Kaggle Titanic Report

C. J. Copley,*

June 18, 2015

Contents

*E-mail:charlescopley@gmail.com

List of Figures

List of Tables

1 Analysis Code

The analysis contained in this document was carried out using the *R* software packages

The R analysis code used in this report can be found on *GitHUB* at `https://github.com/chopley/kaggleCompetition`.

2 Introduction

The *Titanic* sank on the 15 April 1912. Of the total 2224 passengers on board, 1517 people lost their lives on that night.

The high mortality was largely attributed to the limited number of lifeboats on board, which were insufficient for the number of passengers. The sailors stationed at the lifeboats were in the unfortunate position of deciding who should and who should not be allowed onto the lifeboats. Their decisions were largely based on the dominant maritime norm of *'Woman and children first'*.

However, as with any social interactions, other classifications also played a part in deciding who survived the incident:

1. **Wealth**: Wealthier passengers are said to have had an easier time getting onto the lifeboats.

2. **Location**:The location (i.e. cabin) of the passenger on the ship is also likely to play a part, since the lifeboats were located on the upper deck. Passengers in the lower decks may have found it more difficult to get to the top deck during the confusion.

3. **Family Size**: Family members travelling together may have been more likely to survive since they could look after each other. Furthermore, a small family may have been easier to get to the lifeboats than a large one, and people (particularly adults) that were responsible for children may have been less likely to survive than adults without any responsibilities.

Perhaps, a larger family (or the more children in the family?) increases the likelihood of a parent perishing in the disaster? Two children of a certain physical size (likely to be strongly correlated with age) are easier for a single adult to carry, and hence may be more likely to survive. It may be that woman with husbands on board would be more likely to perish, since they would not have wanted to travel without their husband?

4. **Age**: Age of adult passengers would be a classifier. Older woman may be more likely to perish under the conditions than younger woman? Similarly there may be a threshold effect for men.

5. **Ethnicity**: Different ethnicities may have been discriminated against in the lifeboat allocation. This is not included in the data set, but could be inferred from the name.

6. **Legislation/Norms**: Given maritime norms,the captain and crew may be more likely to perish?

7. **Background**: The type of job you did in every day life may played a role? Perhaps doctors, clergy are more likely to stay on board? Or they may be more likely to be able to access the lifeboats? Common knowledge says that the band played while the ship sank. If we can classify the band members, then we know they are more likely to have perished.

8. **Unknown Information**: Some information about passengers is unknown in the data set. Those with unknown information may have been from the subset of people who perished, since their information would possibly be harder to obtain. Those who survived would have been more easy to get information from.

The rest of this report examines the data obtained from the passenger list, with the aim of exploring the issues and features that affected the probability of a passenger surviving. The first part of the report explores the data, evaluating likely variables that may aid in predicting passenger outcomes. In the final sections, the report looks at different classification algorithms that can be used to make robust predictions, including:

- Recursive Partioning
- Support Vector Machine
- Neural Network Solutions
- Random Forest
- Generalized Boosted Regression

3 Description of Data

The data are divided as follows:

```
Training Data: 891 passengers
PassengerId, Survived, Pclass, Name, Sex, Age,
SibSp, Parch, Ticket, Fare, Cabin, Embarked

Test Data: 418 passengers
PassengerId, Pclass, Name, Sex, Age, SibSp,
Parch, Ticket, Fare, Cabin, Embarked

VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)
```

4 Overall Survival Rate

From the aggregate training data in Figure 1, we see that of the 891 passengers, nearly 549 (62%) perished in the accident.

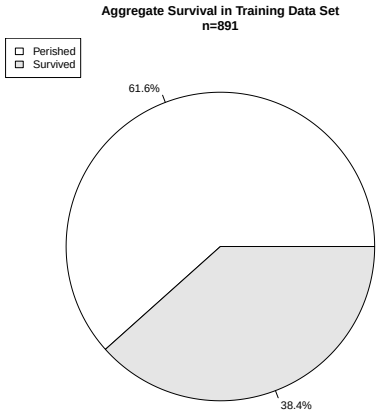
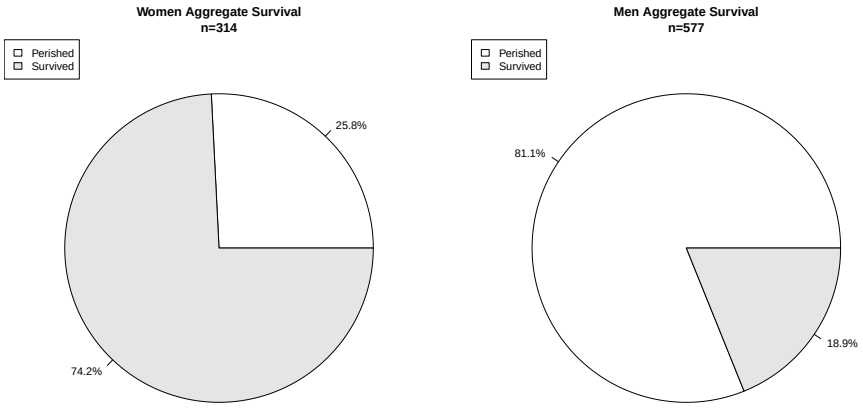


