Causal Inference
Final Project

Or Markovetzki 322861873
Ido Zuckerman 323102830

September 28, 2022

Abstract

Within our work in the course ”Causal Inference” we conducted a study regarding the causal effect of the age of the student at the beginning of their degree, to their probability of succeeding in finishing it in time. This study is relevant and interesting as today, many teenagers all over the world struggle to decide whether they should start studying in the university at an early age, or wait a few years so they will be more ready for this intense time. This v is also very important to those who consider joining the ”Atuda” program.

You can see our project code in https://github.com/idozuck/causal_inference [ZM]

Contents

1 Introduction

In this project, we attempt to estimate the causal effect of the age of a student's enrollment to the academia on the probability of graduating from the degree in the standard time as determined by the university. The general research question that we started with was "How does the age of the student at the beginning of their degree affect the probability of graduating in time?". We wanted to focus our research question, as there are many ways to interpret it, and eventually decided to explore the question ***"What is the causal effect of being an adult (over 21) on the probability of graduating at the predetermined period of time?"***. We chose this research question since it is interesting to see if the meaning of being an adult legally can be expressed in the maturity of the person and its prepares to begin academic studies. In addition, 21 is the age that "separates" students who decided to begin their degree before serving in the army, and after, in Israel. Therefore we can infer from our results whether joining the "Atuda" program can benefit or hurt the student.

1.1 Defining the treatment and the outcome

- *Treatment* - We define the treatment to be being an adult. I.e., $T = 0$ means the record is of a minor (under 21), and $T = 1$ means the record is of an adult (over 21)
- - We define the outcome to be graduated on time: yes or no. I.e., $Y = 0$ means the record is of someone who graduated in time, and $T = 1$ means the record does not meet the later criteria (will be explained in subsection 2.2).

2 The Data

We decided to use the data-set: "Predict Dropout or Academic Success", that you can find here [Hor22]. The data-set contains about 4500 records. The data includes many parameters about the student such as marital status, previous qualification, nationality, etc., moreover, the data include details about their parents, i.e., father's occupation. Also, there are details about the study field that they took that can be one of 17 different fields, and the data even includes the economy of the country at the time the students studied.

2.1 Features

- **Marital status** - The legally defined marital state. Can be of the following: single, married, widower, divorced, facto union, or legally separated.
- **Application mode** - Type of application, like a change of institution, transfer, change, of course, diploma holder, etc.
- **Application order** - The rating that the student gave to the application. 0 - first choice and 9 - last choice.
- **Course** - The number course of study. The course can be one of 17 different courses such as nursing (9500), social service (9238), management (9147), etc.
- **Daytime/evening attendance** - The time in the day that the course occurs, 1 for day time and 0 for the evening.
- **Previous qualification** - The education level of the student. Describe the education level of the student which can be the 12th year of school, bachelor's degree, master's, doctorate, etc.
- **Previous qualification (grade)** - The grade of the previous qualification. Number in the range 0 to 200.
- **Nationality** - The nation of the student. can be Portuguese, German, Spanish, etc.
- **Mother's qualification** - The education level of the student's mother. The attributes are the same as in the "Previous qualification" feature.

