Che Cobb, Joe Schueder, Anthony Egbuniwe
Professor Anthony Tanaydin
SMU-DS 6372
Feb 17, 2019

Graduate Admissions to US Schools for Indian Candidates

Introduction

Every year almost a hundred thousand international students from India apply to American universities for admission to graduate study. However, only a small percentage of these students succeed. Indeed, the decision of whether to admit a student into U. S. graduate programs is an important challenge for both the students and universities.This is due to the fact that most admission procedures involve the rank ordering of students with top-down selection based on TOEFL or GRE scores and College GPA. For Universities, having the ability to predict the success of students when they enter a graduate program is critical in promoting the success of their program.

Section I. Descriptive Statistics

This dataset was created for the prediction of Graduate Admissions from an Indian perspective. It can be thought of as an observational study, thus, conclusion in this analysis will be speculative outside of the data used. “The dataset contains several parameters which are considered important during the application for Masters Programs. This dataset is inspired by the UCLA Graduate Dataset. The test scores and GPA are in the older format. The dataset is owned by Mohan S Acharya. This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.” Mohan S Acharya, Asfia Armaan, Aneeta S Antony. (January 27, 2019). A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019. Retrieved from https://www.kaggle.com/mohansacharya/graduate-admissions#Admission_Predict_Ver1.1.csv

The data was generated by Mr. Acharya based on a UCLA data set sampling and survey.

The parameters included are : 1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of

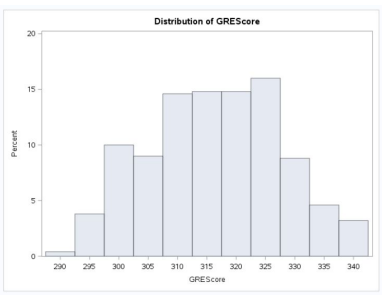
Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1). The data contains five hundred observations.

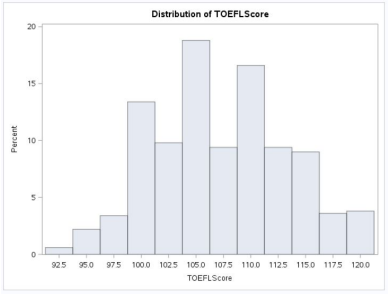
Summary statistics of all the data is shown in Table 1 and Table 2. The data was populated for all rows and all columns. GRE Scores, TOEFL Scores, University Rating, GPA are continuous numeric variables with a somewhat normal looking distributions. Statement of Purpose is a continuous numeric also, though slightly left skewed. Research is a yes, no variable represented by 0 and 1.

