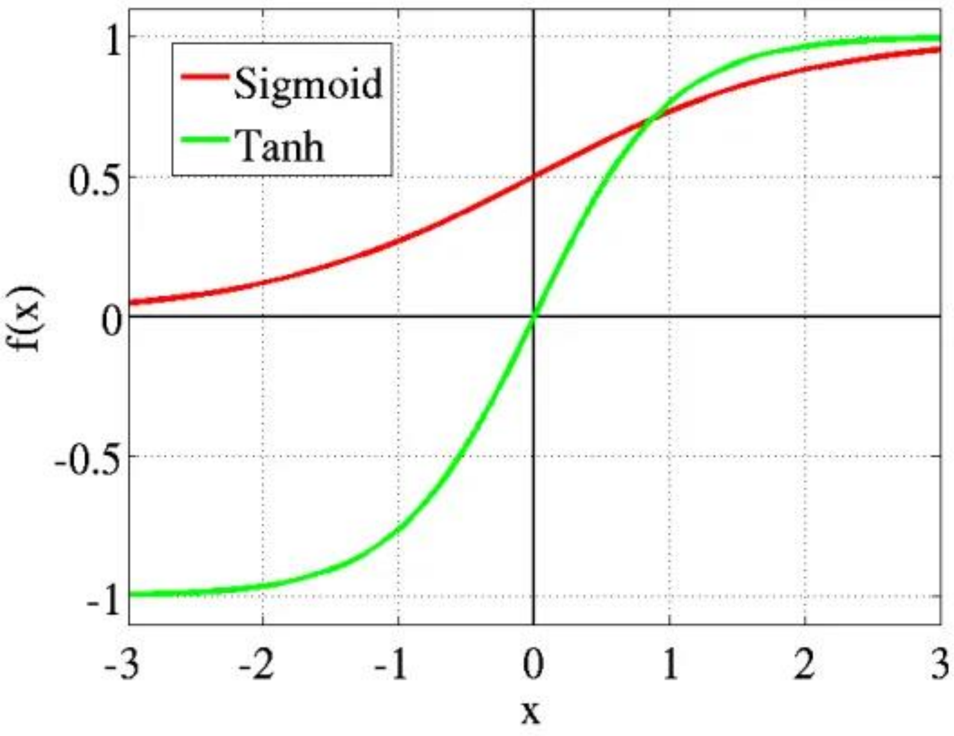
- A shallow is a type of Neural Network with only one hidden layer
  - Input Node
  - Hidden layer
  - Output

- Each node within the hidden layer consists of weights and activations
- Activations play a crucial role in the Neural Networks



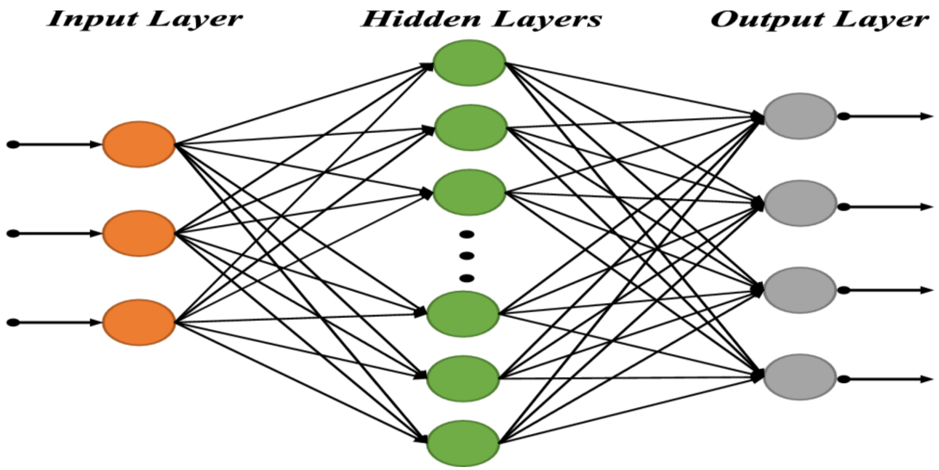
# ACTIVATION METHOD

- To introduce non-linearity in the network, activation functions are necessary
- Sigmoid and Tanh are effective as it results in higher values of gradient during training and higher updates in the weights of the network.
- As the output in this case is either survived or died, we want strong gradients and big learning steps.



