## Random Forests

Read in dataset

titanic = titanic::titanic\_train

Structure and summary

str(titanic)

## '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 : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ 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 : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

summary(titanic)

## PassengerId Survived Pclass Name   
## Min. : 1.0 Min. :0.0000 Min. :1.000 Length:891   
## 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Class :character   
## Median :446.0 Median :0.0000 Median :3.000 Mode :character   
## Mean :446.0 Mean :0.3838 Mean :2.309   
## 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :891.0 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:891 Min. : 0.42 Min. :0.000 Min. :0.0000   
## Class :character 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000   
## Mode :character Median :28.00 Median :0.000 Median :0.0000   
## Mean :29.70 Mean :0.523 Mean :0.3816   
## 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :80.00 Max. :8.000 Max. :6.0000   
## NA's :177   
## Ticket Fare Cabin Embarked   
## Length:891 Min. : 0.00 Length:891 Length:891   
## Class :character 1st Qu.: 7.91 Class :character Class :character   
## Mode :character Median : 14.45 Mode :character Mode :character   
## Mean : 32.20   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

Factor conversion and recoding and imputation

titanic = titanic %>% mutate(Survived = as\_factor(Survived)) %>%   
 mutate(Survived = fct\_recode(Survived, "No" = "0", "Yes" = "1" )) %>%  
 mutate(Pclass = as\_factor(Pclass)) %>% mutate(Sex = as\_factor(Sex)) %>%  
 mutate(Embarked = as\_factor(Embarked)) %>%   
 mutate(Embarked = fct\_recode(Embarked,"Unknown"="","Cherbourg"="C","Southampton"="S","Queenstown"="Q")) %>%  
 select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked)  
  
set.seed(1234)  
imp\_age = mice(titanic, m=5, method='pmm', printFlag=FALSE)  
  
titanic\_complete = complete(imp\_age)   
summary(titanic\_complete)

## Survived Pclass Sex Age SibSp   
## No :549 1:216 male :577 Min. : 0.42 Min. :0.000   
## Yes:342 2:184 female:314 1st Qu.:20.00 1st Qu.:0.000   
## 3:491 Median :28.00 Median :0.000   
## Mean :29.11 Mean :0.523   
## 3rd Qu.:38.00 3rd Qu.:1.000   
## Max. :80.00 Max. :8.000   
## Parch Fare Embarked   
## Min. :0.0000 Min. : 0.00 Southampton:644   
## 1st Qu.:0.0000 1st Qu.: 7.91 Cherbourg :168   
## Median :0.0000 Median : 14.45 Queenstown : 77   
## Mean :0.3816 Mean : 32.20 Unknown : 2   
## 3rd Qu.:0.0000 3rd Qu.: 31.00   
## Max. :6.0000 Max. :512.33

###Building Random Forest Model with Tidymodels

titanic\_recipe = recipe(Survived ~., titanic\_complete) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_model = rand\_forest() %>%   
 set\_engine("ranger") %>%   
 set\_mode("classification")  
  
titanic\_wflow =   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(titanic\_recipe)  
  
set.seed(123)  
titanic\_fit = fit(titanic\_wflow, titanic\_complete)

Check out random forest details

titanic\_fit

## ══ Workflow [trained] ══════════════════════════════════════════════════════════  
## Preprocessor: Recipe  
## Model: rand\_forest()  
##   
## ── Preprocessor ────────────────────────────────────────────────────────────────  
## 1 Recipe Step  
##   
## • step\_dummy()  
##   
## ── Model ───────────────────────────────────────────────────────────────────────  
## Ranger result  
##   
## Call:  
## ranger::ranger(x = maybe\_data\_frame(x), y = y, num.threads = 1, verbose = FALSE, seed = sample.int(10^5, 1), probability = TRUE)   
##   
## Type: Probability estimation   
## Number of trees: 500   
## Sample size: 891   
## Number of independent variables: 10   
## Mtry: 3   
## Target node size: 10   
## Variable importance mode: none   
## Splitrule: gini   
## OOB prediction error (Brier s.): 0.1229052

Predictions

predRF = predict(titanic\_fit, titanic\_complete)  
head(predRF)

## # A tibble: 6 × 1  
## .pred\_class  
## <fct>   
## 1 No   
## 2 Yes   
## 3 No   
## 4 Yes   
## 5 No   
## 6 No

Confusion matrix

confusionMatrix(predRF$.pred\_class, titanic\_complete$Survived, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 531 66  
## Yes 18 276  
##   
## Accuracy : 0.9057   
## 95% CI : (0.8846, 0.9241)  
## No Information Rate : 0.6162   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.7953   
##   
## Mcnemar's Test P-Value : 2.926e-07   
##   
## Sensitivity : 0.8070   
## Specificity : 0.9672   
## Pos Pred Value : 0.9388   
## Neg Pred Value : 0.8894   
## Prevalence : 0.3838   
## Detection Rate : 0.3098   
## Detection Prevalence : 0.3300   
## Balanced Accuracy : 0.8871   
##   
## 'Positive' Class : Yes   
##