Midterm Project Two for Racial Ethnic Representativeness Data

Tobias Boggess, Carl Perry

4/11/2022

The task of data wrangling for this project was fairly straight forward. After reading the data into a data frame it was then filtered by the academic year 2017. This was sufficient for the first two tasks. The third task required the use of mutate to add prediction columns to the data frame.

Task 1

For the first task two separate models were run for the multiple regression analysis. The first model compared the percentage of white students with the percentages of the other ethnic groups. The second model was the predicted percentages based on the geographic market. The following are the estimated model coefficients.

College

$$y_{white} = 90.080 - 0.978X_{hispa} - 0.954X_{black} - 1.239X_{asian} - 0.935X_{amind} - 1.072X_{pacis} - 0.65145X_{twora}$$

Geographic Market

$$y_{white} = 100.435 - 1.187X_{hispa} - 1.007X_{black} - 1.790X_{asian} - 1.283X_{amind} - 0.211X_{pacis} - 0.549X_{twora}$$

There was a negative correlation in both models. This is consistent with the raw data, as most of institutions had a white majority.

Both models had significant P-values, as the summary function reported *** next to the value indicating it was a number small enough as to be effectively zero. The only insignificant value to note is the comparison to Pacific Islanders on the market model being 0.0605 and therefore failing the standard .05 significance.

The college model had a R^2 of 0.8594 with a RSE of 9.047 on 1871 degrees of freedom, the market model had a R^2 of 0.991 with a RSE of 1.528 on 1871 degrees of freedom. The R^2 measures show that both models had a fairly strong fit but the RSE indicates that the college model shows more deviation.

```
##
## Call:
## lm(formula = col_white ~ col_hispa + col_black + col_asian +
##     col_amind + col_pacis + col_twora, data = fryr.cllg.17)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.514  -1.876   2.343   5.164  20.799
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.07997    0.45555 197.738  < 2e-16 ***
## col_hispa   -0.97831    0.01743 -56.134  < 2e-16 ***
## col_black   -0.95399    0.01113 -85.728  < 2e-16 ***
## col_asian   -1.23876    0.03427 -36.144  < 2e-16 ***
## col_amind   -0.93534    0.06547 -14.286  < 2e-16 ***
## col_pacis   -1.07193    0.12351  -8.679  < 2e-16 ***
## col_twora   -0.65145    0.08135  -8.008 2.03e-15 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.047 on 1871 degrees of freedom
## Multiple R-squared:  0.8594, Adjusted R-squared:  0.8589
## F-statistic:  1906 on 6 and 1871 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = mkt_white ~ +mkt_hispa + mkt_black + mkt_asian +
##     mkt_amind + mkt_pacis + mkt_twora, data = fryr.cllg.17)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3241  -0.7835   0.0548   0.6760  11.2242
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 100.434511   0.185447  541.580   <2e-16 ***
## mkt_hispa    -1.187212   0.003623 -327.655   <2e-16 ***
## mkt_black    -1.007227   0.004452 -226.235   <2e-16 ***
## mkt_asian    -1.790438   0.019343  -92.561   <2e-16 ***
## mkt_amind    -1.283475   0.034391  -37.321   <2e-16 ***
## mkt_pacis    -0.210605   0.112132   -1.878   0.0605 .
## mkt_twora    -0.548759   0.052621  -10.429   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.528 on 1871 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.991
## F-statistic: 3.446e+04 on 6 and 1871 DF,  p-value: < 2.2e-16
```

Task 2

For the second task a logistic regression analysis was executed based on the college reported percentages categorized by public, private, and for-profit colleges. The explanatory variables for all the analysis included all of the different ethnicities.