# OPTIMIZATION

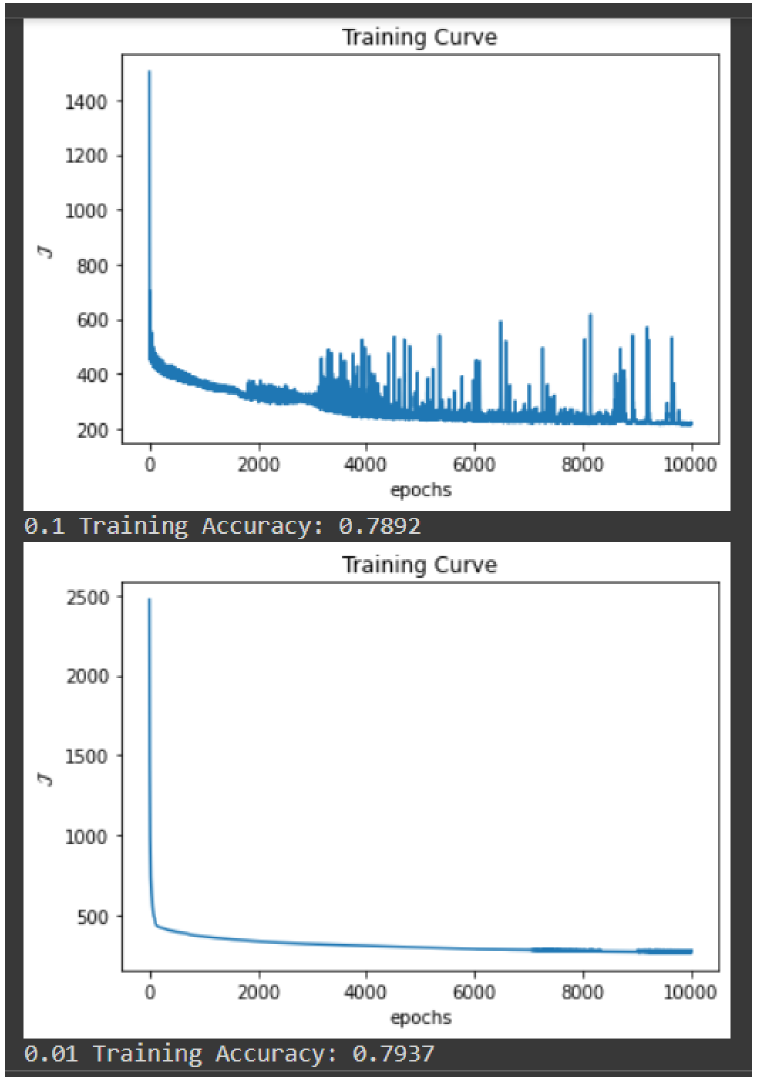
- No Packages!
- Manual optimization
- Quick test by increasing nodes in the neural network to see effect on results




```
10 Training Accuracy: 0.7040
20 Training Accuracy: 0.8117
30 Training Accuracy: 0.7668
40 Training Accuracy: 0.7534
50 Training Accuracy: 0.7758
60 Training Accuracy: 0.7713
70 Training Accuracy: 0.8117
80 Training Accuracy: 0.8072
90 Training Accuracy: 0.7982
100 Training Accuracy: 0.8072
120 Training Accuracy: 0.8072
130 Training Accuracy: 0.8161
140 Training Accuracy: 0.7982
150 Training Accuracy: 0.8117
300 Training Accuracy: 0.7982
```

# OPTIMIZATION

- Same method was carried out for learning rate
- Oscillation occurred on many training curves which indicates dataset may be too small
- Less than 900 rows



# RESULTS

| Nodes | eta | epochs | Accuracy |
|---|---|---|---|
| 30 | 1e-2 | 1e4 | 79.51% |
| 50 | 1e-3 | 1e5 | 80.74% |
| 70 | 1e-2 | 1e4 | 83.41% |
| 150 | 1e-2 | 1e5 | 81.17% |