Figure 1: The aggregate survival rate of all passengers onboard the Titanic. Of the 891 passengers nearly 62% perished in the accident.

5 Gender Effects

From *'Woman and children first'*, we expect that women should have a better chance of survival than men. This is confirmed by the data in Figure 2 where nearly 74% of women survived the accident, while only 19% of men survived the accident. In fact, the effect is so strong that a fairly good prediction of survival could be made using only the gender of the passenger.



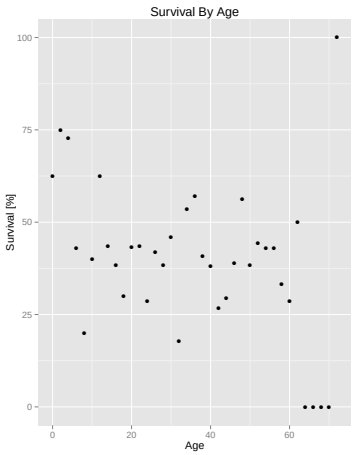
(a) Woman Survival

(b) Men Survival

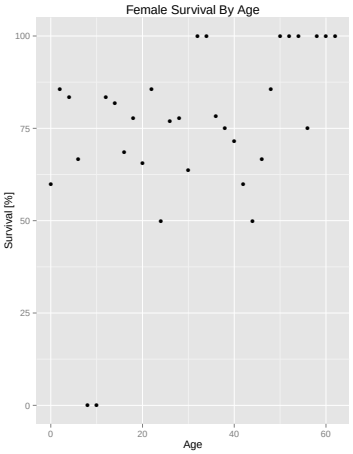
Figure 2: The survival rates of the woman and men aboard the *Titanic* respectively. A woman was much more likely (nearly 74%) to survive than a man (only about 19%)

6 Age Effects

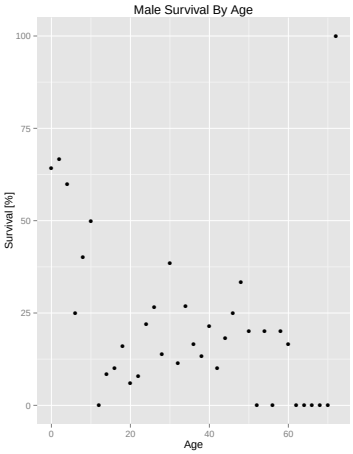
I have presented the difference in survival rates between men and women, however this does not explore the survival rates of children. In Figure 3 we see that children under the age of 10 years old, had much better chances of surviving the incident. The effect is particularly noticeable with boys, as is expected.



(a) Survival by age without distinguishing passenger gender.



(b) Survival by age of women.



(c) Survival by age of men.

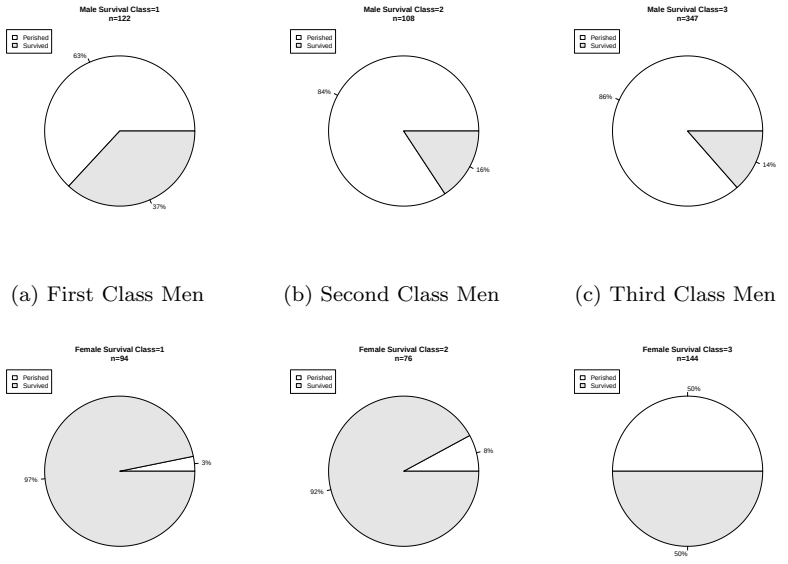
Figure 3: Survival by Age and gender

7 Social Class Effects

This section presents possible correlations between variables that might typically indicate a passengers social class or wealth, and the passengers survival probability.

