Prediction Analysis on College Admission

Jialiang Lyu[1], Siyuan Qiao[2]

[1]Syracuse University

[2]Arizona State University

Keywords: Prediction; Analysis; College; Admission

Abstract: Collect data and get the probability of students being admitted. The first is to understand and clean up the data set to obtain the necessary variables needed to build the simulation process. We have a data file and a data set from a small university website, and after removing the duplicate IDs to obtain the necessary data, we filter out the necessary variables. The second is to calculate multiple variables to get the data, and iteratively prove it. The last is to find the most relevant variables and develop our predictive model. Through repeated verification and argumentation, we have arrived at the result.

1. Introduction

Entrance to university is a matter of great concern to every student and parent. Many schools have different application dates and deadlines, causing students to spend a long time waiting and increasing psychological pressure. When students can estimate their chances of being accepted by the university, they could do more preparation to get higher grades or admission. The hypothesis is to get the percentage of admission through the data set and calculation.

Gale and Shapley (1962) had come up with a new research topic related to college admission, thus, researchers like Kolos Csaba Ágoston, Péter Biró, and Lain McBride (2016) followed their steps to continue with the study [2-4]. For this research, we are going to address the recent issue of applying for college. The problem that we notice is that during each year, colleges receive different information that from a student which contains their contact information, addresses, previous high school GPA, the degrees that they are applying for, and their previous degrees, etc. Unfortunately, colleges take a lot of resources to filter if such students will be getting into the college even though they are enrolled by the college admission office. While not a lot of studies online have addressed the issue, because there are so many distracting factors that could mislead the result. Therefore, our purpose of having this experiment is to provide a prediction model that can help colleges to determine the likelihood of the student getting into the college.

The first thing that needs to make sure is the method of analyzing the research. For the method, the research has been divided into three parts. Through collecting and understanding the data could help filter data. The second part is analyzing the data by constructing the equation and Logistic Regression Method. The third part will be proving the hypothesis and getting results. However, the question is that unable to analyze known results through various interference data. The expected goal is to get the initial model by comparing data and calculations and demonstration. Since our model is analyzing the possibility of a student going to college, therefore, we have three questions: Which factor is highly correlated to if the students are going to college? Which model are we going to use in terms of determining such prediction? What does our result interpret?

The dataset that we are going to use for our research is random college admission data that is provided online that have all the students' information during the few years. From there, the first thing that we need to do is creating the models and figuring out which method are we going to use in this research. The second step is logistic regression and decision tree modeling. Obtaining research data through calculation and comparison and determine the result by comparing the data.

2. Models and methodologies

Previous studies mainly address the calculation of college when taking the candidate's information and calculated such candidates will be selected as their students. Our study is a development of the previous studies. Instead of looking at which factors are important for students to get admitted, we are looking at the possibility of the student going to college after they got admitted.

2.1 Research Logic

The Method that we use is like what Kolos Csaba Ágoston, Péter Biró and Lain McBride (2016) did [2].

Let $A = \{a1, \dots an\}$ be the different variables that are inside of one student's profile, and let C be the college that they had submitted their applications to. We also created binary variables $X = \{0,1\}$ be if such student eventually decides to go to the college. When every variable fits into the model that matches the requirement when x=1, then the student will guarantee to go to this college, otherwise, they are going to the college. In this situation, we named these two extreme conditions as candidates Y and Z, if the student W has less average GPA than $Y\,(a2w < a2y)$, then the possibility of such a student is lower than our perfect condition. In another word, they are not granted that they will be a student for the college.

We also think of a total score calculation, if everything that contributes to candidate Y has a perfect score of 100, and let the set A be the calculation field, and convert every value into digital numbers, if $a1 + a2 + \cdots + an = 100,$ then the possibility will be 100%, as they have a positive correlation, if a1 decrease while other variables remain constant, then the score will be less than 100, so the student will have less possibility of selecting the college.

2.2 Data Mining and Cleaning

As mentioned previously, we have the dataset that provide by a random college that contains all the information that they gained from the year, and we use the data to predict if the student is going to select the college. The dataset that we have contains full-time students, or part-time students (if the candidates are students but pursuing a higher degree or they are new to being as a student).

