Università degli Studi di Firenze

SCUOLA DI ECONOMIA E MANAGEMENT

Corso di Laurea Magistrale in Statistica, scienze attuariali e finanziarie

Analysis of the prediction ability of a university self-evaluation test: Statistical Learning methods for predicting student performance

Mentor:
Leonardo Grilli

Candidate:
Eni Hasa

Academic Year 2016-2017

Contents

List of Figures

List of Tables

Introduction

The School of Economics and Management, University of Florence, uses the self-evaluation test as an instrument to verify the knowledge of students who want to enrol in the three-year degree program. It contributes to the process of orientation towards the choice of the university course. The aim of this study is to evaluate if the self-evaluation test adds information on predicting the student's performance over the variables available before enrolment. The prediction capacity is helpful as it would allow the identification of inactive students or who have low performance. Delays or failures are costs that affect both students and public administration (Grilli et al., 2016).

The measure of the student's performance is based on the number of credits gained after one year. ECTS (European Credit Transfer and Accumulation System) credits represent the workload and learning outcomes of a given course. Credits allow the comparison between different courses of Italian and European universities, through an assessment of the workload required by the student in certain disciplinary areas for the achievement of defined training objectives. They facilitate student mobility between different courses, but also between Italian and European universities. 60 credits represent the equivalent of one year of study or work and a credit usually corresponds to 25 hours of work including lessons, exercises, etc., but also home study. The gain of credits happens when the students pass the exam.

The MIUR (2016)[1], Ministry of Education of the University and Research, uses the gain of at least of 20 credits as a criterion for allocating the share of premiums and the equalization operation of the State Funding Fund. Universities use the threshold of at least 40 credits as a performance indicator for a regular continuation of studies.

We will use the gain of " > 0", " ≥ 20", " ≥ 40" credits as indicators to evaluate student's performance for the academic year 2014/2015. The self-evaluation test is compulsory but does not preclude the enrolment. It consists of 24 written questions multiple responses, one of which it is correct. The topics concern logic, verbal comprehension and mathematics. For each of them, the students will have 20 minutes. For each correct answer the candidates will have a score of 1 point, for the wrong answer -0.25 and

[1] http://attiministeriali.miur.it/anno-2016/dicembre/dm-29122016.aspx

0 points for the response not given. Based on this evaluation, each student will have a final score. Students who obtain a scores equals or more to 8, will be able to take the exams. In case of scores less than 8, students have to study the material indicated by the university. Then, they have to give another self-evaluation test for seeing the improvement in December.

The test can be done on paper, in September, or on the computer in November. According to Article 1 of the announcement[2] on the verification of entry knowledge for those who intend to enrol in the three-year degree program of economics, are excluded those who:

- *Are already in possession of an Italian university degree.*
- *Have done and passed the Cisia Economics test at the consortium universities.*
- *Already enrolled in another study program at the University of Florence, but this student requested the passage to a School of Economics Management provided that they have already supported and passed the verification test at the School (or Faculty) of origin.*

Two problems will be addressed. In the first, we will compare the performance between students who did the test and those who were exempted. We will evaluate whether the difference in performance can be given by the test session. Statistical matching techniques are used for comparison. The second problem regards the evaluation of the capacity of the test in predicting student's performance. We use the methods of the logistic regression and random forest. The results of the two methods, based on the average of prediction error derived from 10-fold cross-validation, are compared.

The thesis is structured as follows. Chapter 1 describes the data selection and the variables used for the analysis. In Chapter 2 we will evaluate if students who were exempted from the test have different performance compared to those who did it. Then we want to assess if the performance of the student is influenced by the period of the test. Making the test in September or later can affect the regularly studying. Chapter 3 refers to the method of logistic regression. We use this method for the prediction of student's performance and to evaluate the possible addition of information of the test score variables. In Chapter 4 we will use the learning method of random forest for comparison with logistic regression. We want to find which method best predicts student's performance and whether the test adds information to predict student's performance.

[2] https://www.economia.unifi.it/upload/sub/test-autovalutazione/bando-test-autovalutazione-2014-15.pdf

Chapter 1

Data description

The original data set is a merge of the administrative career archive and the test archive. We have 978 observations and 56 variables. We delete 19 observations because they are students with a diploma obtained abroad. Also, other 2 observations were deleted because the variable *High school degree* was missing. Most parts of the variables record detailed information of the performance during the academic year. We will use informations on overall performances. The goal is to evaluate the addition of test information on predicting the student's performance, besides the variables already known before the enrolment. We delete also 88 students which didn't take the test in according to Article 1 of the announcement on the verification of entry knowledge. So we limit the analysis on 869 students who took the test and enrolled at the School of Economics and Management in the 2014/2015 academic year. As Grilli et al. (2016), we divide the variables into three big groups and we will consider the **pre-test variables** for each student:

- *gender* = take value 1 if the student is male, otherwise 0.
- *residence* = take value 1 if the student have the residence in "Florence", "Arezzo", "Pisa", "Pistoia", "Prato", otherwise 0. We consider this partition because these cities are neighbouring to the University by one hour by train, so these students don't need to move in Florence. The transfer involves different distractions for far-away students when they have the freedom to live alone.
- *late enrolment* = take value 1 if the age at high school diploma > 19, otherwise 0. We consider this condition to distinguish students who had a regular career in the high school from those who have been rejected at least once in the high school
- High school Type: *Scientific, Humanities, Technical, Other*
- *High school grade* (from 60 to 100)

The **test-variables**, which is the score on Logic, Reading, and Mathematics, are used in chapter 3 and 4 to evaluate if they add information to predict the student's performance.

The **University performance variables**, are the credits gained during the first year. For the prediction capacity of the self-evaluation test, we will consider three different outcomes. Since in the first-year there are 6 exams which have 9 credits each, the outcomes will indicate the number of passed exams. As anticipated in the introduction, to construct the > 0 credits indicator we use the outcome Y_1, which has value 1 if students gives at least one exams, otherwise 0 thus representing inactive students. The binary outcome Y_3 takes value 1 if students have passed at least three exam thus indicating students with low risk of dropping out of studies, otherwise 0. The last outcome, Y_5, represent the indicator ≥ 40 credits. It takes value 1 if students gives at least five exams and it stand for students with a regular continuation of studies, otherwise 0.

Table 1.1: Construction of the binary outcomes that represent the University performance variables

Y_1	≥ 1 exam	> 0 credits
Y_3	≥ 3 exams	≥ 20 credits
Y_5	≥ 5 exams	≥ 40 credits

In figure 1.1 we can see the distribution of the number of exams passed of the 869 students remained. There are 283 students who don't even give an exam, despite they result enrolled and have an active career at the end of the first academic year. 152 students give one exam while the presence of students who give two and three exams is very similar (112 and 114 respectively). Finally, the number of students who give 4, 5 and 6 exams is similar, (83,69 and 5). In this case, as the exams have the same number of credits, the difference in the students' performance is due to the amount of the exams.

Table 1.2 shows the score Test and the proportion of students who give "Y_1", "Y_3", "Y_5", for each variable, in 2014/2015 academic year. The biggest difference is given by the high school grade. The threshold is chosen based on the average grade (77). Students with $HSG \leq 77$ are always worse performing. Only 4.7% of students with lower HSG has given at least 5 exams compared to 23.2% with a score above 77. Furthermore, late-enrolment students have worse performances than those with a regular career.

Lower score of females than males corresponds to a worse performance, thus leading to a relationship between the test score and the carrier indicator. It's a slight relation

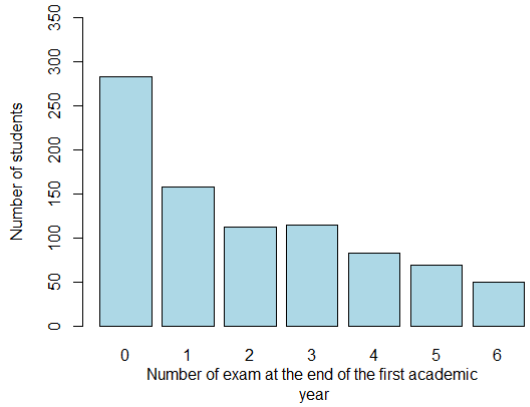

Figure 1.1: Distribution of the number of exams passed at the end of the first academic year 2014-2015, School of Economics and Management, University of Florence.

because there are no significant differences in the variables even though performance at the end of the year is different, as, in the case of late enrolment, there is no difference in score among late students but the performance for Y_5 is very different. There is a greater male presence while the number of far-away. The resident student has a slightly better performance in general against the far-away.

The greater presence of students results from the scientific and technical high school. We can see that the best-performing students are those with scientific studies. This happens because they have a good background in maths and logic. But that's not all. The greatest influence is given by the method of study. In fact, it is observed that similar performances occur from students with humanities studies as they acquire a good methodology of study necessary for text translations from Latin and Greek.

The *Change course* of the variable **Career status** concerns students who switch to another course but always belong to the School of Economics and Management. The "drop out" are students who explicitly renounce the continuation of studies. We can see that despite the drop-out test score does not look different from the other variables, in reality, only23.7% of the students give at least one exam and 2% at least 3 exams. So drop out is strongly related to not giving any exam. It's important to keep in consideration the presence of students who are enrolled but actually decided to stop the studies.

The description of the variables serves to a better comprehension of the phenomenon. We cannot use only these variables for student selection as it discriminates. Furthermore, variables such high school degree isn't fully appropriate to predict the academic performance as it has several limitations. One of these limitations is "the possible mismatch between the competencies evaluated at high school and those required for a given de-

Table 1.2: Score Test and three binary outcomes for the performance results considering each variable at the end of the first academic year 2014/2015. School of Economics and Management, University of Florence

	N	Score Test (max 40)	$\%Y_1$= 1	$\%Y_3$=1	$\%Y_5$=1
All	869	13.35	67.4	36.4	13.7
Gender					
Female	363	12.44	69.1	35.8	11.8
Male	506	14.00	66.2	36.8	15.0
Far-away resident					
Yes	144	12.57	60.4	26.4	09.0
No	725	13.51	68.8	38.3	14.6
HS type					
Scientific	297	14.78	75.4	45.1	21.2
Humanities	67	13.60	80.6	44.8	16.4
Technical	327	12.64	63.6	32.1	09.7
Other	178	12.16	56.2	26.4	07.3
Late-enrolment					
Yes	108	12.25	45.3	14.8	04.6
No	761	13.51	70.6	39.4	15.0
Hs grade					
≤ 77	448	12.54	55.1	20.1	04.7
> 77	421	14.21	80.5	53.7	23.2
Career status					
Active career	718	13.60	75.6	43.2	16.4
Change course	4	14.19	75.0	0	0
Drop out	131	12.07	23.7	01.5	0
Transfer	16	12.44	56.2	25.0	06.3

gree program" (Grilli et al., 2016). The second concerns the diversity of types of higher schools and between the different regional areas. It is necessary to evaluate whether the self-evaluation test adds information to the known variables, for predicting the performances.

Chapter 2

Comparing students' performance

In this chapter, we want to evaluate the performance of students who didn't take the test and those who took it. In the introduction, we anticipated Article 1. Students who have already obtained a degree in Italy are excluded from taking the test. It's unlikely that a graduate person decides to enrol in the three-year period. Students, who have requested a transfer from another University to the School of Economics and Management and who have been recognized as having 18 or more credits, are also excluded. Finally, fall in this category those who requested the transfer from another degree course, provided they have already supported and passed the self-evaluation test at the School (or Faculty) of origin. In this case, the test of the faculty of origin may be different from that of the economy, including different areas of evaluation.

The second goal is to evaluate whether different test sessions lead to different performance. We want to compare students who made the self-evaluation test in September, on paper, with those who made it in the second session on the computer, in November or March. This is because *a priori* we think that students enrolled in September can follow the courses and give the exam at the right time, instead, the students that take the test in later, have followed fewer lessons and have had less time to study regularly.

