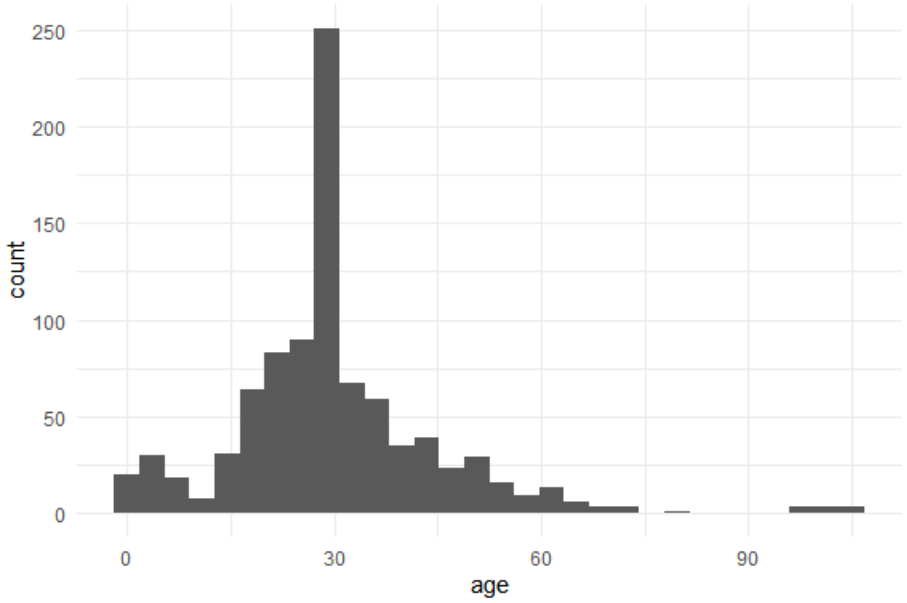
Practice Exam –Titanic Solution

Task 1 – Explore the survival rates by sex, pclass, and embarked (5 points)

Overall, data shows that 39.1% of the 906 passengers on this ship survived. We have information on 10 variables for each passenger. Based on statistical summaries of the percentage of survivors as well as the number of members within each group, we note the following:

- Passengers in higher classes (classes 1 and 2) survived more often than those in a lower class (class 3);
- The standard "women and children first" may have applied in this case, as survivors included 74% of women but only 20% of men;
- Most passengers (655) embarked from Southampton. They had a 34% survival rate, the lowest of the three ports;
- A log transform was applied to correct the large skewness in the fare distribution; and,
- There is an odd spike in mortality rates near the age of 30. The mean age is 30, the max is 105, and the median is 28.



Task 2 – Create a Title variable and simplify the factor levels (7 points)

When fitting predictive models, the input variables cannot have too many factor levels or they become too easy to overfit. The titles included in the passengers' names contain important information. This information has been extracted and stored as a new variable.

Each title has been placed into one of six categories. The expectation is that passengers with authoritative titles had higher social status and were therefore favored over those who did not have titles.

Title	Number of Passengers
Master	41
Miss	188
Mr.	523
Mrs.	131
Officer	23

Task 3 – Create a variable Family Size (2 points)

A new variable, *family_size*, was created using `sibsp`. The number of siblings was added to parch (the number of parents and/or children).

Task 4 – Use Kmeans clustering to look for outliers (10 points)

Some data was reported as unreliable, but it is unclear which specific passenger information has problems. An automated method is used to detect which passengers are different from the rest.

K-means is an unsupervised learning algorithm that seeks to partition data into homogeneous groups based on numeric variables. By minimizing the sum of variances (squared Euclidean distances) within each cluster, groups of observations, known as clusters, are created. In this context, it has been assumed that passenger records with missing values filled in are unique; therefore, they have been placed into a separate cluster.

Because this is such a small data set (< 1000 observations), a higher number of starting iterations, nstart, is needed to reduce the randomness in the k-means algorithm. It has therefore been set to 50.

