The Impact Of Socio-Economic Background
ON
The Access To Higher Education

Monique Alves Cruz

Contents

1 Introduction

In most parts of the world, specially in developed countries, the massive expansion of education has resulted in full access (and completion) of primary and secondary educations. Nowadays, public policy targets have shifted towards achieving mass enrolment in higher education. While socio-economic background impacts student's access to higher education, recent studies (Orr et al., 2017) have shown that the admission systems also play an important role in making higher education accessible to students of all backgrounds.

According to Orr et al. (2017), the admission systems can be divided into four types as seen in Figure (1.1). These admission systems vary in type of selection criteria and type of streams in the secondary education. In Type 1 systems, for instance, there are paths in secondary school that do not lead to a higher education. Moreover, Orr et al. (2017) have found that type 3 systems are usually more inclusive, for both gender and socio-economic background, than type 1 systems. Nonetheless, this behavior can vary significantly among countries with similar systems: While in the Netherlands only 17% of the students belong to a financially not well-off family and 52% belong to a very well-off family, in Germany 27% of the students belong to a financially not well-off[1] family and 27% belong to a very well-off (Hauschildt et al., 2018).

Selection / **Streaming**	(Nearly all) HEIs can select with additional criteria	HEIs cannot select with additional criteria (in normal circumstances)
At least one pathway through the school system does not lead to a qualification enabling higher education entry (to some part of the system)	***Type 4: Double selection*** *Czech Republic, Iceland, Montenegro, Norway, Romania, Serbia, Slovakia, Spain, United Kingdom*	***Type 1: Selection by schools*** *Austria, Belgium, Denmark, Germany, Hungary, Italy, Luxembourg, Netherlands, Poland, Slovenia*
In general, all pathways may lead to higher education entry (in some part of the system)	***Type 2: Selection by HEIs*** *Bulgaria, Croatia, Cyprus, Estonia, Finland, Portugal, Lithuania, Latvia*	***Type 3: Least selection*** *Albania, France, Greece, Ireland, the former Yugoslav Republic of Macedonia, Malta, Sweden, Turkey*

Figure 1.1: Typology of Admission Systems in Europe. Table taken from (Orr et al., 2017).

A good amount of European countries provide free (or low fee) higher education for their citizens. The cost of tuition fees also plays a role in the inclusiveness of the higher education system. As an example, a recent study based on millions of anonymous tax records, in the US, have shown that less than 20% of the students in the bottom 60% are currently studying in one of the Ivy Leagues[2].

In light of the scenario discussed above, Brazil is a good case study because its best universities are public, either funded by the Federal or the State Governments, and the admission system is of type 3 which is usually more inclusive. In this scenario, we would expect the student population in Brazilian universities to reflect the general population in both social and economic conditions.

Up to 2008, in order to enter the undergraduate program of a federal university, one had to take the entrance exam for that particular university. This means that if someone wanted to increase his/her chance to get into an undergrad program, one had to take as many exams as one could. There were two problems with this approach: Overlapping dates and travelling expenses.

[1] It is important to keep in mind that the survey for the 2016-2018 report uses the **not well-off** to **very well-off** metric and not actual income values. In the 2005-2008 report that uses actual income values, the inclusiveness of the countries differs from the most recent report.

[2] https://www.nytimes.com/interactive/projects/college-mobility/harvard-university

In order to make public universities more accessible, in 2009 the government started adopting a single admission exam in most, if not all, Federal Universities and some State universities. Instead of creating a brand new exam for this purpose, they decided to make modifications to an existing exam, the ENEM, and adapt it to this aim.

The *Exame Nacional do Ensino Médio* (ENEM) is an exam created in 1998 to assess the quality of High School education in Brazil. In the beginning, this exam was used only for this purpose. However, over the years, some universities would accept its scores as a way to give some advantage to the applicants that performed well in it. Only in 2009, as explained above, the ENEM became the admission exam for almost all Federal Universities.

When registering for the ENEM, takers are required to fill out a socio-economic form. Therefore, the Ministry of Education, MEC, possesses information for millions of people regarding their performance in the exam and their socio-economic background. Fortunately, all this information, with exception of the takers identities, is public and available at the government website[3].

The aim of this project is to analyze the impact socio-economic background has on the access to higher education in Brazil and to show that, in a country with high inequality inclusion is not achieved even if all the other mechanisms are supposedly in place. We will also make models to predict the chance of a particular person to access higher education based on their social and economic conditions. In Chapter 2, we describe the data and the processes for cleaning and transforming it. Chapter 3 describes the analysis and presents our conclusions.

