

Student dropout analysis based on previously acquired educational achievements: A case of the University of Portalegre

Business Analysis, Business Informatics Ms, Fall 2023.

Izeldeen Nedal Yunis Al Fraijat Danat Semeneev Ieva Žube
Pankaj Chettri Kristaps Eglītis*

2023-12-13

Abstract

In the world of education, the path to success is often visualized as a linear progression, where students follow a predefined journey from kindergarten to graduation. However, the reality is far more complex. There could be various reasons that come along the study program that led students to deviate from this path. These students might encounter different challenges, circumstances, or a lack of proper resources that have led them to drop out of university.

In this dataset provided to us, we will delve deeper into understanding the reasons why students have dropped out of the university, based on the data at our disposal. We will leverage our social knowledge to comprehend the factors that influenced their decision to drop out and work to prevent such occurrences if the issues are within the university's purview. Our goal is to offer solutions, support, and the necessary resources to facilitate students' educational journeys. We will also use the analysis we've conducted on the dropout students to learn from their experiences and chart a unique educational pathway with fewer dropouts.

*. Rīga Technical University

Introduction

Starting from preliminary school we are told that having an education is very important for your future or that without higher education your job possibilities are going to be very limited. While primary education is mandatory, having higher education is not. There are, however, many reasons for which people may want to pursue higher education. According to studies, many factors are materialistic, the most important factor for pursuing higher education is job acquisition (Knutsen 2011). Some other factors may include increased income in the existing job, improved work conditions or increased ability for retirement. Of course, other, more intrinsic factors include seeking for additional knowledge or self-fulfillment (Cortes et al. 2023). There are also factors like meeting new friends, improving social interaction skills or just wanting to make a difference in the world. Of course factors that cannot be ignored are social pressure (Temple 2009), meaning that having friends that want to pursue higher education can influence ones own decision or influence of family members. However, there are people that discontinue their studies prematurely and we are interested to learn what the reasons for such a decision could be. Based on the study and datasets that we used for our research there are multiple factors that influence dropping out.

Nevertheless, pursuing higher education and actually getting the degree has some tangible benefits. According to an OECD – Education at a Glance 2019 research paper (OECD 2019).

“On average across OECD countries, adults with a short-cycle tertiary degree earn 20% more than adults with upper secondary education. The earnings advantage increases to 44% for those with a bachelor’s degree and to 91% for those with a master’s or doctoral degree.”

With this in mind, it is important for government and educational institutions to ensure high level of graduates in society to ensure economic growth and overall increase in well-being. To measure the success of this goal, it is important to set KPI’s, track them and make educated conclusions on what needs to be done or is being done right to reach the goal of higher educated society.

Target Metrics and KPI

In this particular case, KPI’s will be chosen based on datasets of Portugese High Schools but most likely data can be generalised, atleast for Europe, as the region and sociodemographics are not so different. Even though there are many factors that influence the success of graduation, only factors that can be proven by government and educational institutions will be chosen. In order to thwart embezzlement, indicators should be restricted in magnitude and difficult to falsify or manipulate. After rigorous analysis, we propose the following KPIs.

- a. **Student grade improvement compared to support.** Based on the dataset, students who had support had 3x lower dropout rates than students that didn’t have. While it is

not practical to allocate higher amount of money for studying that itself does not generate value, it scoops that it at least a sizeable parts of the dropout students could be held from leaving with a relatively small aid that would make the benefits of studies outweigh those of working/etc. Leaving is commonly associated with very poor grades (otherwise, even a morally disinterested student would opt to formally remain in the university until they are asked to leave due to poor performance). Since a person with infinitesimal grades is a clear candidate for dropping out, one should identify those students with abrupt downward grade dynamics and quench this. In the proposed KPI, the $(grade)_i$ is the mean relative grade change for student j over all their courses at university i at moment t , and the assistance is the mean aid per student (can be 0). If there are no students on their way down, the KPI is guaranteed to be positive.

$$KPI_{1,i,t} = \frac{|\Delta(\overline{grade})_i|}{(\overline{assistance}_i)}$$

This does not depend on the number of courses, because the courses are themselves different difficulties, the important thing that the university (the students too) should look after in this regard, that the situation with grades does drastically deteriorate over time.

- a. **Institutional Improvements.** Although volatile and subjective, as one of the metrics (not KPIs, since it is more difficult to tie this to specific redresses) there could be a longitudinal survey about one's satisfaction with the studies and programme in general in the fashion of a job an exit or quasi-exit interview (when a person does not leave actually, but they are still invited to answer the questions as if they would be leaving). This would allow to track the scale of dropouts due to frustration with the programme (not engaging enough).
- b. **Relative changes in student's grades.** Datasets tell us that the higher the average grade, the lower the dropout rate. Usually students that have low grades are uninterested in the subjects which could be due to having chosen not the right program for them or that the way lectures and information is presented is uninteresting or outdated. Either way this can be improved. Increasing the possibility that the student has chosen the right program for him can be done by introducing more "open days" in higher education institutions and having more upfront information about what can be expected from programs. The overall lecture performance can be improved by taking more time to have up-to-date information presented and teachers having decent motivation of teaching students. This can be achieved by increasing teacher salaries and institutions having more control over teachers and information they present to students.

All these metrics are still vulnerable to misrepresentation, but it is inevitable given the freedom the universities enjoy in managing their study programmes. Still, any manipulation of this metrics can only be temporary and thus is also not in the best interest of the university.

