library("tidyverse")

## ── Attaching packages ─────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.2.1 ✓ purrr 0.3.3  
## ✓ tibble 2.1.3 ✓ dplyr 0.8.4  
## ✓ tidyr 1.0.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0

## ── Conflicts ────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("caret")

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library("ranger")

blood = read\_csv("Blood.csv")

## Parsed with column specification:  
## cols(  
## Mnths\_Since\_Last = col\_double(),  
## TotalDonations = col\_double(),  
## Total\_Donated = col\_double(),  
## Mnths\_Since\_First = col\_double(),  
## DonatedMarch = col\_double()  
## )

blood = blood %>% mutate(DonatedMarch = as.factor(as.character(DonatedMarch))) %>%  
 mutate(DonatedMarch = fct\_recode(DonatedMarch, "No" = "0", "Yes" = "1"))

set.seed(1234)  
train.rows = createDataPartition(y=blood$DonatedMarch, p=0.7, list=FALSE)  
train = blood[train.rows,]  
test = blood[-train.rows,]

fit\_control = trainControl(method = "cv", number = 10)  
set.seed(123)  
rf\_fit = train(x = blood [,-1], y = blood$DonatedMarch, method = "ranger", importance = "permutation", trControl = fit\_control, num.trees = 100)

## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.  
  
## Warning: Setting row names on a tibble is deprecated.

varImp(rf\_fit)

## ranger variable importance  
##   
## Overall  
## DonatedMarch 100.0000  
## Mnths\_Since\_First 0.2748  
## Total\_Donated 0.2576  
## TotalDonations 0.0000

rf\_fit

## Random Forest   
##   
## 748 samples  
## 4 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 674, 674, 673, 673, 673, 673, ...   
## Resampling results across tuning parameters:  
##   
## mtry splitrule Accuracy Kappa  
## 2 gini 1 1   
## 2 extratrees 1 1   
## 3 gini 1 1   
## 3 extratrees 1 1   
## 4 gini 1 1   
## 4 extratrees 1 1   
##   
## Tuning parameter 'min.node.size' was held constant at a value of 1  
## Accuracy was used to select the optimal model using the largest value.  
## The final values used for the model were mtry = 2, splitrule = gini  
## and min.node.size = 1.

Task 3 - The most important variable is DonatedMarch and the least important variable is TotalDonations.

predRF\_train = predict(rf\_fit, newdata = train)  
head(predRF\_train)

## [1] Yes Yes Yes Yes No Yes  
## Levels: No Yes

confusionMatrix(predRF\_train, train$DonatedMarch, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 399 0  
## Yes 0 125  
##   
## Accuracy : 1   
## 95% CI : (0.993, 1)  
## No Information Rate : 0.7615   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0000   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 1.0000   
## Prevalence : 0.2385   
## Detection Rate : 0.2385   
## Detection Prevalence : 0.2385   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : Yes   
##

Task 5 - The accuracy is 100%, the sensitivity is 100% and the specificity is 100%.

Task 6 - The accuracy of the model is 100% so it is better than a naive model. The naive model uses all observations in the majority class which could be overkill when this model above produced 100% accuracy.

predRF\_test = predict(rf\_fit, newdata = test)

confusionMatrix(predRF\_test, test$DonatedMarch, positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 171 0  
## Yes 0 53  
##   
## Accuracy : 1   
## 95% CI : (0.9837, 1)  
## No Information Rate : 0.7634   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0000   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 1.0000   
## Prevalence : 0.2366   
## Detection Rate : 0.2366   
## Detection Prevalence : 0.2366   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : Yes   
##

Task 7 - The model performs great on the testing set as it’s accuracy is 100% as well as the Sensitivity and Specificty.

Task 8 - In the real-world it could be used to predict how much blood they might need during a time when they are normally low on donations to avoid being low on donations. You could use this model in the real world. I really don’t have any concerns about using this model in the real world.