Predicting the Survival of Titanic Passengers

Samuel Gilonis

Abstract

"The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others."

The objective of this paper is to explore the application of machine learning techniques to Kaggle's famous Titanic competition. The result was a gradient boosting algorithm which predicted the test set with 80.9% accuracy, placing it in the 94th percentile for model performance on Kaggle's leaderboard.

Contents

Exploring the data

The data set contains the following features (top five rows given as exaples):

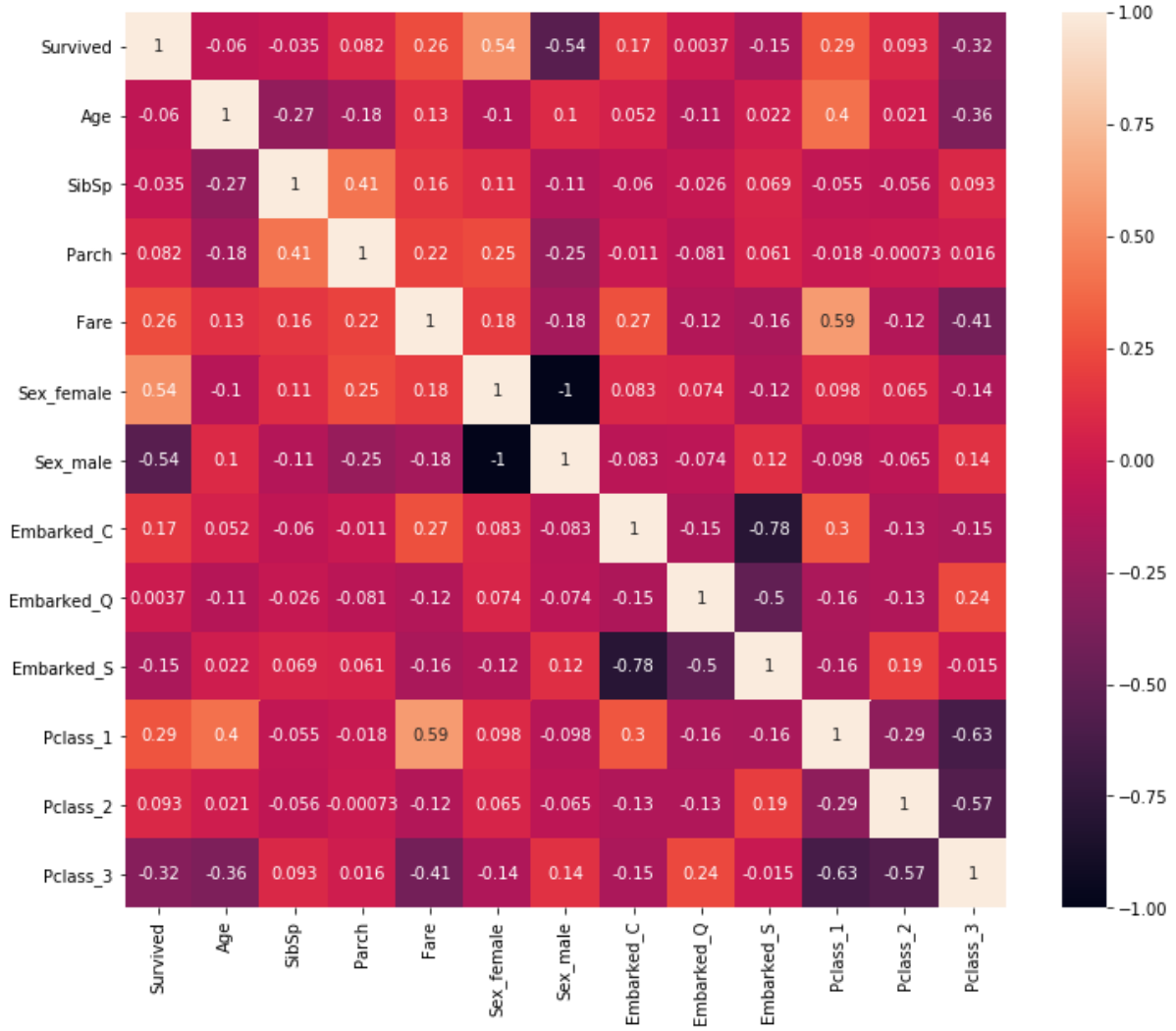
Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

The training data consists of 891 such rows and the test data, 418.

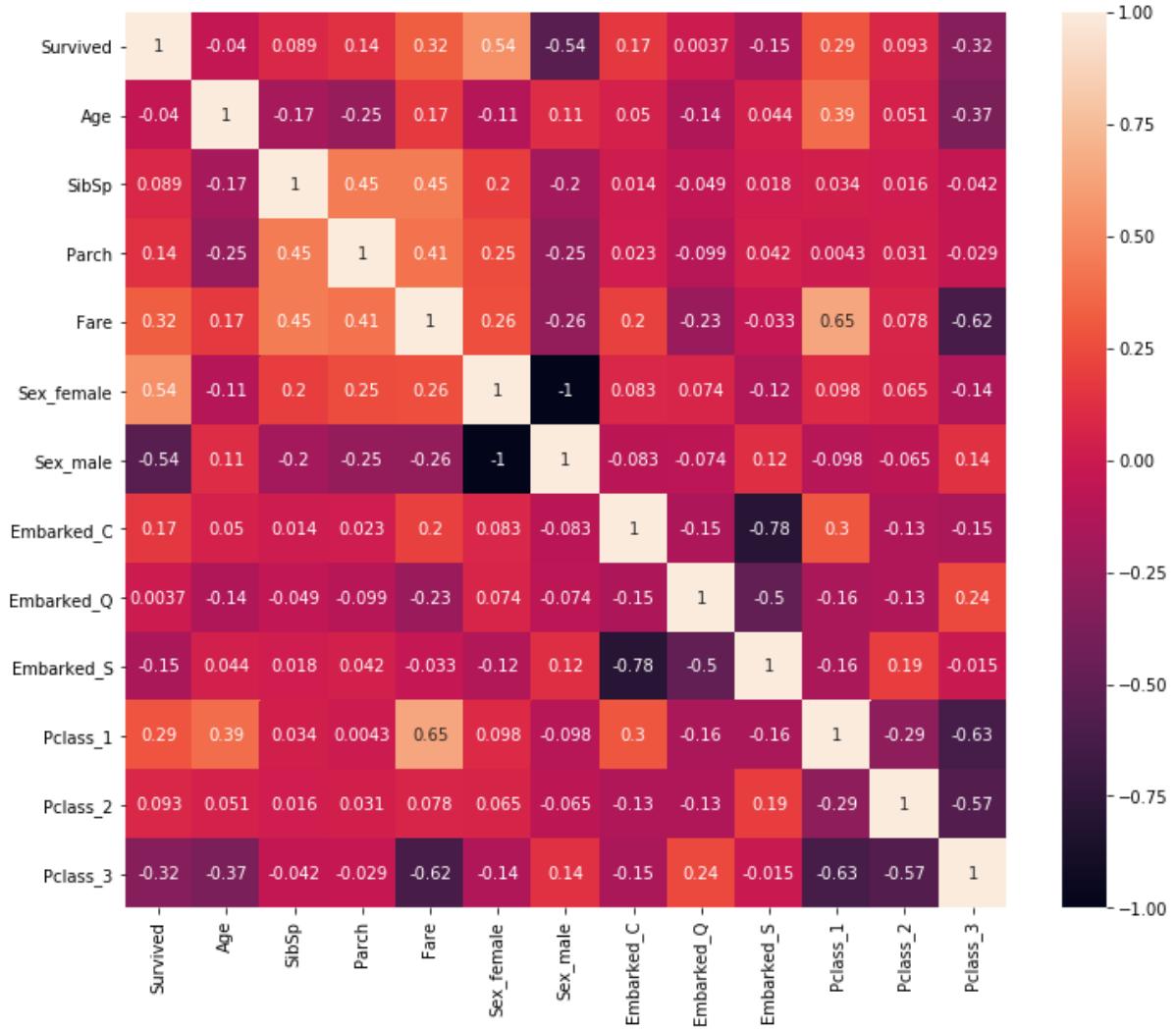
The features are self-explanatory except for perhaps SibSp (the combined number of siblings and/or spouse who are aboard RMS Titanic) and Parch (combined number of parents and children aboard RMS Titanic). E.g. a traveller whose wife, two brothers, both parents and two children were travelling with him would have SibSp = 3 and Parch = 4.