Exploratory Data Analysis

Descriptive Statistics

As we have checked, the dataset does not have zero values, so there is nothing to purge inside it. Later on, we get the basic descriptive statistics, shown below in Tables 3, 1, 2, 4, 5, 6, 7

	Mari. stat.	Appl. mode.	Appl. orde.	Cour.	Dayt. atte.
count	4424	4424	4424	4424	4424
mean	1.18	18.67	1.73	8856.64	0.89
std	0.61	17.48	1.31	2063.57	0.31
min	1	1	0	33	0
25%	1	1	1	9085	1
50%	1	17	1	9238	1
75%	1	39	2	9556	1
max	6	57	9	9991	1

Table 1: Descriptive statistics

	Prev. qual.	Prev. qual. (gra.	Naci.	Moth. qual.	Fath. qual.
count	4424	4424	4424	4424	4424
mean	4.58	132.61	1.87	19.56	22.28
std	10.22	13.19	6.91	15.6	15.34
min	1	95	1	1	1
25%	1	125	1	2	3
50%	1	133.1	1	19	19
75%	1	140	1	37	37
max	43	190	109	44	44

Table 2: Descriptive statistics (cont'd)

	Mother's occupation	Father's occupation	Admission grade	Displaced	Educational special needs
count	4424.00	4424.00	4424.00	4424.00	4424.00
mean	10.96	11.03	126.98	0.55	0.01
std	26.42	25.26	14.48	0.50	0.11
min	0.00	0.00	95.00	0.00	0.00
25%	4.00	4.00	117.90	0.00	0.00
50%	5.00	7.00	126.10	1.00	0.00

	Mother's occupation	Father's occupation	Admission grade	Displaced	Educational special needs
75%	9.00	9.00	134.80	1.00	0.00
max	194.00	195.00	190.00	1.00	1.00

Table 3: Descriptive statistics (cont'd)

'Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approv

count	4424	4424	4424	4424	4424
mean	0.11	0.88	0.35	0.25	23.27
std	0.32	0.32	0.48	0.43	7.59
min	0	0	0	0	17
25%	0	1	0	0	19
50%	0	1	0	0	20
75%	0	1	1	0	25
max	1	1	1	1	70

Table 4: Descriptive statistics (cont'd). Columns, left-to-right: Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrollment

count	4424	4424	4424	4424
mean	0.02	0.71	6.27	8.3
std	0.16	2.36	2.48	4.18
min	0	0	0	0
25%	0	0	5	6
50%	0	0	6	8
75%	0	0	7	10
max	1	20	26	45

Table 5: Descriptive statistics (cont'd). Columns International, Curricular units 1st sem (credited), Curricular units 1st sem (enrolled), Curricular units 1st sem (evaluations).

count	4424	4424	4424	4424
mean	4.71	10.64	0.14	0.54
std	3.09	4.84	0.69	1.92
min	0	0	0	0
25%	3	11	0	0
50%	5	12.29	0	0

75%	6	13.4	0	0
max	26	18.88	12	19

Table 6: Descriptive statistics (cont’d). Columns Curricular units 1st sem (approved), Curricular units 1st sem (grade), Curricular units 1st sem (without evaluations), Curricular units 2nd sem (credited)’

count	4424	4424	4424	4424
mean	6.23	8.06	4.44	10.23
std	2.2	3.95	3.01	5.21
min	0	0	0	0
25%	5	6	2	10.75
50%	6	8	5	12.2
75%	7	10	6	13.33
max	23	33	20	18.57

Table 7: Descriptive statistics (cont’d). Curricular units 2nd sem (enrolled), Curricular units 2nd sem (evaluations), Curricular units 2nd sem (approved), Curricular units 2nd sem (grade)’

The students are from multiple countries, but the overwhelming majority of the students are from Portugal. It would be interesting to see how the students’ admission grade depends on their previous qualification in their home countries, but the samples are scarce. Many students from abroad are from the Overseas Territories where it’s more challenging to get comparable education. However, they and inland Portugal students were naturally given some exemptions, as the dataset states ¹

. For example, the students admitted per Ordinance no. 854 ² were not required to demonstrate the proof of their validity since they received a diploma in secondary education administered in Portuguese (Angola, East Timor, Mozambique, Guinea Equatorial). Students admitted per Ordinance no. 533 ³ were from another university in Portugal with overlapping courses covered recently enough so they were not required to repeat them. Finally, those admitted per Ordinance no. 612 ⁴ came from other countries but had comparable material in their studies and so their points were recalculated with some amortization.

Due to class imbalance , the variability for the Portuguese students is much higher, and while the 3 categories (see Figure 1) with highest grades are natural, i. e. doctors, masters as higher

