Titanic: Random forests

### 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)

library(caret)  
library(dplyr)  
library(ggplot2)  
library(RCurl)

### 1. Reading data

url <- getURL('https://raw.githubusercontent.com/frankwwu/R-Knots/master/Titanic/train.csv')  
train <- read.csv(text = url)   
url <- getURL('https://raw.githubusercontent.com/frankwwu/R-Knots/master/Titanic/test.csv')  
test <- read.csv(text = url)

### 2. Removing NA.

train<-train[, !(colnames(train) %in% c('Name', 'Ticket', 'Cabin'))]  
train <-train %>% na.omit()  
test<-test[, !(colnames(test) %in% c('Name', 'Ticket', 'Cabin'))]  
test <- test %>% na.omit()

### 3. Selecting features

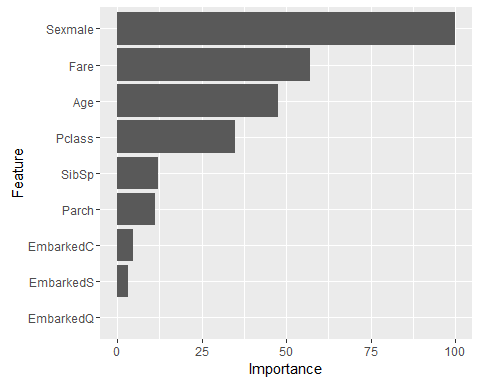
train$Survived <- factor(train$Survived)  
formula = Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked

### 4. Creating random forests model

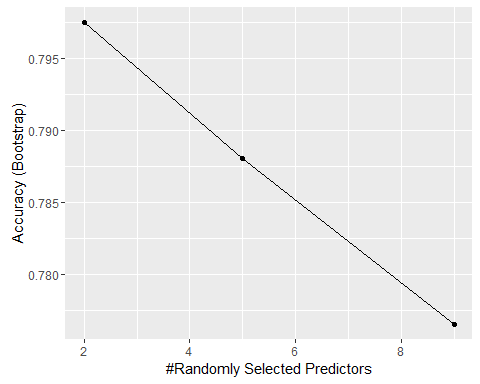
set.seed(1212)  
fit <- suppressMessages(train(formula, data=train, method="rf"))  
fit

## Random Forest   
##   
## 714 samples  
## 8 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 714, 714, 714, 714, 714, 714, ...   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa Accuracy SD Kappa SD   
## 2 0.7974984 0.5672478 0.01628335 0.03383909  
## 5 0.7880676 0.5542570 0.01949362 0.04005035  
## 9 0.7765616 0.5313036 0.02210269 0.04698568  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 2.

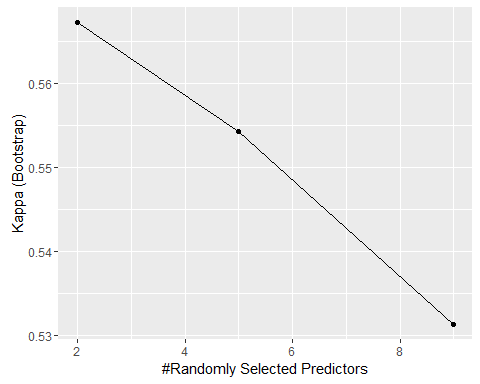
ggplot(varImp(fit))



ggplot(fit, metric = "Accuracy")

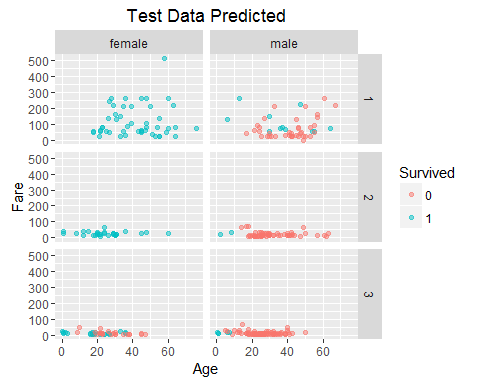


ggplot(fit, metric = "Kappa")



### 5. Predicting with the test data set

Survived <- predict(fit, test)  
r<-cbind(Survived, test)  
ggplot(r, aes(Age, Fare, color=Survived)) +   
 geom\_point(alpha = 0.5) +  
 facet\_grid(Pclass~Sex) +  
 ggtitle("Test Data Predicted")



ggplot(train, aes(Age, Fare, color=Survived)) +   
 geom\_point(alpha = 0.5) +  
 facet\_grid(Pclass~Sex) +  
 ggtitle("Training Data")



total<-rbind(train, r)  
ggplot(total, aes(Age, Fare, color=Survived)) +   
 geom\_point(alpha = 0.5) +  
 facet\_grid(Pclass~Sex) +  
 ggtitle("Total")

