Final Group Project Cardiovascular Disease

Group 1

Kailun Chen

Chenyuan Wang

Zhenwu Yang

Tian Yue

Northeastern University

May 14, 2019

Author Note

This paper was prepared for Predictive Analytics, taught by Justin Rodgers.

**Introduction**

As one of the top 3 mortal diseases in the united states, cardiovascular diseases kill more than 600,000 poeples in every year which is about quarter of patients. Therefore, the researchers tried to fit model in R studio to predict whether a patient exposes to the risk of disease. The etiology of heart disease is arteriosclerosis, hypertension, virus infection, respiratory disease and other causes. The patient's heart disease can be inferred from blood pressure, cholesterol levels, glucose level and bad habits. The dataset is mainly based on factual information, results of medical examination and information given by the patient. In general, according to the binary datasets, the researchers fit tree models or random forests to solve the research question.

**Methods**

**Methodology**

This research will use the functions of R studio including decision tree models( C5.0 Tree model, Boosted Tree model and Cost Matrix Tree Model) and random forests(tune using caret evaluation).

**Dataset & Variables**

The dataset which was recorded by an unknown hospital and downloaded from Kaggle has 70000 observations and 13 variables. In this project, the researchers selected 9 independent variables from the dataset, including gender, Systolic and diastolic blood pressure, cholesterol and glucose levels and habits such smoke, alcohal and active to test different cardiovascular diseases prediction models.

**Analysis Methods**

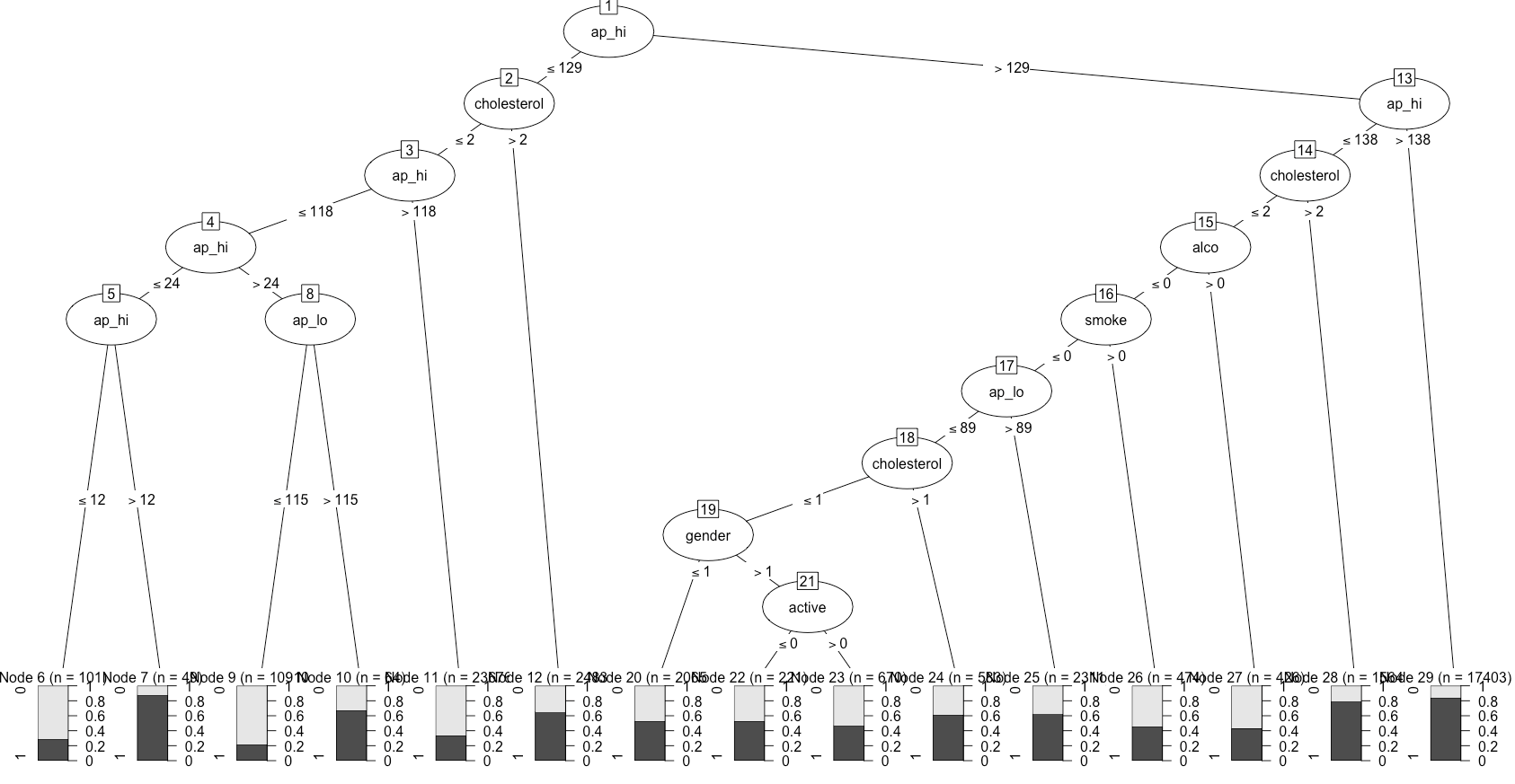
**Tree Model** In this research, researchers will use three models to compare which model is more suitable for testing cardiovascular diseases. These three models are C5.0 Tree model, Boosted Tree model and Cost Matrix Tree Model. Besides, through comparing the predated result of the model and the original result of the target dataset, we will figure out the accuracy of the decision tree model. To improve the accuracy, we also use the boosted model and the cost matri

