CAPSTONE PROJECT

Saibal Bhattacharya

Oct 2, 2020

DATASET

The dataset used for this capstone project is publicly available and was downloaded from https://public.tableau.com/en-us/s/resources.

The original data was compiled by the Integrated Postsecondary Education Data System (IPEDS). IPEDS serves as the primary source for data on colleges, universities, and technical and vocational postsecondary institutions in the United States, and it is part of the National Center for Education Statistics (NCES).

This dataset pertains to about 1534 US universities and colleges for the year 2013. It includes a host of information regarding institute name, location (state and geographic region), status (public or private not-for-profit), religiously affiliated (or not), historically black college (or not) and type of degrees offered. It also includes many undergraduate admission statistics (number of applications, admissions, yield), 25th and 75th percentile ACT and SAT scores, percent of freshmen submitting ACT and SAT scores, tuition and fees (from 2010 to 2014), in-state and out-of-state total price of attendance (2013-14),full- and part-time enrollment, ethnic/racial make-up, percent of in-state, out-of-state, international students, college endowment per FTE (full-time equvalent) enrollment, and percent of freshmen receiving various kinds of financial aid - local, federal, and institutional. Additionally, it also includes the graduation rate (for Bachelor's degree) over 4, 5, and 6 years.

Additionally, this dataset includes some date that are not related to undergraduates such as ethnic/racial makeup of graduate school and total enrollment numbers across the college/university. It also includes some estimated statistics about enrollment for freshmen, undergraduates, graduates, and full- and part-time students.

PROBLEM STATEMENTS

This capstone project attempts to answer the following questions:

1. Can a quick visualization of the dataset be generated with Tableau?

2. Can highly ranked colleges/universities be identified using unsupervised machine learning algorithms? NOTE: College/university rankings have been obtained from 2021 US News (https://www.usnews.com/best-colleges).

3. Can Random Forest and Neural Network regression be used to identify features that affect the 4-yr graduation rate? Compare results from two methods.

4. Can Random Forest regression and Neural Network regression be used to identify parameters that affect the increase in graduation rates between 4 and 6 years? Compare results from two methods.

What does the data look like?

ID_number	Name	Yr	ZIP	County	Longitude	Latitude	Relgious_y_n
100654	Alabama A & M University	2013	35762	Madison County	-86.56850	34.78337	No
100663	University of Alabama at Birmingham	2013	35294-0110	Jefferson County	-86.80917	33.50223	No
100690	Amridge University	2013	36117-3553	Montgomery County	-86.17401	32.36261	Yes
100706	University of Alabama in Huntsville	2013	35899	Madison County	-86.63842	34.72282	No
100724	Alabama State University	2013	36104-0271	Montgomery County	-86.29568	32.36432	No

Applicants	Admission	UG1st_Enrolled	P_submit_SAT	P_submit_ACT	SATRead_25	SATRead_75	SATMath_25	SATMath_75
6142	5521	1104	15	88	370	450	350	450
5689	4934	1773	6	93	520	640	520	650
NA	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
2054	1656	651	34	94	510	640	510	650
10245	5251	1479	18	87	380	480	370	480

SATWrite_25	SATWrite_75	ACTComp_25	ACTComp_75	P_admit	Yield	Tuition2010_11	Tuition2011_12	Tuition2012_13
NA	*NA*	15	19	90	20	5800	6828	7182
NA	*NA*	22	28	87	36	5806	6264	6798
NA	*NA*	*NA*	*NA*	*NA*	*NA*	8360	8720	6800
NA	*NA*	23	29	81	39	7492	8094	8794
NA	*NA*	15	19	51	28	7164	8082	7932

Tuition2013_14	Price_instate_2013_14	Price_outstate_2013_14	State	Region	Status	HBCU	Urbanization
7182	21849	27441	Alabama	South_East	Public	Yes	City: Midsize
7206	22495	31687	Alabama	South_East	Public	No	City: Midsize
6870	*NA*	*NA*	Alabama	South_East	Private not-for-profit	No	City: Midsize
9192	23466	35780	Alabama	South_East	Public	No	City: Midsize
8720	18286	25222	Alabama	South_East	Public	Yes	City: Midsize

What does the data look like (continued)?

Type_of_univ	UG_enroll	Grad_enroll	FTUG_enroll	PTUG_enroll	P_Amer_Indian	P_Asian	P_Af_American	P_Latino
MS_Large_program	4051	969	3799	252	0	0	95	1
Very_High_research_univ	11502	7066	8357	3145	0	5	26	3
BS_Arts_&_Sciences	322	309	202	120	0	0	42	1
Very_High_research_univ	5696	1680	4237	1459	1	4	13	3
MS_Large_program	5356	719	4872	484	0	0	93	1

P_PIslander	P_White	P_2orM_races	P_RaceNA	P_NRAlien	P_Asian_Native_PIslander	P_Women	P_1stUG_instate	P_1stUG_outstate
0	3	0	1	0	0	51	*NA*	*NA*
0	60	3	1	2	5	58	86	13
0	29	0	27	0	1	61	*NA*	*NA*
0	70	2	3	4	4	44	79	14
0	2	1	1	2	0	59	58	37

P_1stUG_foreign	P_1stUG_resNA	Gradrate_4yrs	Gradrate_5yrs	Gradrate_6yrs	P_1yrUG_any_aid	P_1yrUG_Fed_state_grant
NA	*NA*	10	23	29	97	89
1	0	29	46	53	90	79
NA	*NA*	0	0	67	100	90
4	3	16	37	48	87	77
4	0	9	19	25	93	87

P_1yrUG_Fed_grant	P_1yrUG_Pell_grant	P_1yrUG_otherFed_grant	P_1yrUG_state_local_grant	P_1yrUG_institute_grant	P_1yrUG_student_loan
81	81	7	1	32	89
36	36	10	0	60	56
90	90	0	40	90	100
31	31	4	1	63	46
76	76	13	11	34	81

What does the data look like (continued)?

P_1yrUG_Fed_student_loan	P_1yrUG_other_loan	SB_Endowment_per_FTE
89	1	*NA*
55	5	24136
100	0	302
46	3	11502
81	0	13202

NOTE: The original data file contained two columns for college/university endowment funds, namely Endowment_per_FTE_(GASB) and Endowment_per_FTE_(FASB). GASB and FASB are two different accounting methods, and so each college/university reported only one of them. No college/university reported both FASB and GASB. Also, there was some formatting issues in some of the rows as some of the data had a $ sign and some didn't. So these two columns were combined into one column named SB_Endowment_per_FTE.FTE stands for full-time equivalent enrollment.

Data details

The starting dataset had 68 columns (including both categorical and numeric) and 1533 rows.

Dictionary – meaning of column headers

	Name of column (in dataset)	Meaning
1	ID_number	
2	**Name**	**Institution name**
3	Yr	
4	ZIP	
5	County	
6	Longitude	
7	Latitude	
8	**Relgious_y_n**	Is institute religiously affiliated?
9	Applicants	Total undergraduate freshmen applicants
10	Admission	Total undergraduate freshmen admitted
11	UG1st_Enrolled	Undergraduate freshmen enrolled
12	P_submit_SAT	Percent of freshmen submitting SAT scores
13	P_submit_ACT	Percent of freshmen submitting ACT scores
14	SATRead_25	SAT Critical Reading 25th percentile score
15	SATRead_75	SAT Critical Reading 75th percentile score
16	SATMath_25	SAT Math 25th percentile score
17	SATMath_75	SAT Math 75th percentile score
18	SATWrite_25	SAT Writing 25th percentile score
19	SATWrite_75	SAT Writing 75th percentile score
20	ACTComp_25	ACT Composite 25th percentile score
21	ACTComp_75	ACT Composite 75th percentile score
22	P_admit	Percent undergrad freshmen admitted
23	Yield	Undergrad freshmen yield
24	Tuition2010_11	Tuition and fees, 2010-11
25	Tuition2011_12	Tuition and fees, 2011-12
26	Tuition2012_13	Tuition and fees, 2012-13
27	Tuition2013_14	Tuition and fees, 2013-14
28	Price_instate_2013_14	Total price for in-state students living on campus 2013-14
29	Price_outstate_2013_14	Total price for out-of-state students living on campus 2013-14
30	**State**	

	Name of column (in dataset)	Meaning
31	**Region**	Geographic region
32	**Status**	Private or Public
33	**HBCU**	Historically Black College or University - Yes or No
34	**Urbanization**	Degree of urbanization (Urban-centric locale)
35	**Type_of_univ**	Carnegie Classification 2010: Basic
36	UG_enroll	Undergraduate enrollment
37	Grad_enroll	Graduate enrollment
38	FTUG_enroll	Full-time undergraduate enrollment
39	PTUG_enroll	Part-time undergraduate enrollment
40	P_Amer_Indian	Percent of undergraduate enrollment that are American Indian or Alaska Native
41	P_Asian	Percent of undergraduate enrollment that are Asian
42	P_Af_American	Percent of undergraduate enrollment that are Black or African American
43	P_Latino	Percent of undergraduate enrollment that are Hispanic/Latino
44	P_PIslander	Percent of undergraduate enrollment that are Native Hawaiian or Other Pacific Islander
45	P_White	Percent of undergraduate enrollment that are White
46	P_2orM_races	Percent of undergraduate enrollment that are two or more races
47	P_RaceNA	Percent of undergraduate enrollment that are Race/ethnicity unknown
48	P_NRAlien	Percent of undergraduate enrollment that are Nonresident Alien
49	P_Asian_Native_PIslander	Percent of undergraduate enrollment that are Asian/Native Hawaiian/Pacific Islander
50	P_Women	Percent of undergraduate enrollment that are women
51	P_1stUG_instate	Percent of first-time undergraduates - in-state
52	P_1stUG_outstate	Percent of first-time undergraduates - out-of-state
53	P_1stUG_foreign	Percent of first-time undergraduates - foreign countries
54	P_1stUG_resNA	Percent of first-time undergraduates - residence unknown
55	**Gradrate_4yrs**	Graduation rate - Bachelor degree within 4 years, total
56	**Gradrate_5yrs**	Graduation rate - Bachelor degree within 5 years, total
57	**Gradrate_6yrs**	Graduation rate - Bachelor degree within 6 years, total
58	P_1yrUG_any_aid	Percent of freshmen receiving any financial aid
59	P_1yrUG_Fed_state_grant	Percent of freshmen receiving federal, state, local or institutional grant aid
60	P_1yrUG_Fed_grant	Percent of freshmen receiving federal grant aid

	Name of column (in dataset)	Meaning
61	P_1yrUG_Pell_grant	Percent of freshmen receiving Pell grants
62	P_1yrUG_otherFed_grant	Percent of freshmen receiving other federal grant aid
63	P_1yrUG_state_local_grant	Percent of freshmen receiving state/local grant aid
64	P_1yrUG_institute_grant	Percent of freshmen receiving institutional grant aid
65	P_1yrUG_student_loan	Percent of freshmen receiving student loan aid
66	P_1yrUG_Fed_student_loan	Percent of freshmen receiving federal student loans
67	P_1yrUG_other_loan	Percent of freshmen receiving other loan aid
68	SB_Endowment_per_FTE	Endowment per full time equivalent enrollment

NOTE: Categorical columns are shown in **Blue**, while the labels are shown in **Red**.

Focus on undergraduate programs - modification of original data

All colleges/universities have undergraduate programs, but graduate programs vary in size and scope across colleges and universities. For this project, I, therefore, decided to focus on answering questions pertaining to the undergraduate program only.

Thus from the original dataset, I removed all data that didn't focus on undergraduates such as ethnic makeup of graduate programs, ethnic makeup of total enrollment (that included non-undergraduates). I also eliminated estimated enrollment numbers (total, undergraduate, and graduate), since the data base separately included (non-estimated) undergraduate enrollment data. I also excluded the number of degrees and certificates awarded (including Associate, Bachelor, Master's, Doctoral, post baccalaureate, post Master's, and various kinds of certificates) mainly because of large number of zero or missing values for these columns.

Focus on ACTComp scores – for Random Forest and Neural Network models

The original dataset contained ACTComp_75 (75th percentile) and ACTComp_25 (25th percentile), SATMath_75, SATMath_25, SATRead_75, SATRead_25, SATWrite_75, and SATWrite_25.

The NN algorithm used in this project can't handle missing data (NAs) in any column. Thus, a decision had to made regarding whether to include all or some of these important inputs: ACTComp_75 (had 334 NAs), ACTComp_25 (had 334 NAs), SATMath_25 (had 351 NAs), SATMath_75 (had 351 NAs), SATRead_25 (had 364 NAs), SATRead_75 (had 364 NAs), SATWrite_25 (had 819 NAs), and SATWrite_75 (had 819 NAs). It was therefore easy to decide to exclude the SATWrite columns due to large number of NA values. Including ACTComp_75, ACTComp_25, SATMath_25, SATMath_75, SATRead_25, SATRead_75 resulted in 1061 cases while including only ACTComp_75, ACTComp_25 resulted in 1151 cases.

The original dataset is a relatively small sample of 1534 cases (including NAs) for NN training and testing purposes. Thus, I decided to include only ACTComp_75, ACTComp_25 so that I could get as big a sample size to train and test. Moreover, ACTComp_25 showed strong linear correlation with SATMath_25 and SATRead_25, and ACTComp_75 showed strong linear correlation with SATMath_75 and SATRead_75.

References:

1. The neural network regression code and workflow was burrowed from http://uc-r.github.io/ann_regression.
2. The neural network regression code and workflow was burrowed from https://uc-r.github.io/random_forests#prereq.
3. The K-means cluster analysis code and workflow was burrowed from: http://uc-r.github.io/kmeans_clustering#elbow.

QUESTION 1

Question 1: Can a quick visualization of the dataset be generated with Tableau?

Answer: Tableau proved to be very effective in visualizing the dataset. It not only provided a quick display of the data, but I used it effectively to identify features that visually appeared to affect the (4-yr) graduation rate.

The Tableau project link is:
https://public.tableau.com/views/Capstone_Saibal/CapstoneStory?:language=en&:display_count=y&publish=yes&:origin=viz_share_link

Observations:

1. A visual look at the data distribution indicates that more colleges/universities in the NE have higher 4-yr graduation rates. Also, to be noted, that the density of colleges is very high in the NE.
2. Public universities and colleges have higher undergraduate enrollment than private (not-for-profit).
3. Colleges/universities with high ACT scores and low admissions rate have higher graduation rates.
4. Graduation rate increases when percent of part time undergraduate students increases – out-of-school responsibilities may delay graduation for these students.
5. Graduation rate falls when percent of undergraduates with Pell grants increase – these students have resource limitations which may delay their graduation.
6. As tuition increases, in both public and private colleges, 4-yr graduation rate increases – high costs are enough of an incentive to have parents, who are most often paying a share of the costs, to ensure that their wards graduate on time.
7. Also, increase in total cost of attendance correlates with higher graduation rates – students and parents don't want to take on additional debt to fund extra years beyond the 4 years.
8. Schools with higher ACT scores tend to have higher graduation rates.
9. ACT Composite 75th percentile scores correlate strongly with SAT 75th percentile scores in Reading, Writing, and Math.
10. ACT Composite 25th percentile scores correlate strongly with SAT 25th percentile scores in Reading, Writing, and Math.

QUESTION 2

Question 2: Can highly ranked colleges/universities be identified using unsupervised machine learning algorithms?

NOTE: College/university rankings have been obtained from 2021 US News (https://www.usnews.com/best-colleges).

Answer: I used K-means cluster analysis to identify the best colleges/universities. From the results of my analysis, it seems that the K-means clustering proved very effective in identifying these top colleges/universities. Not only do the colleges/universities in the top cluster separate out as a stand alone cluster in box and violin plots, but they rank very high on the US News 2021 colleges survey.

Files:

f_CS_2_Cluster.R, RMD_CS_2_Cluster.Rmd, RMD_CS_2_Cluster.html

Procedural steps for K-means cluster analysis:

a. Load data
b. Initial data processing
c. Check for duplication of college/university name
d. Create dummy variables for categorical variables
e. Scale data
f. Initial analysis - K-means clustering with two clusters
g. Define optimal number of clusters
    i. Elbow method
    ii. Average Silhoutte method
    iii. Gap statistic method
h. Final analysis - extracting results
    i. Gap analysis
    ii. Elbow method
i. Compare selected top colleges/universities with US New 2021 Ranks
j. Summary statistics - compare highly selective colleges/universities with others

NOTE: In this analysis, all features (including numeric and categorical) were used.

b. Initial data processing:

I removed the following features: ID_number, ZIP, County, Longitude, and Latitude, because they contain location identifiers for the colleges and universities. Also, all of the data pertains to the year 2013, and so I also removed the Yr feature.

I also removed the following features: P_1stUG_instate, P_1stUG_outstate, P_1stUG_foreign, P_1stUG_resNA, SATWrite_25, and SATWrite_75 because these columns contained a large number of NAs (missing values), and the K-means algorithm that I used can't have columns with missing values.

Originally, the dataset contained 1534 cases, but I was left with 1066 cases after removing all NAs.

c. Remove duplication of college/university names:

College/university names would be used as row names in this algorithm, and so a check was run of names. It was found that a handful of colleges/universities had non-unique names but they were located in different states. Thus, a new column was created that merged the college/university name with the State, and this was used to name the rows.

Features like the college/University Name and State were removed. College location information contained in the State feature gets included in the Region column (which remained in the dataset).

d. Create dummy variables – for all categorical variables

The function one_hot was used to create dummy variables. These dummy variables were renamed with shorter names.