7.1 Ticket Class

A strong indicator of passenger social status is the class of the ticket purchased. If we look at the breakdown of survival rates (Figure 4), we immediately notice the difference in chance of survival between the different class tickets. For both men and women, passengers travelling first class are much more likely to survive than those travelling in second and third class.



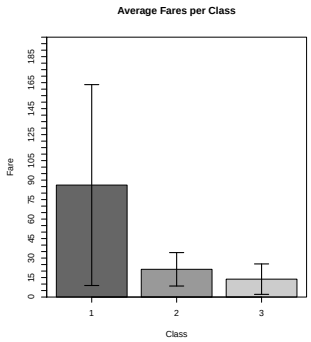
(a) First Class Men (b) Second Class Men (c) Third Class Men

(d) First Class Woman (e) Second Class Woman (f) Third Class Woman

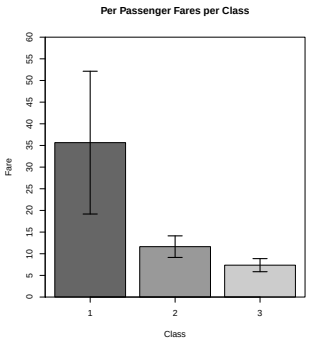
Figure 4: Survival rates by Passenger Class.

7.2 Ticket Price

The fare paid for the ticket can be used as a proxy for the social class of the passenger. The dataset however, shows a very wide range of ticket prices paid by first class passengers in particular (see Figure 5a) which seems suspicious. Looking more closely at the data though, we see that in many cases a number of people travelled together on the same ticket and the Fare represents the price paid for *all the passengers*. A better representation of the ticket price, is the *per passenger* price. This can be calculated by normalising the Fare by the number of passengers travelling on a ticket. This is given in Figure 5b.



(a) Price of each ticket by class. This is the total Fare paid for a ticket, so is the price paid for all passengers travelling on a given ticket.



(b) Price per passenger paid in each class. This is the total ticket Fare normalised by the number of passengers travelling on the ticket.

Figure 5: The ticket fares paid for different class tickets onboard the *Titanic*

To capture the potential effect of the fare on the survival rates, we bin the survival rate by the *per passenger* ticket price. The strong linear correlation between *per passenger* ticket price and survival chances is shown in Figure 6. This suggests that those who paid higher individual ticket prices, were more likely to survive. The average ticket price of First class tickets was $35, so this diagram suggests that even within First Class passengers, those who purchased more expensive tickets had higher probability of survival.

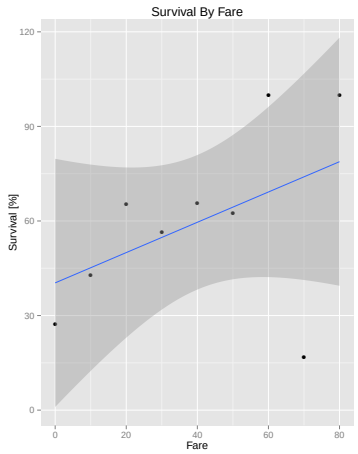


Figure 6: The survival rate of the different *per passenger* ticket prices onboard the *Titanic*. The outlier at $70 is interesting and could be investigated further by looking at the number of tickets, gender etc. this data point actually represents.

7.3 Passenger Title

The title of the passenger encodes a great deal of information about their social standing, and also their age. I have created a data feature based on the passenger title, and divided these into bins that heuristically appear to have different probabilities of survival. Interestingly, not a single Reverend on board the *Titanic* survived the incident. The captain also went down with the ship.

We remind ourselves of the gender division of survival probability given in Table 1.

Gender	Survival Rate
Female	74.2%
Male	18.9%

Table 1: Survival rates of Male and Female Passengers.

Group	Title
Good Male	Major, Sir, Dr, Col
Child Male	Master
Male	Jonkheer, Mr, Don
vBadMale	Rev, Capt
vGoodFemale	the Countess, MMe, Mlle, Lady, Ms
BadFemale	Miss

Table 2: People with different titles appear to have different chances of survival.

