Summer 2020

Exploring and understanding factors that determine first time full time undergraduate enrollment besides high school GPA, Class Rank & SAT/ACT

Pallabi Chatterjee

Exploring and understanding factors that determine first time full time undergraduate enrollment besides high school GPA, Class Rank & SAT/ACT

Pallabi
Chatterjee

2020

Master Professional Report submitted to the Faculty of the Iowa State University, Ames
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN INFORMATION SYSTEMS

Advisory Committee:
Dr. Anthony Townsend

ABSTRACT

Universities, these days, are more and more focused in determining the characteristics that boost their enrollment. These characteristics allow enrollment decision makers and leaders to determine the prospects or applicants having higher propensity to get admitted at their universities and subsequently to better administer their financial awards. In this paper we determine characteristics that influence the possibility of get admitted. For ages, traditional predictors like GPA, Class Rank & SAT/ACT scores have proved to be insightful whether a student could hope to get into a university or not. Although, in the last 10 years, these metrics have started meaning lesser and lesser. For example, in last 10 years SAT has been restructured twice, making it obscure and confusing for enrollment managers to examine, for example, if previous year's accretion in mean scores were an outcome of improved learners or it was merely a diverse test. On top of that, almost half of teens in America are now finishing their high school with a grade A on average, as per recent studies. Hence institutions are compelled to make nonviable choices, allocating a set number of admission slots to an increasing number of pupils who, each and every year, are tougher to distinguish using these traditional systems of measurement. The aim of this paper is to identify and assess factors that determine first time full time undergraduate enrollment, in 4-year public institutions, besides high school GPA, Class Rank & test scores in model-based admission projection for U.S. higher education.

ACKNOWLEDGEMENTS

First of all, I would like to thank my major advisor Dr. Anthony Townsend, for his utmost patience and support throughout the writing of this report.

I would especially like to acknowledge Mr. Greg Forbes of Enrollment and Research Team, under whom I did my research assistantship for two years, which gave me incredible opportunity to learn about various aspects of college enrollments in United States.

I would also like to thank my employer, AKC Marketing & Consulting Inc. and specifically my manager, Phillip Stanhope, for always encouraging me to push beyond my limits and gave me free time to work on this component.

Furthermore, I wish to acknowledge all my faculties, staffs at Iowa State university that immensely helped me in pursuit of knowledge.

Last but not the least, I would also like to thank my mom Mrs. Dalia Chatterjee and my dad Mr. Gautam Chatterjee and my brother Rahul Chatterjee for their love, trust, care and support as I worked towards completing this important personal goal.

Table of Contents

INTRODUCTION

Institutions while providing admissions to students are facing merit crisis: As ACT and SAT scores along with High School GPAs hold less domination, enrollment leaders are looking for other, inescapably more subjective measure.

Each and every year, the organizations constituting enrollment managers on colleges questions its administration about the certain characteristics they contemplate while deciding about enrollment. Although, ACT and SAT scores, HS grade point average and the robustness of a student's high-school syllabus and modules remain the topmost criteria. However now a days there are evidences that other factors are becoming more crucial in order to determine enrollment than they used to: Applicants' "exhibited affinity" towards getting into a specific college, are determined via their visits to the institution or by a college's student teacher ratio. On top of that, enrollment leaders in several colleges conducted surveys that showed that a students' "affordability to bear tuition cost" is of some significance while making a decision to enroll in a college.

Although there are substantial writings on factors predicting enrollment in institutions, majority of the literature discusses traditional predictors. To put in another words, the writings have been inclined to focus more on identifying students tests scores, GPA, students traits rather than the characteristics of Colleges like in state tuition, out of state tuition, whether a college is land grant or not, student faculty ratio. It is commonly thought that the model that best fits the sample data is the best forecasting model. This, however, is often not the case (see Elaine M. Allensworth (2020)). The applicability of ACT/SAT scores have been specifically questioned during last couple of years, strikingly due to discussions that these examinations are biased against some populations (Banerji; Linn, Greenwood, and Beatty). The very reason that caused multiple institutions not wanting these standardized test scores anymore for admission. Another literature also mentioned that "for first year CGPA, neither the SAT nor HSGPA was able to predict successful students" (Kobrin and Michel 6).

Those days are far gone when a student in high school would show up in a university event and expect to be selected by a college merely on the basis of his test scores and grade point average. There is a need to create a balance of thorough excellence of scores and grades which determines the direction of parallel growth.