**Random Forest.** In this part, researchers will use evaluations of tune using caret and grid research. Grid research will define a grid of algorithm parameters to try. Each axis of the grid is an algorithm parameter, and points in the grid are specific combinations of parameters. Because we are only tuning one parameter, the grid search is a linear search through a vector of candidate values.

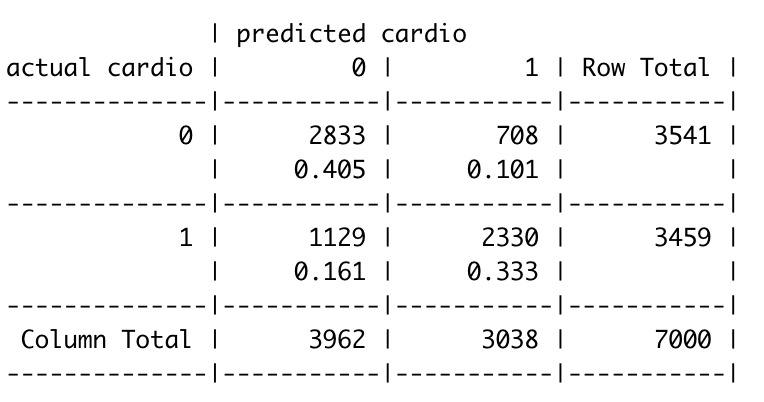
**Results**

**Tree Model**

**Decision tree model**

When cholesterol is less than or equal to 2 and systolic blood pressure greater than 12 and less than or equal to 24, the probability of people getting cardiovascular diseases is the biggest of these variables. 

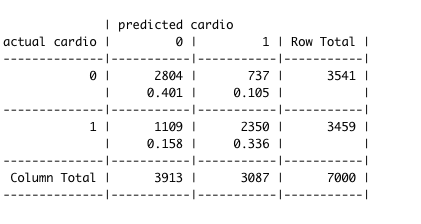
**Figure - 1**



**Table - 1**

In this model, accuracy rate of model is (2833+2330)/7000 = 0.7375714 and predicted Cardio rate is 2330/3459 = 0.6736051.

