Titanic: Machine Learning from disaster

Alex Costa, Giovanni Caminiti | Università Cattolica del Sacro Cuore

12/03/2025

Introduction and data loading

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. In this challenge, we are asked to build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data.

The dataset was taken from Kaggle at this link: https://www.kaggle.com/competitions/titanic and it consist of 891 observations of 12 variables:

```
## Rows: 891
## Columns: 12
## $ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1~
## $ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3~
## $ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Fl~
## $ Sex         <chr> "male", "female", "female", "female", "male", "male", "mal~
## $ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, ~
## $ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0~
## $ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0~
## $ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "37~
## $ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625,~
## $ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", "", "", "", "G6", "C~
## $ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S"~
```

We have: 4 numerical variables, of which 2 continuous (Age, Fare) and 2 discrete (SibSp, Parch); 4 categorical variables (Survived, Sex, Embarked, Pclass); 3 alphanumeric data variables (Name, Ticket, Cabin).

We remove the variables Id, Name, Ticket and Cabin because the first is just an identification and the others are alphanumeric and we are not going to use them in the analysis.

Missing values

First of all we check for the presence of missing values in the dataset.

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch     Fare
##        0        0        0        0      177        0        0        0
## Embarked
##        0
```

We have 177 missing values for Age. We choose to impute the missing values using the Multivariate Imputation by Chained Equations (MICE) technique. It is an iterative method that imputes missing values in each variable using models based on the other variables in the dataset. MICE can use different imputation methods, we use the Predictive Mean Matching which is good to preserve the original data distribution.

Exploratory Data Analysis

Before starting the exploratory data analysis we replace the variables SibSp (number of siblings and spouses aboard) and Parch (number of parents and children aboard) with their sum, which represents the Family Size. We then convert this variable into a categorical by grouping in this way:

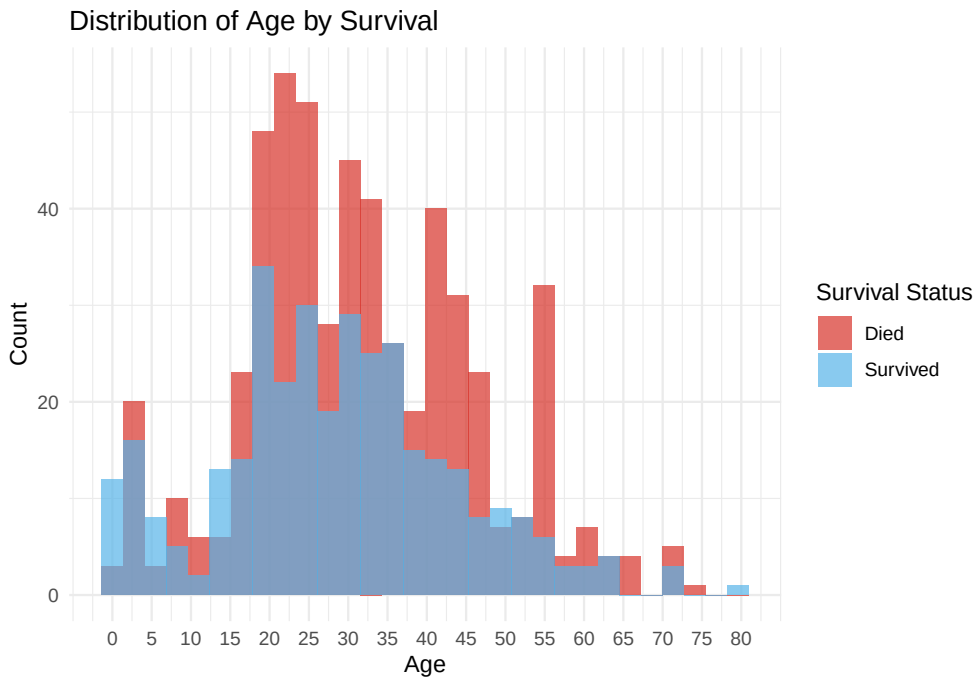
- Single: Family Size = 1
- Small: Family Size = 2 to 3
- Medium: Family Size = 4 to 5
- Large: Family Size $\geq$ 6