Any value of k from 2 – 5 is sufficient as long as the outliers are separated from the others. If k = 2, there are 18 points in one cluster with all remaining points placed in a different cluster.

For k = 2, nstart = 500, the counts are:

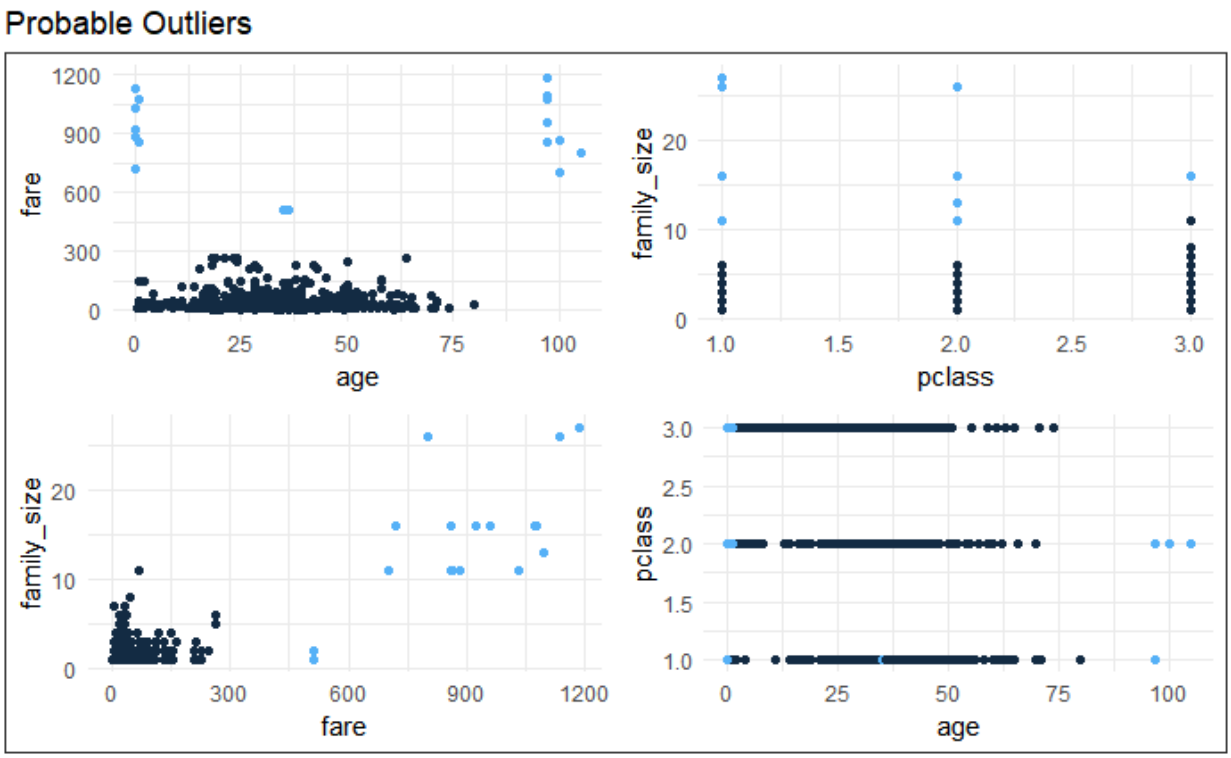
Cluster	Number of Passengers
1	18
2	888

If k = 3 and nstart = 50, there are 18 passengers in the smallest cluster. The following clusters are used.

Cluster	Number of Passengers
1	101

Cluster	Number of Passengers
2	787
3	18

Graphs of the variables used in the clustering are shown below. These demonstrate whether or not a passenger is considered an outlier. The 'fare and age' graph (upper left) shows that all outliers paid a much higher fare (over $600), but their ages are listed as '0'. The graph of 'fare and. family size' (lower left) shows that these outliers have a very large family size. Similar anomalies are evident in the pclass and age graph.