[3]http://inep.gov.br/microdados

2 Data Preparation

MEC provides *cvs* files with data from 1998 to 2017. These files contain information related to the exams scores and the socio-economic background of the takers. They have approximately 150 attributes, however, most of them were not relevant for the purpose of this project. Table (2.1) shows the description of the features selected for the analysis. As it can be seen, there are 5 grades in the database, because the ENEM is comprised by 5 exams: Math, Natural Sciences, Humanities, Reading, and Writing exams.

Table 2.1: Features used in the analysis.

Feature	Description
NU ANO	The year the exam was taken.
SG UF RESIDENCIA	Code of the State the taker lives in.
NU IDADE	Age of the taker.
TP SEXO	Gender of the taker.
TP ESTADO CIVIL	Marital Status.
TP COR RACA	Race/Ethnicity.
TP ST CONCLUSAO	High School Status.
TP ANO CONCLUIU	Year of Graduation.
TP ESCOLA	Type of School.
NU NOTA CN	Grade of the Natural Science Exam.
NU NOTA CH	Grade of the Humanities Exam.
NU NOTA LC	Grade of the Reading Exam.
NU NOTA MT	Grade of the Math Exam.
NU NOTA REDACAO	Grade of the Writing Exam.
Q001	Father's highest level of Education.
Q002	Mother's highest level of Education.
Q005	Number of household members.
Q006	Monthly Income compared to the minimum wage.
Q025	Home access to the Internet.

2.1 Cleaning and Transforming the Data

We will focus on the most recent data set (2017) which contains approximately 6.7 million samples. The features with missing values are shown in Table (2.2) and it can be seen that those with a large number of missing values are the exams' scores. Since these features are the ones we are interested in estimating, there is no way to assign any values to them in an exploratory analysis and we will discard all rows with missing values. This reduces the number of samples to about 4.3 million.

In order to get a better understanding of the impact the attributes have on the grades, modifications were performed to some of them. Numerical attributes, such as age and number of household members were combined into groups and converted into categorical variables (Figure 2.1). Similarly, categorical variables, such as Monthly Income and year of graduation, were merged into existing classes or grouped into new ones (Figure 5.2 in appendix 5).

Although the ENEM is the only selection criterion to enter most public univerisities, the weight used for each of the 5 exams is decided by each institution individually. Usually, the weight varies depending on the major chosen by the applicant. For the purpose of this project, we will use a simple average of the 5 tests to create a final grade. Figure (2.2) shows the distribution of the final grade. It can be seen that the distribution is positively skewed, with the majority of the takers getting grades slightly below 500.

Table 2.2: Number of missing values per feature.

Feature	Count
NU IDADE	101
TP ESTADO CIVIL	271641
NU NOTA CN	2293781
NU NOTA CH	2029913
NU NOTA LC	2029913
NU NOTA MT	2293781
NU NOTA REDACAO	2029913

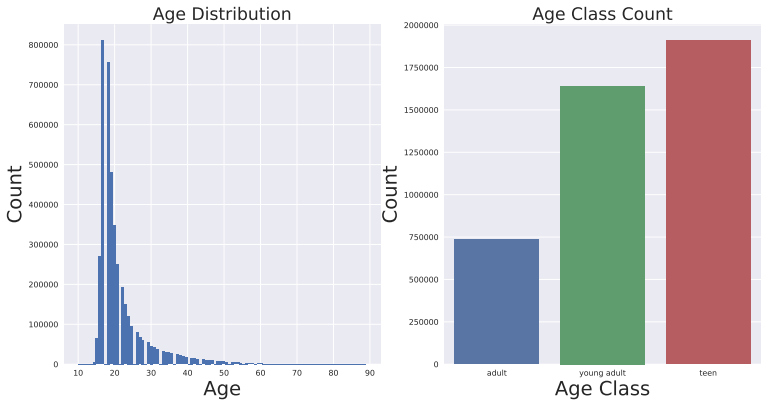

Figure 2.1: Left Panel: Age Distribution. Right Panel: Age class distribution. Age class was created as **teen** - Age ≤ 18, **Young Adult** - $18 <$ Age ≤ 25, and **Adult** - Age > 25.

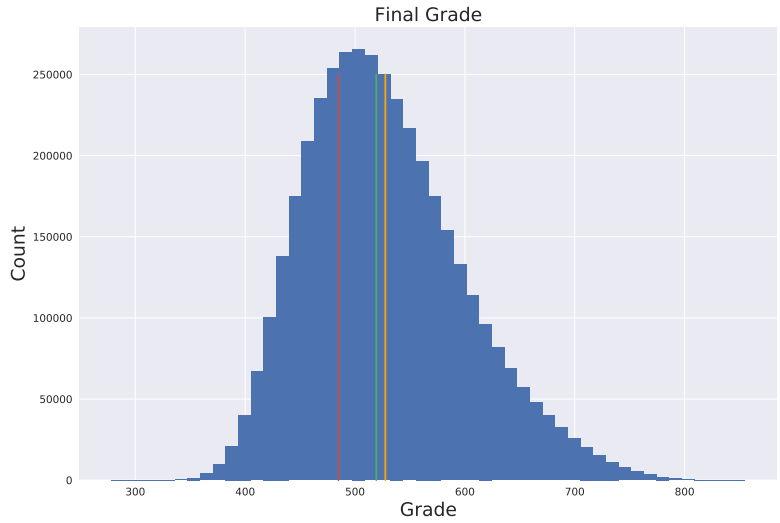

Figure 2.2: Final grade distribution. The orange line represents the mean (527.6), the green line represents the median (519.4), and the red line represents the mode (484.9). The standard deviation is equal to 72.1.