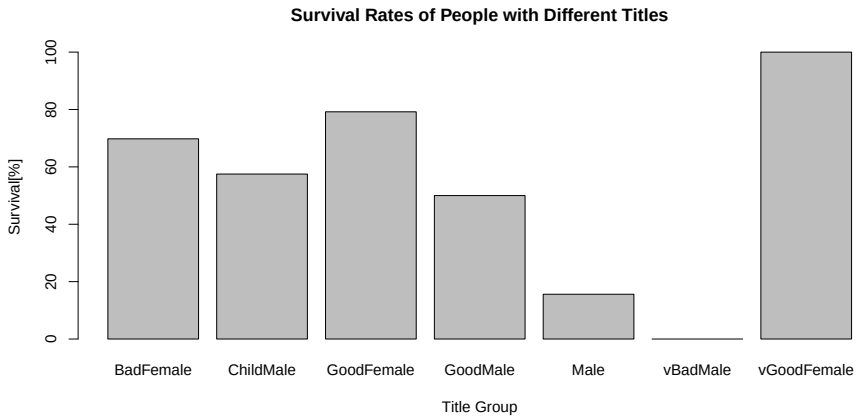


Figure 7: The survival rate of people with different titles onboard the *Titanic*. See Table 2 for the groupings.

7.4 Cabin Location

The cabin location is likely to play some role in the survival expectations, since we would expect higher survival rates from passengers that were able to easily

reach the lifeboats. Unfortunately, many of the passengers do not have cabin information (see Figure 8). However, there may be another effect at play here. Perhaps the passengers that survived, are more likely to have their cabin information recorded, since they were able to retell their stories, and update the information after the incident.

I expect that the cabin location will be a useful feature in the data set.

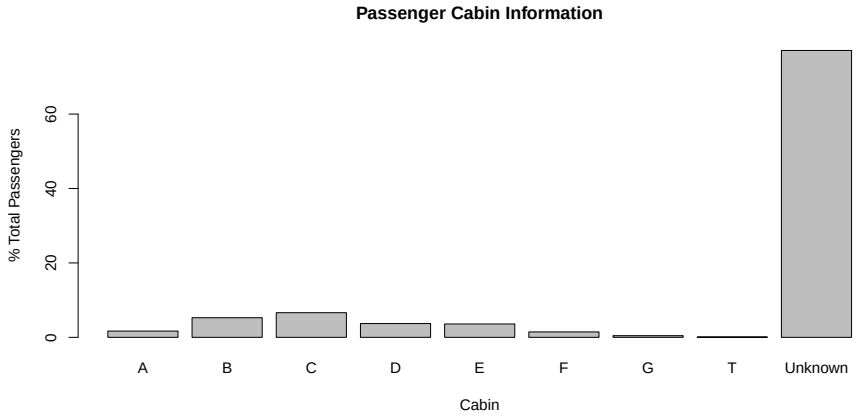


Figure 8: Distribution of Passengers with known Cabin locations. Cabin locations of passengers are not well documented.

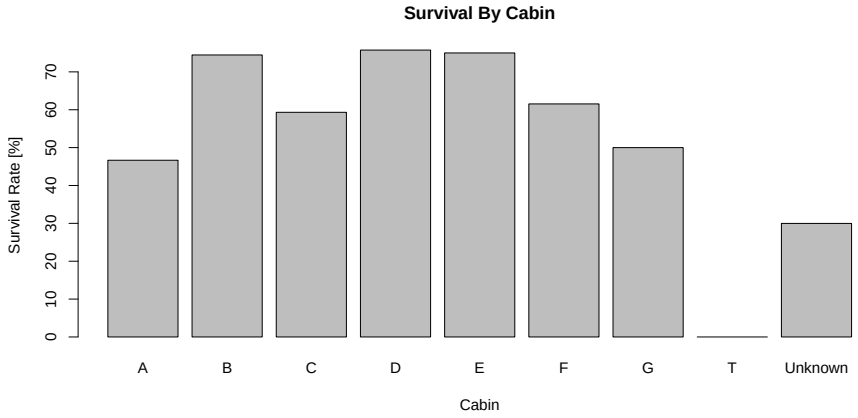


Figure 9: Survival of Passengers with known Cabin locations. The rates of survival amongst those with known cabin locations is slightly higher than average. This may support the notion that cabin location information may have been collected from survivors.

8 Family Size Effects

Families, or couples, travelling together may have had better chances of survival, since they were able to operate as a unit, perhaps resulting in greater persuasive power when trying to get onto the lifeboats. On the other hand, large families are more likely to have a member missing during the collision. This may have led to other family members panicking, delaying their arrival at the lifeboats. In this section I separate the passengers into romantic couples (i.e. a woman and man travelling on the same ticket), and also into group sizes (i.e. the number of passengers travelling on a single ticket, and sharing a surname). More feature selection on this point, may improve the classification somewhat.

