Titantic Survival

Alex Thom

03/12/2019

Introduction

The Titanic was a major catastrophe in 1912 with an estimated 832 passengers and 685 crew members perishing in the disaster. This review of machine learning techniques looks at the best methodologies for predicitng whether a passenger would have survied or died. In order to accomplish this task there is a dataset provided below:

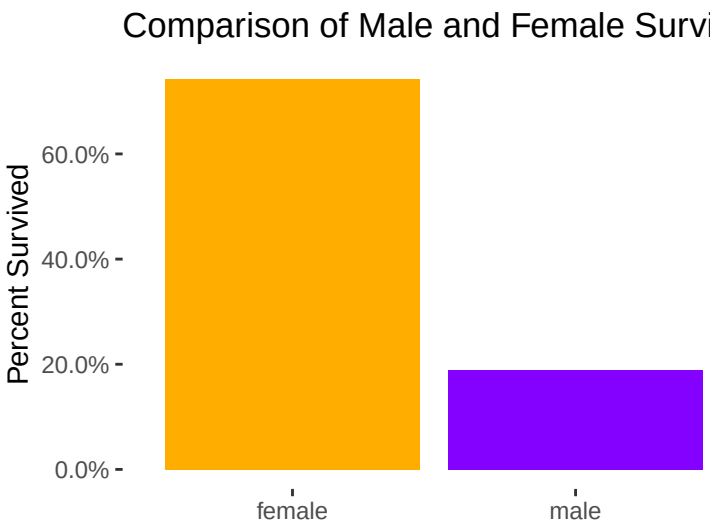
```
str(train)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

The variable that will be predicted is the Survived variable. Although it is currently an integer, it is really a factor as 0 is died and 1 is survied. This will effect the type of model that is used as the target variable being a facotr means this is a classification problem. The missing data is summarised below

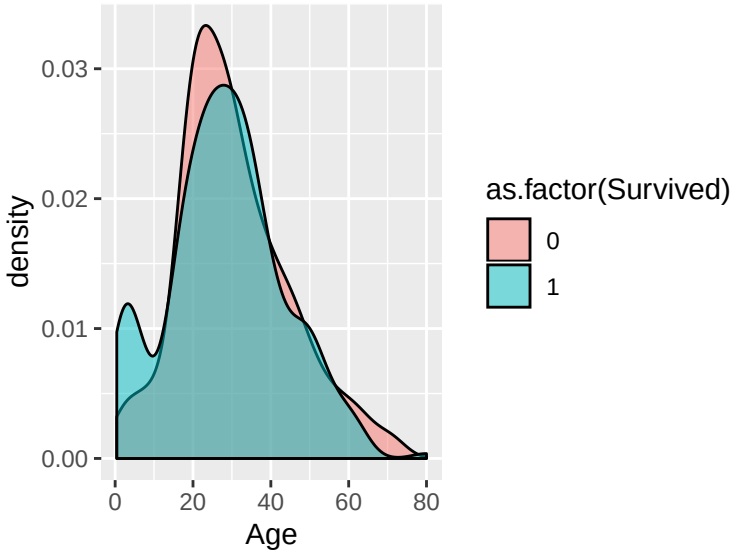
```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0         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

In the training set it looks like there is a lot of missing data for the age variable. This could be something that could be improved on in feature engineering part of this project. Also cabine has a lot of missing data which might be harder to develope into a variabel.



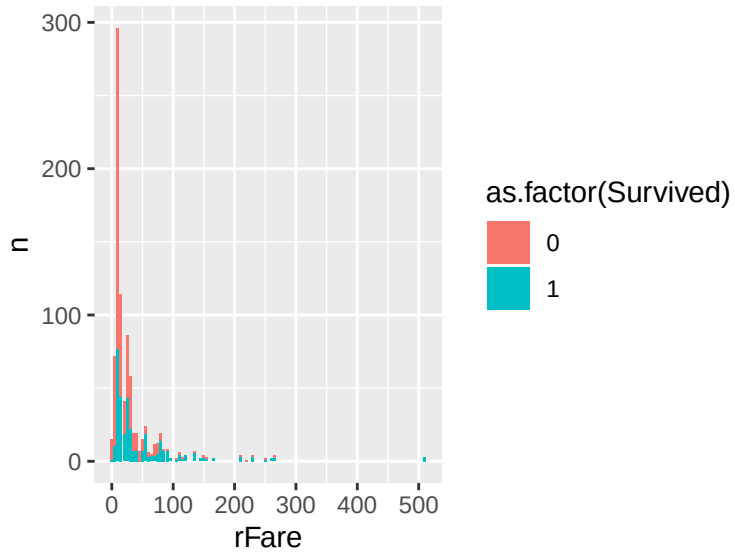
Its clear that females had a significant higher chance of survivl and therefore the Sex variable will have signifancat influneces in the models predictions.

```
## Warning: Removed 177 rows containing non-finite values (stat_density).
```



COmparing how ages effects survival the major feature visible is at the younger ages. Above 10 years old the rates of survival are pretty similar. Howerver, less then 10 years old it looks like there are significantly more likelly to be able to survive. This variable has some missing data however it looks like it has interesting features so will have to create a method in order to geat more details from this feature.

```
ggplot(vis2, aes(x = rFare, y = n, fill = as.factor(Survived))) + geom_col()
```



The lower fares are dominated by people who didnt survive the catastophe. This will be another feature that will have a storng preidctive power.