Causality is related to an action (doing the self-evaluation test) applied to a unit (student). A causal statement assumes that even if a unit was (at a given time) subject to a particular action (active treatment), the same unit could have been exposed to an alternative treatment (control treatment) at the same point in time. So for each unit, the outcome would be observed under the active control and active treatment, and this is called *potential outcome* because in the end only one outcome can be realized and can be observed. The causal effect is to compare these two potential outcomes. For this study, we assume the SUTVA because the potential outcomes for any student don't vary with the treatments assigned to other units, and, for each student, there are no different forms

or versions of each treatment level, which lead to different potential outcomes(Imbens and Rubin, 2015). We denote the observed outcome Y_i^{obs} for a unit $i \in \{1, \ldots, N\}$ in a population of N units:

$$Y_i^{obs} = Y_i(W_i) = \begin{cases} Y_i(0) & if \quad W_i = (0) \\ Y_i(1) & if \quad W_i = (1) \end{cases}$$

where W_i is the treatment indicator, take value 0 for the control treatment, and 1 for the active treatment. For each unit we have the missing potential outcome denoted by Y_i^{mis}:

$$Y_i^{mis} = Y_i(1 - W_i) = \begin{cases} Y_i(1) & if \quad W_i = (0) \\ Y_i(0) & if \quad W_i = (1) \end{cases}$$

For the causal effect, the presence of the missing outcome leads to an inferential problem. The key role is played by the assignment mechanism, which is the process that determines which units receive the treatment and which one takes control treatment. Although the assignment mechanism is an unknown function, because we are in an observational study, we still keep the assumption of *individualiscness*[1] , *probabilisticness*[2] , *unconfoundedness*[3] (Imbens and Rubin, 2015). These assumptions implicate that the assignment mechanism can be interpreted as the division of units into groups, where inside have the same value of the covariates. So we can give a causal interpretation to the comparison of the potential outcomes for the units being submitted at the active and control treatment, for each group. But the approach that divides the population into groups defined by the value of covariates, can create classes in which there are only treated units or just controls. In this way, it becomes impossible to detect the causal effect.

In experimental studies, characterized by the randomization of the treatment assignment vector. covariate balancing is implicitly performed. In observational studies, the treatments were not randomly assigned to experimental units, so the treated and control groups may not directly comparable (Rosenbaum, 1984). Thus, Rosenbaum and Rubin (1983) defined the balancing score, $b(x)$,as a function of the observed covariates, for

[1]The assignment mechanism is said to be individualistic if the individual probabilities, for the unit i, depend on his covariates and his potential outcome
$(p_i(\boldsymbol{X}, \boldsymbol{Y(0}, \boldsymbol{Y(1)} = q(X_i, Y_i(0), Y_i(1))$
and if multiplying individual probabilities, the result is equal to the probability of a particular assignment vector less than a constant of proportionality
$(P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y(0)}, \boldsymbol{Y(1)}) = c\prod_{i=1}^{N} q(X_i, Y_i(0), Y(1))^W (1 - q(X_i, Y_i(0), Y_i(1)))^{1-W_i}$

[2]Probabilistic assignment implies a non-zero probability for each treatment value, for each unit
$(1 \geq p_i(\boldsymbol{X}, \boldsymbol{Y(0}, \boldsymbol{Y(1)}) \geq 0)$

[3]The probability of assignment does not depend on any of the potential outcomes
$(P(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y(0)}, \boldsymbol{Y(1)}) = P(\boldsymbol{W}|\boldsymbol{X}))$

which, by conditioning this function, the conditional distribution of x given $b(x)$ is the same for the treated and the control units :

$$X \perp W | b(x)$$

The propensity score is one balancing score. Rosenbaum and Rubin (1983) define it as a function that shows the propensity towards exposure to the active treatment given the observed covariate, $e(X) = pr(W = 1|X)$. An important propriety of the balancing score is that, if the treatment assignment is strongly ignorable[4] , then it is also strongly ignorable given the balancing score. It allows so to obtain unbiased estimates of average treatment effects. In a randomized experiment, the assignment treatment is known to be strongly ignorable, and this implies to be also ignorable [5], but it is not true the opposite. While, in observational studies, the ignorable assignment treatment is a weak assumption (Rosenbaum, 1984). So, with regard to the unconfoundedness, we cannot observe it from the data, we can only conduct sensitivity analysis. While the propensity score is estimated by the observed data.

[4]The treatment assignment is strongly ignorable given a set of covariate X if $W|(Y(0), Y(1))|X$, and $0 < pr(W = 1|X) < 1$

[5]The assignment mechanism does not depend on the missing potential outcome, but it can depend on the observed outcomes

2.1 Comparison by Test

First of all, the active treatment is the variable *Test* when takes value 1. It represents students who didn't take the test. The group of control is formed by students who did the test and the variable *Test* take value 0.

Table 2.1: Composition of the binary treatment variable "Test", in reference of the 2014/2015 (a.y.). School of Economics and Management, University of Florence

test	0	1
	869	88

In the next table, we can see the number of given exams by Test. Most of the students who didn't take the test didn't even give an exam. On average they give 0.61 exams. The students, who did the test, give an average of 1.96 exams.

Table 2.2: Number of given exams by student who did the test (test=0) and student's who did it (test=1), in reference of the 2014/2015 (a.y.). School of Economics and Management, University of Florence

	0	1	2	3	4	5	6
test=0	283	158	112	114	83	69	50
test=1	60	12	11	2	2	0	1

We use the normalized difference using the notation of (Abadie and Imbens, 2011), calculated as:

$$\text{nor-dif} = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{(S_0^2 + S_1^2)/2}}$$

It provides a measure without a scale for the difference of the two distributions. We use this measure because is useful to see how much we have to adjust for the covariates. For each covariate, the average for the group under treatment and control is displayed, by scrolling along the column of each covariate, and for each one, we calculate the mean. The first step is the difference between the average of the treaties and the average of the controls, for each covariate. In the second step, the variance is calculated for each column of covariates in the respective treatment groups. Finally, the standardized difference is obtained as the ratio of the mean difference and the square root of the sum of the variances of the respective groups divided by number 2.

From table 2.3, on average, in the treated group there are more females, far-away and late enrolment students respect to the control group. The biggest difference is given by the high school degree as the average grade of the treaties is 71 compared to 78 of the control group. Furthermore, there are fewer students treated that coming from the scientific high school. As regards the Humanities and Technical variables, distributions

are almost overlapping because the standardized difference is very small.

Table 2.3: Standardized difference for each variable between the active and control group, in reference of the 2014/2015 (a.y.). School of Economics and Management, University of Florence

	med.t1	med.t0	ST_diff.med
gender	0.489	0.582	-0.188
residence	0.818	0.834	-0.042
late-enrolment	0.193	0.124	0.189
Humanities	0.114	0.077	0.124
Scientific	0.227	0.342	-0.255
Technical	0.364	0.376	-0.026
Other	0.300	0.205	0.210
HS grade	74.22	77.48	-0.286

A logistic model is used to estimate the propensity score, where the dichotomous response variable is the variable that indicates which unit is assigned to the treatment, and vice versa, which one is to the control group. Explanatory variables are: *test, gender, residence, late-enrolment, Humanities, Scientific, Technical, other, High School Grade*. The significance of covariates is not of interest since the model is used to find the propensity score, defined as the individual probability of being assigned to the treatment and is identified by the fitted values of the logistic model.

The aim is therefore to find good estimates of the probability of the assignment to the treatment. If the estimates are good it is expected that the distribution of covariates is the same between the treated group and the control group.

The balancing occurs more easily by comparing the histograms of the treated and control group. We use also to overlap the distribution of the two groups. The distribution of covariates is similar between the group of treaties and the controls. The treated are represented by the blue line in figure 2.1.

We use the matching procedure to improve the balance. Usually, it is applied in the observational study and the treatment variable is not randomly assigned (Ho et al., 2011). When we have a small group of unit's treated, we sample from a large group of potential control. In this way, we form a group of control which has a similar distribution to the treated group (Rosenbaum and Rubin, 1983). This is the reason which led us to consider as treated the student that didn't take the test. It's easier to find students, in a large pool of control, that have similar covariates as the treated.

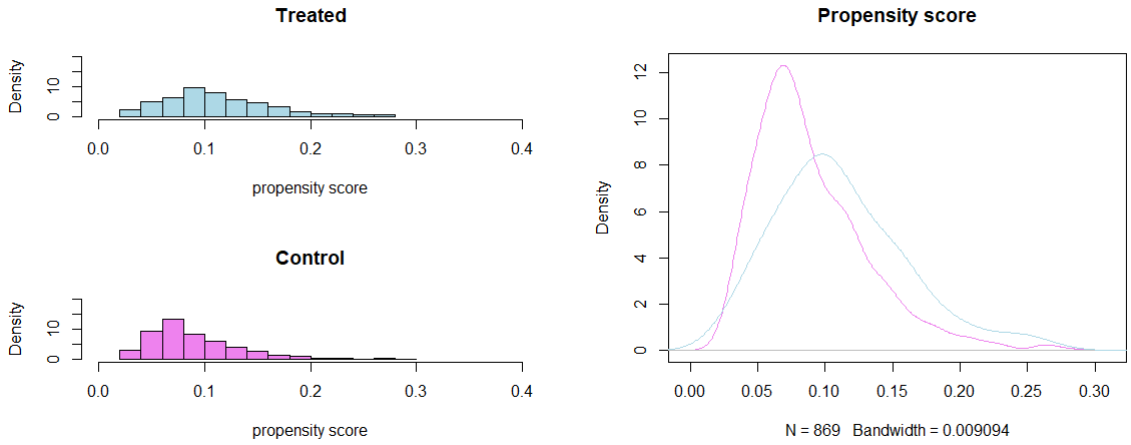

Figure 2.1: Distribution of covariates between the treaties (student who didn't pass the test) and the control group (student who passed it), in the a.y. 2014/2015. School of Economics and Management, University of Florence

In the matching, we use the propensity score as the synthesis of the covariates. We use the packages in R *MatchIt* for this analysis. If $W_i \perp X_i$, we would not need to control for X_i, and so the analysis is reduced to a difference in means of Y for the treated and control groups (Ho et al., 2011).

Stuart (2010) define the *"distance"* as a measure of the similarity between two individuals. The exact distance is determined as:

$$D_{ij} = \begin{cases} 0 & if \quad X_i = X_j \\ \text{inf} & if \quad X_i \neq X_j \end{cases}$$

Rosenbaum and Rubin (1983) showed that exact matching leads the same probability distribution of $\boldsymbol{X}$, for the treated and control groups, because each treated units is matched with all possible control, which has the same value in all the covariates. But it doesn't work very well because it produces few combinations, and discards too many units. The propensity score is defined as

$$D_{ij} = |e_i - e_j|$$

where e_k is the propensity score for individual k (Stuart, 2010). We use the nearest available matching on the estimated propensity score. The propensity score is estimated by using a logit model, as the distribution of the propensity score is approximately normal. The treated and control are randomly ordered. The process begins by matching the first treated unit with the control unit which has the closest propensity score, and then they are removed from the list of treated and control. This step is repeated for

the unmatched treated. Rosenbaum and Rubin (1985) showed that it works very well because it required less computation and it's useful in reducing bias.

The percent balance improvement is calculated:

$$\left(\frac{|a| - |b|}{|a|}\right) \times 100$$

where a is the difference between the treaties mean and controls mean in the original data, before matching, while b is the difference in means after matching (Ho et al., 2011). This difference is done for each covariate.

By looking at the data, intuition leads us that the better choice is given by the *nearest available propensity score matching*, by using the exact matching on the binary covariates, in order to improve the balance.