- **Father's qualification** - The education level of the student's father. The attributes are the same as in the "Previous qualification" feature.
- **Mother's occupation** - The field of work of the student's mother. can be one of the following: student, administrative staff, Armed forces professionals, teachers, etc.
- **Father's occupation** - The field of work of the student's father. The attributes are the same as in the "Mother's occupation" feature.
- **Admission grade** - The admission grade of the student. In the range 0 to 200.
- **Displaced** - A student who lacks a fixed, regular, and adequate nighttime residence. Boolean value, 0 - no, 1 - yes.
- **Educational special needs** - The student has some educational special needs the required adjustments. Boolean value, 0 - no, 1 - yes.
- **Debtor** - A student who has not paid all the payment requirements. Boolean value, 0 - no, 1 - yes.
- **Tuition fees up to date** - A student who paid all the tuition fees. Boolean value, 0 - no, 1 - yes.
- **Gender** - The gender of the student. 0 - female, 1 -male.
- **Scholarship holder** - The student received a scholarship. Boolean value, 0 - no, 1 - yes.
- **Age at enrollment** - The age of the student at the enrollment. Number in the range 17 to 70.
- **International** - The student is an international student. Boolean value, 0 - no, 1 - yes.
- **Curricular units x sem** - The number of units in the x semester can be 1st or 2nd semester. There are 6 different features of this attribute: Credited, Enrolled, evaluations, approved, grade, and without evaluations. The grade is in the range of 0 to 20, all the others are the number of units (Integers).
- **Unemployment rate** - The average unemployment rate in the student's study years.
- **Inflation rate** - The average inflation rate in the student's study years.
- **GDP** - The averaged GDP in the student's study years.
- **Target** - The outcome after the normal duration of the course can be one of the following:
  - Dropout - The student dropped the study before the normal duration of the course.
  - Enrolled - The student is still enrolled in the course at the normal duration of the course.
  - Graduate - The student graduated the course at the normal duration of the course.

2.2 Data Pre-Processing and Analysis

Before beginning with our study, we investigated the data and the features describing it, in order to understand it better and to "clean" it if needed. After studying our data, we decided to remove some features and outliers from the database.

As it is clear to see from the previous subsection, our data contains many features. All of those features might affect a student's success in accomplishing the target of graduating in the predetermined time period. However, many of those features are unsuitable for our assignment due to the fact that they are related to the student's future in the university. Since we are trying to understand the connection between a student's characteristic at the time of enrollment and their condition after a specific amount of time, we cannot use any information about the student that was collected at the end of that time period. Therefore, we cannot dropped the features "debtor" and "tuition fees up to date". Similarly, any feature related to the curricular units of the stunt was dropped as well, as they are directly related to the target and were created after the enrollment time. In addition, we decided to

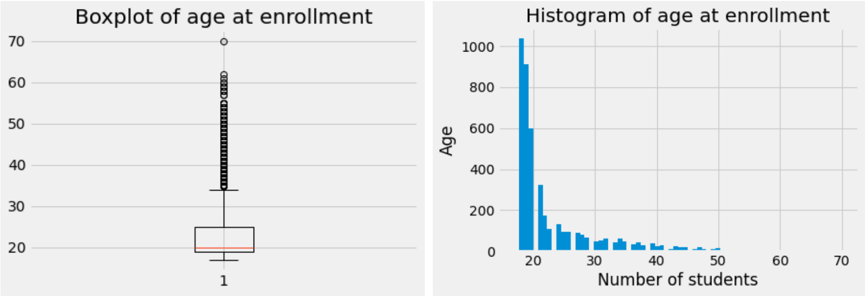

Figure 1: Box-plot and histogram of the age at enrollment

ignore the features "unemployment rate", "inflation rate" and "GDP", under the assumption that since they are not related to the student itself, they are affecting all students the same way. In [RLSM06] it is claimed that perhaps the opposite phenomenon occur, i.e. the success of the students has an effect on the economy. This hypothesis reinforce the necessary to remove these columns, as it suggests they are outcomes of the actions of the students in that period of time. The last features to be removed are "application order" and " application mode" as the first related to the time of the application submission and the later refers to many other features that some of them were removed and some related to the treatment directly (for example, "under 23 years old")

Although we did not want to temper with the records, some students listed in the database had very unique characteristics that we believe are likely to affect the outcome of the study significantly. Since these student were rare in our data-set, we decided to filter out their records. The following outliers were excluded:

- Age at Enrollment, greater than 40.
  As it is possible to see in fig.1, the majority of the students in the data are younger than forty. It is obvious that the lifestyle of a middle-aged person is different in many ways from the lifestyle of a person before the middle-age, what will probably cause a significant changes in the outcome of our study that will conceal the true meaning of it. Therefore, we decided to focus only on students that are younger than 40 years old.