Other variables that we have are Contact ID (Candidate number), Gender, Lifecycle Role(They application stats), student(if such candidate is enrolled into the college as a student), In-State(if they are state residents or not), Campus Visit(How frequently do they visit the campus), Source (Where do they know about the college), Ethnicity(Racial Groups), Age, Average GPA, Highest degree(the current degree that they have), Education (if they went to any academy before), Email Count (How many emails that they send to the college), and Phone Count.

The challenge that we have from the initial dataset is that all the students are recorded in the dataset but not all the students have values in every variable. For example, a student might have every value but he or she doesn't contact the college via emails, then the email Count is empty, or if the student refuses to provide their ethnicity, then the field for that is "null". Fortunately, Tamraparni Dasu and Theodore Johoson (2004) provide a solution for exploratory data mining and cleaning, we can have null values ignored and combine the duplication values (In case of the student has submitted multiple applications) [3]. In Table 1, we have combined every variable and removed duplication, so that one single contact id can have one set of data and contain the sum of all the information.

We began by pulling data directly from the SQL workbench after joining multiple tables by executing various queries; the group thought that gender, academic program, transfer credits, GPA, state, academic degree, class level, college attended, ethnicity, primary role, and life cycle stage would play a significant factor in the model. In addition to gathering existing data, a calculation was included to extract the average GPA from the student's latest education program as opposed to the one contained within the original data - weighted GPA. This resulted in further narrowing the variables to be included in the final design. Later in the initial stages of constructing a usable CSV file, our group assembled a sizable query to grab additional information. The final variables used in the analysis were: ContactID, Student, InState, Gender, Ethnicity, AvgGPA, Highest Degree, EducationInState, SeeksSecondDegree, EmailCount, PhoneCount, DMCount, and AgeWhenApply. Next, we explored both linear and logistic regression given we could assess numeric and categorical target variables. Numerous CSV files were tested on both R studio, Python, Orange, and SPSS to conduct the analyses until we decided on a model that was the best fit for the data produced.

2.3 Hypothesis and Distribution

Once we have the complete dataset, we are going to look at the distributions like age, GPA, etc. to form our hypothesis. The main platform that we run is through Orange, is a great program that can do data analysis be given data set. In Fig.1, we have the age distribution across different.