College Public

$$y_{public} = -11.536 + 0.107x_{wht} + 0.122X_{his} + 0.108X_{blk} + 0.152X_{asn} + 0.212X_{amind} - 0.939X_{pacis} + 0.261X_{twora}$$

College Private

$$y_{private} = 3.718 - 0.027x_{wht} - 0.049X_{his} - 0.041X_{blk} - 0.052X_{asn} - 0.212X_{amind} - 0.050X_{pacis} - 0.050X_{twora}$$

College For-profit

$$y_{profit} = 1.145 - 0.050x_{wht} - 0.014X_{his} - 0.016X_{blk} - 0.046X_{asn} - 0.011X_{amind} + 0.379X_{pacis} - 0.222X_{twora}$$

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = public ~ col_white + col_hispa + col_black + col_asian +
##     col_amind + col_pacis + col_twora, family = "binomial", data = fryr.cllg.17)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5123  -0.8709  -0.6208   1.1907   3.7257
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.53651    1.08094 -10.673  < 2e-16 ***
## col_white     0.10739    0.01131   9.495  < 2e-16 ***
## col_hispa     0.12228    0.01220  10.023  < 2e-16 ***
## col_black     0.10825    0.01149   9.423  < 2e-16 ***
## col_asian     0.15216    0.01658   9.178  < 2e-16 ***
## col_amind     0.21231    0.03916   5.421 5.91e-08 ***
## col_pacis    -0.93905    0.17005  -5.522 3.35e-08 ***
## col_twora     0.26073    0.02748   9.487  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2293.7  on 1877  degrees of freedom
## Residual deviance: 2050.2  on 1870  degrees of freedom
## AIC: 2066.2
##
## Number of Fisher Scoring iterations: 6
```

```
##
## Call:
## glm(formula = private ~ col_white + col_hispa + col_black + col_asian +
##     col_amind + col_pacis + col_twora, family = "binomial", data = fryr.cllg.17)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3693  -1.2460   0.8197   0.9534   2.9705
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.718270   0.592331   6.277 3.44e-10 ***
## col_white   -0.027354   0.006347  -4.310 1.63e-05 ***
## col_hispa   -0.048694   0.007532  -6.465 1.01e-10 ***
## col_black   -0.041382   0.006660  -6.213 5.18e-10 ***
## col_asian   -0.052327   0.011198  -4.673 2.97e-06 ***
## col_amind   -0.217837   0.045555  -4.782 1.74e-06 ***
## col_pacis   -0.050213   0.030953  -1.622   0.1047
## col_twora   -0.049882   0.020253  -2.463   0.0138 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2517.6  on 1877  degrees of freedom
## Residual deviance: 2396.2  on 1870  degrees of freedom
## AIC: 2412.2
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Call:
## glm(formula = forprofit ~ col_white + col_hispa + col_black +
##     col_asian + col_amind + col_pacis + col_twora, family = "binomial",
##     data = fryr.cllg.17)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2019  -0.3929  -0.2918  -0.2419   3.2931
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.144500   0.601730   1.902  0.05717 .
## col_white   -0.046611   0.007043  -6.618 3.64e-11 ***
## col_hispa   -0.014273   0.008014  -1.781  0.07492 .
## col_black   -0.016388   0.006886  -2.380  0.01732 *
## col_asian   -0.045838   0.016051  -2.856  0.00429 **
## col_amind   -0.011090   0.017930  -0.619  0.53624
## col_pacis    0.378508   0.088781   4.263 2.01e-05 ***
## col_twora   -0.222393   0.042625  -5.217 1.81e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1168.32  on 1877  degrees of freedom
## Residual deviance:  967.08  on 1870  degrees of freedom
## AIC: 983.08
##
## Number of Fisher Scoring iterations: 6
```

Task 3

For the purpose of practice, all methods except the Neural network were executed. For the purpose of this write up, it was decided to examine K nearest neighbor (Knn) more closely, as it was the most accurate. A total of three different models were run using Knn. The models used were college percentage, geographic market percentage, and the difference. Total enrollment and all of the racial ethnicity were used in the calculation for each model.

All three models have a reported accuracy above eighty percent indicating that these models predict individuals decently.

Accuracy College Percentage

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.834
```

Accuracy Market Percentage

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.863
```

Accuracy difference Percentage

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy multiclass     0.850
```