3 Analysis

We will divide the analysis in two parts: In part 1 we will investigate the socio-economic background of the exam takers and how they affect their perfomance and in part 2 we will perform predictive models to find those who are most likely to be selected to study in public universities.

3.1 Exploratory Analysis

The profile of the people who took the exam in 2017 was:

- 58.5% of the takers were female.
- The average age of the exam takers was 21 with 44.5% of them being teenagers, 38.3% young adults and 17.2% adults.
- As expected, the vast majority of them were single ($\sim$ 91%).
- The majority of takers (46.1%) declared themselves to be "parda"[1] (mixed-race), 36.7% of the takers were white, followed by 12.6% black, 2.3% asian and 0.6% indigenous. Only 1.7% of the takers did not declare their race/ethnicity.
- 56.7% of the particpants had finished high school, with 14% of them having finished it in the previous year.
- For some unfortunate unknown reason, 68% of the takers did not declare the type of school where they studied, 25.3% studied in public schools, 6.3% in private schools and 0.1% studied abroad.
- Concerning the participant's education background, only 11.7% of the fathers and 17.4% of the mothers had a higher education diploma (undergraduate or graduate).
- 50.3% of the participants had families with 4 to 5 household members (medium size), 35.5% had families with 2 to 3 household members (small), 11.8% had families with more than 6 household members (large) and 2.3% lived alone.
- 64.2% belonged to families earning up to 2 times the minimum wage (class E), 19.6% earned between 2 and 4 times the minimum wage (class D), 12.3% earned between 4 to 10 times the minimum wage (class C), 2.8% earned between 10 and 20 times the minimum wage (class B), and 1.1% earned more then 20 times the minimum wage (class A).
- 36.2% were from the Southeast of Brazil, followed by 34% from the Northeast, 11.2% from the North, 10.5% from the South and 8% from the Center-West.
- Approximately 30% of the takers did not have home access to the Internet.

The most surprising aspect of the profile above, for someone not familiar with Brazil's history, is the large amount of mixed-race participants, specially comparing with the amount of mixed-race people who took the SAT in 2017 (around 3%)[2]. Another interesting difference can be seen for the education background, while at most 17% of the parents had a bachelor's or a graduate degree among the ENEM takers, 48% of the SAT takers had a higher education background.

[1] Parda is the Brazilian word used to designate someone who has a brown skincolor. Differently from most colonies, in Brazil there was a large miscegenation among different races and ethnicities. Although, not all people who identify themselves as pardos are mixed race, the vast majority are.

[2] https://reports.collegeboard.org/pdf/2017-total-group-sat-suite-assessments-annual-report.pdf

Figure (3.1) shows the final grade distribution per Social class, education background, and type of school. The Brazilian public school system has numerous problems and, as a result, the performance of students from public schools is lower in comparison to students from private schools. For this reason, people from lower social classes (D and E), who can usually only rely on public schools, had much worse performaces than those in the upper classes, who can pay for their education.

As shown by many studies (Bar-Haim and Shavit, 2013), students with higher education background are more likely to attain higher education themselves. The reasons why the parents' education has such an impact on the students' access (and also the academic performance) to higher education is still subject to debate (Holmegaard et al., 2017; Thompson, 2017). Figure (3.1) shows a substantial increase in the performance by students with higher education background.

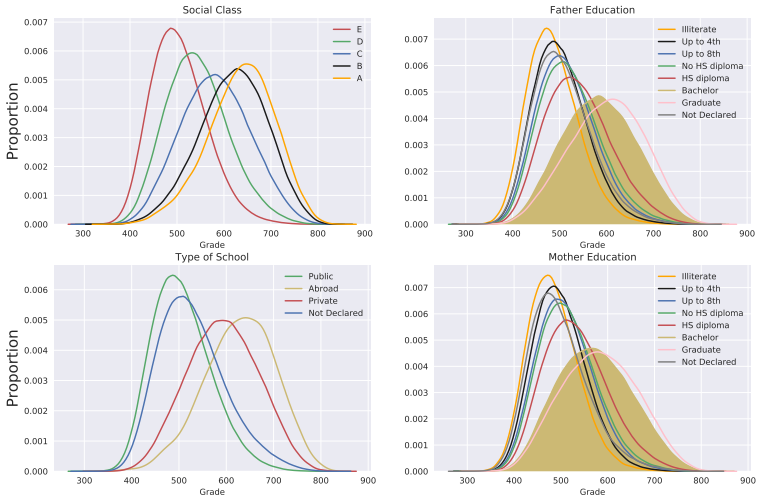

Figure 3.1: Final grade distribution for Social class (upper-left side), Father's education (upper-right size), Mother's education (lower-right side) and type of school (lower-left side).