- Nationality, not Portuguese.
  Students that are not Portuguese might come across other difficulties and struggles during their degree. As shown in fig. 2, there are only a few students that are from different nationalities in our data, that can change significantly the outcomes of our study. Therefore we decided to focus our study only on students that are Portuguese, as this characteristic is not related to what we are hoping to investigate.

- International, international student.
  International students usually have special classes and curricular. As can be seen in fig. 3, we do not have enough data to include those students in our study, and therefore, we decided to exclude them from our work.

The distributions of the remaining features are not irregular.

In fig.4 we can see the distribution of the possible target labels. Note that in our study we will refer to the classes "Dropped" and "Enrolled" as "Did not Graduated". In fig.5 we can see the distribution of the data between adults and minors.

2.3 Challenges

While working with the data we ran into several challenges that unfortunately we had to dill with:

1. The variance between the different ages - As discussed before, we used the rows in the data where the age at enrollment where in the range of 17 to 40, and as treatment was divided into two disjoint sets: the first set of "age at enrollment" below 21 - minors, and the other one of 21 and above - adults, which means that the minors span over 4 years from 17 to 20, and the adults span over 20 years from 21 to 40. That can cause the adults' attribute values to be more varied, while

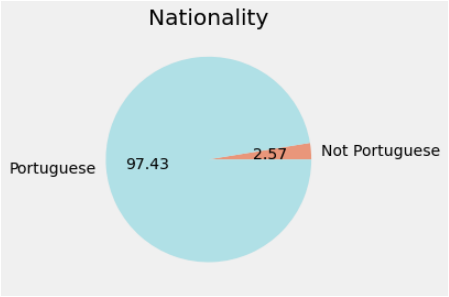

Figure 2: Nationality types distribution

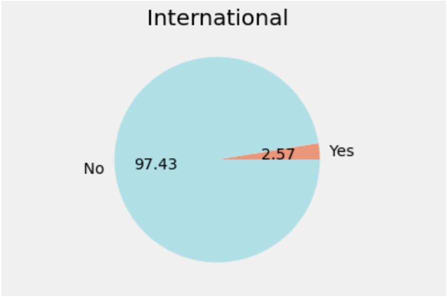

Figure 3: International types distribution

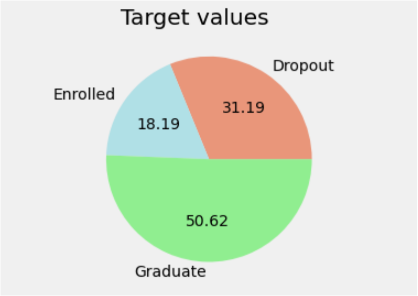

Figure 4: Target types distribution

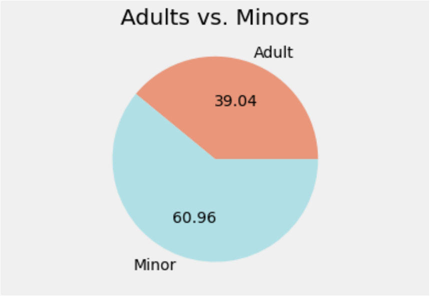

Figure 5: Segmentation by age

the minors' values are more similar to each other, hence the variance of all the other attributes might be lower.

2. Unbalanced data - There are more Adults than minors. as you can see in fig. 5 there are about 61% minors and about 39% adults, that can cause the data be unbalanced and as consequence compromise the results.

3 Assumptions

We have seen in class that in order to properly conduct a trial, we must make sure that our data meet several criteria. These criteria are assumptions that ensure us the outcomes of our calculations are not related to some other parameters, that we did not take into account, so we can truly answer our research question. The assumptions are: Stable Unit Treatment Value Assumption, consistency, ignorability and common support. In the following subsections we will explain why our data does not violate the assumptions above.