Correlations

Pearson's correlations measure how well each feature pair fit a straight line when plotted together. The correlation matrices below show some features that have been extracted/aggregated from the data as well as those in the original dataset:



Spearman's rank correlations measure whether there is any monotonic relationship between the feature pairs. This is of more interest to us as we are interested in any relationship, not just linear ones:

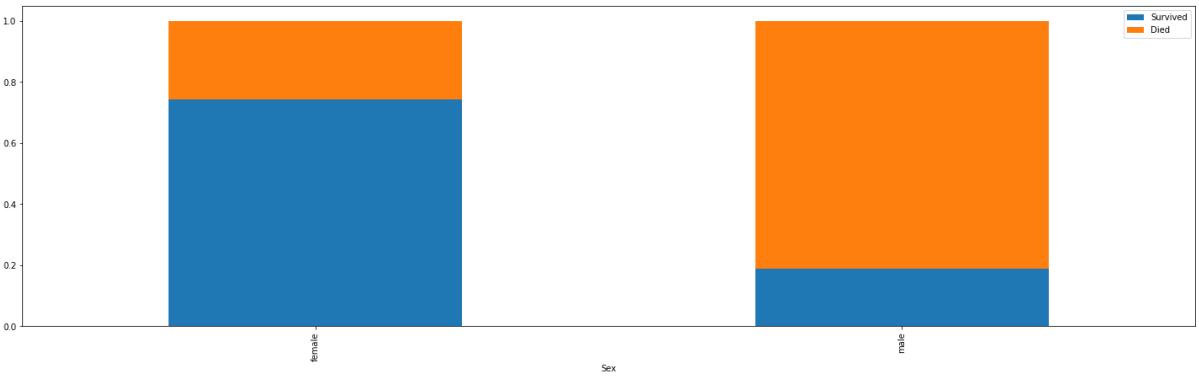


We can see immediately that sex, passenger class and fare paid stand out (although we can see that fare paid and passenger class are highly collinear).

It is worth looking at each feature in turn.

Sex

Sex is the most powerful predictor of survival on the Titanic. In fact, it is not trivial to model survivorship better than an algorithm which declares all females to survive and all males to perish.



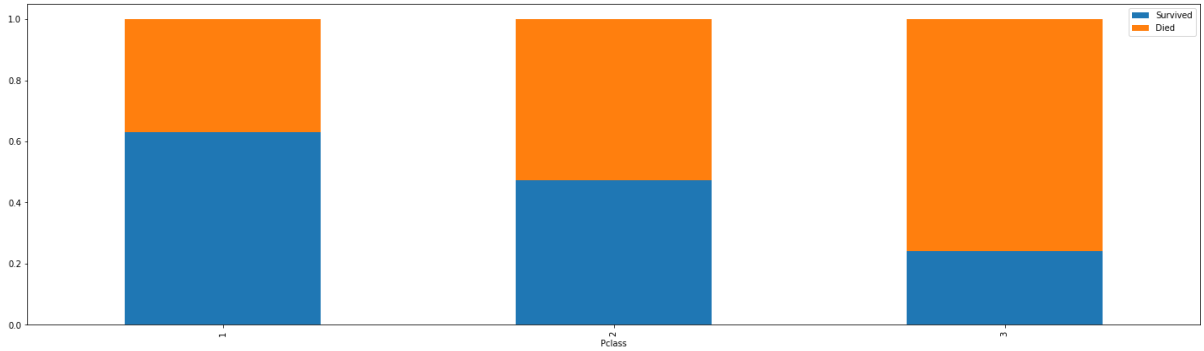
We can see that 81% of males died compared with only 19% of females.

The Phi coefficient is a measure of the association between binary variables which can take values between ±1 (indicating perfect agreement or disagreement, 0 indicating no relationship at all).

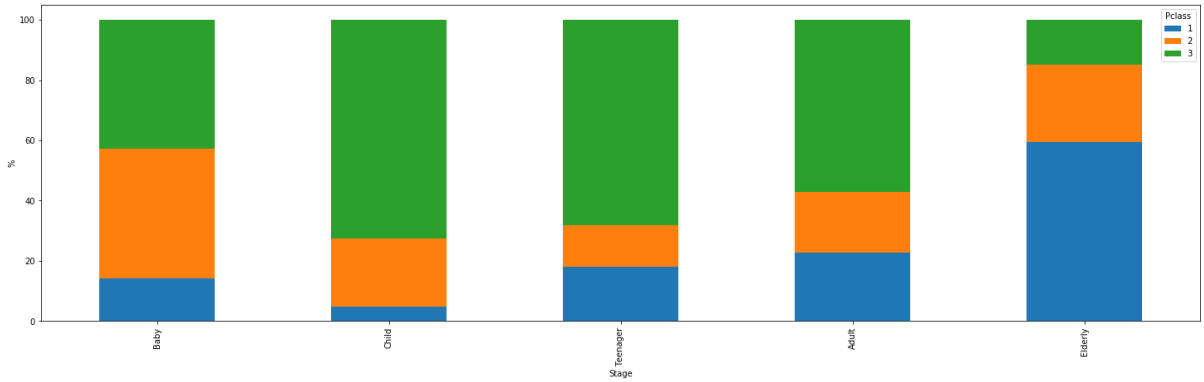
The Phi coefficient for sex and survival (1 being male, 0 being female) is -0.54.

Pclass

Passenger class is also a potent linear predictor of whether a passenger would have survived.

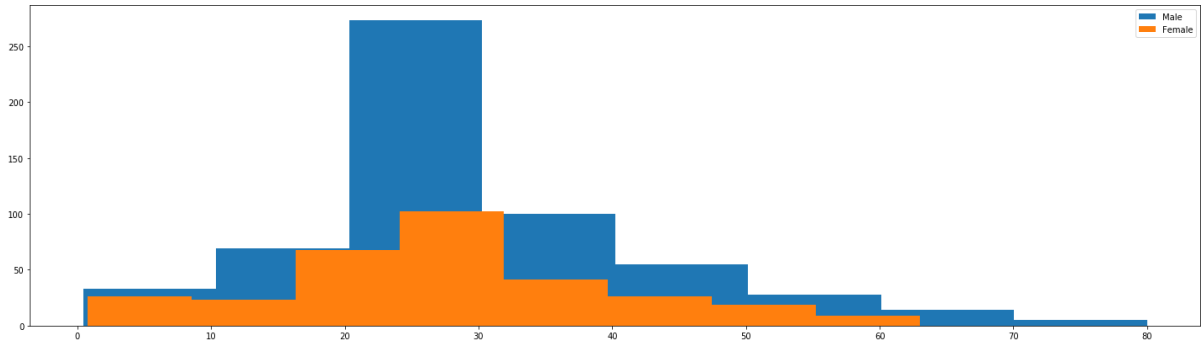


It is interesting to note that Pclass correlates quite strongly with age, with older travellers likely to travel in greater luxury. As we will see, this does not lead to greater survival probability amongst older travellers.