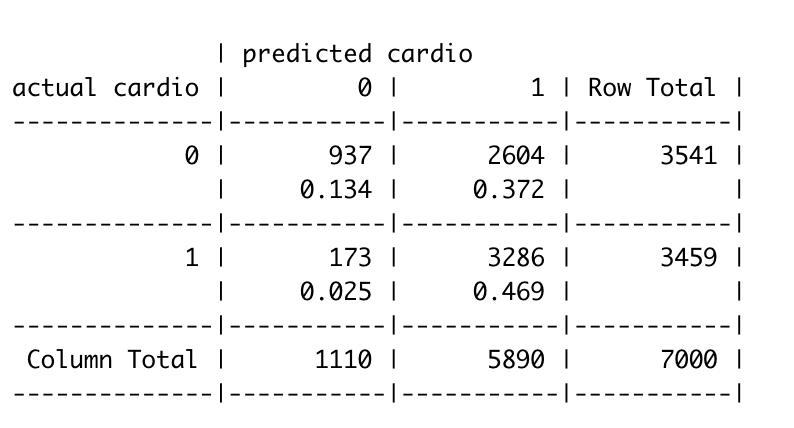
**Boosted model**



**Table - 2**

In this model, accuracy rate of model is (2804+2350)/7000 = 0.7362857 and predicted Cardio rate is 2350/3459 = 0.6793871.

**Cost matrix model**



**Table - 3**

In this model, accuracy rate of model is (937+3286)/7000 = 0. 6032857 and predicted Cardio rate is 3286/3459 = 0.9499855.

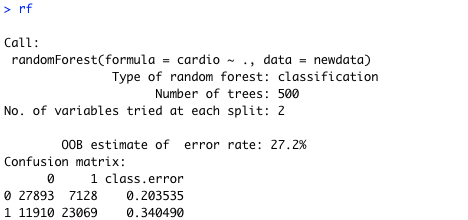
|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy rate of model** | **Cardio rate** |
| Decision tree model | 73.76% | 67.36% |
| Boosted model | 73.63% | 67.94% |
| Cost matrix model | 60.33% | 94% |

**Table - 4**

According to the table, we can find the decision tree model has the highest accuracy. However, for cardiovascular diseases, the no default accuracy is more important for them to predict cardiovascular diseases. Thus, the cost matrix model is a good choice for researchers to fit the model and predict the results.

