California Community College Completion Project

Aeron Zentner

04/01/2020

Contents

Introduction

The state of California has the largest two-year college system with 114 colleges and has a set of focused goals to increase degree and certificate attainment and transfer to four-year universities with the end goal to support economic and social mobility. To ensure that colleges are making progress towards meeting these goals the California Community College Chancellor's Office (CCCCO) has established a set of accountability metrics. These metrics focus on a range of student achievement and institutional effectiveness outcomes. The outcomes are centralized around degree and certificate attainment and transfer to four-year universities and are noted in the system as the completion rate metric.

Research Question

The research question of the study was "What institutional characteristics, student characteristics, and academic performance outcomes influence completion rates?"

The aim of the study was to develop a model to determine which combination of variables could best predict completion rates. The completion rate was based on the percentage of degree, certificate and/or transfer-seeking students starting first-time in 2011-12 tracked for six years through 2016-17 who completed a degree, certificate or transfer-related outcomes.

Data

All the data was collect from the various data calls from the (CCCCO) public data mart (https://datamart.ccco.edu/DataMart.aspx) on March 29 and 30, 2020. The comprehensive data file of college data, student characteristics, and courses outcomes is stored in Github and includes that data that contributed the raw data for the model. All data was anonymized to avoid any bias in the study. Note only 113 of the 114 college were included as one college did not have data reporting for the completion rate.

The data variables of the study include college size, annual unduplicated headcount, full-time faculty, student race, student sex, traditional students, course load, financial aid, course success and retention.

College Size, Headcount, and Full-time Faculty

College size (size) was tabulated based on the CCCCO's framework associated the amount of full-time equivalent students (FTES) per institution. The data was collected over the six-year period from July 2011 to June 2017 to determine the average FTES per college. The colleges were categorized into three sizes (Small (1) = < 10,000 FTES, Medium (2) 10,000-20,000 FTES; Large (3) >20,000 FTES).

College unduplicated headcount (hc) represents the unduplicated number of students which attended a college during an academic year (July 1 to June 30) from July 2011 to June 2017.The data was collected from the CCCCO data mart in an aggregated form and synthesized by mean annual headcount prior to entering the data into the model.

The data was collected from the CCCCO data mart over the six-year period from fall 2011 to fall 2016 and calculated to determine the average full-time faculty (ftf) per college per year during the timeframe.

Student Characteristics

Student race data was collected for the 2011 cohort year and utilized the CCCCO's establish race categories of African American/ black (aabl), American Indian/ Alaskan Native (aiak), Asian/ Filipino (asian), Hispanic (hisp), Pacific Islander (paci) + Multiple Ethnicity / Other (multi), and white non-Hispanic (white). The data was calculated based on proportion of unduplicated headcount by race for the cohort year.

Student sex data was collected for the 2011 cohort year and utilized the CCCCO's establish sex categories of female (female), male (male), and other/unknown (xgender). The data was calculated based on proportion of unduplicated headcount by sex for the cohort year.

Traditional student data was collected for the 2011 cohort year and utilized the CCCCO's establish age categories to determine the number of students which met the definition of being 24 or younger (trad). The data was calculated based on proportion of unduplicated headcount by traditional students for the cohort year.

Full-time equivalent student course load proportions (ftesps) was collected form the annual full-time equivalent student (FTES) number divided by unduplicated headcount to estimate the average ftesps over the six-year period.

Financial Aid (finaid) data was from the CCCCO data mart under the California Promise tuition waiver grant collected. The California Promise uses an income-based criterion to award the grants and spans further than Title IV federal financial aid. The variable is typically used for flagging low-income students in other statewide projects. The data was calculated based on proportion of unduplicated headcount by California Promise grant recipients across the six-year period.

Course Outcomes

Course success rates (s1117) were collected from the student outcomes data sets in the CCCCO data mart. Course success is calculated based on the number of students that receive a passing grade (A, B, C, Pass) out of the total student population. This metric was calculated across the six-year period to determine and overall rate of success by college during the timeframe.

Course retention rates (r1117) were collected from the student outcomes data sets in the CCCCO data mart. Course retention is calculated based on the number of students that receive a grade (A, B, C, D, F, No Pass, Pass) other than a withdraw (W) grade out of the total student population. This metric was calculated across the six-year period to determine and overall rate of retention by college during the timeframe.