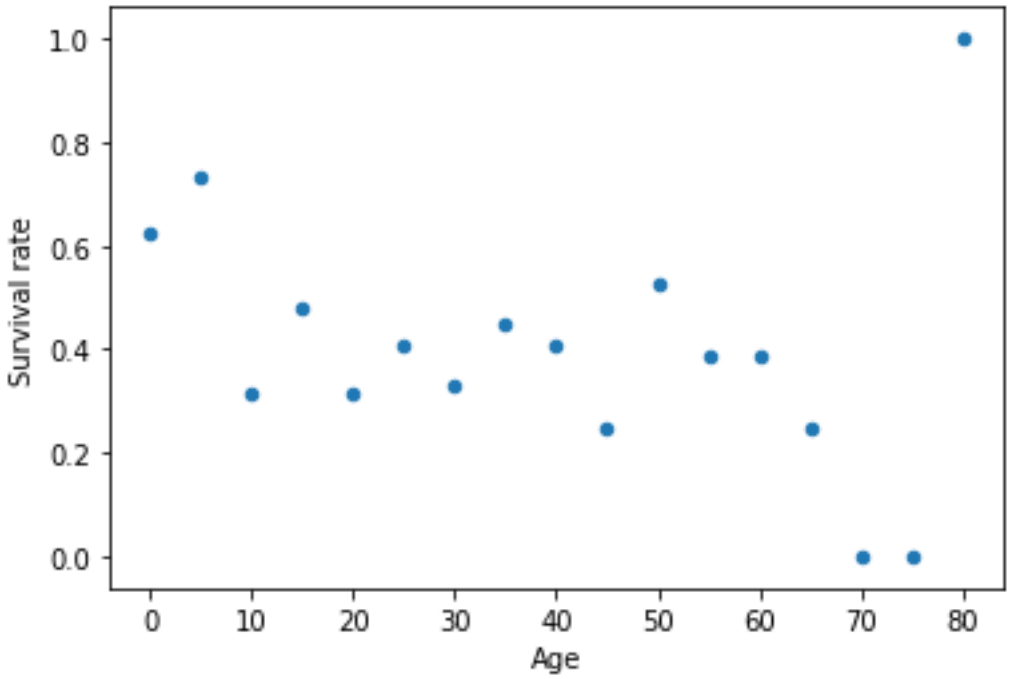
Age

We can see how the age of the passengers is distributed in the below histogram. Both distributions are skewed to the right.

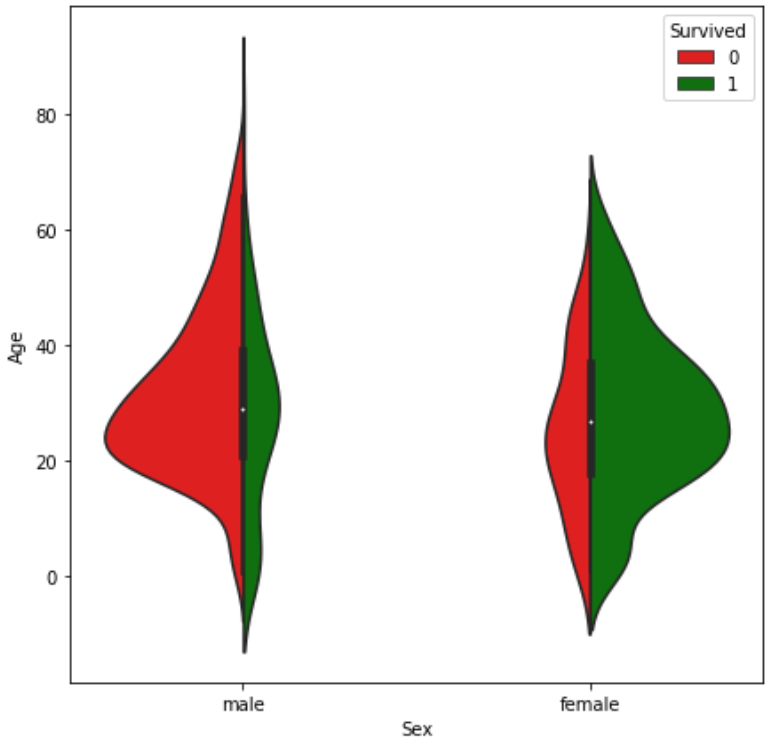


The mean age for the passengers is 30.14 and 27.93 years for men and women respectively. The median age for both sexes is 28.

By rounding each passengers age to the nearest five years we can see a weak correlation between age and probability of surviving:



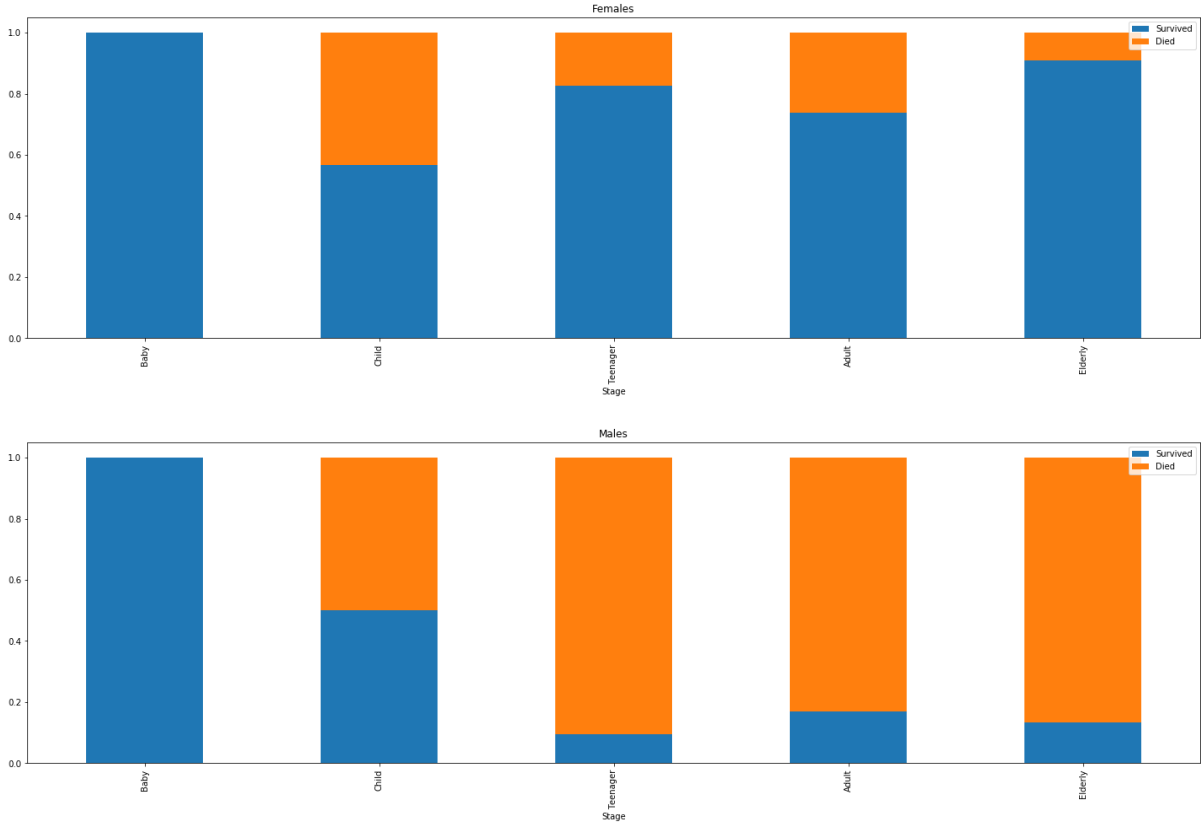
The below violin plot gives some insight into how age and sex interplay to affect an individual's chance of surviving. Violin plots are similar to box plots but show the probability density distribution at different values.



There are many things that we can see in the above plot but the key observation is that the effect of age on one's likelihood of surviving depends on sex. We can infer:

- Babies survive
- Males were more likely to survive than die as infants
- Males were more likely to die at any age thereafter
- Males over 70 were almost certain to die
- Females were more likely survive than die at any age
- Females were most vulnerable when they were children
- Females over 65 were almost certain to survive

We can also reduce age to a categorical variable and observe many of the same trends. Additionally, we can observe that male teenagers were particularly vulnerable. We might speculate that they neither had the attributes of grown men that would help them survive nor benefitted from any "women and children first" boost to their survival probability. It is also important to note that this demonstrates a non-linear relationship between age, gender and survivorship. This recommends including age in this categorical form rather than as a continuous variable. As the correlation between age and survival is quite weak, age appears to be a candidate for feature elimination as it is likely that the useful information is already contained in a passenger's title.