1. Link to the dataset description: <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>

2. Link to the source document: <https://dre.tretas.org/dre/106607/portaria-854-B-99-de-4-de-outubro>

3. Link to the source document: <https://dre.tretas.org/dre/104726/portaria-533-A-99-de-22-de-julho>

4. Link to the source document: <https://dre.tretas.org/dre/51542/portaria-612-93-de-29-de-junho>

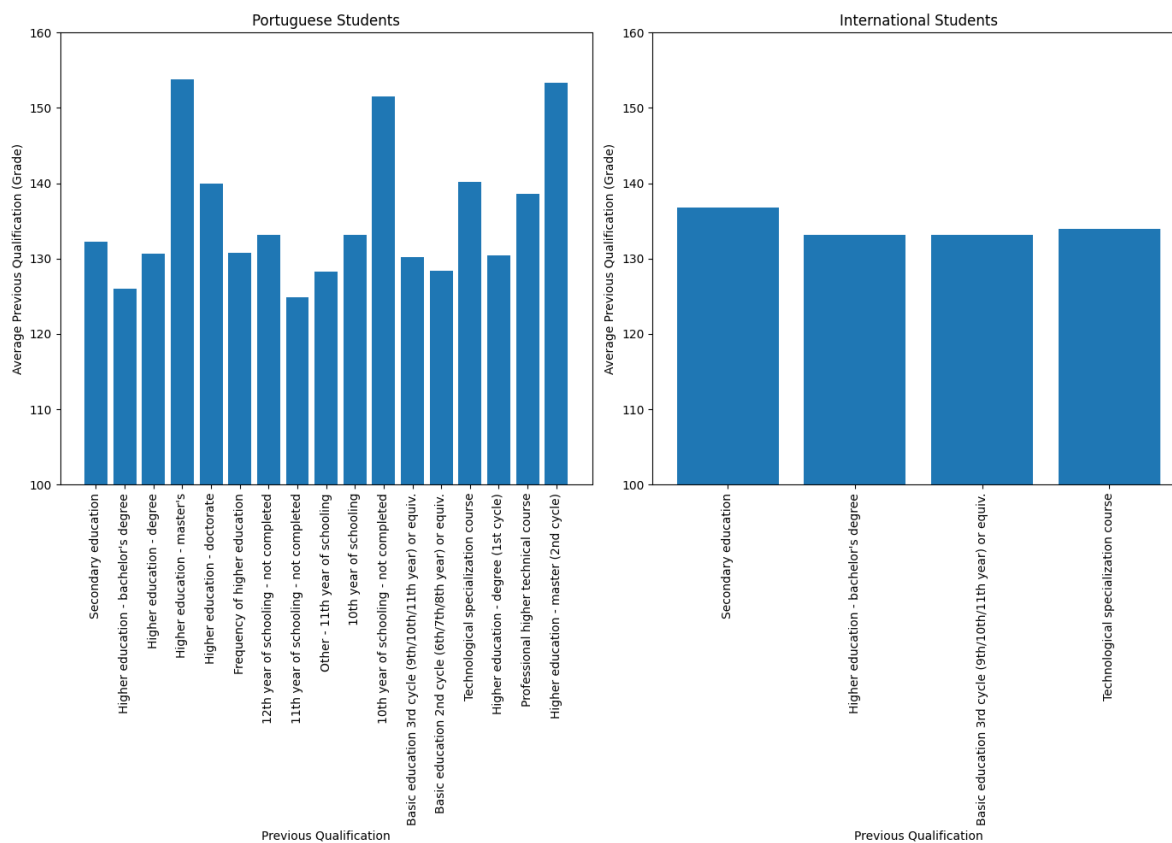


Figure 1: Relative graduation points for students with different education backgrounds

education, the 3rd is unintuitive (the 10 classes) and we tend to explain it as self-selection and high correlation with other indicators (those entering the university in the 10th grade are more motivated then dwelling in schools in 11th and 12th grades).

Also, there is a drastic imbalance over yet another crucial factor: age. Students of age are far less ubiquitous, can have far more incentives to abandon studies and smaller potential for apprehension of material. Indeed, this is clearly shown on the next graph [2](#)

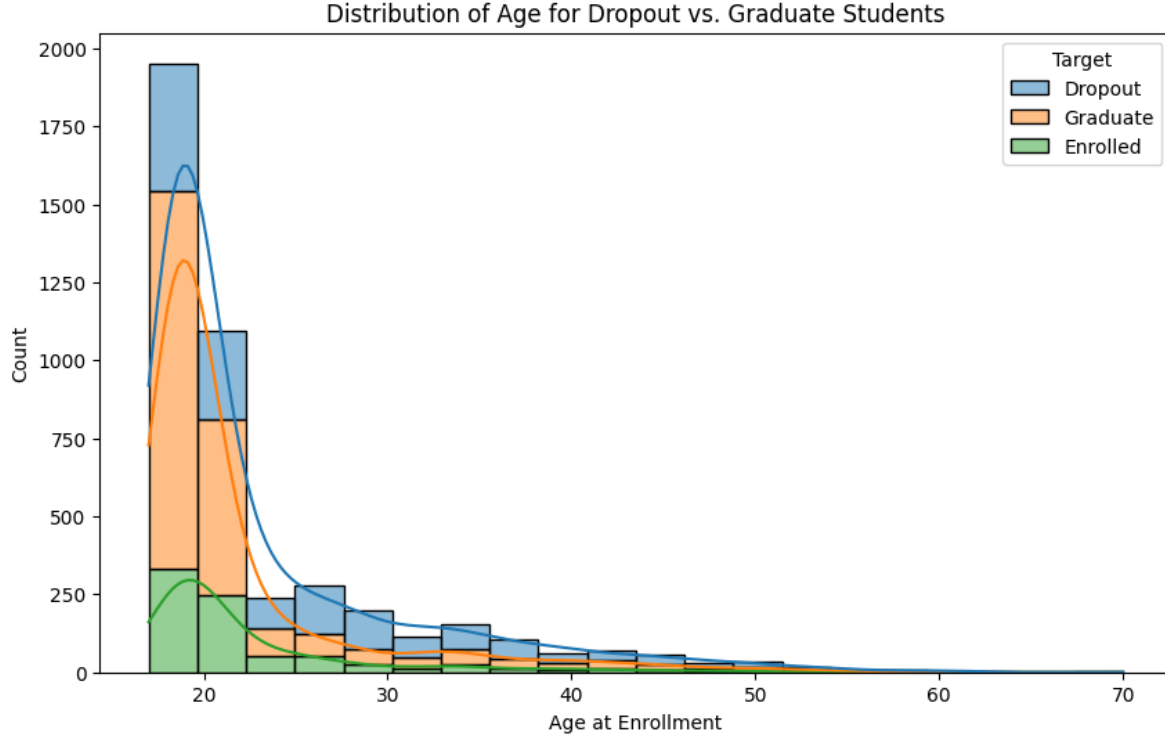


Figure 2: Distribution of age for dropout and graduate student

Q. v. the sizes of the bins for dropout students differ far less than the total size for the name of the student.

If the hypothesis about some external factors is correct , the target variable should be much dependent on previous grades,

The datapoint cloud on Table [3](#), however, shows that this rule has a lot of exceptions.

We can draw the following observations:

- The **distribution of admission grades** is roughly normal with most students scoring between *120 and 160 marks*.
- The **distribution of previous qualifications** (grades) is also the same with most of them having grades in between *120 and 160*.