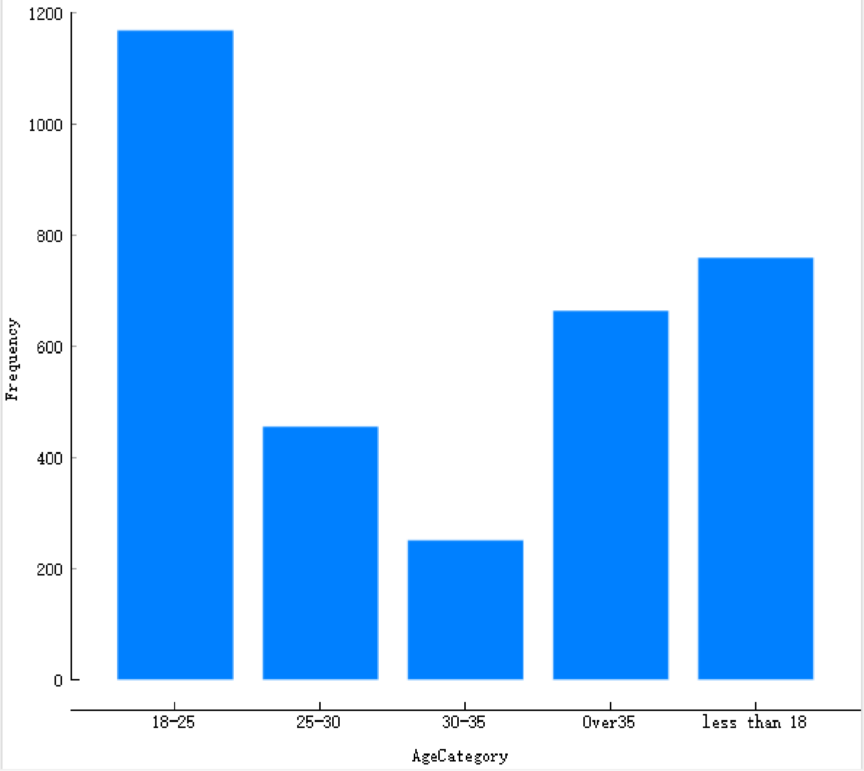

Figure 1. The Age Distribution By given Contact ID

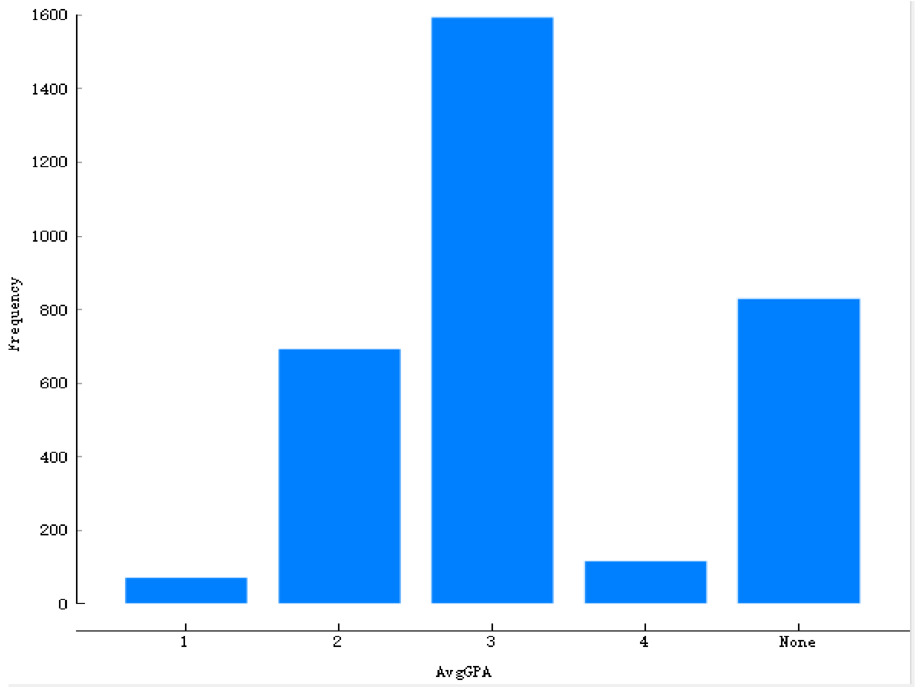

Figure 2. The GPA Distribution By given Contact ID

Model	AUC	CA	F1	Precision	Recall
Tree	0.646	0.741	0.732	0.728	0.741
Logistic Regression	0.821	0.775	0.756	0.760	0.775

Figure 3. Logistic Regression and Decision Trees

		Predicted		
Actual		0	1	Σ
	0	2177	199	2376
	1	546	383	929
	Σ	2732	582	3305

Figure 4. Confusion Matrix

Contact IDs, as shown in Fig.2, we have the GPA distribution. Therefore, we can have formed our hypothesis. In Fig.1, the most frequent age distribution lays between the ages of 18-25, we assume that the students who are 18-25 years old have a higher possibility of going to college, so it can be a positive variable considering our score calculation. While in Fig.2, most students who have a 3.0 out of 4.0 scale have a better chance of getting into college. In general, we interpret that, the target that we are looking for is a student W who is 18-25 years old and has a 3.0 GPA.

2.4 Correlation Calculation

After we interpret our hypothesis, the next step is testing our hypothesis and finding out the highly correlated variables which contribute to our research. In Table 2, we find out the correlation as selecting the student as the dependent variable, and other variables as our independent variables. From the table, we interpret that, variables like Table 1.

Table 1. Variable Type

Gender	In State	Campus Visit	In State	Age	Highest Degree	Phone Count	Email Count
Female. Male	0(No),1(Yes)	0(No),1(Yes)	0(No),1(Yes)	18 or above	1,2,3,4	0…n	0…0

Table 1.2 Continues

Ethnicity	Source	Seeks second Degree
String	Sting	String

In-State, Campus Visit, In State, Age, GPA, Highest Degree, Phone Count, Email Count have a higher significate level below 0.05, while other variables like Gender, Ethnicity, Sources have greater significate level above 0.05. Thus, we can remove these variables and only looking at the variables that have a higher significance level. While looking at Table 2. We also notice our hypothesis is relevant, even though the majority doesn't mean all the cases, but it can represent the most.