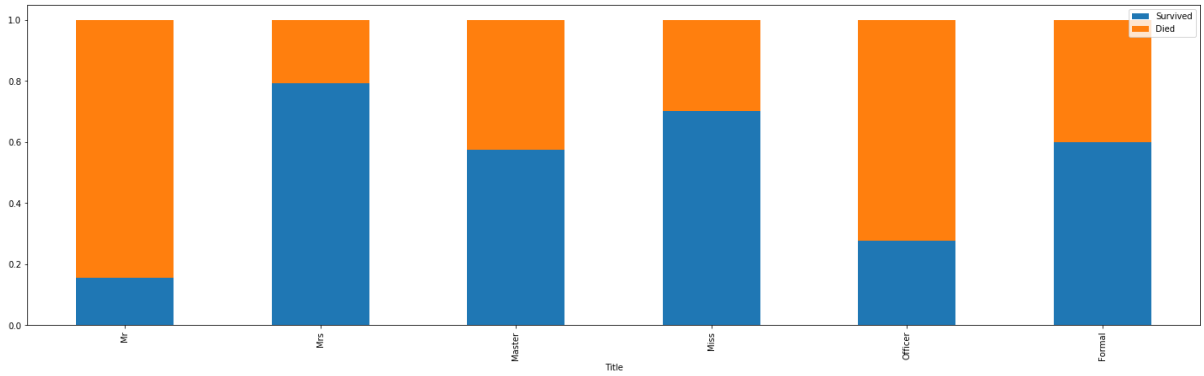
A few heuristics are emerging that could allow us to construct a rudimentary decision tree: we would barely need to worry about our grandmother, travelling in first-class, boarding RMS Titanic (93.3%

survival rate). However, our male teenage relative, travelling on the cheap, would be of great concern (5.55% survival rate).

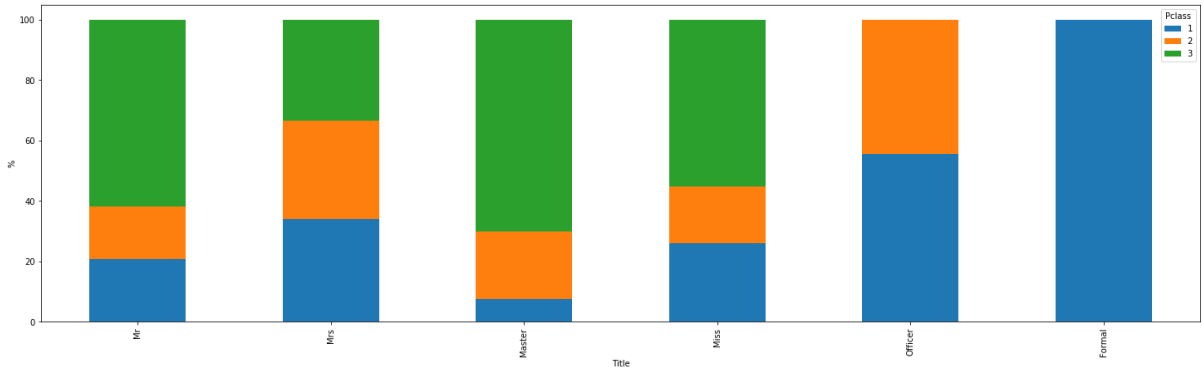
Name

I have standardised similar titles (e.g. ‘Mlle’ and ‘Ms’ have been amended to ‘Miss’). All military ranks or titles that denote a profession (e.g. ‘Dr’) have been amended to ‘Officer’ and all formal honorifics (e.g. ‘Sir’, ‘Lady’) have been amended to ‘Formal’.

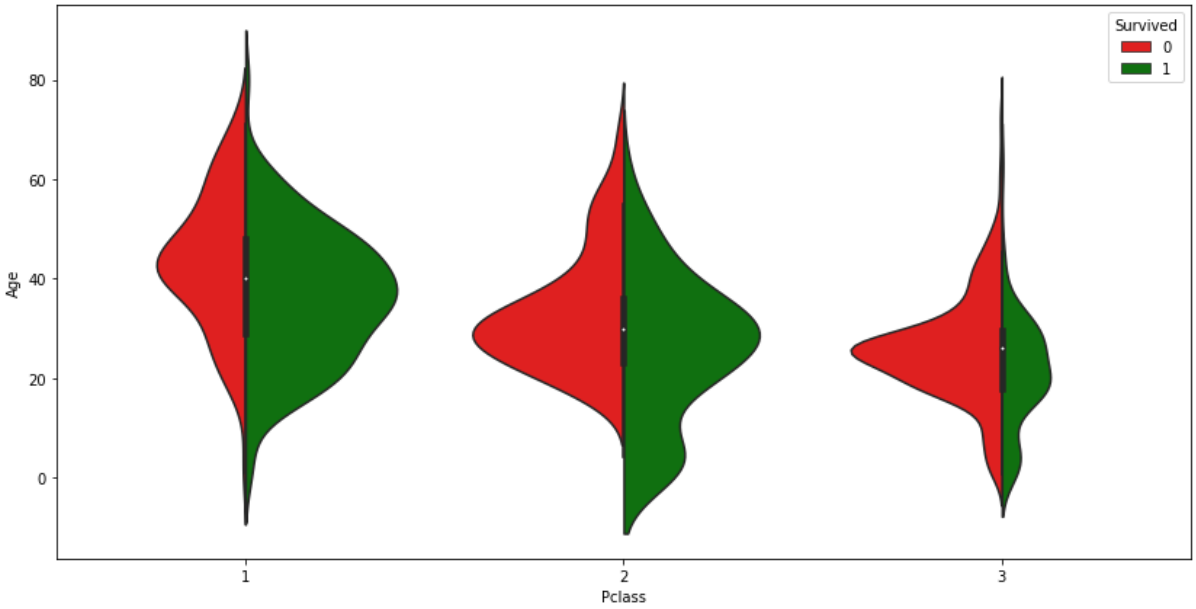
There is some useful information contained in a passenger’s name. Their title will correlate highly with gender, age and Pclass. It could, therefore, potentially serve to reduce the dimensions of the data.



The above plot demonstrates a lot of what we have already seen: men were more likely to die than women, young females were more vulnerable than older females and older males were more vulnerable than younger males. It is interesting that ‘officers’ seem to fair quite poorly as we might expect a favourable correlation with Pclass.



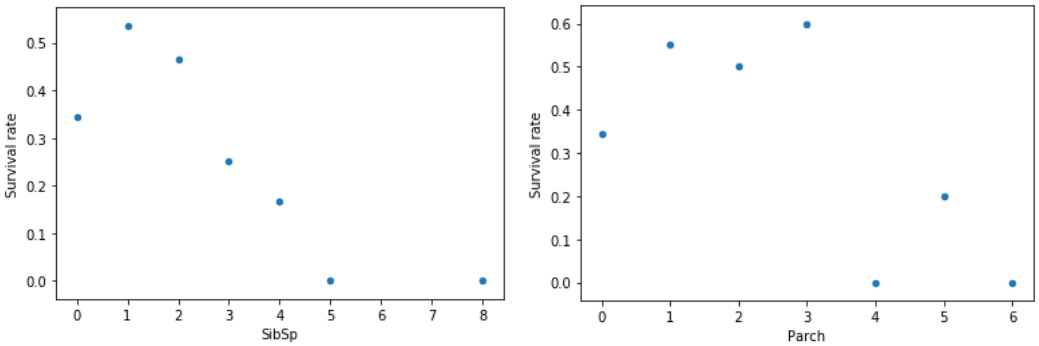
The officers were 94% male with a median age of 50. We may have expected their first- and second-class tickets to have sheltered them from the inferences we have drawn about the impact of age and sex on survival odds and we but they fared only slightly better than the Misters (with a median age of 28). The below violin plot demonstrates that the benefits of being in first-class (in terms of survival) were mostly enjoyed by the under-fifties.