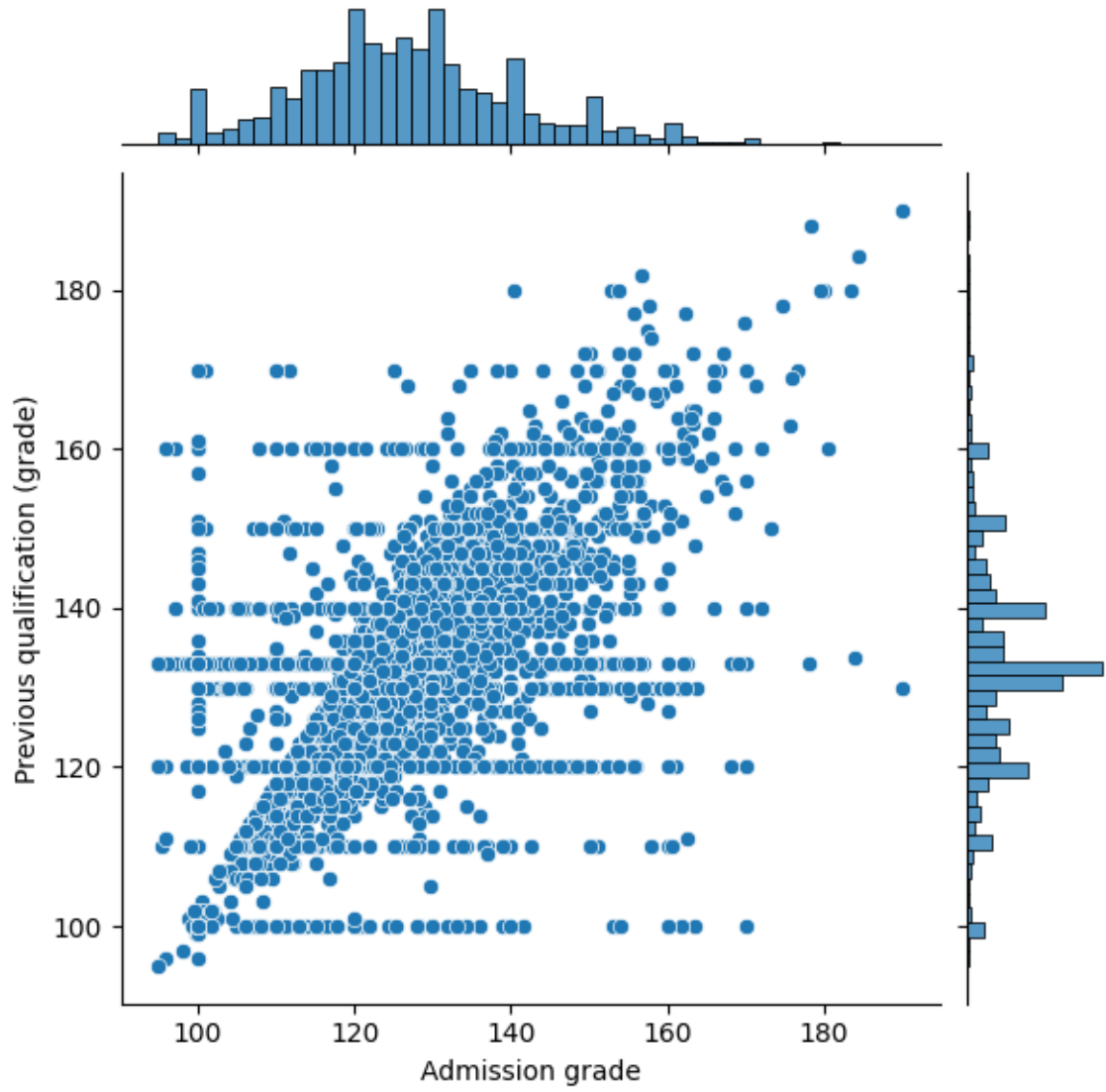


Figure 3: Joint and marginal distributions of current admission and previous qualification grades

- There is seen a **positive correlation** between admission grade and previous qualification grade indicating students with higher previous qualifications tend to have higher admission grades.

This was the visualization for the few quantitative columns, which shows the natural interconnection between the curricularly accrued units in the 1st and the 2nd year, which are in turn mostly unrelated to the admission grade. This is understandable since the grades are commonly based on the successfulness of the local program and student's toil, while the students' backgrounds are commonly different and this puts them into inequitable positions when passing the admission exams.

In these previous graphs, we considered quantitative columns that are more or less exogenous to the dataset (e. g. age and the previous qualification grade are not influenced by the the current grade of the students).

However, the majority of columns of this dataset are qualitative and they are at least partially endogenous as stem from the decisions during the study. For this, we need to propose a mechanism of influence, then formulate and test a hypothesis via an analysis of discriminate groups.

We also consider the impact of scholarships and other compensations in academic support, which should alleviate the complications associated with adaptations in new environment.

We see that having debt is always a serious impediment against studies because it gives wrong incentives towards directly making money in the short run instead of focusing solely on one's studies that could aid to make altogether greater money in the long run.

In different studies, it is quite common to compare the academic success of a student with the academic successes of their parents as this has both direct and indirect effects, s. a. i. e. both are connected to welfare, but also it can be that there is another channel of knowledge transmission to the younger generation.

Observations : * The bar chart shows that mother's occupation is quite influential. This influence is greater the pa's due to traditional effect, and we distinctly see that students whose mothers are 'white collars' dropout significantly more rarely than those whose mothers are more engaged in physical labor.

- This also may suggest the mother's occupation can influence student retention, emphasizing the need for financial support and family engagement.

Data correlation table (quantitative columns only)

In the remaining part, we examine the correlations of purely endogenous temporal variables. This does not give a scoop about the source of causation and is not a good predictor, but exhibits an analysis of autocorrelation inside the quasi-temporal data.

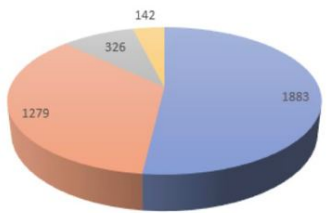
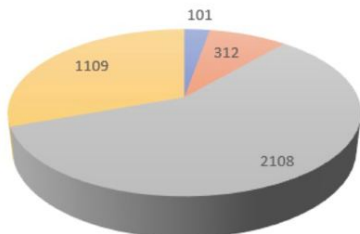
No	Reason	Based on data explanation example	Visual representation										
1	Student doesn't like what they study	If the student is learning their first choice (TOP3), they might have bigger motivation to graduate	<p>Application order - 0 first choice, 9- last</p> <p>Application choice based on order</p>  <table><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>TOP3 graduates</td><td>1883</td></tr><tr><td>TOP3 dropouts</td><td>1279</td></tr><tr><td>4-9 graduates</td><td>326</td></tr><tr><td>4-9 dropouts</td><td>142</td></tr></tbody></table>	Category	Count	TOP3 graduates	1883	TOP3 dropouts	1279	4-9 graduates	326	4-9 dropouts	142
Category	Count												
TOP3 graduates	1883												
TOP3 dropouts	1279												
4-9 graduates	326												
4-9 dropouts	142												
2	Has a debt, that doesn't allow student to focus on studies because of money issues (small income, already big debt, stress about credit)	People with debt have 3 times more dropouts than graduates. People who don't have debt have the biggest number of graduates	<p>Students and their debt</p>  <table><thead><tr><th>Category</th><th>Count</th></tr></thead><tbody><tr><td>Have debt- Graduates</td><td>101</td></tr><tr><td>Have debt- Dropouts</td><td>312</td></tr><tr><td>No debt- Graduates</td><td>2108</td></tr><tr><td>No debt- Dropouts</td><td>1109</td></tr></tbody></table>	Category	Count	Have debt- Graduates	101	Have debt- Dropouts	312	No debt- Graduates	2108	No debt- Dropouts	1109
Category	Count												
Have debt- Graduates	101												
Have debt- Dropouts	312												
No debt- Graduates	2108												
No debt- Dropouts	1109												