8.1 Couples

Romantic couples (see Figure 10) appear to have better probability of survival than non-romantic couples.

Figure 10: Travelling together in a couple increased probability of survival

8.2 Family Units

A family size of four (see Figure 11)seems to be the optimal size for survival chances.

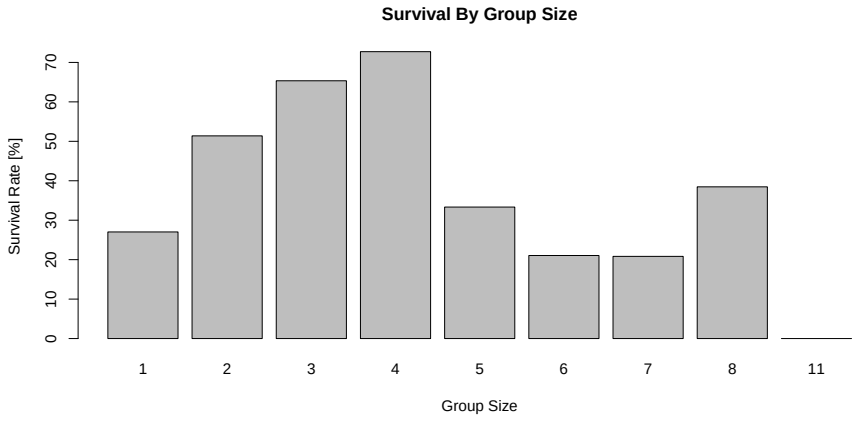


Figure 11: Group Size vs Survival Chances. Groups of four people on a ticket were most likely to survive, however this may be strongly correlated with class, since most of the multi-passenger tickets appear to be First Class tickets.

9 Ethnicity

The ethnicity of people is often a predictor in cases of social interaction. This may have played a role onboard the *Titanic*, particularly since the decision as to who would be allowed onto the lifeboats was made by sailors, many of whom may have had similar ethnicities. Unfortunately, the ethnicity of the passengers is not directly captured, however it can be inferred by correlating the names of the passengers against a census list of names that includes ethnicity.

The U.S Census of 2000 (United States Census Bureau 2000) [1] has a list of common U.S. surnames, together with ethnicity percentages. I have used this to arrive at the following ethnicity distribution onboard the *Titanic*. The ethnic labels are taken directly from that used by the U.S. Census.

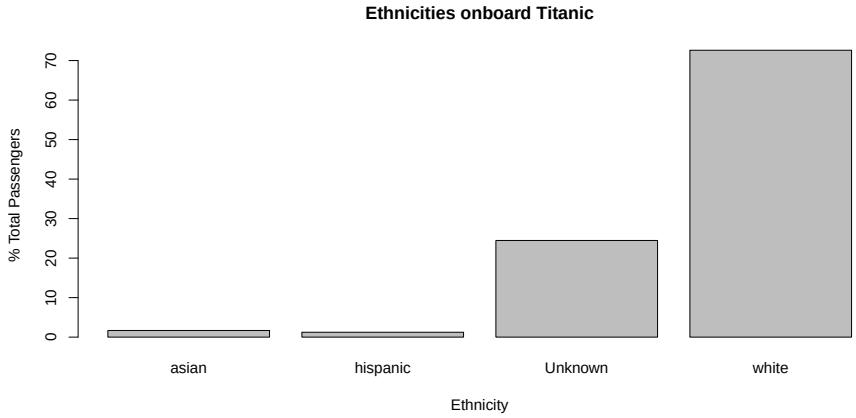


Figure 12: Ethnicity of passengers onboard the Titanic using names and ethncities taken from the US Census 2000. http://www.census.gov/topics/population/genealogy/data/2000_surnames.html. There are 25% of Passengers where the ethnicity was not classified, since the surnames were not present in the US Census database.