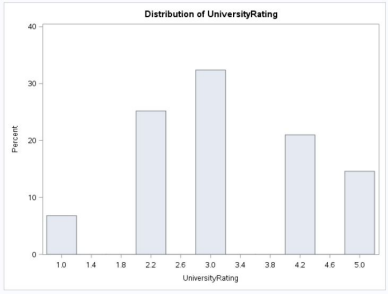
Table 1

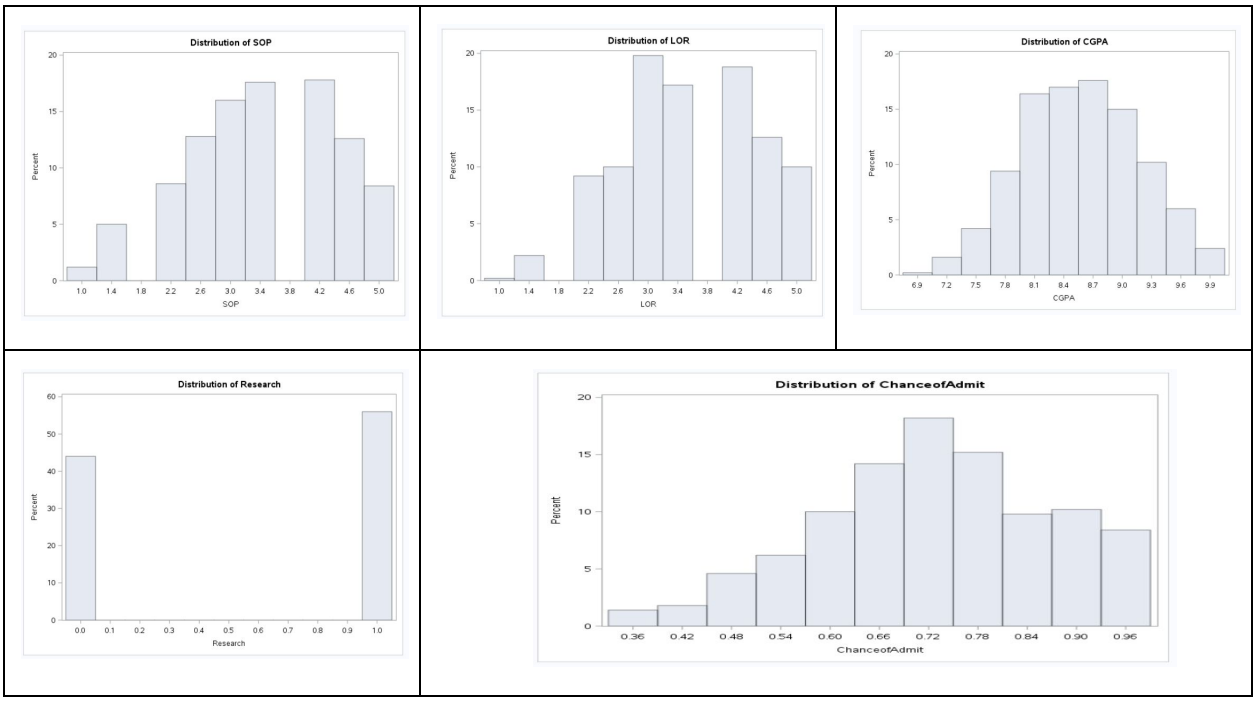
Variable	Observations	Mean	Median	Standard Deviation
GRE Scores	500	316.47	317	11.95
TOEFL Scores	500	107.19	107	6.058
University Rating	500	3.11	3	1.14
Statement of Purpose	500	3.374	3.5	.991
Letter of Recommendation	500	3.48	3.5	.925
Undergraduate GPA	500	8.576	8.56	.60481
Research Experience	500	.56	1	.4968
Chance of Admit	500	.72174	.72	.1411

Table 2









An exploratory analysis was done to check the relationships between the variables contained in the data. A graphical representation is shown in Image 1. As can be seen many of the variables have linear relationships with each other. For all these relationships the relationship is positive: as one increases, so does the other.

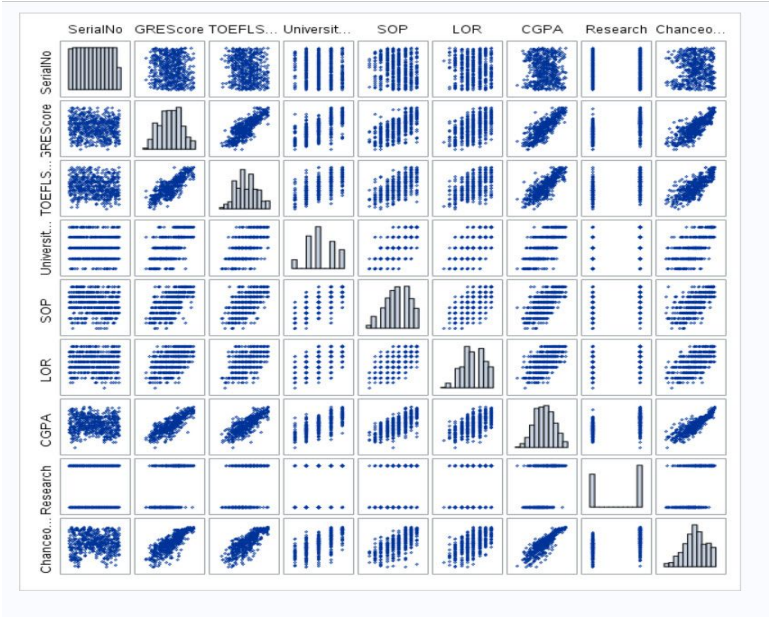


Image 1 - Variable Relationships

Based on how the values of the variables are distributed, the sample size and the positive relationships between the variables it seems like a multiple linear regression model will help predict the likelihood of acceptance into a U.S. University.

Section II. Analysis

The objective is to identify which factors play the most important role in predicting what will help a student get accepted into graduate school. The outcome will be to be able to predict the response variable 'Chance of Admit (ranging from 0 to 1)'. A zero represents no chance of admittance while a 1 represents a hundred percent chance of admission. Regression analysis was chosen to create the predictive model. The potential student should be able to input their academic profile details to determine whether they can expect admittance. A sampling of fit statistics for each method is shown below. Although ultimately all have closely matching statistics, predictor variables and estimates, Lasso chose the most accurate model based on CV press with the least number of predictors. With this data set there was very little differentiation between these models and any could be reasonably used.