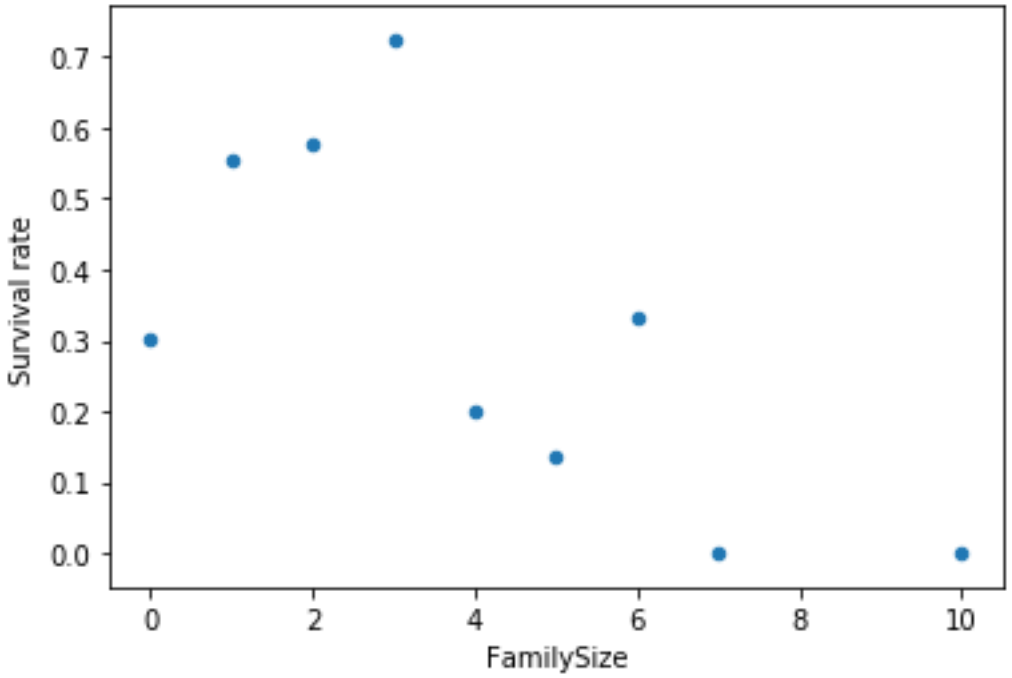
We could also possibly detect the origin of the names and then look to see whether any group was less likely to survive than another. This has not been attempted in this approach.

SibSp and Parch

SibSp (the combined number of siblings and/or spouse who are aboard RMS Titanic) and Parch (combined number of parents and children aboard RMS Titanic) both exhibit negative Pearson correlations with survival rate (-0.85 and -0.7 respectively).



We can aggregate these variables into one ‘Family Size’ variable:



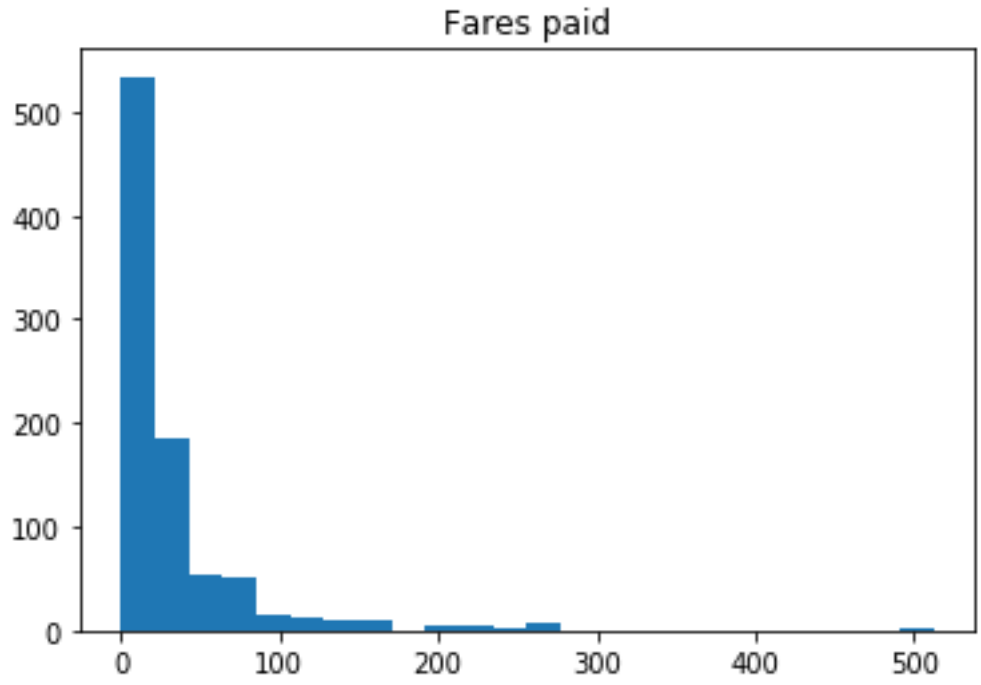
While the variance in these scatter plots should make as cautious of drawing too strong an inference, it is interesting to note that the above plots show decreasing odds of survival with additional family members (we can easily imagine individuals who would otherwise have survived being encumbered by children or elderly relatives) but passengers with no accompanying family members at all were also highly likely to perish. This implies a non-linear impact with a high boost to survival odds from not being alone but then a penalty for each additional party member. This recommends discretising this variable.

Ticket

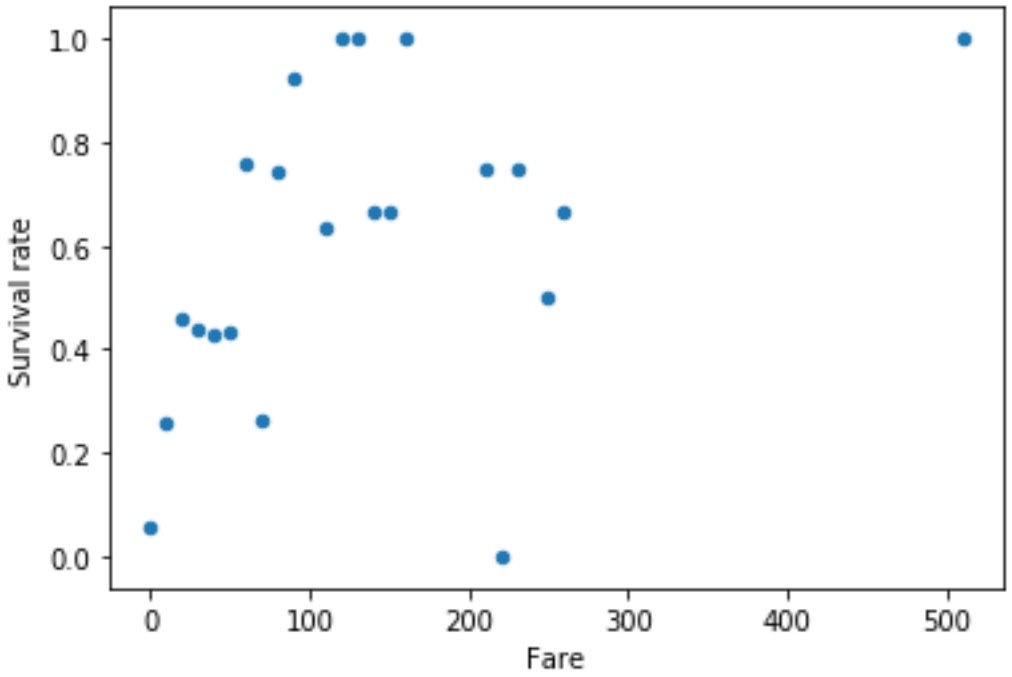
We could look for consecutive ticket numbers that might indicate fellow travellers. It seems plausible that clustering the passengers into parties that were travelling together would convey information on survival chances. This has not been attempted in this approach.

Fare

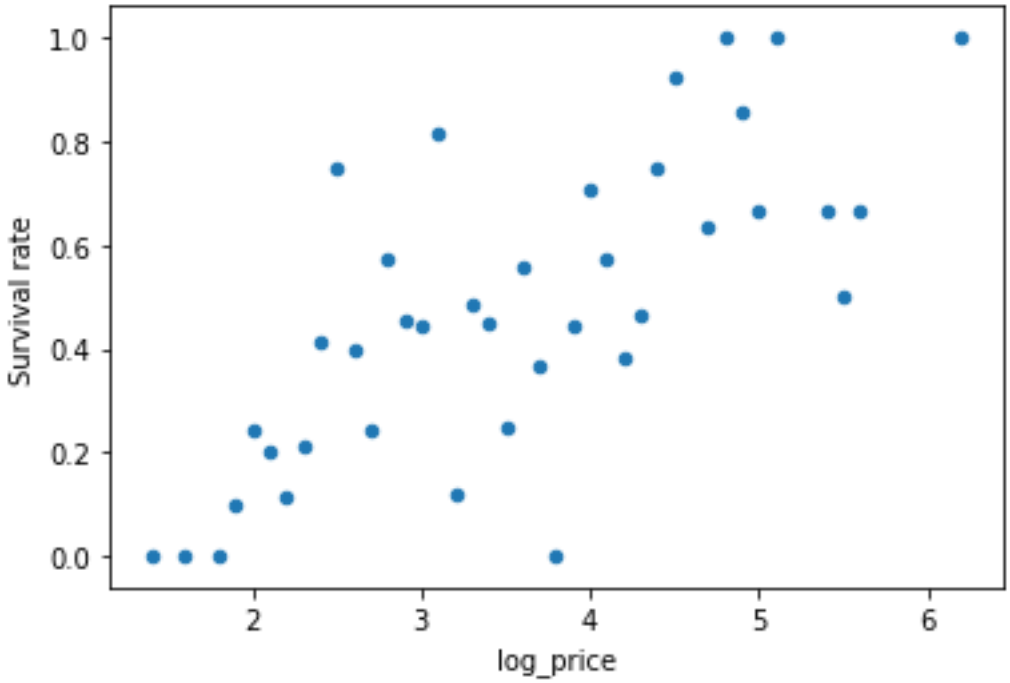
A histogram of the fares paid by each passenger shows a highly non-linear distribution:



We can plot the survival rate for each fare paid (rounded to nearest £10):



Even rounding the fares considerably, the pattern is not clear. We can take logs of the fares paid (and round to 1 d.p.) in order to eliminate the non-linear dynamics and effects of the skewed distribution when visualising the relationship:

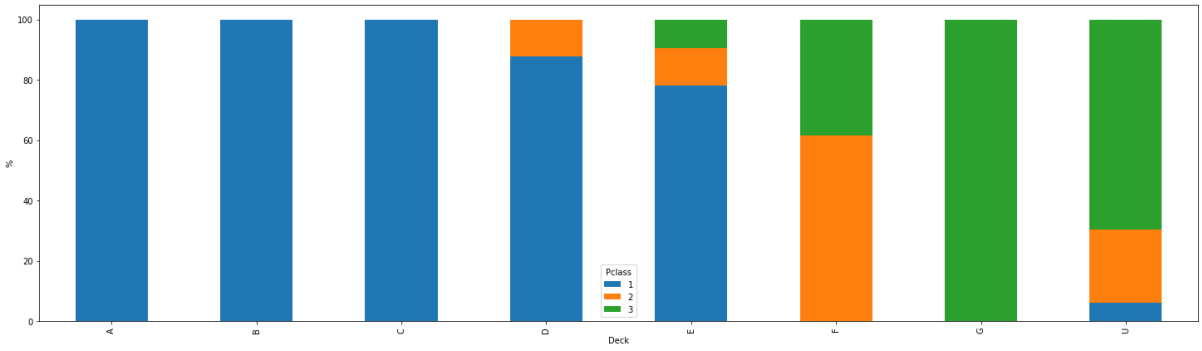


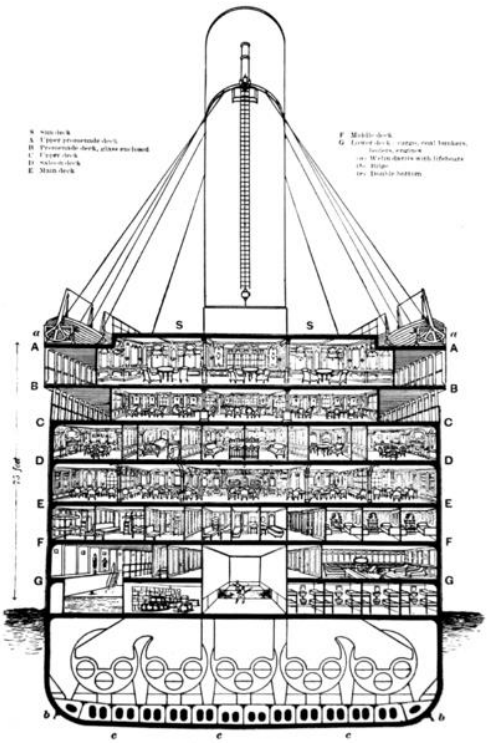
The Pearson's correlation of the log of fare paid with survival rate, excluding those who paid no fare, is 0.73.

Cabin

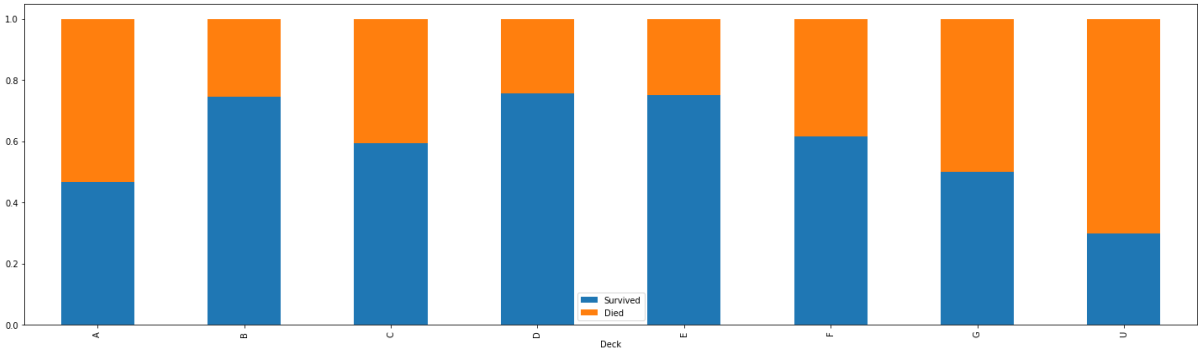
A passenger's cabin contains two pieces of information: the deck they were on and the room number. The room number may correspond to position fore/aft on the ship and so could conceivably be transformed to a useful purpose. This has not been attempted in this approach.

The deck that they were on is highly likely to be pertinent to a passenger's chances of survival. Unfortunately, most of the cabin information is missing and the passengers for which the cabin is known were mostly travelling first class. This can be seen in the below bar plot ('U' indicates unknown cabin).





However, we can see from the below proportions of survivors by deck that a passenger was more likely to survive from Deck F or G than Deck A despite the fact that Deck A was a first-class deck whereas F and G were second- and third-class decks.