3.1 Stable Unit Treatment Value Assumption - SUTVA

The SUTVA assumption refers to two different aspects: 1. The potential outcomes for any unit do not vary with the treatments assigned to other units. 2. For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. Since our data-set was originated from a university, it is clear that the first aspect of the SUTVA assumption holds. As students ourselves, we can testify that whether our peers started their degree before or after the age of 21, does not affect our experience and studies at all. Moreover, we do not know our peers age in many cases, and therefore, it cannot even affect our feelings regarding this subject. Unfortunately, the second aspect of the SUTVA assumption does not necessarily hold. As shown in fig.1, although the treatment group only includes subjects that are (legally) adults, the range of their ages is quite wide. This fact could be considered as a violation of the SUTVA assumption, or not. We do believe that the age of the students affects their ability to finish the degree, but we also have a strong belief that the main difference is caused by being adults/minors. Therefore, since we are considering a binary treatment, and choosing only to explore the differences between adults and minors, there is no violation of the SUTVA assumption.

3.2 Consistency

The consistency assumption states that for a unit that receives treatment T, we observe the corresponding potential outcome $Y_t = TY_1 + (1 - T)Y_0$. Our study meets the criteria for this assumption as well. The data is taken from [RMBM21] that was created from data-sets of high education institutions. Because the universities take their data from the Ministry of Interior, that keeps the correct records and statistics about all people, we know for sure that the our data is correct and reliable, specifically, we know the age (treatment) of the subjects is correct. In addition, the data-set belongs to the universities and that means they can directly report the student status.

3.3 No Unmeasured Confounders - Ignorability

The ignorability assumption hold that the potential outcomes are independent of treatment assignment, conditioned on features, i.e. $(Y_0, Y_1) \perp\!\!\!\perp T|X$.
According to [OE17, LP02, EGH09], stress and motivation can have a huge influence on academic success, those features are mental measurements that are impossible to measure by the university, which our data are taken from, and in general, those measurements are hard to evaluate and not accurate so if we had them in our data it can compromise our results because they aren't accurate. Furthermore, all of our features are personal features such as accomplishments and details about their life, without any mental data, adding mental features can change the essence of the data-set. Hence, we can conclude that as we found in our literature review, there aren't any unmeasured confounders, because as we found all the confounders that we didn't take into account are mental and aren't accurate.

3.4 Common Support

The common support assumption states, that every set of features could belong to a subject from the treated or the untreated group. At first, it was obvious to us that there is no reason for this assumption to be violated. There are no restrictions stated by the university that can prevent from someone to take a specific course, receiving a scholarship, attending evening classes, etc. As mentioned before, the "application mode" contains information regarding the age of the students sometimes, and therefore, some values of this feature cannot seen with $T = 0$, but this is not a problem since (as explained in subsection 2.2) we removed this feature from our data. Parents' occupation and qualification are also not a problem for this assumption. Even the feature of "marital status" is not violating the common support, since the minimum age in Portugal for marriage is 16, and our data only contains records of people that are 17 or older, i.e., finished high school. Hence, the common support assumption is sustained. Fig.6 shows histogram based on the propensity score of the treated-adults group and the untreated-minors group, based on the results of a Logistic-Regression classifier. Note that the process of choosing this classifier for calculating the propensity score is specified in subsection 4.1. It can be seen from the graph that although the histograms do not overlap exactly, there are treated records that are classified as untreated and vice versa.

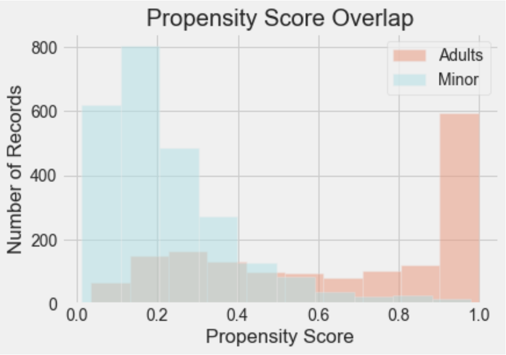

Figure 6: ROC curve of the methods

4 ATE Estimation - Setup

Average Treatment Effect (ATE) is the goal of our project. The ATE is defined as the difference between the expected outcome when receiving the treatment and the expected outcome when not receiving the treatment. I.e.,

$$ATE = E[Y_1] - E[Y_0] = E[Y_1 - Y_0]$$

The ATE is in fact, the answer to our research question. We want to understand the difference between the probability of graduating a degree on time as adults and as minors, and this measure is the answer. In this section we will present the methods we will be using in order to estimate the ATE value with respect to our data.