Table 2.4: Synthesis matching *Test*: Nearest available propensity score using the exact matching on binary covariates

Percent Balance Improvement	Mean Diff.
distance	98.55
gender	100
residence	100
late enrollement	100
Humanities	100
Scientific	100
Technical	100
Other	100
HS grade	97.98

Sample sizes	Control	Treated
All	869	88
Matched	82	86
Unmatched	787	2
Discarded	0	0

In Table 2.4 we can see how the difference between the mean of the treated and the mean of the control group after matching is 0, thus leading a percentage improvement in the balance of a 100%. Only two treated remains unmatched. The matching is done *with replacement* because it gets better matches. Without replacement, the order in which we match the treated is important because if some control was already matched, they could be the better matched for the treated units unmatched (Imbens and Rubin,

2015). The matching with replacement allows the use of the control units more than once. The advantage is that the bias is reduced; the problem is that we can often take the same control unit, increasing the variability (Abadie and Imbens, 2006). The nearest neighbour matching method in MatchIt is by default a *greedy* matching. It matches the closest control for each treated unit, one at a time. We use the order specified by default is *largest* because by first it matches the units treated with the highest values of the distance, units that are the most difficult to pairing. This happens because we can imagine the distribution of the propensity score of the left-handed controls compared to the treaties. The remaining treated units are then matched. We have defined the option *exact* on the dichotomous variable on which to perform the exact matching within the nearest neighbour matching. Only the matches that match exactly on the covariates will be allowed. Within the matches that match on the variables in exact, the match with the closest distance measure will be chosen (Ho et al., 2011).

In figure 2.2 we can observe the distribution of the propensity score between the treaties and the controls analysed before and after marching. The situation in terms of balancing has improved considerably.

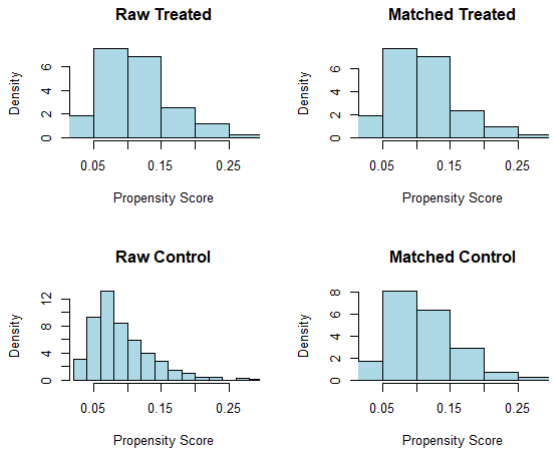

Figure 2.2: Histogram: before and after matching on the treatment variable *Test*

Through the qq.plot, in the case of good balancing between treaties and controls, we obtain points that are situated on the straight line. We use this chart as the compression of adjusting it's faster. Figure 2.3 illustrates for each covariate the balancing before and after the matching.

Now that we have two groups of similar students in terms of covariates, we can compare them to identify the effect of the self-evaluation test. For each students, the effect of treatment is $Y_i(1) - Y_i(0)$. As only one outcome is observable, for the treaties

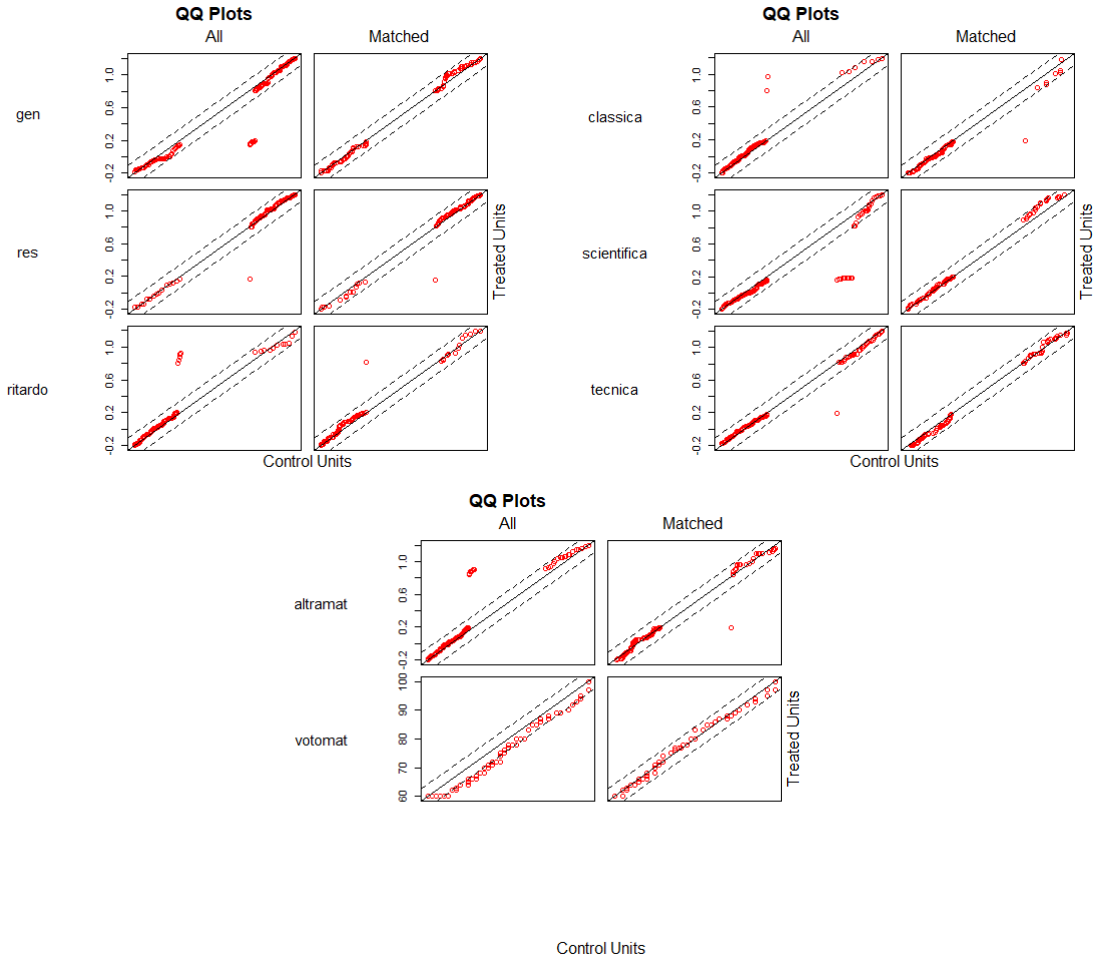

Figure 2.3: QQplot: before and after matching on the treatment variable *Test* for each covariate

we use the outcome of the matched control. We can focus on the most two estimand most used in the literature. Using the notation of (Imbens, 2004), the first one is the population average treatment effects (PATE), which is the difference of the expected number of exams after taking or not the self-evaluation test:

$$\tau^p = E[Y(1) - Y(0)]$$

Alternatively, the second estimand is population average treatment effect on the treated:

$$\tau^p_T = E[Y(1) - Y(0)|W = 1]$$

The first parameter shows the effect of randomly assigning students for doing or not the self-evaluation test. We will use the second estimand as it focuses on the effect on those who didn't take the test (Caliendo and Kopeinig, 2008).

As Becker et al. (2002) illustrate in the paper, let $C(i)$ be the set of control unit matched to the treated unit (i), and let denote the Y_i^T and Y_j^C the observed outcome of the treated and control unit. Using the nearest neighbour matching with an estimand value of propensity score, ps, define $C(i)$ as:

$$C(i) = min_j|p_i - p_j|$$

The number of control matched, with observation i in the set of the treated units, is defined $N_i^C = 82$, while the number of treated is $N^T = 86$. As we used matching with replacement, some controls are in the matched sample more than once, we define the weight as:

$$w_{ij} = \begin{cases} \frac{1}{N_i^C} & j \in C(i) \\ 0 & otherwise \end{cases}$$

The ATT is obtained by averaging the difference between the outcome of the treated units and the outcome of the matched control:

$$\begin{aligned} \tau^M &= \frac{1}{N^T} \sum_{i \in T} \left[Y_i^T - \sum_{j \in C(i)} w_{ij} Y_j^C \right] \\ &= \frac{1}{N^T} \left[\sum_{i \in T} Y_i^T - \sum_{i \in T} \sum_{j \in C(i)} w_{ij} Y_j^C \right] \\ &= \frac{1}{N^T} \sum_{i \in T} Y_i^T - \frac{1}{N^T} \sum_{j \in C(i)} w_{ij} Y_j^C \end{aligned}$$

We need to find the variance in order to calculate the confidence interval. The necessary assumptions for the estimate of the variances are fixed weights and independent outcomes across the units:

$$\begin{aligned} Var(\tau^M) &= \frac{1}{(N^T)^2} \left[\sum_{i \in T} Var(Y_i^T) + \sum_{j \in C} w_j^2 Var(Y_j^C) \right] \\ &= \frac{1}{(N^T)^2} \left[N^T Var(Y_i^T) + \sum_{j \in C} w_j^2 Var(Y_j^C) \right] \\ &= \frac{1}{N^T)} Var(Y_i^T) + \frac{1}{(N^T)^2} \sum_{j \in C} w_j^2 Var(Y_j^C) \end{aligned}$$

From the computational point of view, we can apply this formula, or, we can fit a linear model with the respective response variables of the student's performance and with the only *test* as independent variables. In the absence of covariates, the coefficient of the average effect of the treatment is an OLS estimator and it is identical to the difference of the sampling averages of Y (Neyman estimator). Therefore, the coefficient is unbiased (Imbens and Rubin, 2015). In Table 2.5 we can see the average treatment effect on student's who didn't take the test. We fit three model, each for the different performance

indicators Y_1, Y_3, Y_5.

Table 2.5: Average treatment effect on the students who didn't take the test, by using the performance indicators of the 2014/2015 (a.y.), School of Economics and Management, University of Florence.

Outcome	Coefficient	Estimate	Std. Error	Pr(>|t|)
Y_1	test	-0.289	0.085	***0.000934
Y_3	test	-0.209	0.055	***0.000183
Y_5	test	-0.081	0.034	*0.016855

With $\alpha = 5\%$ all the coefficients are statistically significant. The conclusion for the comparison by test leads to the assertion that students, who didn't take the self-evaluation test at the School of Economics and Management, in the 2014/2015 a.y., have:

- a lower probability of 29% to give at least one exam than the students who did the test
- a lower probability of 21% to give at least three exams than the students who did the test
- a lower probability of 8% to give at least five exams than the students who did the test

The test is different for each faculty and it contributes to the process of orientation towards the choice of the university course. If one student passes the test in a particular faculty, this doesn't mean that he/she has the knowledge necessary to access directly to another faculty. In fact, in the School of Economics and Management, students that don't pass the test at first attempt have to fill gaps with the materials indicated by the university. Such students must pass the next test in order to take exams during the academic year. This means that the test should always be done.

2.2 Comparison by Test Session

The goal, in this section, is to verify if the Test Session (September or Later) can influence the student's performance. The treatment variable is *Test Session*. It takes value 1 if the test is done later on computer, otherwise, 0 if it is done on paper in September.

Table 2.6: Composition of the binary treatment variable *Test Session*, in reference of the 2014/2015 (a.y.). School of Economics and Management, University of Florence

Test Session	0	1
	722	147

In the next table, we can see how many exams give students per session. On average, students who did the test in September give 2.09 exams. Students who did it later give on average 1.30. The difference between the two groups is lower than the case studied in the previous section.