```
##  Survived Pclass      Name              Sex           Age        SibSp
##  0:549    1:216   Length:891         female:314   Min.   : 0.42   0:608
##  1:342    2:184   Class :character   male  :577   1st Qu.:20.75   1:209
##           3:491   Mode  :character                Median :29.00   2: 28
##                                                   Mean   :30.47   3: 16
##                                                   3rd Qu.:40.00   4: 18
##                                                   Max.   :80.00   5:  5
##                                                                   8:  7
##  Parch        Fare        Embarked     Family          Fsize
##  0:678   Min.   :  0.00   C:170    Min.   : 1.000   Single:537
##  1:118   1st Qu.:  7.91   Q: 77    1st Qu.: 1.000   Small :263
##  2: 80   Median : 14.45   S:644    Median : 1.000   Medium: 44
##  3:  5   Mean   : 32.20            Mean   : 1.905   Large : 47
##  4:  4   3rd Qu.: 31.00            3rd Qu.: 2.000
##  5:  5   Max.   :512.33            Max.   :11.000
##  6:  1
```

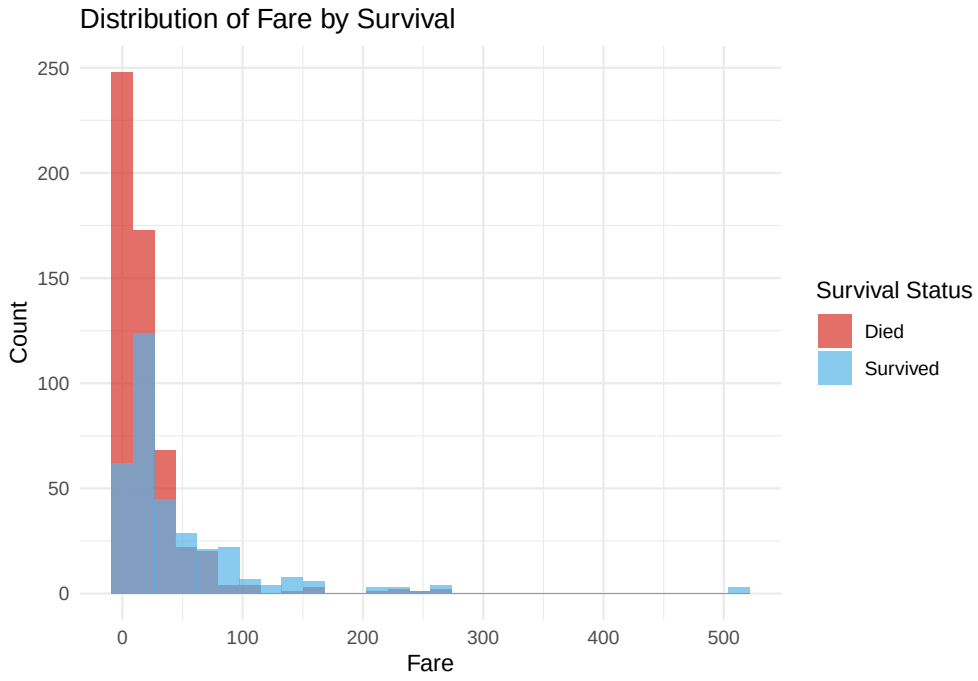
Form the summary of the data we get some early information about passengers in dataset:

- around 38% of the 891 passengers survived
- 65% were male
- the mean age is around 30 years and there are few elderly passengers and children
- 60% of the passengers traveled without any family member aboard
- fares varied significantly with few passengers paying a very high price with respect to others (mean = 32, max = 512).

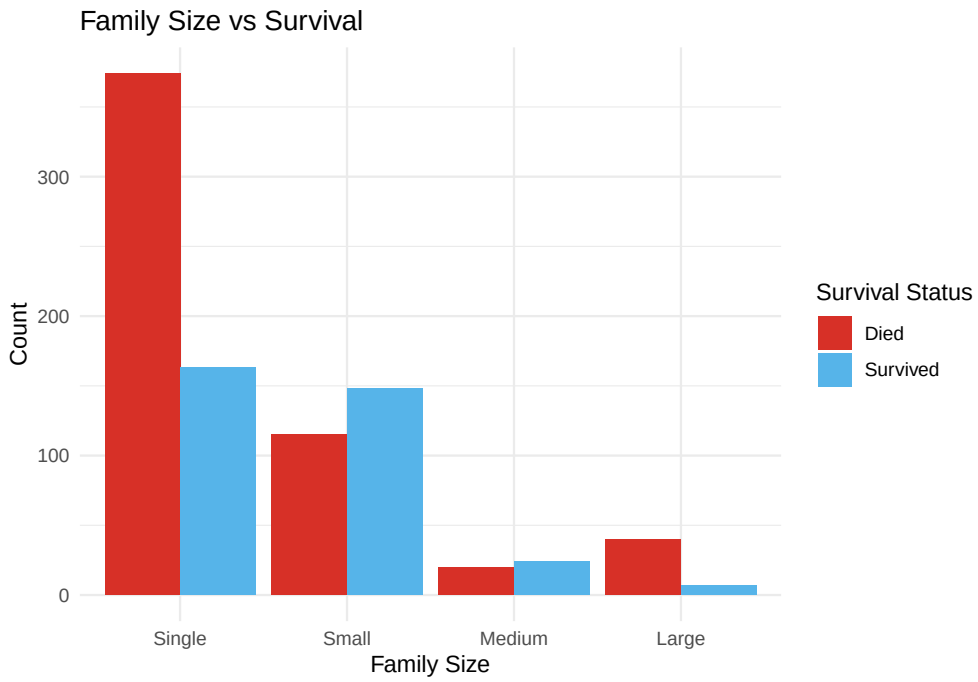