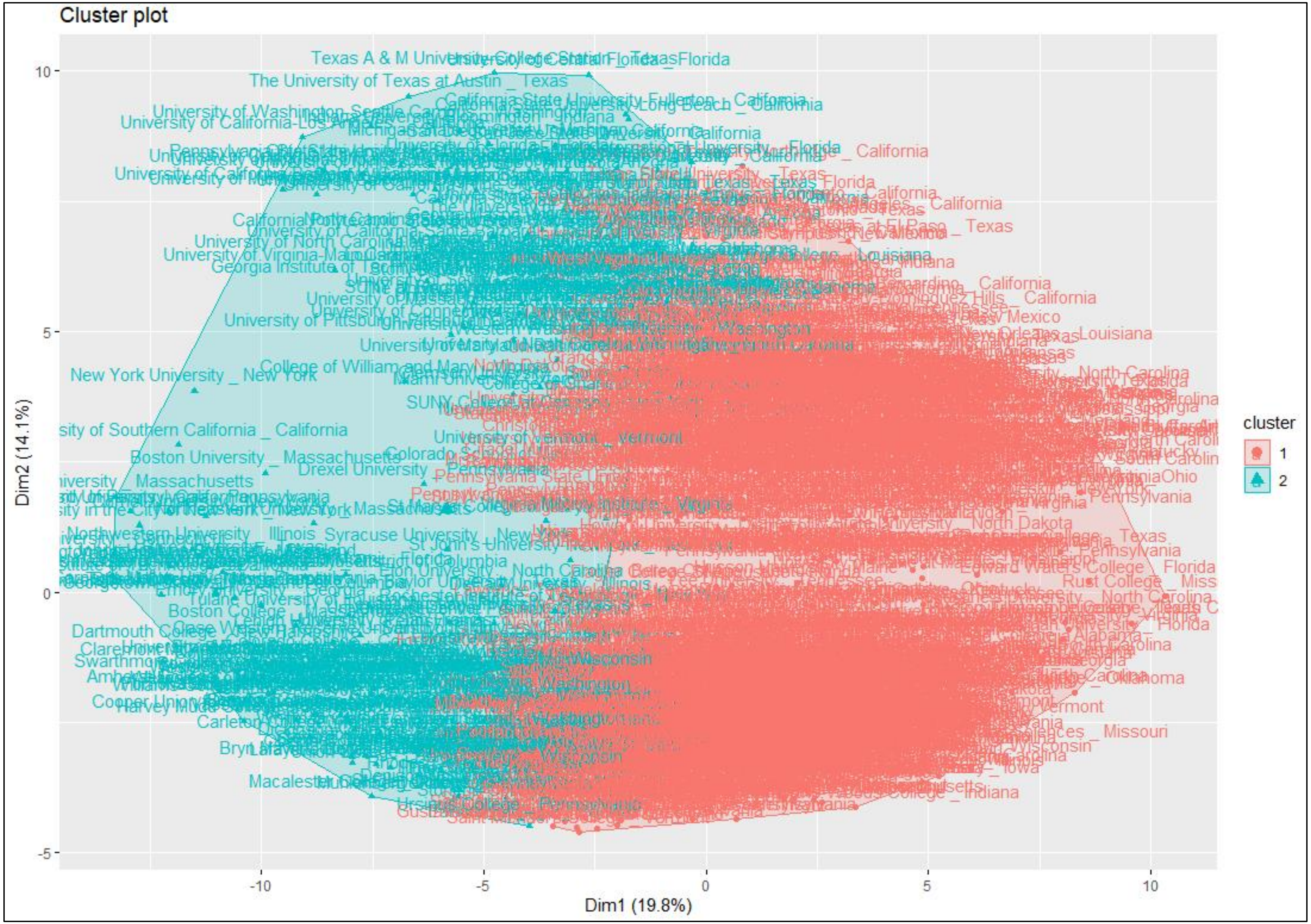
e. Data scaling

All features were scaled using the scale function.

f. Initial analysis - K-means clustering with two clusters

Each cluster (**Figure 2.1**) contained large number of universities - so none of the clusters contained top colleges/universities exclusively.

Figure 2.1


Cluster plot
The University of Texas at Austin _ Texas
New York University _ New York
Boston University _ Massachusetts
Dartmouth College
Edward Waters College _ Florida
cluster
1
2
Dim2 (14.1%)
Dim1 (19.8%)
10
5
0
-5
-10
-5
0
5
10

g. Find optimal number of clusters:

I used three of the most popular methods to find optimal clusters: Elbow & Silhouette methods, and Gap statistics.

g(i). Elbow method:

6 clusters appear to be optimum (**Figure 2.2**) - as the (bend of) the knee occurs near it.

g(ii). Average Silhouette Method

2 clusters appear to be optimum, with 4 clusters coming in as second optimal number of clusters (**Figure 2.3**).

g(iii). Gap analysis

The highest value for gap = 0.959 occurs for cluster = 10. So ten clusters appear to be optimum (**Figure 2.4**).

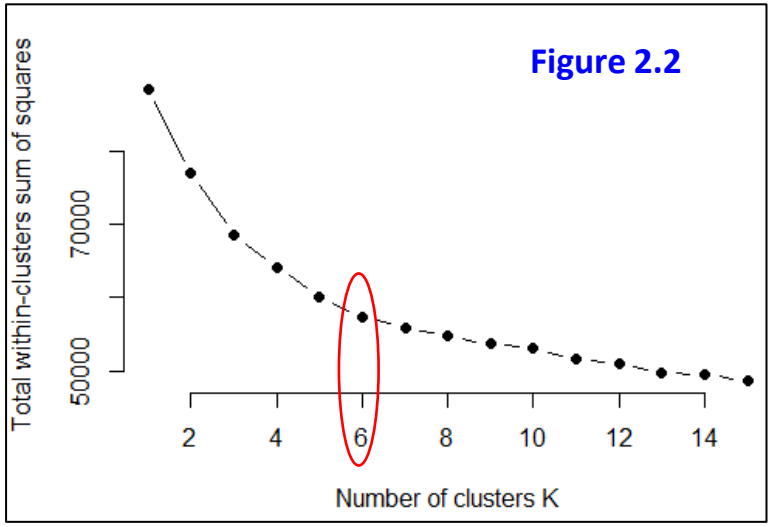


Figure 2.2

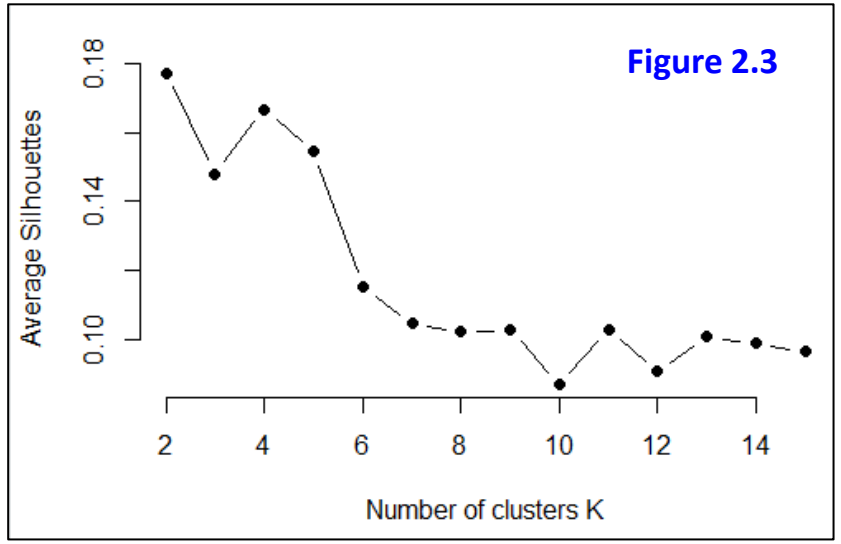


Figure 2.3

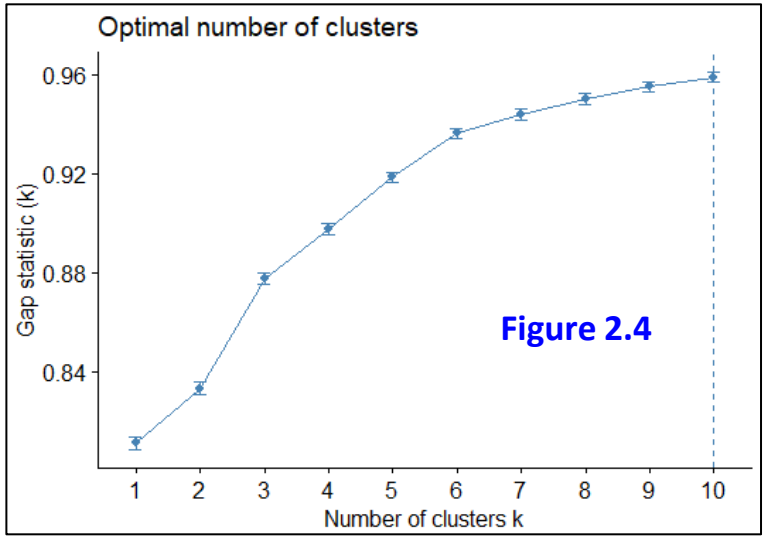


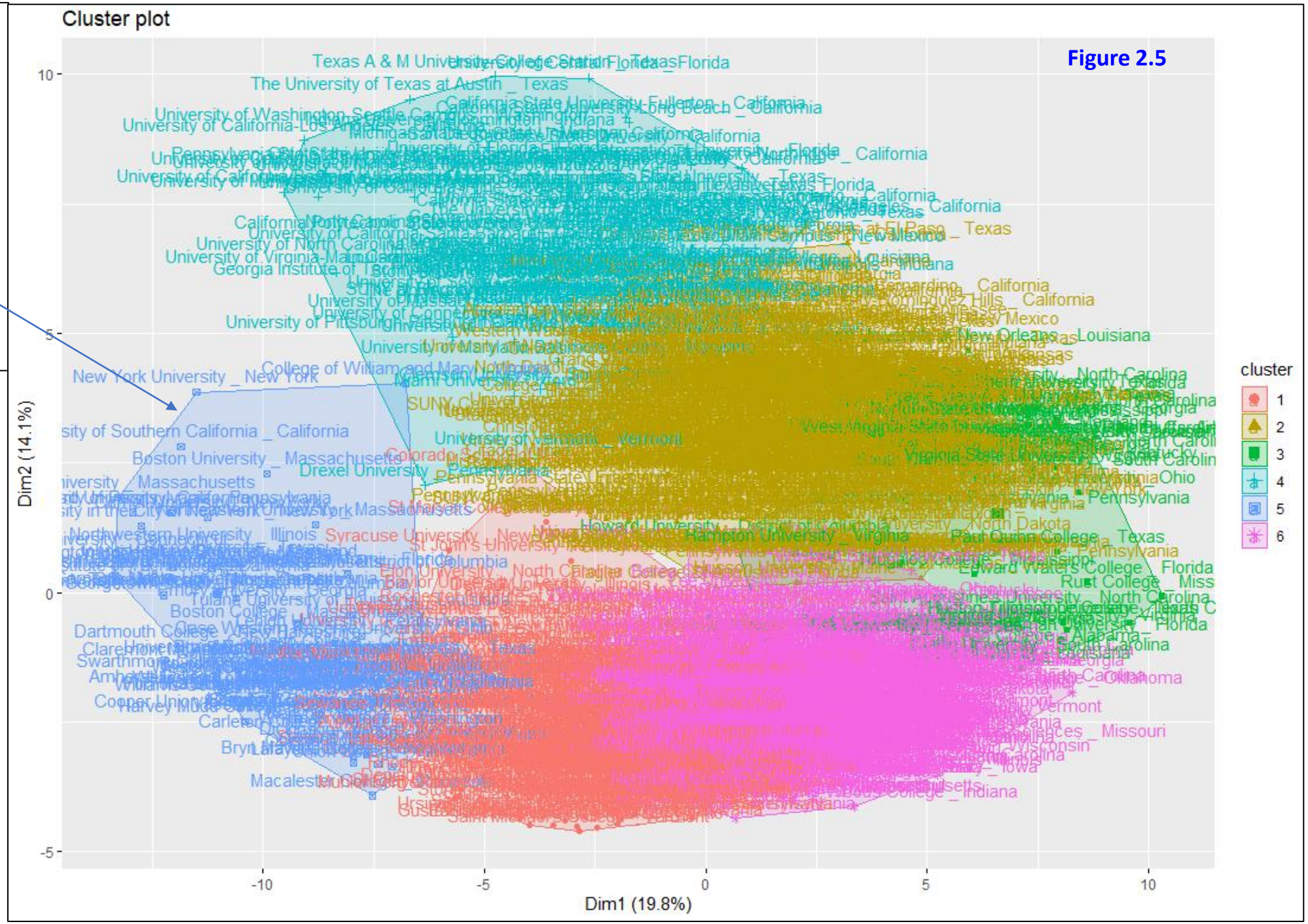
Figure 2.4

h. Final analysis - extracting results

h(i): Elbow method - The top colleges/universities are all in cluster 5, and this cluster has 80 members.

Cluster 5 is highlighted in Blue in **Figure 2. 5**.

Figure 2.5

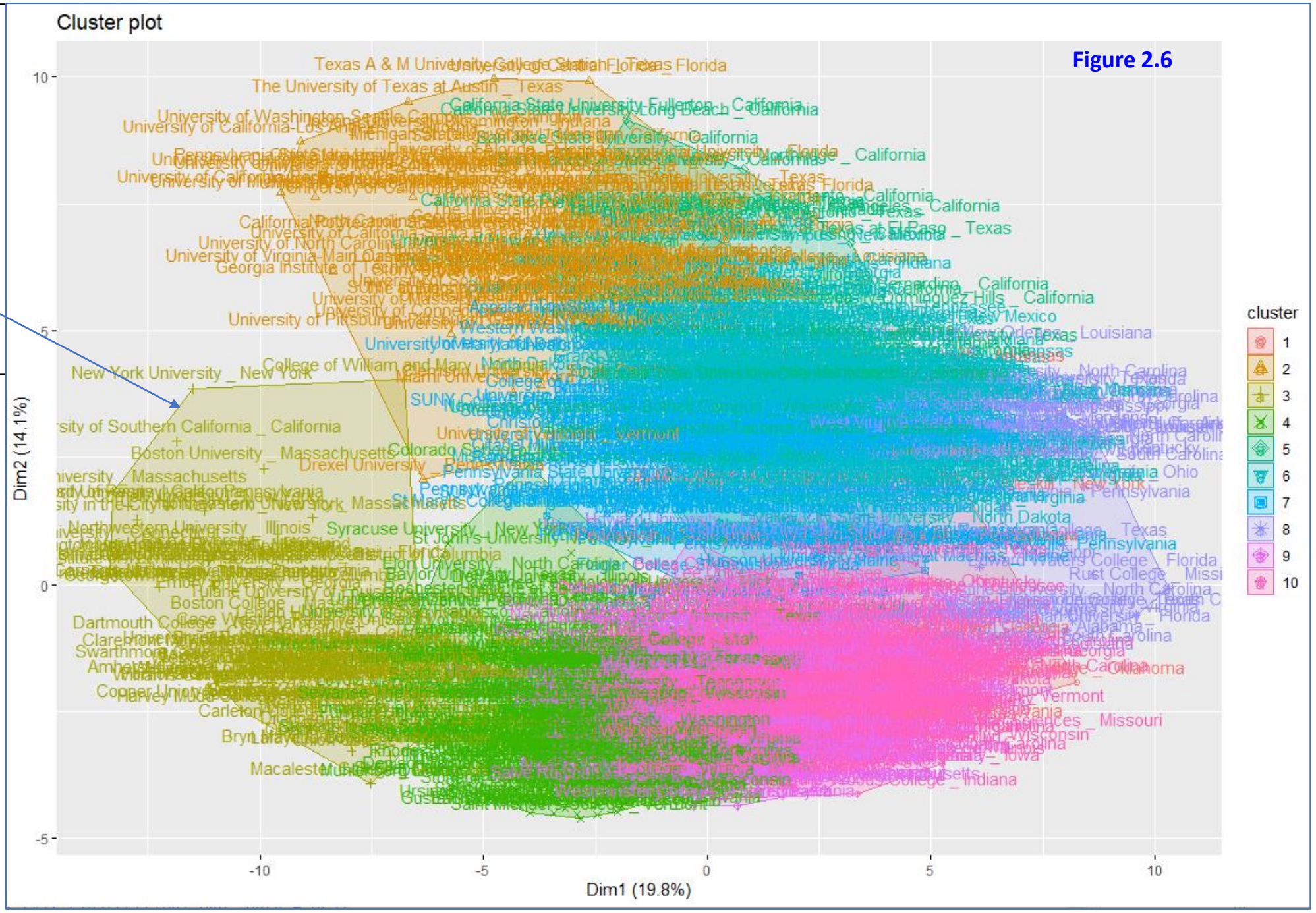


h. Final analysis - extracting results

h(ii): Gap Analysis - The top colleges/universities are all in cluster 3, and this cluster has 80 members.

Cluster 3 is highlighted here in **Figure 2. 6**.

Figure 2.6



i. Compare selected top national universities with US New 2021 Ranks : Both Gap and Elbow analysis show that the cluster containing top colleges/universities have 80 members. The top colleges/universities clustered by the above analyses includes national universities and national liberal arts colleges. **Figure 2.7** shows that Elbow and Gap analyses resulted in the selection of the exactly same set of highly ranked National Universities. US News 2021 ranks for these universities are also displayed, and it shows that the K-means algorithm does a good job in selecting all of the top 19 ranked National Universities. Beyond the top 19 ranked universities, the algorithm additionally selected some more universities between rank 20 and 66, but it did miss out on selecting every university within this range.

This may be indicative of a couple of reasons:

a) The K-means model needs additional tuning to make it select all the missing universities between rank 20 and 66.

b) The top 19 universities are head and shoulders better than the remaining, and it is easy for the algorithm to cluster them together without missing a single one.

Recommendation: Isolate the cluster of top ranked colleges/universities (with 80 members) and perform a new K-means cluster analysis to find a group that just includes the top 20 universities and top 15 colleges.

Figure 2.7

National University - Elbow analysis	US News 2021 Rank	National University - Gap analysis
Princeton University, New Jersey	1	Princeton University, New Jersey
Harvard University, Massachusetts	2	Harvard University, Massachusetts
Columbia University in the City of New York, New York	3	Columbia University in the City of New York, New York
Massachusetts Institute of Technology, Massachusetts	4	Massachusetts Institute of Technology, Massachusetts
Yale University, Connecticut	4	Yale University, Connecticut
Stanford University, California	6	Stanford University, California
University of Chicago, Illinois	6	University of Chicago, Illinois
University of Pennsylvania, Pennsylvania	8	University of Pennsylvania, Pennsylvania
California Institute of Technology, California	9	California Institute of Technology, California
Johns Hopkins University, Maryland	9	Johns Hopkins University, Maryland
Northwestern University, Illinois	9	Northwestern University, Illinois
Duke University, North Carolina	12	Duke University, North Carolina
Dartmouth College, New Hampshire	13	Dartmouth College, New Hampshire
Brown University, Rhode Island	14	Brown University, Rhode Island
Vanderbilt University, Tennessee	14	Vanderbilt University, Tennessee
Rice University, Texas	16	Rice University, Texas
Washington University in St Louis, Missouri	16	Washington University in St Louis, Missouri
Cornell University, New York	18	Cornell University, New York
University of Notre Dame, Indiana	19	University of Notre Dame, Indiana
Emory University, Georgia	21	Emory University, Georgia
Georgetown University, District of Columbia	23	Georgetown University, District of Columbia
University of Southern California, California	24	University of Southern California, California
Carnegie Mellon University, Pennsylvania	26	Carnegie Mellon University, Pennsylvania
New York University, New York	30	New York University, New York
Tufts University, Massachusetts	30	Tufts University, Massachusetts
University of Rochester, New York	34	University of Rochester, New York
Boston College, Massachusetts	35	Boston College, Massachusetts
Tulane University of Louisiana, Louisiana	41	Tulane University of Louisiana, Louisiana
Boston University, Massachusetts	42	Boston University, Massachusetts
Brandeis University, Massachusetts	42	Brandeis University, Massachusetts
Case Western Reserve University, Ohio	42	Case Western Reserve University, Ohio
Lehigh University, Pennsylvania	49	Lehigh University, Pennsylvania
Northeastern University, Massachusetts	49	Northeastern University, Massachusetts
University of Miami, Florida	49	University of Miami, Florida
Rensselaer Polytechnic Institute, New York	53	Rensselaer Polytechnic Institute, New York
Santa Clara University, California	53	Santa Clara University, California
Villanova University, Pennsylvania	53	Villanova University, Pennsylvania
George Washington University, District of Columbia	66	George Washington University, District of Columbia
Southern Methodist University, Texas	66	Southern Methodist University, Texas
Cooper Union for the Advancement of Science and Art, New York	no rank available	Cooper Union for the Advancement of Science and Art, New York

Figure 2.8

i. Compare selected top national liberal arts colleges with US New 2021 Ranks : Both Gap and Elbow analysis show that the cluster containing top colleges/universities have 80 members. The top colleges/universities clustered by the above analyses includes national universities and national liberal arts colleges. **Figure 2.8** shows that Elbow and Gap analyses resulted in the selection of the exactly same set of National Liberal Arts Colleges. US News 2021 ranks for these Liberal Arts Colleges are also displayed, and it shows that the K-means algorithm does a good job in selecting all of the top 15 ranked National Liberal Arts Colleges. Beyond the top 15 ranked universities, the algorithm additionally selected some more colleges between rank 20 and 63, but it miss out on selecting every college within this range.