4.1 Propensity Score Estimator

The propensity score, $e(x) = \Pr(T = 1|X = x)$, is the probability of a given point in the data to be in the treatment group. This probability is used in several ATE estimation methods, so we tested a few estimators on our data using 5-folds validation, in order to minimize our calculations' error. We compared five models: Logistic Regression, Random Forest, Ada Boost, Bagging and Extra Trees. Fig.7 shows the ROC curves that were created for the each of these models. Form the graph, we can see that the Logistic Regression model stands out from the others in the direction of the point of $(0,1)$. The AUC values as shown in table.1 supports the observation from the graph, as it shows that Logistic regression provided with the highest score. Due to that, we decided to use this model as our estimator for the propensity score.

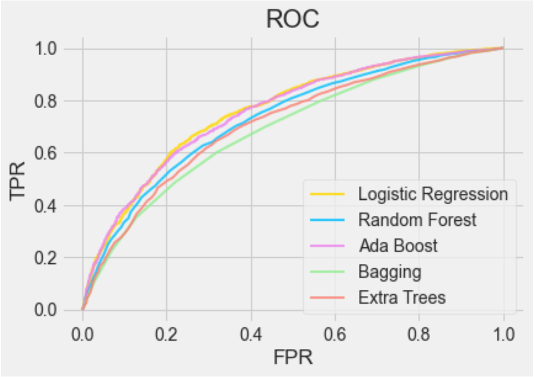

Figure 7: ROC curve of the methods

Method	ROC AUC
Logistic Regression	**0.7601**
Random Forest	0.7338
Ada Boost	0.7567
Bagging	0.7034
Extra Trees	0.7096

Table 1: Area under ROC curve for every method

4.2 Methods

- **IPW** - Inverse Probability Weighting with propensity scores. This method calculates the ATE using estimation of the propensity, as below:

$$\widehat{ATE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{t_i y_i}{e(x_i)} - \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(1-t_i) y_i}{1-e(x_i)}$$

  where:

$$\hat{e}(x) = p(T = 1|X = x), 1 - \hat{e}(x) = p(T = 0|X = x)$$

- **S-Learner** - In this method, we fit a model $f(X,t)$ with t as a feature on the entire sample: $\hat{y} \approx f(x,t)$. Then predict on all the data the ATE as follows:

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$$

  According to [kjy22], the S-learner can be biased toward zero.

- **T-Learner** - n this method, we fit a two different models $\hat{Y_0} \approx f_0(X)$ and $\hat{Y_1} \approx f_1(X)$, one model we fit on the treatment group (Adults) and the other model we fit on the others (Minors). Then we predict the ATE on all the data as follows:

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f_1(x_i) - f_0(x_i)$$

  According to [kjy22], when there are no common trends between the response under control and the response under treatment, and the treatment effect is very complex, T-Learner performs well.

- **Matching** - In this method, we decided to use 1-NN matching for calculating the ATE, which calculates the ATE by dividing the data into two groups, the treated (Adults) and the not treated (Minors), and calculate the ATE by the sum of the distance between the label of each point to the closest point from the other group.
  That is,

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\widehat{ITE}(i)$$

  $$\widehat{ITE}(i) = t_i \cdot (y_i - y_{j(i)}) + (1 - t_i) \cdot (y_{j(i)} - y_i)$$

  where:

  $$j(i) = argmin_{j s.t. t_i \neq t_i} dist(x_j, x_i)$$

- **Doubly Robust** - In this method, as in the T-Learner, we calculate the ATE by fitting two different models, one on the adults, and one on the minors, then predict on all the data, and finally using the formula below where $\hat{\mu_0}$ and $\hat{\mu_1}$ are the predictions of the two models.

  $$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_i \cdot (y_i - \hat{\mu_1}(x))}{\hat{e}(x)} + \hat{\mu_1}(x) + \frac{(1 - t_i) \cdot (y_i - \hat{\mu_0}(x))}{1 - \hat{e}(x)} + \hat{\mu_0}(x)\right)$$

5 Results

In this section, we implemented and tested the methods from Methods on our data in order to drew conclusions and determine the answer regarding our study question.

