The Titanic

Predicting survivors

The Titanic

- The Titanic was a British passenger liner, operated by the White Star Line.
- She was on her maiden voyage from Southampton to New York City.
- Four days into this voyage she sank in the North Atlantic Ocean in the early hours of the 15th April 1912.
- Of the 2,240 passengers and crew on board, more than 1,500 lost their lives.
- It is believed the lack of lifeboats was a key reason behind the enormous loss of life.

Source: https://www.history.com/topics/early-20th-century-us/titanic
Titanic image: https://www.businessinsider.com/titanic-shipwreck-disappearing-dive-reveals-2019-9
Map image: https://risk-engineering.org/concept/Titanic

Can we reliably predict if someone survived the Titanic?

Data

Sourced from Kaggle as a csv file, containing information on 1,309 passengers
https://www.kaggle.com/datasets/vinicius150987/titanic3

Dataset contains 14 fields

Field	Number of null values
Passenger class, (1st, 2nd, 3rd)	0
Whether they survived	0
Name	0
Sex	0
Age	263
Sibsp (Number of Siblings/Spouses Aboard)	0
Parch (Number of Parents/Children Aboard)	0
Ticket number	0
Fare amount	1
Cabin number	1014
Port of embarkation (Cherbourg, Queenstown, Southampton)	2
Life boat number	823
Body identification number	1188
Home & destination (where they lived and destination)	564

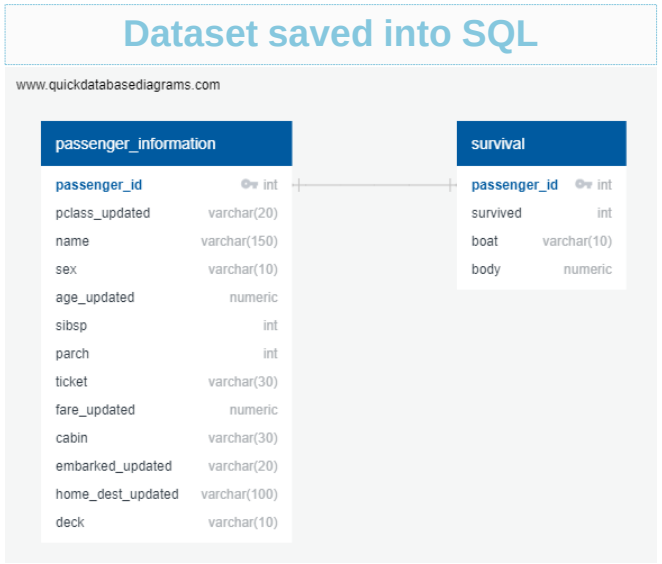
Low quality dataset

- At 1,309 records the dataset is incomplete as there were 2,208 passengers and crew on the Titanic.
- Data for the Titanic is split amongst various historical records, with discrepancies common.
- Additionally, some information is scare due to it being collected around the time of sinking (such as lifeboat numbers), or due to not all bodies having been recovered (body identification number).