Task 5 – Select which variables should be used in modeling (3 points)

My assistant has already excluded *passengerID*. This is logical because as each passenger has a unique ID and therefore no useful information can be obtained from including this as a variable. *Name* and *ticket* were also excluded because these variables are too unique to be useful for modeling purposes.

Task 6 – Fit a Random Forest (7 points)

Because our goal is to understand the risk factors, the number of variables must be reduced to make the results easier to understand. Variable importance is a measure of their relative predictive power. We fit a random forest in order to get the 7 most important variables.

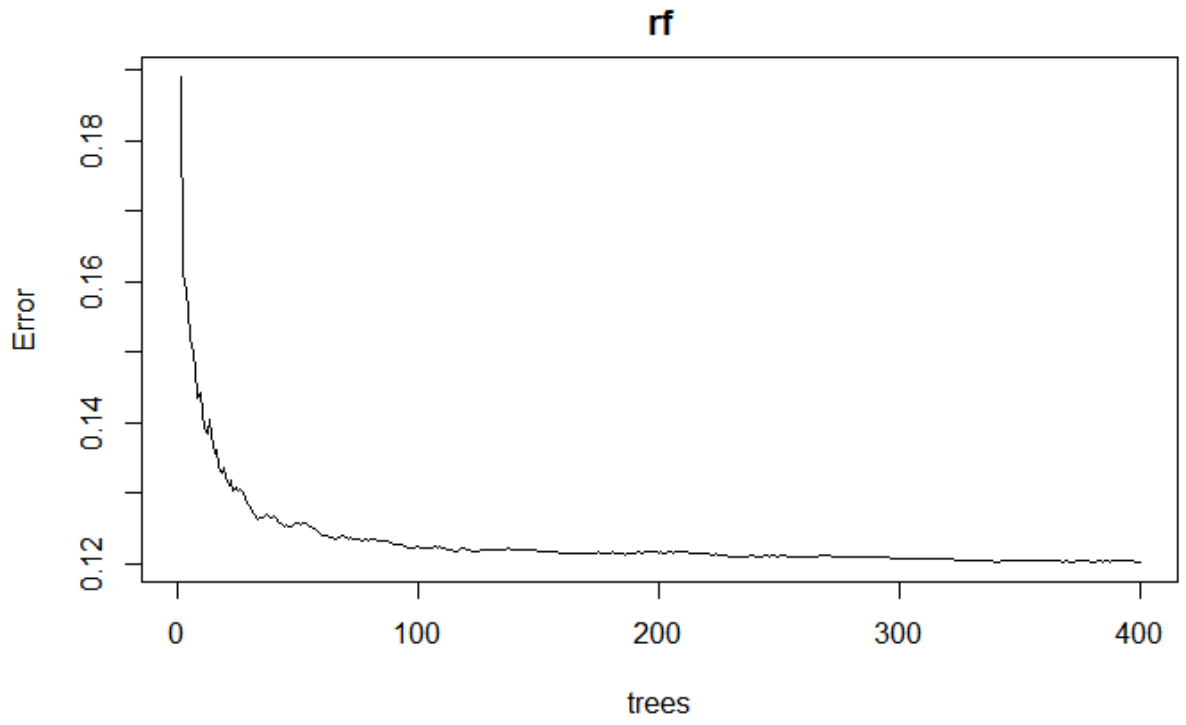
As the name implies, a forest is made up of *trees*. Separate trees are fit to sampled datasets using bootstrapped samples of the observations. The final prediction is an average of all trees.

AUC is often preferred over Accuracy for binary classifications when there are more observations in one class than in another. In our case, the training data shows 260 people who survive and 420 people who die. Just focusing on overall accuracy would be biased towards making good predictions about those that die (because there are more of them) but less accurate predictions about those who survive. For example, if we predict that everyone will die, the accuracy would be 420/(420+260) = 62%.