5.1 Code

To examine the methods that we presented in the Methods subsection we implemented them in the attached python notebook ”methods.ipynb”, then we run a bootstrap with 1000 repetitions to achieve confidence intervals of the ATE of each method. The methods were implemented by the formulas from Methods and we used the SK-Learn library for the implementation of the logistic regression. In addition, we used also matplotlib library in order to sketch box plots of the results.

5.2 Methods’ Results

Using our code as described above, we calculated the ATE based on the methods. In fig.8 We can see that the values and areas of the boxplots of the S-Learner, T-Learner and Matching methods have a similar values’ range. The results in table.2 support the results from the boxplots. The confidence intervals of these three methods are very similar to each other, lower bound around -0.26 and upper bound around -0.16, with average and median around -0.21. The fact that is these three methods yielded values that are very similar to each over and a median that is very close to the average, is leading us to believe that we can trust these results and that our trial was conducted properly. We are not surprised that the methods IPW and Doubly-Robust yielded boxplots and confidence intervals that are far from the rest and each other; as we have seen in the lectures, when we have some features that are biased towards the treated/untreated group, the calculations of the propensity score are less accurate and therefore methods that are using this score are more vulnerable. In our data-set, we have seen that the features ”daytime/evening attendance” and ”marital status”, for example, are imbalanced, as the values ”married”/”widow”/”divorced” and ”evening” that belongs to the feature respectively, are shown in adults’ records mostly. IPW and Doubly-Robust are both methods that are based on the propensity score, and we believe that this is the cause of their results being a bit abnormal. Therefore we concluded that the ATE value is approximately -0.21. I.e., the probability of a minor to finish the degree in the predefined time period is greater than the probability of an adult in 0.21! This result is very interesting since we thought the probability of succeeding increases as a person becomes more mature, but we are happy to wrong :).

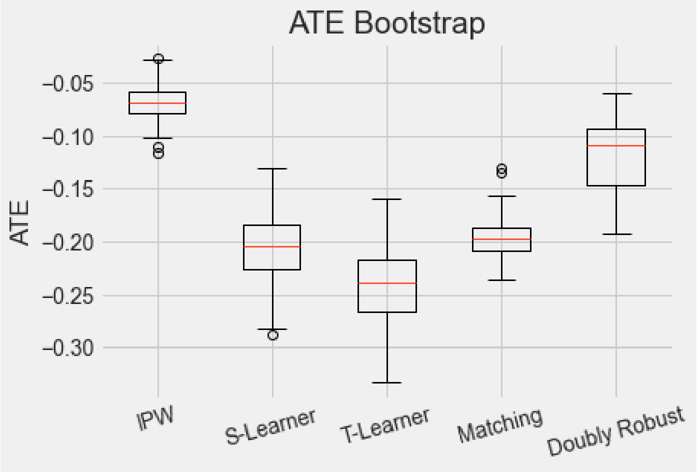

Figure 8: Bootstrap results of ATE

Method	Average	CI Lower Bound	CI Upper Bound
IPW	-0.0677	-0.1017	-0.0292
S-Learner	-0.2062	-0.2754	-0.1516
T-Learner	-0.2410	-0.2995	-0.1794
Matching	-0.1954	-0.2300	-0.1571
Doubly Robust	-0.1178	-0.1788	-0.0668

Table 2: ATE statistics for every learner using bootstrap.

6 Possible Weaknesses

While working on our project, we noticed several weaknesses in the data-set and the process that might compromise our final results. In the following section we will view those weaknesses and explain why we decided to continue with our project the way it is.