Figure 4: Student mobility and financial burden as indicators and drivers of their motivation and impediments

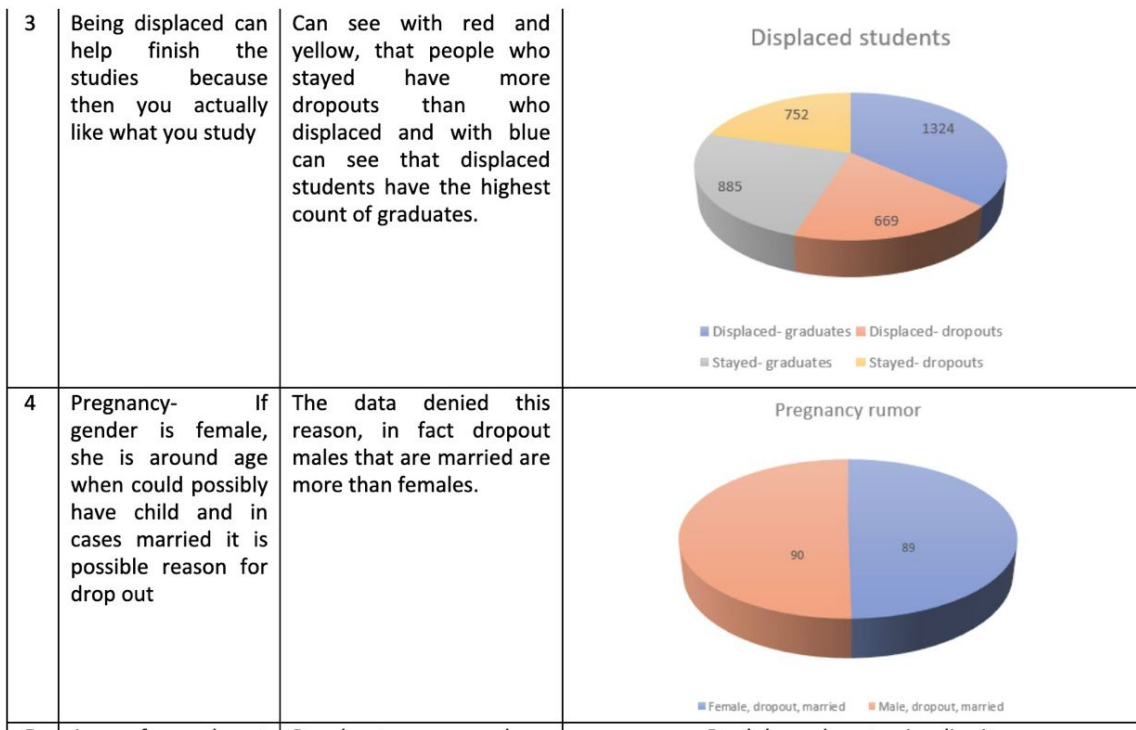


Figure 5: Student inter-university mobility and health conditions proxies as indicators and drivers of their motivation towards learning and impediments (Part 2)