Table 2. Pearson Correlation

		Student	InState	CampusVisit	EducationInState
Student	Pearson Correlation	1	.048**	.151**	.159**
	Sig. (2-tailed)		0.006	0.000	0.000
	N	3305	3305	3305	3305
InState Table 2(Continued)	Pearson Correlation	.048**	1	.066**	.360**
	Sig. (2-tailed)	0.006		0.000	0.000
	N	3305	3305	3305	3305
CampusVisit	Pearson Correlation	.151**	.066**	1	.043*
	Sig. (2-tailed)	0.000	0.000		0.013
	N	3305	3305	3305	3305
EducationInState	Pearson Correlation	.159**	.360**	.043*	1
	Sig. (2-tailed)	0.000	0.000	0.013	
	N	3305	3305	3305	3305
EmailCount	Pearson Correlation	.306**	0.024	.121**	.081**
	Sig. (2-tailed)	0.000	0.164	0.000	0.000
	N	3305	3305	3305	3305
PhoneCount	Pearson Correlation	.243**	-0.019	.198**	0.034
	Sig. (2-tailed)	0.000	0.263	0.000	0.051
	N	3305	3305	3305	3305
DMCount	Pearson Correlation	.071**	-0.014	.098**	-0.002
	Sig. (2-tailed)	0.000	0.410	0.000	0.925
	N	3305	3305	3305	3305

EmailCount	PhoneCount	DMCount
.306**	.243**	.071**
0.000	0.000	0.000
3305	3305	3305
0.024	-0.019	-0.014
0.164	0.263	0.410
3305	3305	3305
.121**	.198**	.098**
0.000	0.000	0.000
3305	3305	3305

.081**	0.034	-0.002
0.000	0.051	0.925
3305	3305	3305
1	.413**	.058**
	0.000	0.001

Note: * and ** represent digits after the decimal point is omitted.

3. Logistic Regression and Decision Tree

As the method explained previously, the next stage is to analyze the predictive data by using different models to testify and answer our questions.

Logistic Regression is a predictive model that is used to predict a set of outcomes. While it contains factors like AUC (Area Under the curve), F1 score, and AC score which simulate the outcomes by valuing different weights of the variable. David G. Kleinbaum and Mitchel Klein (2010) discuss the importance of the Logistic Regression and how it can benefit us from the result [5]. The logic that we used in the model is that since we have a set of A and given a college C, thus, we consider each correlated variable, and the weight of each variable to interpret the equation:

$$Y^\wedge = a1 * w1 + a2 * w2 + \cdots . + an * wn$$

While Y^ stands for the scores of the prediction and w stands for the weight of each variable. The same logic will also apply to the decision modeling.

Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, Steven and D. Brown (2004) talk about another model: Decision Tree, which is also a predictive model that interprets the possible outcome, but less great than the logistic regression model while analyzing the outcomes because compared to logistic regression, it has fewer variables to consider [6]. For our research questions, we expect that by comparing both models, we can have the result as accurately as possible. Thus, during the modeling process, we are going to use two models, and by looking at the result, we can decide if we want to have to use the existing data for our prediction rather than re-editing variables.

4. Models and Results

In Fig.3, we construe our logistic regression. The AUC is 0.821 on a 1 scale which means that about 82% of our data can be explained by the logistic regression. While the decision tree model tells us that 0.646 of 1 scale which interprets that only 64.6% of data can be explained by the model. The measure that measures both models is that scores lay down between 0.5 and 1, the closer to 1, the better the model is, thus, we interpret that the logistic regression is accurate enough for us to do the prediction. After we finish the Logistic regression and decision tree modeling, by considering the score of each model, we have our result both from the confusion matrix and the predictive model.

4.1 Confusion Matrix and Predictive Model

We construe a confusion matrix that predicts the cases of students who possibly belong to a category. Fig. 4 represents an overall view of the interpreted result, where it represents the true positive, true negative, false, and true negative cases. As we go deep into the result, we construe a predictive model of students who are likely to go into college and export it as a CSV file. Table 3 shows a view of the possibility in percentages of a student. For example, student 1 has a logistic score of 0.82 and 0.17 which indicates that this student has a 79% of going to college while there is still a 17% chance that he or she will choose a different c