Table 2.7: Number of given exams by students who did the test in September (session=0) and the students who did it later (session=1) , in reference of the 2014/2015 (a.y.). School of Economics and Management, University of Florence

	0	1	2	3	4	5	6
Test Session=0	212	133	95	98	75	61	48
Test Session=1	71	25	17	16	8	8	2

Table 2.8: Standardized difference for each variable between the active and control group, in reference of the 2014/2015 (a.y.). School of Economics and Management, University of Florence

	med.pc	med.paper	ST_diff.med
gender	0.592	0.580	0.023
residence	0.728	0.856	-0.319
late-enrolment	0.245	0.010	0.391
Humanities	0.112	0.069	0.160
Scientific	0.347	0.341	0.013
Technical	0.259	0.400	-0.305
Other	0.279	0.190	0.211
HS grade	75.94	77.74	-0.156

Table 2.8 show the results of the normalized differences. On average, in the treated group there are less resident student and fewer students that coming from the Technical High school. There is a higher presence of students from Scientific and Other High School. Furthermore, the treated group have on average a lower diploma score than the control group.

A logistic model is used to estimate the propensity score, where the dichotomous response variable is the variable that indicates which unit is assigned to the treatment, and vice versa, which one is to the control group. Explanatory variables are: *test, gender, residence, late-enrolment, Humanities, Scientific, Technical, other, High School Grade.* As in the previous sections, we use the nearest available propensity score matching, by using the exact matching on the binary covariates. From the results of table 2.9, we can see that all the treated were matched. The match works very well because all the dichotomous variables have been matched with a balance improvement of 100% and the HSG has a 96.72% balance improvement.

Table 2.9: Synthesis matching *Test Session*: Nearest available propensity score using the exact matching on binary covariates

Percent Balance Improvement	Mean Diff.
distance	99.87
gender	100
residence	100
late-enrolment	100
Humanities	100
Scientific	100
Technical	100
Other	100
HS grade	96.72

Sample sizes	Control	Treated
All	722	147
Matched	123	147
Unmatched	599	0
Discarded	0	0

In figure 2.4 We compute the distribution of the propensity score in the treaties and the controls before and after matching. In the histogram, we can observe that matching has greatly improved the balance. Figure 2.5 illustrate for each covariate the balancing before and after matching.

We proceed on calculating the treatment effect. For treaties, we use the outcome of the matched control, and vice versa. In this section, we use the ATT estimand as we are interested in the effect on those who did the self-evaluation later. We fit a linear model for each responses variables, constructed in Table 1.2, with only the independent variable *Test Session.* The results are shown in table 2.10.

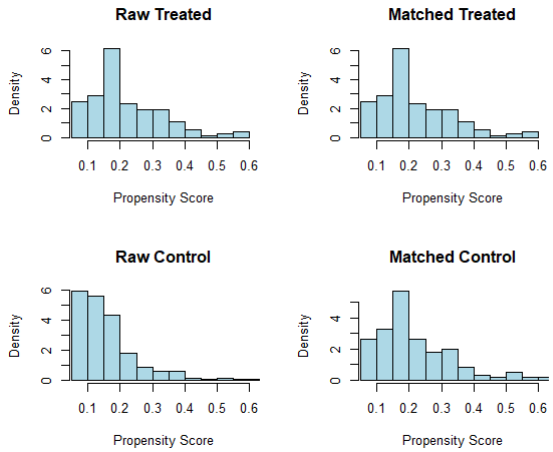

Figure 2.4: Histogram: before and after matching on treatment variable *Test Session*

Table 2.10: Average treatment effect on the students who did the test in November or March, by using the performance indicators of the 2014/2015 (a.y.), School of Economics and Management, University of Florence

Outcome	Coefficient	Estimate	Std. Error	Pr(>|t|)
Y_1	session	-0.088	0.061	0.146
Y_3	session	-0.088	0.054	0.104
Y_5	session	-0.07	0.036	0.062

With $\alpha = 5\%$ all the coefficients are not statistically significant. There are no performance differences at the end of the year between students who did the test Later or those who did it in September.

Matching techniques allowed the comparison between students with similar characteristics. The results lead to the conclusion that the test session does not affect the student's performance. The reason is not for the test period itself but is look elsewhere.

In Italy, non-enrolled subjects can participate in lessons. We suppose so that students who did the test later have been able to participate in the lessons and therefore get a preventive idea. This means that thanks to the lessons open to the public, students continued to attend the lessons and study regularly waiting for the next test. This has led to a similar performance at the end of the academic year.

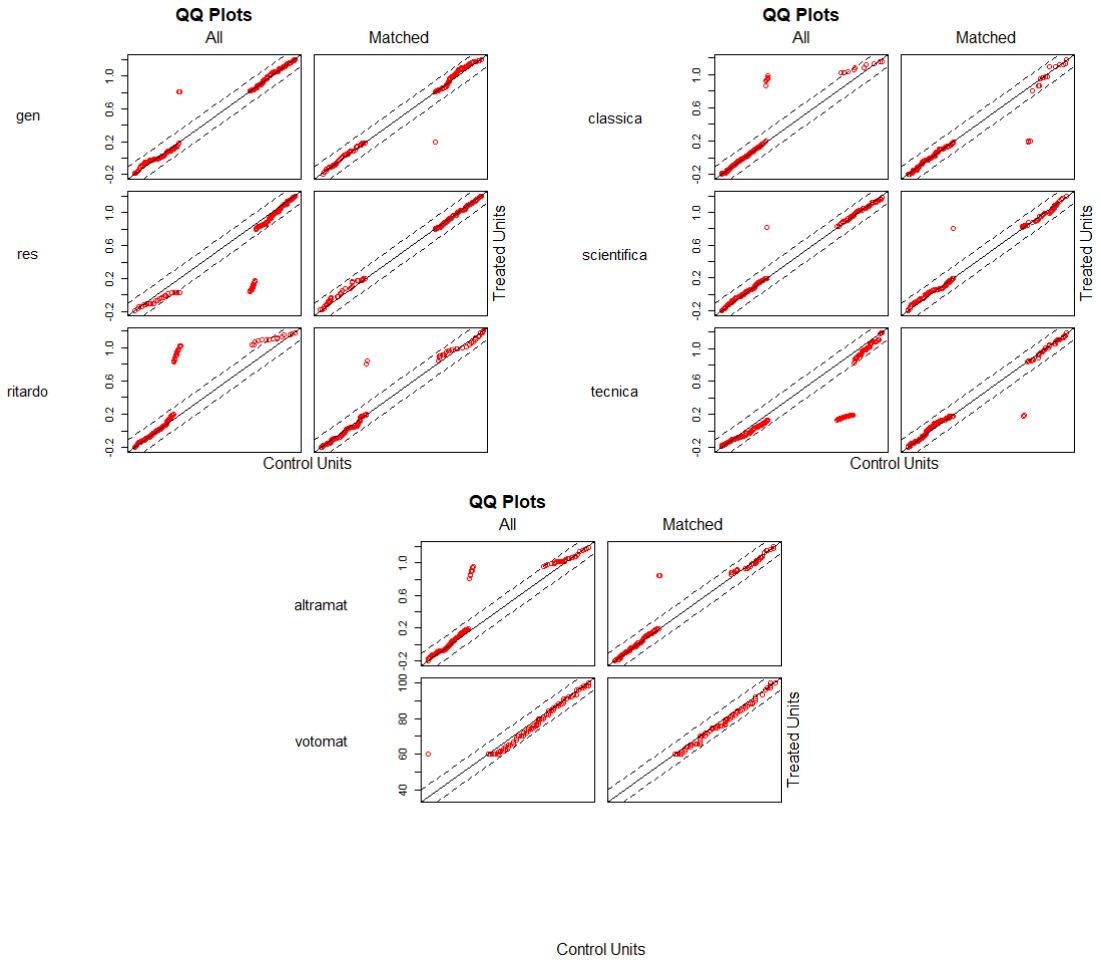

Figure 2.5: QQplot: before and after matching with propensity score estimated on the treatment variable *Type*

Chapter 3

Logistic Regression

In this chapter, we use the logistic regression model to predict the student's performance using the **pre-test** variables. We want to see also how much information can add the **test** variables on predicting the results.

We consider the three different outcomes constructed in Table 1.1 of the previous chapter. We cannot apply the linear regression model because with the binary dependent variable we obtain the *linear probability model*. The response is $y = 1$ if the student pass the exam and $y = 0$ if he doesn't. It is called linear probability because in the conditional expectations we can see how a unit change in X_p always result in the same change in the probability (Scott Long, 1997)

$$E(y_i|\boldsymbol{X}_i) = [1 \times Pr(y_i = 1|\boldsymbol{X}_i)] + [0 \times Pr(y_i = 0|\boldsymbol{X}_i)] = Pr(y_i = 1|\boldsymbol{X}_i)$$

But with the binary outcome some assumptions are violated. The first concerns the variance of the response:

$$Var(y|\boldsymbol{X}) = Pr(y = 1|\boldsymbol{X})[Pr(y = 0|\boldsymbol{X})] = \boldsymbol{X}\boldsymbol{\beta}(1 - \boldsymbol{X}\boldsymbol{\beta})$$

which implies an heteroscedastic model. So the model will have inefficient estimate of β and biased standard error. Furthermore, the model predict values of y that are greater than 1 or negative thus leading probabilistic prediction nonsensical.

To overcome these problems, we can use the generalized linear models. These models have three part components (Agresti, 2015):

- The *random component* is formed by the observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ that are independent and identically distributed, with distribution belonging to the exponential family.

- The *linear predictor*. The expression linear refers to the parameters, while explanatory variables can be non linear functions is expressed in matrix form

$$\boldsymbol{\eta} = \boldsymbol{X\beta}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$,$X_{n\times p}$ is the matrix of the explanatory variables and $\boldsymbol{\beta}$ is the parameter vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$.

- The linear predictor is related to $E(y)$ through a monotone and differentiable *link function* $[g(\cdot)]$. Let $\mu_i = E(y_i), i = 1, \ldots, n$.

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}$$

The GLM are an extension of the linear model, developed thanks to the software improvement. With these methods some fundamental hypotheses are attenuated or changed. The relation between the explanatory variables and the response through a *link* function is non-linear. The hypothesis of homoschedasticity and normality of the observations are also relaxed. In classical linear model, the response variable is assumed normal distributed while in GLM the dependent variable is a random variable whose distribution belongs to the exponential family.

In our dataset, the variables response has binomial distributions which belong to the exponential family. There are a large choice of link functions and the three most used in practise are (McCullagh and Nelder, 1989):

1. Logit or logistic function: $g_1(\pi) = log\{\pi/(1-\pi)\}$
2. Probit or inverse Normal function:$g_2(\pi) = \boldsymbol{\phi}^{-1}(\pi)$
3. Complementary log-log function: $g_3(\pi) = log\{-log(1-\pi)\}$

The first two functions are almost linearly related over the interval $0.1 \leq \pi \leq 0.9$, while complementary log-log is different for values of π close to 0 or 1. As in our dataset the observed proportion of students who give at least one exam, at least three or five is within the range of $[0.1; 0.9]$ we can exclude the third function.

$$\%(Y_1 = 1) = 75.3 \qquad \%(Y_3 = 1) = 40.6 \qquad \%(Y_5 = 1) = 15.3$$

In order to make a prediction logit and probit are equivalent. However, for the greater notoriety and simplicity of interpretation we choose to adopt the logit link.
The model parameters for logistic regression has two formulations (Agresti, 2015):

$$\pi_i = \frac{exp\big(\sum_{j=1}^{p} \beta_j x_{ij}\big)}{1 + exp\big(\sum_{j=1}^{p} \beta_j x_{ij}\big)} \qquad \text{or} \qquad logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij}$$

The first formulations ensure that the probability will not be less than 0 or greater than 1. The parameters are estimated by maximum likelihood techniques. The estimation equations of a GLM are not in closed form so the solution is found by an iterative process (Menard, 2002). The default method in R uses iteratively reweighted least squares (IWLS).