We go deeper in the exploratory analysis by looking at some plots.



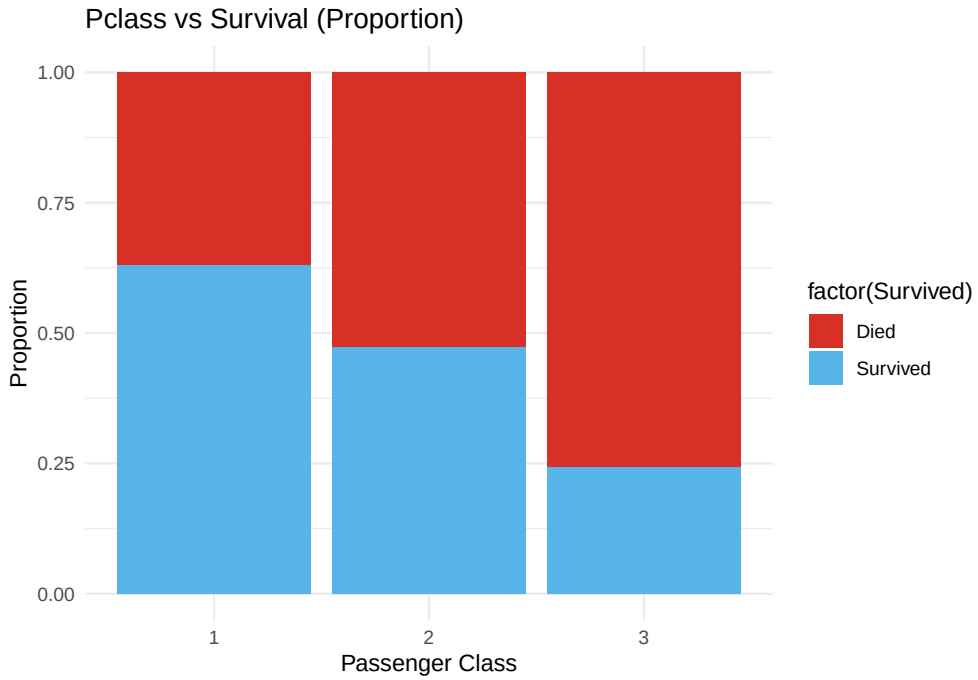
From this histogram we observe that: - Infants (age 0-1 approximately) had very high survival rate - The oldest passenger (Age = 80) survived - A large proportion of 18-25 and 40-45 year old passengers did not survive



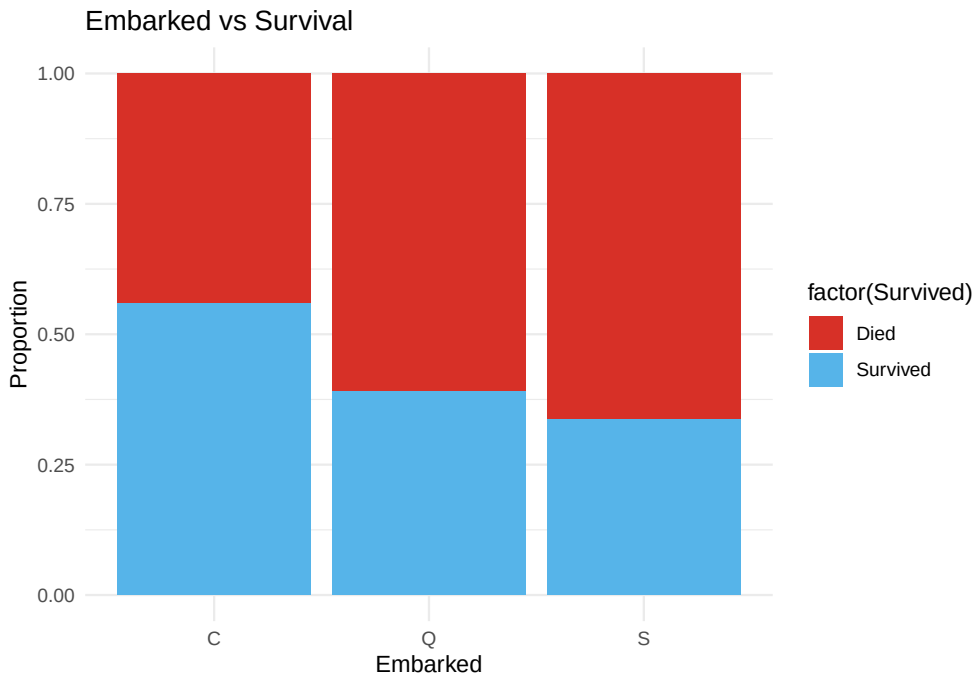
We can see that the higher the fare, the higher the proportion of survival.



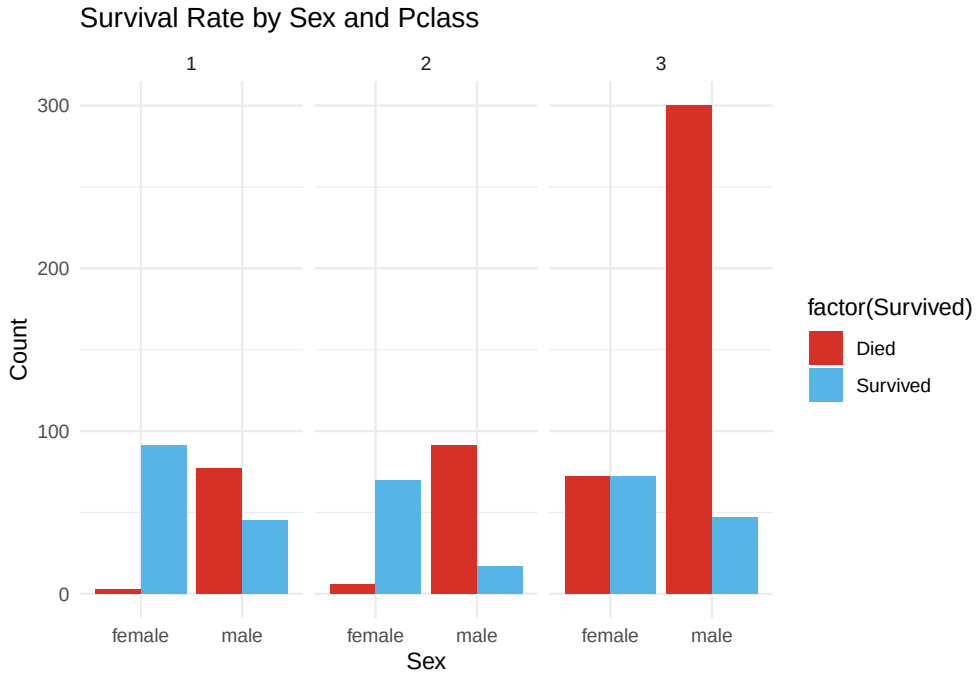
Singles and people with a large number of family components on board have a larger mortality rate with respect to people with 1 to 4 members aboard.



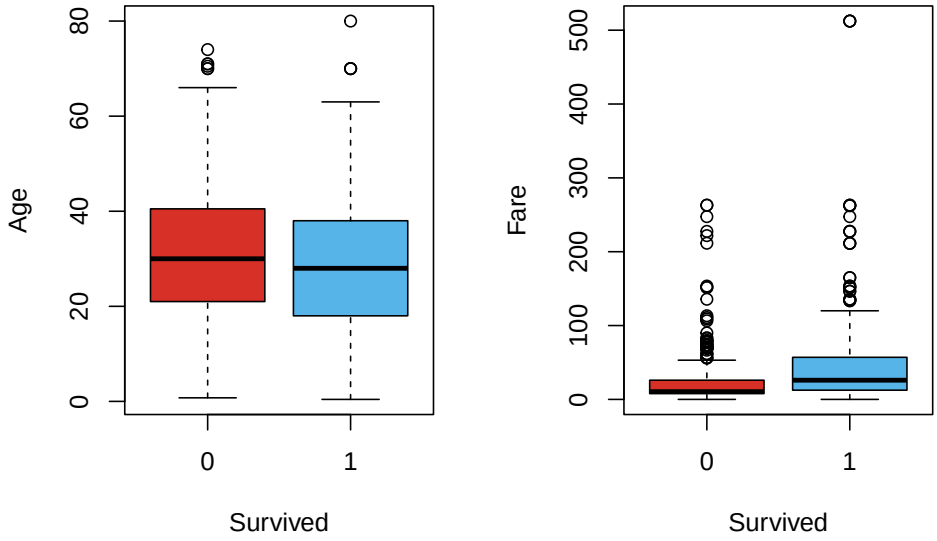
Clearly the proportion of survived is higher for higher classes. This is consistent with what we have seen for fare.



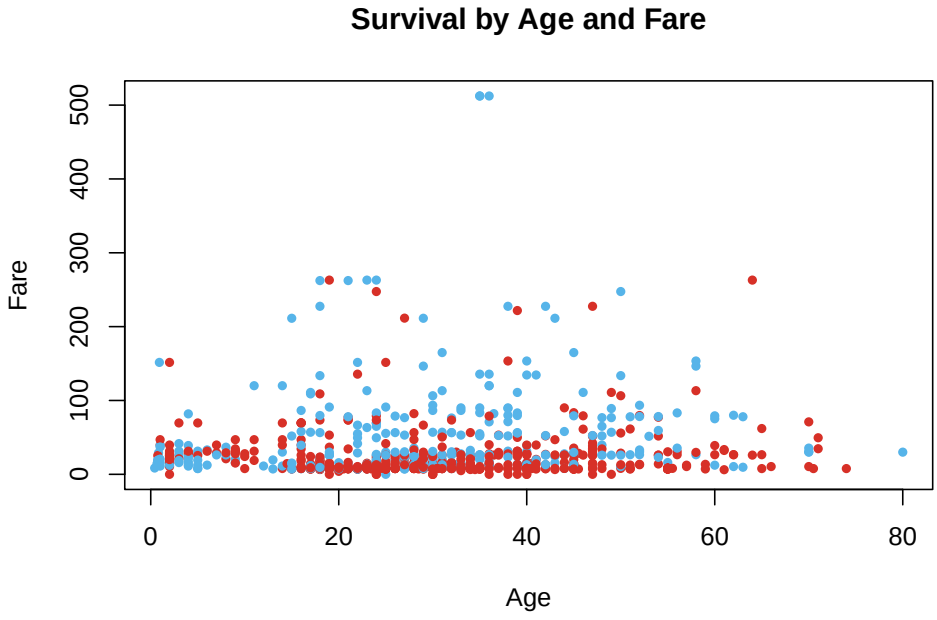
The port where passengers where embarked seems to be also correlated with the survival, indeed people embarked in Cherbourg have an higher survival rate with respect to the ones embarked in Queenstown and Southampton.



The survival rate is different for the combinations of sex and class. But the most important facts remain the same: females are more likely to survive than men and passengers in first class are more likely to survive than the others. Nevertheless we will include an interaction between sex and age because we will see that it is significant.



Fare seems to have a strongest effect than age on survived.



We check the linear correlation between the numerical covariates, which are only age and fare if we consider family grouped in categories:

```
## [1] 0.06278591
```

Age and Fare are not significantly linearly correlated.

Exploratory analysis summary:

- Women had an higher survival rate than men
- Survival rate was higher for higher classes
- Small and medium families had an higher survival rate then singles and large families
- Very young and very old passengers had an higher survuval rate than the others

Data preparation

We remove the variables SibSp and Parch and we keep Fsize. We then scale the numerical covariates.

Then we split the dataset in train set (80%) and test set (20%).

Now the data are ready to be used for implementing our models.

Logistic regression

The first model we consider is a logistic regression with Survived as responde and all the other variables as covariates (without interactions):

```
##
## Call:
## glm(formula = Survived ~ ., family = "binomial", data = train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.84009    0.38948   7.292 3.05e-13 ***
## Sexmale     -2.57790    0.22767 -11.323  < 2e-16 ***
## Pclass2     -1.00783    0.33622  -2.998  0.00272 **
## Pclass3     -2.20306    0.34593  -6.369 1.91e-10 ***
## EmbarkedQ   -0.02501    0.42705  -0.059  0.95329
## EmbarkedS   -0.46821    0.27469  -1.705  0.08828 .
## FsizeSmall   0.10923    0.23239   0.470  0.63835
## FsizeMedium -0.39340    0.45087  -0.873  0.38291
## FsizeLarge  -2.03173    0.58097  -3.497  0.00047 ***
## Age         -0.53241    0.11982  -4.443 8.85e-06 ***
## Fare         0.12592    0.13232   0.952  0.34129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 948.95  on 712  degrees of freedom
## Residual deviance: 619.80  on 702  degrees of freedom
## AIC: 641.8
##
## Number of Fisher Scoring iterations: 5
```

The coefficients estimated by the model suggest:

- The intercept estimate is significant and equal to 2.84, meaning that when all covariates are at their reference levels, the probability of survival is approximately 95%.
- Sexmale estimate is significant and indicates that being male significantly reduces the log-odds of survival compared to females. This confirms that gender is a strong predictor of survival.
- Pclass: Passengers in second class have significantly lower survival odds than first-class passengers and third-class passengers have even lower survival odds. This suggests a strong effect of socioeconomic status on survival.
- Embarked: at signifcance level 5% there is not a significant difference between the ports of embarkation.
- Fsize: large families have significantly lower survival odds, possibly due to difficulties in evacuating together. Small and Medium size families have not a significantly difference in the survival odds with respect to singles.
- Age: Older passengers have significantly lower survival odds, indicating that younger individuals had a better chance of survival.
- Fare: the effect of fare on survival is not statistically significant, likely due to its correlation with passenger class.

The model fits well but could potentially be improved by feature selection or interaction terms.

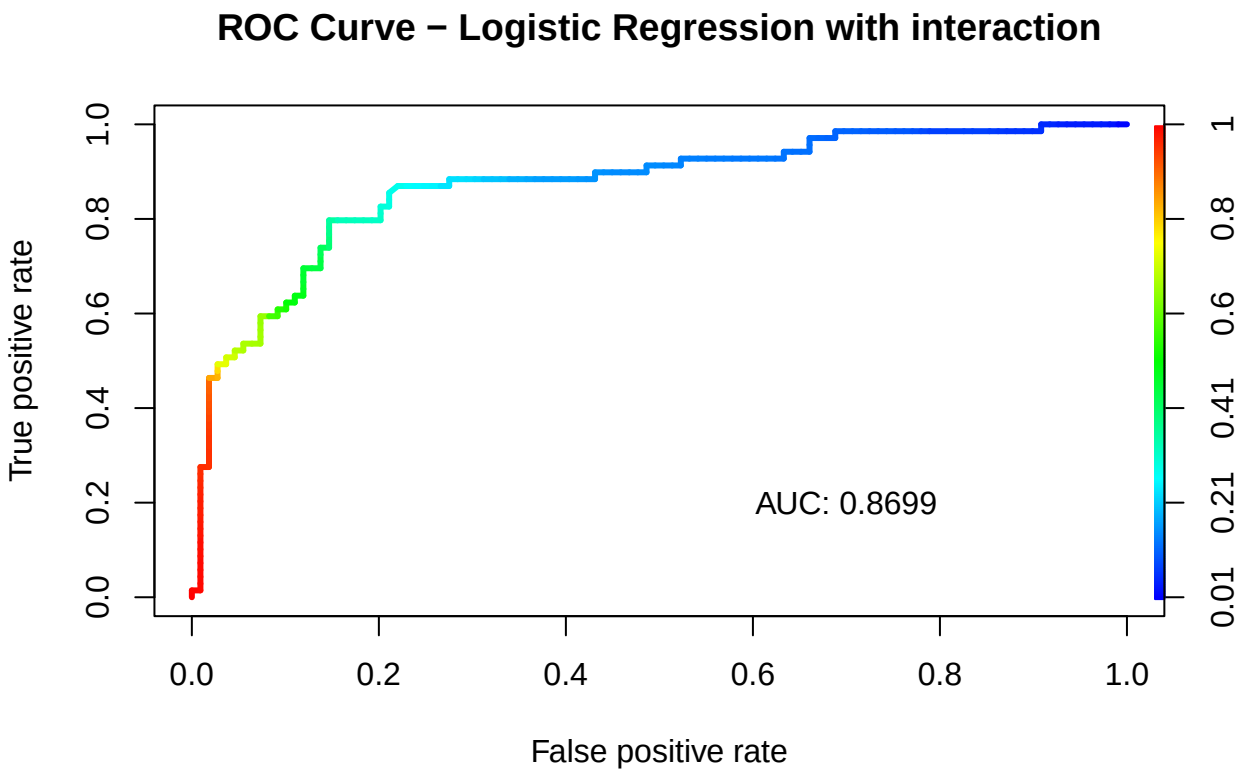
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 99 22
##          1 10 47
##
##                Accuracy : 0.8202
##                  95% CI : (0.7558, 0.8737)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 1.645e-09
##
##                   Kappa : 0.6088
##
##  Mcnemar's Test P-Value : 0.05183
##
##             Sensitivity : 0.9083
##             Specificity : 0.6812
##          Pos Pred Value : 0.8182
##          Neg Pred Value : 0.8246
##              Prevalence : 0.6124
##          Detection Rate : 0.5562
##    Detection Prevalence : 0.6798
##       Balanced Accuracy : 0.7947
##
##        'Positive' Class : 0
##
```

The model has a good overall accuracy of 82.02%, which is much better than just predicting the most frequent class (No Information Rate of 61.24%). The sensitivity is very high (91%), meaning the model is good at detecting Class 0 instances, but the specificity is lower (68%), meaning the model could improve at identifying Class 1 instances.