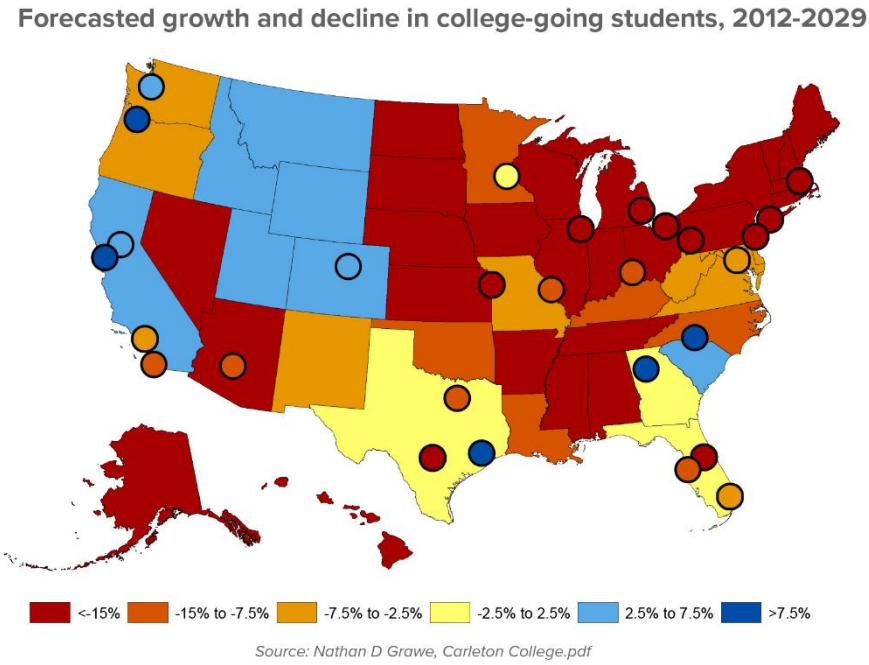
A substantial amount of colleges all across USA are of similar opinion regarding standardized tests scores, not making an important criterion as college enrollment or to drop them altogether. There are substantial justifications that resonate these schools' judgements: for example, family income and these standardized scores are highly correlated Dahl and Lochner (2012).

While HSGPA and ACT SAT test scores do hold importance if we are trying to estimate enrollments and student success for different populations and in different programs, unfortunately we are not doing so. Regardless of the considerable diversification in the construction, purposeful outcomes, anticipation, traits, and standards of variety of programs and colleges all across USA, we have a strangely lesser and lesser way for estimating enrollment and success.

Importance of POE (prediction of enrollment)

Predicting enrollment in universities and colleges has become more important than ever. As per CDC data, back in 2008 during great recession, the natality rate plummeted which decision makers accredited to the US economy. Despite the US economy has eventually recoiled in the past 10 years since, the natality rate has not.

The result of a dwindling U.S. natality rate is extensive, one of the most important things is how the institutions across US could be affected by a nosedive in the age group of ready-to-college students in the overall citizens. This emerging enrollment crisis is gradually becoming the talk of the town among presidents' top advisors and boards of trustees all over US, as college enrollment leaders look to get in front of this challenge.



The future for enrollment in colleges and universities looks bleak. The high school graduates in united states is predicted to persist comparatively flat for many upcoming years before jumping up a little bit in the middle of next 5 years. However, the worrisome part is from 2026 to 2031 the rate of high school graduates is estimated to plummet by 9 %.

Further on the general numbers undergoing haul, the classes of high school graduates would see more and more diversity. We would see a less of white students and more of Hispanic students, as per US census data, and a broader scope of intellectual capabilities. Students' financial need would soar high whereas their family income might remain stagnant. In another words, the upcoming decade will be tempestuous for higher ed enrollment.

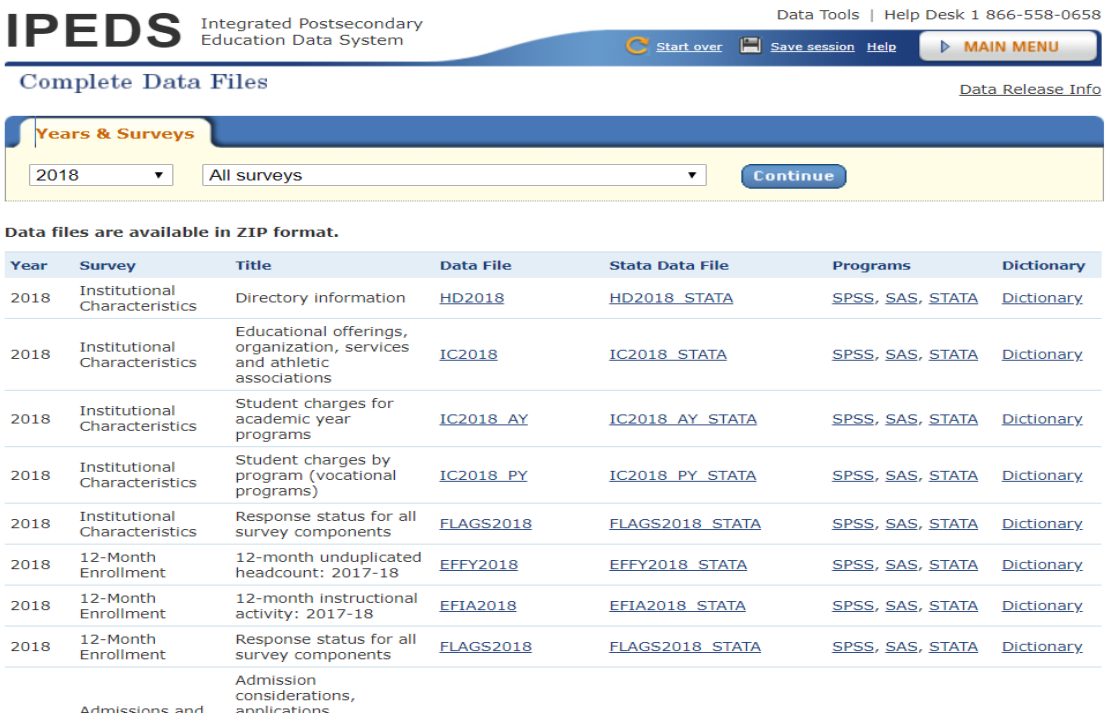
The aim of this research is to study, identify and assess what factors affect enrollment in 4 or more-year traditional schools beyond traditional predictors. What relationship do *in state, out of state tuition* has on enrollment. What role do *retention rate, land grant or non-land grant, financial aid* play in determining enrollment in a schools in USA? So that colleges and universities prepare and take measures beforehand for a possible decline in tuition dollars in the coming years.