- **Unbalanced Features** - As described in The Data section, there are some features that are unbalanced between the treated/untreated groups, i.e., one of the groups has a tendency for specific values in some features, or the some values have a tendency for a specific group. For example, in the ”marital status” feature, most of the control group (minors) have a value of ”single”, unlike the treatment group (adults) that have significantly higher rate of the value ”married”. Unfortunately, unbalanced features can affect our calculations, especially in the propensity score estimations, that influence the IPW and the Doubly-Robust methods. We overcome this weakness by using several methods, and as mentioned in the previous section, we received promising results from the S-Learner, T-Learner and Matching methods, that were not supposed to be affected by this weakness.

- **Unmeasured Confounders** - As discussed in the subsection No Unmeasured Confounders - Ignorability, there are more features that could have been measured and be confounders. Most of them are mental related, e.g., stress and motivation, and therefore are hard to evaluate and measure, and more importantly, might change during the period between the beginning of the predetermined time period and its end. Of course, it is always possible to find more confounders to explore, but we believe that we included all reasonable the confounders in our data. Hence, we can conclude that as we found in our literature review, there aren't any unmeasured confounders.

- **Variance Between Different Age Groups** - The age range of the untreated group, minors, is between 17 and 20, while the age range of the treated group, adult, is between 21 and 40. Similarly to what was discussed in the subsection Challenges, the wide age range of the treated group can lead us to inaccurate results, because we cannot know if the older subjects caused the big difference we have seen in the results. We could have chosen the group of adults to be in the range of 21 to 30, and the results might have been changed significantly. But, we decided to use this range after all so our study will match our research question, as this wide age range is what we wanted to test.

7 Discussion

Our project focused on a subject that is very close to us - the odds of succeeding in the university. More specifically, the subject of our project is the age's influence on the probability of graduating in time. Our research question was ***"What is the causal effect of being an adult (over 21) on the probability of graduating at the predetermined period of time?"*** and we tested using a data containing more than 4000 records that is originated from a Portugal and was taken from [Hor22], [RMBM21]. We tested our question using five of the methods learned in class with bootstrap, and received some very interesting results; the causal affect of being an adult is bad, that is, the probability of finishing a degree on time is lower as on adult than as minor. We believe that the reason for this phenomena is that when starting the academic studies as a minor, the knowledge gained in high school is still "fresh", and the studying skills have not faded yet. Although the weaknesses we pointed out, the results we received are very conclusive, and we do not doubt them. As explained in previous sections, every weakness was taken under consideration, and we examined our results properly and carefully in order to avoid wrong conclusions. In future work, we would like to examine the same data and the same research question, using different methods, measures and a non-binary treatment. These changes might pure some more light on the matter and will help us understand more deeply the causality behind our results. We enjoyed working on this project very much and it exposed us to new methods and aspects that can help contribute to us in many other projects.

References

[EGH09] Daniel Eisenberg, Ezra Golberstein, and Justin B Hunt. Mental health and academic success in college. *The BE Journal of Economic Analysis & Policy*, 9(1), 2009.

[Hor22] Ankan Hore. Predict Dropout or Academic Success. https://www.kaggle.com/datasets/ankanhore545/dropout-or-academic-success, 2022.

[kjy22] kjytay. T-learners, s-learners and x-learners. https://statisticaloddsandends.wordpress.com/2022/05/20/t-learners-s-learners-and-x-learners/, 2022.

[LP02] Elizabeth A Linnenbrink and Paul R Pintrich. Motivation as an enabler for academic success. *School psychology review*, 31(3):313–327, 2002.

[OE17] Patrick Owusu and George Essel. Causes of students' stress, its effects on their academic success, and stress management by students, 2017.

[RLSM06] Francisco O. Ramirez, Xiaowei Luo, Evan Schofer, and John W. Meyer. Student Achievement and National Economic Growth. *American Journal of Education*, 113(1):1–29, 2006.

[RMBM21] Valentim Realinho, Jorge Machado, Luís Baptista, and Mónica V. Martins. Predict students' dropout and academic success. https://doi.org/10.5281/zenodo.5777340, 2021.

[ZM] Ido Zuckerman and Or Markovetzki. Our project in github. https://github.com/idozuck/causal_inference.