[1] http://www.census.gov/topics/population/genealogy/data/2000_surnames.html

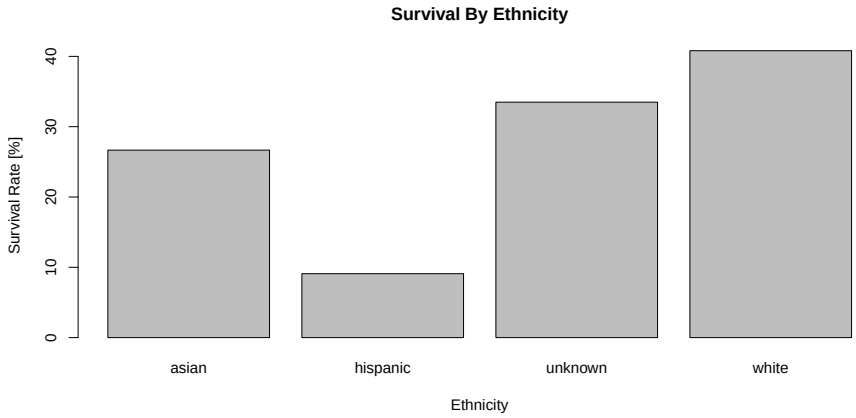


Figure 13: Survival by Ethnicity. Low counts of Asian and Hispanic travellers make the values unreliable, however a more extensive name database, could reveal patterns within European sub-populations. I was not able to easily find such a database though.

10 Prediction of Survivors using Data Features

I have used five different alogrithms to classify the data. Each has been trained using cross-validation in order to improve robust classification and to avoid over-fitting.

10.1 Data Cleaning and Feature Engineering

Before running the models I have done the following cleaning operations:

1. **Missing Passenger Age**:Where passenger ages aren't available, I used the average age of other passenger with the same title.
2. **Cabin Information**:Where available I extracted the cabin information of the passengers.
3. **Title extraction**: Extracted the title from the name.
4. **Ethnicity**: Best-guess of ethnicity using US Census 2000 data.
5. **Fare**: Corrected for the fact that the fare given in the data is price paid for all passengers travelling on the same ticket

Using the cleaned data I created the following additional features for the dataset:

1. **EthnicFeat**: Ethnicity of the passenger (Asian, Hispanic, Unknown, White) using passenger surname and correlating with the US Census 2000 database
2. **GroupFeat**: Number of passengers travelling on the same ticket.
3. **MFCoupleFeat**: Is the passenger travelling with only their spouse/partner?
4. **FamilyFeat**: Is the passenger travelling with a family (i.e. are the other surnames on the ticket the same?)?
5. **TitleFeat**: The title of the passenger.
6. **CabinFeat**: Where available, extract which deck the passenger was on (A,B,C,D,E,F,G,T,Unknown).
7. **PriceFeat**: Binned data of the passenger fare prices.
8. **FarePassenger**: Price paid for each individual passenger.
9. **AgeFeat**: Binned passenger age.

I also used the following variables from the original data set as predictors:

1. **Pclass**: The class the passenger was travelling
2. **Age**: Age of the passenger
3. **Embarked**:Which port did the passenger embark from?

10.2 Recursive Partioning (RPART)

I tuned the recursive partitioning tree as following:

```
Define the model used for fitting
formula <-Survived ~ Pclass + Sex + FarePassenger+ Age + Embarked
    + EthnicFeat + TitleFeat + GroupFeat + MFCoupleFeat +
    FamilyFeat + CabinFeat

Tune the hyper-parameters using a grid search-
tunedRpart<-tune.rpart(formula, data=as.data.frame(trainFeat),
    minsplit = c(25,50,75,100),minbucket = c(10,20,30,40,50), cp =
    c(0,0.1,0.2,0.3,0.4,0.5), maxcompete = c(0,1), maxsurrogate = c
    (0,1), usesurrogate = c(0,1), xval = c(0,1,2,3,4))

Fit the model to the training set
fit <- rpart(formula, data=as.data.frame(trainFeat),
     method="class",control=rpart.control(minsplit=100,
     cp=0,minbucket=10,maxcompete=0,maxsurrogate=0,usesurrogate=0,
         xval=0))

Calculate accuracy on the training set
Prediction <- predict(fit, as.data.frame(trainFeat), type = "class"
    )
results.matrix <- confusionMatrix(Prediction, trainFeat$Survived)
accuracyRpart<-results.matrix$overall[1]

Make a prediction using the test set
PredictionRpart <- predict(fit, as.data.frame(testFeat), type = "
    class")
submit <- data.frame(PassengerID = testFeat$PassengerId, Survived =
     (PredictionRpart))
write.csv(submit, file = "submitCJCRpart.csv", row.names = FALSE)
```

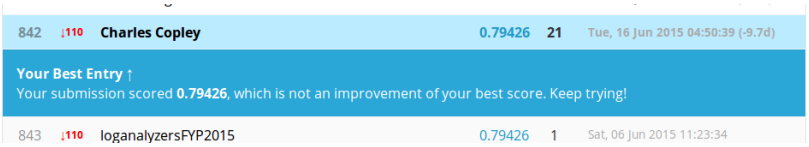


Figure 14: Recursive partitioning result on test data.