I look at the documentation `?randomForest` to see descriptions of the ntree and mtry parameters.

ntree	Number of trees to grow. To ensure that every input row gets predicted at least a few times, this parameter should not be set too low.
mtry	Number of variables randomly sampled as candidates at each split.

A cross-validation shows that the error rate flattened off after about 200-500 trees and so I chose to use a value of 200 for ntree.



The value of 300 for `mtry` which your assistant has set is illogical because there are only 15 predictor variables in the data. Therefore, it is impossible for any given tree's split to consider more than 15 variables. I set mtry to 4, following a rule-of-thumb stating that a good starting value is the square root of the number of predictor variables, which turned out to be 3.9 in this data. (Any value between 2 – 15 would be okay).

The default metrics were:

```
[1] "Accuracy"
[1] 0.7699115
[1] "AUC"
[1] 0.7966208
```

After updating the parameters to ntree = 200 and mtry = 4, both accuracy and AUC improved (increased).

```
[1] "Accuracy"
[1] 0.8097345
[1] "AUC"
[1] 0.8329579
```

Task 7 – Select the Top 7 Features (7 points)

Variable importance is a way of measuring how each variable impacts a model's predictions. This is a numeric measure between 0 and 100 based on relative importance of the variable, where 0 means unused, and 100 means most important.

The top 7 variables in order of importance were:

1) Sex_female
2) Pclass
3) Fare
4) Age
5) Family_size
6) Title_Master
7) Title_Miss

Task 8 – Fit a Logistic Regression and interpret the signs of the coefficients (4 points)

A logistic regression is a GLM. As survival is binary (a person can either survive or not), a binomial family is appropriate. A logit link is the most common for binary classifications.

At first, the logit was not fitting because the variables used were linearly dependent, and so the model matrix was rank-deficient. This is because the family_size variable is a linear combination of *parch* and *sibsp*. Removal of *parch* fixed this issue.

The coefficients are below.

term	**estimate**
<chr>	<dbl>
(Intercept)	0.479061065
sex_female	3.087536939
pclass	-1.108434232

age	-0.004398902
fare	0.138923333
title_Master	2.403650920
family_size	0.003569770
title_Miss	-0.155543080

The sign indicates the direction of change in the log odds. As the log odds increase, the probability increases, and as the log odds decreases, the probability decreases. Therefore, the sign of the coefficient indicates the effect of each variable on the probability.

The probability of survival:

- Is higher for women;
- Is lower for lesser passenger classes (pclass = 2 or 3);
- Decreases as age increases;
- Increases as fare increases;
- Is higher for passengers with the title "Master";
- Decreases as family size increases; and,
- Is higher for passengers with the title "Miss".

Task 9 – Calculate the probabilities of a person of age 10, 30, and 60 surviving (9 points)

Logistic regression is a type of generalized linear model which models the probability of survival. A binomial response family is used, since passengers can either survive or die. The logit is a common choice for the link function.

The odds of surviving reflect the probability of surviving divided by the probability of dying. The logit link means that the linear predictor is a log of the odds, and so the coefficients apply to the log odds and not to the probability itself.

To convert from log odds to probability, p is the probability of survival, z is the linear predictor, and z = log (p/(1-p)). To find the probabilities for different ages, the values of sex of fare, family_size, and sibsp are held constant at their mean and the sex_female, pclass and title_Miss have been held constant at their mode. The tables below show these calculations.

		Average Passengers with		
Term	**Coefficient**	**Age = 10**	**Age = 30**	**Age = 60**
(Intercept)	0.5	1	1	1
sex_female	3.1	0	0	0
pclass	-1.1	3	3	3
age	0.0	**10**	**30**	**60**
fare	0.1	2.7	2.7	2.7
title_Master	2.4	0	0	0
family_size	0.0	1	1	1
title_Miss	-0.2	0	0	0

	Age = 10	Age = 30	Age = 60
Linear Predictor	-2.51	-2.60	-2.73
Survival Probability	7.5%	6.9%	6.1%

Task 10 –Validate the Logistic Regression and compare the AUC and Accuracy to the random forest in Task 5 (4 points)

A worse performance may reasonably be expected because the random forest has 14 variables so only keeping 7 of them will result in less information. In addition, the random forest automatically captures interaction effects and non-linear transformations whereas the logit does not. For example, if there are non-linear patterns in age or fare, or in interactions between pclass and family_size then these factors would not be replicated in the logit. Due to these conditions, the p-values are large for some variables.