Features such as age, gender and family size present slightly larger advantages for some of their categories. The race/ethnicity variable shows that white people usually have the best performances in the ENEM, however, this is the result of the fact that the Brazilian upper classes are mainly comprised by white people.

There were only approximately 238,000 positions offered by all universities accepting the ENEM in 2017. This means that approximately 5.5% of the exam takers were selected to study in a public institution. People who did really well on ENEM are likely to do well on the entrance exam for the top 2 universities (USP and UNICAMP) in Brazil that do not use the ENEM as a selection criterion. Thus, they are likely to choose these universities instead of the Federal ones, making possible for more people with lower grades to pass. There were about 18,000 positions in these two universities. If we assume that all students who were selected to them did not enroll in any other university, around 6% of the ENEM takers were accepted in public universities in 2017.

The percentile 94 of our sample corresponds to a grade close to 650. We now divide our sample in two class:

Table 3.1: New category: MT Class

Class	Criterion	Selection Status
0	Final Grade less than 650	Not Selected
1	Final Grade equal or larger than 650	Selected

Figure (3.2) shows the proportion of people with a particular feature belonging to class 1. We can see that while approximately 50% of people in social class A belong to class 1, less than 10% of social classes D and E achieved the same. If the system was really fair and only based on personal effort and merit, the chances for all social classes would be around 6%. Figure (3.3) shows the composition of the accepted group (class 1) for some selected features. It is evident, from the behavior seen in these Figures, the advantages that people from upper classes have.

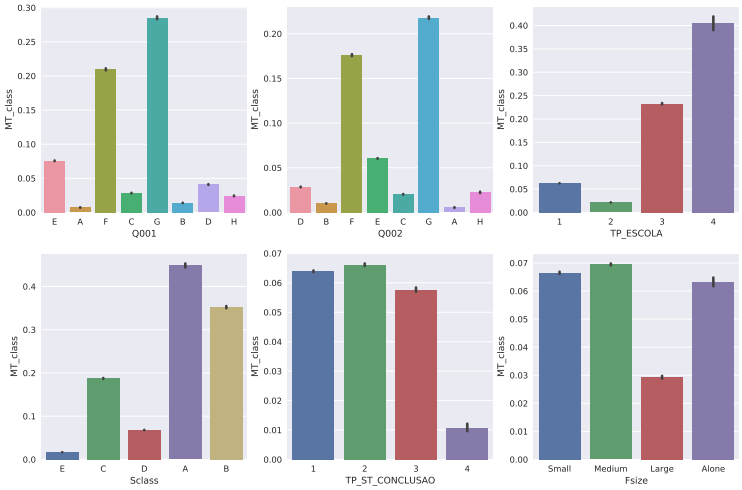

Figure 3.2: Proportion of people in class 1 for six features.

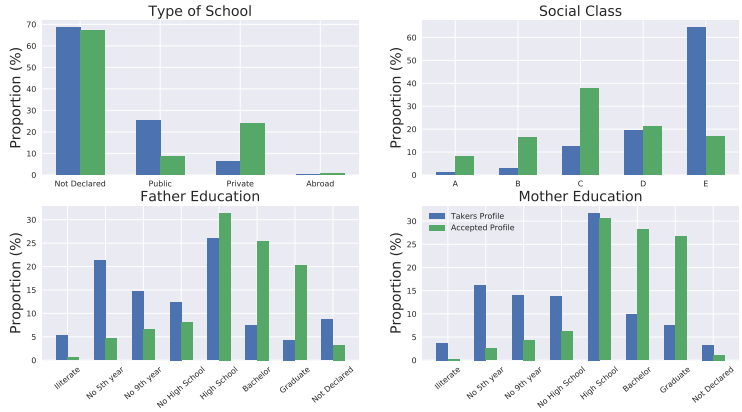

Figure 3.3: Composition of class 1 for some selected features in comparison to all the applicants.

3.2 Predictive Analysis

The analysis in the previous Section clearly shows the advantages that some social conditions have in comparison to others. There is a clear correlation between the socio-economic background of a particular person and his/hers performance in the ENEM. The next question to be answered is to what extent these advantages can be detected. Would we be able to correctly predict whether a person was selected or not based only on the socio indicators we have available for this analysis?