Data was cleaned and saved to a SQL database

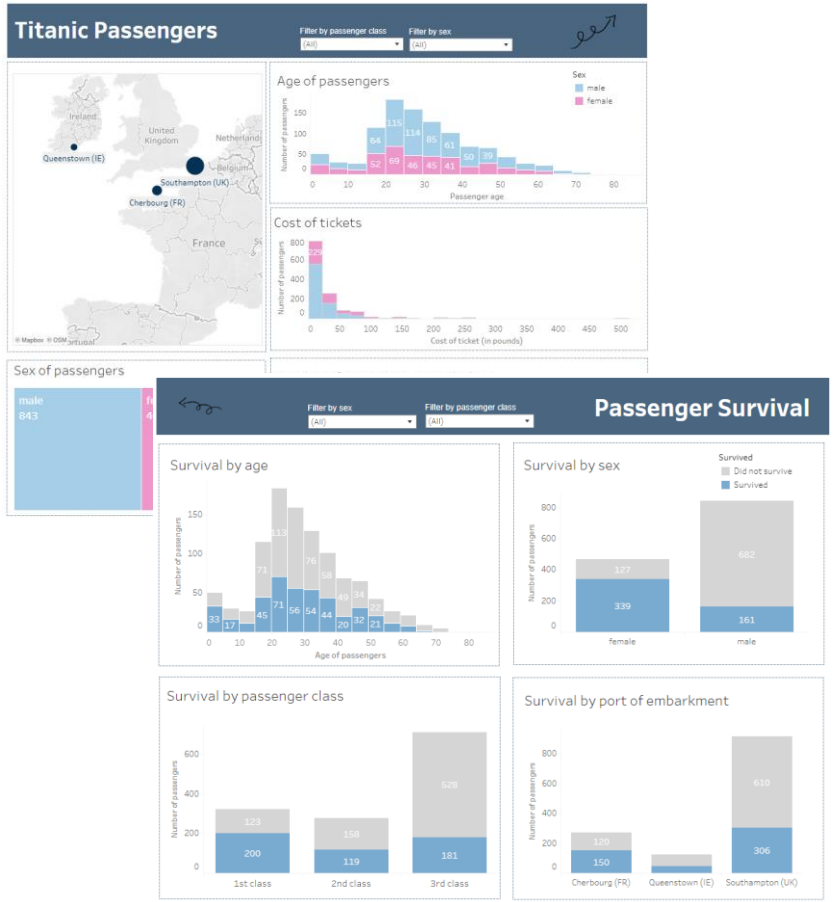
Age 263 null values	- An age value was calculated for those missing an age. It was calculated average age based on the passenger's sex and class they were travelling in.
Cabin 1014 null values	- Cabin information was not able to be obtained for the missing values. Null values were replaced with 'Unknown' in the data set.
Home and destination 564 null values	- Missing values were replaced with 'Unknown'
Fare & Embarked 1 null value | 2 null values	- Cabin information was not able to be obtained for the missing values. Null values were replaced with 'Unknown' in the data set.
Boat & Body 823 null values | 1188 null values	- These fields were not able to be updated and were not going to be used in the analysis given the model was looking to predict survival, so they were left as they were.

Dataset saved into SQL



Exploratory analysis

- Initial analysis of the data was undertaken in python to look for any patterns, anomalies and check assumptions. Summary statistics were reviewed, and charts created.
- Data visualisations were there developed in Tableau and are available via a webpage.
- Charts utilised were largely stacked bar charts to show differences between groups (such as males/females, whether they survived, passenger class). Histograms were used to show the distribution of data in relation to age and ticket prices.



Binary classification models

Predicting whether someone survived the Titanic lends itself to supervised machine learning classification model. As the outcome of the passengers is known, the model needs to predict whether they survived.

Logistic Regression
- It has good accuracy for simple data sets. - It is easier to implement, interpret and is efficient to train. - It is fast a classifying unknown records. - Less inclined to over-fitting but it can overfit in high dimensional datasets.

Random Forest
- Robust against overfitting because each weak classifier is trained on different pieces of data. - Robust to outliers and noise in the data. - Random forest ranks the independent variables in order of importance helping aid interpretation of the model.

Logistic regression

Focused on feature engineering

Given the outcome of the Titanic, in tuning the model an aspect focused on was fields that may indicate status and wealth. In doing this new fields were created from existing ones.

Deck they were staying on

- The first letter in a cabin number referred to the deck the passenger was on. This letter was extracted and saved to a new field.
- It also helped to further clean the data as often a family travelling together would have multiple cabin numbers assigned.

Number of family members onboard

- Given the cost of tickets, an assumption was that wealthier people may travel with a greater number of family members.
- A field was calculated by adding the columns sibsp (which is the number of siblings/spouses aboard) and parch (the number of parents/children aboard) together.