Project Setup

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## Warning: package 'ggplot2' was built under R version 3.6.3
## Warning: package 'dplyr' was built under R version 3.6.3
## Warning: package 'forcats' was built under R version 3.6.3
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.6.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
## Loading required package: lattice
##
## Attaching package: 'caret'
## The following objects are masked from 'package:DescTools':
##
##     MAE, RMSE
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.6.3
##
## Attaching package: 'data.table'
## The following object is masked from 'package:DescTools':
##
##     %like%
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:DescTools':
##
##     Recode

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
library(readxl)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.6.3

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Data Import and Partitioning

The original data was stored in a Microsoft Excel file data imported into R. The data was primarily categorized as double date with the exception to college id which was categorical.

A summary of the data was created to review the means and data distribution. Additionally, the standard deviation for completion rate (ycomp) was calculated along with a scatter plot of the ycomp outcomes.

```
# Calculate summary data for the overall dataset
summary(College)
```

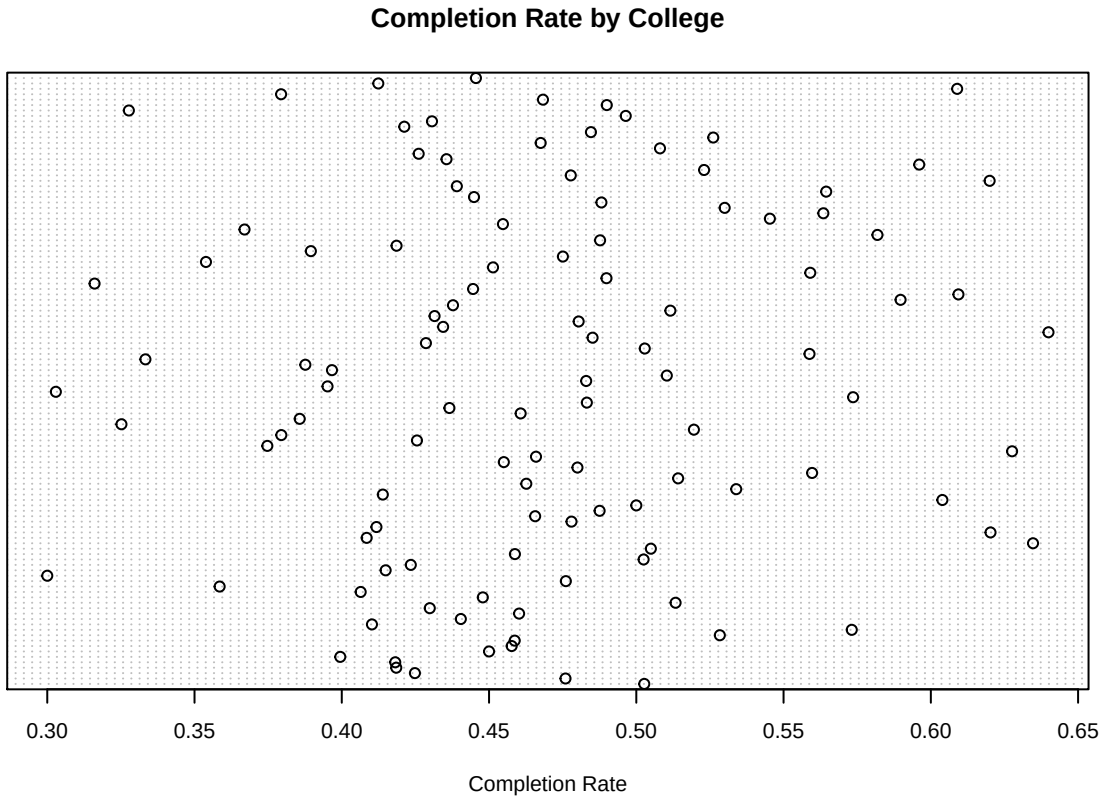
```
##  CollegeCode             size             hc            fteps
##  Length:113         Min.   :1.000   Min.   : 2789   Min.   :0.2583
##  Class :character   1st Qu.:1.000   1st Qu.:11651   1st Qu.:0.4551
##  Mode  :character   Median :1.000   Median :16698   Median :0.5137
##                     Mean   :1.504   Mean   :20147   Mean   :0.5076
##                     3rd Qu.:2.000   3rd Qu.:27381   3rd Qu.:0.5696
##                     Max.   :3.000   Max.   :72992   Max.   :0.6838
##       trad            finaid           female            male
##  Min.   :0.2012   Min.   :0.1929   Min.   :0.2021   Min.   :0.3098
##  1st Qu.:0.5002   1st Qu.:0.4211   1st Qu.:0.5153   1st Qu.:0.4187
##  Median :0.5771   Median :0.5353   Median :0.5442   Median :0.4464
##  Mean   :0.5603   Mean   :0.5198   Mean   :0.5322   Mean   :0.4568
##  3rd Qu.:0.6252   3rd Qu.:0.6269   3rd Qu.:0.5693   3rd Qu.:0.4735
##  Max.   :0.7476   Max.   :0.7880   Max.   :0.6877   Max.   :0.7966
##     xgender             aabl               aiak              asian
##  Min.   :0.000000   Min.   :0.006314   Min.   :0.0003007   Min.   :0.006615
##  1st Qu.:0.002153   1st Qu.:0.031987   1st Qu.:0.0030487   1st Qu.:0.054541
##  Median :0.008689   Median :0.049043   Median :0.0045973   Median :0.090909
##  Mean   :0.010940   Mean   :0.080693   Mean   :0.0069770   Mean   :0.125125
##  3rd Qu.:0.014189   3rd Qu.:0.094551   3rd Qu.:0.0075244   3rd Qu.:0.166729
##  Max.   :0.090892   Max.   :0.490852   Max.   :0.0713421   Max.   :0.453207
```

```
##       hisp              paci                multi              white
##  Min.   :0.1004   Min.   :0.0001002   Min.   :0.02313   Min.   :0.01603
##  1st Qu.:0.2295   1st Qu.:0.0029047   1st Qu.:0.06398   1st Qu.:0.21861
##  Median :0.3317   Median :0.0044888   Median :0.08672   Median :0.30903
##  Mean   :0.3522   Mean   :0.0056140   Mean   :0.09048   Mean   :0.33888
##  3rd Qu.:0.4668   3rd Qu.:0.0065792   3rd Qu.:0.10864   3rd Qu.:0.45180
##  Max.   :0.9065   Max.   :0.0213264   Max.   :0.19883   Max.   :0.72841
##      r1117             s1117              ftf              ycomp
##  Min.   :0.7886   Min.   :0.6249   Min.   :0.1521   Min.   :0.3000
##  1st Qu.:0.8605   1st Qu.:0.6987   1st Qu.:0.2498   1st Qu.:0.4212
##  Median :0.8701   Median :0.7196   Median :0.3042   Median :0.4626
##  Mean   :0.8736   Mean   :0.7161   Mean   :0.3026   Mean   :0.4688
##  3rd Qu.:0.8879   3rd Qu.:0.7333   3rd Qu.:0.3460   3rd Qu.:0.5104
##  Max.   :0.9515   Max.   :0.8042   Max.   :0.5071   Max.   :0.6400
```

```r
# Calculate standard deviation for completion rate (ycomp)
sd(College$ycomp)
```

```
## [1] 0.07496524
```

```r
# Create a scatter plot of the completion rate data (ycomp)
dotchart(College$ycomp,labels=row.names(College$CollegeCode),cex=.7,
         main="Completion Rate by College",
         xlab="Completion Rate")
```