Table 3. Interpret Result

gender	In State	Campus Visit	Age Category	Avg GPA	Highest Degree	Education In State	Seeks Second Degree	Email Count
F Female Male Other Perfer\ Not\ To\ Identify U	0 1	0 1	18-25 25-30 30-35 Over35 less\ than\ 18	1 2 3 4 None	1 3 4 5 NULLS	0 1	Higher Lower NULL Same	continuous
Male	0	0	Over35	None	NULL	0	NULL	3
Female	1	0	Over35	None	NULL	1	NULL	3
Male	0	0	30-35	2	NULL	0	NULL	6
Female	0	0	Over35	None	NULL	0	NULL	1
Male	0	0	30-35	2	3	1	Same	2
Male	0	0	18-25	3	NULL	0	NULL	1
Female	0	0	Over35	3	3	1	Lower	10
Male	1	0	25-30	2	3	0	Same	1
Female	0	0	25-30	3	NULL	0	NULL	3
Female	1	0	18-25	3	3	1	Lower	6
Female	1	0	25-30	3	3	0	Lower	12
Female	1	0	Over35	3	NULL	0	NULL	2
Male	0	0	Over35	2	NULL	0	NULL	7
Female	1	0	25-30	2	NULL	0	NULL	5
Female	1	0	Over35	None	NULL	0	NULL	14
Female	0	0	18-25	3	NULL	0	NULL	1
Female	1	0	Over35	3	NULL	1	NULL	7
Female	0	0	Over35	3	NULL	1	NULL	3
Male	1	1	25-30	2	3	1	Same	3
Female	1	0	30-35	3	NULL	0	NULL	2
Female	0	0	25-30	2	NULL	0	NULL	1
Female	0	0	30-35	3	NULL	0	NULL	1
Female	0	0	25-30	2	3	1	Same	17
Male	0	0	Over35	None	NULL	0	NULL	1
Female	0	0	30-35	2	NULL	0	NULL	0
Female	0	0	Over35	3	NULL	0	NULL	2
Female	0	0	25-30	2	3	0	Lower	22
Female	0	1	18-25	4	NULL	0	NULL	0
Female	0	0	18-25	None	NULL	0	NULL	15
Female	0	0	18-25	3	NULL	0	NULL	0
Female	1	0	30-35	3	NULL	1	NULL	1
Male	0	0	18-25	3	NULL	0	NULL	2
Female	0	0	25-30	4	NULL	0	NULL	5

Table 3.1 Variable Value

Phone Count	DM Count	Ethnicity	source	Student	Logistic Regression	Logistic Regression (0)	Logistic Regression (1)
continuous	continuous	African\ American, \ non-Hispanic Asian Black\ Or\ African\ American Black\ or\ African\ American Hispanic Hispanic\ or\ Latino Hispanic/Latino Hispanics\ of\ any\ race Mexican, \ Mexican\ American, \ Or\ Chicano Native\ American\ Gila\ River\ Pima Native\ A	AASHE\ Conference ACT\ Scores ACT\ Scores\ Received ACT/SAT\ College\ Report AMERICORPS APPLYWEB Adventure\ Magazine American\ Art\ Therapy\ Association Application\ Common\ App Application\ On\ Line Application\ Paper\ Application Arizona\ Sneak\ Peek\ E	0 1	0 1	continuous	continuous
				class	meta	meta	meta
0	0	White	Walk In	1	0	0.823316	0.176684
0	0	Hispanics of any race	Local Person	0	0	0.697374	0.302626
0	0	White	Internet	0	0	0.640819	0.359181
0	0	Black or African American	Application Paper Application	0	0	0.807142	0.192858
0	0	White	Word of Mouth	0	0	0.589959	0.410041
0	0	Two or More Ethnicities	Attend RDP	0	0	0.770568	0.229432
0	0	Two or more races	Other	1	0	0.870496	0.129504
0	0	White	Word of Mouth	1	0	0.627325	0.372675
2	0	White	NOLS	0	0	0.548088	0.451912
1	0	White, non-Hispanic	Word of Mouth	1	0	0.704793	0.295207
1	0	Hispanics of any race	Application Common App	1	0	0.631668	0.368332
1	0		Application On Line	0	0	0.635929	0.364071
0	0		PC Web Request	1	0	0.712666	0.287334
1	0	White	Attend RDP	0	0	0.615012	0.384988
0	0	White	Application Paper Application	1	0	0.520715	0.479285
0	0		ACT/SAT College Report	0	0	0.858724	0.141276
3	0	Native American White Mtn Apache	Friend	1	1	0.499544	0.500456
4	0	White	Read It Here	1	0	0.52836	0.47164
3	1	White	Application Paper Application	0	1	0.282072	0.717928