Title

- The title someone holds can also be an indicator of status and wealth.
- Passenger names included titles, so they were extracted and grouped together in a way to help indicate wealth.

```
# Creating a function to classify the titles into different groups
def title_grouped(title):
        if title == 'Miss' or title == 'Mlle' or title == 'Ms':
            return "Miss"
        elif title == 'Mrs' or title =='Mme':
            return "Mrs"
        elif title == "Mr" or title == 'Rev':
            return "Mr"
        elif title == "Master":
            return "male_child"
        elif title == "Col" or title == "Major" or title == "Capt" or title == 'Dr':
            return "Professional"
        elif title == "Don" or title == "Jonkheer" or title == "Sir":
            return "Noblility"
        elif title == "the Countess" or title == "Lady" or title == "Dona":
            return "Noblility"
        else:
            return title
```

Model 1

Balanced accuracy
0.7707

- Initial model run to get an understanding of performance, the feature engineered field ‘deck’ was included, and the sample was stratified.
- Fields included: survived, sex, sibsp, parch, age, embarked, fare, deck, pclass

The balanced accuracy score of the model is 0.7707389162561576

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	183	20
Actual - Survived	45	80

	precision	recall	f1-score	support
Did not survive	0.80	0.90	0.85	203
Survived	0.80	0.64	0.71	125
accuracy			0.80	328
macro avg	0.80	0.77	0.78	328
weighted avg	0.80	0.80	0.80	328

- The model had an accuracy score of 77% (based on balanced accuracy_
- The model is better in predicting those who did not survive, with a recall score of 90%.
  - Meaning it correctly predicts someone did not survive 90% of the time. Compared to only correctly predicting a survivor 64% of the time.
- The model achieves 80% precision for both groups, meaning when it does make a prediction that someone does or does not survive it is right 80% of the time.

Model 1 - tuned

Balanced accuracy
0.7796

- The data set appears to have a slight imbalance with 62% (n=809) of passengers not surviving and 38% (n=500) surviving. Given this, I wanted to understand the impact of using random over sampling on the data

The balanced accuracy score of the model is 0.7796453201970444

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	172	31
Actual - Survived	36	89

	precision	recall	f1-score	support
Did not survive	0.83	0.85	0.84	203
Survived	0.74	0.71	0.73	125
accuracy			0.80	328
macro avg	0.78	0.78	0.78	328
weighted avg	0.79	0.80	0.79	328

- The overall accuracy of the model lifted marginally from 77% to 78%.
- What was notable is that recall for survivors improved from 64% to 71%. Which means this model has improved its ability to correctly predict a survivor.
- Unfortunately there was a dip in recall for those not surviving (from 0.90 to 0.84) and in precision for survivors (from 0.80 to 0.74)

Model 2

Balanced accuracy
0.7747

- Optimisation - inclusion of the number of family members the passenger was travelling with as a new field, the original fields it was built from (sibsp and parch) were dropped.

```
The balanced accuracy score of the model is 0.77471921182266
```

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	170	33
Actual - Survived	36	89

```
                precision    recall  f1-score   support

Did not survive       0.83      0.84      0.83       203
       Survived       0.73      0.71      0.72       125

       accuracy                           0.79       328
      macro avg       0.78      0.77      0.78       328
   weighted avg       0.79      0.79      0.79       328
```

- Including the created field did not improve the accuracy of the model, with the results very similar to the previous model.

Model 3

Balanced accuracy
0.7806

- Optimisation made - reinstate sibsp and parch and include the feature engineered field called ‘title’, which was extracted from passenger's names.

The balanced accuracy score of the model is 0.7805714285714285

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	174	29
Actual - Survived	37	88

	precision	recall	f1-score	support
Did not survive	0.82	0.86	0.84	203
Survived	0.75	0.70	0.73	125
accuracy			0.80	328
macro avg	0.79	0.78	0.78	328
weighted avg	0.80	0.80	0.80	328

- The inclusion of the title field had a positive impact on the accuracy of the model lifting it from 77% to 78%.
- Precision and recall remained similar to model 1.

Model 3 - tuned

Balanced accuracy
0.8028

- Optimisation - as a well trained model is be better at predicting unseen cases, the train/test split was changed to 80/20, to provide more cases for training, while balancing the small number of cases in the dataset.