Data Description

The variable “Total Fall Enrollment” (numerical) data is from 2018 IPEDS datasets. IPEDS is a classification of survey components that gathers statistics from about 6,400 organizations that deliver postsecondary education across the United States.

IPEDS is an organization of interconnected studies and surveys conducted every year, which collects data from each and every college, academia, and technical institution in USA and other that partakes in the federal learner economic aid programs. Higher Education Act, as revised, entails that organizations which partake in federal student financial aid plans account statistics on admissions, curriculum accomplishments, completion rates, teaching staff and workforce, funds, official costs, and learner financial aid. These statistics have become accessible to learners and parents, and to scholars. IPEDS offers fundamental information required to explain — and evaluate patterns via in — postsecondary learning in USA, in terms of pupils registered, financial aid applied, staff in a job, cash exhausted, and grades and diplomas received. Congress, federal bureaus, state administrations, education benefactors, skilled and qualified relations, private companies, media, learners, and their parents, count on IPEDS records for rudimentary info. IPEDS data is utilized at national and state level intended for strategy assessment and expansion and at the official stage for yardstick and cohort evaluation. It also shapes the established testing framework for further NCES higher education inspections.

So, the first step of the research was to collect relevant data from IPEDS for analysis. This was done building a script in Python in order to programmatically access the data recursively from various pages. I also leveraged the application programmable interface (API) provided by NCES. The python script collected the needed data of variety of factors such as enrollment, cost of attendance, financial aid etc. on various traditional and nontraditional colleges.



IPEDS Integrated Postsecondary Education Data System

Data Tools | Help Desk 1 866-558-0658

Start over | Save session | Help | MAIN MENU

Complete Data Files

Data Release Info

Years & Surveys

2018 | All surveys | Continue

Data files are available in ZIP format.

Year	Survey	Title	Data File	Stata Data File	Programs	Dictionary
2018	Institutional Characteristics	Directory information	HD2018	HD2018_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Educational offerings, organization, services and athletic associations	IC2018	IC2018_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Student charges for academic year programs	IC2018_AY	IC2018_AY_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Student charges by program (vocational programs)	IC2018_PY	IC2018_PY_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Response status for all survey components	FLAGS2018	FLAGS2018_STATA	SPSS, SAS, STATA	Dictionary
2018	12-Month Enrollment	12-month unduplicated headcount: 2017-18	EFFY2018	EFFY2018_STATA	SPSS, SAS, STATA	Dictionary
2018	12-Month Enrollment	12-month instructional activity: 2017-18	EFIA2018	EFIA2018_STATA	SPSS, SAS, STATA	Dictionary
2018	12-Month Enrollment	Response status for all survey components	FLAGS2018	FLAGS2018_STATA	SPSS, SAS, STATA	Dictionary
	Admissions and	Admission considerations, applications				

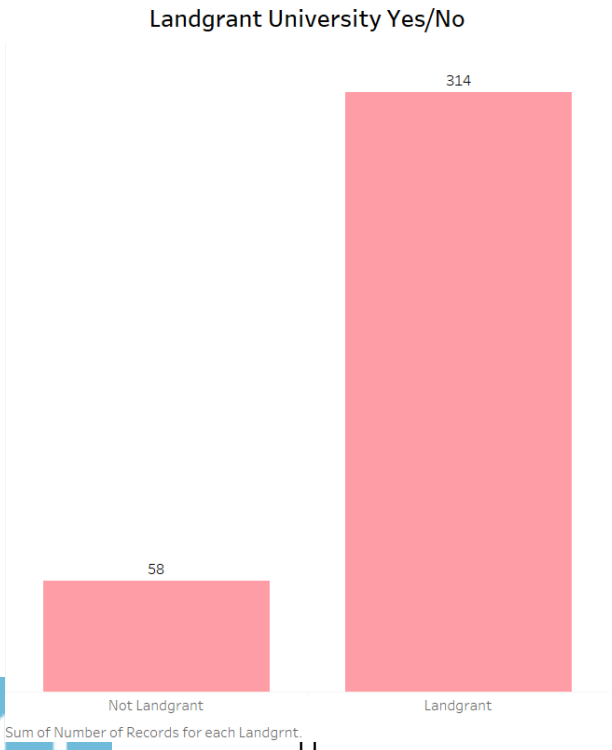
I preprocessed and transformed several datasets with various components of interests into one digestible format: a single csv file. An excerpt of data looks like this below:

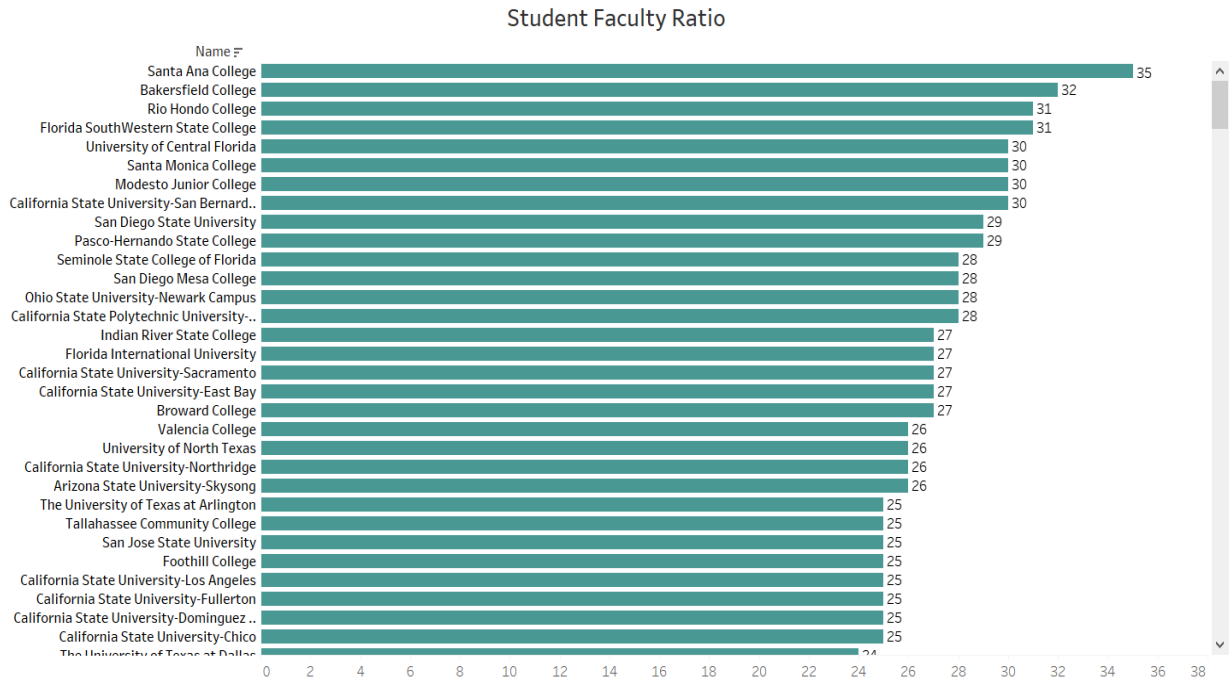
Name of the Institution	Land grant	Retention Rate	Student Faculty Ratio	In State Tuition	Out State Tuition	Percent of Stdnt Having Financial Aid	Total Fall Enrollment
California State University-Los Angeles	1	79	29	6383	17543	91	3675
Georgia Institute of Technology	1	97	20	12212	32404	76	2873
Iowa State University	0	88	19	8219	21583	88	6145
Michigan State University	0	91	17	14062	37890	67	8005

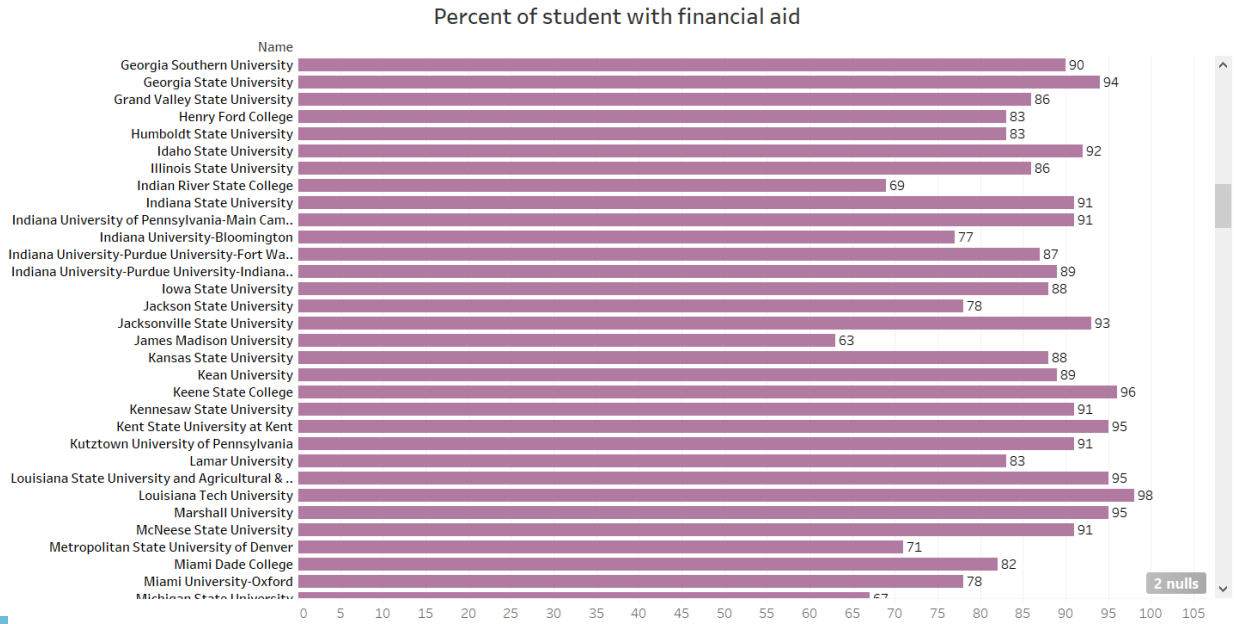
My sample consists of 300 randomly chosen, 4 or more-year public and private (for nonprofit) schools of United States of America. It has the following variables of interest:

1. **control** - whether a school is public or private-not for profit (categorical)
2. **landgrnt** - whether a school is a Land Grant or Not a Land Grant school (categorical)
3. **ft_ret_rate16** - full time retention rate for the fall of 2018 (unit in percentage) (quantitative)
4. **st_fac_ratio** - student faculty ratio (quantitative)
5. **grand_total** - total undergraduate first time full time enrollment for the fall of 2018 (dependent variable)
6. **in_st_tuifee** - in state tuition and fees (quantitative)
7. **out_st_tuifee** - Out of state tuition and fees (quantitative)
8. **prcnt_std_UG_any_aid** - percent of undergraduate students granted any tuition aid (quantitative)
9. **Total fall enrollment** – total fall enrollment for the year 2018 (dependent variable)

Below is an overview of the data:







Methodology

This paper builds a regression model to examine and explain the relationship and predict the number of students enrolled in higher education institutions by use of regression analysis, more precisely multiple linear regression, as a part of methodology.

Multiple Linear Regression:

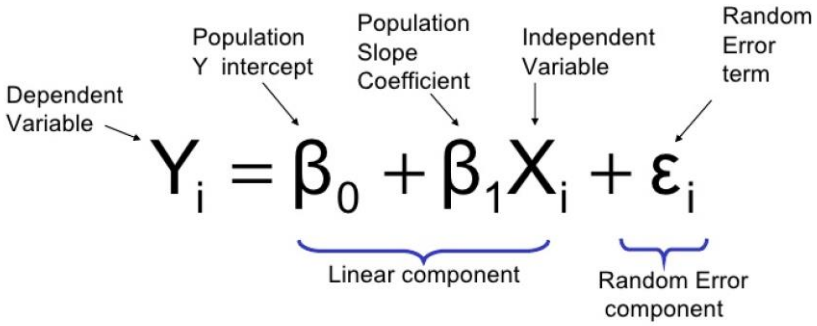
Multiple linear regression helps modelling linear relationship between dependent variable and its predictors. It is, in fact, the add-on of ordinary least-squares (OLS) regression that requires more than one independent variable.

We started of with building four models as shown below in R Studio. In the first model initially centered variables at the mean shifting the scale in order to avoid multicollinearity (Model A) as well as tried to see the interaction between Retention Rate and Instate Tuition, Retention Rate and Out of state tuition, Retention rate and percentage of students receiving financial aid in every college and Land grant and Percent of student aid. In Model B we removed all the interactions however just retained the centered variable. Model C is limited to being 2 predictor variable and Model D has one interaction between retention and out of state and four other predictors in order to maintain lower complexity.

	Model A			Model B			Model C			Model D		
Variable	Coef	pvalue	Confint	Coef	pvalue	Confint	Coef	pvalue	Confint	Coef	pvalue	Confint
Landgrant	-1160.00	5.06E-08	(-1561.00,-749.57)	-1110.00	1.67E-09	(-1464.00,-759.13)	-1253.73	5.33E-10	(3469.55,4168.06)	-1077.00	3.31E-09	(-1425.10,-728.20)
Retention Rate	88.39	6.00E-05	(45.63,131.20)	58.92	2.48E-10	(41.17,76.68)	96.99	2.00E-16	(-1639.30,-868.16)	73.80	2.00E-16	(57.30,90.26)
Student to Faculty Ratio	118.20	2.10E-08	(77.70,158.71)	123.00	8.98E-09	(81.99,164.03)				139.60	1.38E-11	(100.4,178.9)
In State Tuition	0.09	2.54E-01	(-0.07,0.25)	-0.08	1.96E-02	(-0.14,-.013)						
Out State Tuition	-0.01	7.43E-01	(-0.09,0.06)	0.10	1.66E-10	(.07,.126)				0.06	2.52E-05	(0.03,0.0)
Percent Aid	-72.07	1.79E-04	(-109.40,-34.67)	-24.08	1.95E-03	(-39.25,-8.91)						
Retention*In State	-0.01	4.25E-03	(-0.02,-0.003)									
Retention*Out State	0.01	3.63E-04	(0.003,0.01)							0.004	5.88E-05	(0.002,0.006)
Retention*Percent Aid	1.26	1.20E-01	(-.32,2.85)									
Student Fact*Percent Aid	-1.59	4.26E-01	(-5.50,2.33)									
Landgrant*Retention	-23.04	3.44E-01	(-70.90,24.80)									
Landgrant*In State	-0.18	3.76E-02	(-0.35,-0.01)									
Landgrant*Out State	0.11	9.81E-03	(0.03,0.19)									
Landgrant*Percent Aid	69.26	6.93E-04	(29.50,109.04)									
sqret_rate16	1.43	5.47E-02	(-0.03,2.89)									
Intercept	3610.00	2.00E-16	(3197.10,4018.80)	3701.00	2.00E-16	(3385.64,4018.81)	3818.81	2.00E-16	(3469.55,4168.06)	3534.00	2.00E-16	(3212.04,3855.40)
Radj	0.59			0.54			0.41			0.54		
F-Test	33.54	2.20E-16		67.49	2.20E-16		117.20	2.20E-16		80.87	2.20E-16	

Since the Adjusted R square came out to be similar for all the models also after a series of tests such as centering the variables, performing backwards elimination on the full model, trying various interactions we choose Model D, that performs fairly well with lower complexity.