Prediction example Using college percent.

```
## 
## Call:
## kknn(formula = as.factor(fourcat) ~ total_enrollment + col_white +     col_hispa + col_black + col_as
## 
## Response: "nominal"
##                     fit prob.For.Profit prob.Private.More.Selective
## 1 Private Non-Selective      0.08818637                   0.1140304
##   prob.Private.Non.Selective prob.Private.Selective prob.Public.More.Selective
## 1                  0.4348051              0.3629782                          0
##   prob.Public.Non.Selective prob.Public.Selective
## 1                         0                     0
```

All Code

```
# Name: Tobias Boggess, Carl Perry
# Date: April 5, 2022
# Description: Three tasks, the first one is to carry out a multiple regression
# analysis with a minimum of three explanatory variables. The second task is to
# carry out a logistic regression analysis with two or more explanatory
# variables. The third task is to carry a machine learning classification procedure
# to predict the fourcat.


################################################################################
#                               Loading Libraries                              #
################################################################################

library(tidyr)
library(dplyr)
library(ggplot2)
library(rattle)
library(mclust)
library(yardstick)
library(rpart)
library(randomForest)
library(kknn)

# Loading data into r from csv worksheet
my.file <- file.choose()
fryr.cllg <-
  read.csv(my.file,
           header = TRUE,
           sep = ",",
           stringsAsFactors = FALSE)
fryr.cllg.17 <- filter(.data = fryr.cllg, year == 2017)

################################################################################
#                                   Task One                                   #
################################################################################


# Uses mkt explanatory variables in this multiple regression analysis
# Try to predict white ethnicity percent in college demographics
# fryr.cllg.colwht.reg <-
#   lm(
#     col_white ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.colwht.reg)
#
#
# # Best variables to predict white ethnicity percent for college demographics
# fryr.cllg.colwht.regshort <-
#   lm(col_white ~ mkt_white + mkt_hispa + mkt_black + mkt_twora,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.colwht.regshort)
#
#
# # Try to predict hispanic ethnicity percent in college demographics
# fryr.cllg.colhsp.reg <-
#   lm(
#     col_hispa ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.colhsp.reg)
#
#
# # Best variables to predict hispanic ethnicity percent for college demographics
# fryr.cllg.colhsp.regshort <-
#   lm(col_hispa ~ mkt_white + mkt_hispa + mkt_black,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.colhsp.regshort)
#
#
# # Try to predict black ethnicity percent in college demographics
# fryr.cllg.colblk.reg <-
#   lm(
#     col_black ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.colblk.reg)
#
#
# # Best variables to predict black ethnicity percent for college demographics
# fryr.cllg.colblk.regshort <-
#   lm(col_black ~ mkt_asian + mkt_black + mkt_twora,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.colblk.regshort)
#
#
# # Try to predict asian ethnicity percent in college demographics
# fryr.cllg.colasn.reg <-
#   lm(
#     col_asian ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.colasn.reg)
#
```

```
#
# # Best variables to predict asian ethnicity percent for college demographics
# fryr.cllg.colasn.regshort <-
#   lm(col_asian ~ mkt_white + mkt_black + mkt_asian,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.colasn.regshort)
#
#
# # Try to predict American Indian ethnicity percent in college demographics
# fryr.cllg.colamd.reg <-
#   lm(
#     col_amind ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.colamd.reg)
#
#
# # Best variables to predict American Indian ethnicity percent for college demographics
# fryr.cllg.colamd.regshort <-
#   lm(col_amind ~ mkt_amind + mkt_pacis + mkt_twora,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.colamd.regshort)
#
#
# # Try to predict Pacific Island ethnicity