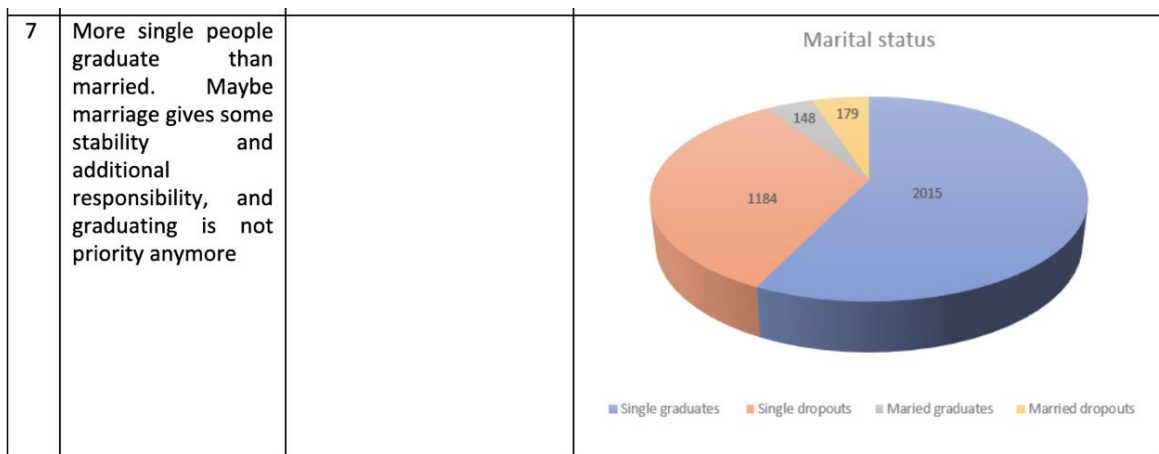


Figure 6: Marital status as distractor from studying

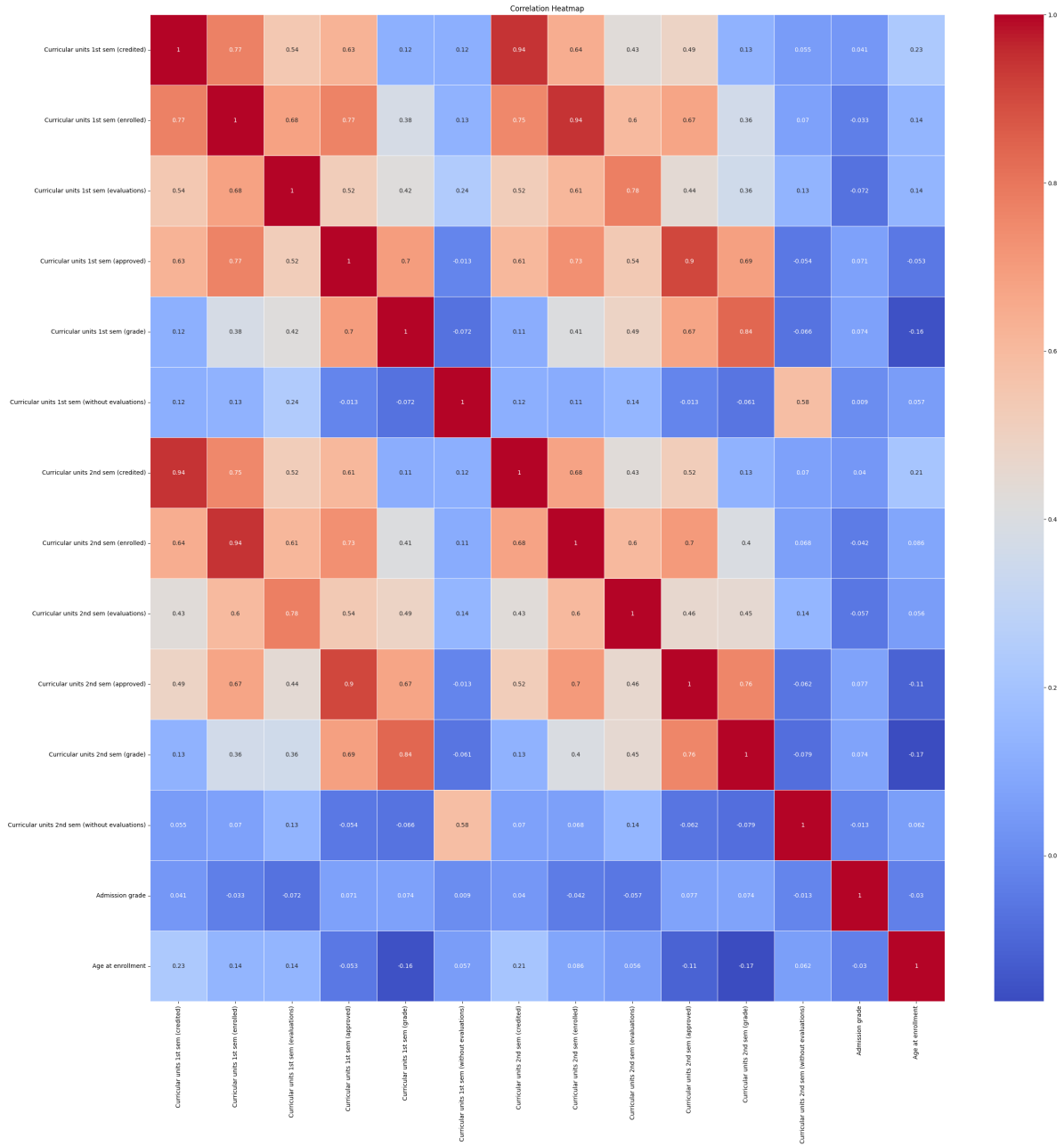
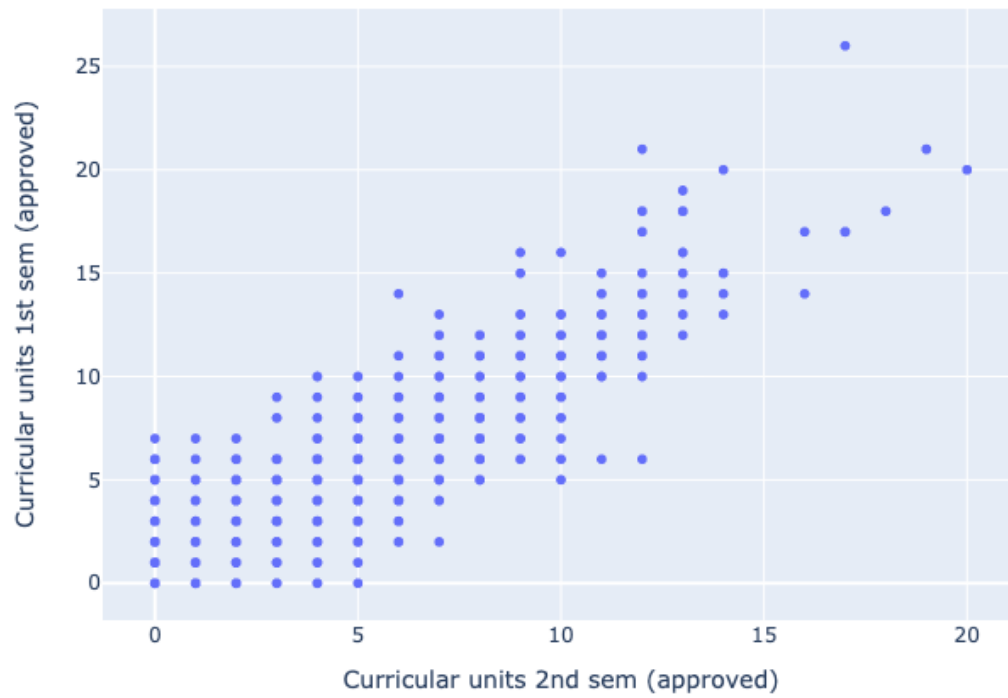
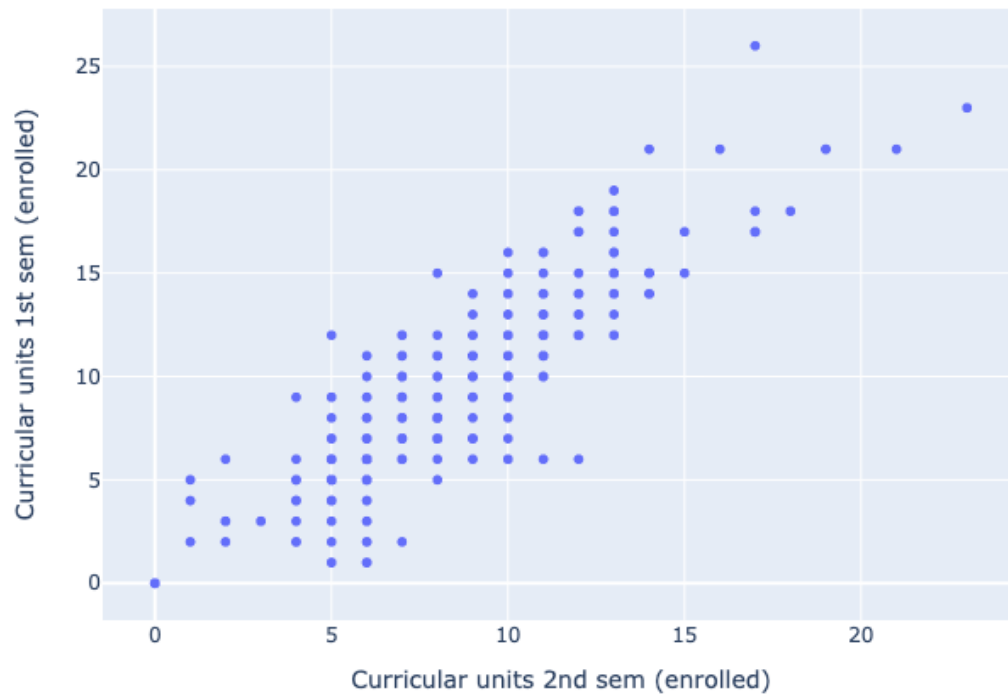
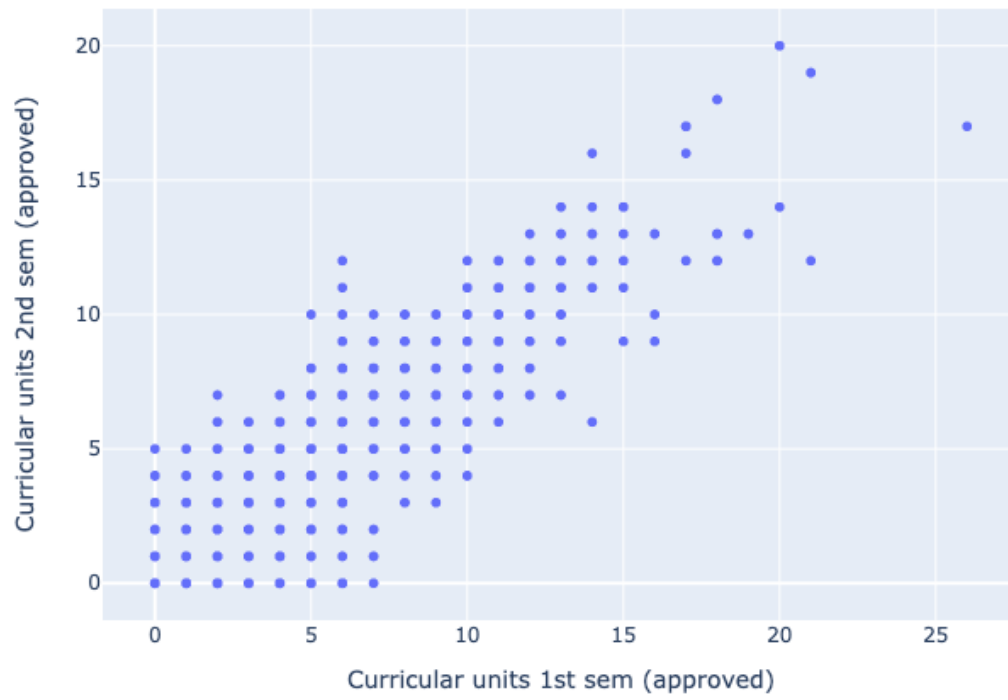