In order to answer this question, we have used logistic regression and naive bayes to make predictions for this classification problem. For this purpose, we selected features with the largest internal variations (as seen in Figure 3.2) to be used in the models:

- Social Class - (**Sclass**);
- Father's Education - (**Q001**);
- Mother's Education - (**Q002**);
- Type of School - (**TP ESCOLA**);
- Region;

- Race/Ethnicity - (**TP COR RACA**);
- Home access to the internet - (**Q025**)

The features were enconded using the One-Hot Enconding method, increasing the number of variables to 38 and the sample was divided in 70% for training and 30% for testing.

Since we are dealing with imbalanced classes (class 0 constains 94% of the sample and class 1 only 6%), we used a resampling technique to create a training set with balanced classes. There are two ways to resample the dataset: oversampling and undersampling. Oversampling consists in artificially increasing the minority class until balance is achieved. This technique is recommended for small sample sizes. In our case, because the sample size is large, undersampling is the recommended approach.

Among the ways to resample the dataset, random undersampling is a fast and simple technique. We used the python library *Imbalanced-learn* (Lemaitre et al., 2017) to perform random undersampling. The final training set had approximately 350,000 samples.

In terms of model performance, both logistic regression and Naive Bayes provided an accuracy of about 0.79. Nevertheless, when dealing with classification problems, accuracy is not sufficient to evaluate the model. Before applying undersampling, for instance, both classifiers had 94% accuracy, however they incorrectly labelled all elements from class 1. Figure (3.4) shows the confusion matrix, in proportion, for the logistic regression. The model was able to correctly label 79% of the samples in both classes. Table (3.2) shows the true positive rate (recall), the true negative rate (specificity) and the precision for all models used in this analysis. The same result was obtained for the Naive Bayes. Both models predicted around 20% of false positives, resulting in a low precision value.

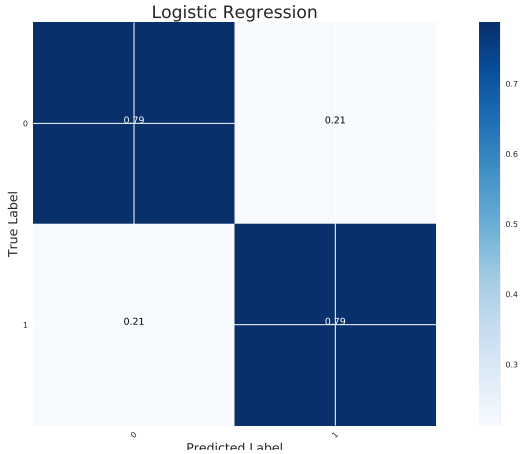

Figure 3.4: Confusion Matrix in Proportion for the Logistic Regression.

One caveat of random undersampling is the fact that information might be lost in the process of reducing the majority class. In order to minimize this effect, we have used the **Easy Ensemble**[3] and **Balance Cascade**[4] tasks of the imblearn library to create 23 balanced training datasets. We, then, used a hard voting system to decide the final class[5].

Using the **Easy Ensemble** datasets, we obtained similar results to that of a single dataset and were not able to decrease the number of misclassified samples from either of the classes. On the other hand, using **Balance Cascade** we were able to slightly reduce the number of misclassification in class 0, however, it increased the number of false negatives.

Table 3.2: Metrics for Classification.

Model	Multiple Datasets	Method	Precision	Recall	Specificity
Logistics Regression	No	—	0.21	0.79	0.79
Naive Bayes	No	—	0.22	0.78	0.78
Logistics Regression	yes	Easy Ensemble	0.21	0.79	0.79
Logistics Regression	yes	Balance Cascade	0.27	0.73	0.83

[3]**Easy Ensemble** uses random undersampling to create each dataset.

[4]**Balance Cascade** Create an ensemble of balanced sets by iteratively under-sampling the imbalanced dataset using an estimator. In our case, we chose to use logistic regression as the estimator

[5]In the hard voting system only the labels are used to decide which class is the most likely. A soft voting system takes the probabilities into consideration.

Figure (3.5) shows the composition of the misclassified samples and the composition of the samples correctly labelled as class 1. The vast majority of false negative records belong to social classes D and E. Moreover, most of them do not have a higher education background. We can conclude that these records represent people who beat the odds and that it is unlikely that they can be correctly assigned by the models using the current features. Similarly, the false positive records represent people that have the social and economic conditions to perform well (for instance, middle and upper classes), but did not in spite of that. It can also be observed in Figure (3.5) that the elements of class 1 that our model can correctly assign have the expected values for all features, as described in the exploratory analysis.