percent in college demographics
# fryr.cllg.colpcs.reg <-
#   lm(
#     col_pacis ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.colpcs.reg)
#
#
# # Best variables to predict Pacific Island ethnicity percent for college demographics
# fryr.cllg.colpcs.regshort <-
#   lm(col_pacis ~ mkt_asian + mkt_pacis + mkt_twora,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.colwht.regshort)
#
#
# # Try to predict multiracial ethnicity percent in college demographics
# fryr.cllg.coltwa.reg <-
#   lm(
#     col_twora ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
#       mkt_amind + mkt_pacis + mkt_twora,
#     data = fryr.cllg.17
#   )
#
# summary(fryr.cllg.coltwa.reg)
#
#
# # Best variables to predict multiracial ethnicity percent for college demographics
# fryr.cllg.coltwa.regshort <-
#   lm(col_twora ~ mkt_white + mkt_amind + mkt_twora,
#      data = fryr.cllg.17)
#
# summary(fryr.cllg.coltwa.regshort)


# Predict col_white based on other col_* variables
fryr.cllg.colwht.reg1 <-
  lm(
    col_white ~  col_hispa + col_black + col_asian +
      col_amind + col_pacis + col_twora,
    data = fryr.cllg.17
  )

# Summary of multiregression model using college variables
summary(fryr.cllg.colwht.reg1)

# Plots the residuals using the built in function
plot(fryr.cllg.colwht.reg)


# Predict mkt_white based on other mkt_* variables
fryr.cllg.mktwht.reg1 <-
  lm(
    mkt_white ~ + mkt_hispa + mkt_black + mkt_asian +
      mkt_amind + mkt_pacis + mkt_twora,
    data = fryr.cllg.17
  )

# Summary of multiregression model using market variables
summary(fryr.cllg.mktwht.reg1)

# Plots the residuals using built in R
plot(fryr.cllg.mktwht.reg1)


################################################################################
#                                   Task Two                                   #
################################################################################


# Public variable predictions
# College percent
fryr.cllg.logreg <-
  glm(
    public ~ col_white + col_hispa + col_black + col_asian +
```

```r
      col_amind + col_pacis + col_twora,
    data = fryr.cllg.17,
    family = "binomial"
  )

# Summary of the logistic regression model using the college percent variables
summary(fryr.cllg.logreg)

# Market percent
fryr.cllg.mktlogreg <-
  glm(
    public ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
      mkt_amind + mkt_pacis + mkt_twora,
    data = fryr.cllg.17,
    family = "binomial"
  )

# Summary of the logistic regression model using the market percent variables
summary(fryr.cllg.mktlogreg)


# Private variable predictions
# College percent
fryr.cllg.privlogreg <-
  glm(
    private ~ col_white + col_hispa + col_black + col_asian +
      col_amind + col_pacis + col_twora,
    data = fryr.cllg.17,
    family = "binomial"
  )

# Summary of the logistic regression model using the college percent variables
summary(fryr.cllg.privlogreg)

# Market percent
fryr.cllg.privmktlogreg <-
  glm(
    private ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
      mkt_amind + mkt_pacis + mkt_twora,
    data = fryr.cllg.17,
    family = "binomial"
  )

# Summary of the logistic regression model using the market percent variables
summary(fryr.cllg.privmktlogreg)


# For profit variable predictions
# College percent
fryr.cllg.proflogreg <-
  glm(
    forprofit ~ col_white + col_hispa + col_black + col_asian +
      col_amind + col_pacis + col_twora,
    data = fryr.cllg.17,
    family = "binomial"
  )

# Summary of the logistic regression model using the college percent variables
summary(fryr.cllg.proflogreg)

# Market percent
fryr.cllg.profmktlogreg <-
  glm(
    forprofit ~ mkt_white + mkt_hispa + mkt_black + mkt_asian +
      mkt_amind + mkt_pacis + mkt_twora,
    data = fryr.cllg.17,
    family = "binomial"
  )

# Summary of the logistic regression model using the market percent variables
summary(fryr.cllg.profmktlogreg)



################################################################################
#                                   Task Three                                 #
################################################################################