**Random Forest**

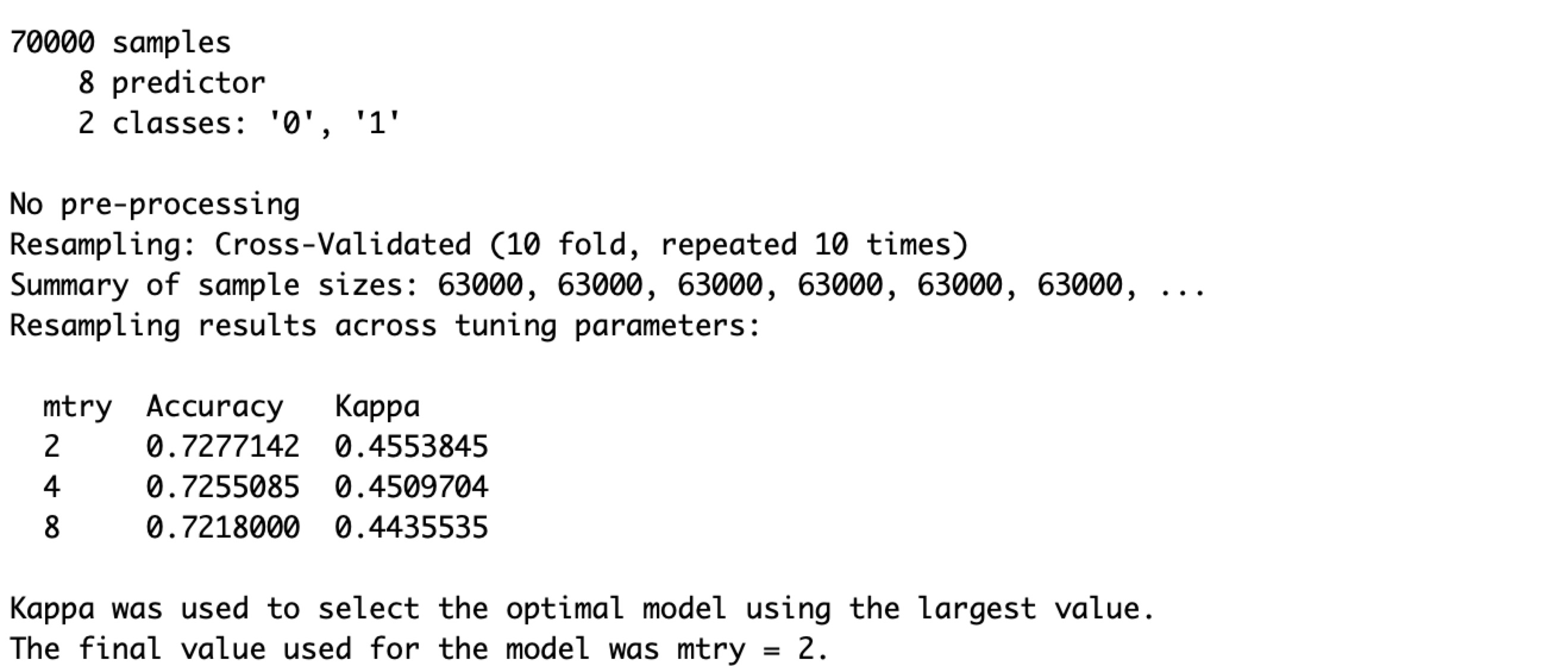
In order to find a precise model, the reaserchers used random forest mode to develop further study. Random forests focus on ensembles of decision trees and combines the base principles of bagging with random feature selection to add additontionala diversity to the decision tree model, therefor, it could handel extremely large dataset.



**Output - 1**

The output notes that the random forest included 500 trees and tried 2 variabe which is calculated by sqrt(p) and is used to limit the features so that each splits. The project totally has 8 features and in order to make each feature has a chance to appear in several models, the random forest model used a large number of trees to train the dataset and the number of variable tried at each split is calculated by sqrt(p) and is used to limit the features occurs from tree-to-tree. The error rate 27.2% is unexpected low, the researchers believe it is because the causes of Cardiovasculer disease are complex and varied and 8 variables are not enough to build a precise model.

The researches compare the auto-tuned random forest to the best auto-tuned boosted C5.0 model to identify a optimal model.



**Output - 2**

A screenshot of a cell phone

Description automatically generated

**Output - 3**

The Output – 1and Output – 2 are the summarized results indicating that for both approaches with mtry equals to 2, the model is the most precise. Because the Kappa value of random forest about 0.4553845 is slightly higher than the Kappa value of boosted C5.0 model about 0.4552804, the random forest value is the best choice.

**Conclusion**

**Tree Model**

The

**Random Forest**

For

For the decision tree models, the cost matrix model has the highest accuracy rate of predicting cardio diseases.

For the random forest, after using the tune algorithm of the Caret packages in R, when the “mtry” is 2, we got the highest accuracy (0.728)and kappa value(0.455).

The accuracy rate and the kappa values of the random forest are good but not very well. We believe fit models with factors that will lead to cardio disease is not as easy as we believe. The causes of Cardio disease are complex and varied.

References

Aiken, L. S., West, S. G., & Reno, R. R. (1991). Multiple regression: Testing and interpreting interactions. Sage.