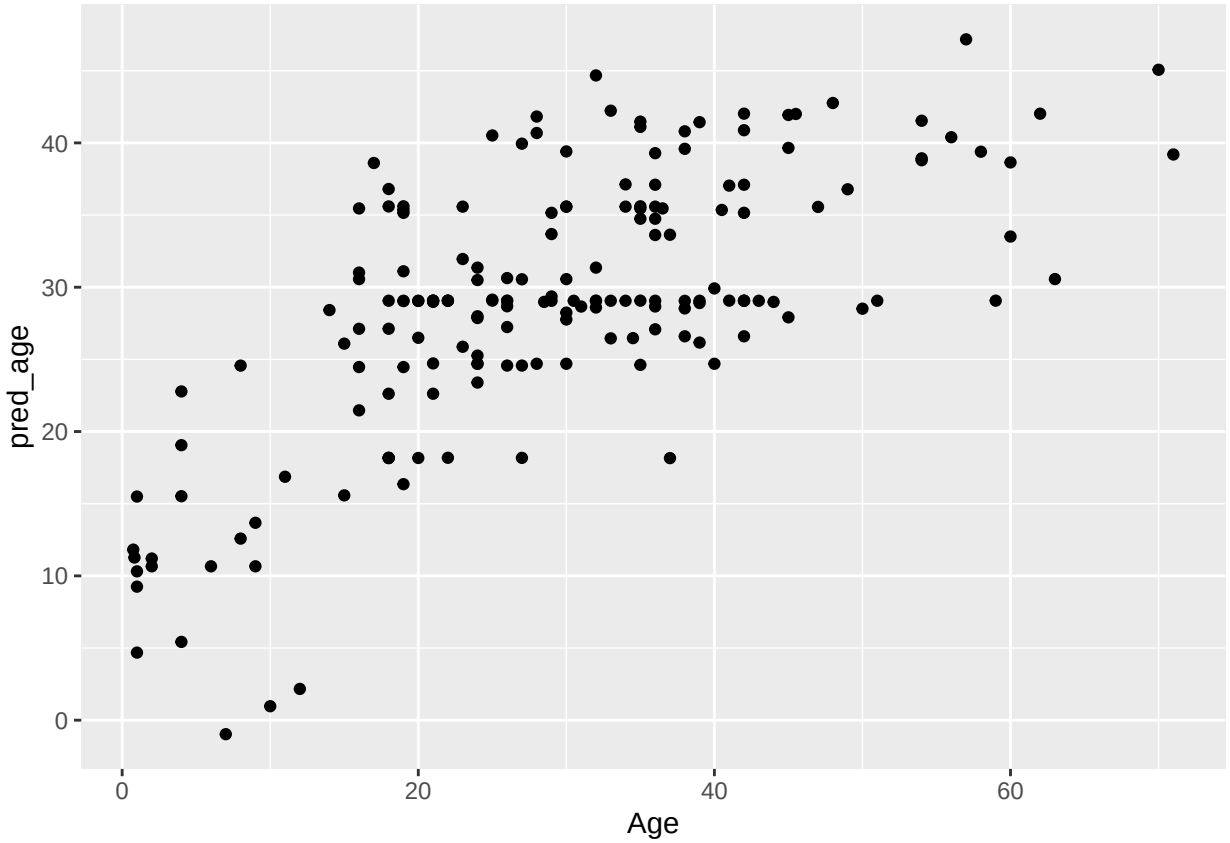
Methods

I am going to develope 3 models and compare them and there performance for this specific task. The 3 models will be random forest classifier, support vector machine and a logistic regression. In the litreature a random forest model has been cited in over 900 articles on this subject compared to around 5000 for support vector machine and only 100 for logistic regression. This gives some insight into likely success othe model. Before i get ot that I need to do some feautre engineering particluarly on the age column. Can i come up with reasonble way to impute this data?