################################  Decision Trees  ##############################

# First using college percent ethnicities and total enrollment to predict
# the fourcat variable
fryr.cllg.tree.1 <-
  rpart(
    fourcat ~ total_enrollment + col_white + col_hispa + col_black + col_asian +
      col_amind + col_pacis + col_twora,
    data = fryr.cllg.17,
    control = rpart.control(cp = 0.05)
  )

# Used to plot the decision tree of college explanatory variables
par(xpd = TRUE)

plot(fryr.cllg.tree.1, compress = TRUE)
text(fryr.cllg.tree.1, use.n = TRUE)
par(xpd = FALSE)

# Summary of decision tree to predict the fourcat variable based on total
# enrollment and col_* explanatory variables.
summary(fryr.cllg.tree.1)

# Used for getting the predictions based on the above model
preds <- predict(fryr.cllg.tree.1, type = "class")

# Used to compare the truth (fourcat) with the prediction (predType)
```

```r
fryr.cllg.17one <- mutate(fryr.cllg.17, predType = preds)

# Actually gets the accuracy of the predictions from the model above
accuracy(
  data = fryr.cllg.17one,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Shows the predictions in a table with the true values
conf_mat(data = fryr.cllg.17one, truth = fourcat, estimate = predType)


# Second using market percent ethnicities and total enrollment to predict the
# type of school (fourcat)
fryr.cllg.tree.2 <-
  rpart(
    fourcat ~ total_enrollment + mkt_white + mkt_hispa + mkt_black + mkt_asian +
      mkt_amind + mkt_pacis + mkt_twora,
    data = fryr.cllg.17,
    control = rpart.control(cp = 0.05)
  )

# Generates plot of the decision tree for the market explanatory variables
par(xpd = TRUE)

plot(fryr.cllg.tree.2, compress = TRUE)
text(fryr.cllg.tree.2, use.n = TRUE)
par(xpd = FALSE)

# Provides the summary of the model above. Will provide how accurate the model is.
summary(fryr.cllg.tree.2)

# Used to find the predictions to add to the data frame
preds2 <- predict(fryr.cllg.tree.2, type = "class")

# Adds the predictions to the data frame
fryr.cllg.17two <- mutate(fryr.cllg.17, predType = preds2)

# Illustrates accuracy of the model above
accuracy(
  data = fryr.cllg.17two,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Shows results of the predictions in a table format with the true values
conf_mat(data = fryr.cllg.17two, truth = fourcat, estimate = predType)


# Third using difference between college and market percent ethnicities with
# total enrollment to predict fourcat
fryr.cllg.tree.3 <-
```

```
  rpart(
    fourcat ~ total_enrollment + dif_white + dif_hispa + dif_black + dif_asian +
      dif_amind + dif_pacis + dif_twora,
    data = fryr.cllg.17,
    control = rpart.control(cp = 0.05)
  )

# Graphs the decision tree for the above model
par(xpd = TRUE)

plot(fryr.cllg.tree.3, compress = TRUE)
text(fryr.cllg.tree.3, use.n = TRUE)
par(xpd = FALSE)

# Provides a summary of the model above
summary(fryr.cllg.tree.3)

# Predictions to add to data frame
preds <- predict(fryr.cllg.tree.3, type = "class")

# Adds prediction to the data frame
fryr.cllg.17thr <- mutate(fryr.cllg.17, predType = preds)

# Shows the accuracy of the predictions compared to the true data
accuracy(
  data = fryr.cllg.17thr,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Shows predictions and true values in a table
conf_mat(data = fryr.cllg.17thr, truth = fourcat, estimate = predType)

# Using shown variables from each decision tree with the best Pr values
fryr.cllg.tree.4 <-
  rpart(
    fourcat ~ total_enrollment + dif_pacis + mkt_pacis,
    data = fryr.cllg.17,
    control = rpart.control(cp = 0.05)
  )

# Plotting of the decision tree above
par(xpd = TRUE)

plot(fryr.cllg.tree.4, compress = TRUE)
text(fryr.cllg.tree.4, use.n = TRUE)
par(xpd = FALSE)

# Summarizes the decision tree, will obtain the equation of given tree
summary(fryr.cllg.tree.4)

# To be used to add predictions to data frame
preds <- predict(fryr.cllg.tree.4, type = "class")
```

```
# adds predictions to data frame
fryr.cllg.17for <- mutate(fryr.cllg.17, predType = preds)

# Shows the accuracy of the predictions to the true values
accuracy(
  data = fryr.cllg.17for,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Displays predictions and true values in a table