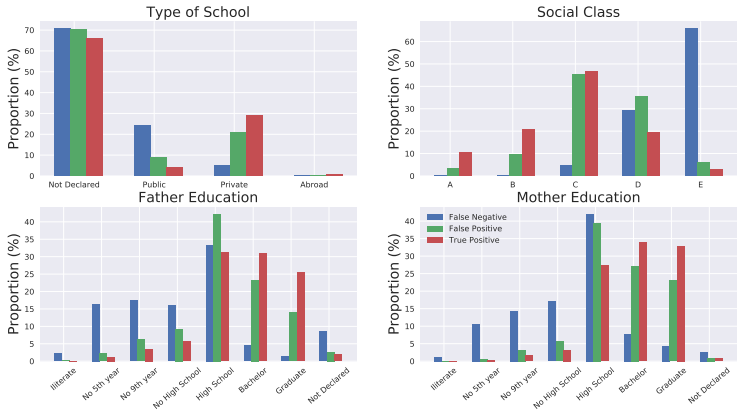

Figure 3.5: Composition of three predicted classes: Blue - False negative records, Green - False positive records, and Red - True negative records.

4 Summary and Conclusions

We have analyzed the 2017 data for the ENEM which is the exam used as a selection criterion for the public universities in Brazil. The data consists of social and economic information of the takers together with information related to their perfomance in the exam. We have performed the necessary cleaning, filtering, and transformation of the data.

The analysis was divided into exploratory and predictive analysis. The initial exploratory analysis already reveals that socio-economic background has an important impact on the access to higher education in Brazil. We have found the features that play a larger role in the takers' performance: social class, father's highest education, mother's highest education, type of school, region, access to internet and race/ethnicity. It is important to point it out that race/ethniticy is not an entirely independent variable, because the composition of social classes in terms of this variables varies significantly. For instance, while more than 70% of takers belonging to social classes A and B are white, they represent only 28% of takers from class E.

For the predictive analysis, we resampled our training dataset to deal with the imbalance between the classes. We were able to predict correctly 80% of the sample from both classes. The 20% left are mostly those not following the general trend and it is unlikely to be correctly predicted by models using the features described in the exploratory analysis. This result strenghts our initial argument that inclusion in the access to higher education cannot be currently achieved in Brazil in spite of all mechanims in place. This shows that full access to primary and secondary education alone is not sufficient if the quality of the education provided is poor.

A huge improvement of the public primary and secondary educations is the only possible option to get the country in course to a fair, merit based, admission system. Meanwhile, affirmative actions might help increase the inclusiveness and fill the gaps of inequality.

5 Appendix A: The Data

In this Appendix we present additional information concerning the data and its transformation.

Figure (5.1) shows the evolution of the number of people who signed up to take the ENEM over the years. We can see three years highlighted in this Figure, representing changes that affected significantly the number of participants. In 2001, for instance, the number of places where the exam could be taken almost doubled. In 2004, the *ProUni* was created and allowed the exam to be used as a criterion for getting scholarships, funded by the government, for private universities.

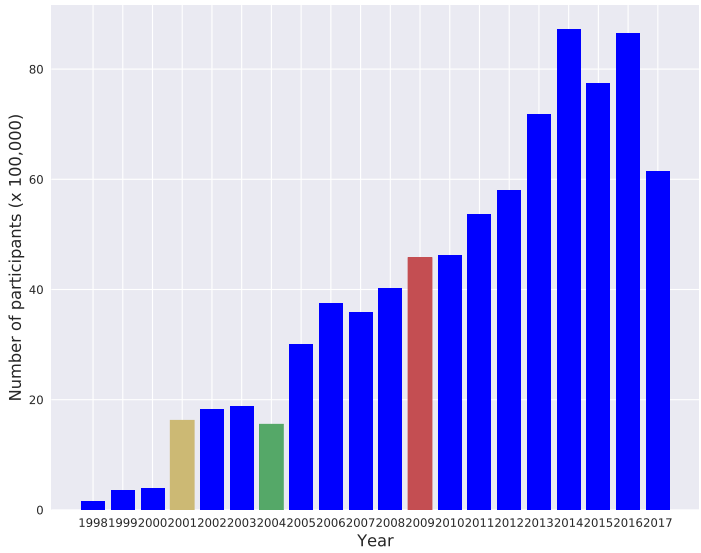

Figure 5.1: Number of people who signed up to take the ENEM per year. Highlighted are the years when major changes happened.

The Brazilian agency responsible for the census is called IBGE (Instituto Brasileiro de Geografia e Estatística). It divides the social classes in Brazil in 5 categories as seen in Table (5.1).

Table 5.1: Definition of Social Class.