3.1 Model fitting

We work on the entire dataset with 869 students who did the self-evaluation test and enrolled in the in the 2014/2015 a.y., School of the Economics and Management, University of Florence.

The response variable of interest is the amounts credits earned in the next solar year. For this reason, the binary responses built in Table 1.1 are used as the university performance indicators. In table 3.1 we display the coefficients, which give the change in log odds of the response variables for a one unit increase in the predictor variable. There are two models for each of the three outcomes. In the first model there are only *pre-test* variables. In the second model, we also added the test variables to observe the significance of the coefficients.

In each model there aren't interactions because, with lrt test, they didn't add any information, thus obtaining a parsimonious model.

Table 3.1: Logistic regression: output for each of the three binary responses constructed in Table 1.1. First model with only *pre-test* variables and the second with *pre-test* variables and the *test-variables*. In reference of the 2014/2015 (a.y.), School of Economics and Management, University of Florence.

	Y_1	Y_1	Y_3	Y_3	Y_5	Y_5
Pre-Test						
(Intercept)	***-5.81	***-5.89	***-9.76	***-10.15	***-12.77	***-13.07
gender	0.06	0.01	*0.39	0.21	**0.67	0.50
residence	**0.60	*0.54	***0.85	**0.73	*0.77	0.58
late-enrolment	**-0.65	**-0.62	**-1.01	**-0.95	-0.76	-0.63
Humanities	***1.34	***1.32	**1.09	**0.95	*1.13	0.95
Scientific	***1.09	***0.97	***1.19	***0.87	***1.66	**1.20
Technical	0.29	0.31	0.21	0.23	0.17	0.15
HS grade	***0.07	***0.07	***0.10	***0.09	***0.11	***0.10
Test						
Logic		-0.01		*0.11		0.05
Reading		0.03		0.04		0.02
Mathematics		*0.10		***0.18		***0.27

The asterisks indicate the significance (0'***';0.001'**',0.01'*')

The intercept is the estimated baseline log odds when all independent variables are set to 0, or the reference category in case of categorical variables. The reference categories are *female, students far-away, students with regular studies at the high school, Other high school*.

Scientific High school, the High school grade and the Mathematics score obtained in the self-evaluation test are always statistically significant. As the number of exams increases, in these coefficients increases the change in log odds. For a unit increase in Mathematics test score the log odds of Y_1 increase by 0.10, for Y_3 increase by 0.18 and for Y_5, increase by 0.27. The Logic test score is significant only for Y_3 while Reading test score is always meaningless.

Students with an irregular career at the high school have significant negative influence on Y_1 and Y_3 respect students with a regular career. The resident students have always a positive influence on the far-away but result not significant only for Y_5 with test scores. Students from Humanities High school have always a positive effect respect the students for Other High School, but as the number of examinations increases, the coefficients lose significance until they result meaningless for Y_5 with test score variables. The effect of males on the outcomes, respect the females, increasing as increase the number of exams. For Y_1 we can see a very small effect, while in the models with only pre-test variables, the coefficient results statistically significant for Y_3 and Y_5. When we add test scores variables, coefficients lose significance.

3.2 ROC Curve

In our dataset we are in the presence of *ungrouped data*[1] where each observation y_i, results from a single Bernoulli trial. We can cross-classifies the binary response y with the binary prediction $\hat{y}$ in a *classification table*. There are four possible outcomes:

- *True Positive (TP)* if $y = 1$ is classified as $\hat{y} = 1$
- *False Negative (FN)* if $y = 1$ is classified as $\hat{y} = 0$
- *True Negative (TN)* if $y = 0$ is classified as $\hat{y} = 0$
- *False Positive (FP)* if $y = 0$ is classified as $\hat{y} = 1$

The estimates of a **true positive** rate (or sensitivity) of a classifier is:

$$\text{tp rate} = \frac{\texttt{Positives correctly classified}}{\texttt{Total positives}} = P(\hat{y} = 1 | y = 1)$$

[1]There is also a second form of binary data: *grouped data*. Here there is a set of observations which have the same value of the explanatory variables.

The estimates of a **false positive rate** (or specificity) of a classifier is:

$$\text{fp rate} = \frac{\texttt{Negatives incorrectly classified}}{\texttt{Total negatives}} = P(\hat{y} = 1|y = 0)$$

We use the ROC graphs to see which the best trade-off between the two rates is. On the Y axis is plotted *tp rate* while on the X axis is plotted the *fp rates* . The point $(0, 1)$ represents the perfect classification. The diagonal line, $y = x$ represent a no realistic classifier because it guesses to get half of the positives and half of the negatives. If the classifier is below the diagonal, it performance worse (Fawcett, 2006).

We need to select the cut-off π_0 because the classification table depends on it for the prediction. When π_0 is near 1 all the point are near$(0, 0)$ because predictions $\hat{y}_i = 0$, otherwise, when π_0 is near 0 all the point are near$(1, 1)$ because predictions $\hat{y}_i = 1$ (Agresti, 2015). We need to select π_0 which is the better at identifying likely positives than at identifying likely negatives. So the better predictive power is given where there is the greater area under the ROC curve (AUC) (Fawcett, 2006).

We fit the model on the entire data set, with pre-test and test variables for each binary responses referred to table 1.2. With *pROC* package we print the best cut-off. In table 3.2 we can see how the threshold decreases as the proportion of zeros contained in the response variable increases.

Table 3.2: Best cutoff estimated with the Roc curve. For the two models with the respective three binary responses constructed in Table 1.2.

Model	**Cutoff**
Y_1 (pre-test)	0.738
Y_1 (pre-test and test)	0.765
Y_3 (pre-test)	0.372
Y_3 (pre-test and test)	0.383
Y_5 (pre-test)	0.141
Y_5 (pre-test and test)	0.117

In figure 3.1 we can observe that the Roc curve moves away from the diagonal to the top of the quadrant as the number of exams increases. For Y_1 the ROC curve have a poor predictive power and the area under the ROC curve is equalled for both models. While, for Y_3 we can observe how the area increases slightly by the inclusion into the model of the test score variables, with a decrease in False Positive rate and an increase in True Positive rate. Finally, for Y_5 we can see an improvement in the area thanks to the insertion of the test variables, but, the FPr increases and decreases the TPr compared to the Y_5 model with only pre-test variables.

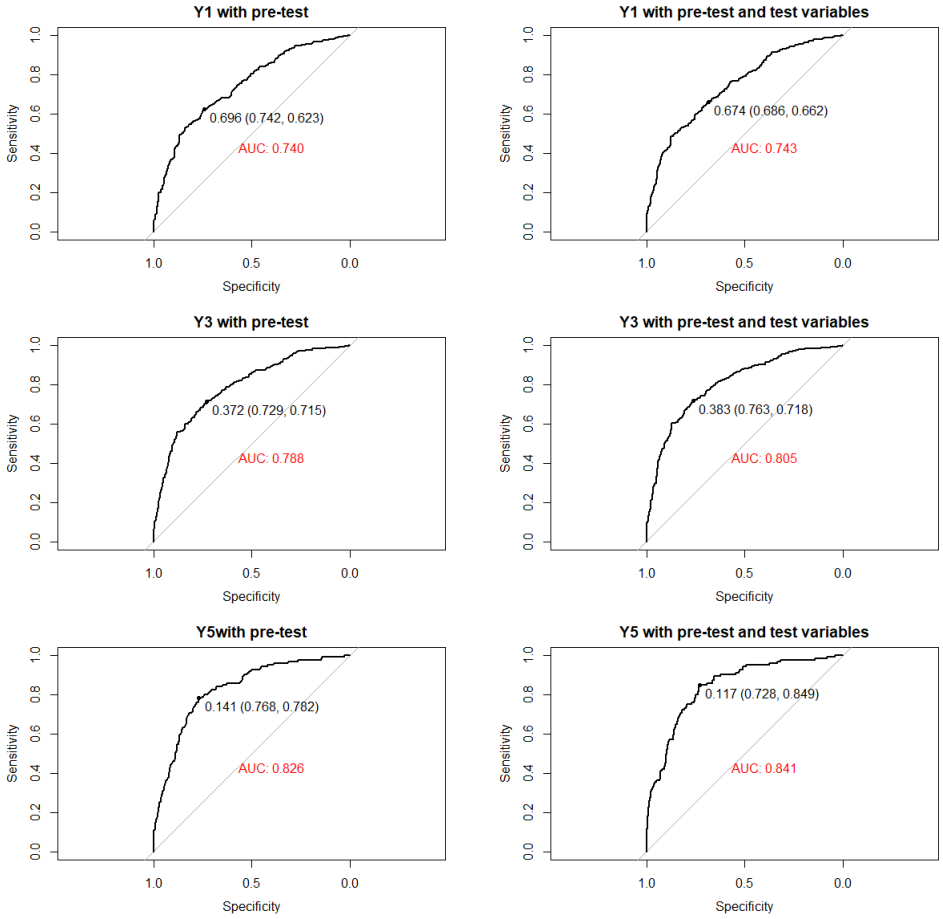

Figure 3.1: Roc curves for each of the three binary GLM, in reference to the predictions on the data used to fit the model with background and test variables. School of Economics and Management, University of Florence, 2014/2015 (a.y.).

3.3 Results

Instead of predicting from the model fitted on the entire dataset, we prefer using k-fold cross-validation. The learning method is to divide the data set in k equal-sized parts and the model is fitted in the $K-1$ parts. We predict the kth part that we left out as the test set and we calculate the prediction error (Hastie et al., 2009). For each the predicted kth we choose the best cut-off between sensitivity and specificity, $\hat{\pi}_0$. The prediction for the response variables y_i is $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi_0$, otherwise $\hat{y}_i = 0$ if $\hat{\pi}_i \leq \pi_0$.

Since we are in a classification problem, the prediction error is calculated as follows:

$$PE = 1 - ACCURACY = 1 - \left(\frac{TP + TN}{TP + FN + TN + FP} \right)$$

This process is done for each k=10 parts. We take the average of the 10 PE values and we use it to evaluate which model produces the lowest prediction error. Another goal

is to verify if the addition of test variables (score on Logic, Reading and Mathematics) leads to a decrease in prediction error.

Table 3.3: Average of the 10 prediction error for the model with only *pre-test* variables and for the model with in addition the *test* variables. The three binary responses are used for each model. School of Economics and Management, University of Florence, 2014/2015(a.y)

	Y_1	Y_3	Y_5
PE (pre_{test})	0.312	0.229	0.239
PE($pre_{test} + test$)	0.310	0.219	0.229

In table 3.3 we can see the average of the prediction error computed on each of the one-tenth parts of the data. In the first model, we used only the *pre-test* variables and we fitted it for each of the binary responses. In the second model in addition to the pre-test variables, we also added test variables, score in Reading, Logic and Mathematics. The lowest average of the estimated prediction error is always for Y_3. Students who give at least three exams are most likely to be predicted. Test variables add little information in general. We can observe how the earning in terms of forecast error for Y_1 is really small. This is due to the difficulty of predicting students who are able to give at least one exam based on the covariates. The addition of information about the test score gets an improvement of 1% prediction error for students who give at least 3 exams. For predicting students who give at least 5 exams, the improvement in forecast error is always 1%. However, with the test variables, the predicted error is 0.229 compared to the 0.219 of the Y_3. This means that it is more difficult to predict students who give at least five exams because there are few students in this category, about 13.7%, on the date set.