10.3 Support Vector Machine (SVM)

```
Define the model used for fitting
formula <-Survived ~ Pclass + Sex + FarePassenger+ Age + Embarked
    + EthnicFeat + TitleFeat + GroupFeat + MFCoupleFeat +
    FamilyFeat + CabinFeat

Tune the hyper-parameters using a grid search-
tuned <- tune.svm(formula, data=trainFeat, gamma = 10^(-3:3),
        cost = 10^(-2:4))

Fit the model to the training set
fitSVM <- svm(formula, data = as.data.frame(trainFeat), type="C-
    classification", kernel="radial", probability=T, gamma=0.1,
    cost=1)
```

Calculate accuracy on the training set

```
testFeat$Survived<-NULL
PredictionSVM <- predict(fitSVM, data=as.data.frame(trainFeat),
    type="C-classification")
results.matrix <- confusionMatrix((PredictionSVM), trainFeat$
    Survived)
accuracySVM<-results.matrix$overall[1]
```

Make a prediction using the test set

```
PredictionSVM2 <- predict(fitSVM, newdata=as.data.frame(testFeat),
    type="C-classification")
submit <- data.frame(PassengerID = testFeat$PassengerId, Survived =
     (PredictionSVM2))
write.csv(submit, file = "submitCJCSVM.csv", row.names = FALSE)
```

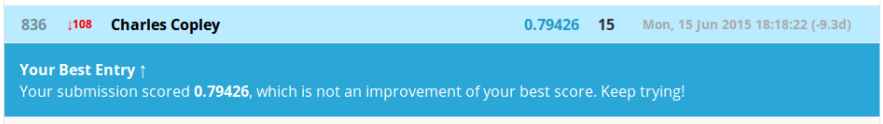


Figure 15: Support Vector Machine result on test data.

10.4 Neural Network Solutions (NNET)

Define the model used for fitting

```
formula <-Survived ~ Pclass + Sex + FarePassenger+ Age + Embarked
    + EthnicFeat + TitleFeat + GroupFeat + MFCoupleFeat +
    FamilyFeat + CabinFeat
```

Tune the hyper-parameters using a grid search-

```
tuned <- tune.nnet(formula, data=trainFeat, size=c(5,10,15,30),
    decay=c(0,0.005,0.010),MaxNWts= 20000)
```

Fit the model to the training set

```
fitNnet <- svm(formula, data = as.data.frame(trainFeat), type="C-
    classification", size=5, decay=0.005)
```

Calculate accuracy on the training set

```
PredictionNnet <- predict(fitNnet, data=as.data.frame(trainFeat),
    type="C-classification")
results.matrix <- confusionMatrix((PredictionNnet), trainFeat$
    Survived)
accuracyNnet<-results.matrix$overall[1]
```

Make a prediction using the test set

```
PredictionNnet <- predict(fitNnet, newdata=as.data.frame(testFeat),
     type="C-classification")
submit <- data.frame(PassengerID = testFeat$PassengerId, Survived =
     (PredictionNnet))
write.csv(submit, file = "submitCJCNNet.csv", row.names = FALSE)
```

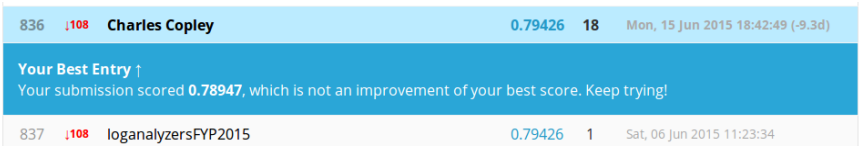


Figure 16: Neural network result on test data.

10.5 Random Forest (FOREST)

Define the model used for fitting

```
formula <-Survived ~ Pclass + Sex + FarePassenger+ Age + Embarked
    + EthnicFeat + TitleFeat + GroupFeat + MFCoupleFeat +
    FamilyFeat + CabinFeat
```

We need to define the output as a factor for FOREST

```
trainFeatForest<-trainFeat
testFeatForest<-testFeat
trainFeatForest$Survived<-as.factor(trainFeatForest$Survived)
testFeatForest$Survived<-as.factor(testFeatForest$Survived)
```

Tune the hyper-parameters using a grid search-

```
tuned <- tune.randomForest(formula, data=trainFeat, ntree=c
    (50,500,5000,50000))
```

Calculate accuracy on the training set

```
fitForest<-randomForest(formula, data=trainFeatForest, nTree=20000)
PredictionForest <- predict(fitForest, as.data.frame(
    trainFeatForest))
results.matrix <- confusionMatrix((PredictionForest),
    trainFeatForest$Survived)
accuracyForest<-results.matrix$overall[1]
```