The balanced accuracy score of the model is 0.8028395061728395

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	137	25
Actual - Survived	24	76

	precision	recall	f1-score	support
Did not survive	0.85	0.85	0.85	162
Survived	0.75	0.76	0.76	100
accuracy			0.81	262
macro avg	0.80	0.80	0.80	262
weighted avg	0.81	0.81	0.81	262

- The change in the test/train split had a positive impact on the model, increasing accuracy to its highest level (80%).
- It also resulted in the highest recall score for predicting survivors, meaning that when the passenger is a survivor it correctly predicts it 76% of the time.

Random Forest

Model 1

Balanced accuracy 0.7864

- As a starting point, the fields from model 3 in logistic regression were used.
- Number of trees set to 500

```
The balanced accuracy score of the model is 0.786423645320197
The accuracy score of the model is : 0.8079268292682927
```

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	178	25
Actual - Survived	38	87

```
Classification Report
              precision    recall  f1-score   support

           0       0.82      0.88      0.85       203
           1       0.78      0.70      0.73       125

    accuracy                           0.81       328
   macro avg       0.80      0.79      0.79       328
weighted avg       0.81      0.81      0.81       328
```

- The model has an accuracy score of 79%.
- It performs similarly on precision and recall to strongest performing model from logistic regression. Except for recall of survivors, which is slightly lower in this mode at 70% (compared to 76%).

Model - tuned

Balanced accuracy 0.7840

- Increased the number of trees to 2,000

```
The balanced accuracy score of the model is 0.783960591133005
The accuracy score of the model is : 0.8048780487804879
```

	Predicted - Did not survive	Predicted - Survived
Actual - Did not survive	177	26
Actual - Survived	38	87

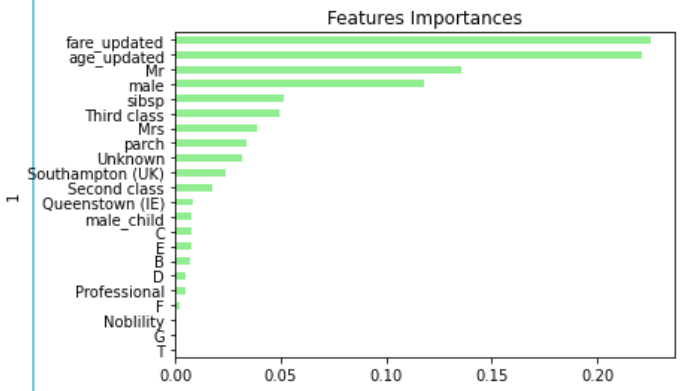
```
Classification Report
              precision    recall  f1-score   support

           0       0.82      0.87      0.85       203
           1       0.77      0.70      0.73       125

    accuracy                           0.80       328
   macro avg       0.80      0.78      0.79       328
weighted avg       0.80      0.80      0.80       328
```

- The inclusion of additional trees in the forest did not improve accuracy or recall and precision.

Feature importance



- Fare and age are the most important features feeding into the model at 22% importance each.
- Followed by Mr, (14%), male (11%), siblings and spouses (5%) and third class (5%).

Findings

Logistic regression had the strongest result	- Both logistic regression and random forest were used, with logistic regression having a slightly higher accuracy score (79% for random forest and 80% for logistic regression).

Optimisation improved the accuracy of the model	- Optimising the logistic regression model improved the accuracy from 77% to 80%.
	- Changes made were:
	  - Inclusion of the deck the passenger was staying on, as well as their title - as both can be indications of wealth/status.
	  - Random oversampling was used due to the slight imbalance in the data set.
	  - The train test split was changed to 80/20 to increase the models training, so it would be better at predicting unseen cases.

Model was more accurate at predicting those who did not survive	- The logistic regression model was more accurate in predicting those who did not survive with higher precision and recall scores. - In optimising the model, recall for predicting survivors increased from 64% in the initial model to 76% in the final model. Meaning the ability of the model to predict an actual survivor had improved.
Fare paid and age were key factors	- The random forest model showed the importance of features in the model with the top 5 features accounting for 75% of the importance. - They are: fare paid 23%, age 22% , Mr 14%, male 12%, siblings and spouses 5%.