Class	Definition
A	Income $> 20 \times$ MW
B	$10 \times$ MW $<$ Income $\leq 20 \times$ MW
C	$4 \times$ MW $<$ Income $\leq 10 \times$ MW
D	$2 \times$ MW $<$ Income $\leq 4 \times$ MW
E	Income $< 2 \times$ MW

Brasil is divided in five regions: North, Northeast, Center-West, Southeast, and South. Figure (5.3) shows the location of each region. These regions differ not only geographically, but also economically. The Southeast region has the largest GDP of the country and the largest population. The state of São paulo alone has more than 20% of the total Brazilian population. The North, on the other hand, is the poorest region in Brazil and the least populated. We expect that the economic differences between regions will be reflected in the quality of

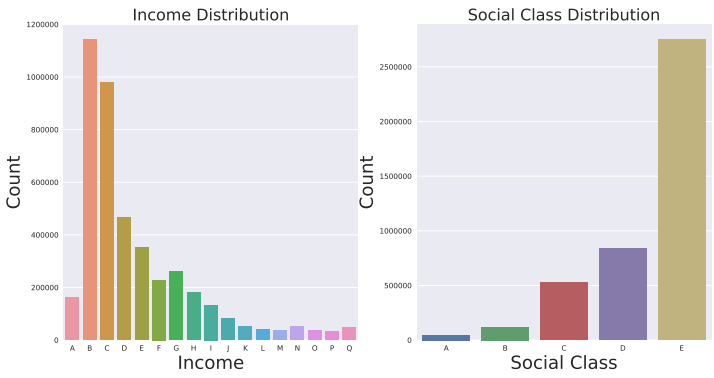

Figure 5.2: Left Panel: Income Distribution. Each letter represents a range of income. For example, A represents incomes smaller than the minimum wage of that year. Right Panel: Social class distribution.

the public and private school and, thus, in the performance of the exam takers.

Figure 5.3: The geographic location of the five regions in Brazil and the states that belong to each of them.

Another interesting aspect is related to the race/ethnicity proportion per region. The predominantly white population in the South is the result of the migration of Italians and Germans during the first and second world wars to Brazil.

6 Appendix B: Exploratory Analysis

In this appendix we present additional figures showing the impact of socio-economic background on the final grade.

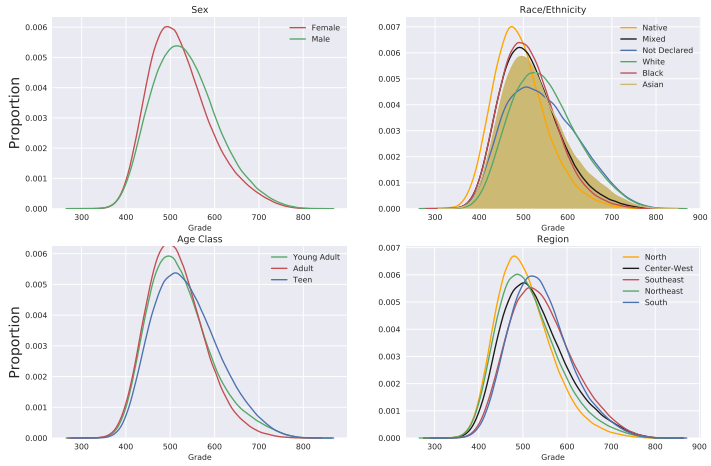

Figure 6.1: Final grade distribution for Sex (upper-left side), Race/Ethnicity (upper-right size), Region (lower-right side) and Age (lower-left side).

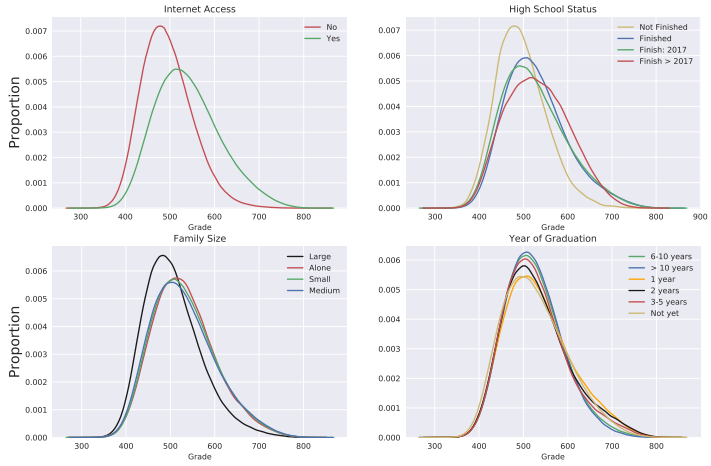

Figure 6.2: Final grade distribution for Internet access (upper-left side), High School Status (upper-right size), Year of Graduation (lower-right side) and Family Size (lower-left side).