The data was split into a 25/75 for training and testing data. The data was also reviewed and validated to ensure the partitioning was correct.

```r
# Training set will be 25% of College dataset
set.seed(3, sample.kind="Rounding")
```

```
## Warning in set.seed(3, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```r
test_index <- createDataPartition(y = College$ycomp, times = 1, p = 0.25, list = FALSE)
testing<- College[-test_index,]
training<- College[test_index,]

# Validate the data in the training set
dim(training)
```

```
## [1] 29 20
```

```r
# Validate the data in the testing set
dim(testing)
```

```
## [1] 84 20
```

Model Development

Seven multivariate models were developed to measure the level of influence that college data, student characteristics, and courses outcomes have on the predictability of completion rates.

Model 1

Model 1 assessed how college size, headcount and traditional student proportions predicted completion rates.

```r
# Set up multivariate model 1 with college size, headcount and traditional student proportions
mod1<-lm(ycomp ~ size + hc + trad, data = training)

# Execute model 1
summary(mod1)
```

```
##
## Call:
## lm(formula = ycomp ~ size + hc + trad, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.112609 -0.049561 -0.009515  0.032539  0.130152
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.689e-01  7.107e-02   3.784 0.000862 ***
## size        -2.721e-02  3.811e-02  -0.714 0.481872
## hc           2.922e-06  1.833e-06   1.594 0.123430
## trad         3.082e-01  1.218e-01   2.531 0.018037 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0651 on 25 degrees of freedom
## Multiple R-squared:  0.3021, Adjusted R-squared:  0.2183
## F-statistic: 3.607 on 3 and 25 DF,  p-value: 0.0272
```

```r
# Set up prediction for model 1
predictions1<- predict(mod1, testing)

# Create a residual map for model 1
resid(mod1)
```

```
##             1             2             3             4             5
##  0.0523526836 -0.0089750808  0.0224138581 -0.0002519854 -0.1126089521
##             6             7             8             9            10
##  0.0126465637  0.1301522855 -0.0592326056 -0.0369648158  0.0186891463
##            11            12            13            14            15
## -0.0841503583  0.0255126118 -0.0510374818 -0.0502447549 -0.0120899728
##            16            17            18            19            20
## -0.0365957725  0.0997708099 -0.0095145591 -0.0266843917  0.0719347677
##            21            22            23            24            25
## -0.0156348457 -0.0376180869 -0.0745786411  0.0479341971  0.1276910985
##            26            27            28            29
## -0.0495611861 -0.0501040620  0.0325389324  0.0742105977
```

```r
# Calculate RMSE for model 1
RMSE(testing$ycomp, predictions1)
```

```
## [1] 0.07418886
```

```r
#Assess model 1 accuracy
sigma(mod1)/mean(College$ycomp)
```

```
## [1] 0.1388805
```

Model 2

Model 2 assessed how college size, headcount, traditional student proportions, and full-time equivalent student course load proportions predicted completion rates.

```r
#Set up multivariate model 2 by adding ft equivalent status per student
mod2<-lm(ycomp ~ size + hc + trad + fteps, data = training)

#Execute model 2
summary(mod2)
```

```
##
## Call:
## lm(formula = ycomp ~ size + hc + trad + fteps, data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10863 -0.04231 -0.01101  0.03697  0.13508
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.865e-01  8.614e-02   3.325  0.00283 **
## size        -2.142e-02  4.173e-02  -0.513  0.61242
## hc           2.597e-06  2.056e-06   1.263  0.21880
## trad         3.497e-01  1.661e-01   2.105  0.04593 *
## fteps       -8.571e-02  2.285e-01  -0.375  0.71094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.06625 on 24 degrees of freedom
## Multiple R-squared:  0.3062, Adjusted R-squared:  0.1905
## F-statistic: 2.647 on 4 and 24 DF,  p-value: 0.05819
```

```
#Set up prediction for model 2
predictions2<- predict(mod2, testing)

#Create a residual map for model 2
resid(mod2)
```

```
##           1           2           3           4           5           6
##  0.040541195 -0.006590544  0.023856133 -0.006200715 -0.108633068  0.013386859
##           7           8           9          10          11          12
##  0.135076131 -0.052723505 -0.035184028  0.017433282 -0.083124683  0.026983713
##          13          14          15          16          17          18
## -0.056761718 -0.051162068 -0.020104990 -0.034549252  0.095857962 -0.011010479
##          19          20          21          22          23          24
## -0.028816380  0.080378056 -0.017046710 -0.039058959 -0.077617351  0.044080937
##          25          26          27          28          29
##  0.124366665 -0.046901142 -0.042315175  0.036973196  0.078866640
```

```
#Calculate RMSE for model 2
RMSE(testing$ycomp, predictions2)
```

```
## [1] 0.07483938
```

```
# Assess model 2 accuracy
sigma(mod2)/mean(College$ycomp)
```

```
## [1] 0.1413308
```

