810 Team Project

Bo Li U24425931

2/24/2021

update.packages(ask = FALSE, checkBuilt = TRUE) tinytex::tlmgr_update()

install.packages(c("data.table", "ggplot2", "ggthemes", "scales", "rpart","randomForest","glmnet","gbm"))

install.packages("xgboost") install.packages("rpart.plot") install.packages("ipred")

library(data.table) library(ggplot2) library(ggthemes) library(glmnet) theme_set(theme_bw()) library(MASS) library(rpart) library(rpart.plot) library(ipred) library(randomForest)

dd_train <- fread("C:/Users/boli0/Downloads/train.csv") dd_test <- fread("C:/Users/boli0/Downloads/test.csv") set.seed(810)

Build single decision tree classification

fit = rpart(price_range ~ .,method = "class", data = dd_train,control = rpart.control(minsplit = 1) , parms = list(split = "information"))

print(fit)

summary(fit)

plot(fit) # Build decision tree while cp = 0.01000000 # E1 is 0.183.

fit2 <- prune(fit, cp = 0.01000000) rpart.plot(fit2, type = 4, branch = 0, extra = 2)

print(fit2)

CFit1 <- predict(fit, dd_train, type = "class")

ConfM1 <- table(dd_train$price_range, CFit1)

(E1 <- (sum(ConfM1) - sum(diag(ConfM1)))/sum(ConfM1))

Utiltize bagging from ipred package to build combined decision tree 1

Bagging classification trees with 25 bootstrap replications, E2 is 0.75.

BagM1 <- bagging(price_range ~., data = dd_train, nbagg = 25, coob = TRUE, control = rpart.control(minsplit = 1))

CFit2 <- predict(BagM1, dd_train, type = "class")

ConfM2 <- table(dd_train$price_range, CFit2)

(E2 <- (sum(ConfM2) - sum(diag(ConfM2)))/sum(ConfM2))

Random forest

E3 is 0.9995.

rFM <- randomForest(price_range ~., data = dd_train, importance = TRUE, proximity = TRUE)

print(rFM)

summary(rFM)

plot(rFM)

Fit <- predict(rFM, dd_train)

ConfM3 <- table(dd_train$price_range, Fit)

(E3 <- (sum(ConfM3) - sum(diag(ConfM3)))/sum(ConfM3))

barplot(rFM$importance[,2],main = "Feature Importance Barplot")

varImpPlot(x = rFM, sort = TRUE, n.var = nrow(rFM$importance), main = "Feature Importance ScatterPlot")