The conclusion is that the addition of test scores information doesn't help to predict the student's performance. The reason behind the analysis made in this chapter is that knowing the student's score in mathematics is partly due to background variables, such as the origin of the Scientific High school or the High School grade. So the test variables give similar information already known before the students give the test. It would be necessary to modify the form of the test as the Reading portion is not significant while Logic is only important for Y_3. The part of mathematics depends on the High school of origin. The test lacks the motivational part, which is fundamental to continuing the studies. Many students choose the School of Economic and Management in the absence of other opportunities. But this does not lead to great interest in the study.

Chapter 4

Random Forest

Ensembles methods are learning machines that allow improving the predictive performance, instead of the result that would have been achieved with respect to unique learning algorithms. Recently, the interest in the research area is motivated by technological development, that allows fast implementations. Resampling methods are used to generate different hypotheses, for examples bootstrap techniques is used to generate different training sets and the learners algorithm are applied in the subset of the data. These techniques are useful because if we use the only algorithm, like decision trees, it product unstable results as they are sensitive to small changes in the training set (Valentini and Masulli, 2002).

In this chapter, we use the random forest classifier to predict the student performance using the **pre-test** variables. As in the previous chapter, we want to see how much information can add the **test** variables on predicting the student's performance, in addition to the **pre-test** variables.

4.1 Theoretical background

A *Classification Tree* is used to predict a qualitative response. Different variables subsets are used at different tree levels instead of using all the variables contained in the data set jointly to create a decision rule. Recursive binary splitting is used to grow a classification tree. The approach begins on the top of the tree with all the predictors and then splits the predictor's space. Different measures are used for making the binary splits:

- *Classification error rate* represents the proportion of training observations, $\hat{p}_{mk}$, in the mth region that are from the kth class.

$$E = 1 - max_k(\hat{p}_{mk})$$

- *Gini Index* measures the total variance across the K classes.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- *Cross-entropy*

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$$

Gini index and the cross-entropy are used more for evaluating the quality of the split. The Classification error rate is used when the goal is the accuracy of the prediction. According to the chosen split criterion, the N units are progressively subdivided into a series of disjointed subgroups that have a degree of homogeneity greater than the initial set. At each step of the process, the heterogeneity of the groups is reduced compared to the previous step. The terminal nodes of the tree have a degree of homogeneity such as it can be attributed to one of the modes of the Y response variable. In this way, we have a classification rule that allows us to classify test set observations and to calculate the classification error rate. Trees are easy to explain to the people and they can be displayed graphically but despite these advantages, they suffer from high variance, because the fit of each tree are very different and there isn't the same level of accuracy (James et al., 2013).

Bagging is a statistical learning method that aggregates many trees for improving the predictive performance and for reducing the variance. From the learning set,$L = (y_n, \boldsymbol{x}_n)$ where $n = (1, \ldots, N)$, bagging takes repeated bootstrap sample $L_{B,1}, \ldots, L_{B,K}$ and construct classifiers, $h(\boldsymbol{x}, L_k)$, that vote to form the bagged predictor, $\{\varphi(\boldsymbol{x}, L^{(B)})\}$. The kth predictor $\varphi(\boldsymbol{x}, L_{k,B})$ is based on the kth bootstrap learning set, thus defining K predictors. The aim is to use $\{L^{(B)}\}$ to get a better predictor that one single predictor. The proportion of times that the estimated class differs from the true class is the bagging misclassification rate (Breiman, 1996a). The Out-of-bag estimation is used to estimate the misclassification rate. Each bootstrap sample leaves out about 37% of the observations. The ith observation response is predicted by using each of the trees in which that observation resulted in OOB. To obtain a single OOB prediction we obtain a majority vote for a single prediction (Breiman, 1996b).

The disadvantage of bagging is the high correlation between the trees, since, the use of all explanatory variables involves splitting the tree at the beginning with the most influential variables, thus making similar trees (James et al., 2013).

One way to overcome this disadvantage is by using Random Forest. The difference is in the selection of the number of explanatory variables because the process begins with the random selection of m variable, $m < p$ where p is the total number of explanatory

variables (in the bagging m = p). Each new bootstrap sampling is done with replacement from the original data set and the tree growth are not pruned.

One of the main advantages concerns the accuracy that is good as AdaBoost, if not better. It is faster than bagging and boosting and it gives useful estimates of error, the strength of the classifier, correlation and variable importance. But it is relatively robust to outliers and noise (Breiman, 2001).

Classification Algorithm for Random Forest

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\boldsymbol{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\left\{T_b\right\}_1^B$.

To make a prediction at a new point x:
Let $\hat{C}_b(x)$ be the class prediction of the bth random forest tree.
Then $\hat{C}_{rf}(x) = mayority\ vote\ \left\{\hat{C}_b(x)\right\}_1^B$

(Hastie et al., 2009)

From the strong law of the large numbers and the tree structure, the random forest do not overfit as more trees are added, but produce a limiting value of the generalization error (Breiman). It can be shown through the definition of the margin function in a training set drawn at random from the distribution of the random vector $Y, \boldsymbol{X}$:

$$mg(\boldsymbol{X}, \boldsymbol{Y}) = av_k I(h_k(\boldsymbol{X}) = Y) - max_{(j \neq Y)} av_k I(h_k(\boldsymbol{X} = j) \tag{4.1}$$

where $I(.)$ is the indicator function, $h_1(\boldsymbol{x}), \ldots, h_k(\boldsymbol{x})$ are en ensemble of classifiers. The margin measure how many times the average number of votes in the right class, $\boldsymbol{X}, Y$, exceeds the average of each other's classes. When the margin increase, increase also the

credibility of the classification. So the error is given by:

$$PE^* = P_{\boldsymbol{X},Y}(mg(\boldsymbol{X}, Y) \leq 0) \tag{4.2}$$

Since in random forest $h_k(\boldsymbol{X} = h(\boldsymbol{X}, L_k)$, it follows that for all the sequences of $L_1, \ldots, L_k$, the PE^* converges to:

$$P_{(\boldsymbol{X},Y)}(P_L(h(\boldsymbol{X}, L) = Y) - max_{(j \neq Y)} P_L(h((\boldsymbol{X}, \boldsymbol{L}) = j) \leq 0 \tag{4.3}$$

The strength of the individual classifiers in the forest, and the correlation between them in terms of the raw margin functions, are the two elements that affects the generalization error. The strength is understood as the expected values of the margin function of a random forest. When the number of random variables used in each tree decrease, it reduces both the correlation and the strength. So an optimal value is found by using the out of bag error rate in a range of different values of the number of variables, and the best number is one that minimizes the error. As the random forest are an extension of bagging, we can use the out-of-bag method for estimate the generalization error, (PE^*) (Breiman, 2001).

4.2 Fitted classifier

We use the *randomForest* R package for the prediction of the student's performance. As in Chapter 3, we use one model with only *pre-test* variables without iteration so to allow the comparison of the results between the random forest and logistic regression. In the second model, the test variables are added to the pre-test variables for assessing the possible addition of information. We apply the random forest classifier on the 869 students who did the self-evaluation test and enrolled in the 2014/2015 academic year, School of Economics and Management, University of Florence. The student's performances are shown by the indicators constructed in Table 1.1 which represents a number of credits gained at the end of the next solar year. We have two models to evaluate for three binary responses thus obtaining six models to estimate.

The first parameter of the random forest is *ntree*, which represent the number of trees to grow. In our models we grow 500 trees, default number in random forest implementation in R. The second parameter is the number of variables randomly sampled as candidates at each split, *mtry*. The default values for classification is $\sqrt{p}$ where p is number of variables in $\boldsymbol{X}$.

4.2.1 Imbalanced Data

Decision trees are sensitive to imbalanced classes. As Random forests are built on decision trees, the classifier focuses on the maximal accuracy prediction of the majority class. We may obtain an optimal accuracy but, in case of fewer positive class in the data, we will have all instances negative predicted. This issue is present in many real situations on real-world applications and, for this reason, it has caught the attention of many researchers (**?**). In general, sampling-based approach and cost-based approach are used to overcome this problem. The basic idea of sampling methods is to simply adjust the proportion of the classes in order to increase the weight of the minority class observations within the model. It can be done with *under-sampling*, which randomly eliminate chosen cases of the majority class to decrease their effect on the classifier. All cases of the minority class are kept. Through *over-sampling* in the minority class, all existing observations are taken and copied. Extra observations are then added by randomly sampling with replacement from this class. These methods can, therefore, be combined with every appropriate classifier. In ensembles method, different techniques are used in the literature. One extension of the over-sampling is *Over-Bagging*. It consists in the combination of sampling with the bagging approach.

In this study, the random forest learner is fused with the over-sampling bagging for imbalance correction. We use the *mlr* packages as it allows to integrate learners with new functionality. We use 10-fold cross-validation. For each $K-1$ fold, we grow 500 trees on the 500 bootstrap sample. For each bootstrap sample, minority class observations are oversampled with a given rate. We decide to triple the smaller classes. The majority class cases are bootstrapped with replacement to increase variability between training data sets during iterations. Then, in each tree the split consider a random sampled of m predictors from the full set of p predictors. We use the kth fold as a test set in order to predict the classes which had the majority vote. For Y_1, Y_3 the average prediction error in 10-fold cross-validation is very high. For Y_5 with only pre-test variables the PE mean is 0.162, while, with the addition of test scores variables the PE mean is 0.154.

Many factors, such as the choice of the rate for over-sampling without criterion and the resampling of the same data, make this method, in my opinion, not reassuring.

An idea to improve the balance is by choosing the threshold. We change the default values of *cut-off*. In the random forest, the *cut-off* is a vector of length equal to a number of classes. The "winning" class for an observation is the one with the maximum ratio of the proportion of votes to cut-off. The default is $1/k$ where k is the number of classes. In our study, the indicators of student's performance are binary, which involve two classes. For the predicted classes the majority vote is applied with $cutoff = (k, 1-k) = (0.5, 0.5)$.

We choose the cut-off values in order to keep in consideration the proportion of

students which pass at least one exam, at least three and five exams, in reference with table 1.2. Each pair of cut-off represents a possible threshold for the classification.

The idea is:

- If the proportion of $y = 1$ is $\geq 50\%$ then the worst mistake we can make is the False Negative ratio. We search for the threshold that has a minimum of FN. If there are different cut-off which presents the same amount of FN we choose the thresholds that have the highest accuracy value.
- If the proportion of $y = 1$ is $< 50\%$ then the worst mistake we can make is the False Positive ratio. We search for the threshold that has a minimum of FP. If there are different cut-off which presents the same amount of FP we choose the thresholds that have the highest accuracy value.

The procedure begins with the selection of the first pair of cut-off. We define the resampling with the 10-fold cross-validation and we specify the classification random forest learner with the predict.type set to predict probabilities, which give the matrix of class probabilities for the kth fold used as a test. For each kth fold we calculate the relative index across the 10-fold cv, for the respective cut-off.

Table 4.1: Improve of the probability thresholds for class imbalances by selection of the best cut-off among the 10-fold cross validation. The procedure is done for each variables response for the model with *pre-test* variables and for the model with *pre-test* and *test* variables.

Model	Best cut-off
Y_1 with pre-test variables	(.25,.75)
Y_1 with pre-test and test variables	(.35,.65)
Y_3 with pre-test variables	(.65,.35)
Y_3 with pre-test and test variables	(.55,.45)
Y_5 with pre-test variables	(.90,.10)
Y_5 with pre-test and test variables	(.70,.30)

From table 1.2 the proportion of $Y_1 = 1$ is 67.4%. We generate 9 cutoff vectors in order to form a 9×2 matrix. The cut-off is a sequence from 0.1 to 0.50 for the first column, by an increment of 0.05. For the second column, we made a sequence from 0.90 to 0.50 with an increment of 0.05. From table 1.2 the proportion of $Y_3 = 1$ is 36.4% and for $Y_5 = 1$ is 13.7%. We generate 9 cut-off vectors in order to form a 9×2 matrix. The cut-off is a sequence from 0.50 to 0.90 for the first column, by an increment of 0.05. For the second column, we made a sequence from 0.50 to 0.10 with an increment of 0.05.