Multiple Linear Regression Equation:



PredEnrollment = 3534 - 1077(Landgrant1) + 73.80(centered retention rate) + 139.60(centered stud-fac ratio) + 0.55(centered out of state tuition) +0.0041(centered retention*centered out of state tuition)

Hypotheses Test Results

Landgrant:

Ho: b1=0

Ha: b1≠0

Test statistic=-6.079

P-value = 3.31e-09

Reject Ho since 3.31e-09 < 0.05.

Centered Retention Rate:

Ho: b2=0

Ha: b2≠0

Test statistic=8.817

P-value = <2e-16

Reject Ho since <2e-16< 0.05.

Centered Student-Faculty Rate:

Ho: b3=0

Ha: b3≠0

Test statistic=7.003

P-value = 1.38e-11

Reject Ho since 31.38e-11 < 0.05.

Centered Out of State Tuition:

Ho: b4=0

Ha: b4≠0

Test statistic=4.273

P-value = 2.52e-05

Reject Ho since 2.52e-05 < 0.05.

Centered Retention Rate*Centered Out of State Tuition:

Ho: b5=0

Ha: b5≠0

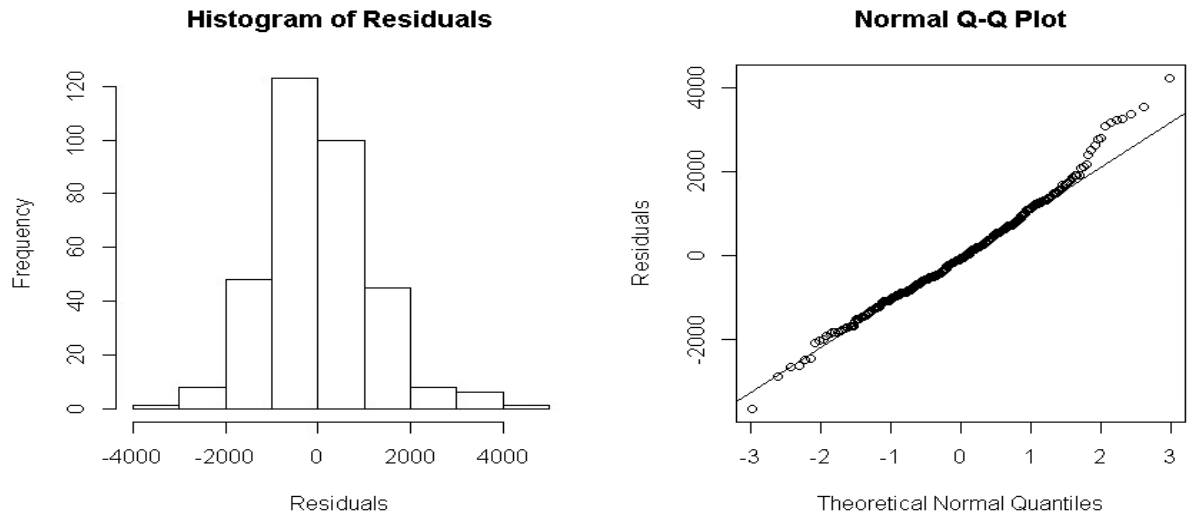
Test statistic=4.070

P-value = 5.88e-05

Reject Ho since 5.88e-05 < 0.05.

Landgrant, Centered Retention Rate, Centered Student- Fac Rate, Centered Out of State Tuition, and Centered Retention Rate*Centered Out of State Tuition are all significant predictors of first time, full-time undergraduate enrollment.