# EVALUATION FOR BEST PARAMETERS

| Precision | Recall | F1 | Accuracy |
|---|---|---|---|
| 76.47 | 79.27 | 77.84 | 83.41 |

# SUMMARY

- Key Points
  - Feature Selection
  - Standardization
  - Activation method
  - Optimization
- Results
  - Positive results
- Use case effectiveness:
  - The NTSB can take confidence in these results and build on this model to further aid their analysis in future investigations

# APPENDIX

| | Sex | Age | SibSp | Parch | Fare | Title_Miss | Title_Mr | Title_Mrs | Cabin_B | Cabin_U |
|---|---|---|---|---|---|---|---|---|---|---|
| Sex | 1.000000 | 0.100952 | -0.114631 | -0.245489 | -0.182333 | -0.691548 | 0.867334 | -0.552686 | -0.109689 | 0.140391 |
| Age | 0.100952 | 1.000000 | -0.267801 | -0.184627 | 0.125602 | -0.303490 | 0.205098 | 0.203422 | 0.106245 | -0.279046 |
| SibSp | -0.114631 | -0.267801 | 1.000000 | 0.414838 | 0.159651 | 0.084945 | -0.250489 | 0.059941 | -0.034538 | 0.040460 |
| Parch | -0.245489 | -0.184627 | 0.414838 | 1.000000 | 0.216225 | 0.102514 | -0.333905 | 0.221318 | 0.056498 | -0.036987 |
| Fare | -0.182333 | 0.125602 | 0.159651 | 0.216225 | 1.000000 | 0.120829 | -0.183766 | 0.105665 | 0.386297 | -0.482075 |
| Title_Miss | -0.691548 | -0.303490 | 0.084945 | 0.102514 | 0.120829 | 1.000000 | -0.599803 | -0.207996 | 0.065664 | -0.045347 |
| Title_Mr | 0.867334 | 0.205098 | -0.250489 | -0.333905 | -0.183766 | -0.599803 | 1.000000 | -0.479363 | -0.114673 | 0.137319 |
| Title_Mrs | -0.552686 | 0.203422 | 0.059941 | 0.221318 | 0.105665 | -0.207996 | -0.479363 | 1.000000 | 0.061767 | -0.121660 |
| Cabin_B | -0.109689 | 0.106245 | -0.034538 | 0.056498 | 0.386297 | 0.065664 | -0.114673 | 0.061767 | 1.000000 | -0.433053 |
| Cabin_U | 0.140391 | -0.279046 | 0.040460 | -0.036987 | -0.482075 | -0.045347 | 0.137319 | -0.121660 | -0.433053 | 1.000000 |

# STANDARDIZED

```python
continuous = ['Age', 'Fare', 'Parch', 'SibSp', 'FamilySize']

for var in continuous:
    titanic[var] = titanic[var].astype('float64')
    titanic[var] = titanic[var] - np.average(titanic[var]) / (np.std(titanic[var]))


titanic.head(5)
```

| | Sex | Age | SibSp | Parch | Fare | Title_Miss | Title_Mr | Title_Mrs | Cabin_B | Cabin_U | FamilySize | Ticket_XXX | Cabin_T | Cabin_G | Ticket_WEP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 19.844483 | 0.525455 | -0.473674 | 6.601578 | 0 | 1 | 0 | 0 | 1 | 0.81889 | 0 | 0 | 0 | 0 |
| **1** | 0 | 35.844483 | 0.525455 | -0.473674 | 70.634878 | 0 | 0 | 1 | 0 | 0 | 0.81889 | 0 | 0 | 0 | 0 |
| **2** | 0 | 23.844483 | -0.474545 | -0.473674 | 7.276578 | 1 | 0 | 0 | 0 | 1 | -0.18111 | 0 | 0 | 0 | 0 |
| **3** | 0 | 32.844483 | 0.525455 | -0.473674 | 52.451578 | 0 | 0 | 1 | 0 | 0 | 0.81889 | 1 | 0 | 0 | 0 |
| **4** | 1 | 32.844483 | -0.474545 | -0.473674 | 7.401578 | 0 | 1 | 0 | 0 | 1 | -0.18111 | 1 | 0 | 0 | 0 |