We make this process for each dataset used to estimate the respectively six models. We can see how close are the best cut-off chosen for the random forest with the cut-off

estimated in ROC curve, from table 3.2. Since In logistic regression we draw the ROC curve among the observed response variables and the predicted probabilities, the excess of the threshold indicates the probability of assigning $\hat{y} = 1$, in the random forest we consider the majority vote for the class. The random forest cut-off can be compared if we observe the second element of the cut-off vector chosen. The difference concern the prediction since for the random forest we use the same cut-off in the 10-fold cross validation, for each model. For the logistic regression, we chose the best cut-off between sensitivity and specificity, $\hat{\pi}_0$. each the predicted kth.

For every model, we use 500 trees,$mtry = \sqrt{p}$ and the respective cut-off. In table 4.2 we can see the estimation of the False positive and False negative index from the confusion matrix. The random forest confusion matrix was constructed with the respective cut-off in Table 4.1. The confusion matrix, for models estimated with logistic regression, was constructed with the ROC curve. The big difference concern the FN for Y_1 and for Y_5. Since Y_3 represent the most balanced class, the index is similar to the two methods.

Table 4.2: Comparison for classifier fitted on entire dataset: False Positive rate and False Negative rate between random forest and logistic regression. The output for the two models with the respective three binary responses constructed in Table 1.2.In reference of the 2014/2015 (a.y.), School of Economics and Management, University of Florence

	FP_{rf}	FN_{rf}	FP_L	FN_L
Y_1 (pre-test)	0.512	0.200	0.180	0.452
Y_1 (pre-test and test)	0.403	0.326	0.124	0.517
Y_3 (pre-test)	0.150	0.452	0.271	0.285
Y_3(pre-test and test)	0.165	0.415	0.237	0.282
Y_5 (pre-test)	0.093	0.571	0.232	0.218
Y_5 (pre-test and test)	0.084	0.605	0.272	0.151

4.2.2 Variable importance

We use the Variables Importance measures for interpreting the results as we have a large number of trees. The *Mean Decrease Accuracy* is based on a permutation of the variables in the OOB data. For each tree, the prediction error on the out-of-bag portion of the data is recorded. After the permutation of the variables is done the same process of recording. The difference between the two is then averaged over all trees, and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done. So we can see how worse the model performs without each variable, so a high decrease in accuracy would be expected for very predictive variables. The second measure is the total decrease in node impurities, which is measured by Gini index, from splitting on the variable, averaged over all trees.

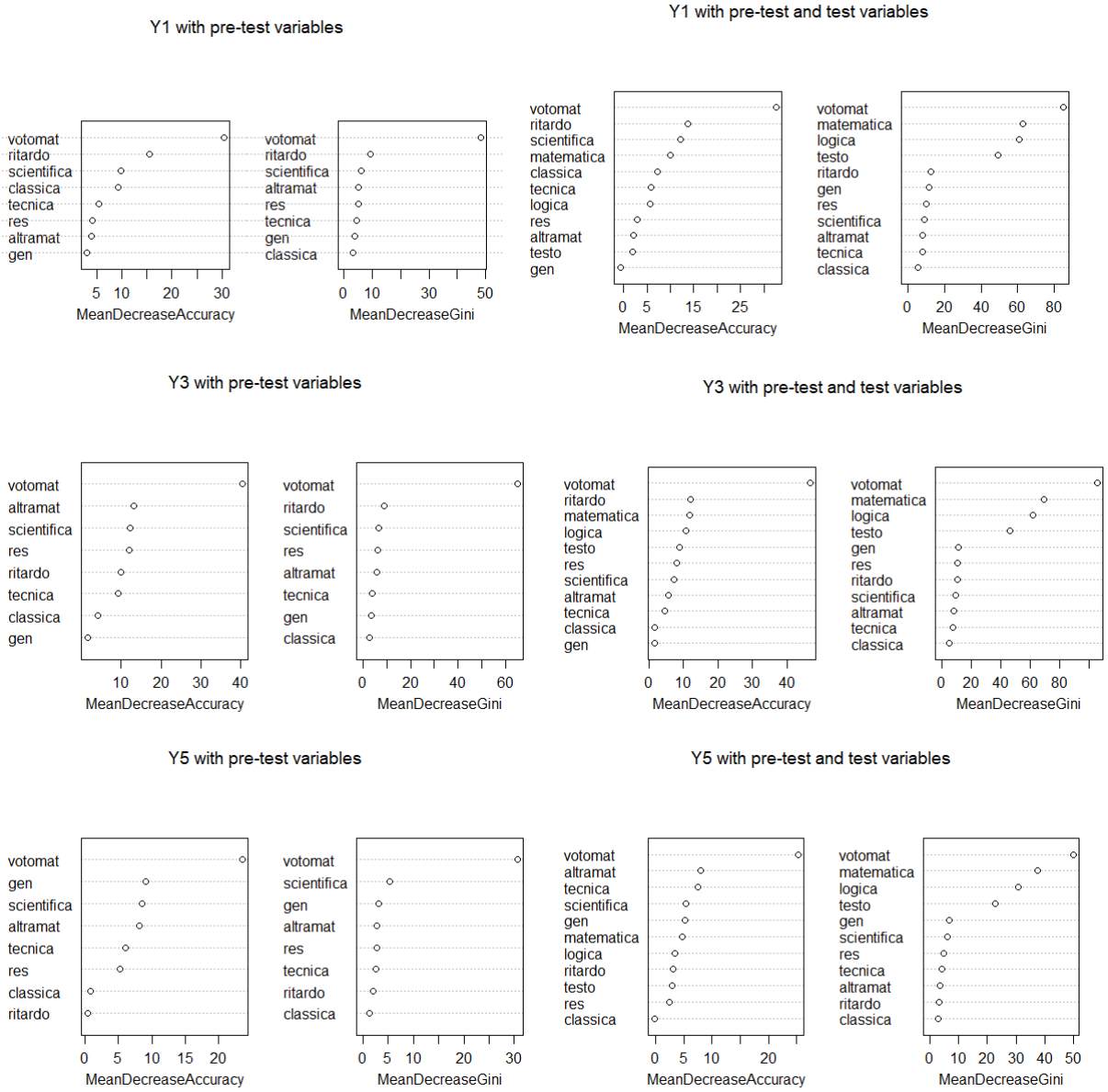

Figure 4.1: Plot for variable importance measures. The most important variables on predicting number of exams passed is High School grade. Follows test-score variables.

In all the six models the most important variables is High School grade (votomat). Gini index show that the test variables have an important role in growing the trees, especially for Y_3 and Y_5. When we fit the model with only the pre-test, the gini index give them low values, while with the decrease Accuracy the important variables have different role for the different response binary in reference.

4.2.3 Error rate across the 500 decision tree

In the next plot, we can see the error rate across the 500 decision tree, for the different classes (coloured) and out-of-bag samples (black) over the amount of tree.

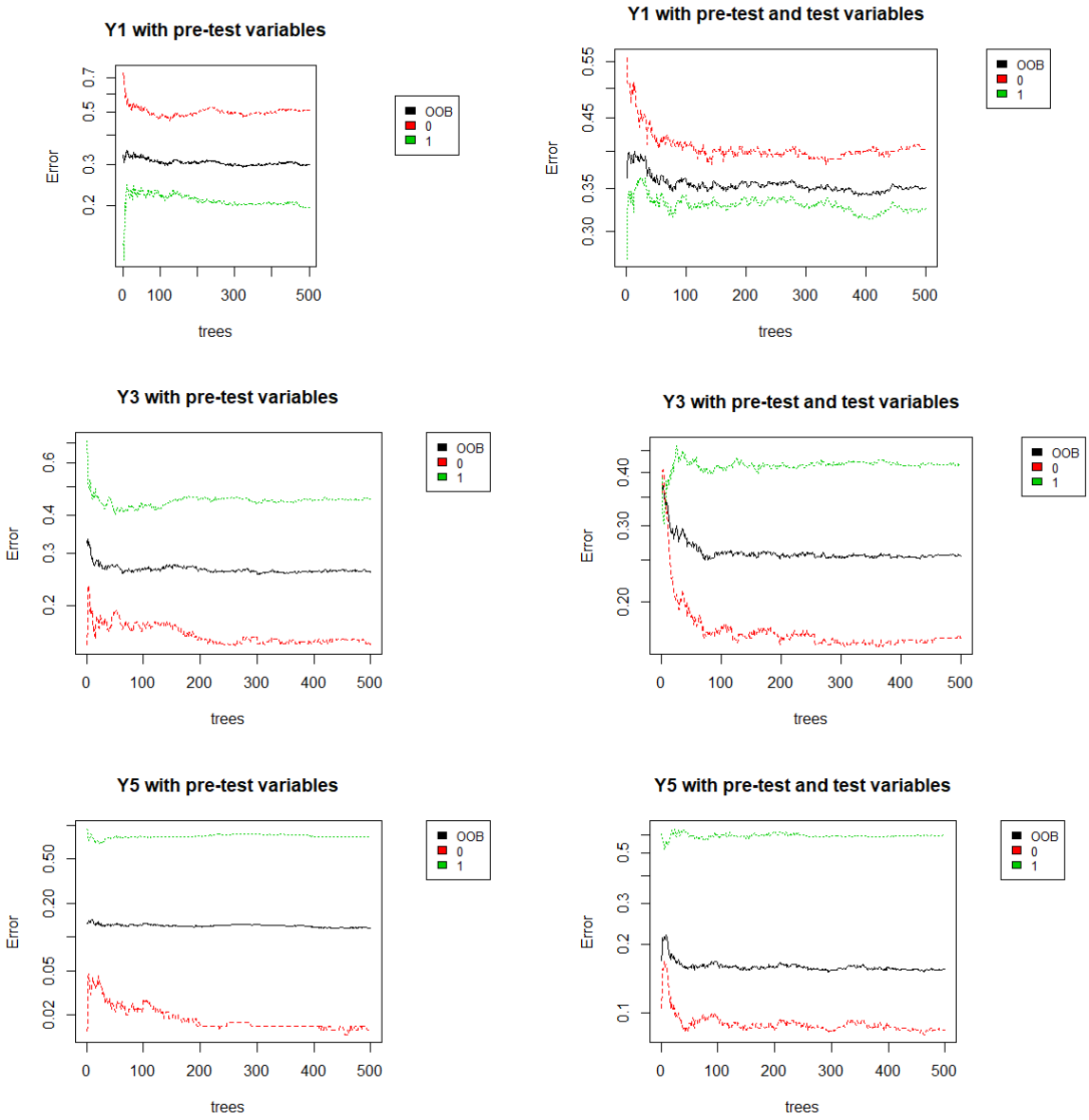

Figure 4.2: Error rate across the 500 decision tree

Despite the first cut-choice criterion has focused on decreasing false negatives, we can observe for Y_1 that the error rate for $y_1 = 0$ is much higher than $y_1 = 1$. Instead, the situation is reversed for Y_3 and Y_5, where $y = 1$ is most difficult to assign. Furthermore, the inclusion of test variables decreases the error level for Y_1 and Y_3. The plot seems to indicate that after 100 decision trees for Y_1, there is not a significant reduction in error

rate, while, Y_3 and Y_5 needs between 200 and 400 decision trees.

4.2.4 Margin function

We plot the margin histogram of predictions from a random forest classifier. The margin of a data point is defined as the proportion of votes for the correct class minus maximum proportion of votes for the other classes. Thus under majority votes, positive margin means correct classification and vice versa.