Make a prediction using the test set

```
PredictionForest <- predict(fitForest, newdata=as.data.frame(
    testFeatForest))
submit <- data.frame(PassengerID = testFeat$PassengerId, Survived =
     (PredictionForest))
write.csv(submit, file = "submitCJCforest.csv", row.names = FALSE)
```

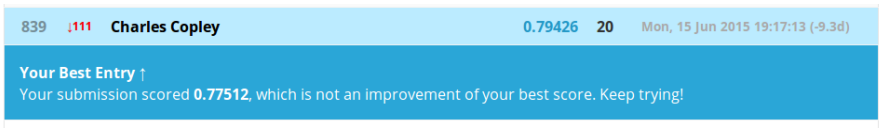


Figure 17: Random forest result on test data.

10.6 Generalized Boosted Regression (GBM)

Define the model used for fitting

```
formula <-Survived ~ Pclass + Sex + FarePassenger+ Age + Embarked
    + EthnicFeat + TitleFeat + GroupFeat + MFCoupleFeat +
    FamilyFeat + CabinFeat
```

Tune the hyper-parameters using a grid search-

```
fitControl <- trainControl(method = "repeatedcv",number = 10,
    repeats = 10)
gbmGrid <- expand.grid(interaction.depth = c(1, 5, 9), n.trees =
    (1:30)*50, shrinkage = 0.1, n.minobsinnode = 10)
gbmFit2 <- train(formula, data = as.data.frame(trainFeat), method =
     "gbm", trControl = fitControl, verbose = FALSE, tuneGrid =
    gbmGrid)
```

Calculate accuracy on the training set

```
fitBoost<-gbm(formula,data= as.data.frame(trainFeat), n.trees=50,
```

```
        interaction.depth=9, shrinkage=0.1,n.minobsinnode=10,
            distribution="gaussian")
PredictionBoost <- predict(fitBoost, as.data.frame(trainFeat),n.
    trees=50)
results.matrix <- confusionMatrix(round(PredictionBoost), trainFeat
    $Survived)
accuracyBoost<-results.matrix$overall[1]

Make a prediction using the test set
PredictionBoost <- predict(fitBoost, newdata=as.data.frame(testFeat
    ),n.trees=50,interaction.depth=9, shrinkage=0.1,n.minobsinnode
    =10,distribution="gaussian")
submit <- data.frame(PassengerID = testFeat$PassengerId, Survived =
     round(PredictionBoost))
write.csv(submit, file = "submitCJCgbm.csv", row.names = FALSE)
```

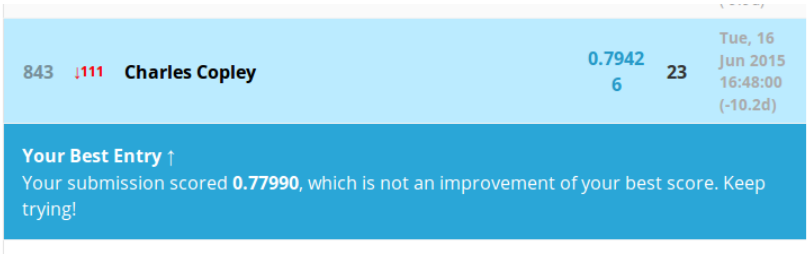


Figure 18: Generalized Boost Regression result on test data.

10.7 Comparison

Algorithm	Training Accuracy	Test Accuracy
RPART	0.851	0.794
SVM	0.844	0.794
NNET	0.834	0.789
FOREST	0.933	0.775
GBM	0.891	0.780

We see that the RPART and SVM algorithms work fairly well. The Neural network works well on the test set, given the training set accuracy. The Random Forest and GBM seem to suffer a little from over-fitting to the training set. This could probably be improved with a little work on the parameter settings.

11 Summary

This document works through the process of fitting a predictive model using classification algorithms. It explains simple feature engineering, as well as the importance of understanding of the data set (see the Fare correction required in Figure 5, and careful cleaning of the data to extract maximum information. It would be possible to do more with this data, however that is beyond the scope of this document.

The cleaned and feature-enriched data, were analysed using five different algorithms. The results of the algorithms on the *Kaggle* leaderboard are shown, and a comparison table of the different algorithms is given.

The code for the analysis was written in *R* and is available on *GITHub* at https://github.com/chopley/kaggleCompetition

References

Kuhn, M. & Johnson, K. (2013), *Applied Predictive Modelling*, Springer.

United States Census Bureau (2000), 'United States Census 2000'.