Model 3

Model 3 assessed how college size, headcount, traditional student proportions, full-time equivalent student course load proportions, course success rates, and retention rates predicted completion rates.

```
# Set up multivariate model 3 by adding course success and retention rates
mod3<-lm(ycomp ~ size + + hc + trad + fteps + s1117 + r1117, data = training)

# Execute model 3
summary(mod3)
```

```
##
## Call:
## lm(formula = ycomp ~ size + +hc + trad + fteps + s1117 + r1117,
##     data = training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10060 -0.01376 -0.00430  0.02108  0.06924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.170e+00  5.636e-01  -2.075 0.049874 *
## size         4.332e-02  3.965e-02   1.093 0.286380
## hc          -1.290e-06  1.979e-06  -0.652 0.521253
## trad         7.972e-01  1.770e-01   4.505 0.000176 ***
```

```
## fteps       -4.642e-01  2.284e-01  -2.033 0.054331 .
## s1117        2.470e+00  5.668e-01   4.357 0.000252 ***
## r1117       -4.614e-01  5.048e-01  -0.914 0.370600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05065 on 22 degrees of freedom
## Multiple R-squared:  0.6282, Adjusted R-squared:  0.5268
## F-statistic: 6.196 on 6 and 22 DF,  p-value: 0.0006368
```

```
# Set up prediction for model 3
predictions3<- predict(mod3, testing)

# Create a residual map for model 3
resid(mod3)
```

```
##             1             2             3             4             5
##  0.0592801870  0.0111079634 -0.0063909488 -0.0895713371 -0.1005973397
##             6             7             8             9            10
## -0.0576288740  0.0654113664 -0.0233291997 -0.0134028600  0.0058659794
##            11            12            13            14            15
## -0.0236662117  0.0210825741 -0.0054594559 -0.0008732717 -0.0137570940
##            16            17            18            19            20
##  0.0105172327  0.0546663621 -0.0128830496 -0.0085733852  0.0601788210
##            21            22            23            24            25
## -0.0508150739  0.0028860084 -0.0122196131  0.0450726746  0.0668332138
##            26            27            28            29
## -0.0620662694  0.0133917477 -0.0042996281  0.0692394813
```

```
# Calculate RMSE for model 3
RMSE(testing$ycomp, predictions3)
```

```
## [1] 0.08470107
```

```
# Assess model 3 accuracy
sigma(mod3)/mean(College$ycomp)
```

```
## [1] 0.1080531
```

Model 4

Model 4 assessed how college size, headcount, traditional student proportions, full-time equivalent student course load proportions, course success rates, retention rates and financial aid proportions predicted completion rates.

```
#Set up multivariate model 4 by adding proportion of financial aid
mod4<-lm(ycomp ~ size + + hc + trad + fteps + s1117 + r1117 + finaid, data = training)

#Execute model 4
summary(mod4)
```

```
##
## Call:
## lm(formula = ycomp ~ size + +hc + trad + fteps + s1117 + r1117 +
##     finaid, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.090937 -0.019547 -0.001096  0.032178  0.075649
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.548e-01  6.039e-01  -1.415 0.171583
## size         3.385e-02  3.966e-02   0.853 0.403076
## hc          -1.337e-06  1.947e-06  -0.687 0.499894
## trad         7.747e-01  1.750e-01   4.428 0.000234 ***
## fteps       -4.333e-01  2.259e-01  -1.918 0.068793 .
## s1117        1.920e+00  6.966e-01   2.757 0.011821 *
## r1117       -2.693e-01  5.176e-01  -0.520 0.608357
## finaid      -1.363e-01  1.036e-01  -1.316 0.202441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04983 on 21 degrees of freedom
## Multiple R-squared:  0.6565, Adjusted R-squared:  0.5421
## F-statistic: 5.735 on 7 and 21 DF,  p-value: 0.0008253
```

```
#Set up prediction for model 4
predictions4<- predict(mod4, testing)

#Create a residual map for model 4
resid(mod4)
```

```
##           1           2           3           4           5           6
##  0.053023656  0.032177596 -0.008726844 -0.088769304 -0.090937393 -0.072568909
##           7           8           9          10          11          12
##  0.060798801 -0.023550629 -0.012263485 -0.001096014 -0.015535433  0.022178886
##          13          14          15          16          17          18
## -0.019547039  0.004213722 -0.004833488  0.016298634  0.041146754 -0.004166534
##          19          20          21          22          23          24
## -0.004371577  0.051470191 -0.027699551  0.008506562 -0.041213213  0.054289298
##          25          26          27          28          29
##  0.075648776 -0.061303092  0.014657878  0.004743878  0.037427875
```

```
#Calculate RMSE for model 4
RMSE(testing$ycomp, predictions4)
```

```
## [1] 0.07457377
```

```
#Assess model 4 accuracy
sigma(mod4)/mean(College$ycomp)
```

```
## [1] 0.106301
```