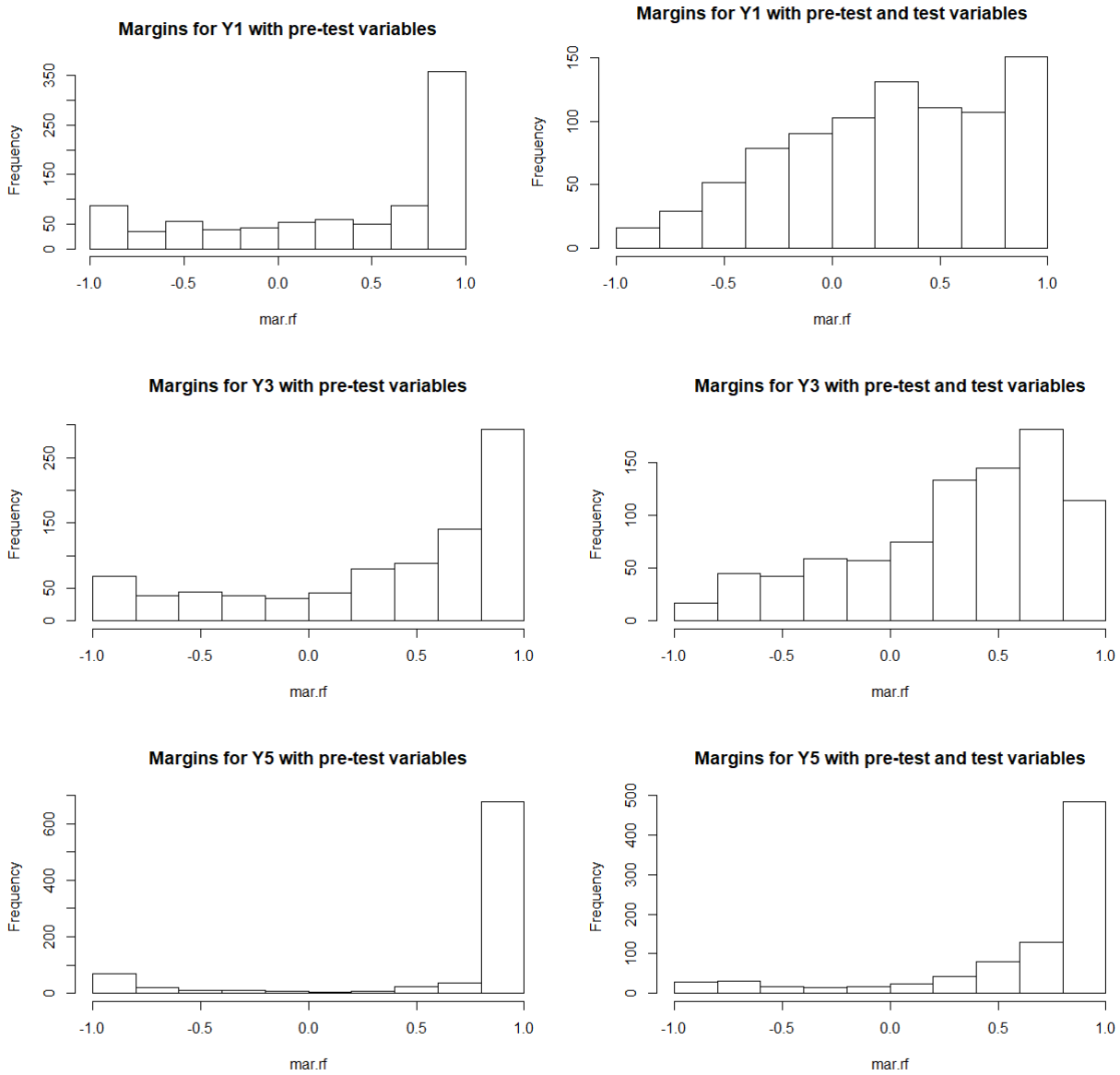

Figure 4.3: Histogram of Margin function for each model fitted with random forest classifier.

Values greater than 0 correspond to correct prediction. The model with test-score variables are right-skewed, thus indicating that the majority of observations was correctly classified. The worst concern about Y_1 with only the pre-test variables.

4.3 Results

We can compare random forests predictions with logistic regression by doing ten repetitions of 10-fold cross-validation, using the *errorest* functions in the *ipred* package . In the implementation we include each cut-off defined in 4.1 as we can include additional parameter to model. In the first model we used only *pre-test* variables. In the second model we included also the *variables*

Table 4.3: Average of the 10 prediction error for the random forest classifier, with only *pre-test* variables and with in addition the *test* variables. The three binary responses are used for each model. Random forest classier is implemented with default number of trees, default mtry and the respective cutoff in Table 4.1. School of Economics and Management, University of Florence, 2014/2015(a.y)

	Y_1	Y_3	Y_5
PE (pre_{test})	0.300	0.265	0.1791
PE($pre_{test} + test$)	0.349	0.257	0.1611

The lowest average in prediction error is always for Y_5 which means that students who give at least five exams are most likely to be predicted. We get a 2% improvement by adding test variables, compared to the logistic regression for which we had a 1% improvement. This means that the criteria used for defining the thresholds were good for handling the class imbalance.
The addition of information about the test score gets an improvement of 1% prediction error for students who give at least 3 exams. Another difference with logistic regression concerns Y_1. Here we can observe how the inclusion of test variables get worse the prediction error.

Thanks to the variable importance we saw how adding test variables are considered important for tree growth and with the margin function, we observed as the majority of observations were correctly classified for all performance indicators. We've found that they improve the prediction error but the improvement rate is really low considering the context. We don't consider to be satisfactory the gain of 2% in predicting the student's performance. We, therefore, confirm what we said in the conclusions of logistic regression analysis. The score in the self-evaluation test depend on the pre-test variables, so it gives similar information.

Conclusions

This study was focused on the self-evaluation test of the School of Economics and Management, University of Florence, academic year 2014/2015. The aim was to verify the capacity of the self-evaluation test to predict student's performance at the end of the first year. Two problems were assessed. The first concerns differences in performance between students who had to take the test and students who were exempted. In the second problem, we evaluated the predictive ability of the test for students who took the test.

The first performance comparison was made between students who were exempted from the test and those who had to do it. We used the nearest available propensity score matching with exact matching on the binary covariates, with replacement, in order to obtain two groups similar in terms of covariates. On average, the student's who didn't take the test have a lower probability of 29% to give at least one exam as compared to students who did the test. Furthermore, they have a lower probability of 21% and 8% to give respectively three and five exams. The conclusion of this part of the analysis is that exempted students from the test have worse performances than those who did it. Performance does not depend on the test itself. The difference is due to the fact that the exempted students are different from the others.

Next, we investigated if different test sessions lead to different performance. We compared students who made the self-evaluation test in September and those who made it in the second session (November or March). We used the same techniques of nearest available propensity score matching with exact matching on the binary covariates. The results of the analysis tell us that there aren't performance differences at the end of the year between students who did the test in September or those who did it later. Since all people can attend the lessons, this has allowed non-enrolled students to know better the subject and wait for the next test.

In the second part we evaluated the predictive ability. First we considered logistic regression. We used two sets of explanatory variables. In the first there are only pre-test variables while in the second, test score variables were included in order to see how much

information they can add for predicting the results. The predictive ability was assessed using 10-fold cross-validation (CV). For each predicted kth part, we have chosen the best cut-off between sensitivity and specificity using the ROC curve. The largest average of the prediction error (0.31) regards students who pass at least one exams, for both sets of explanatory variables. For students who give at least five exams, the average of the prediction errors(PE) with only pre-test variables was 0.239 while if we add the test score variables, the average of PE improves of 1%, 0.229. We obtain the lowest average of PE with only pre-test variables on predicting students who passed at least three exams, with 0.229. In the second set of variables, the PE improves to 0.219 for the prediction of students who passed at least three exams. We got a 1% improvement in the predicted error by inserting test score variables, but it doesn't mean that the self-evaluation test helps to predict students performances. The test score gives similar information already known before that student gives the self-evaluation test.

A random forest classifier was implemented for the prediction of the student's performance. As in logistic regression, we used one model with only pre-test variables and a second model where test variables were included. For each model, the number of trees and the number of variables randomly sampled as candidates at each split are set with default values, respectively 500 and square root of the total number of explanatory variables. The cut-off was chosen to minimize the error rates based on 10-fold CV. In variable importance plots, the most important variable is High School grade. Gini index shows that the test variables have an important role in growing the trees, especially for the indicators the indicator of passing at least 3 and 5 exams. With margin function we saw that the model with test-score variables are right-skewed, thus indicating that the majority of observations was correctly classified. The worst concern about the indicator of passing at least one exam with only the pre-test variables. We can compare random forests predictions with logistic regression by doing ten repetitions of 10-fold cross-validation. The lowest average in prediction error is always for students who passed at least 5 exams, 0.179 with only pre-test variables and 0.161 for the model which include the test score variables. We got a 2% improvement by adding test variables improvement. The addition of information about the test score gets an improvement of 1% prediction error for students who give at least 3 exams. For students who gives at least one exam, the inclusion of test variables gets worse the prediction error.

In reference to the methodology used, we observe that Variable Importance plot in random forest allows the comparison between continuous and binary covariates, in the same scale of measurement. We can see which variables are more important in predicting the outcomes. In logistic regression, we can only observe which independent variables are statistically significant but we cannot make a comparison between the two different

types of covariates.

Besides, independent variables no statistically significant added in the logistic regression model do not compromise error prediction. In the random forest, the inclusion of new variables can improve or worsen the prediction error as, in each tree, the split considers a random sample of m predictors from the full set of predictors. Random forest is a powerful method for prediction but it may have unsatisfactory performance for unbalanced outcomes. Logistic regression is a standard method that can handle this issue through easier computational techniques. In this study, the logistic standard method has similar performance over the more complex random forest method.

In conclusion, the addition of test scores variables yield a modest gain in the prediction ability, in the range of $1\% - 2\%$. Thus, the information provided by the pre-enrolment test is largely redundant.

Bibliography

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *econometrica 74*(1), 235–267.

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Becker, S. O., A. Ichino, et al. (2002). Estimation of average treatment effects based on propensity scores. *The stata journal 2*(4), 358–377.

Breiman, L. (1996a). Bagging predictors. *Machine learning 24*(2), 123–140.

Breiman, L. (1996b). Out-of-bag estimation.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys 22*(1), 31–72.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters 27*(8), 861–874.

Grilli, L., C. Rampichini, and R. Varriale (2016). Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts. *Statistical Modelling 16*(1), 47–66.

Guo, L., Y. Ma, B. Cukic, and H. Singh (2004). Robust prediction of fault-proneness by random forests. In *Software Reliability Engineering, 2004. ISSRE 2004. 15th International Symposium on*, pp. 417–428. IEEE.

Hastie, T., R. Tibshirani, and J. Friedman (2009). Overview of supervised learning. In *The elements of statistical learning*. Springer.

Ho, D. E., K. Imai, G. King, E. A. Stuart, et al. (2011). Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software 42*(8), 1–28.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

McCullagh, P. and J. A. Nelder (1989). Generalized linear models, no. 37 in monograph on statistics and applied probability.

Menard, S. (2002). *Applied logistic regression analysis*, Volume 106. Sage.

MIUR (2016). Criteri di ripartizione della quota premiale e dell'intervento perequativo del fondo di finanziamento ordinario (ffo) delle università statali per l'anno 2016. `http://attiministeriali.miur.it/anno-2016/dicembre/dm-29122016.aspx`, urldate=03-09-2017.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association 79*(385), 41–48.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenbaum, P. R. and D. B. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39*(1), 33–38.

Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences 7*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics 25*(1), 1.

Valentini, G. and F. Masulli (2002). Ensembles of learning machines. In *Italian Workshop on Neural Nets*, pp. 3–20. Springer.