conf_mat(data = fryr.cllg.17for, truth = fourcat, estimate = predType)



################################  Random Forest  #################################

# All possible variables to predict fourcat with using a max of three of the
# explanatory variables in each tree
fryr.cllg.forest <-
  randomForest(
    as.factor(fourcat) ~ total_enrollment + col_white + col_hispa + col_black +
      col_asian + col_amind + col_pacis + col_twora + mkt_white + mkt_hispa +
      mkt_black + mkt_asian + mkt_amind + mkt_pacis + mkt_twora + dif_white +
      dif_hispa + dif_black + dif_asian + dif_amind + dif_pacis + dif_twora,
    data = fryr.cllg.17,
    ntree = 5000,
    mtry = 3
  )

# to add predictions to data frame
forest.preds <- predict(fryr.cllg.forest, type = "class")

# adds predictions to data frame
fryr.cllg.17fiv <- mutate(fryr.cllg.17, predType = forest.preds)

# Gives the accuracy of the predictions
accuracy(
  data = fryr.cllg.17fiv,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Shows the predictions and true values in table
conf_mat(data = fryr.cllg.17fiv, truth = fourcat, estimate = predType)

importance(fryr.cllg.forest)


# Most important variables based on above random forest models
fryr.cllg.forest1 <-
  randomForest(
    as.factor(fourcat) ~ total_enrollment + mkt_asian + mkt_amind,
    data = fryr.cllg.17,
    ntree = 1000,
    mtry = 2
  )

# predictions to add to data frame
forest.preds1 <- predict(fryr.cllg.forest1, type = "class")

# adds predictions to data frame
fryr.cllg.17six <- mutate(fryr.cllg.17, predType = forest.preds1)

# Shows the accuracy of the predictions based on the true values
accuracy(
  data = fryr.cllg.17six,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Displays predictions in a table with true values
conf_mat(data = fryr.cllg.17six, truth = fourcat, estimate = predType)

# Will show the most important variables
importance(fryr.cllg.forest1)


##############################  K Nearest Neighbor  ############################
set.seed(123)

# College percent nearest neighbor with 7 nearest neighbors. Uses explanatory
# variables such as total enrollment and col_* to predict type of college
# (fourcat)
fryr.cllg.knn <-
  kknn(
    as.factor(fourcat) ~ total_enrollment + col_white + col_hispa +
      col_black + col_asian + col_amind + col_pacis + col_twora,
    train = fryr.cllg.17,
    test = fryr.cllg.17,
    k = 7
  )

# Gets the predictions
knn.preds <- fitted(fryr.cllg.knn)

# Add predictions to data frame
fryr.cllg.17svn <- mutate(fryr.cllg.17, predType = knn.preds)

# Shows the accuracy of the model based on the predictions made
accuracy(
  data = fryr.cllg.17svn,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
  )
```

```r
# Market percent nearest neighbor with 7 nearest neighbors. Uses explanatory
# variables such as total enrollment and col_* to predict type of college
# (fourcat)
fryr.cllg.knn1 <-
  kknn(
    as.factor(fourcat) ~ total_enrollment + mkt_white + mkt_hispa +
      mkt_black + mkt_asian + mkt_amind + mkt_pacis + mkt_twora,
    train = fryr.cllg.17,
    test = fryr.cllg.17,
    k = 7
  )

# Gets the predictions
knn.preds <- fitted(fryr.cllg.knn1)

# Add predictions to data frame
fryr.cllg.17eght <- mutate(fryr.cllg.17, predType = knn.preds)

# Shows the accuracy of the model based on the predictions made
accuracy(
  data = fryr.cllg.17eght,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)

# Difference percent nearest neighbor with 7 nearest neighbors. Uses explanatory
# variables such as total enrollment and col_* to predict type of college
# (fourcat)
fryr.cllg.knn2 <-
  kknn(
    as.factor(fourcat) ~ total_enrollment + dif_white + dif_hispa +
      dif_black + dif_asian + dif_amind + dif_pacis + dif_twora,
    train = fryr.cllg.17,
    test = fryr.cllg.17,
    k = 7
  )

# Gets the predictions
knn.preds <- fitted(fryr.cllg.knn2)

# Add predictions to data frame
fryr.cllg.17nine <- mutate(fryr.cllg.17, predType = knn.preds)

# Shows the accuracy of the model based on the predictions made
accuracy(
  data = fryr.cllg.17nine,
  truth = as.factor(fourcat),
  estimate = as.factor(predType)
)
```