Table 3

	Lasso(Remove Research)	Forward(all variables)	Stepwise(all variables)
Root MSE	.05913	.058	.058
Adj- R-square	.8280	.833	.836
AIC	-1853.42	-1867	-1867
SBC	-2227.48	-2246	-2246
CV Press	1.43	1.393	1.39

Linear regression models have some assumptions that must be met to ensure it is the right model to use. See Image 2 for assumption analysis. The Residual, Studentized Residual, and QQ Plot of residuals show the residuals are tighter at the higher end of prediction(close to 1) than the lower end(close to 0). Log transformations

were attempted to improve the outcome, but this did not improve the residual plots. This most likely means that there are additional predictors that were not included in the data set. These could include factors such as age of the applicant, actual school sought, and work experience. As for Influential Points examining the scatterplot of the data, it appears that there are not significant outliers. Cook's D is also low for this data. This indicates that there are not any influential data points.

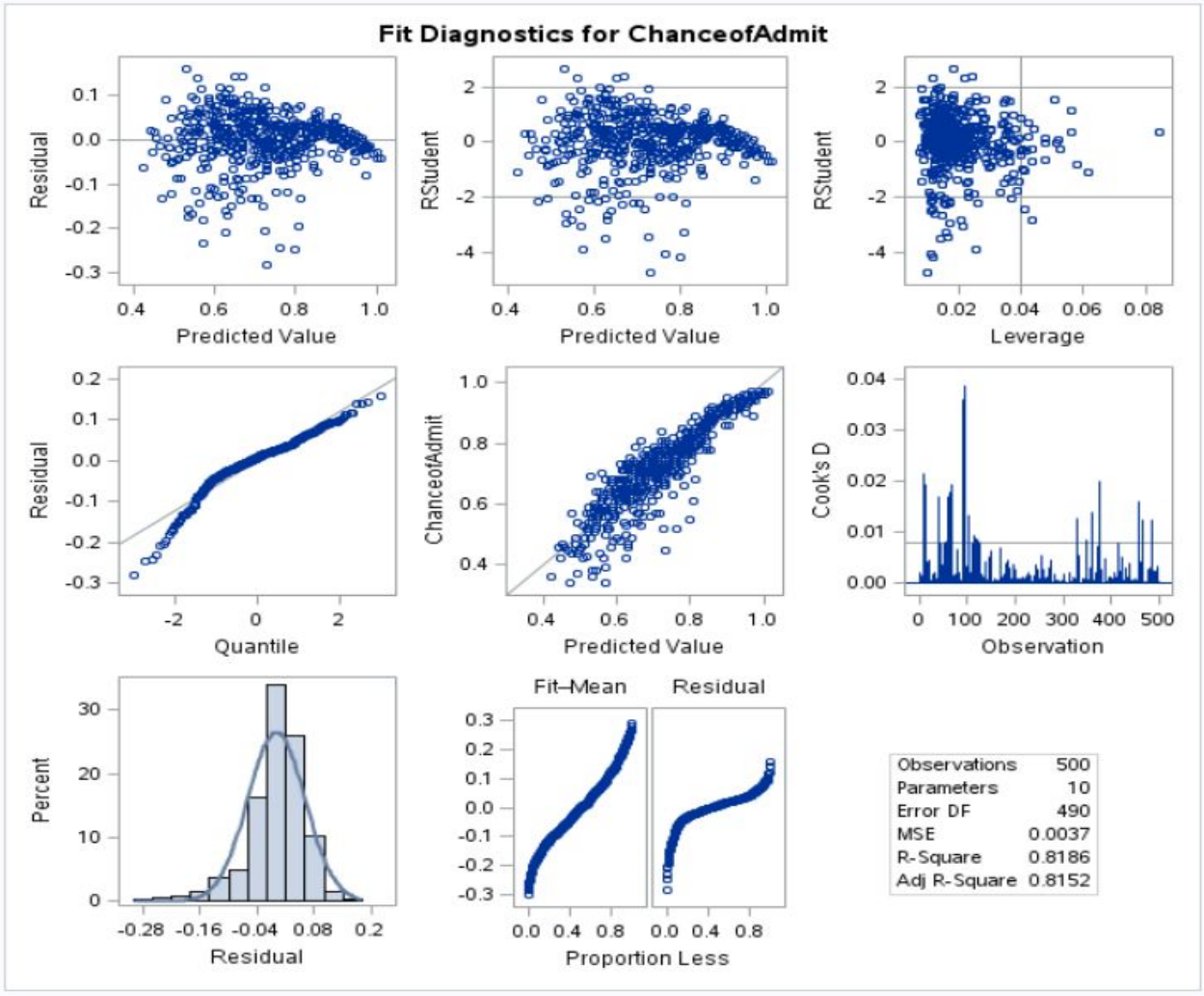


Image 2

Image 3 shows a report produced in SAS showing the order in which variables were selected:The GLMSELECT Procedure:

The GLMSELECT Procedure

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	ASE	Test ASE	CV PRESS
0	Intercept		1	0.0203	0.0183	8.2664
1	CGPA		2	0.0084	0.0088	1.7083
2	GREScore		3	0.0055	0.0064	1.5376
3	TOEFLScore		4	0.0047	0.0057	1.5133
4	LOR		5	0.0042	0.0052	1.4412
5	SOP		6	0.0036	0.0046	1.4402
6	UniversityRating_5		7	0.0034	0.0045	1.4303*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	UniversityRating_4	1.4485	>	1.4303

Image 3

From Image 3, it is observed that the regression included all variables in the model. It can be noticed that the variables related to CGPA, GRE Score & TOEFL Score had the most correlation with the target variable(Chance of Admit). We can see the addition of variables at each step and the growth of the coefficients from the coefficient progression curve given by SAS(Image 4). Additionally it can be observed that the effect on prediction error increase as the model grows more complex.

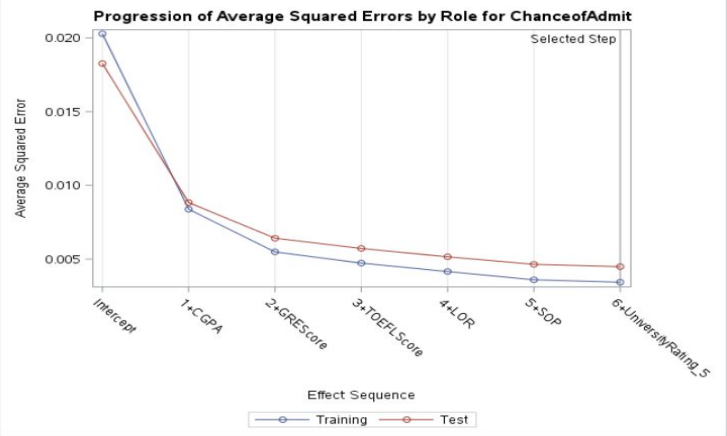


Image 4

The selected model and confidence intervals, based on Cross Validation, are shown in Table 5. The Parameter Estimates section shows the predictors and their contribution to the response variable.

Table 5

The GLMSELECT Procedure
Selected Model

The selected model, based on Cross Validation, is the model at Step 6.

Effects:	Intercept UniversityRating_5 GREScore TOEFLScore SOP LOR CGPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	6	6.73981	1.12330	321.24
Error	393	1.37424	0.00350	
Corrected Total	399	8.11405		

Root MSE	0.05913
Dependent Mean	0.72248
R-Square	0.8306
Adj R-Sq	0.8280
AIC	-1853.42612
AICC	-1853.05784
SBC	-2227.48587
ASE (Train)	0.00344
ASE (Test)	0.00449
CV PRESS	1.43033

Cross Validation Details

Index	Observations		CV PRESS
	Fitted	Left Out	
1	343	57	0.3129
2	325	75	0.3598
3	312	88	0.2585
4	298	102	0.2861
5	322	78	0.2129
Total			1.4303

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-1.380791942	B	0.10444286	-13.22	<.0001	-1.586003058	-1.175580827
GREScore	0.002353318		0.00048879	4.81	<.0001	0.001392941	0.003313695
SOP	0.002632009		0.00462862	0.57	0.5699	-0.006462383	0.011726401
TOEFLScore	0.002674558		0.00088289	3.03	0.0026	0.000939840	0.004409276
UniversityRating 1	-0.018944493	B	0.01739436	-1.09	0.2766	-0.053121223	0.015232237
UniversityRating 2	-0.029914609	B	0.01251585	-2.39	0.0172	-0.054505975	-0.005323243
UniversityRating 3	-0.022525949	B	0.01043608	-2.16	0.0314	-0.043030931	-0.002020967
UniversityRating 4	-0.019110301	B	0.00958508	-1.99	0.0467	-0.037943232	-0.000277370
UniversityRating 5	0.000000000	B	.	.	.	.	.
LOR	0.017588788		0.00417936	4.21	<.0001	0.009377104	0.025800471
CGPA	0.119054050		0.00983268	12.11	<.0001	0.099734634	0.138373467