###Feature Engineering

Age

In order tofill in the missing age data I am going to create a model which can predict the age. This will be a simple linear model. Looking at the passengers name variable I think if the title is master or miss that will signify what age the passenger is. Therefore I will extract the the title and use all the other available columns to estimate the age of a passneger



The summary of actual age against predicted age seems to do ok around the median age. However at either extremes there seems to be some innacuracies. This could be due to the modeling method decided and the limited data meaning the model doesnt feneralize well to the extremes. I will use this to impute the missing values however iw ill train two models and compare if it actually benefits it.

Also the title variable now it has been created I will use that in the model i create

###Random Forest

A random forest is a learning method that builds on the decision tree format. Anumber of random decision trees are are pruduced and the prediction is based on the average of all the decision trees.

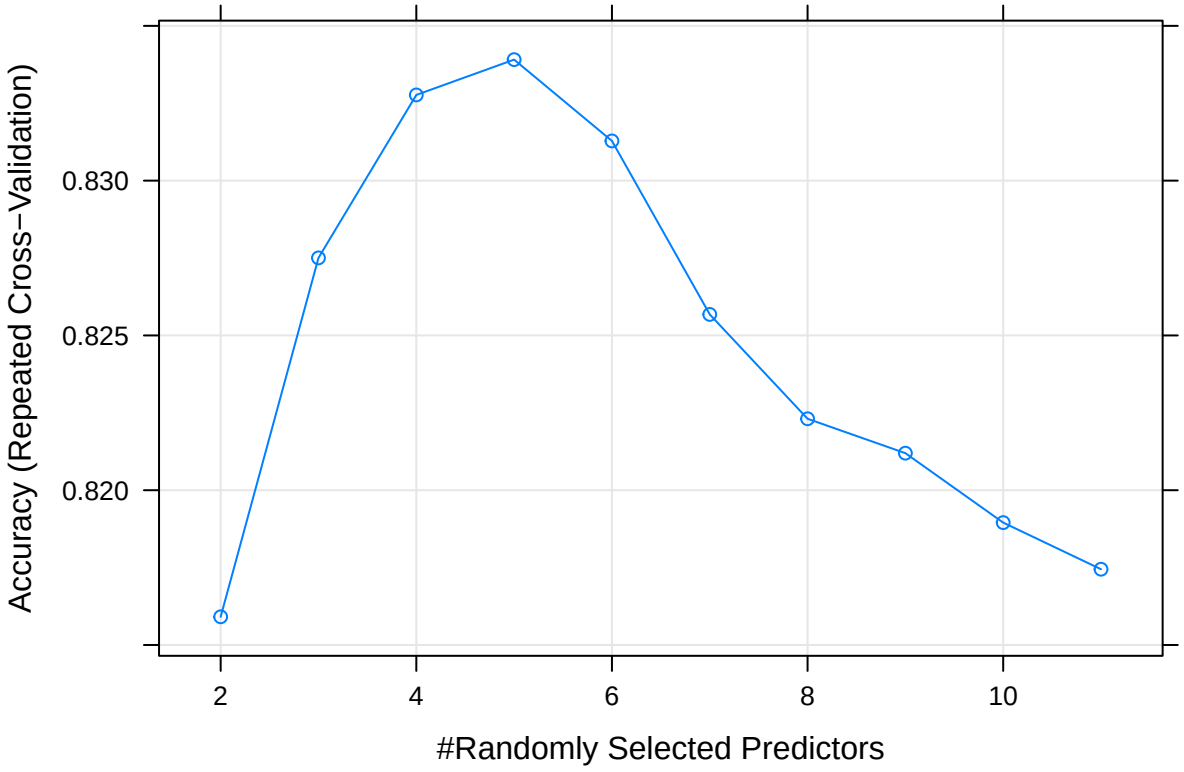
tuning parameters

mtry - the number of variables randomly sampled for each split ntress - which is the number of trees

```
## note: only 10 unique complexity parameters in default grid. Truncating the grid to 10 .
```

```
## Random Forest
##
## 891 samples
##   6 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 802, 803, 802, 801, 801, 802, ...
## Resampling results across tuning parameters:
```

```
##
##   mtry  Accuracy   Kappa
##    2    0.8159102  0.5990178
##    3    0.8275046  0.6193914
##    4    0.8327692  0.6324361
##    5    0.8339094  0.6378275
##    6    0.8312834  0.6347033
##    7    0.8256780  0.6250260
##    8    0.8223073  0.6193364
##    9    0.8211962  0.6178483
##   10    0.8189532  0.6129690
##   11    0.8174468  0.6095875
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 5.
```



Highest accurarcy for mtry is 5 so that is what i will be selecting for the final model.

Support Vector Machines

The second model which i will look at is the support vector machines. It is a non probalistic binary classifier. It learns from the training set the attributes that cause a particular result and then assigns new data to one category or the other based on that.

It has two tuning paremters:

c

sigma

###logistic regression

Logisitic regression gives the probability of a certain event occuring. The data is trained on events and gives a probability of the classification based on the variables.

##Results