Accuracy and AUC have both decreased (gotten worse).

RF

```
[1] "Accuracy"
[1] 0.8097345
[1] "AUC"
[1] 0.8329579
```

Logit

```
[1] "Accuracy"
[1] 0.800885
[1] "AUC"
[1] 0.8302461
```

Task 11 – Fit an Elastic Net with all pairwise interactions (6 points)

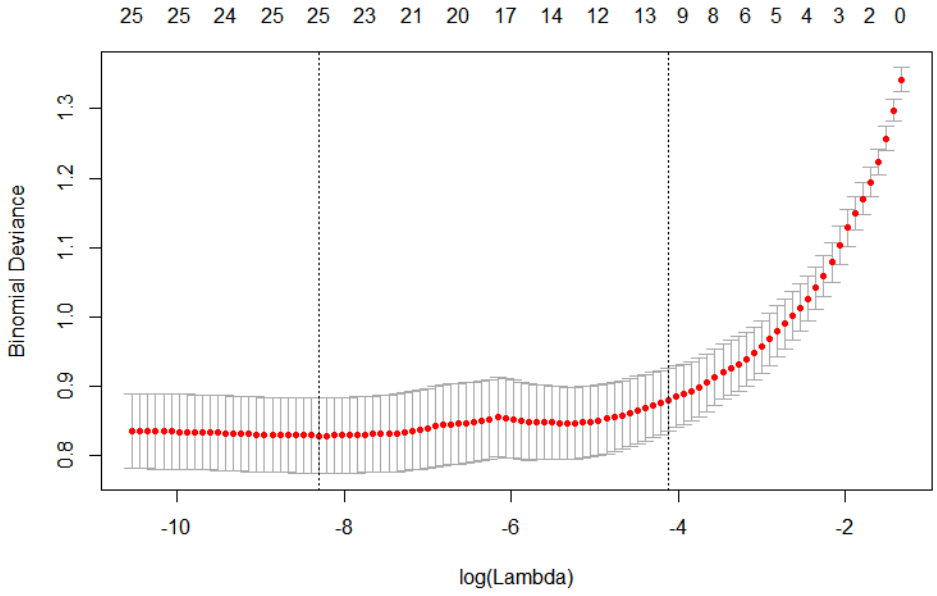
An elastic net with alpha = 1 is known as a Lasso. This is a special type of GLM that performs variable selection automatically by forcing some coefficients to be zero. This is done through the use of a penalty term lambda. As lambda increases, some variables are dropped from the model and the sum of the absolute value of the coefficients decreases.

An interaction is when the impact of a variable on the response changes depending on the value of another predictor variable. It is unclear which combinations of interactions will be predictive, but by looking at all 21 possibilities, the Lasso can decide which ones should remain in the model. I use cross-validation to estimate how the training performance changes for different values. The error metric is a binomial deviance, high when the target is likely to be a good fit and low when this is unlikely. This is not an exhaustive search, because three-way, four-way (and higher) interactions are possible.