There is still room for improvement, so We try to add the interaction between Sex and Class

```
##
## Call:
## glm(formula = Survived ~ Age + Fare + Embarked + Fsize + Sex *
##     Pclass, family = "binomial", data = train)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.17601    0.79142   5.277 1.32e-07 ***
## Age             -0.61696    0.12923  -4.774 1.80e-06 ***
## Fare             0.06533    0.13046   0.501  0.61654
## EmbarkedQ        0.06864    0.40853   0.168  0.86658
## EmbarkedS       -0.52850    0.27942  -1.891  0.05857 .
## FsizeSmall       0.17050    0.24018   0.710  0.47777
## FsizeMedium     -0.33693    0.47972  -0.702  0.48246
## FsizeLarge      -1.93343    0.60761  -3.182  0.00146 **
## Sexmale         -3.99187    0.76013  -5.252 1.51e-07 ***
## Pclass2         -0.99398    0.92743  -1.072  0.28383
## Pclass3         -4.11526    0.81096  -5.075 3.88e-07 ***
## Sexmale:Pclass2 -0.50465    0.97560  -0.517  0.60497
## Sexmale:Pclass3  2.40666    0.80801   2.979  0.00290 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 948.95  on 712  degrees of freedom
## Residual deviance: 589.30  on 700  degrees of freedom
## AIC: 615.3
##
## Number of Fisher Scoring iterations: 6
```

The addition of the interaction term improved the model fit, indeed both residual deviance and AIC decreased. The coefficients changed as well as their significance level. The main difference is that Pclass 2 is no more significant as well as the interaction between male and class 2. On the other hand, Pclass 3 and its interaction with male are significant, meaning that for males in 3rd class, the effect of being male on survival is significantly different compared to males in 1st class.



The model performs well in separating the two classes, as indicated by the AUC of 0.8699. The curve is well above the diagonal (which represents random guessing), confirming strong predictive power.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 93 15
##          1 16 54
##
##                Accuracy : 0.8258
##                  95% CI : (0.762, 0.8785)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 5.551e-10
##
##                   Kappa : 0.6341
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8532
##             Specificity : 0.7826
##          Pos Pred Value : 0.8611
##          Neg Pred Value : 0.7714
##              Prevalence : 0.6124
##          Detection Rate : 0.5225
##    Detection Prevalence : 0.6067
```

```
##       Balanced Accuracy : 0.8179
##
##        'Positive' Class : 0
##
```

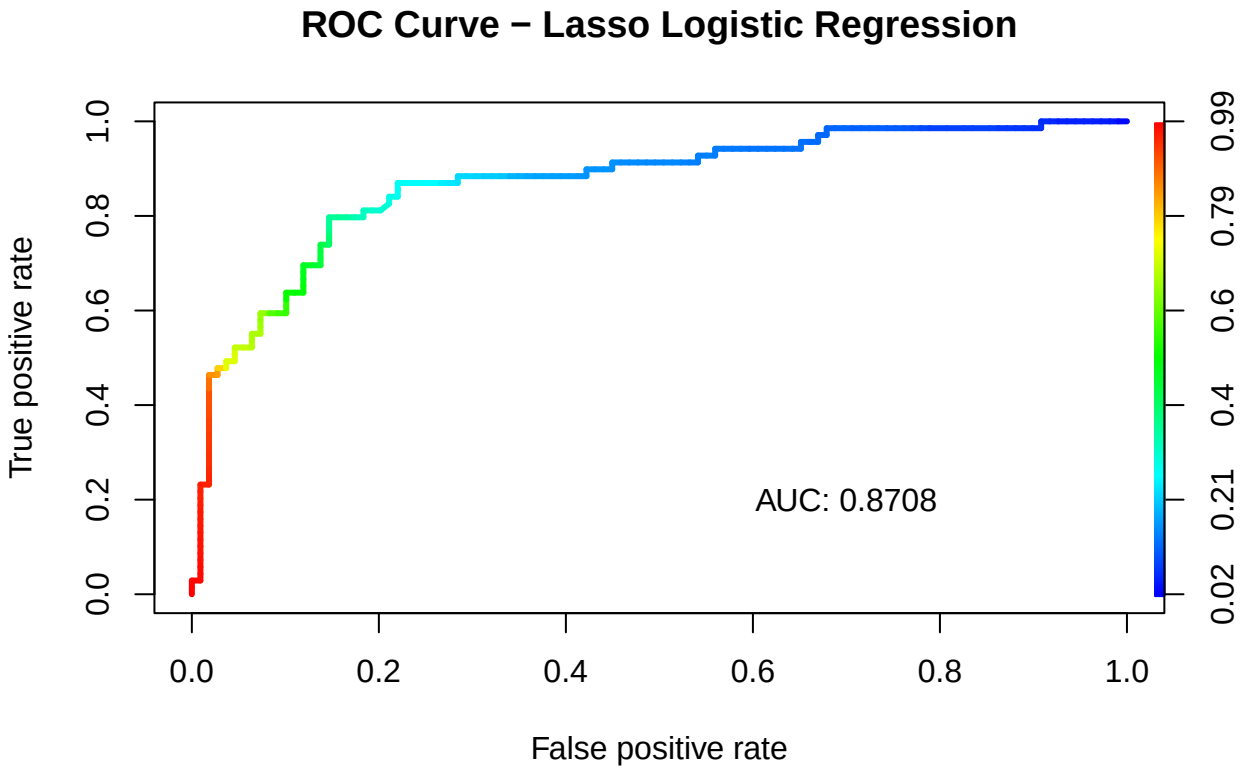
The model with interaction (0.8258) has a slightly better accuracy than the one without interaction (0.8202), indicating a marginal improvement in overall classification. The model with interaction sacrifices some sensitivity but improves specificity. It is slightly better overall.

Now we apply Lasso to the last model to see if it shrinks to zero some of the coefficients improving the model's generalization ability.

```
## [1] "Best Lambda: 0.0025"
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
## (Intercept)      3.17080357
## Age             -0.54959550
## Fare             0.09156319
## EmbarkedQ        .
## EmbarkedS       -0.52321419
## FsizeSmall       0.17035411
## FsizeMedium     -0.25584643
## FsizeLarge      -1.75964807
## Sexmale         -3.00763111
## Pclass2          .
## Pclass3         -3.00274845
## Sexmale:Pclass2 -1.40915687
## Sexmale:Pclass3  1.32039173
```

Lasso shrinks the coefficients of EmbarkedQ and Pclass2 to zero, which were not significant in the previous model, but did not shrink other coefficients that were not significant.



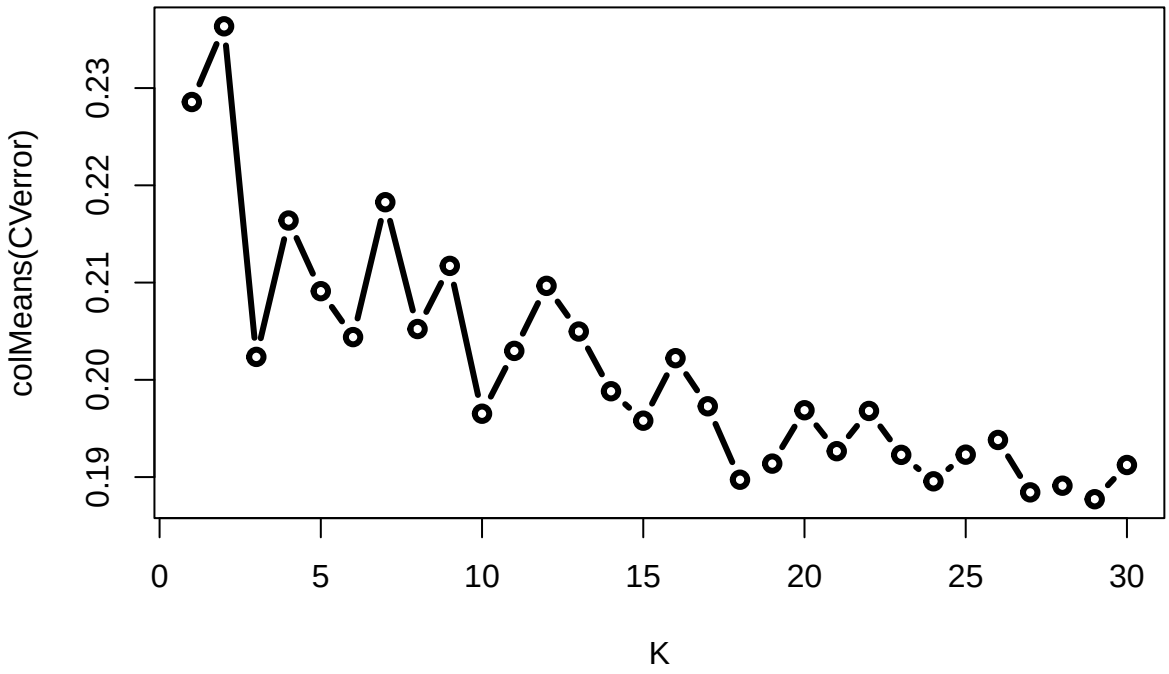

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 93 15
##          1 16 54
##
##                Accuracy : 0.8258
##                  95% CI : (0.762, 0.8785)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 5.551e-10
##
##                   Kappa : 0.6341
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8532
##             Specificity : 0.7826
##          Pos Pred Value : 0.8611
##          Neg Pred Value : 0.7714
##              Prevalence : 0.6124
##          Detection Rate : 0.5225
##    Detection Prevalence : 0.6067
##       Balanced Accuracy : 0.8179
##
##        'Positive' Class : 0
```

##

The model with Lasso has exactly the same accuracy, sensitivity and specificity, and a AUC slightly higher (but almost equal). So we have the same performance with 2 coefficients "removed", hence we would choose this as the final logistic regression model.

KNN

We use KNN with all the covariates and the interaction: categorical covariates can be used if they are properly encoded, in our case we used one-hot encoding. We choose the best value of the hyperparameter k using k-fold cross validation



The best is K = 18, so we use this value and make predictions on the test set.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 96 27
##          1 13 42
##
##                Accuracy : 0.7753
##                  95% CI : (0.7068, 0.8343)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 2.766e-06
```

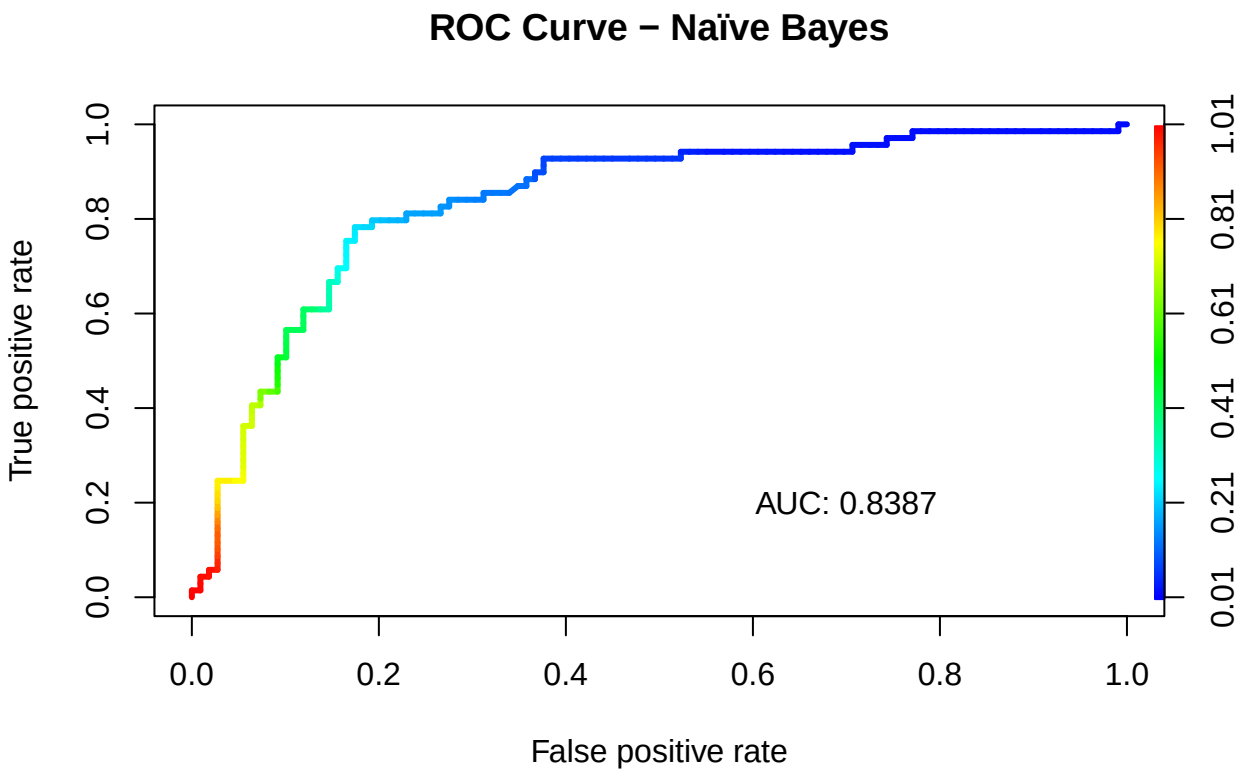
```
##
##                   Kappa : 0.5084
##
##  Mcnemar's Test P-Value : 0.03983
##
##             Sensitivity : 0.8807
##             Specificity : 0.6087
##          Pos Pred Value : 0.7805
##          Neg Pred Value : 0.7636
##              Prevalence : 0.6124
##          Detection Rate : 0.5393
##    Detection Prevalence : 0.6910
##       Balanced Accuracy : 0.7447
##
##        'Positive' Class : 0
##
```

Lasso logistic regression remains the best model so far, with higher accuracy and specificity.

Naïve Bayes

We also use a Naïve Bayes classifier, without the interaction term because this method relies on the naïve assumption that all predictors are conditionally independent given the class label. This means that it does not account for interactions between variables. The predictors are assumed to have a Gaussian prior distribution.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 99 36
##          1 10 33
##
##                Accuracy : 0.7416
##                  95% CI : (0.6707, 0.8042)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 0.0001932
##
##                   Kappa : 0.4152
##
##  Mcnemar's Test P-Value : 0.0002278
##
##             Sensitivity : 0.9083
##             Specificity : 0.4783
##          Pos Pred Value : 0.7333
##          Neg Pred Value : 0.7674
##              Prevalence : 0.6124
##          Detection Rate : 0.5562
##    Detection Prevalence : 0.7584
##       Balanced Accuracy : 0.6933
##
##        'Positive' Class : 0
##
```



Accuracy is 74% and specificity is very low (48%), so the best is still Lasso logistic regression.