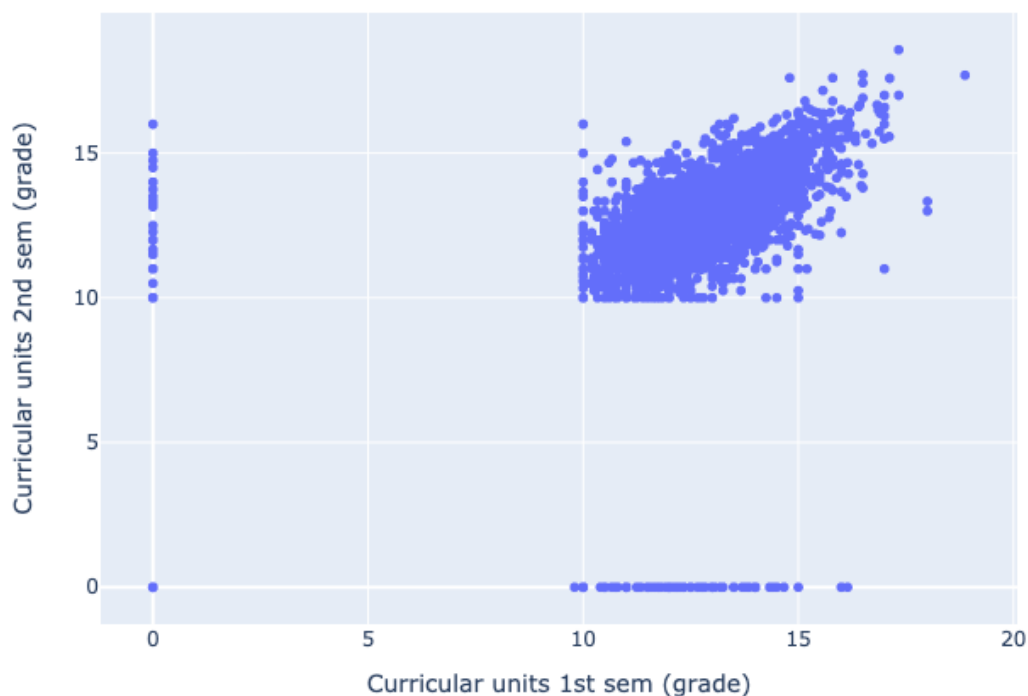
Figure 7: Data correlation table (quantitative columns are represented only since there to compute true correlations between quantitative columns it is necessary to OHE-encode them, which would burst no. of columns to many thousands, but the values of the correlations will be statistically insignificant due to low cardinality of 90% of classes)



We can see that the points for the 1st semester and 2nd semester are correlated which shows that one's marks are primary drivers of success and exhibit sizeable correlations







Data Mining

In this matrix for correlations, we already see high correlations between many values. Hence, if we (certainly) consider qualitative variables in our data mining analysis, we must reduce the number of variables because the true dimensionality of the initial space is too high and virtually all ways of embedding are too costly and prohibitive given a relatively small amount of datapoints in this dataset. First our common step would be to dispose of multicollinear columns.