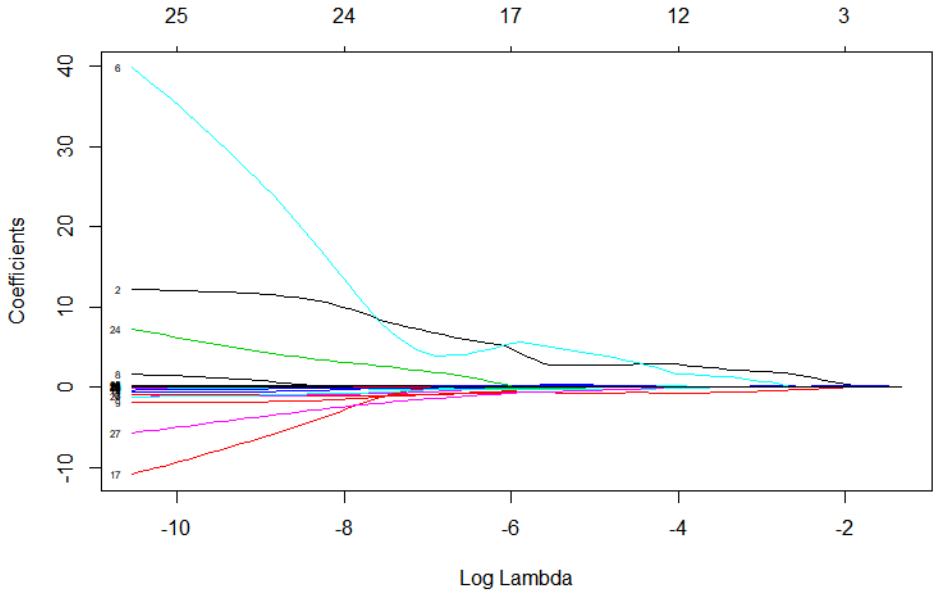
(Bonus): After considering all pairwise interactions, there are (p + p – choose - 2) terms where p is the number of original variables.

Task 12 – Choose the value of Lambda so that between 5-10 variables remain (6 points)

The graph below shows that the best value of log (lambda) is about -10, which means that lambda is about exp (-10). This is a very small number which results in a low penalty and hence a complex model with many variables.



The code template was configured to create a graph of the coefficients against the percent of deviance explained on the x-axis. As it was important to see lambda values, lambda is shown on the x-axis instead.



Each curve represents a variable's coefficient. This graph shows the path of the coefficients against the log of the penalty term lambda. As lambda gets larger, the penalty term increases and the flexibility of the model gets smaller. I chose the value of Log Lambda = -4, or lambda as exp (-4), which is larger than the CV results as fewer variables are included than would have been the case if the lambda from the cross validation had been used. Using the CV min lambda would have resulted in too many variables.

The resulting variables and interactions are:

- Sex_female
- Pclass
- Fare
- Title_master
- Sex_female and fare
- Sex_female and family_size
- Pclass and age
- Plcass and family_size
- Fare and family_size

Task 13 – Compare the AUC with the Logit's AUC (6 points)

The AUC with the Lasso is higher (better) while the accuracy is lower. Note that although these metrics have not been penalized by the model's complexity as would have been the case with AIC or BIC; the error is on the test set. Therefore, it would have accounted for models overfitting on the training data.

Logit

```
[1] "Accuracy"
[1] 0.800885
[1] "AUC"
[1] 0.8302461
```

Lasso

```
[1] "Accuracy"
[1] 0.8097345
[1] "AUC"
[1] 0.8375887
```

Task 12 – Executive summary (24 points)

In order to successfully market and price this specialty life product, we need to understand the risks associated with travel cruises. Many different types of people choose to go on cruises, and some have a much higher likelihood of death, including drowning or experiencing another unfortunate life-ending circumstance. By segmenting high-risk from low-risk policies we can improve the profitability of Life Jacket™.

We used historical data from the Titanic, where the fate of each passenger is known. Based on 10 different variables, we built a predictive model to estimate the probability that an individual passenger would survive.

The key drivers of survival in order of importance are:

1. The passenger's gender
2. Their ticket class
3. The fare (ticket price)
4. Their age
5. The size of their family on board
6. Whether they have the title "Master"
7. Whether they have the title "Miss"

As you think about marketing this product, use these factors to identify which customers are high risk vs. low risk. Each of these factors reflects whether a passenger is more or less likely to survive.