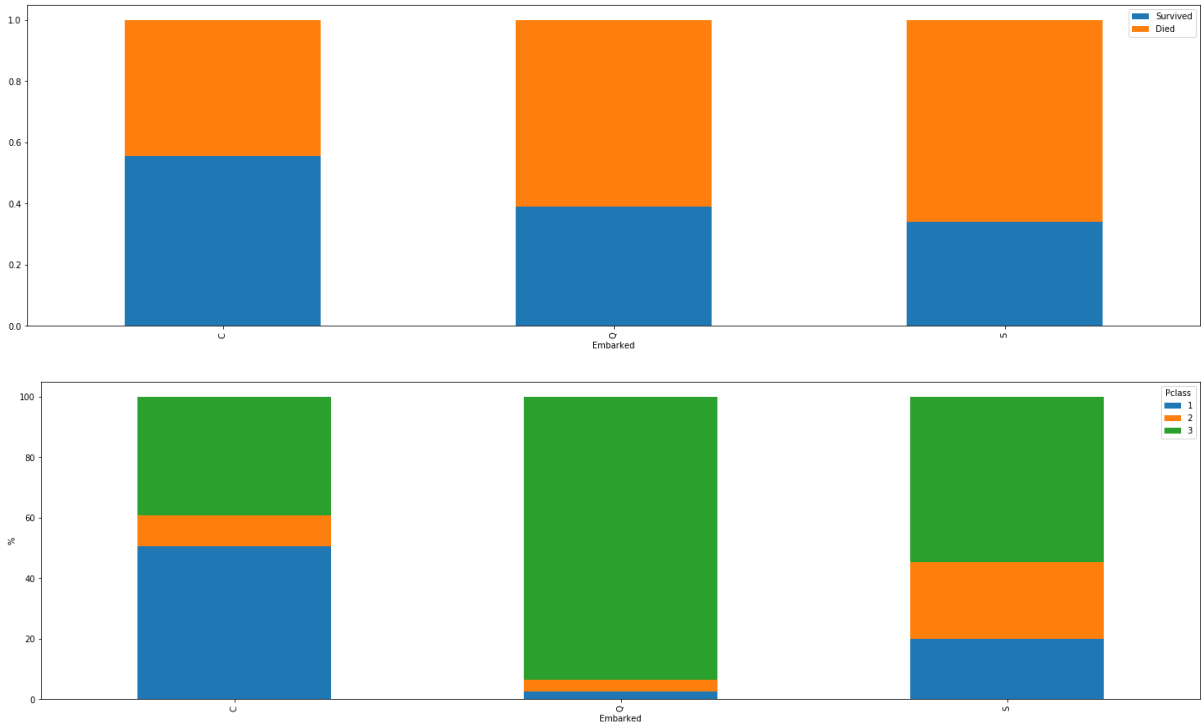
We cannot rule out the possibility that the deck a passenger was on had no relationship with a passenger's odds of survival and that the observed differences were pure chance, especially since the number of data points for each deck is quite small.

Let us model survival on Deck A (fifteen first-class passengers) as a binomial distribution with probability of survival equal to 60%. We would expect nine survivors and in reality, there were seven.

Our p-value, the probability of getting seven or fewer survivors by chance, is 0.2131 Likewise, the probability of getting 11 survivors was 0.2173 giving a combined p-value of 0.4304. Therefore, across Decks A-E (which are almost exclusively first-class), there was an approximately 94% chance that we would see a result as extreme as that which we see on Deck A.

Embarked

Passengers embarked at Southampton, Cherbourg and Queenstown. Passengers embarking at Cherbourg were significantly more likely to survive than those who embarked at Queenstown who were, in turn, marginally more likely to survive than those who embarked at Southampton.



Passenger class is the most obvious distinction but the demographic makeup of each consignment of passengers also varies in the distribution of age and sex.

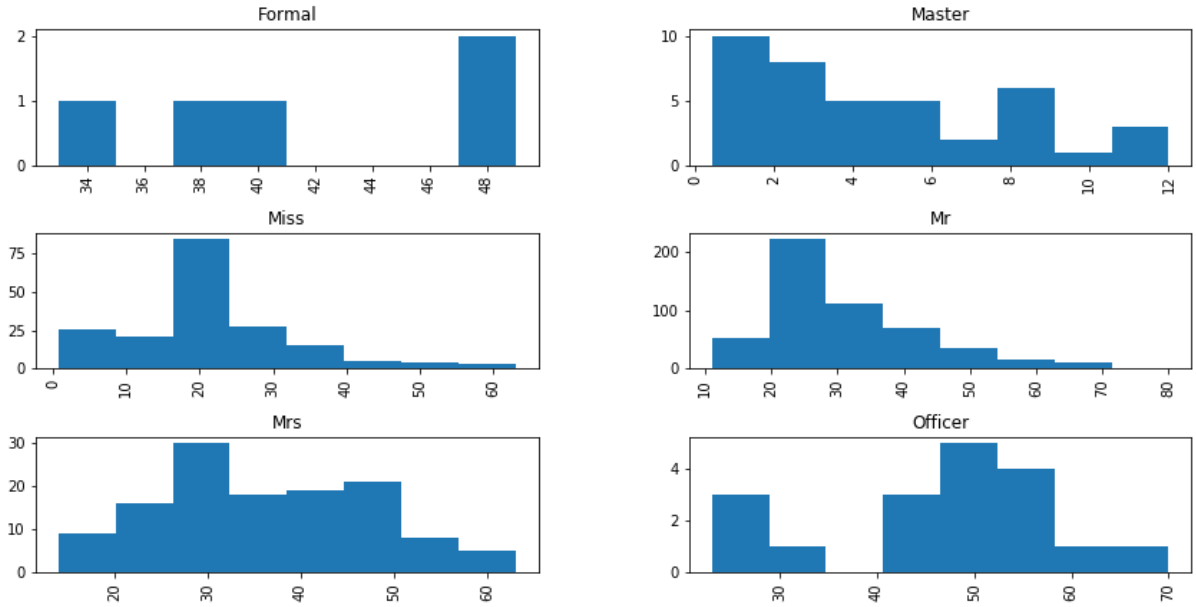
Pre-processing

A number of measures were taken to pre-process the data.

Imputing missing values

Some values were missing from the ‘Age’, ‘Fare’ and ‘Embarked’ features. Age and fare paid both have skewed distributions so I have imputed the median values for both.

We have seen that sex and passenger class are correlated with age. Title is also a useful predictor:



As is family size. However, for the sake of simplicity we shall group passengers by only sex, passenger class and title. We shall then take the median ages of these groups and impute those values.

Likewise, the median fare for each passenger class, embarkation point and family size was calculated and then used to impute missing values.

Only two passengers were missing the details of where they embarked so they were assumed to have embarked at Southampton, the most popular boarding point.

A more sophisticated method might be to look at the modal boarding point in relation to the fare these passengers paid.

Feature engineering

- The deck that the passenger boarded on was extracted from the Cabin feature.
- The passenger's title was extracted from their name using a REGEXP. Unusual titles where lumped together.
- Parch and SibSp were aggregate into one 'Family Size' feature and then this was discretised into a 'Family Type' variable (with values 'Lone', 'Small' (<5 members) and 'Large').
- A 'stage of life' feature was created with possible values: 'Baby (<1 years), 'Child' (<13 years), 'Teenager' (<18 years), 'Adult' (<50 years) and 'Elderly' (≥ 50 years).
- A 'FamilySurvived' feature was created by taking the survival odds of somebody with the same surname as the passenger. This exhibits high pearson correlation with survival odds in the test data but this is unsurprising. This was feature was removed as it severely harms the predictive power of the model.