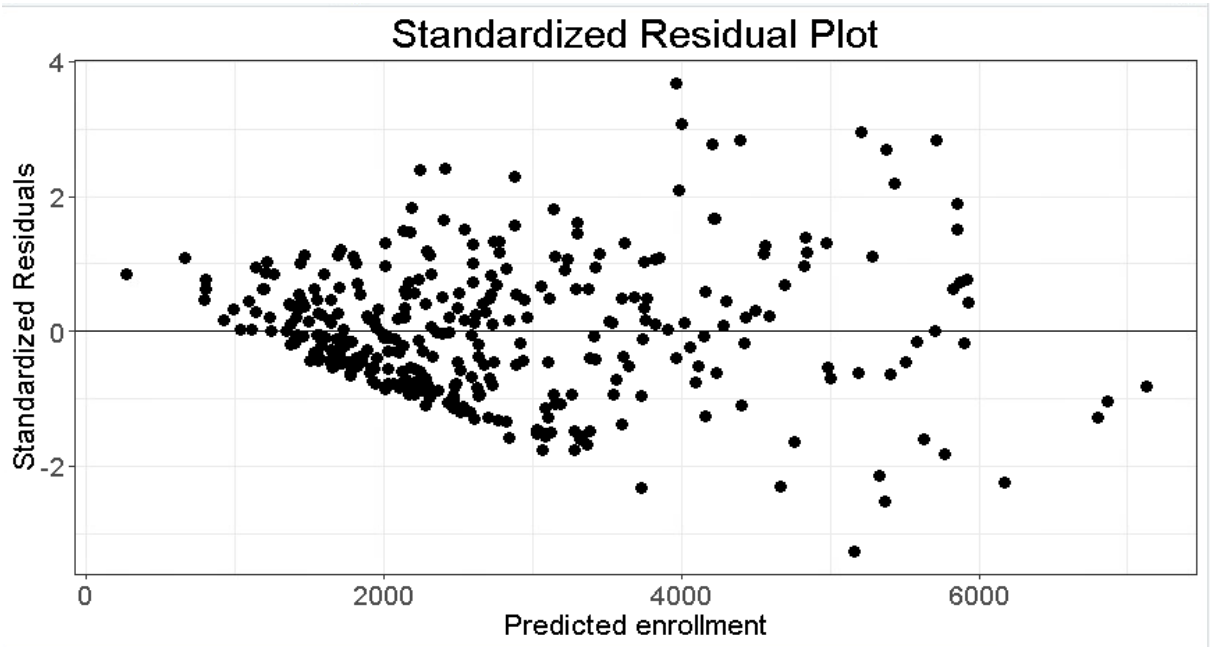
Assumptions



The normality hypothesis is assessed depending on residuals and can be estimated employing a **QQ-plot** by linking the residuals to "ideal" normal observations along the 45-degree line.

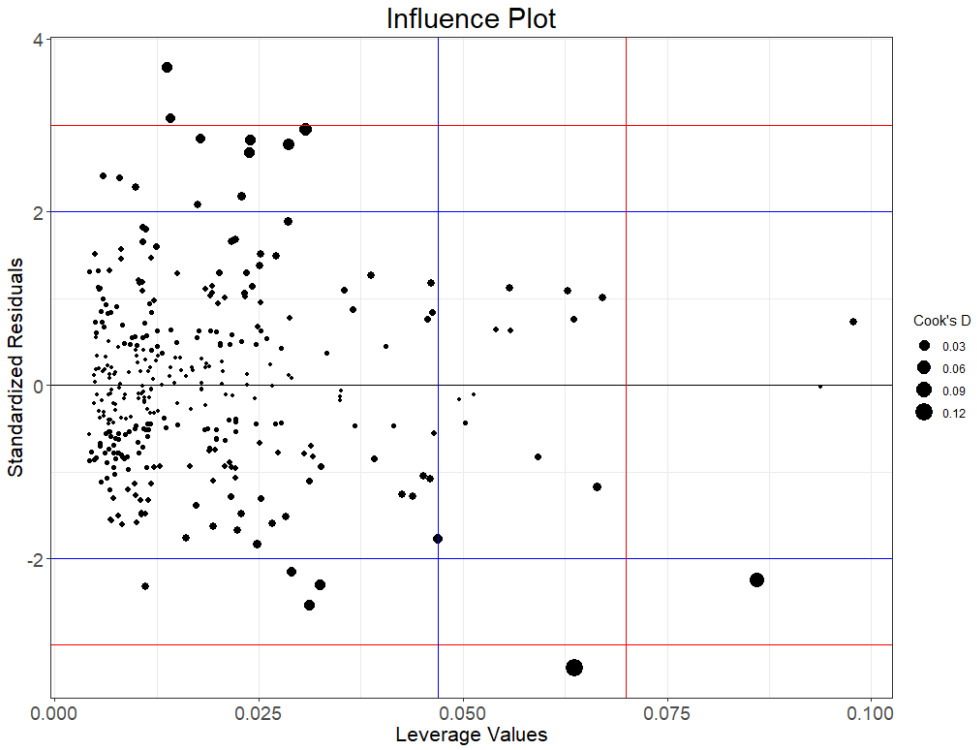
As you can see from the figure above that my program automatically flagged a couple of data points that have large residuals (observations more than 2300). However, apart from those numbers, observations remain fairly alongside the 45-degree line in the QQ-plot, so **we may accept that normality maintains here.**

The above histogram plot is shown here to see if the variance in my data is, fairly, normally distributed. Usually, a bell-shaped histogram that is uniformly dispersed all over the place zero implies that the normality assumption is expected to be accurate. If the histogram shows that arbitrary error is not normally dispersed, it indicates that the model's core hypotheses may possibly have been flouted. So as you can see from the graph above **Histogram of the Residuals demonstrating that the variation is normally distributed.**



As shown above that the residuals in the plot are not evenly distributed showing a clear shape. The plot demonstrates "heteroscedasticity," suggesting that the residuals become larger as the estimation shifts from insignificant to sizable (or from sizable to insignificant) which suggests this model has scope for improvement.