Random Forest

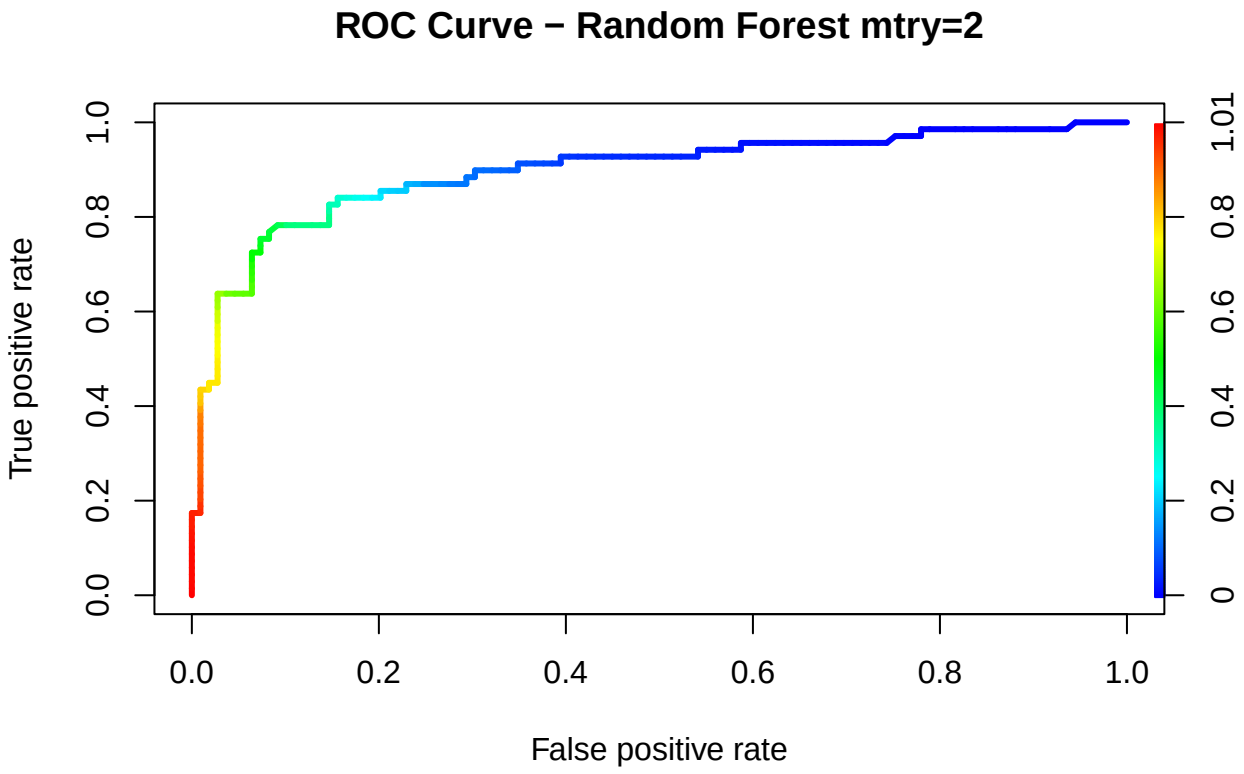
We run a cross validation for the random forest to choose the number of variables that are randomly sampled as candidates at each split.

The optimal value of mtry is 2 in terms of Out of Bag error, while the cv error is slightly better for mtry = 3. We use fit the final random forest with mtry = 2 and we increase the number of trees to 9000 because the accuracy increases and the computational time does not increase much.

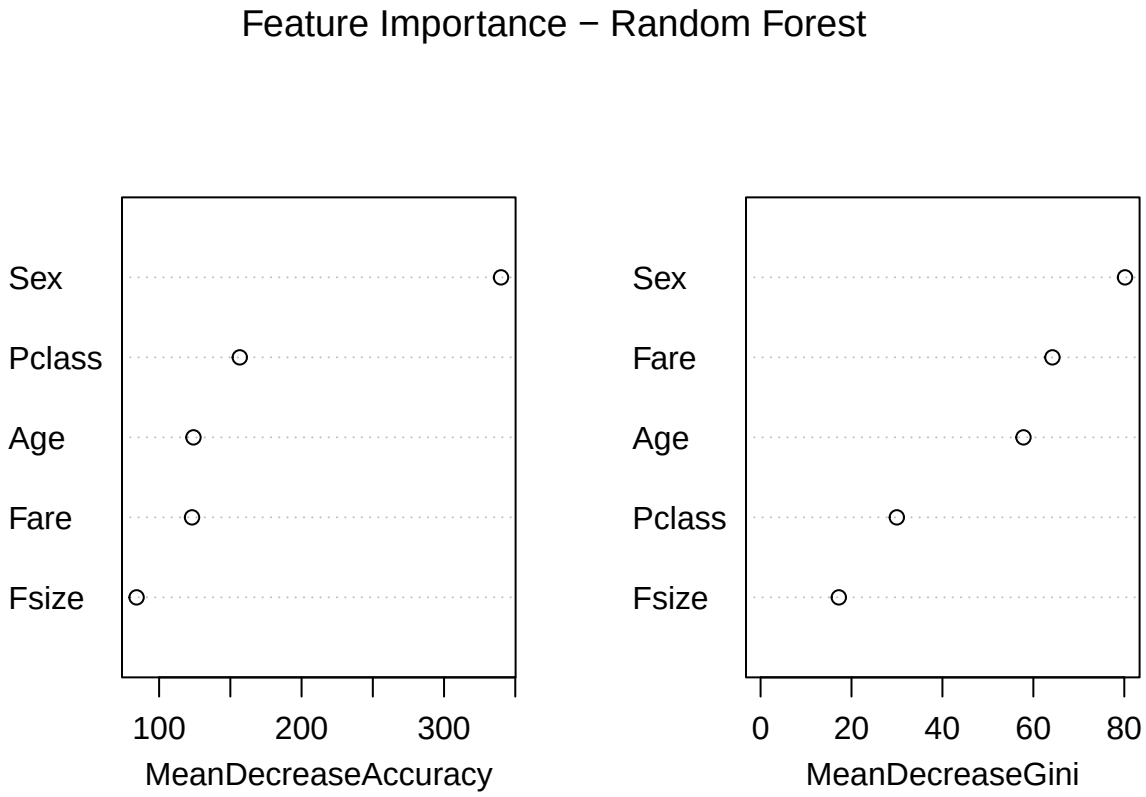
```
##
## Call:
##  randomForest(formula = Survived ~ Age + Fare + Fsize + Sex *      Pclass, data = train, mtry = 2, n
##                Type of random forest: classification
##                      Number of trees: 9000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 17.81%
## Confusion matrix:
##     0   1 class.error
## 0 402  38  0.08636364
## 1  89 184  0.32600733
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction   0   1
##          0 102  19
##          1   7  50
##
##                Accuracy : 0.8539
##                  95% CI : (0.7933, 0.9023)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 1.346e-12
##
##                   Kappa : 0.6822
##
##  Mcnemar's Test P-Value : 0.03098
##
##             Sensitivity : 0.9358
##             Specificity : 0.7246
##          Pos Pred Value : 0.8430
##          Neg Pred Value : 0.8772
##              Prevalence : 0.6124
##          Detection Rate : 0.5730
##    Detection Prevalence : 0.6798
##       Balanced Accuracy : 0.8302
##
##        'Positive' Class : 0
##
```