Lastly, we can observe the final regression coefficients of the model parameters. Chance of Admittance = -1.381 + (.00235 X GREScore) + (.00267 X TOEFLScore) + (.011X University Rating 5) + (.00263 X Statement of Purpose) + (.0176 X Letter of Recommendation) + (.12 X Grade Point Average). Per the website College Countdown, the student should only consider applying to a school if their chance of admittance is greater than 50% and a 'sure thing' acceptance to be greater than a 75% chance. Understand your Chances of Acceptance. (February 17,2019). Retrieved from https://www.collegecountdown.com/choosing-the-right-college-for-you/understand-your-chances-of-acceptance.html If the student has a limited amount of resources for application fees, they should only consider applying to those schools with at least a 50% chance of admittance, but even better to apply for those with a 75% chance of admittance or higher. For example, a student with a GRE score of 298, TOEFL of 98, University Ranking of 2, SOP 1.5, LOR of 2.5, and CGPA of 7.5 has a predicted chance of admittance of .49 may want to be very selective about which schools they apply. While a person with GRE of 328, TOEFL of 119, University Rating of 5, SOP of 5, and

LOR of 4.5, and CGPA of 9.7 has a chance of admittance of .97 and can reasonably expect to be accepted to most schools they apply.

Section III. Interpretation & Conclusion

The team found that variables in the related to CGPA, GREScore & TOEFLScore showed the maximum correlation in predicting the target variable of Chance of Admit. Some variables in the original predictor set, like University Rating and SOP(Statement of Purpose) are less impactful and could be discarded from the selected model to maintain the bias-variance trade off at the optimum point. Since this dataset is limited to the Indian student perspective, we must limit the inference gained from this study to only the subjects of this sample.

Appendix:

```
FILENAME REFFILE '/home/jschueder0/Admission_Predict_Ver1.1.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=Admission4;
        GETNAMES=YES;

RUN;

proc print data=Admission4;
run;

proc sgscatter data=Admission4;
matrix SerialNo GREScore TOEFLScore UniversityRating SOP LOR CGPA
Research ChanceofAdmit / diagonal=(histogram);
run;

/* Split data randomly into test and training data*/
proc surveyselect data=Admission4 out=TrainTest seed = 123
samprate=0.8 method=srs outall;
run;
```

```
ods graphics on;
proc glmselect data=traintest plots=all seed=123;
partition ROLE=selected(train='1' test='0');
class UniversityRating Research;
model ChanceofAdmit  = GREScore TOEFLScore UniversityRating SOP LOR CGPA
Research
/ selection=LASSO( choose=CV stop=CV) CVdetails;
run;

ods graphics on;
proc glmselect data=traintest plots=all seed=123;
partition ROLE=selected(train='1' test='0');
class UniversityRating Research;
model ChanceofAdmit  = GREScore TOEFLScore UniversityRating SOP LOR CGPA
Research
/ selection=LASSO( choose=adjrsq stop=adjrsq) CVdetails;
run;


ods graphics on;
proc glmselect data=traintest plots=all seed=123;
partition ROLE=selected(train='1' test='0');
class UniversityRating Research;
model ChanceofAdmit  = GREScore TOEFLScore UniversityRating SOP LOR CGPA
Research
/ selection=forward( choose=CV stop=CV) CVdetails;
run;

ods graphics on;
proc glmselect data=traintest plots=all seed=123;
partition ROLE=selected(train='1' test='0');
class UniversityRating;
model ChanceofAdmit  = GREScore TOEFLScore SOP LOR CGPA Research
/ selection=stepwise( choose=CV stop=CV) CVdetails;
run;

ods graphics on;
proc glmselect data=traintest plots=all seed=123;
```

```
partition ROLE=selected(train='1' test='0');
class UniversityRating;
model ChanceofAdmit  = GREScore TOEFLScore SOP LOR CGPA
/ selection=LASSO( choose=CV stop=CV) CVdetails;
*/modelAverage tables=(EffectSelectPct(all) ParmEst(all)) alpha=0.1;  */ ;
run;
quit;

data Admission4;
set Admission4;
logUR = log(UniversityRating);
logCOA = log(ChanceofAdmit);
logLOR = log(LOR);
logGPA = log(CGPA);
logSOP = log(SOP);
run;



run;
quit;
ods graphics offproc glm data=Admission4  plots=(all);
ods graphics on;
proc glm data=Admission4  plots=(all);
class UniversityRating;
model ChanceofAdmit  = GREScore SOP         TOEFLScore UniversityRating LOR
CGPA / solutionclparm clparm ;
output out=new p=yhat r=resid stdr=eresid;
run;
quit;
```