The probability of survival:

- Is higher for women;
- Is lower for lesser passenger classes (pclass = 2 or 3);
- Decreases as age increases;
- Increases as fare increases;
- Is higher for passengers with the title "Master";
- Decreases as family size increases; and,
- Is higher for passengers with the title "Miss".

The next logical step would be to inquire *why* passengers in higher classes survived more often. Is the belief of "women and children first" the reason why these passenger groups were more likely to survive? Why did passengers with more expensive tickets have a higher likelihood of survival? Did they have better access to life vests and lifeboats? Could older people have been at greater risk because of poor health? These are all questions which are relevant to Life Jacket™.

These factors are not independent. We built a model which can account for multiple effects simultaneously. For example, the effect of gender on survival probability was dependent upon the ticket price. Women who had families had different survival rates than those who did not. Passengers that had first-class tickets had the best chance at surviving, but this probability was affected by the size of their families and their age. Looking at ticket price alone only explains part of the story as families may have paid different ticket prices or gotten a group discount.

For the typical passenger, increasing age lowers the probability of surviving, as can be seen in the following survival rates:

10 year old children - 7.5%;

30 year old adults - 6.9%; and

60 year old adults- 6.1%.

We began with an observational analysis of the data. We looked at the survival rates by sex, pclass, and embarkation point. We found that passengers in classes 1 and 2 were more likely to survive than those in class 3, that women and children were more likely to survive, and that most passengers embarked from Southhampton.

Information about passenger titles, such as Dr. or Mrs., was obtained from the list of names. We also found the family size of each passenger based on the number of parents and children cited.

The data is more than 100 years old, and so not all records are trustworthy. We used an automated method to remove suspect records, known as kmeans clustering. We assigned each passenger to a group based on age, fare, class, and family size. Then we used an algorithm to find the groups which were most similar to one another. We found that there were 18 passengers who were very different from the others, with such anomalies as extremely high fares or very large families. As removing these would have substantially impacted the small data set, we created an outlier flag.

There were a lot of variables to consider. To simplify, we used a random forest (RF) to determine which were most predictive of survival. This is a data-mining method which is easy to calibrate and provides a measure of the importance of each variable. The RF has the advantage of automatically detecting which variables are predictive, removing those that are not, detecting interaction effects, and is robust to outliers and missing values. The disadvantages of this technique are mainly lower interpretability and slow training time (both irrelevant to the current case). We only considered the top 7 variables rated by importance.

Once we had reduced the number variables to 7, we fit a logistic regression (Logit). This is a type of linear model which considers multiple inputs simultaneously and predicts the survival probability. We looked at the accuracy, the proportion of passengers which survived, as well as the AUC, a statistical measure that balances accuracy on both the survivors and those who died. In a blind holdout test, this model predicted which passengers would survive with 80% accuracy.

We then considered interaction effects, when the variables act together to influence the probability of survival. The logit alone does not detect interaction effects and so we tested all 21 variables and interactions and then used a penalization term (also known as a Lasso) to remove those not needed. As the number of variables used is controlled by a single chosen parameter, 5-10 variables remained out of the original 21. We used cross validation to make these decisions. By adding interaction effects, the AUC improved from 0.830 to 0.837.

This data set of Titanic passengers had fewer than 1000 records. Therefore, the analysis could have been heavily biased if passengers were not randomly chosen. We relied on several advanced machine learning algorithms but did not always use cross-validation. Especially when working with a small data set, using 5-fold or 10-fold cross validation may lead to better model results than using separate train/test splits.

Because this incident happened so long in the past, the results would likely be different for tourists today. There were also only three ports considered, and so the findings could vary for other geographies.

Our investigation has identified the most important risk factors for passengers onboard the Titanic. Such an event is a complex process involving many people interacting over several hours, and so the

goal of predicting who will live and who will die is challenging. Safety standards of cruise ships have improved in the last hundred years, and it is possible to extend this study by applying similar statistical analyses to recent cruise experiences. It would also be beneficial to obtain further insight into passengers with missing data.