Model 5

Model 5 assessed how college size, headcount, traditional student proportions, full-time equivalent student course load proportions, course success rates, retention rates, financial aid proportions, and student sex proportions predicted completion rates.

```
#Set up multivariate model 5 by adding the proportions of sex
mod5<- lm(ycomp ~ size + + hc + trad + fteps + s1117 + r1117 + finaid + male +female +xgender, data = t

#Execute model 5
summary(mod5)
```

```
##
## Call:
## lm(formula = ycomp ~ size + +hc + trad + fteps + s1117 + r1117 +
##     finaid + male + female + xgender, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079280 -0.017057  0.002325  0.013968  0.075255
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.124e-01  1.006e+00   0.112  0.91222
## size         2.992e-02  3.984e-02   0.751  0.46183
## hc          -1.090e-06  1.988e-06  -0.548  0.59002
## trad         7.927e-01  2.102e-01   3.772  0.00129 **
## fteps       -3.950e-01  2.319e-01  -1.703  0.10480
## s1117        2.042e+00  7.064e-01   2.891  0.00937 **
## r1117       -2.746e-01  5.244e-01  -0.524  0.60657
## finaid      -1.227e-01  1.041e-01  -1.179  0.25308
## male        -1.042e+00  8.832e-01  -1.180  0.25255
## female      -1.143e+00  8.750e-01  -1.307  0.20693
## xgender             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04988 on 19 degrees of freedom
## Multiple R-squared:  0.6887, Adjusted R-squared:  0.5412
## F-statistic:  4.67 on 9 and 19 DF,  p-value: 0.002289
```

```
#set up prediction for model 5
predictions5<- predict(mod5, testing)
```

```
## Warning in predict.lm(mod5, testing): prediction from a rank-deficient fit may
## be misleading
```

```
#Create a residual map for model 5
resid(mod5)
```

```
##             1             2             3             4             5
##  0.0053006180  0.0281386675 -0.0166341907 -0.0778284877 -0.0771148001
##             6             7             8             9            10
## -0.0792796192  0.0513613574 -0.0170567277 -0.0061443321  0.0032314431
##            11            12            13            14            15
## -0.0207656441  0.0089902025 -0.0001085812  0.0137018713  0.0139678536
##            16            17            18            19            20
## -0.0004809077  0.0597070447 -0.0058940991  0.0023253314  0.0494975499
##            21            22            23            24            25
## -0.0210758245  0.0066129054 -0.0404060958  0.0624443836  0.0752551448
##            26            27            28            29
## -0.0657371887  0.0119987970 -0.0060636069  0.0420569352
```

```
#Calculate RMSE for model 5
RMSE(testing$ycomp, predictions5)
```

```
## [1] 0.07430355
```

```r
#Assess model 5 accuracy
sigma(mod5)/mean(College$ycomp)
```

```
## [1] 0.1063986
```

Model 6

Model 6 assessed how college size, headcount, traditional student proportions, full-time equivalent student course load proportions, course success rates, retention rates, financial aid, student sex proportions, and student race proportions predicted completion rates.

```r
#Set up multivariate model 6 by adding the proportions of race
mod6<- lm(ycomp ~ size + + hc + trad + fteps + s1117 + r1117 + finaid + male +female +xgender + aabl + a

#Execute model 6
summary(mod6)
```

```
##
## Call:
## lm(formula = ycomp ~ size + +hc + trad + fteps + s1117 + r1117 +
##     finaid + male + female + xgender + aabl + aiak + asian +
##     hisp + paci + multi + white, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.088332 -0.026158  0.003507  0.024707  0.057181
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.539e-01  1.460e+00  -0.242   0.8123
## size         3.741e-02  4.504e-02   0.831   0.4212
## hc          -9.633e-07  2.423e-06  -0.398   0.6974
## trad         7.482e-01  3.065e-01   2.441   0.0297 *
## fteps       -3.409e-01  2.619e-01  -1.301   0.2157
## s1117        1.869e+00  9.862e-01   1.895   0.0805 .
## r1117       -8.211e-02  5.863e-01  -0.140   0.8908
## finaid      -1.281e-01  1.242e-01  -1.032   0.3210
## male        -7.179e-01  1.276e+00  -0.563   0.5832
## female      -7.201e-01  1.206e+00  -0.597   0.5606
## xgender             NA         NA      NA       NA
## aabl         6.250e-02  2.217e-01   0.282   0.7825
## aiak         5.615e-01  3.008e+00   0.187   0.8548
## asian        1.688e-01  1.522e-01   1.109   0.2876
## hisp        -1.433e-02  1.310e-01  -0.109   0.9146
## paci         4.332e+00  4.139e+00   1.047   0.3143
## multi       -9.788e-02  4.469e-01  -0.219   0.8300
## white               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04961 on 13 degrees of freedom
## Multiple R-squared:  0.7893, Adjusted R-squared:  0.5461
## F-statistic: 3.246 on 15 and 13 DF,  p-value: 0.01961
```

```
#set up prediction for model 6
predictions6<- predict(mod6, testing)
```

```
## Warning in predict.lm(mod6, testing): prediction from a rank-deficient fit may
## be misleading
```

```
#Create a residual map for model 6
resid(mod6)
```

```
##             1             2             3             4             5
##  8.678646e-05  1.435463e-03 -1.408885e-02 -4.832598e-02 -8.833165e-02
##             6             7             8             9            10
## -4.447557e-02 -5.275433e-03  8.109622e-03 -2.984600e-02 -2.615764e-02
##            11            12            13            14            15
## -3.293676e-02  6.674803e-03 -3.795032e-02  2.114531e-02  2.736449e-02
##            16            17            18            19            20
## -5.065399e-03  5.718142e-02  2.075235e-02  5.456243e-03  3.955818e-02
##            21            22            23            24            25
##  3.506753e-03  2.470728e-02 -1.623237e-02  3.913259e-02  4.724542e-02
##            26            27            28            29
## -3.784714e-02  3.397167e-02  1.044405e-02  3.976069e-02
```

```
#Calculate RMSE for Model 6
RMSE(testing$ycomp, predictions5)
```

```
## [1] 0.07430355
```

```
#Assess model 6 accuracy
sigma(mod6)/mean(College$ycomp)
```

```
## [1] 0.1058326
```