The accuracy is the highest so far, 85%, and sensitivity and specificity are also very good. OOB Error Rate is 17.81%, indicating strong generalization. We check the importance of the predictors with 2 metrics: the first measures how much accuracy decreases when a feature is removed and the second measures the purity improvement in decision trees due to each feature.



Sex is the most influential feature in both metrics, confirming that gender played a major role in survival. Age and Fare have high Gini importance but low Accuracy decrease, meaning they help with tree splits, but removing them may not drastically reduce accuracy. Family Size is the least important in both metrics, suggesting that having family aboard was less critical compared to gender, age, and class.

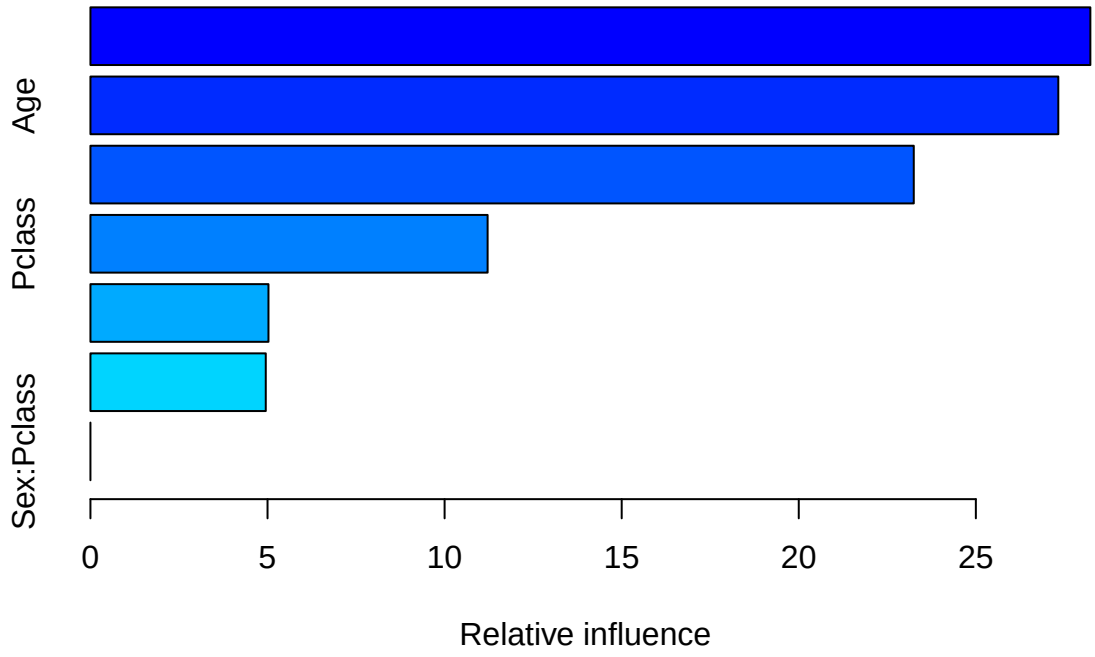
Boosting

The last method we implement is boosting with cross validation to tune the number of trees, the maximum depth of each tree and the shrinkage

```
## [1] "Best hyperparameters:"
```

```
##    Var1 Var2 Var3
## 40    5  500 0.05
```

The best hyperparametrs found with 10-fold cross-validation are ntrees = 500, interaction depth = 5, shrinkage = 0.05. We train the model with those values and evaluate its performance on the test set.




```
##                   var   rel.inf
## Fare             Fare 28.233207
## Age               Age 27.328403
## Sex               Sex 23.247456
## Pclass         Pclass 11.214287
## Embarked     Embarked  5.026403
## Fsize           Fsize  4.950245
## Sex:Pclass Sex:Pclass  0.000000
```

The relative influence of the variables in boosting is completely different from the one of random forest. Here fare is the most important while sex is in third place. Fsize remains the least influential one.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 98 19
##          1 11 50
##
##                Accuracy : 0.8315
##                  95% CI : (0.7682, 0.8833)
##     No Information Rate : 0.6124
##     P-Value [Acc > NIR] : 1.804e-10
##
##                   Kappa : 0.6373
```

```
##
##  Mcnemar's Test P-Value : 0.2012
##
##             Sensitivity : 0.8991
##             Specificity : 0.7246
##          Pos Pred Value : 0.8376
##          Neg Pred Value : 0.8197
##              Prevalence : 0.6124
##          Detection Rate : 0.5506
##    Detection Prevalence : 0.6573
##       Balanced Accuracy : 0.8119
##
##        'Positive' Class : 0
##
```

Accuracy is 83%, so random forest remains the best.

Conclusion

In this analysis, we applied multiple machine learning models to predict passenger survival on the Titanic, in particular: Logistic Regression, KNN, Naïve Bayes, and methods based on trees like Random Forest and Boosting.

These are the main findings we have achieved:

1. Feature Importance: The most influential predictors were Fare, Age, and Sex, confirming that wealth and gender played a critical role in survival chances. Passenger Class, Port of Embarkation, and Family Size had a lower impact, and the interaction term Sex:Pclass had a small contribution to the models.
2. The Random Forest model achieved the best performance with an accuracy of 85.4%, even if with a small unbalance between sensitivity (93.6%) and specificity (72.5%). This was a common behavior between all the models since the training set is a little bit unbalanced (38% survived). Logistic regression and boosting also performed well.
3. Insights: Women and wealthier passengers had higher survival probabilities. The low importance of Embarked suggests departure location had minimal impact.

The following is a summary table of the metrics of all the models used:

Table 1: Model Performance Comparison

Model	Accuracy	Sensitivity	Specificity
Logistic Regression (LASSO)	0.8258	0.8532	0.7826
KNN	0.7753	0.8807	0.6087
Naïve Bayes	0.7416	0.9083	0.4783
Random Forest	0.8539	0.9358	0.7246
Gradient Boosting	0.8315	0.8991	0.7246