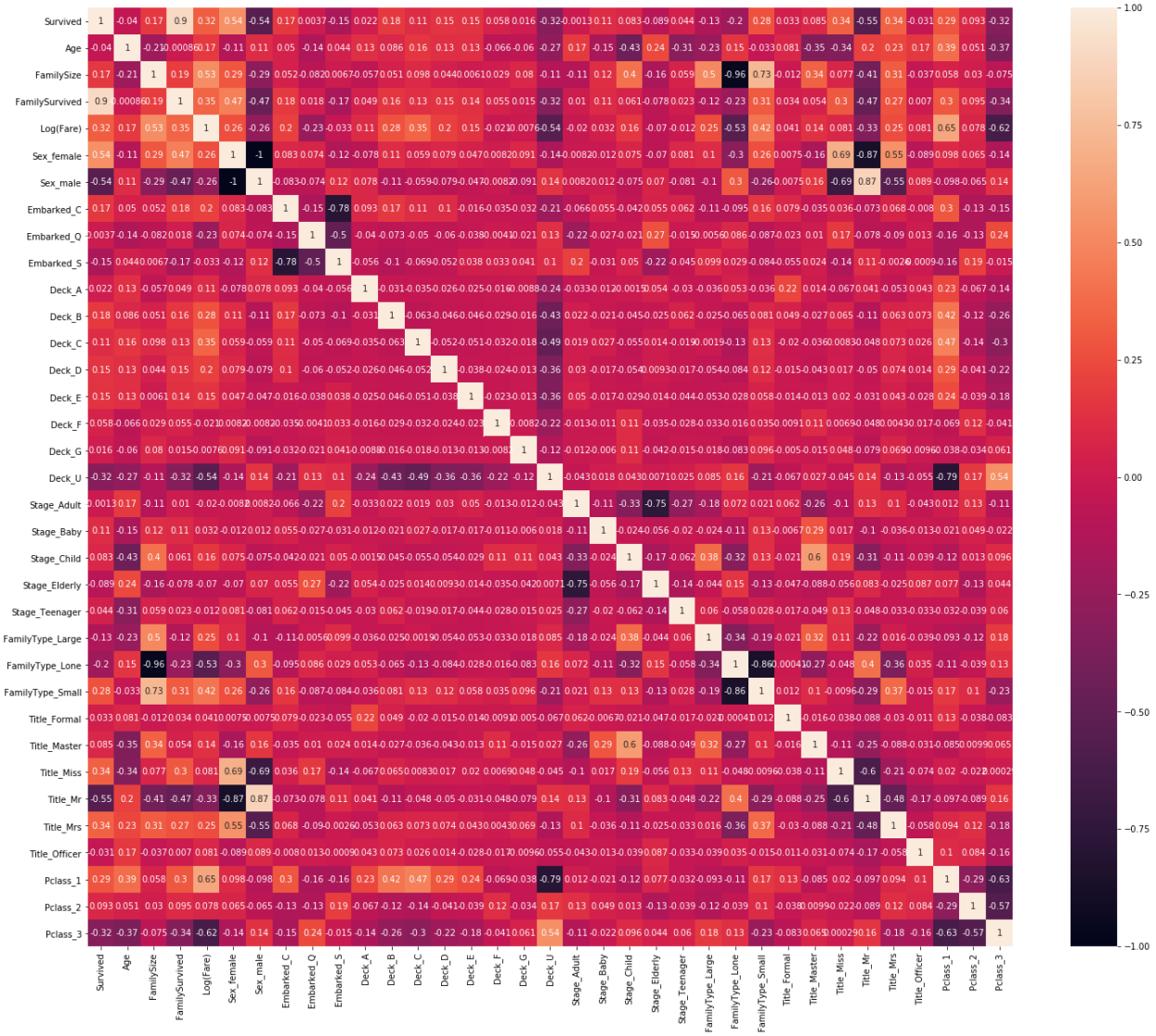
Encoding categorical data

All categorical features were encoded into n-1 binary vectors, where n = number of unique values, using Scikit-learn's OneHotEncoder. The first dummy variable derived from each feature is dropped in order to avoid the dummy variable trap (e.g. we do not need two binary variables to denote the sex of the passenger).

Feature scaling

The features were scaled using Scikit-learn's MinMaxScaler to remove the impact of varying magnitudes between the independent variables.

After the pre-processing it is worth plotting another correlation matrix (Spearman Rank) in order to see what features we should focus on:



Model-selection

A wide range of models were tested. I have included the results of some very crude, manual decision trees (e.g. all women survive, all men die) to benchmark model performance against.

Model	Accuracy	Brier score	Area under ROC curve
*No survivors!	0.614	0.386	0.500
*39 % chance	0.534	0.466	0.508
*Women survive	0.785	0.215	0.777
*Misters die	0.794	0.206	0.802

Logistic Regression	0.812	0.188	0.814
Guassian Naive Bayes	0.753	0.247	0.765
Bernoulli Naive Bayes	0.785	0.215	0.795
RBF SVM	0.821	0.179	0.813
Linear SVM	0.825	0.175	0.821
Polynomial SVM	0.807	0.193	0.800
Random Forest	0.789	0.211	0.783
Extra Tree	0.789	0.211	0.781
Neural Network	0.798	0.202	0.790
XGBoost	0.830	0.170	0.820

Accuracy

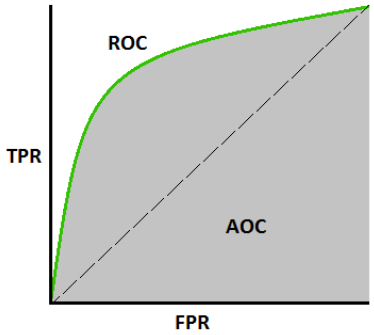
Accuracy represents the number of correct guesses divided by the total.

Brier score

A Brier score is essentially the mean squared error and measures the quality of the guess, taking the probability given into account.

Area under the ROC curve

An ROC curve plots sensitivity (true positive rate) against 1 – specificity (the false positive rate) at different threshold-probabilities. If we require the model to be 100% sure before classifying a survivor then we would expect very low sensitivity and a very low FPR. The corollary to this is that if we have a very low threshold-probability, we would expect lots of true-positives but also lots of false-positives.



A well performing model will, at an appropriate threshold probability, identify true positives without admitting false-positives.

We can see that the ‘No survivors!’ model has area equal to 0.5 under the ROC curve. This is exactly what we should expect from a useless algorithm as it predicts survival for the same proportion of passengers who survived as who died. Therefore, the true-positive rate equals the false positive rate. The same should apply to the model that gave each passenger a 39% chance of survival, it is probably a quirk of the training-test split that has made it appear marginally superior.

Selecting a model

There are a couple of interesting observations here. I have included some very crude decision trees against which to benchmark the performance of the other algorithms. A classifier which states that

nobody lives will fair better than one that gives each passenger a 39% chance of survival. This is to be expected since the odds of a correct guess in the latter model would be:

$$0.61^2 + 0.39^2 \approx 52\%$$

Whereas the odds of a correct guess if we state that all passengers will die are 61%. However, this illustrates that adding complexity/information to a model does not necessarily enhance predictive power. Likewise, we can see that the relatively simple Logistic Regression delivers the same returns as the far more complex and opaque neural net (with a lower Bier score and greater area under the ROC curve indicating that the passengers that it identified as likely to die were more likely to die and that there were fewer false positives). Perhaps more startling is that the NN barely performs better than a decision tree which states that all 'Misters' die. It is a reminder about the need for considering whether a given task is really a machine learning problem and that data-scientists should not be men-with-hammers, to whom everything looks like a nail.

The optimized distributed gradient boosting, XGBoost, algorithm delivers the best results therefore this model will be used on the true test data.

Gradient boosting is a technique for classification problems that converts an ensemble of weak learners into a stronger one. Each subsequent model is trained by the residuals (the difference between the predicted and true values) of the ensemble.

Feature selection

Recursive feature elimination

Recursive feature elimination seeks to remove dependent and collinear independent variables. It does this by recursively removing features and testing the F1 score of the model output. Features that are removed without degrading the F1 score are eliminated.

Removing features in this way had a deleterious effect on model performance. Therefore, some features were removed manually. The objective was to remove collinearity and noise. To this end the age and stage of life features were removed.

The final feature list is as follows:

```
Log(Fare)
Sex_male
Embarked_Q
Embarked_S
Deck_B
Deck_C
Deck_D
Deck_E
Deck_F
Deck_G
Deck_U
FamilyType_Single
FamilyType_Small
Title_Master
Title_Miss
Title_Mr
Title_Mrs
Title_Officer
Pclass_2
Pclass_3
```

Hyperparameter tuning

In order to tune the model, Stratified K-fold cross-validation was performed with a range of hyperparameters.

Hyperparameter	Description	Range of values	Final value
n_estimators	The number of decision trees.	100-1000	750
max_depth	The maximum depth of a tree. Used to control over-fitting.	3-8	3
min_child_weight	Defines the minimum sum of weights of all observations required in a child. Used to control over-fitting.	1	1
gamma	A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split.	0-0.1	0
subsample	Denotes the fraction of observations to be randomly samples for each tree	0.5-1	0.6
colsample_bytree	Denotes the fraction of columns to be randomly samples for each tree.	0.5-1	0.6
reg_alpha	L1 regularization term on weight.	0.01-0.1	0.1
learning_rate	Makes the model more robust by shrinking the weights on each step	0.01-0.2	0.05

Results

The gradient boosting algorithm performed well. The model predicted the survivors of the test set with 80.9% accuracy, placing it in the 94th percentile for model performance on Kaggle's leader board.