Model 7

Model 7 assessed how college size, headcount, traditional student proportions, full-time equivalent student course load proportions, course success rates, retention rates, financial aid, student sex proportions, student race proportions and full-time faculty proportion predicted completion rates.

```
#Set up multivariate model 7 by adding the proportion of full-time faculty
mod7<- lm(ycomp ~ size + + hc + trad + fteps + s1117 + r1117 + finaid + male +female +xgender + aabl + a

#Execute model 7
summary(mod7)
```

```
##
## Call:
## lm(formula = ycomp ~ size + +hc + trad + fteps + s1117 + r1117 +
##     finaid + male + female + xgender + aabl + aiak + asian +
##     hisp + paci + multi + white + ftf, data = training)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.086434 -0.027235  0.001337  0.023205  0.057435
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.632e-01  1.577e+00  -0.294   0.7739
```

```
## size         4.046e-02  4.830e-02   0.838   0.4186
## hc          -1.148e-06  2.620e-06  -0.438   0.6690
## trad         7.913e-01  3.615e-01   2.189   0.0491 *
## fteps       -3.651e-01  2.885e-01  -1.266   0.2297
## s1117        1.967e+00  1.095e+00   1.796   0.0977 .
## r1117       -9.770e-02  6.118e-01  -0.160   0.8758
## finaid      -1.215e-01  1.316e-01  -0.924   0.3738
## male        -6.554e-01  1.347e+00  -0.486   0.6355
## female      -6.858e-01  1.259e+00  -0.545   0.5959
## xgender             NA         NA      NA       NA
## aabl         8.150e-02  2.422e-01   0.336   0.7423
## aiak         7.734e-01  3.234e+00   0.239   0.8150
## asian        1.865e-01  1.729e-01   1.078   0.3021
## hisp        -3.430e-03  1.427e-01  -0.024   0.9812
## paci         3.971e+00  4.530e+00   0.876   0.3980
## multi       -9.104e-02  4.647e-01  -0.196   0.8480
## white               NA         NA      NA       NA
## ftf         -6.534e-02  2.596e-01  -0.252   0.8056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0515 on 12 degrees of freedom
## Multiple R-squared:  0.7904, Adjusted R-squared:  0.5108
## F-statistic: 2.828 on 16 and 12 DF,  p-value: 0.03723
```

```
#set up prediction for model 7
Predictions7<- predict(mod5, testing)
```

```
## Warning in predict.lm(mod5, testing): prediction from a rank-deficient fit may
## be misleading
```

```
#Create a residual map for model 7
resid(mod7)
```

```
##             1             2             3             4             5
##  0.0013374261  0.0011582234 -0.0147534685 -0.0472661083 -0.0864335167
##             6             7             8             9            10
## -0.0436303761 -0.0062148967  0.0034932468 -0.0279799024 -0.0272352001
##            11            12            13            14            15
## -0.0326334677  0.0060723585 -0.0392297780  0.0228776610  0.0288916515
##            16            17            18            19            20
## -0.0050694253  0.0574345436  0.0133202227  0.0007020305  0.0408619525
##            21            22            23            24            25
##  0.0074624136  0.0232053932 -0.0143260344  0.0394815485  0.0481172046
##            26            27            28            29
## -0.0376835605  0.0388288044  0.0097427385  0.0394683152
```

```
#Calculate RMSE for Model 7
RMSE(testing$ycomp, predictions5)
```

```
## [1] 0.07430355
```

```
#Assess model 7 accuracy
sigma(mod7)/mean(College$ycomp)
```

```
## [1] 0.1098645
```