0	0		Application On Line	0	0	0.630392	0.369608
1	0	White		0	0	0.689239	0.310761
0	0		Phi Theta Kappa	0	0	0.677828	0.322172
2	0		Previously Attended Prescott College	1	0	0.546736	0.453264
0	0	White	Current Prescott College Student	0	0	0.663028	0.336972
2	0	Hispanics of any race	Application On Line	1	0	0.762104	0.237896
0	0	White	Word of Mouth	1	0	0.585347	0.414653
3	0	White, non-Hispanic	Internet	0	0	0.521974	0.478026
1	0	White	Friend	0	0	0.611478	0.388522
5	0	White	College Fair	0	0	0.869862	0.130138
0	0	Black or African American	Internet	0	0	0.869112	0.130888
0	0	Hispanics of any race	Application Paper Application	0	0	0.648181	0.351819
3	0		Application Common App	0	0	0.784773	0.215227
7	0	White	Internet	1	1	0.473696	0.526304

5. Conclusion

In conclusion, the results which were calculated through the necessary variables of the data show there is 82.1% accuracy of the model created in this research. To make the experimental results more rigorous. Then the necessary variable is education, email, phone, and DM (direct message) count, and age. This can help improve the accuracy of our experimental results and successfully prove the hypothesis through this experiment. Data analysis allows us to process and integrate data faster. By changing the variables of the data and analyzing the data, the probability of a student's application success can be obtained. Using these data can help students clarify their goals. Our paper creates a possible solution for colleges to consider after selecting different students and waiting for students' feedback, we want to ensure their colleges don't need to waste resources and consume time while waiting for the students and eventually lost the students. In the future, we would like to design more studies that help colleges not only be limited by string variables, but also the machine can learn themselves while adding more non-relational variables into the dataset, and the machine can automatically convert the variables to strings and make the prediction.

We are also interested in building a predictive model for students after they get into the universities. For example, the finical aid system will ideally input the students' stats and information to predict their future GPA and if they are on the edge of applying for the finical aid, including a cheating system than can find out the similarities of student's paper to prevent the academic dishonesty of the students. Even though we think our model is relevant to answer the questions that we come up with and a good way of predicting a student's stats, but we also acknowledge that a prediction model can be a consideration to an issue, but not the true result in real life. Thus, the result doesn't mean the college should give up on students if they have a low score in the model. Also, the data that we find is mainly from a simulated college for experimental purposes. In real life, the issues can be much more complicated with unexpected changes. Thus, our result is trying to give a solution to the issue but not giving a guideline to the issue.

References

[1] Atkinson, R. C., &amp; Geiser, S. (2009). Reflections on a century of college admissions tests. Educational Researcher, 38 (9), 665 – 676. https://doi.org/10.3102/0013189x09351981.

[2] Biró, P., &amp; McBride, I. (2014). Integer programming methods for special college admissions problems. Combinatorial Optimization and Applications, 429–443. https://doi.org/10.1007/978-3-319-12691-3_32.

[3] Dasu, T., &amp; Johnson, T. (2003). Exploratory data mining and data cleaning. Wiley-Interscience.

[4] Gale, D., &amp; Shapley, L. S. (1961). College admissions and the stability of Marriage. Rand Corporation.

[5] Kleinbaum, D. G., &amp; Klein, M. (2011). Logistic regression: A self-learning text. Springer.

[6] Ragsdale, C. T. (2004). Spreadsheet Modeling &amp; Decision Analysis: A practical introduction to management science. Thomson/South-Western.

[7] Sawyer, R. (2013). Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. Applied Measurement in Education, 26 (2), 89 – 112. https://doi.org/10.1080/08957347.2013.765433.

[8] Zhou, J. (2020, February 11). University of California 1994-2016 admission stats - dataset by Jimzhou. data. world. Retrieved November 24, 2021, from https: //data.world/jimzhou/university-of-california-1994-2016-admission-stats.

[9] Sternberg, R. J. (2010). College Admissions for the 21st Century. United States: “Harvard University Press”.

[10] Baker, D. L., Leonard, B. (2016). Neuroethics in Higher Education Policy. United Kingdom: “Palgrave Macmillan US”.