This may be indicative of a couple of reasons:

a) The K-means model needs additional tuning to make it select all the missing colleges between rank 20 and 63.

b) The top 15 colleges are head and shoulders better than the remaining, and it is easy for the algorithm to cluster them together without missing a single one.

National Liberal Arts Colleges - Elbow analysis	US News 2021 Rank	National Liberal Arts Colleges - Gap analysis
Williams College, Massachusetts	1	Williams College, Massachusetts
Amherst College, Massachusetts	2	Amherst College, Massachusetts
Swarthmore College, Pennsylvania	3	Swarthmore College, Pennsylvania
Pomona College, California	4	Pomona College, California
Wellesley College, Massachusetts	4	Wellesley College, Massachusetts
Claremont McKenna College, California	6	Claremont McKenna College, California
Bowdoin College, Maine	6	Bowdoin College, Maine
Carleton College, Minnesota	9	Carleton College, Minnesota
Hamilton College, New York	9	Hamilton College, New York
Middlebury College, Vermont	9	Middlebury College, Vermont
Washington and Lee University, Virginia	9	Washington and Lee University, Virginia
Grinnell College, Iowa	13	Grinnell College, Iowa
Vassar College, New York	13	Vassar College, New York
Colby College, Maine	15	Colby College, Maine
Davidson College, North Carolina	15	Davidson College, North Carolina
Haverford College, Pennsylvania	15	Haverford College, Pennsylvania
Wesleyan University, Connecticut	20	Wesleyan University, Connecticut
Colgate University, New York	20	Colgate University, New York
Barnard College, New York	22	Barnard College, New York
University of Richmond, Virginia	22	University of Richmond, Virginia
Harvey Mudd College, California	25	Harvey Mudd College, California
Colorado College, Colorado	25	Colorado College, Colorado
Macalester College, Minnesota	27	Macalester College, Minnesota
Scripps College, California	28	Scripps College, California
Kenyon College, Ohio	28	Kenyon College, Ohio
Bryn Mawr College, Pennsylvania	28	Bryn Mawr College, Pennsylvania
Soka University of America, California	28	Soka University of America, California
Bucknell University, Pennsylvania	34	Bucknell University, Pennsylvania
Skidmore College, New York	36	Skidmore College, New York
Oberlin College, Ohio	36	Oberlin College, Ohio
Occidental College, California	40	Occidental College, California
Lafayette College, Pennsylvania	40	Lafayette College, Pennsylvania
Trinity College, Connecticut	44	Trinity College, Connecticut
Union College, New York	44	Union College, New York
Dickinson College, Pennsylvania	47	Dickinson College, Pennsylvania
Whitman College, Washington	47	Whitman College, Washington
Gettysburg College, Pennsylvania	54	Gettysburg College, Pennsylvania
Reed College, Oregon	63	Reed College, Oregon
College of William and Mary, Virginia	no rank available	College of William and Mary, Virginia
Bentley University, Massachusetts	no rank available	Bentley University, Massachusetts

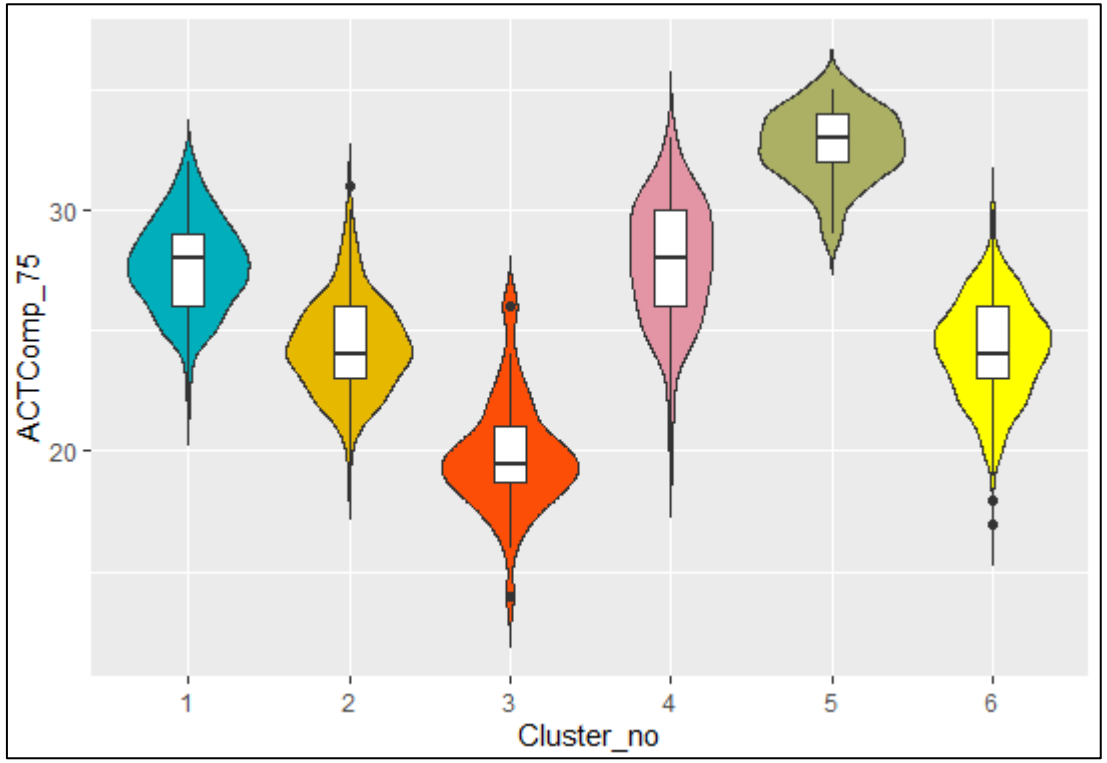
j. Summary statistics - compare highly selective colleges/universities with others:

Both Gap and Elbow analysis show that the cluster containing top colleges/universities have 80 members. So 6 clusters (from Elbow analysis) were selected to generate summary statistics on a few metrices because I found that fewer clusters (than 10 as predicted from the Gap analysis) resulted in distinctly different clusters. Doing the same analysis for 10 clusters (from Gap analysis) resulted in some clusters with overlapping ranges for metric values.

j1. Metric – ACT Composite 75th percentile

	Cluster_no	ACTComp_75.mean	ACTComp_75.max	ACTComp_75.min	ACTComp_75.median	ACTComp_75.sd
1	1	27.71245	32	22	28.0	1.886584
2	2	24.39841	31	19	24.0	1.957712
3	3	19.77083	26	14	19.5	2.271372
4	4	27.85000	33	20	28.0	2.512092
5	5	32.63750	35	29	33.0	1.469037
6	6	24.19774	30	17	24.0	2.049085

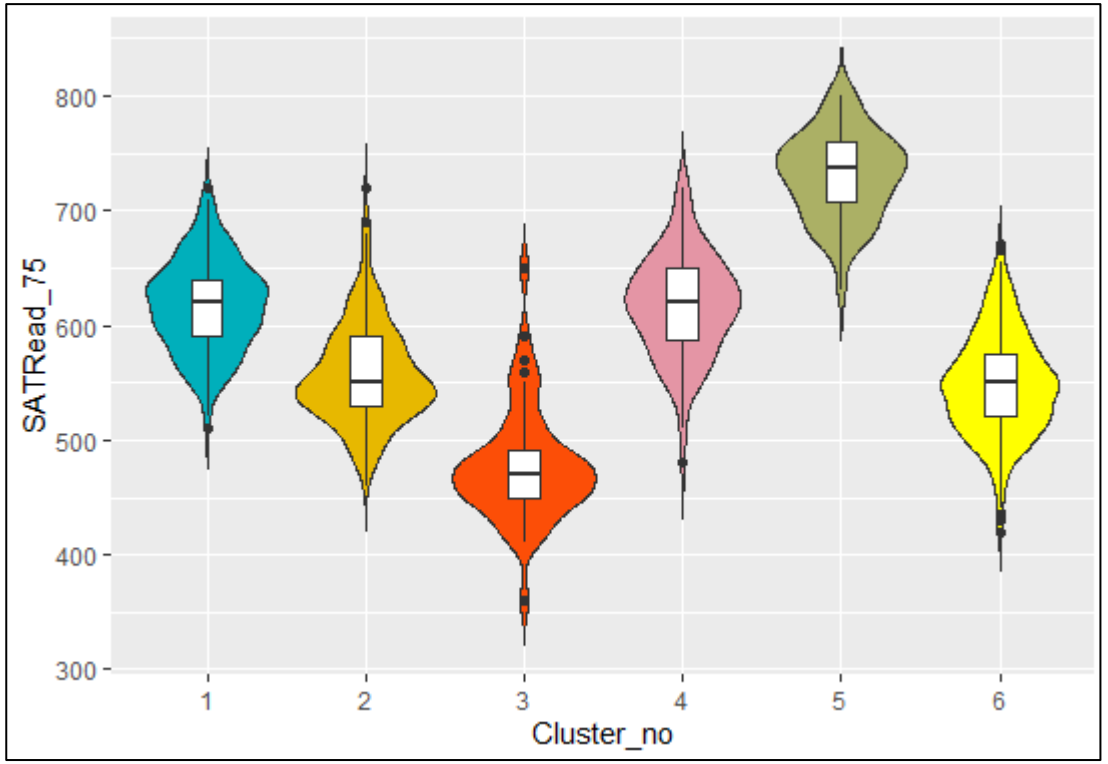
Cluster 5 (consisting of top colleges/universities) has the highest ACTComp_75 scores and its violin plot is shaped differently from others. The violin plot indicates that the ACTComp_75 scores for these colleges/universities are cloistered around the median.



j2. Metric – SAT Reading 75th percentile

	Cluster_no	SATRead_75.mean	SATRead_75.max	SATRead_75.min	SATRead_75.median	SATRead_75.sd
1	1	617.3605	720	510	620.0	40.38808
2	2	557.5378	720	460	550.0	43.44516
3	3	478.2083	650	360	470.0	50.66660
4	4	617.6500	720	480	620.0	45.10065
5	5	730.4750	800	630	736.5	37.70571
6	6	548.1328	670	420	550.0	44.98255

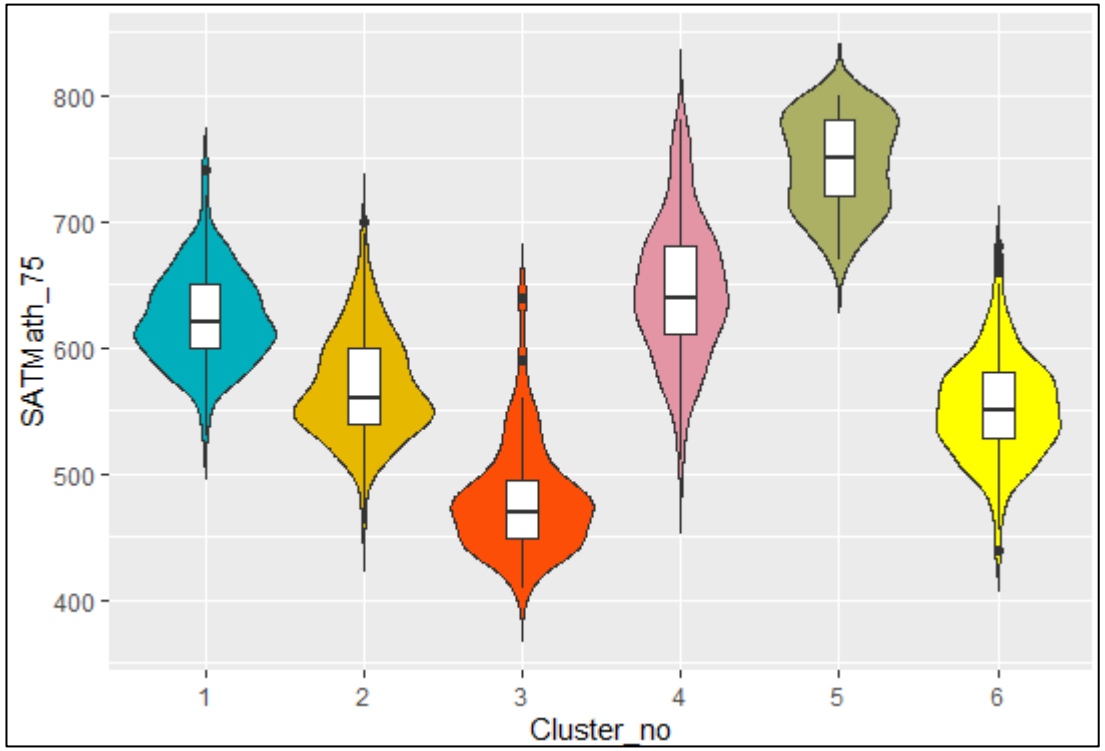
Cluster 5 (consisting of top colleges/universities) has the highest SATRead_75 scores.



j3. Metric – SAT Math 75th percentile

	Cluster_no	SATMath_75.mean	SATMath_75.max	SATMath_75.min	SATMath_75.median	SATMath_75.sd
1	1	623.8112	740	530	620	37.12965
2	2	569.1554	700	460	560	41.22014
3	3	480.8750	640	410	470	44.38594
4	4	646.3800	780	510	640	55.02007
5	5	745.2000	800	670	750	36.60511
6	6	551.7797	680	440	550	41.04060

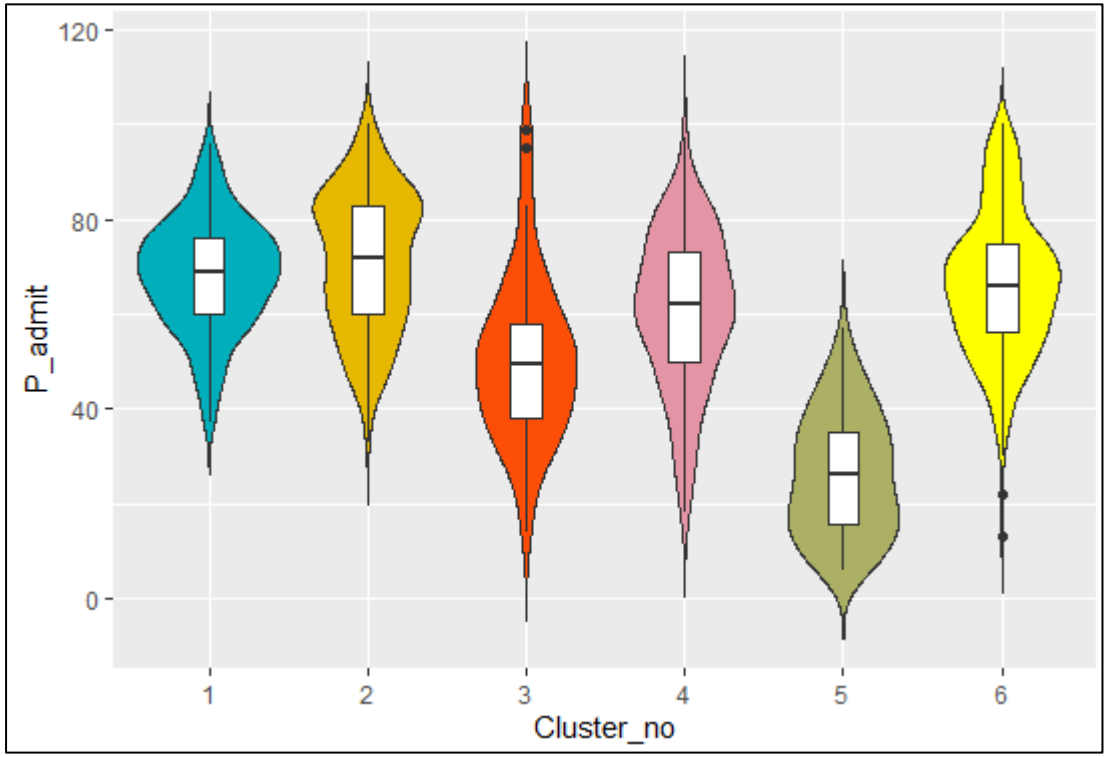
Cluster 5 (consisting of top colleges/universities) has the highest SATMath_75 scores and its violin plot is shaped differently from others. The violin plot indicates that the SATMath_75 scores for these colleges/universities are cloistered around the median. .



j4. Metric – Percent admitted

	Cluster_no	P_admit.mean	P_admit.max	P_admit.min	P_admit.median	P_admit.sd
1	1	68.25751	96	37	69.0	11.91394
2	2	71.25100	100	33	72.0	14.71451
3	3	50.58333	99	14	49.5	17.69461
4	4	60.61000	97	18	62.0	16.48010
5	5	26.01250	57	6	26.0	12.86787
6	6	66.40395	100	13	66.0	15.20448

Cluster 5 (consisting of top colleges/universities) has the lowest admissions rate (P_admit, i.e., percent admit). The table above shows that some of the most selective colleges/universities amongst cluster 5 admit only 6% of the applicants.