Model Results

Based on the assessment of the seven models the result showed that mod7 had the lowest RMSE at 0.07430355. The table below provides the model outputs.While the overall model is considered statatically significant with the greatest influence being associated with tradional student proporion.

Overall Data Analysis

A multivariate analysis was conducted on the overall dataset to determine which variables influenced completion rates. The overall model showed that the higher proportionalities and rates of traditional students, course success, and Asian students showed a statistically significant positive impact on completion rates. In contrast, the higher proportionalities of financial aid, African American/ Black and Hispanic student showed a statistically significant negative impact on completion rates.

```
#Set up multivariate model 8 which included all variables in the College overall dataset
mod8<- lm(ycomp ~ size + + hc + trad + fteps + s1117 + r1117 + finaid + male +female +xgender + aabl + a

#Execute model 8
summary(mod8)
```

```
##
## Call:
## lm(formula = ycomp ~ size + +hc + trad + fteps + s1117 + r1117 +
##     finaid + male + female + xgender + aabl + aiak + asian +
##     hisp + paci + multi + white + ftf, data = College)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.120819 -0.024703  0.000797  0.023917  0.109321
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.632e-01  3.619e-01   0.727 0.468818
## size         5.097e-03  1.518e-02   0.336 0.737764
## hc          -7.920e-08  7.735e-07  -0.102 0.918659
## trad         2.618e-01  6.429e-02   4.072 9.59e-05 ***
## fteps        6.864e-02  7.401e-02   0.927 0.356028
## s1117        4.809e-01  2.362e-01   2.036 0.044521 *
## r1117       -2.575e-01  2.394e-01  -1.076 0.284796
## finaid      -2.031e-01  4.165e-02  -4.878 4.25e-06 ***
## male         4.668e-02  3.273e-01   0.143 0.886903
## female       7.245e-02  3.300e-01   0.220 0.826678
## xgender             NA         NA      NA       NA
## aabl        -1.936e-01  6.410e-02  -3.020 0.003235 **
## aiak        -2.233e+00  5.940e-01  -3.760 0.000292 ***
## asian        1.273e-01  5.642e-02   2.257 0.026275 *
## hisp        -1.397e-01  3.827e-02  -3.651 0.000426 ***
## paci        -7.196e-01  1.196e+00  -0.602 0.548779
## multi       -3.517e-02  1.340e-01  -0.263 0.793482
## white               NA         NA      NA       NA
## ftf          5.296e-02  6.374e-02   0.831 0.408147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04147 on 96 degrees of freedom
```

```
## Multiple R-squared:  0.7377, Adjusted R-squared:  0.694
## F-statistic: 16.87 on 16 and 96 DF,  p-value: < 2.2e-16
```

Limitations and Recommendations for Future Research

While the data models provide a variety of factors that can influence completion, the models are limited by the data available. These limitations are also associated with the cultural, social, economic, political, and technological shifts that have happened over the past three years. Therefore, future research should be conducted to include these factors in addition to the movement of CCCCO statewide projects focused on student equity, AB-705 (diminishing all remedial courses), structured guided pathways, and the California student-centered performance funding formula. These recent shifts will have tremendous implications on the two-year colleges and their outcomes.

Conclusion

The aim of the study was to understand which institutional characteristics, student characteristics, and academic performance outcomes influence completion rates. The research used publicly available data to measure a variety of variables over a six-year period to determine which factors were could predict completion rates. Model 7 to have the highest rate of predictability with the lowest RMSE (0.07430355). The overall model (Model 8) provided insight on identifying student characteristics proportionality and course outcome rates, which influence completion rates.