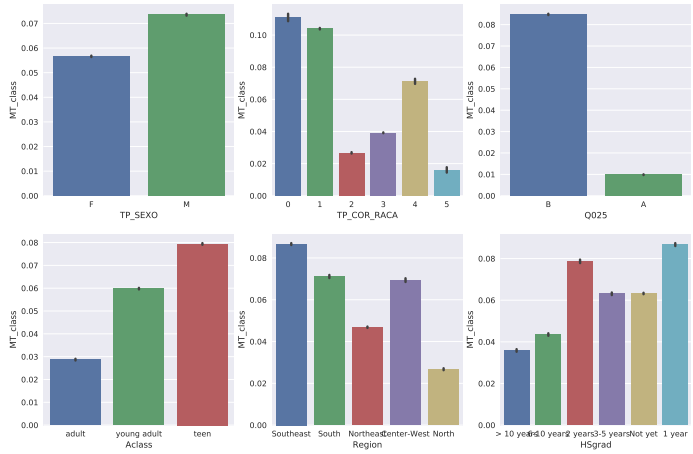

Figure 6.3: Proportion of people in class 1 for the other six features.

7 Appendix C: Predictive Analysis

In this Appendix we present additional information concerning the predictive analysis.

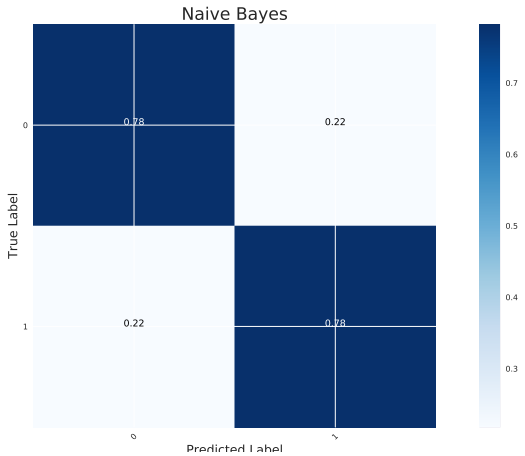

Figure 7.1: Confusion Matrix, in Proportion, for Naive Bayes.

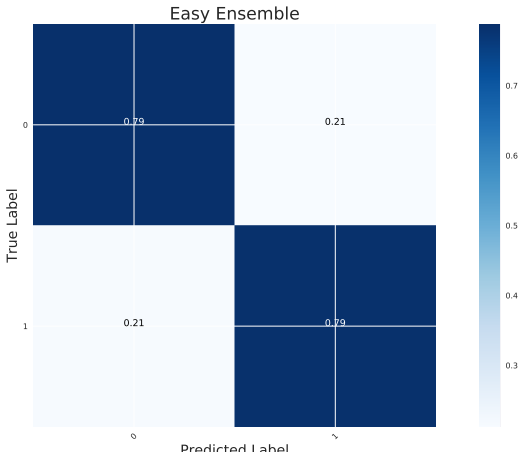

Figure 7.2: Confusion Matrix, in Proportion, for logistic regression and datasets created by **Easy Ensemble**.

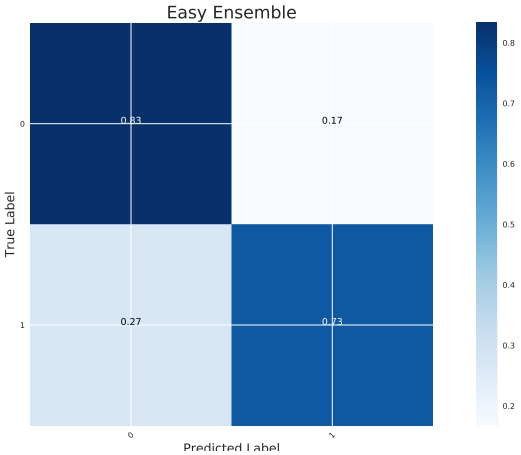

Figure 7.3: Confusion Matrix, in Proportion, for logistic regression and datasets created by **Balance Cascade**.

Bibliography

E. Bar-Haim and Y. Shavit. Expansion and inequality of educational opportunity: A comparative study. 31: 22–31, 03 2013.

K. Hauschildt, E. M. Voegtle, and C. Gwosc. *Social and Economic Conditions of Student Life in Europe. EUROSTUDENT VI 2016-2018. Synopsis of Indicators.* W. Bertelsmann Verlag GmbH & Co. KG, 2018.

H. T. Holmegaard, L. M. Madsen, and L. Ulriksen. Why should European higher education care about the retention of non-traditional students? *European Educational Research Journal*, pages 3–+, 2017.

G. Lemaitre, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18:1–5, 2017.

D. Orr, A. Usher, C. Haj, G. Atherton, and I. Geanta. *Study on the impact of admission sysytems on higher education outcomes: Volume I: Comparative report. Education and Training.* Luxembourg: Publications Office of the European Union., 2017.

R. Thompson. *Explaining Inequality? Rational action Theories of educational decision making.* In A. Mountford-Zimdars & N. E. Harrison (Eds.), *Society for Research into Higher Education (pp. 67-84).*, 2017.