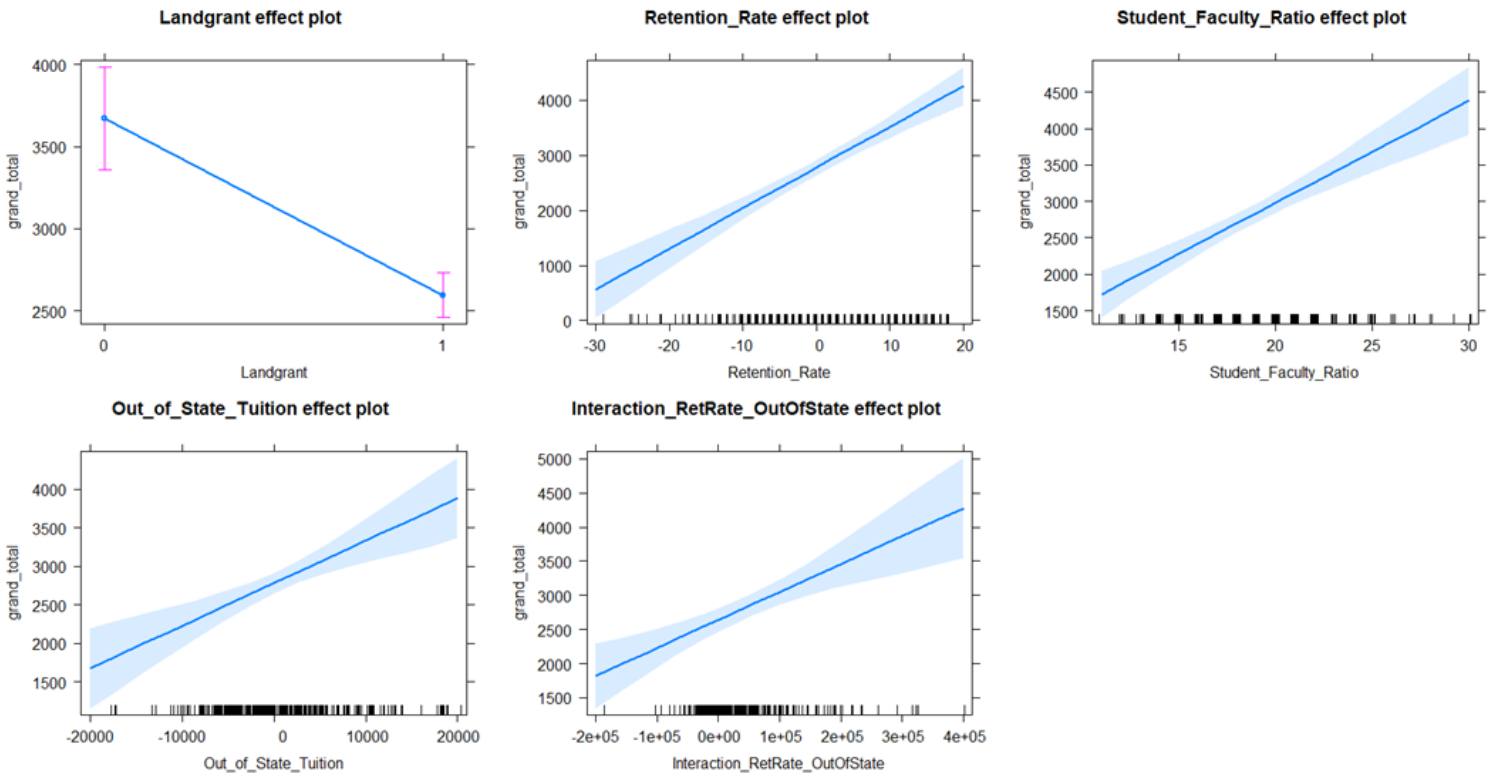
Influence Plot



- **Residuals**
  - # Somewhat High=14
  - # Unusually High =3
- **Leverage**
  - # Somewhat High=12
  - # Unusually High =3
- **Cook's D**
  - No influential points

An influence plot illustrates the outliers, leverage, and influence of every single case.
The plot demonstrates the residual on the perpendicular axis, leverage on the plane axis, and the point volume is the square root of Cook's D number, a measurement of the effect of the point.

Graph Summarizing Our Chosen Model



Based on our research, we have learnt that fall enrollment in universities have positive linear relationship with retention rate, student-faculty ratio, in/out state tuition. We could successfully build a model that showed that besides traditional predictors enrollment leaders do need to take into account the above mentioned variables which would provide them with better insights. However further research is required to analyze bigger and broader sample along with the possibility of having nonlinear relationship of the variables in order to observe enrollment trends.

Conclusion

In this study I casted light on the relationship between nontraditional predictors and fall enrollment. I used multiple linear regression approaches to infer the level of association between enrollment and such aspects. Specifically, I approached predicting enrollment numbers at universities from a regression point of view where I must recognize the chances of enrollment for a collection of colleges. Using this methodology, I directly determined the number of enrolled students that would register without identifying them personally. The findings show that my recommended models can forecast enrollment with consistent accurateness using only a small set of traits related to institution characteristics.

Future Scope

Predicting human behavior is not always straightforward. A good extent of effort is done in exploring and forecasting enrollment, but all these findings have used traditional variables like ACT SAT test scores and GPA. There is a clear need to look beyond the age-old traditional variables and start examining other factors that influences enrollment. Other than educational aspects, there are significant amounts of factors that perform substantial part in projection, which incorporates institutional attributes. Proper data modelling methods are required to assess, scrutinize, and deduce these factors for prediction. Thus, enhancing the key vector with qualitative standards may boost the precision rate of projection as well.

References

https://nces.ed.gov/ipeds/DFR/2019/ReportHTML.aspx?unitid=180902

Broyles, Susan G. (1994). Integrated Postsecondary Education Data System : IPEDS. [Washington, DC] :U.S. Dept. of Education, Office of Educational Research and Improvement, Educational Resources Information Center : National Center for Education Statistics.

https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx

https://ademos.people.uic.edu/Chapter12.html

https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

https://www.theatlantic.com/education/archive/2018/05/college-admissions-gpa-sat-act/561167/

https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx

https://knocking.wiche.edu/nation-region-profile

https://hechingerreport.org/college-students-predicted-to-fall-by-more-than-15-after-the-year-2025/

https://www.bloomberg.com/opinion/articles/2019-05-30/college-enrollment-bust-is-headed-this-way-by-2026

https://www.usnews.com/news/education-news/articles/2019-05-30/nationwide-college-enrollment-is-down-again