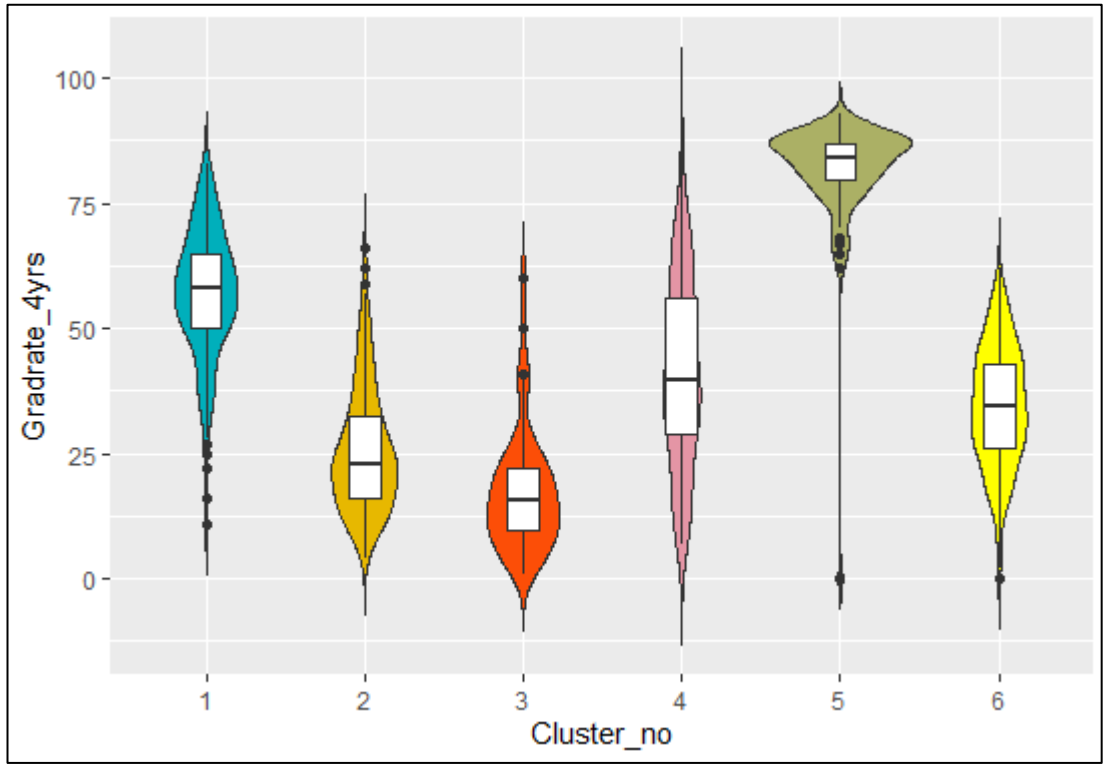


j5. Metric – 4-yr graduation rate

	Cluster_no	Gradrate_4yrs.mean	Gradrate_4yrs.max	Gradrate_4yrs.min	Gradrate_4yrs.median	Gradrate_4yrs.sd
1	1	57.18884	83	11	58.0	12.85860
2	2	25.74502	66	4	23.0	12.60185
3	3	17.41667	60	1	15.5	12.36129
4	4	41.05000	86	7	39.5	18.76348
5	5	81.72500	93	0	84.0	11.39728
6	6	34.44068	62	0	34.5	12.12065

Cluster 5 (consisting of top colleges/universities) has the highest 4-yr graduation rate(Gradrate_4yrs). The violin plot clearly indicates that vast majority of colleges/universities in this cluster maintain a 4-yr graduation rate in excess of 75%. The table above shows that some of the most selective colleges/universities amongst cluster 5 admit only 6% of the applicants.

There is one college/university in this cluster, the Northeastern University, where the 4-yr graduation rate is denoted as 0, and this skewers the violin plot for this cluster. I think, this is a case of missing information that got tabulated as a zero value rather than NA.



Overall comments: The K-means unsupervised algorithm did a good job in clustering a group of 80 elite colleges/universities that ranked in the top 60 or so in the US New 2021 college ranking. Summary statistics for this cluster clearly show that it stands out from the remaining 980 colleges/universities in the US.

QUESTION 3a
(Random Forest Regression)

Question 3: Can Random Forest and Neural Network regression be used to identify features that affect the 4-yr graduation rate? Compare results from two methods.

Answer 3a: I used Random Forest regression to identify variables (features) that affect the 4-yr graduation rate in US colleges/universities.

Files:

f_CS_3a_RF_GradR_4yrs.R, RMD_CS_3a_RF_GradR_4yrs.Rmd, RMD_CS_3a_RF_GradR_4yrs.html

Procedural steps for random forest regression:

a. Load data
b. Initial data processing
c. Split data into Training and Testing sets
d. Random Forest (RF) regressions:
    i. Run 1: Initial RF run with default values
    ii. Run 2: Tune mtry (using randomForest::tuneRF)
    iii. Run 3: Full grid search with ranger()
e. Apply best random forest model on test data
f. Variable importance plot

NOTE: In this analysis, all features (including numeric and categorical) were used.

Focus on ACTComp scores:

The original dataset contains ACTComp_75 (75th percentile) and ACTComp_25 (25th percentile), SATMath_75, SATMath_25, SATRead_75, SATRead_25, SATWrite_75, and SATWrite_25. The random forest (RF) algorithm used in this project can't handle missing data (NAs) in any column. Thus, a decision had to made regarding whether to include all or some of these important inputs namely: ACTComp_75 (334 NAs), ACTComp_25 (334 NAs), SATMath_25 (351 NAs), SATMath_75 (351 NAs), SATRead_25 (364 NAs), SATRead_75 (364 NAs), SATWrite_25 (819 NAs), and SATWrite_75 (819 NAs). It was indeed easy to decide to exclude the SATWrite columns due to large number of NA values. Including ACTComp_75, ACTComp_25, SATMath_25, SATMath_75, SATRead_25, SATRead_75 resulted in 1061 cases while including only ACTComp_75 and ACTComp_25 resulted in 1151 cases. The original dataset is a relatively small sample set of 1534 cases (with NAs) for RF training and testing purposes. Thus, I decided to include only ACTComp_75 and ACTComp_25 so that I could get as big a sample size as possible to train and test. Moreover, ACTComp_25 showed strong linear correlation with SATMath_25 and SATRead_25, and ACTComp_75 showed strong linear correlation with SATMath_75 and SATRead_75.

b. Initial data processing:

I removed the following features: ID_number, ZIP, County, Longitude, and Latitude, because they contain location identifiers for the colleges and universities. Also, all of the data pertains to the year 2013, and so I also removed the Yr feature. Additionally, I removed all rows where the ACTComp_25 and ACTComp_75 had missing values, and the columns Gradrate_5yrs and Gradrate_6yrs because my label was Gradrate_4yrs.

I also removed the following features: P_1stUG_instate, P_1stUG_outstate, P_1stUG_foreign, P_1stUG_resNA, SATWrite_25, and SATWrite_75 because these columns contained a large number of NAs (missing values), and the RF algorithm that I used can't have columns with missing values.

Originally, the dataset contained 1534 cases, but I was left with 1151 cases after removing all NAs.

c. Split data into Training and Testing sets

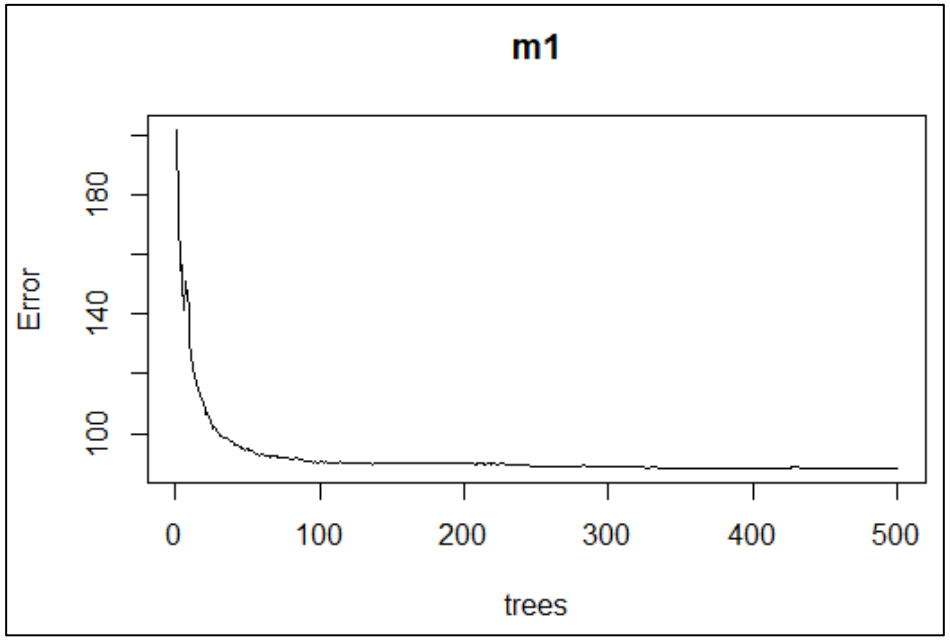
A training set was created using 70% of the data while the remaining 30% of the data was used to create a test set.

d. Random Forest (RF) regressions

d(i). Run 1: Initial RF run with default values

The initial RF run was made using default values present in the randomForest() function. This resulted in a mean of squared residuals = 88.45, and it was able to explain 79.42% of the variance. A plot of the error rate vs. number of trees (**Figure 3a.1**) showed that the error rate stabilized with around 100 trees but it continued to decrease (slowly) until around 300 trees. The lowest error was achieved when using 472 trees which resulted in average error of 9.39% for predicting Gradrate_4yrs.

Figure 3a.1



d. Random Forest (RF) regressions

d(ii). Run 2: Tune mtry (using randomForest::tuneRF)

In this run, I modified one of the primary tuning parameters, i.e., mtry which is the number of candidate variables to select from at each split, in order to improve the performance of the RF model. I started with mtry = 5, and increased it by 1.5 until the OOB error stopped improving by 1%.

The Min OOB Error (when improve = 0.01) is at mtry = 10 as seen in **Figure 3a.2**.

d(iii). Run 3: Full grid search with ranger()

I performed a larger grid search across several hyperparameters in this run. I created a grid, looped through each hyperparameter combination, and evaluated the model. This is where randomForest becomes quite inefficient since it does not scale well, and so I used ranger() as it is faster than randomForest().

The code applied 500 trees as it looped through each hyperparameter combination as previous work showed that 500 trees was aplenty to achieve a stable error rate.

The top 10 performing models all have RMSE values around 9.4 (**Figure 3a.3**). These results show that models with deeper trees (node_size = 3-7 observations in terminal node), mtry = 22, and sample size = 0.8 perform best.

So far, the best RF model - retains columnar categorical variables and uses mtry = 22, terminal node_size of 5 observations, and a sample size of 80%.

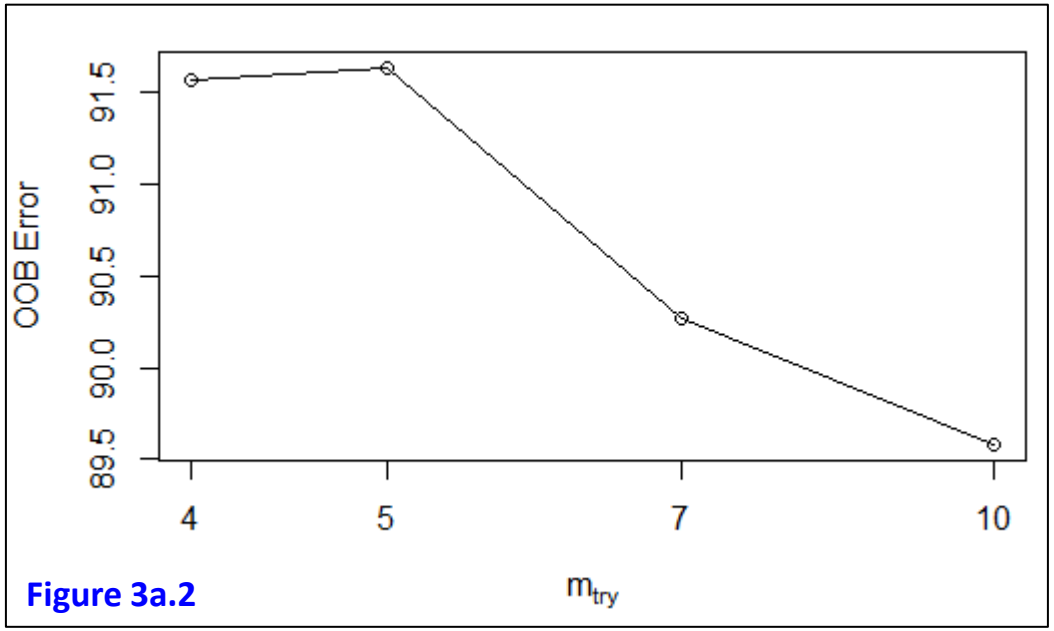


Figure 3a.2

Figure 3a.3

	mtry	node_size	sampe_size	OOB_RMSE
1	22	5	0.800	9.394130
2	19	7	0.800	9.394936
3	28	3	0.800	9.405324
4	19	3	0.800	9.406792
5	22	3	0.800	9.407071
6	19	5	0.800	9.408117
7	25	3	0.700	9.410705
8	25	5	0.800	9.410762
9	25	5	0.632	9.411979
10	22	7	0.700	9.412912

e. Apply best random forest model on test data

The best model is repeated to the **training data** get a better expectation of the error rate.

The Expected error ranges between ~9.35-9.48% with a most likely of 9.42% (**Figure 3a.4**).

Figure 3a.5 provides a perspective to the above error, as it shows that the 4-yr graduation rate varies between 0 and 90%, with a most likely value between 20-30%.

The best random forest model was applied on the **test data**, and the expected error was found to range between ~9.6 to 9.8% with a most likely value of ~9.7% (**Figure 3a.6**).

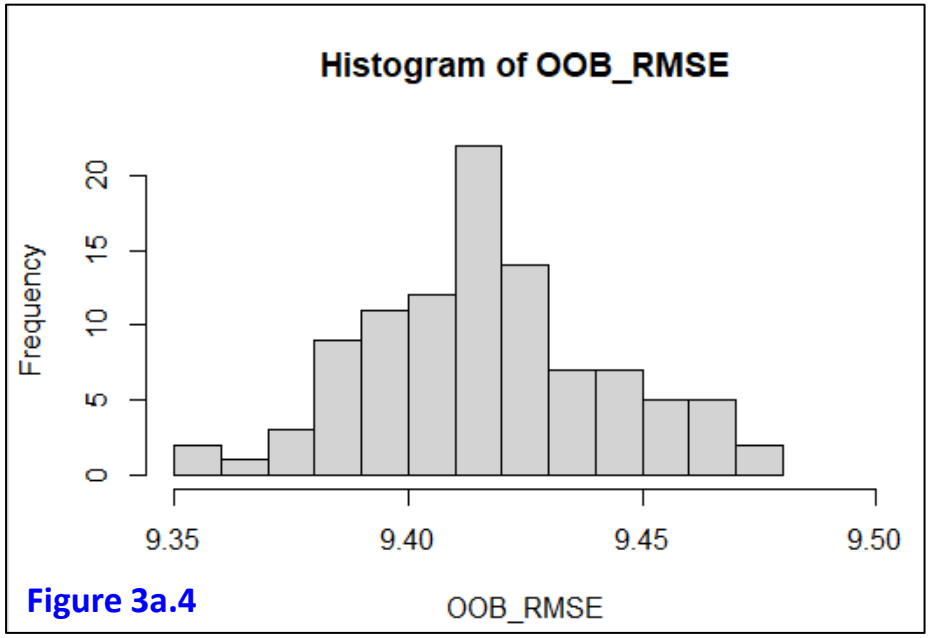


Figure 3a.4

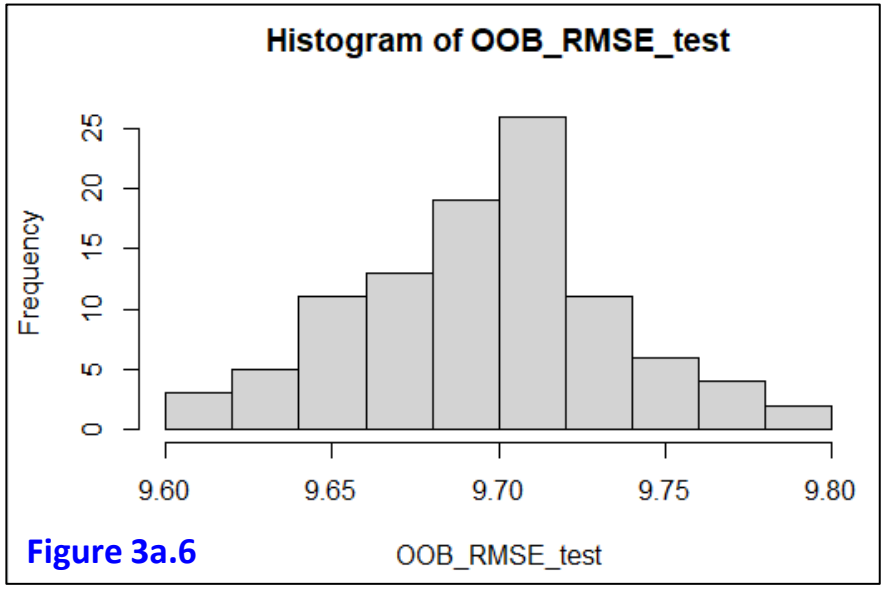


Figure 3a.6

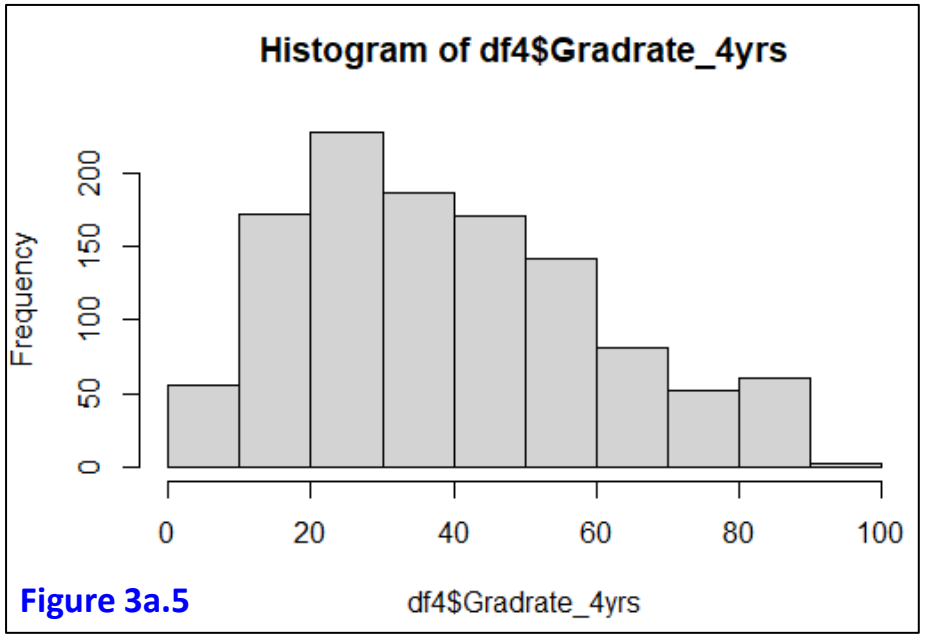


Figure 3a.5

f. Variable importance plot

Variable importance is measured by decrease in MSE when a variable is used as a node split. The remaining error after a node split is known as node impurity, and a variable that reduces this impurity is considered more important. The reduction in MSE for each variable across all the trees is accumulated, and the variable with the greatest accumulated impact is considered important or impactful.

The variable importance plot (**Figure 3a.7**) shows that the top two variables that affect 4-yr graduation rates are: P_1yrUG_Pell_grant (i.e., percent of 1st yr undergraduates with Pell grants) and P_1yrUG_Fed_grant. Students on Pell grants and Federal grants mostly come from economically disadvantaged families, and so may have to balance school with many other family responsibilities which might cause a delay in their graduation.

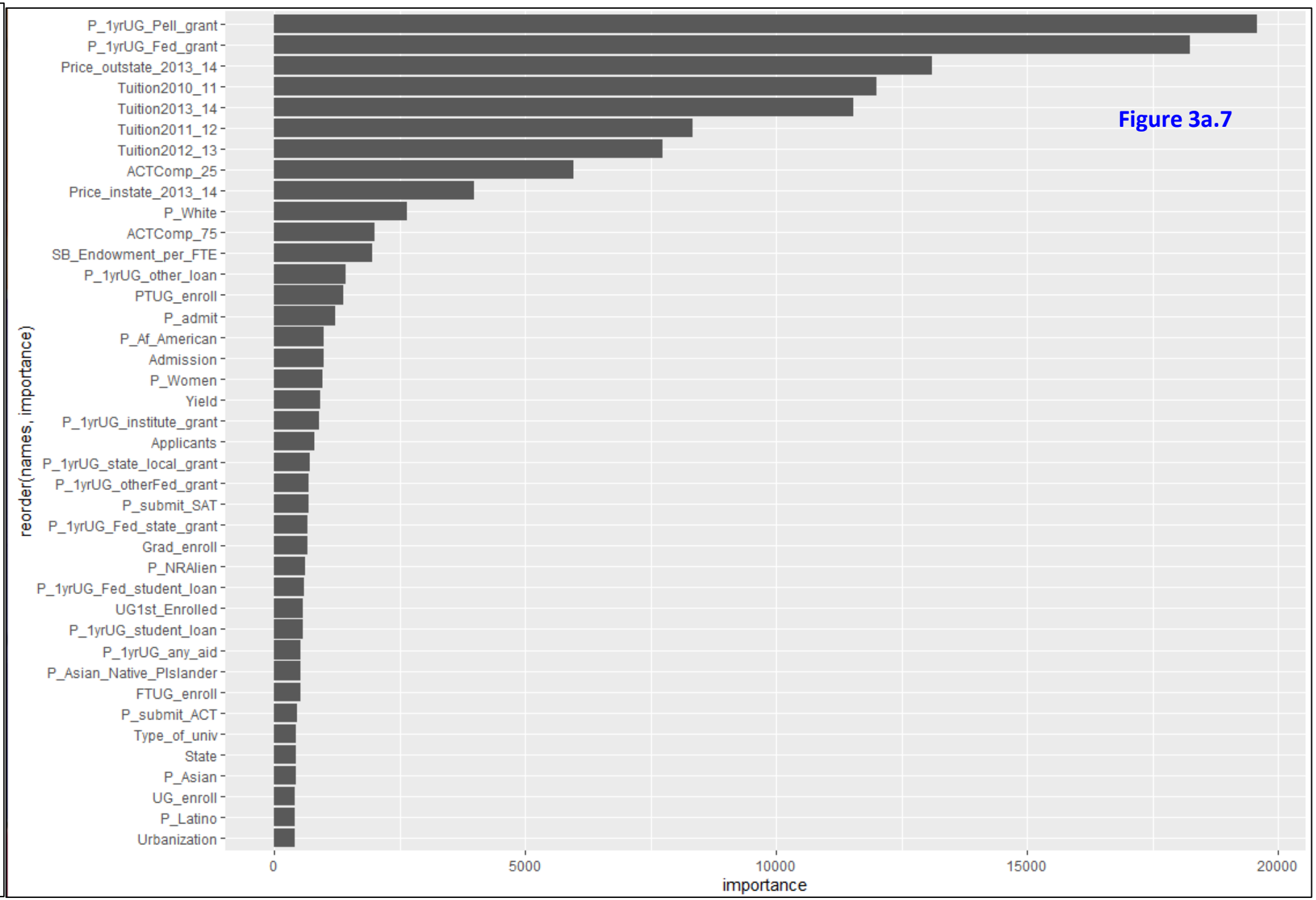


Figure 3a.7

f. Variable importance plot

Figure 3a.7 also shows that Price_outstate_2013_14, Tuitions (for 2010_11, 2013_14, 2011_12, and 2012_13) also affect 4-yr graduation rates. I think these variables influence the graduation rates because for schools with high tuitions, parents try to ensure that their kids graduate on time (within 4-yr window) so that they are not further burdened with even higher expenses. The out-of-state tuition for state universities are at par with expensive private colleges and universities, and so parents paying out-of-state prices (tuition + expenses) are burdened enough so as to ensure that their kids graduate on time. As expected ACTComp_25 and ACTComp_75 scores also seem to affect the 4-yr graduation rate. Schools with high scores under these categories will enroll motivated kids who want to graduate in 4 years.

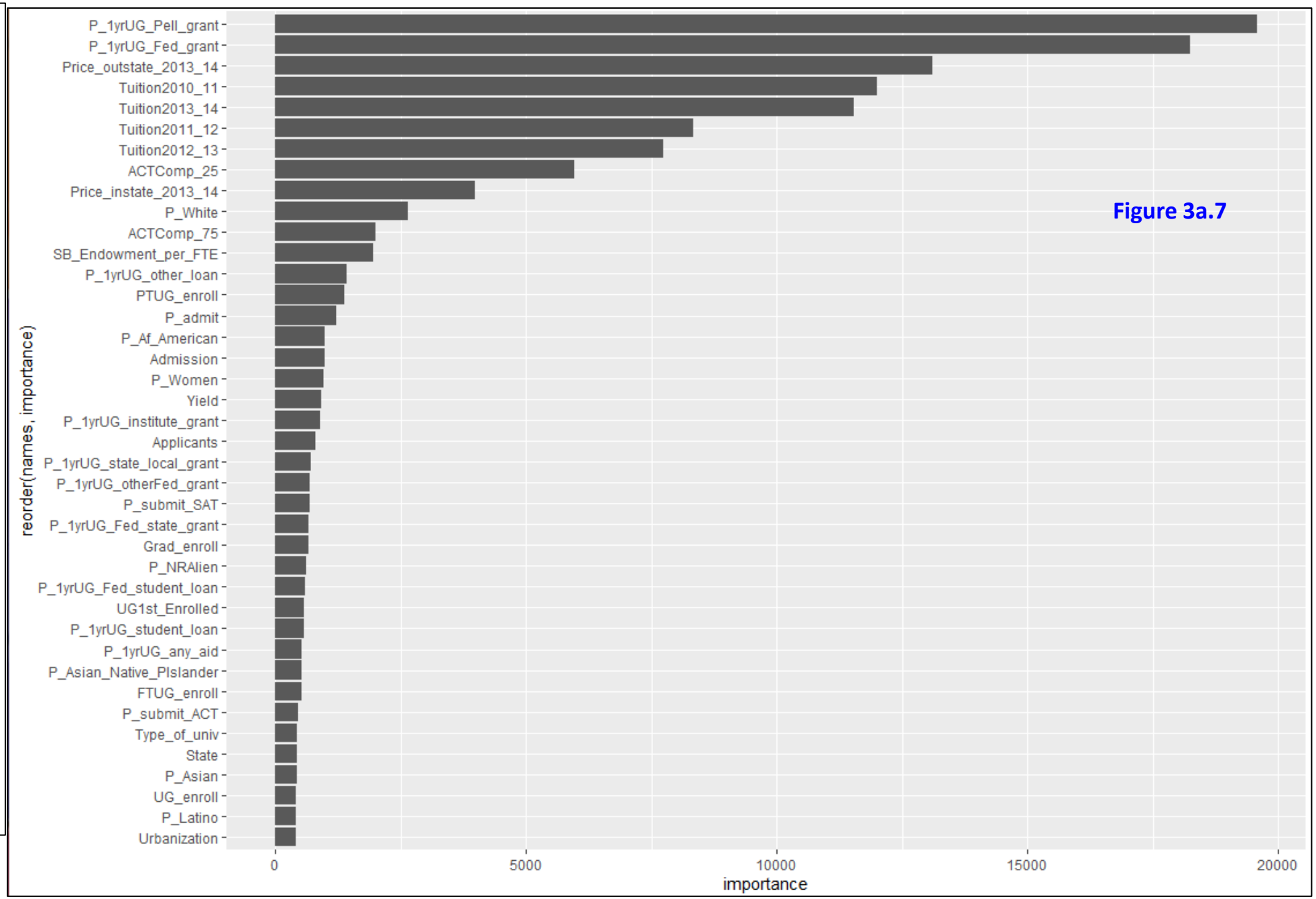


Figure 3a.7

QUESTION 3b
(Neural Network Regression – only numeric parameters)

Question 3: Can Neural Network regression be used to identify features that affect the 4-yr graduation rate? How does it compare with that obtained using RF?

Answer: I used Neural Network (NN) regression to identify variables (features) that affect the 4-yr graduation rate in US colleges/universities. The NN model helped identify important variables that affect 4-yr graduation rate. However, the NN model with both numeric and categorical parameters would crash due to convergence issues when I used the tanh activation function for which I had to rescale the data from a scale of [0,1] to a scale of [-1,1] using the rescale package. Thus, I also ran a NN regression model using only the numeric features, and all my NN model converged.

Files:

f_CS_3b_NN_GradR4yrs.R, RMD_CS_3b_NN_GradR4yrs.Rmd, RMD_CS_3b_NN_GradR4yrs.html

Procedural steps for random forest regression:

a. Load data
b. Initial data processing
c. Remove categorical variables and scale numeric data
d. Split data into Training and Testing sets
e. ANN regressions:
    i. Run 1: 1-hidden layer with 1 neuron
    ii. Run 2: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, logistic activation function
    iii. Run 3: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, tanh activation
    iv. Run 4: 1-Hidden Layer, 1-neuron, tanh activation function
f. Compare results - identify run with least test error
g. Variable importance plot
    i. Garson plot
    ii. Olden plot
h. Compare top 15 features (affecting 4-yr graduation rate) with RF and previous NN models

NOTE: In this analysis, only numeric features were used.

Focus on ACTComp scores:

The original dataset contains ACTComp_75 (75th percentile) and ACTComp_25 (25th percentile), SATMath_75, SATMath_25, SATRead_75, SATRead_25, SATWrite_75, and SATWrite_25. The random forest (RF) algorithm used in this project can't handle missing data (NAs) in any column. Thus, a decision had to made regarding whether to include all or some of these important inputs namely: ACTComp_75 (334 NAs), ACTComp_25 (334 NAs), SATMath_25 (351 NAs), SATMath_75 (351 NAs), SATRead_25 (364 NAs), SATRead_75 (364 NAs), SATWrite_25 (819 NAs), and SATWrite_75 (819 NAs). It was indeed easy to decide to exclude the SATWrite columns due to large number of NA values. Including ACTComp_75, ACTComp_25, SATMath_25, SATMath_75, SATRead_25, SATRead_75 resulted in 1061 cases while including only ACTComp_75 and ACTComp_25 resulted in 1151 cases. The original dataset is a relatively small sample set of 1534 cases (with NAs) for RF training and testing purposes. Thus, I decided to include only ACTComp_75 and ACTComp_25 so that I could get as big a sample size as possible to train and test. Moreover, ACTComp_25 showed strong linear correlation with SATMath_25 and SATRead_25, and ACTComp_75 showed strong linear correlation with SATMath_75 and SATRead_75.

b. Initial data processing:

I removed the following features: ID_number, ZIP, County, Longitude, and Latitude, because they contain location identifiers for the colleges and universities. Also, all of the data pertains to the year 2013, and so I also removed the Yr feature. Additionally, I removed all rows where the ACTComp_25 and ACTComp_75 had missing values, and the columns Gradrate_5yrs and Gradrate_6yrs because my label was Gradrate_4yrs.

I also removed the following features: P_1stUG_instate, P_1stUG_outstate, P_1stUG_foreign, P_1stUG_resNA, SATWrite_25, and SATWrite_75 because these columns contained a large number of NAs (missing values), and the NN algorithm that I used can’t have columns with missing values.

Originally, the dataset contained 1534 cases, but I was left with 1151 cases after removing all NAs.

c. Remove categorical variables and scale numeric data:

Removed all categorical features including: Relgious_y_n, State, Region, Status, HBCU, Urbanization, Type_of_univ. All numeric features were scaled.

d. Split data into Training and Testing sets:

Randomly extracted (without replacement) 80% of the observations to build the Training data set. The remaining 20% made up the test data set.

e. ANN regressions

e(i): Run 1: 1-hidden layer with 1 neuron

I constructed a 1-hidden layer ANN with 1 neuron, the simplest of all neural network and trained it on the training data set. The test error was found to be 1.064 while training error = 4.3602. So test error is smaller than the training error.

e(ii): Run 2: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, logistic activation function

I tried to improve the network by modifying its basic structure and hyperparameters, and so added depth to the hidden layer of the network. In this case, the test error was found to be 2.517 while the training error was 2.755. So test error is slightly smaller than training error.

e(iii): Run 3: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, tanh activation

In this run, I changed the activation function from logistic to the tangent hyperbolicus (tanh) to determine if these modification can improve the test data set SSE. For using the tanh activation function, I had to rescale the data from a scale of [0,1] to a scale of [−1,1] using the rescale package. As a result, I obtained a test error = 8.555 and a training error = 10.344.

e(iv): Run 4: 1-Hidden Layer, 1-neuron, tanh activation function

I modified the regression hyper-parameters again to see if I could reduce the testing errors. As a result, I obtained a test error = 7.218 and a training error = 17.361.

f. Compare results - identify run with least test error :

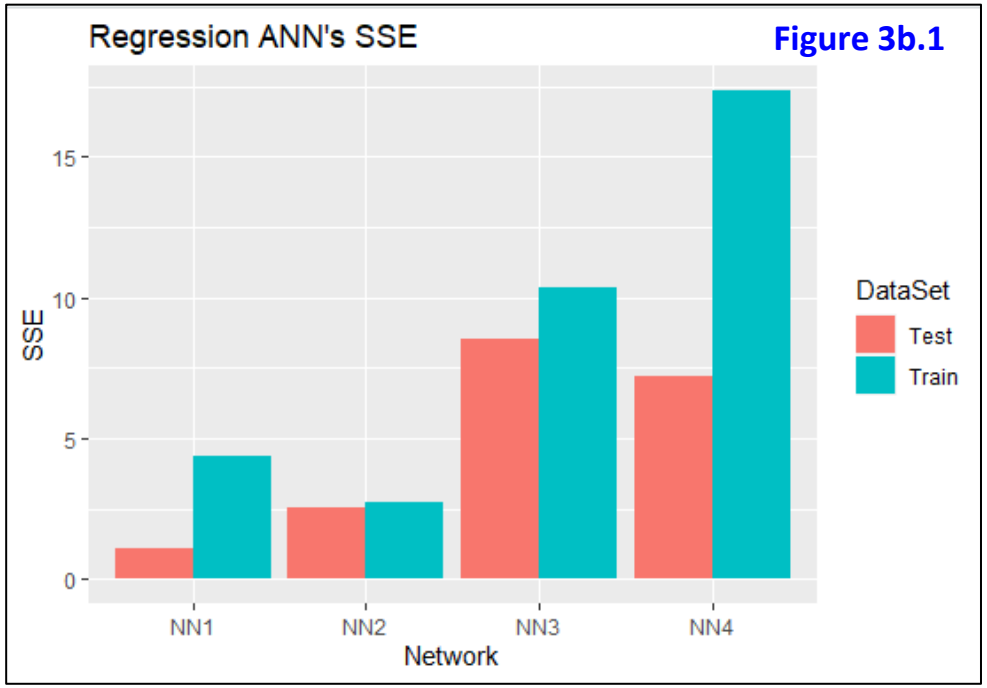
Figure 3b.1 compares the training and test errors for Run 1 (NN1), Run 2 (NN2), Run 3 (NN3), and Run 4 (NN4). It becomes clearly apparent that Run 1 has lower test errors.

g. Variable importance plot

The garson() function was used to plot the importance of each variable, and its results are indicative of the relative importance of each feature on a scale from zero to one.
NOTE: Only neural networks with one hidden layer and one output node can be evaluated with the garson() function.

The variable importance plot (using garson() function) is shown on the next page (**Figure 3b.2**)

The olden() function was also used to create a relative variable importance plot (**Figure 3b.3**). The advantage of the Olden plot is that it shows the relative contributions of each connection weight in terms of both magnitude and direction as compared to Garson's algorithm which only considers the absolute magnitude. Also, the Olden's algorithm is capable of evaluating neural networks with multiple hidden layers and response variables.



g(i). Variable importance plot – Garson plot

Figure 3b.2 shows that the top 5 features that significantly affect the 4-yr graduation rate include: P_Asian_Native_PIslander (i.e., % undergraduate enrollment - Asian/Native Hawaiian/Pacific Islander), PTUG_enroll (part time undergraduate enrollment), P_Asian (% undergraduate enrollment – Asian), SB_Endowment_per FTE (endowment per full time equivalent enrollment), and P_Af_American (% undergraduate - Black or African American).

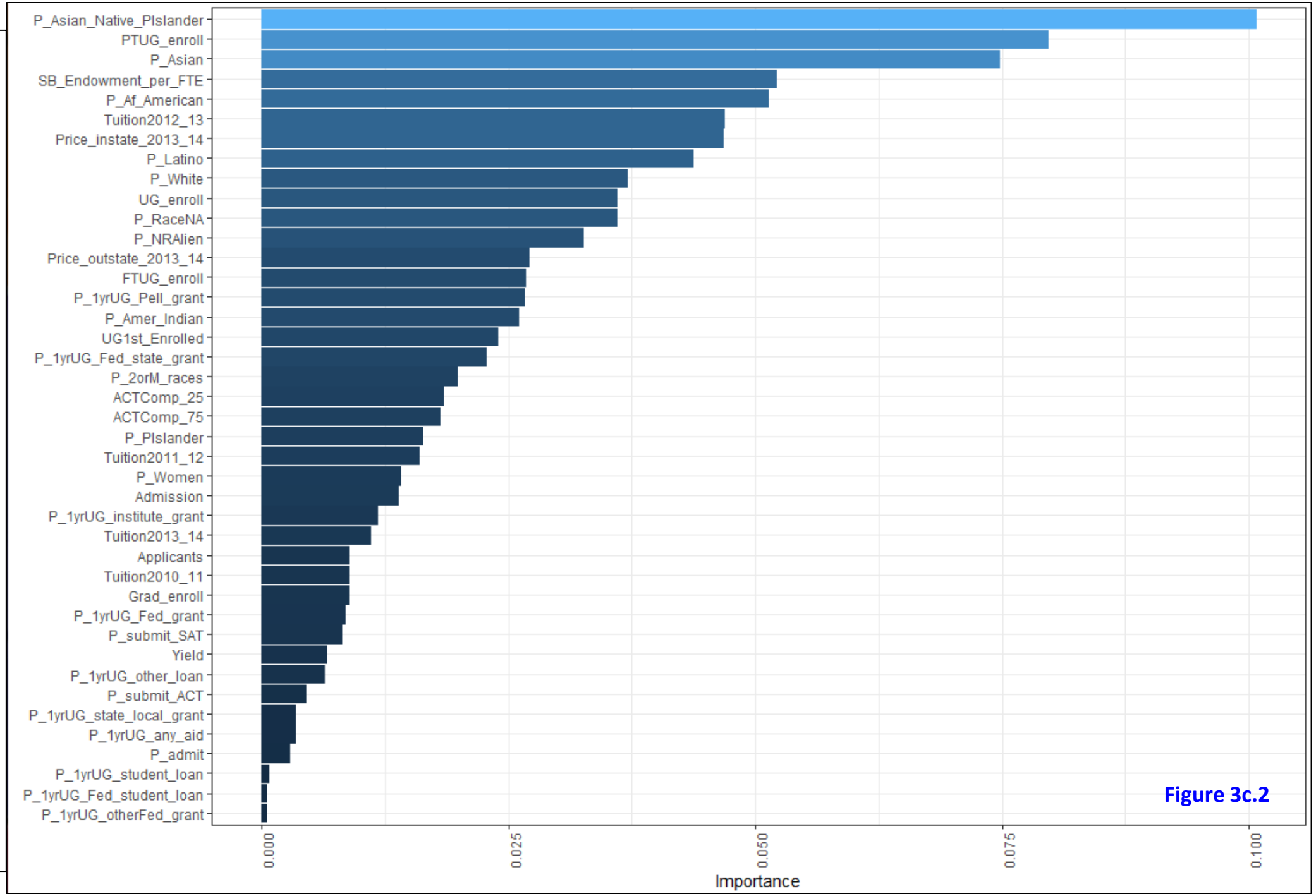


Figure 3c.2

g(ii). Variable importance plot – Olden plot

Figure 3b.3 shows that features such as : P_Asian (% undergraduate enrollment – Asian), SB_Endowment_per FTE (endowment per full time equivalent enrollment), Tuition 2012_13, and Price_outstate_2013_14 affect the 4-yr graduation rate positively, i.e., higher values of each of these parameters result in higher 4-yr graduation rates.

Figure 3b.3 also shows that features such as P_Asian_Native_PIslander (i.e., % undergraduate enrollment - Asian/Native Hawaiian/Pacific Islander), PTUG_enroll (part time undergraduate enrollment), P_Af_American (% undergraduate - Black or African American), Price_instate_2013_14, and P_Latino (i.e., % undergraduate – Latino) affect the 4-yr graduation rate negatively.

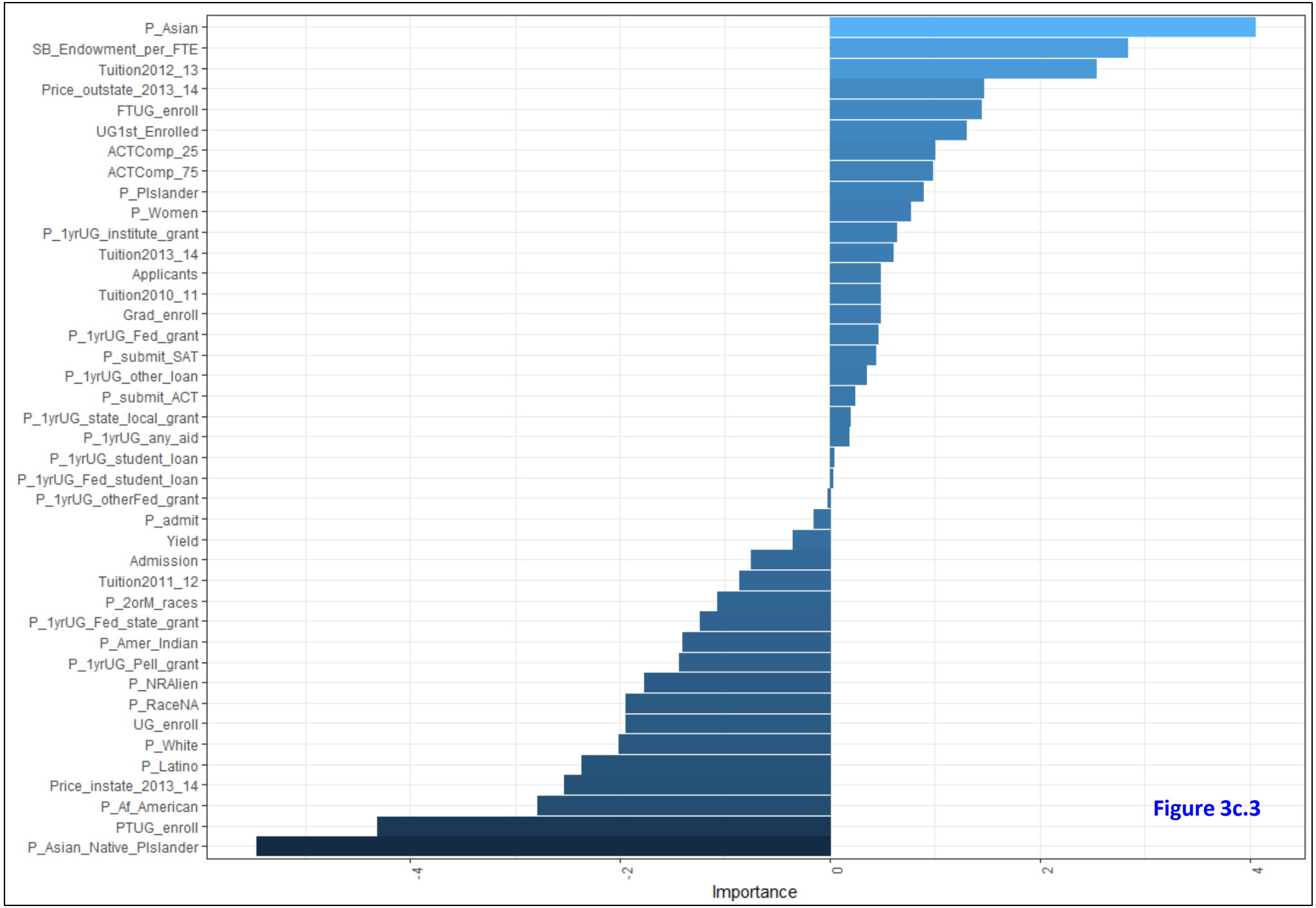


Figure 3c.3

h. Compare top 15 features (affecting 4-yr graduation rate) with RF model

Figure 3b.4 compares the top 15 features between the RF (3a) and the NN (3b) models.

The cells that are filled in blue show the common features between the RF and the NN (with only numeric parameters) models. Seven out of the top 15 driver features that affect the 4-yr graduation rate are common between these two models.

FINAL NOTE

RF and NN models are black box models – there are no known (phyics-based) laws that govern the relationship between the multiple input features and the label (here, the increase in graduation rates between 4 and 6 years). Thus, like in any social science research, it is difficult to establish causality between a feature and the label – one can only provide educated guesses as to why certain features rank high on the variable importance plot.

Figure 3b.4

	3a. RF model	3b. NN - numeric only
1	P_1yrUG_Pell_grant	P_Asian_Native_Pislander
2	P_1yrUG_Fed_grant	PTUG_enroll
3	Price_outstate_2013_14	P_Asian
4	Tuition2010_11	SB_Endowment_per_FTE
5	Tuition2013_14	P_Af_American
6	Tuition2011_12	Tuition2012_13
7	Tuition2012_13	Price_instate_2013_14
8	ACTComp_25	P_Latino
9	Price_instate_2013_14	P_White
10	P_White	UG_enroll
11	ACTComp_75	P_RaceNA
12	SB_Endowment_per_FTE	P_NRAlien
13	P_1yrUG_other_loan	Price_outstate_2013_14
14	PTUG_enroll	FTUG_enroll
15	P_admit	P_1yrUG_Pell_grant

QUESTION 4a
(Random Forest Regression)

Question 4a: Can Random Forest regression be used to identify features that affect the increase in graduation rates between 4 and 6 years?

Answer: I used Random Forest regression to identify variables (features) that affect the increase in graduation rates between 6 and 4 years in US colleges/universities. The RF model did a good job in identifying important variables that affect the increase in graduation rates between 4 to 6 years.

NOTE: The graduation rate can either increase or remain constant between 4 to 6 yrs.

Files:

f_CS_4a_RF_GradR_diff_6to4yrs.R, RMD_CS_4a_RF_GradR_diff_6to4yrs.Rmd, RMD_CS_4a_RF_GradR_diff_6to4yrs.html

Procedural steps for random forest regression:

a. Load data
b. Initial data processing
c. Split data into Training and Testing sets
d. Random Forest (RF) regressions:
    i. Run 1: Initial RF run with default values
    ii. Run 2: Tune mtry (using randomForest::tuneRF)
    iii. Run 3: Full grid search with ranger()
e. Apply best random forest model on test data
f. Variable importance plot

NOTE: In this analysis, all features (including numeric and categorical) were used.

Focus on ACTComp scores:

The original dataset contains ACTComp_75 (75th percentile) and ACTComp_25 (25th percentile), SATMath_75, SATMath_25, SATRead_75, SATRead_25, SATWrite_75, and SATWrite_25. The random forest (RF) algorithm used in this project can't handle missing data (NAs) in any column. Thus, a decision had to made regarding whether to include all or some of these important inputs namely: ACTComp_75 (334 NAs), ACTComp_25 (334 NAs), SATMath_25 (351 NAs), SATMath_75 (351 NAs), SATRead_25 (364 NAs), SATRead_75 (364 NAs), SATWrite_25 (819 NAs), and SATWrite_75 (819 NAs). It was indeed easy to decide to exclude the SATWrite columns due to large number of NA values. Including ACTComp_75, ACTComp_25, SATMath_25, SATMath_75, SATRead_25, SATRead_75 resulted in 1061 cases while including only ACTComp_75 and ACTComp_25 resulted in 1151 cases. The original dataset is a relatively small sample set of 1534 cases (with NAs) for RF training and testing purposes. Thus, I decided to include only ACTComp_75 and ACTComp_25 so that I could get as big a sample size as possible to train and test. Moreover, ACTComp_25 showed strong linear correlation with SATMath_25 and SATRead_25, and ACTComp_75 showed strong linear correlation with SATMath_75 and SATRead_75.

b. Initial data processing:

I removed the following features: ID_number, ZIP, County, Longitude, and Latitude, because they contain location identifiers for the colleges and universities. Also, all of the data pertains to the year 2013, and so I also removed the Yr feature. Additionally, I removed all rows where the ACTComp_25 and ACTComp_75 had missing values.

I also removed the following features: P_1stUG_instate, P_1stUG_outstate, P_1stUG_foreign, P_1stUG_resNA, SATWrite_25, and SATWrite_75 because these columns contained a large number of NAs (missing values), and the RF algorithm that I used can’t have columns with missing values.

I created a new column called Gradrate_diff_6to4yrs which was the difference in graduation rates between 6 and 4 years, and this became my Label. After creation of this label, I removed the columns: Gradrate_4yrs, Gradrate_5yrs, and Gradrate_6yrs.

Originally, the dataset contained 1534 cases, but I was left with 1151 cases after removing all NAs.

c. Split data into Training and Testing sets

A training set was created using 70% of the data while the remaining 30% of the data was used to create a test set.

d. Random Forest (RF) regressions

d(i). Run 1: Initial RF run with default values

The initial RF run was made using default values present in the randomForest() function. This resulted in a mean of squared residuals = 41.83, and it was able to explain 47.42% of the variance. A plot of the error rate vs. number of trees (**Figure 4a.1**) showed that the error rate stabilized with around 100 trees but it continued to decrease (slowly) until around 300 trees. The lowest error was achieved when using 100 trees which resulted in average error of 6.45% for predicting the increase in graduation rates between 4 and 6 years.

d(ii). Run 2: Tune mtry (using randomForest::tuneRF)

In this run, I modified one of the primary tuning parameters, i.e., mtry which is the number of candidate variables to select from at each split, in order to improve the performance of the RF model. I started with mtry = 5, and increased it by 1.5 until the OOB error stopped improving by 1%.

The Min OOB Error (when improve = 0.01) is at mtry = 4 as seen in **Figure 4a.2**.

Figure 4a.1

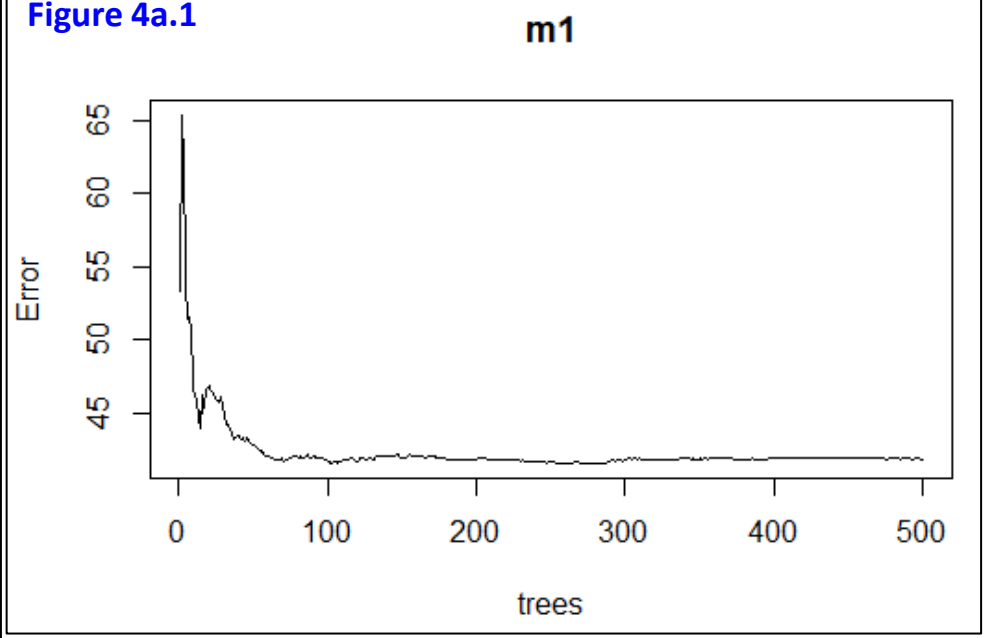
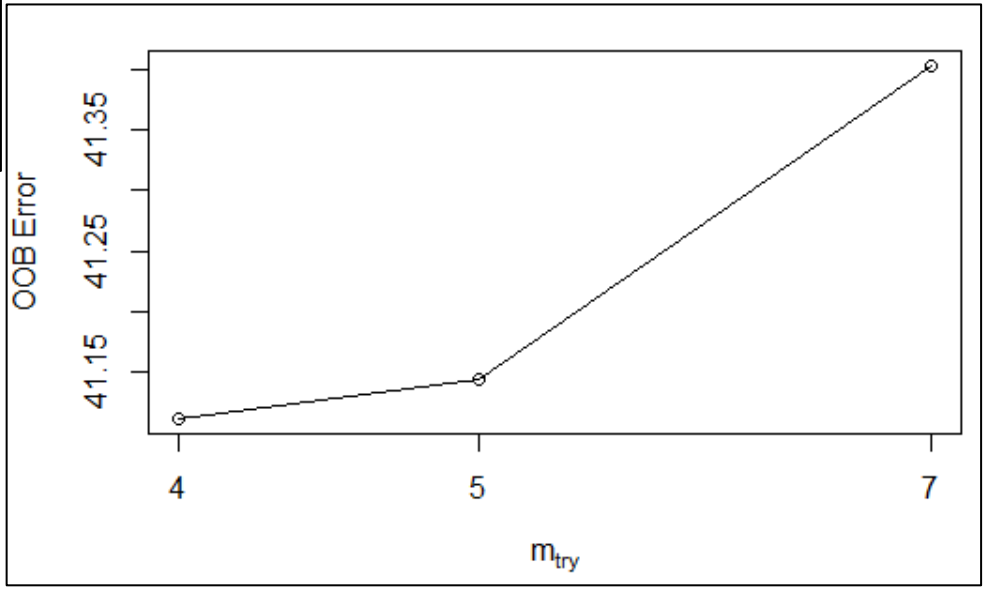


Figure 4a.2



d. Random Forest (RF) regressions

d(iii). Run 3: Full grid search with ranger()

I performed a larger grid search across several hyperparameters in this run. I created a grid, looped through each hyperparameter combination, and evaluated the model. This is where randomForest becomes quite inefficient since it does not scale well, and so I used ranger() as it is faster than randomForest().

The code used 300 trees as it looped through each hyperparameter combination as previous work showed that 300 trees was aplenty to achieve a stable error rate.

The top 10 performing models all have RMSE values around 6.37 (**Figure 4a.3**). These results show that models with deeper trees (node_size = 3-5 observations in terminal node), mtry = 14, and sample size = 0.7 perform best.

So far, the best RF model - retains columnar categorical variables and uses mtry = 14, terminal node_size of 3 observations, and a sample size of 70%.

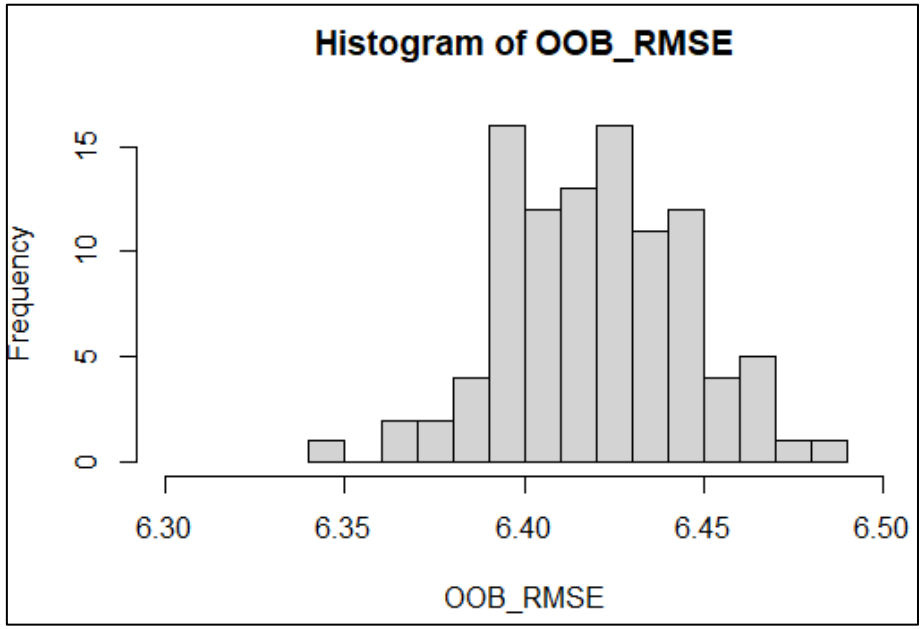
e. Apply best random forest model on test data

The best model is repeated on the training data to get a better expectation of the error rate. The Expected error ranges between ~6.35-6.49% with a most likely of 6.42% (**Figure 4a.4**).

Figure 4a.3

	mtry	node_size	sampe_size	OOB_RMSE
1	14	3	0.700	6.358926
2	6	5	0.632	6.369172
3	6	5	0.550	6.371672
4	6	3	0.700	6.372336
5	18	9	0.800	6.372842
6	10	7	0.800	6.373940
7	10	7	0.700	6.375683
8	18	9	0.550	6.377285
9	6	7	0.632	6.381480
10	14	7	0.700	6.381843

Figure 4a.4



d. Apply best random forest model on test data

Figure 4a.5 provides a perspective to the above error, as it shows that the increase in graduation rate between 4 and 6 years varies between 0 and 40%, with a most likely value between 10-20%.

The best random forest model was applied on the test data, and the expected error was found to range between ~6.4 to 6.55% with a most likely value of ~6.8% (**Figure 4a.4**).

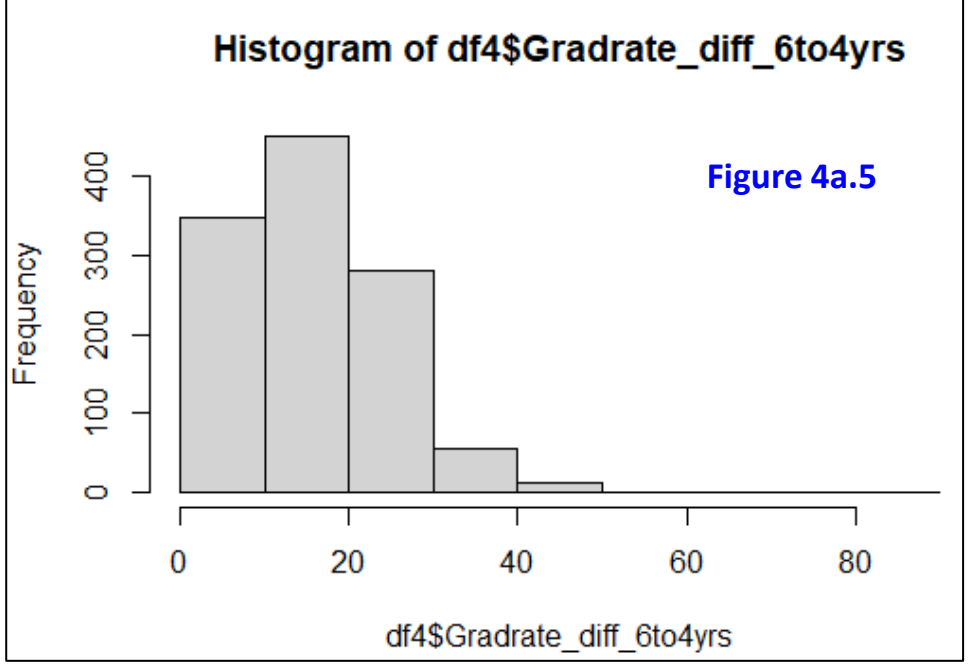
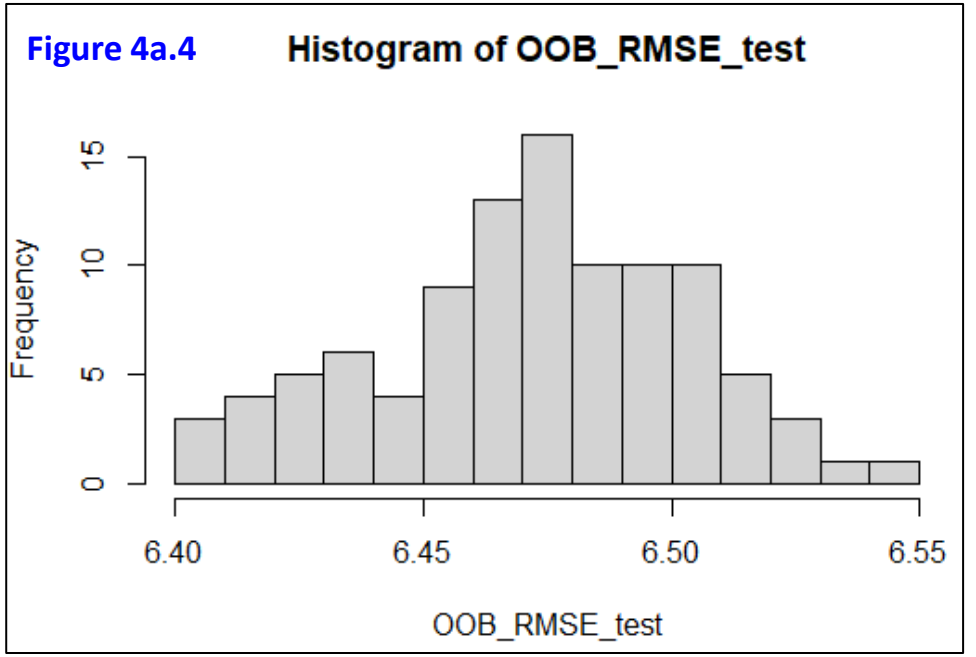


Figure 4a.5

Figure 4a.4



f. Variable importance plot

Variable importance is measured by decrease in MSE when a variable is used as a node split. The remaining error after a node split is known as node impurity, and a variable that reduces this impurity is considered more important. The reduction in MSE for each variable across all the trees is accumulated, and the variable with the greatest accumulated impact is considered important or impactful.

The variable importance plot (**Figure 4a.6**) shows that the top four variables that affect the increase in graduation rates between 4 and 6 years are all related to tuition: Tuition2010_11, Tuition2012_13, Tuition2011_12, and Tuition2013_14. It is understandable that parents paying high tuition want to have their kids graduate as early as possible. So if they miss graduating in 4 years, then parents ensure that they graduate in the 5th or 6th year.

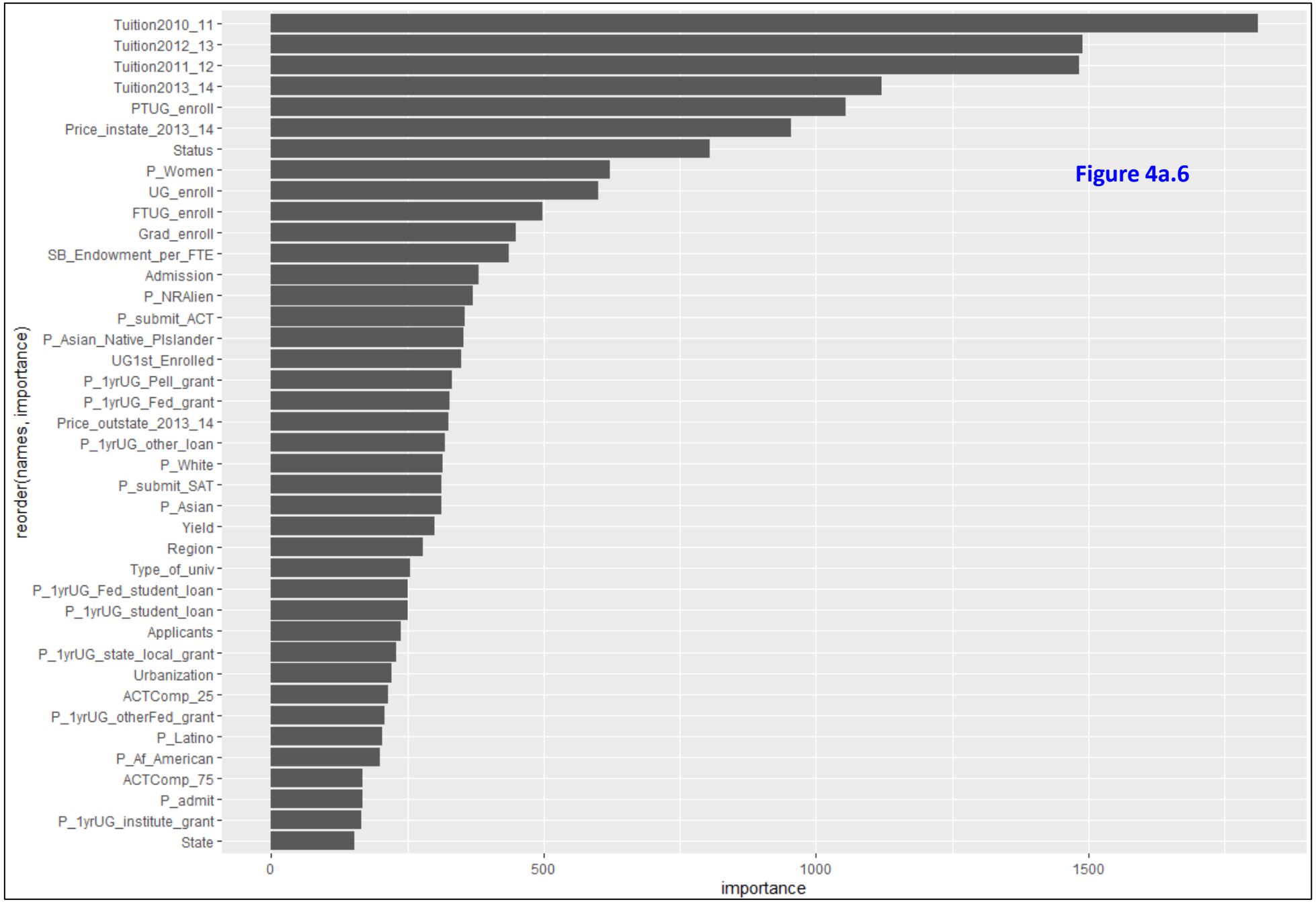


Figure 4a.6

f. Variable importance plot

Another important variable is the PTUG_enroll (i.e., part time undergraduate enrollment). It is again understandable that part time students often have to juggle multiple responsibilities outside school and so their graduation may be delayed beyond the typical 4 years.

I am unsure about how to explain the role that Price_instate_2013_14 plays in determining increase of graduation rate between 4 and 6 years other than that it also refers to money being spent on school, and delaying graduation beyond 4 years adds to the money spent in college. So higher this price, the more incentive a student has to graduate on time (in 4 years).

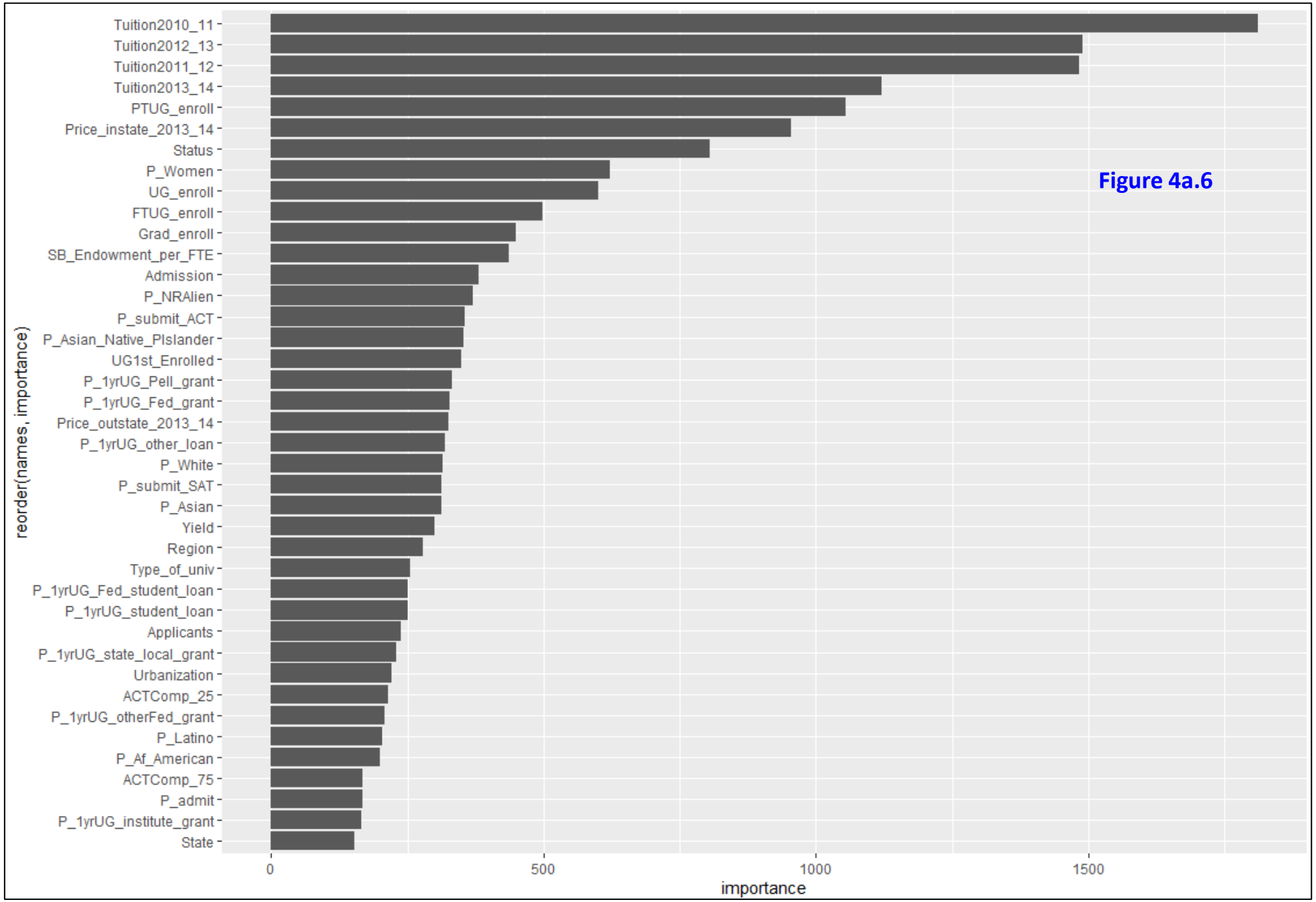


Figure 4a.6

QUESTION 4b
(Neural Network Regression – only numeric parameters)

Question 4b: Can Neural Network regression be used to identify features that affect the increase in graduation rates between 4 and 6 years? How do the top 15 variables compare with that predicted by the RF model?

Answer: I ran a NN regression model using only the numeric features in this case because the NN model including all features (numeric and categorical) would crash due to convergence issues when I used the tanh activation function.

Files: f_CS_4b_NN_GradR_diff_6to4yrs.R, RMD_CS_4b_NN_GradR_diff_6to4yrs.Rmd, RMD_CS_4b_NN_GradR_diff_6to4yrs.html

Procedural steps for random forest regression:

a. Load data
b. Initial data processing
c. Remove categorical variables and scale numeric data
d. Split data into Training and Testing sets
e. ANN regressions:
    i. Run 1: 1-hidden layer with 1 neuron
    ii. Run 2: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, logistic activation function
    iii. Run 3: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, tanh activation
    iv. Run 4: 1-Hidden Layer, 1-neuron, tanh activation function
f. Compare results - identify run with least test error
g. Variable importance plots - on run with least error
    i. Garson plot
    ii. Olden plot
h. Compare top 15 features (affecting increase in graduation rates between 4 and 6 years) with RF models

NOTE: In this analysis, only numeric features were used.

Focus on ACTComp scores:

The original dataset contains ACTComp_75 (75th percentile) and ACTComp_25 (25th percentile), SATMath_75, SATMath_25, SATRead_75, SATRead_25, SATWrite_75, and SATWrite_25. The random forest (RF) algorithm used in this project can't handle missing data (NAs) in any column. Thus, a decision had to made regarding whether to include all or some of these important inputs namely: ACTComp_75 (334 NAs), ACTComp_25 (334 NAs), SATMath_25 (351 NAs), SATMath_75 (351 NAs), SATRead_25 (364 NAs), SATRead_75 (364 NAs), SATWrite_25 (819 NAs), and SATWrite_75 (819 NAs). It was indeed easy to decide to exclude the SATWrite columns due to large number of NA values. Including ACTComp_75, ACTComp_25, SATMath_25, SATMath_75, SATRead_25, SATRead_75 resulted in 1061 cases while including only ACTComp_75 and ACTComp_25 resulted in 1151 cases. The original dataset is a relatively small sample set of 1534 cases (with NAs) for RF training and testing purposes. Thus, I decided to include only ACTComp_75 and ACTComp_25 so that I could get as big a sample size as possible to train and test. Moreover, ACTComp_25 showed strong linear correlation with SATMath_25 and SATRead_25, and ACTComp_75 showed strong linear correlation with SATMath_75 and SATRead_75.

b. Initial data processing:

I removed the following features: ID_number, ZIP, County, Longitude, and Latitude, because they contain location identifiers for the colleges and universities. Also, all of the data pertains to the year 2013, and so I also removed the Yr feature. Additionally, I removed all rows where the ACTComp_25 and ACTComp_75 had missing values, and the columns Gradrate_5yrs and Gradrate_6yrs because my label was Gradrate_4yrs.

I also removed the following features: P_1stUG_instate, P_1stUG_outstate, P_1stUG_foreign, P_1stUG_resNA, SATWrite_25, and SATWrite_75 because these columns contained a large number of NAs (missing values), and the NN algorithm that I used can’t have columns with missing values.

Originally, the dataset contained 1534 cases, but I was left with 1151 cases after removing all NAs.

c. Remove categorical variables and scale numeric data:

Removed all categorical features including: Relgious_y_n, State, Region, Status, HBCU, Urbanization, Type_of_univ. All numeric features were scaled.

d. Split data into Training and Testing sets:

Randomly extracted (without replacement) 80% of the observations to build the Training data set. The remaining 20% made up the test data set.

e. ANN regressions

e(i): Run 1: 1-hidden layer with 1 neuron

I constructed a 1-hidden layer ANN with 1 neuron, the simplest of all neural network and trained it on the training data set. The test error was found to be 0.741 while training error = 2.508. So test error is smaller than the training error.

e(ii): Run 2: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, logistic activation function

I tried to improve the network by modifying its basic structure and hyperparameters, and so added depth to the hidden layer of the network. In this case, the test error was found to be 0.876 while the training error was 1.786. So test error is smaller than training error.

e(iii): Run 3: 2-Hidden Layers, Layer 1: 4-neurons, Layer 2: 1-neuron, tanh activation

In this run, I changed the activation function from logistic to the tangent hyperbolicus (tanh) to determine if these modification can improve the test data set SSE. For using the tanh activation function, I had to rescale the data from a scale of [0,1] to a scale of [−1,1] using the rescale package. As a result, I obtained a test error = 16.158 and a training error = 7.115.

e(iv): Run 4: 1-Hidden Layer, 1-neuron, tanh activation function

I modified the regression hyper-parameters again to see if I could reduce the testing errors. As a result, I obtained a test error = 23.681 and a training error = 21.519.

f. Compare results - identify run with least test error :

Figure 4b.1 compares the training and test errors for Run 1 (NN1), Run 2 (NN2), Run 3 (NN3), and Run 4 (NN4). It becomes clearly apparent that Run 1 has lower test errors.

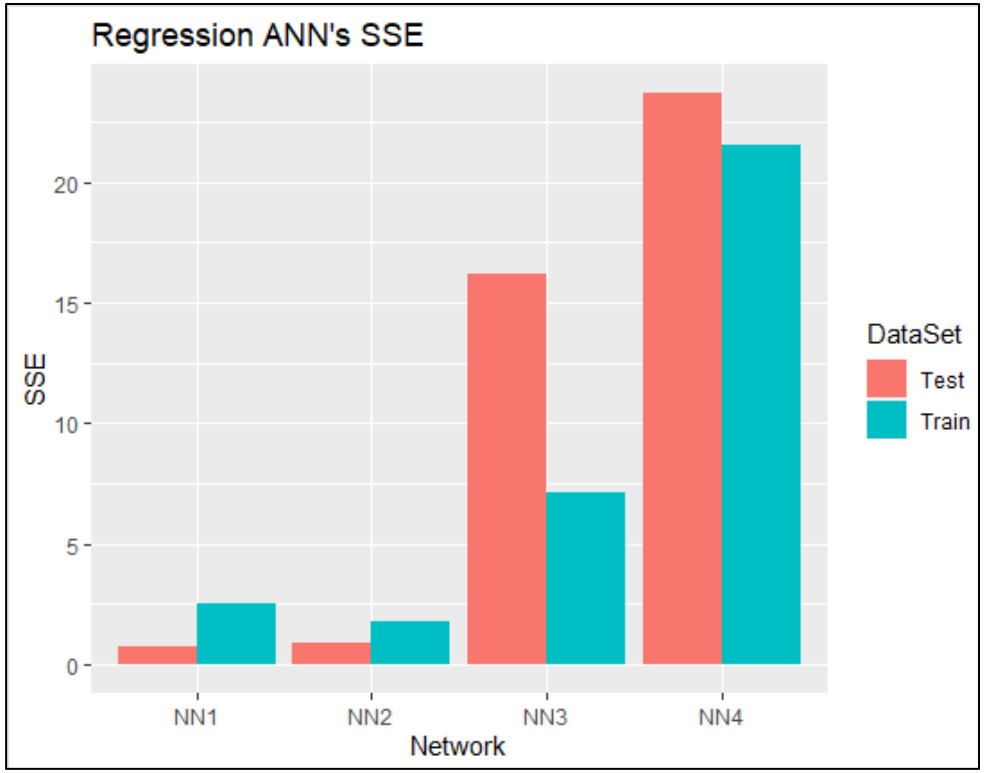
g. Variable importance plot

The garson() function was used to plot the importance of each variable, and its results are indicative of the relative importance of each feature on a scale from zero to one.
NOTE: Only neural networks with one hidden layer and one output node can be evaluated with the garson() function.

The variable importance plot (using the garson()) is shown on the next page (**Figure 4b.2**)

The olden() function was also used to create a relative variable importance plot (**Figure 4b.3**). The advantage of the Olden plot is that it shows the relative contributions of each connection weight in terms of both magnitude and direction as compared to Garson's algorithm which only considers the absolute magnitude. Also, the Olden's algorithm is capable of evaluating neural networks with multiple hidden layers and response variables.

Figure 4b.1



g(i). Variable importance plot – Garson plot

Figure 4b.2 shows that the top 5 features that significantly affect the increase in graduation rates between 4 and 6 years, and these include: UG1st_Enrolled (undergraduate freshmen enrolled), FTUG_enroll (full time undergraduate enrollment), P_Asian_Native_PIslander (% undergraduate enrollment - Asian/Native Hawaiian/Pacific Islander), P_Asian (% undergraduate enrollment – Asian), and P_PIslander (% undergraduate enrollment - Native Hawaiian or Other Pacific Islander).

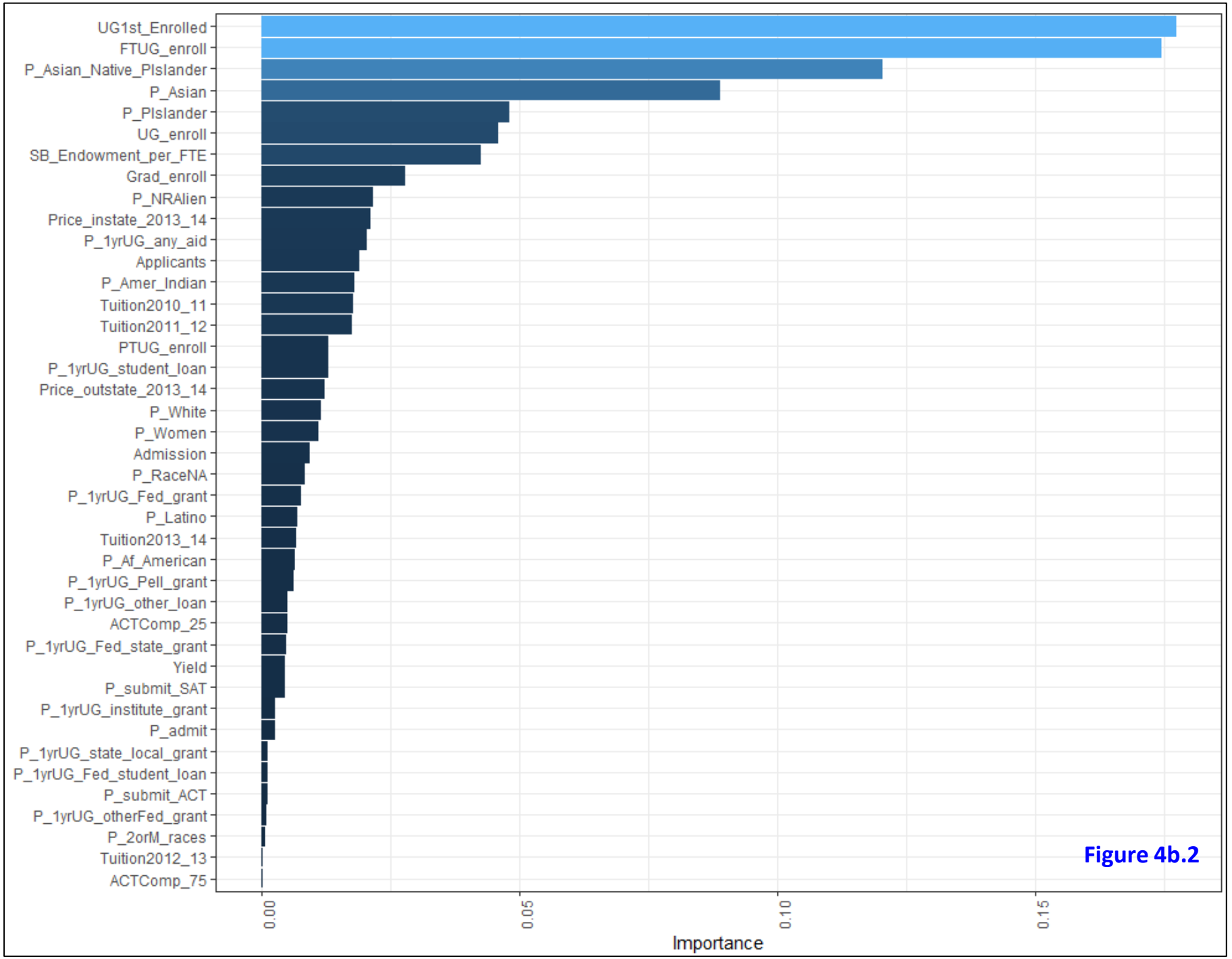
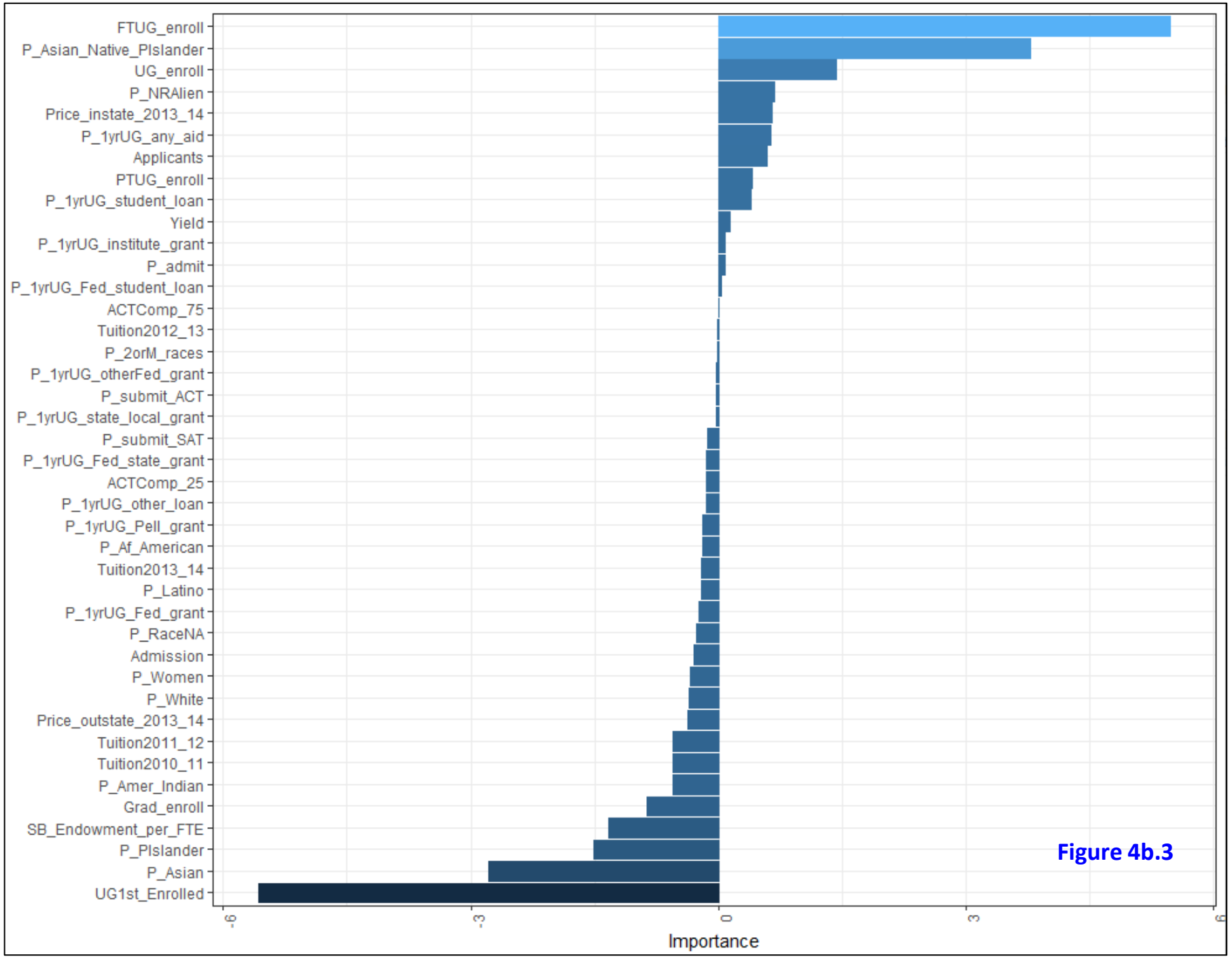


Figure 4b.2

g(ii). Variable importance plot – Olden plot

Figure 4b.3 shows that features such as : FTUG_enroll (full time enrollment), P_Asian_Native_PIslander (% undergraduate enrollment - Asian/Native Hawaiian/Pacific Islander), UG_enroll (undergraduate enrollment), P_NRAlien (% undergraduate enrollment - Nonresident Alien), and Price_instate_2013_14 (total price for instate students) affect the increase in graduation rate from 4 to 6 years positively, i.e., higher values of each of these parameters result in higher increases in graduation rates.

It is understandable that increased number of students will try to graduate within 6 yrs in schools with higher number of full time undergrads (FTUG_enroll), and that students in schools with high in-state attendance price tag (Price_instate_2013_14) will also want to complete their degrees soon after completion of 4 yrs. Also many students who are Nonresident Alien (P_NRAlien) have financial and other limitations which prevent them for graduating in 4 years, but these same financial limitations motivate them to graduate within 6 years so as not to add to their stressed finances.



g(ii). Variable importance plot – Olden plot

For schools with high enrollment of undergraduate freshmen (UG1st_Enrolled), most will graduate on time (in 4 years) and so result in lower increases in graduation rates between 4 and 6 years. This analysis also indicates that schools with a higher percentage of students of Asian origin (P_Asian) tend to show lesser increases in graduation rates beyond 4 years. This maybe because most Asian kids graduate within 4 years and so there are less of them to graduate in the 5th or 6th years, or those you don't graduate within 4 years take more than 6 years to graduate. This analysis also shows that if the school's endowment (per full time equivalent enrollment), i.e., SB_Endowment_per_FTE is high, then perhaps the school has enough resource to help at risk students, and so most kids graduate within 4 years, and the increase in graduation rate between 4 and 6 years is not high.

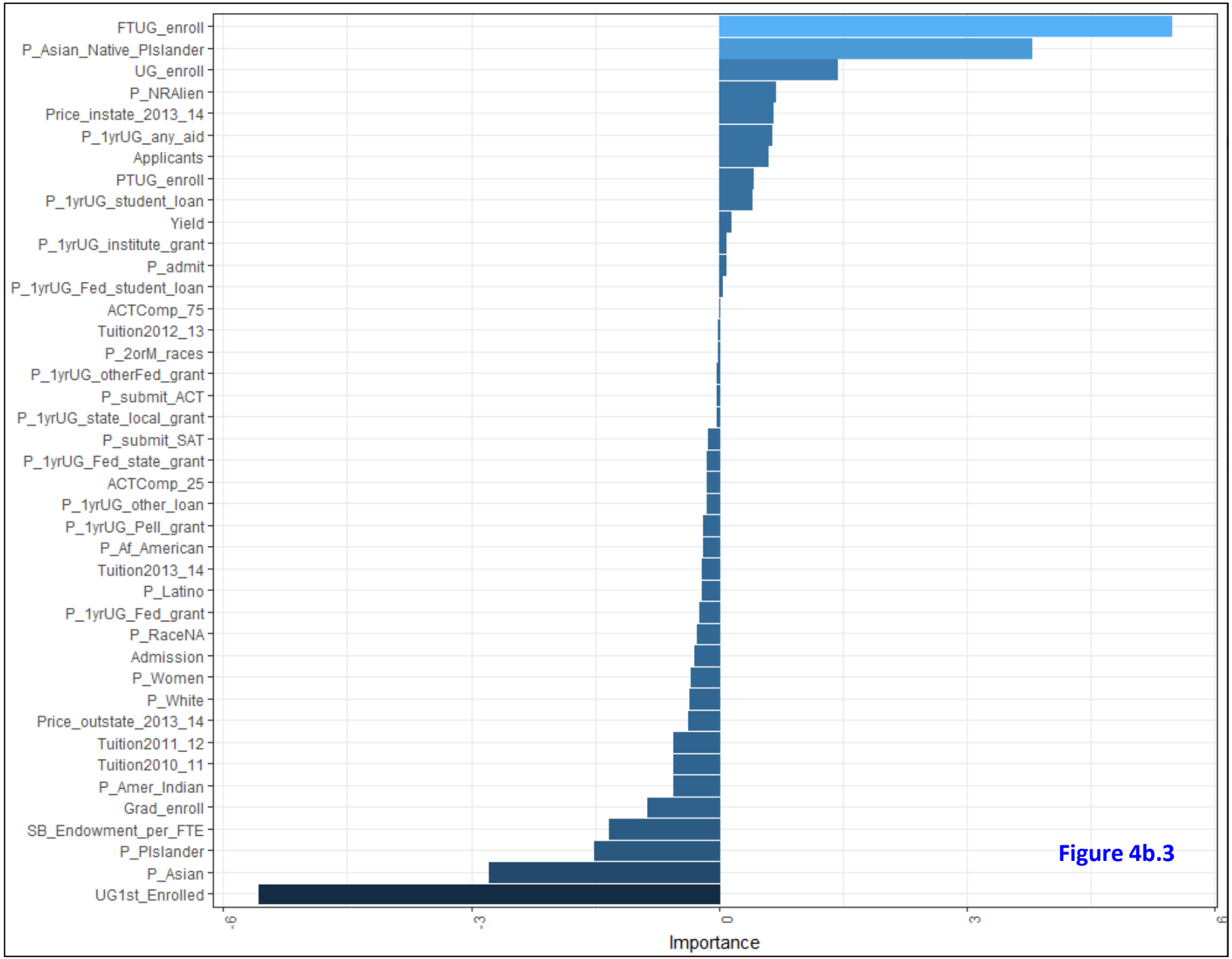


Figure 4b.3

h. Compare top 15 features (affecting 4-yr graduation rate) with RF model

Figure 4b.4 compares the top 15 features between the RF (4a) and the NN (4b) models.

The cells highlighted in yellow contain common features between the NN and RF models. It seems that out of top 15 variables, 8 features are shared between NN and RF model.

FINAL NOTE

RF and NN models are black box models – there are no known (phyics-based) laws that govern the relationship between the multiple input features and the label (here, the increase in graduation rates between 4 and 6 years). Thus, like in any social science research, it is difficult to establish causality between a feature and the label – one can only provide educated guesses as to why certain features rank high on the variable importance plots, and that is what I have tried to do with my explanations.

Figure 4b.4

	RF model - 4a	NN model (numeric) - 4b
1	Tuition2010_11	UG1st_Enrolled
2	Tuition2012_13	FTUG_enroll
3	Tuition2011_12	P_Asian_Native_PIslander
4	Tuition2013_14	P_Asian
5	PTUG_enroll	P_PIslander
6	Price_instate_2013_14	UG_enroll
7	Status	SB_Endowment_per_FTE
8	P_Women	Grad_enroll
9	UG_enroll	P_NRAlien
10	FTUG_enroll	Price_instate_2013_14
11	Grad_enroll	P_1yrUG_any_aid
12	SB_Endowment_per_FTE	Applicants
13	Admission	P_Amer_Indian
14	P_NRAlien	Tuition_2010_11
15	P_submit_ACT	Tuition_2011_12