High dimensionality prevents intuitive DBSCAN threshold setting and some inferior algorithms as TSNE.

After we perform the PCA, we select estimators from various standard families that are independently fine-tuned and then, by F1 measure, the model that is most precise in predicting the outcome is rendered. The results of the best models are given in leaderboard below in Table 8.

<IPython.core.display.HTML object>

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
catboost	CatBoost Classifier	0.7290	0.8626	0.7290	0.7099	0.7150	0.5443	0.5491
gbc	Gradient Boosting Classifier	0.7293	0.8674	0.7293	0.7124	0.7147	0.5426	0.5494
et	Extra Trees Classifier	0.7271	0.8606	0.7271	0.7068	0.7105	0.5377	0.5446
lightgbm	Light Gradient Boosting Machine	0.7248	0.8591	0.7248	0.7044	0.7092	0.5357	0.5413
rf	Random Forest Classifier	0.7255	0.8644	0.7255	0.7041	0.7088	0.5357	0.5420
lda	Linear Discriminant Analysis	0.7190	0.8560	0.7190	0.7047	0.6984	0.5163	0.5310
ada	Ada Boost Classifier	0.7071	0.8237	0.7071	0.6952	0.6956	0.5077	0.5134
qda	Quadratic Discriminant Analysis	0.7113	0.8471	0.7113	0.6939	0.6892	0.5023	0.5166
knn	K Neighbors Classifier	0.6945	0.8128	0.6945	0.6716	0.6792	0.4882	0.4917
ridge	Ridge Classifier	0.7125	0.0000	0.7125	0.6846	0.6676	0.4912	0.5196
dt	Decision Tree Classifier	0.6237	0.7070	0.6237	0.6283	0.6251	0.3907	0.3914
lr	Logistic Regression	0.4754	0.5142	0.4754	0.3960	0.4313	0.0909	0.0956
nb	Naive Bayes	0.4997	0.5850	0.4997	0.3821	0.3679	0.0198	0.0399
dummy	Dummy Classifier	0.4994	0.5000	0.4994	0.2494	0.3326	0.0000	0.0000
svm	SVM - Linear Kernel	0.3043	0.0000	0.3043	0.2231	0.2482	0.0001	-0.0007

Table 8: Results of fitting estimators of different families

<IPython.core.display.HTML object>

Thus, the best model by F1 measure is CatBoostClassifier, which is renowned for scoring fairly well on tabular data, while ordinary GBC is the most second to prime and the most robust one, featuring best conventional recall, accuracy, and precision metrics.

However, while all top models in Table 8 demonstrate significant improvement over a dummy classifier and other simplistic models such as Logistic Regression, the scores are still which indicates that reduction of dimensionality, which is inevitable under given class imbalance, has come at a price of variance loss, or, alternatively, all the covariates do not explain sufficiently well the outcome of studies: in academic success, as in life, a lot depends on the proper characteristics of a person which are difficult to elicit and much is undetermined. After all, it is a matter of principle whether to continue studying despite all ordeals.

Conclusion

With this analysis, we have some valuable insights some crucial factors like Academic support, socioeconomic factors, previous qualifications, and others play a significant role in student retention.

The observed patterns imply a lot to stress in the lives of students and their associates. First, we strive to insentivize parents to improve their labour efficiency and

pursue greater career so that ultimately they could dedicate more time to their children's education, and proactively stir their self-propelled interest. Additionally, we could provide financial assistance to those who are struggling to pay with if this is contemporaneous with a significant degradation in their university marks, as this subrogates the stimuli for a person in an age where they are most perceptive to knowledge and is a good predictor of a dropout. Also importantly, we could teach the students, especially going on their second studies, that it is quite unlikely that they are going to get high grades or exit the university without proper time management and confirmation that they assign top priority to their studies. They are also advised to make that clear to all their relatives and stakeholders who might underestimate the effects of such a change. Although this could result in a reduction of enthusiastic entrants, this would increase at least the KPI of retention and arguably also increase the KPI on number of diplomas issued, because with fewer but more motivated students the university would have more time to dedicate to most obstinate pending alumni.

Addressing these factors carefully can effectively lead to dropout rates reduction and improve overall student outcomes

References

- Cortes, Sylvester, Alma Agero, Elena Maria Agravante, Janelyn Arado, Cynthia Anne Arbilon, Eddalin Lampawog, Arlene Fe Letrondo, et al. 2023. "Factors influencing students' intention to enroll in Bachelor of Science in Biology: A structural equation modelling approach." Publisher: Cogent OA _eprint: <https://doi.org/10.1080/2331186X.2023.2273635>, *Cogent Education* 10 (2): 2273635. ISSN: null, accessed December 11, 2023. <https://doi.org/10.1080/2331186X.2023.2273635>. <https://doi.org/10.1080/2331186X.2023.2273635>.
- Knutsen, David. 2011. "Motivation to Pursue Higher Education." *Ed.D. Dissertations*, https://digitalcommons.olivet.edu/edd_diss/26.
- OECD. 2019. *Education at a Glance 2019: OECD Indicators* [in en]. Education at a Glance. OECD. ISBN: 978-92-64-80398-5 978-92-64-43527-8 978-92-64-88811-1 978-92-64-64514-1, accessed December 11, 2023. <https://doi.org/10.1787/f8d7880d-en>. https://www.oecd-ilibrary.org/education/education-at-a-glance-2019_f8d7880d-en.
- Temple, Shawn. 2009. "Factors that Influence Students' Desires to Attend Higher Education." *Seton Hall University Dissertations and Theses (ETDs)*, <https://scholarship.shu.edu/dissertations/420>.