**Appendix**

**R Coding**

dataset <- read.csv2("/Users/mac/Documents/Analytics/ALY6020/Final/cardio.csv")

myvars <- c("gender","ap\_hi","ap\_lo","cholesterol","gluc","smoke","alco","active","cardio")

newdata <- dataset[myvars]

#Part A

#Tree Model

#Prepare data

str(newdata)

table(newdata$cardio)

set.seed(666)

train\_sample <-sample(70000, 63000)

str(train\_sample)

cardio\_train <- newdata[train\_sample,]

cardio\_test <- newdata[-train\_sample,]

#No.1 method Build tree model

prop.table(table(cardio\_train$cardio))

prop.table(table(cardio\_test$cardio))

installed.packages("C50")

library(C50)

Tree\_model <- C5.0(cardio\_train[-9],factor(cardio\_train$cardio))

Tree\_model

summary(Tree\_model)

plot(Tree\_model)

cardio\_pred <- predict(Tree\_model,cardio\_test)

library(gmodels)

CrossTable(cardio\_test$cardio, cardio\_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,dnn = c('actual cardio','predicted cardio'))

(2781+2305)/7000

#accuracy rate 0.7265714

2305/3541

#cardio rate 0.6509461

#No.2 method

#Boosted model of decision tree for 10 boosted

cardio\_boost10 <- C5.0(cardio\_train[-9], factor(cardio\_train$cardio), trials = 40)

cardio\_boost10

summary(cardio\_boost10)

cardio\_boost10\_pred <- predict(cardio\_boost10, cardio\_test)

CrossTable(cardio\_test$cardio, cardio\_boost10\_pred, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,dnn=c('actual cardio','predicted cardio'))

(2745+2334)/7000

#accuracy rate 0.7255714

2334/3541

#cardio rate 0.6591358

#No.3 method

#Cost martix model of decision tree

matrix\_dimensions <-list(c("0","1"),c("0","1"))

names(matrix\_dimensions) <- c("predicted", "actual")

matrix\_dimensions

error\_cost <-matrix(c(0,1,3,0), nrow = 2, dimnames = matrix\_dimensions)

error\_cost

cardio\_cost <- C5.0(cardio\_train[-9], factor(cardio\_train$cardio),

costs = error\_cost)

cardio\_cost\_pred <- predict(cardio\_cost, cardio\_test)

CrossTable(cardio\_test$cardio, cardio\_cost\_pred,

prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,

dnn = c('actual cardio', 'predicted cardio'))

(822+3328)/7000

#accuracy rate 0.5928571

3328/3542

#cardio rate 0.9395822

#Part B

#Random Forest

library(randomForest)

set.seed(123)

newdata$cardio <- factor(newdata$cardio)

rf<- randomForest(cardio ~ ., data = newdata)

rf

#evaluate #1

library(caret)

ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 10)

grid\_rf <- expand.grid(.mtry = c(2, 4, 8))

set.seed(123)

m\_rf <- train(cardio ~ gender+ap\_hi+ap\_lo+cholesterol+gluc+smoke+alco+active, data = newdata, method = "rf",

metric = "Kappa", trControl = ctrl,

tuneGrid = grid\_rf)

# picture:33-1-2-2

#mtry =2, accuracy = 0.7277142, kappa = 0.4553845

#mtry =4, accuracy = 0.7255085, kappa = 0.4509704

#mtry =8, accuracy = 0.7218000, kappa = 0.4435535

#evaluate #2

#grid search

# to define a grid of algorithm parameters to try

# Each axis of the grid is an algorithm parameter, and points in the grid are specific combinations of parameters.

# Because we are only tuning one parameter, the grid search is a linear search through a vector of candidate values.

ctrl <- trainControl(method = "repeatedcv",number = 10, repeats = 3, search = "grid")

set.seed(123)

tunegrid <- expand.grid(.mtry=c(1:15))

rf\_gridsearch <- train(cardio ~ . ,data = newdata, method = "rf", metric = "Accuracy", turnGrid = tunegrid, trContr0l = ctrl)

print(rf\_gridsearch)

plot(rf\_gridsearch)

#mtry= 2/5/8, accurancy = 0.7276203 / 0.7207180 / 0.7179817, Kappa = 0.4552804 / 0.4414862 / 0.4360141

References:

Lantz, B. (2015). Machine Learning with R (2nd Edition). Birmingham: Packt Publishing Ltd.

**Dataset citation**

Ulianova, S. (2019, January 20). Cardiovascular Disease dataset. Retrieved May 14, 2019, from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset