TOR VERGATA
UNIVERSITY OF ROME

Università degli Studi di Roma "Tor Vergata"

Facoltà di Economia

Laurea/Bachelor of Arts

in

Global Governance

The effects of gender identity and immigration background on youth unemployment within a recent socio-economic landscape in Italy

Candidate: Arina Lopukhina

Supervisor: Barbara Guardabascio, Ph.D.

A.Y. 2021-2022

Dedications:

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Barbara Guardabascio, for the continuous support, availability to help, and invaluable patience.

Also, sincere thanks to Laura, Benjamin, Daniel, and the rest of the LIN11 crew for the cherished time spent together at the office and their contribution to my professional and daily life.

Finally, my appreciation goes to Jane, Alessandra, and Anna for the unwavering support and belief in me.

Table of Contents

Introduction

High unemployment rates have been a persistent problem in Italy over the past few decades. More so, young people remain one of the most fragile categories, which are most susceptible to political and economic fluctuations. Although the Italian government has repeatedly tried to introduce public policies to improve employment opportunities for younger generations, such as the Jobs Act of 2015, most of them have failed to bring about positive changes. Furthermore, the COVID-19 pandemic has added a new layer of complexity as, on one hand, it opened possibilities for remote employment, giving access to a wider job market, but on the other hand it jeopardized service and tourist industries which historically comprise a significant part of the Italian economy, with over 15 million people being employed in this sector (Statista Research Department, 2021).

However, within the overall fragile cluster of young people, there are minority groups that exhibit even more difficulties in the attempts to find employment. Historically, unemployment rates are significantly higher among women compared to men, which is primarily a result of the deep-rooted socio-cultural norms as well as influenced by childcare responsibilities. On average, ISTAT reports that unemployment rates are roughly 5-10% higher among young women compared to men, depending on the age group. Furthermore, people with non-Italian backgrounds, be they first- or second-generation immigrants, continue to face discrimination while seeking employment due to linguistic and socio-cultural barriers, aggravated by xenophobic tendencies (Fernandez Macias 2018).

This study aims at looking into the recent state of youth unemployment and the above-listed minority groups within it, attempting to trace whether there is indeed a large gap between the employment status of Italian men compared to women and people with immigrant backgrounds. The primary research question concerns the explanation behind the disparity in unemployment rates between the above-mentioned groups: do social factors such as respondents' age, gender, education, parental background, and country of origin significantly affect their chances of gaining employment? This research is meaningful not only because it would be a useful read for recent Italian university graduates who are about to enter the labor force, but the topics of socio-demographic analysis of Italian youth and their ability to gain employment are rather overlooked, so the paper adds value to the variety of social research, as the subject of unemployment is oftentimes addressed strictly from the economic perspective.

The first part of the paper examines the historical context of youth unemployment in Italy and the overall trends, as well as discusses the causes and repercussions of the skyrocketing unemployment rates. The latter chapters present the results of both inferential and classification analysis carried out on European Social Survey (ESS) data. ESS is a unique initiative that allows to access extensive socio-economic data about thousands of randomly selected respondents across Europe. ESS is conducted in bi-yearly multiple rounds, first performed in 2001 and continuing until today. Unfortunately, Italy has not participated in the rounds consistently due to organizational issues, hence the most recent available data is reflected in ESS rounds 8 and 9 from 2016 and 2018 respectively (ESS, 2018).

For the study over 50 variables are considered, ranging from basic descriptions of age, the highest level of education attained, and country of origin, to more extensive information such as household composition and parents' level of education, perception of their residential area, presence of physical or mental disabilities, and so on. Despite the issues and limitations which came along with using real-world survey data as well as the lack of economic data, the results allowed us to pinpoint several critically influential factors which will be explored further in the paper.

Chapter 1: Exploration of Youth Unemployment in Italy in the 21st century

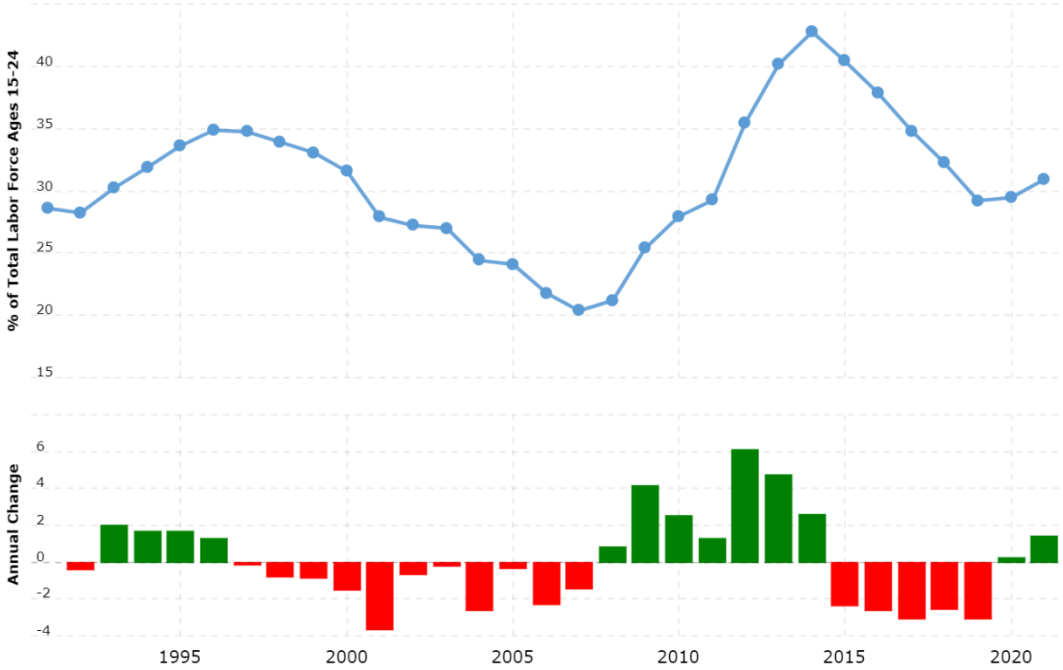
1.1 Background context

To begin the discussion, it is important to first define who is considered a young unemployed person. According to the International Labor Organization (ILO), youth unemployment is classified as the share of the labor force ages 15-24 without work but available for and seeking employment. Additionally, *Not in Education, Employment or Training* (NEET) is a separate definition that is used to characterize the situation of many young persons, aged between 15 and 29, in Europe, which are no longer a part of the education system but are yet not assimilated within the workforce. Nonetheless, background information is provided for those aged 15-34, while the two main age groups that are studied in more detail are the official youth unemployment rate referring to those between 15 and 24, and the NEET rates for individuals ages 15-29. Therefore, Italy's National Institute of Statistics, ISTAT (Istituto Nazionale di Statistica), keeps track of youth from age 15 to 34, breaking them into ranges from 15 to 24, 15 to 34, and 25 to 34. Lastly, another important term to consider within the framework of unemployment measurements is *inactive population*, which refers to young people who are: housekeepers, unregistered unemployed, or out-of-the-labor force but looking for a job, out-of-the-labor force not looking for a job but available to work, and those out of the labor force but not currently available to work (Bradley, Migali & Navarro Paniagua, 2020). In other words, it accounts for all those who even might be in the suitable age group but have a certain lifestyle that does not permit them to become employed, those who are simply not interested in finding a job due to various reasons, or individuals missing from the public records.

According to the World Bank (2022), Italy's youth unemployment rates have been fluctuating between roughly 20% and 40% (see Figure 1) throughout the past two decades, which began to rapidly rise as the result of the 2007-2008 Financial Crisis and worsened as a consequence of forthcoming political events. The following European debt crisis (2009 - the late 2010s) further aggravated the Italian economy, leading to youth unemployment peaking at an unprecedented rate of 43% in 2014. Apart from the crisis which brought Italy's industrial capacity down by nearly 25%, the European Union was also going through the accommodating inflows of refugees as a result of warfare in Syria, the Afghan war, Boko

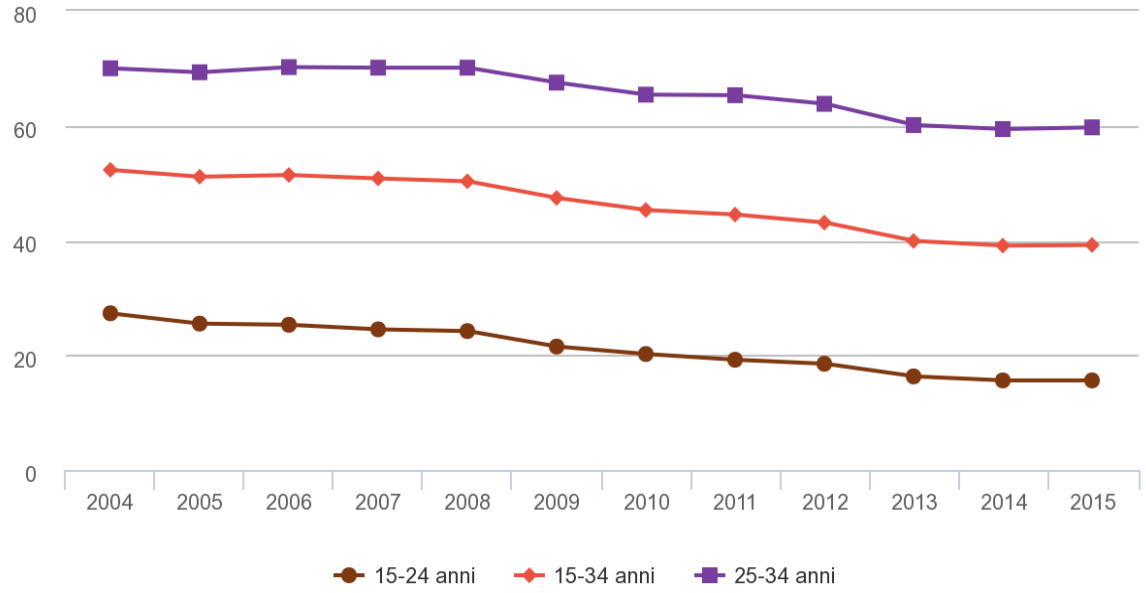
Haram insurgencies in Nigeria, and other global conflicts which occurred from 2014 onwards. Thousands of displaced refugees entered the EU through Italy in a span of a few months, with roughly 500,000 people receiving asylum and remaining in the country (Cirillo et. al. 2017). Such an influx in the foreign resident population added pressure on the already unstable labor market.

Figure 1: Historical youth unemployment rates in Italy (ISTAT)



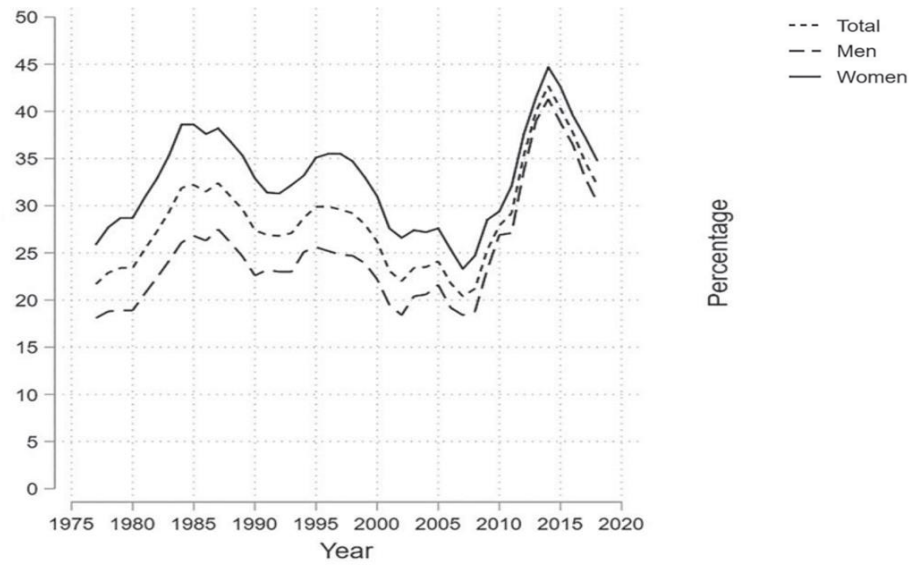
So, what are the trends among young people? Naturally, as age rises, the employment rates tend to be higher as well, as more and more individuals manage to secure jobs. On the surface level, it can be correlated with many different factors, from receiving higher education or gaining more extracurricular experiences to having a wider network of connections or ageism. Nevertheless, starting from the mid-twenties, most Italian people obtain a job. This is also reflected in Figure 2 below, which showcases ISTAT's data among age groups between 15 and 34, as one can notice that the employment rate among 25+'s is roughly 70%, while those of ages 15-24 is half as high.

Figure 2: Youth unemployment rates by age cohort (ISTAT)



Among those unemployed, there is a large disproportion among men and women, which also translates to older age cohorts. As can be seen in Figure 3, the gap is gradually closing: from as much as 15% in the 1990s to as little as 5% now. Nonetheless, the disparity remains.

Figure 3: unemployment rates, age 15-24 with gender differentiation (ISTAT)



Lastly, an immigrant background of an individual is shown to negatively contribute to their achievements in education and hence employment attainment (Azzolini et. al., 2012). Although migrants in Southern European countries have a higher rate of participation in the

labor market compared to the Central European ones, their jobs are rather characterized as low-paid and manual. Moreover, the Italian agricultural sector is known for the overexploitation of seasonal migrant workers at extremely low pay, which is borderline modern-day slavery (Scaturro 2021).

1.2 Insufficient labor policy-making

Italy has been struggling with high unemployment rates among both young people and adults long before the economic crisis largely due to its flawed labor policies, protectionist welfare system, and poorly adapted education system. The process of reforming Italian labor legislation began in the late 1990s with ‘‘Pacchetto Treu’’ (law 196/1997) attempting to ease the Employment Protection Legislation (EPL) and introducing temporary and para-subordinate contracts and reducing firing restrictions. In 2003, this initiative has been further extended with ‘‘Legge Biagi’’ (Law n.30/2003), which added several other contract types for project collaboration, staff leasing, and other non-standard contractual forms. Although seemingly successful at first, Lucidi and Kleinknecht (2010) point out that the gain in employment rates was largely attributable to the fact that the policies helped to rather formalize previously informal jobs as opposed to creating more opportunities. Given Italy's long-lasting struggle with the value of the shadow economy, it was a meaningful step. Nonetheless, Malgarini et al. (2013) claim that untying employers' hands did show a short-term rise in employment rates, however, had much more serious repercussions on workers' productivity. Temporary contracts turn out to have negative effects on employee motivation and labor productivity since although it might have become easier to find new employment, it is also associated with higher risks of losing it due to a lack of legal protection. Boeri and Garibaldi (2007) similarly detect a positive effect of temporary contracts on employment and a negative effect on labor productivity.

The underlying issues of labor market liberalization have shown themselves during the recession times a decade after. Due to the economic limitations brought upon Italy by the 2008 crisis, small firms and corporations alike chose to cut the labor force expenses, leaving thousands unemployed. Most of those laid off, unsurprisingly, were temporary workers, which led to a reversal in the trend of non-standard forms of contracts. According to ISTAT,

since the introduction of temporary contracts in the late 1990s, the share of young people employed with a temporary contract tripled from 20% to 60%, and it keeps growing.

Therefore, the Italian government tried to eliminate the discrepancy between permanent workers with an open-ended contract and temporary ones by introducing "Legge Fornero" in 2012. It aimed to weaken Article 18 of the previous Law 300 legislation and hence limit workers' protection if deemed necessary by the court. In more recent years, the 2014 Jobs Act (Decree Laws Nos. 148, 149, 150, and 15) completely abolished the article, thus harshly lifting firing restrictions. However, making it easier to fire a permanent contractor does not fully help temporary workers to secure their jobs, as it still allows for rotating interns and short-term workers in an attempt to cut down labor costs. Needless to say, young people, do not stand a high chance of obtaining long-term contracts at the beginning of their career. By definition, they are often more disadvantaged than more mature employment seekers in terms of both demand-side and supply-side factors. Not only do new job seekers tend to have fewer connections and relevant soft skills to boost their applications, such as creating a high-quality resume or cover letters as well as leaving a good impression at job interviews, but also firms often prefer to hire more experienced workers. Certainly, the overall inability to predict whether the long-term productivity of a young employee would be successful makes them perfect candidates for temporary contracts (Bradley et.al. 2019). Given Italy's traumatic experience of the recent economic crisis and the speedy termination of non-permanent job contracts, it is safe to infer that young people are the most fragile group as their employment terms make them primary candidates for lay-offs in economically challenging situations, such as the one the world is currently facing.

Moreover, Lucidi and Kleinknecht (2010) bring up the issue of the uneven wealth distribution between the North and South of Italy. To be clear, the t is rather a duality between Northern (Piedmont, Aosta Valley, Lombardy, Trentino-South Tyrol, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Liguria) and Central (Tuscany, Marche, Umbria, Lazio) administrative regions, opposed to the Southern ones (Abruzzo, Molise, Campania, Apulia, Basilicata, Calabria) and the Islands (Sicily and Sardinia). The southern provinces have struggled to develop at the same rate as their northern counterparts due to lower industrialization, and the gap is only widening. Southern Italy continues to have fewer

opportunities for education and work, as the rate of labor productivity in the South is on average 20% lower than in the North, and the unemployment gap is as high as 30%. According to Cannari and Franco (2010), bureaucratic offices, as well as health, educational, and legal institutions, show lower efficacy than the ones in northern provinces. The South continues to face struggles with low accessibility by public transportation, lack of infrastructure, and inefficient public institutions, and these issues are further aggravated by the continuous issue related to organized crimes.

As a result of the labor market liberalization, the employment opportunities dualism between the North and the South has gotten even worse - most of the new jobs were created in the already-advantageous regions, while the Southern provinces did not gain any new benefits. It is worth noting that Lucidi and Kleinknecht (2010) emphasize that the social groups most affected by the gap are women and young people. As can be seen in Figure 3, not only the dualism is terrifying, but also concerning the rest of the European population, Italy's unemployment stably outperforms the EU average every year. Desperate to find a job, people are forced to relocate to regions that hold more employment opportunities – big cities like Rome or Milan, and northern provinces in general. It is estimated that at least two million southern Italians moved out to other regions solely to find better employment. Similarly, younger people tend to relocate to Northern and Middle Europe as there is a shortage of skilled labor which translates into potentially receiving higher salaries than one could find in Italy for similar positions. Italy has been experiencing a brain drain for over a decade, with thousands of young Italian leaving to seek better employment and living standards elsewhere, which makes the country lose approximately 14 million euros in human capital investment (Paolazzi, 2017). During the COVID-19 pandemic, around 100,000 returned to the country, slightly reversing the outflow, but it is likely to be temporary as the pandemic has come to end abruptly. Thus, labor market liberalization only worsened the conditions of the most fragile group, the youth, by solidifying the use of temporary contracts (most of which are claimed to be involuntary, as noted by Cirillo et. al., 2016).

Additionally, in terms of the gender gap explored at the beginning of the chapter, the Italian government has relied on the *pink quotas* (*le quote rosa*) to ensure that women have more access to day-to-day jobs as well as have a higher representation in the government

and corporate leadership. Promulgated in 2011 as the Law 120 of 12, the quotas were set to ensure women's representation across different sectors, targeting 40%. According to Pastore and Tommaso (2016), the policy has granted more women unprecedented access to jobs, turning out to be effective, although mostly for women in high-skilled job positions. Furthermore, Italy is among the top EU countries which achieved or came close to achieving the benchmark. Lastly, the quotas are not set in stone and the legislation could be overturned, so it is questionable whether they will continue to have a long-term effect.

However, while there are some mediocre attempts to diminish the gender gap on the legislative level, Italy continues to disregard the issues related to discrimination against ethnic and religious minorities. It has been over thirty years since the adoption of the UN's Convention on Protection of the Rights of All Migrant Workers and Members of their Families (CRMW), yet Italy has neither signed nor ratified it. Moreover, thanks to the *Decreto Sicurezza* introduced by the infamous Matteo Salvini in 2018, most of the humanitarian obligations outlined by the Geneva Convention have been lifted. The decree has been partially mitigated with the 2020 *Legge 18 (n.173)* which outlines broader opportunities for migrants and refugees to convert their asylum-seeking residence permits into work ones (Scaturro 2021). Nonetheless, the Italian government continues to disregard opportunities for combatting discrimination and closing the disadvantage gaps when it comes to minorities. Most of the legislation that now constitutes the pillars for protection against hate crimes and discrimination are the laws introduced over two decades ago, as, for instance, *Legge Mancino* (1993) against racism and discrimination against one's faith. In the meantime, new proposals like DDL Zan, which mainly aimed to tackle inequalities faced by the Italian LGBT+ community but also touched upon other forms of discrimination, keep getting vetoed largely due to the prevalence of the conservative senators in the Italian government.

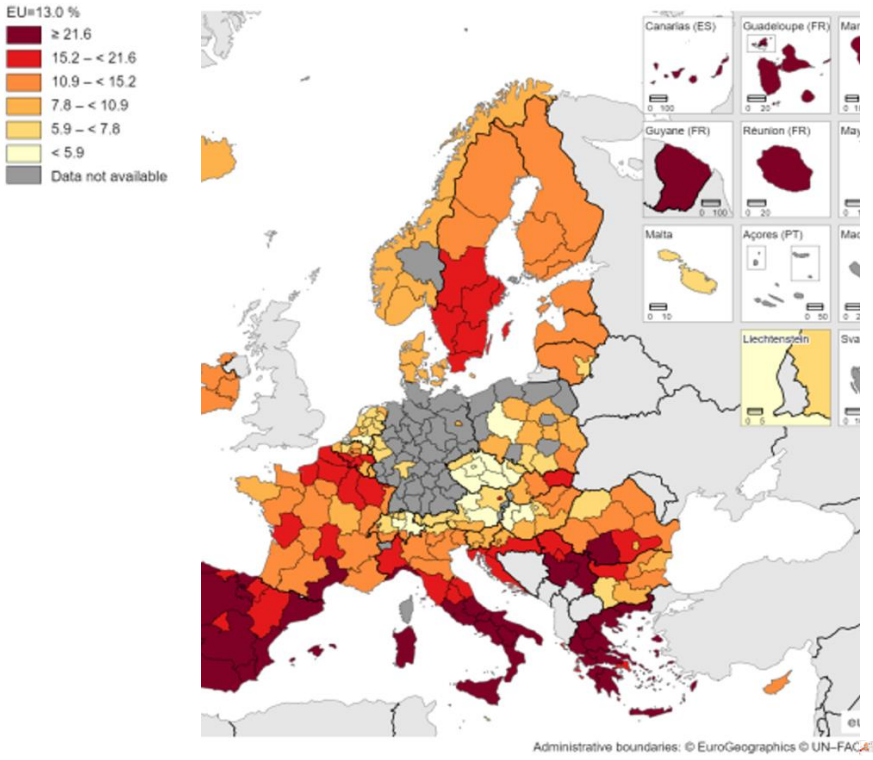


Figure 4: Youth unemployment by region (EUROSTAT)

1.3 Protectionist welfare state flows

The Italian bureaucratic system continues to struggle, and the unemployment benefits matter is not an exception. IZA Institute of Labor Economics infers that not only the logic of distributing the benefits is rather unbalanced, but the lack of controllership adds a new layer of complexity. For instance, most unemployment benefits address experienced workers who find themselves in a situation of suddenly losing employment. Hence, if ones do not have a history of employment valid in Italy, which is mostly the case for young job seekers and immigrants, one cannot request unemployment insurance, such as NASpI or DIS-COLL (Pacifico et. al. 2018).

Logically, when a person struggles to find a job, they would certainly want an extra source of income, which can be found in unemployment benefits. The only appropriate way to receive governmental support is to request *reddito di cittadinanza* (RDC), which is a monthly allowance for low-income households, Italy's most rigorous attempt to eliminate poverty up to today. Although it was not intended as an unemployment welfare payment, RDC has quickly become a way for those who nearly lose hope in finding a job to gain extra

income by simply filling out a request to INPS, Italy's national social security organization. There is no thorough background check, and the process is rather simple, so even if a person is not looking for a job or is in education, they can still play unemployed and receive an allowance of several hundred euros a month. For instance, there are cases of young students requesting RDC while being simultaneously enrolled in an educational institution, while falsely claiming that they pursue job hunting, which is far away from reality. Moreover, RDC has been known for recurring multi-million euro tax frauds, which once again confirms the insufficiency of the system regulations. Such abuses and misallocations only further aggravate the already scarce funding distribution as RDC continuously fails to reach the intended recipients (Maitino et. al., 2022).

1.4 Education system drawbacks

The two main challenges related to the Italian education system are the completion rate and the mismatch of skills. According to the European Commission report, Italy has one of the lowest higher education attainment rates in the EU, coupled with an extremely high university drop-out rate of 45%. College students are largely unmotivated to graduate on time, with only one-third managing to finish their degrees within the first graduation round, while the average for a Bachelor's degree stands at 5.1 years instead of 3 (Montanari et.al. 2015). In absence of university education, there are also no vocational training opportunities in Italy, which results in further gatekeeping of job opportunities for those without a college degree.

However, even people with a degree report that it is usually not enough to obtain a job solely based on academic achievements. Almost half of the young graduates shared that they lacked adequate training to be hired for one or more positions, while one-third also pointed out the lack of professional experiences, such as mentorship or internship (Montanari et.al., 2015). These numbers are scarily high and point to the deep-root cause of insufficient preparation within educational institutions. Most Italian universities appear to have an outdated teaching approach that does not let students complete enough practice before joining the workforce. As internships and fellowships are quite scarce (especially in the South) and demand a high initial standing, many students are only able to secure their first-ever internship in the last year of their degree or after graduation. Most likely, proper career

services and pre-graduation preparation could help steer students toward the right path to employment. On the side of the firms, the use of personal connections and nepotism shall be limited in the Italian labor market and the process of securing a job has to be more transparent.

Moreover, for minority groups, the education system does not fully tackle the integration of migrants (especially those with Middle Eastern and African origins), as even second-generation ones continue to experience difficulties with being fully integrated within the system compared to Italian students. Although the trend is slowly improving, support from the public bodies is still absent in Southern Europe as well as Central Europe (Fernandez Macias 2018). The only European countries that show progressive improvements are Sweden and the UK thanks to their extensive focus on integrating inclusive teaching within secondary and tertiary education as well as introducing scholarship opportunities that are catered to immigrant minorities.

1.5. Relevant outcomes of persistently high youth unemployment

The prolonged struggle to find employment has adverse consequences for both the national economy as well as the affected individuals. On a socio-psychological level, Bell & Blanchflower (2010) claim that unemployed individuals are prone to becoming demotivated and developing mental and physical illnesses more frequently than their employed peers, which might also result in lower life expectancy. Unemployed people are more likely to develop mental health issues, such as depression and anxiety, due to the limited ability to be in control of their lives and pursue their true aspirations, as well as the piling pressure from societal expectations. This also affects their close family and friends, as they are usually the ones to take care of the struggling individual in both economic and non-tangible terms. In some cases, joblessness can also become the primary reason for committing suicide.

In terms of the country's perspective, unemployment adds pressure on public health services due to the above-mentioned rising probability of developing health issues caused by prolonged and unsuccessful job searches. Moreover, an increase in suicide rates among young adults translates into an undesired loss of a fully capable labor force and overall human capital. On a different note, a rise in unemployment correlates with higher crime

rates, which once again challenges the quality of life not only of the ones directly affected by unemployment and members of their households but society as a whole. It is extremely relevant in the case of Italy, which has struggled with combatting organized crime for generations. Lastly, those who are exhausted by unsuccessful attempts to find a job but luckily remain motivated and in good health decide to leave the country and seek employment elsewhere, which results in further loss of human capital for the nation as a whole.

Chapter 2: Regression analysis

To look at the research question from a different angle, two types of analysis were carried out in an attempt to see if there is any correlation between the socio-economic features of the respondent or the household they live in and the likelihood of them getting a job. Several binary logistic regression models were constructed to carry out inferential analysis and understand which of the variables describing the sample can be seen as drivers of differences in the unemployment status of the subjects. Secondly, a random forest was performed for carrying out supervised classification. The results of the two models were finally compared to spot any interesting patterns. All of the analysis was carried out in R (see the full code in the Appendix).

The goal of this study is to define the variables, if any, among socio-economic characteristics such as gender, age, immigration background, educational background, or family history, that significantly allow us to discern between young people that can attain a job and those who don't. For this particular aim, logistic regression appeared to be the best choice as it suited the binary outcome (employed or unemployed) and allowed us to look into a wide variety of different factors which could influence the dependent variable. Moreover, it will let us understand the extent and "direction" of this impact.

Conducting a random forest analysis will provide an additional outlook on the most significant variables outlined by logistic regression analysis, potentially confirming that some of them are indeed meaningful. The logic behind choosing a classification tree as a classification algorithm lies in the fact that the results are in general easily interpretable and, mostly, in the possibility of observing which of the features the algorithm relies on to split between the observations. The decision of relying on ensemble methods was driven by the increase in the stability of the results their deployment insures, thus making them more reliable. Moreover, one of the advantages of classification trees and random forests is that these are in general good at handling wide datasets, which came in handy considering the initial high number of variables considered and limited sample size (James, Witten, Hastie & Tibshirani, 2015).

2.1. Methodology

Before diving deeper into the step-by-step process, it is important to outline the theoretical background of the statistical methods utilized in the study in order to ensure full understanding of the consequent analysis. Logistic regression, the first approach undertaken for the data analysis, is a statistical method which helps to determine a binary outcome based on the prior observation of the dataset. It is used for estimating the relationship between one or more independent variables and the dependent one (James, Witten, Hastie & Tibshirani, 2015). In the case of this study, the desired outcome is knowing whether a person is employed or unemployed, and how it is affected by various socio-demographic variables, if at all. The model can be represented with the following function:

(1.1) $$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

If the function is slightly modified, we can derive the quantity called *odds* (to the left). It is a value from 0 to infinity which represents the likelihood of a particular outcome.

(1.2) $$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

By taking the logarithm of both sides of the formula 1.2, the left side is now modified into log-odds:

(1.3) $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

From the formula, it can be noticed that in logistic regression, an increase or decrease by one unit of X results into a respected change in log-odds by β1. Hence, regardless of the value of X, if β1 is positive then increasing X will be associated with increasing p(X), and if β1 is negative then increasing X will be associated with decreasing p(X) (James, Witten, Hastie & Tibshirani, 2015, p. 132). This terminology will be later used to interpret the results produced by the R code.

Lastly, another statistical method consulted in the study is random forest classification. It is essentially an ensemble learning method for defining non-linear relationships as it can classify new objects based on the input vector, and it can be attributed as one of the supervised-learning techniques (Shah, Patel, Sanghvi & Shah, 2020). However, the random

forest approach is fundamentally different from logistic regression which measures the statistical significance of given variables with respect to probability.

Random forest is a combination of decision trees, which split into "branches" if there is a significant distinction between variables. Each time the algorithm splits the data, a random sample of *m* predictors is chosen to split the full set of *p* predictors. Usually, *m* would be equal to the square root of the total number of *p* predictors at each iteration , p. 320). Therefore, it is very versatile and easy-to-follow, and it can perform on a wide variety of data values, which makes random forest good candidate for our dataset with an extensive amount of variables. However, it has to be noted that some claim the algorithm can be biased when it comes to categorical data, which constitutes the majority of our variables (James, Witten, Hastie & Tibshirani, 2015).

2.2 Dataset and variables

This section aims to describe the data dimensions and the way the dataset was compiled. The data used for the study comes from the European Social Survey (ESS), a socio-demographic survey that collects over 1000 data points per respondent not only regarding their basic demographic profile, but also the state of their well-being, political views, or media preferences, as well as information about their household and living arrangements. The data is randomly sampled across multiple European countries every two years and is publicly available for research purposes.

Since the beginning of the ESS data collection twenty years ago, Italy has participated in four rounds out of ten, with the most recent ones dating to 2016 (ESS Round 8) and 2018 (ESS Round 9). As the paper focuses on the socio-economic landscape of Italy within the past decade, these two rounds were chosen for the construction of the data frame, since not only do they contain the most up-to-date information but also because the survey has greatly evolved over the years, so the latter rounds contain more extensive, quality responses.

Given the focus of the study on Italy, only data for respondents who filled out the response in the given country is considered in the analysis. Moreover, the dataset initially contained information about people of all ages, from 15 to over 80, which had to be edited. In the end, only the responses of individuals between the ages of 15 and 34 were kept in the frame, largely due to the above-discussed definitions of the calculation of youth unemployment rates and the need to observe the inactive population up until the mid-30s.

Therefore, the study only focuses on the recent snapshot of the young people's attempts to find employment in 2016-2018 and does not take into consideration the comparison between younger and older generations, as well as the past historical practices.

The total number of observations before data cleaning was 1,276, which decreased to 637 after eliminating survey participants with a high share of missing values contained in their responses. This was done to prevent potential disruptions in the analysis results. Although such a significant reduction slightly increased the margin of error, the sample size was sufficient to produce insightful results.

Overall, from roughly a thousand available variables, the list was narrowed down to sixty-three, featuring a mix of numeric and categorical values. The full list of the original dataset is featured in Table 1 below. Some of the below-mentioned variables can be automatically reduced by eliminating the ESS-related information, such as round number, country, weights, and so on, while the rest can be broken down into several groups.

This data is purely informative and does not give us any practical knowledge about the respondents' employment journey. An important disclaimer is needed, however: although ESS moderators recommend using weights with the collected data to better mirror the demographic structure of the country considered, this statistical analysis does not consider the coefficients. The main reason for this is the potential distortion in logistic regression analysis results as the weights would alter the coefficients.

Apart from the survey-specific variables, the first main group of variables refers to the demographic profile: age, gender, household composition, domicile type (whether it is a city, small town, or a village), close relationships (father, mother, partner), and all of them but age and year of birth are strictly categorical. Only a small number of respondents had a non-Italian background, however, it was still representative of various ethnic minorities.

The second group is related to education, as it remains one of the main deciding factors when it comes to employment seeking (Montanari et.al., 2015). The key numeric education-related variable is years spent in education, then the supporting categorical ones are levels of parents' education, which can help understand the family's socio-economic status, and, finally whether the respondent is currently pursuing any type of education. The third major group of variables is set to elaborate on the employment status of the respondent: it looks into whether the individual is currently seeking employment and for how long, whether they had previously managed to secure a job, or if they currently have a job. Lastly, the remaining

miscellaneous variables concerning the respondent's perception of their quality of life and whether they have ever encountered discrimination for any reason.

Table 1: Variable names and descriptions (*ESS Codebook, ESS9-2018*, 2019)

	VARIABLE NAME	DESCRIPTION	GROUP
1	essround	Number of the survey round	Survey information
2	edition	survey edition	Survey information
3	idno	ID of the respondent	Survey information
4	cntry	County where the survey is conducted	Survey information
5	dweight	Assigned weight	Survey information
6	pspwght	Assigned weight	Survey information
7	pweight	Assigned weight	Survey information
8	aesfdrk	Being afraid of walking home in the dark (due to crime)	Quality of life
9	blgetmg	Belong to minority ethnic group in country	Demographic
10	cntbrthc	Country of birth	Demographic
11	ctzshipc	Country of citizenship	Demographic
12	dscrage	Exeperienced agism	Quality of life
13	dscrdsb	Exeperienced ablism	Quality of life
14	dscretn	Exeperienced ethnic discrimination	Quality of life
15	dscrgnd	Exeperienced gender discrimination	Quality of life
16	dscrgrp	Exeperienced minority group discrimination	Quality of life
17	dscrnap	Not experienced discrimination	Quality of life
18	dscrntn	Exeperienced nationalism	Quality of life
19	dscrrce	Exeperienced racism	Quality of life
20	dscrsex	Exeperienced sexism	Quality of life
21	happy	Feels happy	Quality of life
22	health	Feels healthy	Quality of life
23	hlthhmp	Mental or physical health hampered	Quality of life
24	livecnta	Year when came to reside in Italy	Demographic
25	lnghom1	1st language spoken at home	Demographic
26	lnghom2	2nd language spoken at home	Demographic
27	rlgdgr	Is religious	Demographic
28	rlgdnm	Name of the religion	Demographic
29	hhmmb	Number of people living regularly as member of household	Demographic
30	gndr	Gender	Demographic
31	rshipa2	2nd person in household: relationship to respondent	Demographic
32	rshipa3	3rd person in household: relationship to respondent	Demographic
33	agea	Age	Demographic
34	anctry1	First ancestry, European Standard Classification of Cultural and Ethnic Groups	Demographic
35	domicil	Describing the area of residence	Demographic
36	dsbld	Has a disability	Demographic
37	edctn	Last 7 days: in education	Education
38	edulvlb	Highest education level, respondent	Education
39	edulvlfb	Highest education level, father	Education
40	edulvlmb	Highest education level, mother	Education
41	edulvlpb	Highest education level, partner	Education
42	eduyrs	Years in education	Education
43	emplrel	Employment relationship	Employment
44	hswrk	Mainly carries out housework, looking after children	Demographic
45	isco08	ISCO work industry code	Employment
46	lvgptnea	Ever lived with a partner outside of marriage	Demographic
47	maritalb	Marital status	Demographic
48	mnactic	Main activity within past 7 days	Employment
49	pdjobev	Ever had a paid job	Employment

50	pdwrk	The work is paid	Employment
51	pdwrkp	Partner is employed	Employment
52	tporgwk	Type of organization working for	Employment
53	uemp12m	Unemployed withing past 12 months	Employment
54	uemp3m	Unemployed withing past 3 months	Employment
55	uempla	Unemployed, actively looking	Employment
56	uempli	Unemployed, not looking for a job	Employment
57	wkhct	Total contracted hours per week in main job overtime excluded	Employment
58	wrkac6m	Has had a work contract for 6 months	Employment
59	wrkctra	Work contract type	Employment
60	nacer2	Job code	Employment
61	yrbrn	Year born	Demographic
62	new_vars.hincsrca	Main source of household income	Demographic
63	new_vars.hinctnta	Household's total net income, all sources	Demographic

2.3 Data preprocessing

The initial step of the statistical analysis is preprocessing. The chosen data set has turned out to be very challenging to format due to two reasons – a high number of categorical variables for relatively low sample size and the presence of missing values. Social survey data is generally hard to process because most variables are broken down into multiple categories, with only a few of them being binary. Moreover, the values indicating the industry code of the job, language spoken at home, or country of birth other than Italy had a myriad of sub-categories, some of which were only represented by one or two respondents. At the first attempt to build a model, variables with many sub-categories were left untouched, but it was later deemed necessary to combine them into fewer groups.

Secondly, eliminating missing variables was hampered by the coding of the responses. Instead of putting N/A or a similar standardized value for whenever a response was missing, oftentimes each variable had a unique code for reporting non-response. Moreover, non-responses were also classified into different types: example, respondents' conscious refusal to respond, respondent's having doubts over picking a most suitable response, and sometimes certain questions would be ignored or omitted due to their inapplicability. For example, if a person responded that they do not have a job at the moment, all the consequent questions regarding their job type would not be addressed, and hence it would result in encoding with a distinct value.

Therefore, to clean variables of missing values, a tedious procedure had to be established. It was necessary to go to the ESS Data Protocol booklet, search for each of the variables one

by one, understand the range of potential responses, then manually eliminate the codes related to N/A – usually, a "777" or "888". This action had to be performed for each of the sixty variables to avoid having regression results disrupted by the missing values. Data cleaning was performed both in Excel and R.

2.4 First model preparation

Before moving to the discussion of building the target variable, it is important to point out a major disclaimer. The goal of the exercise was to conduct inferential analysis, meaning that we wanted to see if there were any variables related to the socio-demographic background that would show to have a significant impact on the outcome of being unemployed. As there is no goal of forecasting future unemployment rates or making any predictive analysis, the given dataset is not initially split into testing and training sets for the logistic regression analysis, the significance of variables is analyzed using the whole dataset.

Once the data was polished and the terms are clarified, it was finally time to start building the target. As reflected in the previous chapter, the goal of the study is to trace if there is any noticeable influence on youth unemployment rates, hence the desired outcome would revolve around the differences between employed and unemployed individuals. To build the target vector, *uempla* column was utilized, which is a binary categorical variable that reflects whether a person is unemployed and is searching for a job or otherwise. Provided that there was a similar variable, *uempli*, which pointed out who is an unemployed person but not seeking any employment at the moment, it was important to separate the two. *Uempla* refers to those included in the calculation for unemployment, while *uempli* refers to inactive population (as defined by ILO) or could simply run in parallel with *eductn* – classifying an individual as a student who is not actively searching for a job.

Therefore, respondents who are associated with positive *uempli* were eliminated from the study to be able to focus solely on the respondents with a defined desire to find a job. Filtering by *uempli = 0* (see the code in the Appendix) allowed us to only leave respondents who are interested in finding a job, or at least have not declared otherwise. As a result, our sample size reduces by roughly 50 individuals, leaving us with 1,276 observations in total, among which 185 are classified as unemployed and looking for a job.

Next, we checked for missing variables again and discovered that there were still some inconsistencies with variables related to the country of birth and citizenship, which remained with over 200 missing values. Given a large amount of missing data, it was decided to eliminate the two, as there were other variables alluding to the immigration background as well as there was no option to impute them. With respect to other variables, they were complete and ready for the regression analysis.

Before creating the model, it was necessary to control whether all of the initially selected variables were suitable and should have been included. Namely, we excluded variables that could be deemed too similar to the target variable, such as unemployment in the past 3 to 12 months, work contract details, industry name, and so on. Then, we removed the variables which were deemed redundant in the context, such as ESS round data, weights, and similar. Lastly, it occurred that the dataset had only two people who reported to have some form of disability, so we excluded them as well due to insufficient material. Also, disabled people must be facing different difficulties with finding employment, so it would not make sense to compare them with an average respondent (Pacifico et. al., 2018).

Finally, after eliminating more variables and adjusting the categorical codes to be binary (0 and 1) whenever possible and factorizing the categorical variables, we tried to run the first glm model in R, using all of the remaining 30 variables. The result was rather unfortunate – the model simply did not converge.

2.4 Second model preparation

As we still had a lot of variables in the dataset, the initial guess was that there must be too many categories within some of them to allow the model to function. Therefore, we proceeded to build a new data frame. First, we grouped up languages into fewer categories – Italian, Italian dialect, European, and non-European. Furthermore, the obtained highest level of education has been reworked. Now, instead of using the ISCED classification as in ESS questionnaire, we split the responses between Low level (no school – middle school), Medium (some middle school-high school diploma), and High (Bachelor's degree and above). Nonetheless, this change did not help to make the model run, and the algorithm once again did not converge.

Another suggestion remaining was that we failed to exclude more variables that are too similar to the target. Therefore, we tried running a Type II ANOVA to try to understand

whether some variables would be highlighted for being extremely correlated with the output. The ANOVA test determines whether two or more different groups have any relationship between each other. In other words, ANOVA shows how much variance is added to a sample by all of the distinct factors (James, Witten, Hastie & Tibshirani, 2015, p. 132).

Indeed, the ANOVA for the failed glm algorithm pointed out only one variable *mnactic,* which is responsible for the answer to the question regarding the person's current main activity (can be seen from the results of the ANOVA in Table 2). Thus, it indeed gave us the same information about the subject as the output variable and had to be excluded in the latter models to obtain more accurate results.

Table 2: Results of Type II ANOVA.

```
## Response: X2$uempla
##                    LR Chisq Df Pr(>Chisq)
## aesfdrk               0.00  5          1
## blgetmg               0.00  3          1
## dscrage               0.00  1          1
## dscretn               0.00  1          1
## dscrgnd               0.00  1          1
## dscrgrp               0.00  2          1
## dscrnap                     0
## dscrntn               0.00  1          1
## dscrrce               0.00  1          1
## dscrsex               0.00  1          1
## hlthhmp               0.00  2          1
## rlgdgr                0.00 12          1
## hhmmb                 0.00  8          1
## gndr                  0.00  1          1
## agea                  0.00  1          1
## domicil               0.00  4          1
## edctn                 0.00  1          1
## eduyrs                0.00  1          1
## hswrk                 0.00  1          1
## lvgptnea              0.00  4          1
## maritalb              0.00  4          1
## mnactic             264.73  5     <2e-16 ***
## pdwrkp                0.00  1          1
## new_vars.hincsrca     0.00  9          1
## new_vars.hinctnta     0.00  9          1
## lnghom1ITA            0.00  1          1
## lnghom1DIAL           0.00  1          1
## lnghom1EURO           0.00  1          1
## edulvlbLOW            0.00  1          1
## edulvlbHIGH           0.00  1          1
## edulvlfbLOW           0.00  1          1
## edulvlfbIGH           0.00  1          1
## edulvlmbLOW           0.00  1          1
## edulvlmbIGH           0.00  1          1
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.5 Third model preparation

Now that the model was adjusted to exclude the variable which corrupted the results, the algorithm finally converged. The results appeared to be quite mixed: the p-value of some of the variables was significantly low to infer that there was some correlation between some socio-demographic parameters and the outcome variable. As seen in Table 3, the model has pointed out the significance of 9 variables, and borderline significance (p-value of 0.1) of 4 variables. As expected, an increase in unit age would result in a decrease in the odds of *uempla* taking the value of 1 opposed to *uempla* taking the value of 0, which signifies a lower probability of being unemployed and actively seeking a job. Similarly, a low level of education (no high school diploma) signaled higher odds of being unemployed. Surprisingly, the model did not indicate any significance for parental levels of education, which was deemed one of the determining factors in the literature review in Chapter 1. Ironically, none of the variables related to experiencing any type of discrimination have shown any significance.

Another interesting outcome is seeing that a high level of religious commitment (the respondents were asked to rank the strength of their faith from 0 to 10, 10 being the strongest) is associated with a positive correlation with being unemployed, which could be potentially explained by the rigidity of religious following. Also, living in a countryside/village (*domicil4*) is borderline significant for implying a higher odd of being unemployed as well, hence supporting the assumption about labor opportunities disproportion between rural and urban areas.

Moreover, with the increase in the number of members of the household (*hhmmb)*, the odds of remaining unemployed increased as well. Given that the variable describing the obligation to perform house chores and/or look after children (*hswrk1*) has shown to be highly statistically significant and results in negative odds of being unemployed and looking for a job, it could be implied that in the case of having large households with children, certain respondents assume responsibilities over house-making and potentially drop out from the pool of active job seekers.

Lastly, following education courses within the past 7 days (*edctn*) has shown a very low p-value together with negative odds of being unemployed and seeking employment. This might indicate a controversial result since it could signal insufficient reporting among

respondents, as some of them might be involved in training while remaining open to pursuing job opportunities, however, this contradicts the definition of youth unemployment and result interpreting harder.

To double-check the results, Type II ANOVA was performed to confirm whether the results of the model were accurate. As seen in Table 4, ANOVA results also highlighted the high significance of Chi-squared coefficients for age, attainment of education, as well as presence of house-making duties.

Table 3: Coefficients output for Model 3.

```
## glm(formula = X3$uempla ~ ., family = binomial(link = "logit"),
##     data = X3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8263  -0.5913  -0.1468  -0.0295   4.1383
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -29.06415 3936.67450  -0.007  0.99411
## aesfdrk2       -0.25218    0.36316  -0.694  0.48743
## aesfdrk3        0.46234    0.48439   0.954  0.33984
## aesfdrk4       -0.15017    0.67154  -0.224  0.82305
## aesfdrk7      -15.60500 6522.63866  -0.002  0.99809
## aesfdrk8        0.74298    1.57691   0.471  0.63753
## blgetmg2        0.13638    0.99127   0.138  0.89057
## blgetmg7      -17.17335 6522.63871  -0.003  0.99790
## blgetmg8        0.94983    1.69235   0.561  0.57463
## dscrage1      -16.76082 2649.82851  -0.006  0.99495
## dscretn1      -14.50837 2559.46582  -0.006  0.99548
## dscrgnd1      -16.06238 2784.07123  -0.006  0.99540
## dscrgrp2       -0.18916    1.00742  -0.188  0.85106
## dscrgrp7        0.96861    1.84578   0.525  0.59974
## dscrgrp8        0.11002    1.82212   0.060  0.95185
## dscrnap1             NA         NA      NA       NA
## dscrntn1        0.31043    1.81660   0.171  0.86432
## dscrrce1       -1.34205 2769.35108   0.000  0.99961
## dscrsex1      -13.49580 3223.89573  -0.004  0.99666
## hlthhmp2       12.89886 3509.29433   0.004  0.99707
## hlthhmp3       12.17920 3509.29417   0.003  0.99723
## rlgdgr1        -0.23113    0.76050  -0.304  0.76119
## rlgdgr2        -1.15023    0.88301  -1.303  0.19270
## rlgdgr3        -0.10954    0.65459  -0.167  0.86710
## rlgdgr4        -0.15468    0.71261  -0.217  0.82816
## rlgdgr5         0.37318    0.54412   0.686  0.49282
## rlgdgr6         0.28165    0.55728   0.505  0.61328
## rlgdgr7         0.45436    0.55329   0.821  0.41153
## rlgdgr8         0.59890    0.62143   0.964  0.33517
## rlgdgr9        -0.60323    0.98171  -0.614  0.53890
## rlgdgr10        1.50056    0.75741   1.981  0.04757 *
## rlgdgr77      -15.07185 4075.05212  -0.004  0.99705
## rlgdgr88      -15.83129 3054.78220  -0.005  0.99587
## hhmmb2          1.36766    0.72627   1.883  0.05968 .
## hhmmb3          1.41749    0.64286   2.205  0.02745 *
```

```
## hhmmb4          0.98500    0.64460   1.528  0.12649
## hhmmb5          1.36665    0.69904   1.955  0.05058 .
## hhmmb6          2.28244    0.99895   2.285  0.02232 *
## hhmmb7          4.05552    1.76859   2.293  0.02184 *
## hhmmb8        -12.47988 4513.67738  -0.003  0.99779
## hhmmb77       -15.23027 2163.66352  -0.007  0.99438
## gndr1          -0.13394    0.31347  -0.427  0.66918
## agea           -0.12005    0.03948  -3.040  0.00236 **
## domicil2       -0.21586    0.98657  -0.219  0.82681
## domicil3        0.82684    0.57125   1.447  0.14778
## domicil4        1.03217    0.55257   1.868  0.06177 .
## domicil5        0.61468    0.95559   0.643  0.52006
## edctn1         -5.86514    1.13026  -5.189 2.11e-07 ***
## eduyrs          0.02101    0.07862   0.267  0.78930
## hswrk1         -3.75296    1.53596  -2.443  0.01455 *
## lvgptnea2      -0.54216    0.43983  -1.233  0.21770
## lvgptnea6      -1.85700    0.80092  -2.319  0.02042 *
## lvgptnea7     -14.94755 6522.63900  -0.002  0.99817
## lvgptnea8     -18.58078 6522.63872  -0.003  0.99773
## maritalb2     -15.20843 1934.86296  -0.008  0.99373
## maritalb3       2.70722    1.63189   1.659  0.09713 .
## maritalb4     -15.53720 4202.32929  -0.004  0.99705
## maritalb5     -18.58005 6522.63877  -0.003  0.99773
## maritalb6       0.09908    0.43795   0.226  0.82102
## lnghom1ITA1    16.90899 1783.88821   0.009  0.99244
## lnghom1DIAL1   16.51904 1783.88834   0.009  0.99261
## lnghom1EURO1   16.31282 1783.88827   0.009  0.99270
## edulvlbLOW1     1.09066    0.48396   2.254  0.02422 *
## edulvlbHIGH1    0.07350    0.56307   0.131  0.89615
## edulvlfbLOW1    0.17964    0.35396   0.508  0.61180
## edulvlfbIGH1   -0.02046    0.70438  -0.029  0.97683
## edulvlmbLOW1   -0.04019    0.34725  -0.116  0.90785
## edulvlmbIGH1    0.09038    0.78554   0.115  0.90840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.03  on 606  degrees of freedom
## Residual deviance: 361.84  on 540  degrees of freedom
## AIC: 495.84
##
## Number of Fisher Scoring iterations: 17
1.     pR2(mod3)
2.     ## fitting null model for pseudo-r2
3.     ##          llh      llhNull           G2     McFadden         r2ML         r2CU
4.     ## -180.9175922 -265.0146892  168.1941940    0.3173299    0.2420144    0.4155575
```

Table 4: ANOVA output for Model 3.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: X3$uempla
##             LR Chisq Df Pr(>Chisq)
## aesfdrk        3.568  5  0.6131299
## blgetmg        0.801  3  0.8491700
## dscrage        1.179  1  0.2776437
## dscretn        0.147  1  0.7013949
## dscrgnd        0.637  1  0.4248669
## dscrgrp        0.576  2  0.7496066
## dscrnap               0
## dscrntn        0.029  1  0.8647371
```

```
## dscrrce        0.000  1  0.9999251
## dscrsex        0.050  1  0.8237442
## hlthhmp        0.478  2  0.7875832
## rlgdgr        12.468 12  0.4088723
## hhmmb         12.578  8  0.1272281
## gndr           0.183  1  0.6689282
## agea           9.659  1  0.0018845 **
## domicil        5.704  4  0.2223431
## edctn         78.035  1  < 2.2e-16 ***
## eduyrs         0.071  1  0.7899213
## hswrk         11.963  1  0.0005426 ***
## lvgptnea       6.766  4  0.1487723
## maritalb       5.311  5  0.3791376
## lnghom1ITA     2.428  1  0.1192098
## lnghom1DIAL    1.671  1  0.1960624
## lnghom1EURO    1.556  1  0.2122363
## edulvlbLOW     5.107  1  0.0238264 *
## edulvlbHIGH    0.017  1  0.8962043
## edulvlfbLOW    0.259  1  0.6107882
## edulvlfbIGH    0.001  1  0.9768136
## edulvlmbLOW    0.013  1  0.9078822
## edulvlmbIGH    0.013  1  0.9088036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.6. Fourth model preparation

The peculiar results of the previously discusses model, especially concerning rather low R-squared value of 24% and adjusted R-squared of 42%, inspired further modifications to the model. For the next model, values with relatively high p-values and Chi-squared values have been eliminated, bringing the total number of considered variables down to 20.

However, it did not yield much difference. The model returned the same list of significant variables as the one before, with most of them remaining in the same p-value ranges as before. Running ANOVA once again did produce slightly different results, with variables attributed to Italian or European languages spoken at home showing higher Chi-square values. This could potentially hint at the discriminatory notes against people of non-European origins which was outlined by Fernandez Macias (2018), however, given the persistently low R-squared and adjusted R-squared (23% and 41% respectively), such interpretation seems rather too far-fetched.

Table 5: Coefficients output for Model 4.

```
## Call:
## glm(formula = X4$uempla ~ ., family = binomial(link = "logit"),
##     data = X4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6281  -0.5812  -0.1521  -0.0415   4.0754
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.772e+01  1.628e+03  -0.011   0.9913
## rlgdgr10      1.511e+00  7.370e-01   2.050   0.0404 *
## hhmmb2        1.227e+00  7.080e-01   1.734   0.0830 .
## hhmmb3        1.386e+00  6.164e-01   2.249   0.0245 *
## hhmmb5        1.269e+00  6.798e-01   1.867   0.0618 .
## hhmmb6        2.237e+00  9.475e-01   2.361   0.0182 *
## hhmmb7        3.843e+00  1.720e+00   2.234   0.0255 *
## agea         -1.102e-01  3.808e-02  -2.895   0.0038 **
## domicil4      9.190e-01  5.147e-01   1.785   0.0742 .
## edctn1       -5.674e+00  1.113e+00  -5.097 3.45e-07 ***
## hswrk1       -3.737e+00  1.530e+00  -2.442   0.0146 *
## lvgptnea6    -1.786e+00  7.836e-01  -2.279   0.0227 *
## edulvlbLOW1   1.024e+00  4.679e-01   2.189   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.03  on 606  degrees of freedom
## Residual deviance: 369.72  on 558  degrees of freedom
## AIC: 467.72
##
## Number of Fisher Scoring iterations: 17
##          llh      llhNull           G2     McFadden         r2ML         r2CU
## -184.8605659 -265.0146892  160.3082465    0.3024516    0.2321027    0.3985383
```

Table 6: ANOVA output for Model 4

```
## Analysis of Deviance Table (Type II tests)
##
## Response: X4$uempla
##             LR Chisq Df Pr(>Chisq)
## dscrnap        0.213  1  0.6444502
## rlgdgr        14.240 12  0.2856351
## hhmmb         13.110  8  0.1081295
## gndr           0.174  1  0.6764391
## agea           8.715  1  0.0031555 **
## domicil        6.515  4  0.1638683
## edctn         76.865  1  < 2.2e-16 ***
## eduyrs         0.025  1  0.8736340
## hswrk         12.026  1  0.0005246 ***
## lvgptnea       6.470  4  0.1666744
## maritalb       5.052  5  0.4095737
## lnghom1ITA     4.527  1  0.0333663 *
## lnghom1DIAL    2.944  1  0.0861841 .
## lnghom1EURO    2.802  1  0.0941649 .
## edulvlbLOW     4.805  1  0.0283849 *
## edulvlbHIGH    0.100  1  0.7517339
```

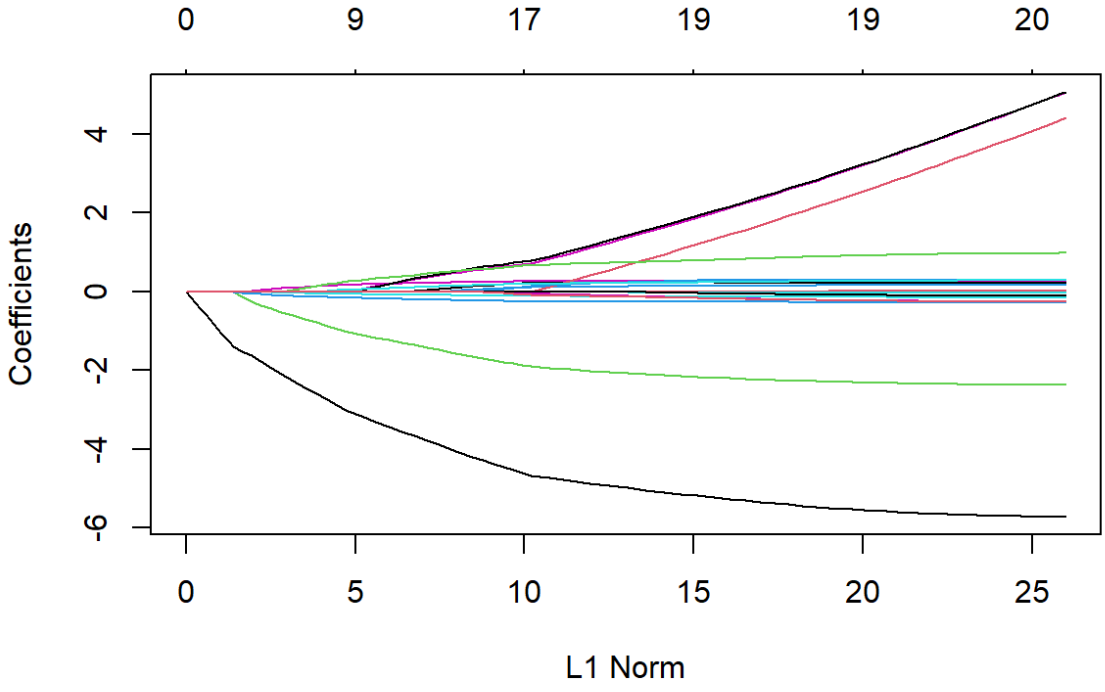
```
## edulvlfbLOW    0.044  1  0.8340451
## edulvlfbIGH    0.015  1  0.9016370
## edulvlmbLOW    0.020  1  0.8862428
## edulvlmbIGH    0.011  1  0.9166339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.7. *Lasso logistic regression*

As both of the converged models issued warnings regarding the presence of fitted probabilities numerically 0 or 1, indicated the need for exploring other statistical methods, as well as the possibility that another type of regression analysis would suit the data better. Unlike ridge regression, lasso regression forces the more redundant coefficients to converge to 0, thus allowing us to select only the most meaningful variables for further consideration (James, Witten, Hastie & Tibshirani, 2015).

The lasso model was created using the latest version of the dataset, employed for the creation of Model 4. As the result, none of the coefficients were shrunk to 0, implying that all of them had some level of significance for the model. However, there was a large number of variables showing borderline low insignificance (straight lines parallel to the x-axis depicted in Figure 5), which was also showcased in the previous models. Therefore, the lasso model failed to exclude any of the variables.

Figure 5: Lasso Regression results plot



Lastly, as the literature reviewed in Chapter 1 suggested that men had on average higher chances of being employed compared to women, in order to control for any differences in outcome related to the gender, a Chi-squared test was performed specifically on the variables *gndr*, which is a variable for gender (either man or woman), and *uempla*, the categorical variable for describing whether the respondent is unemployed or not. This statistical method allows to understand whether the two categorical variables are independent of one another (Lane, n.d.). Although ANOVA results captured the essence of most variables' relationships, an additional check is conducted on the specific variables in order to rule out inconsistencies related to the amount of variables in the models. Similarly, the Chi-squared test of relationships betweel *uempla* and variables related to foreign origins (*cntbrthc*) and discrimination based on non-Italian background (*dscrntn*, *dscrrce*) was performed, as the literature review pointed out that Italian residents with foreign background have harder time searching for a job compared to the natives.

Despite the theoretical foundation, the individual Chi-squared results repeated the outcome of the previous statistical methods, showing no particular significance between the above-mentioned variables and the employment status. Therefore, according to the outcomes of the statistical methods described above, the dataset does not exhibit strong relationships between factors suggested by the examined literature.

At this point, the above-mentioned approaches started to seem rather repetitive, which inspired the search for alternative analytical methods. The next chapter, therefore, outlines the use of a Random forest classification algorithm.

Chapter 3: Random Forest classification.

This section aims to describe the process of building random forests on the dataset deployed for the final logistic regression model, composed of 20 predictors and the target *uempla*. As previously mentioned, the rationale behind relying on classification trees and random forests is gaining a different view on which factors are relevant in dividing the employed from the unemployed in our sample. To achieve this, the metric we are going to observe is feature importance, which will be used to rank variables from least to most relevant in dividing the observations so that the obtained clusters are as different as possible between them and as homogeneous as possible within.

3.1 Random Forest parameter tuning.

The library used to perform classification with random forests was *randomForests* (Liaw, 2002), based on the article "Random Forests" published by L. Breiman in issue 45 of *Machine Learning* in 2001.The parameters which we are going to tune in our study are:

- *mtry*: the number of variables considered for each tree
- *ntrees*: the number of trees generated

Before performing the modeling, the dataset was split into the training and testing sub-sets at the rates of 80 and 20 respectively. The following are the combinations that were considered.

Table 7: Random Forests models summary

	mtry	*ntrees*	***accuracy***
Attempt 1	5	50	0.5626374
Attempt 2	10	50	0.5846154
Attempt 3	19	50	0.5846154
Attempt 4	10	100	0.578022
Attempt 5	10	500	0.5802198

4.2 Discussion of the model results

The model accuracy is used to compare the performance of the model on our sample. The accuracy is the *test accuracy*, as obtained by computing the percentage of true positive and true negative guesses of the algorithms on a test set accounting for the 20% of the total population. Looking at Table 7, we immediately notice that, by increasing the number of factors considered by each tree from 5 (*n*/4, being *n* the number of variables in the model) to 5 (*n*/2) and keeping *ntree* unchanged, the accuracy increased by circa 2%. The maximum value at which we can set *mtry* is 19, although doing so does not increase the accuracy of the model. Hence, we decide to keep the number of trees to 10, thus favoring the generation of less correlated trees (Breiman, 2001).

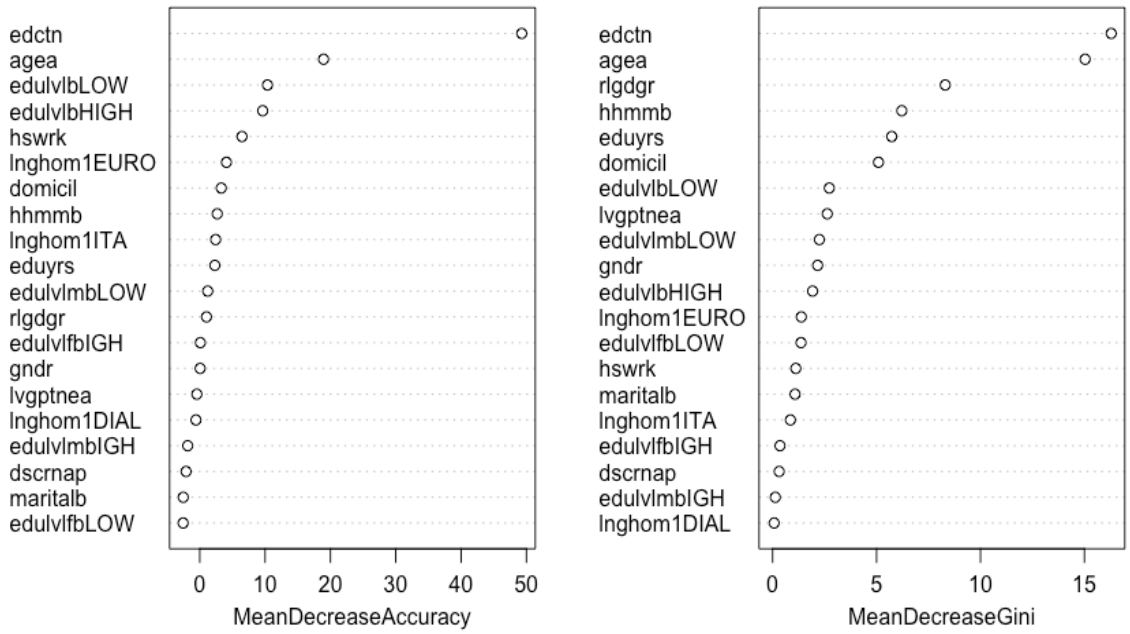
By increasing the number of trees from 50 to 100 we notice a decrease in the model performance from 58.46% to 57.8%. The accuracy also decreases when setting the number of trees to 500. Hence, we decide that *Attempt 2* is the model which performs best on the data at disposal.

Table 8: Confusion Matrix

Actual/Guess	*employed*	*unemployed*	***class error***
employed	61	15	0.1973684
unemployed	10	66	0.1315789

From Table 8 above, we can infer that the class errors of the model are 19.74% and 13.16% for the unemployed and the employed, respectively, meaning that random forests performs slightly better at classifying the employed as such.

Figure 6: Feature Importance



The graphs above were obtained by deploying the *VarImpPlot* method while using R. The first shows the mean decrease in the model accuracy obtained by randomly permuting one of each variable, while the latter measures the total decrease in node impurity that results from splits over that variable, averaged over all trees (James, Witten, Hastie & Tibshirani, 2015). Even though the two measures of importance are different, if we look at the first five most important variables in both graphs of Figure 6, we immediately notice that both *edctn* and *agea* seem to play an important role in classifying subjects as belonging to one group rather than the other. Additionally, strong religious affiliation, a high number of people in the household, engagement in housework, as well as domicile structure seem to possess higher significance than the other parameters as well. In particular, the fact that a subject has been enrolled in education during the past seven days seems to have a higher impact compared to the other features.

To validate the results, we shall investigate the relationship with the target *uempla*. The result is indeed compelling, as Figure 7 indicates that there is a high number of individuals who demonstrate enrollment into some unspecified mode of education while not being classified as unemployed and not interested in looking for a job (as everyone responded belonging to that group was eliminated at the very beginning of statistical analysis). The similar observation was also spotted earlier during the analysis of the logistic

regression model outcomes. Although slightly confusing, the variable shall nonetheless remain present in the analysis as it did not meet any of the previously established criteria for exclusion.

Figure 7: target VS *edctn* variable

target/edctn	*enrolled in education in past 7 days*	*otherwise*
employed	319	192
unemployed	95	1

Study limitations

Logically, the results of the statistical analysis cannot be considered obsolete due to the presence of significant limitations. Firstly, the dataset contained real-world data with quite a lot of missing variables potentially due to human error and misunderstanding among respondents, which certainly affected the analysis. Moreover, the sample size should have preferably been larger as the results are theoretically extrapolated onto the entire Italian population. Due to the need to exclude certain variables solely based on the lack of data consistency, the total number of observations in the models was reduced to around 600 respondents, which is estimated to have a rather influential margin of error – approximately around 4%. Moreover, as mentioned at the beginning of Chapter 2, the ESS weights were excluded for the fear of them meddling with glm coefficients, however, this might have taken a tall on the accuracy of the correlations and whether they can be inferred on a wider Italian population of young people.

Secondly, the data fitting is not expected to be high due to the initial selection of variables for analysis. It is clear that the socio-demographic conditions alone cannot explain respondents' employment or unemployment, and this study is not aimed at accounting for all of the possible aspects that can influence the dependent variable. In fact, without economic-oriented data such as, for instance, GDP fluctuations, the results are very unlikely to be complete. As the initial goal of the study was to conduct inferential analysis and see if the produced model results demonstrate any significant correlations between any of the variables in considered combinations, it has overall been fulfilled.

Lastly, the dataset is merged from two rounds of ESS, which could potentially add some inconsistencies to the responses, as the answers were recorded two years apart from each other (in 2016 and 2018). So there is a plausibility of the difference in context interfering with the results, even thought the offered questions and answers did not change over the years. This decision to merge results from two different rounds was mainly inspired by the interest of having a larger dataset of values related solely to younger age cohorts of 15 to 34 y.o. to help reduce the margin of error.

Conclusions

Overall, the carried out statistical analysis has produced quite interesting outcomes. Firstly, given the concerns over the validity of logistic regression model results, the decision to use the classification method of the random forest has proven to be worthy. The results of the last glm model were largely intact with the outcomes of the random forest and they highlighted almost the same set of variables: involvement in education, low level of highest achieved education, age, household composition, housework, and type of domicile. Additionally, extremely high dedication to religion also showed to be significant. Consequently, an increase in age and involvement in education would suggest a decrease in the number of unemployed respondents who are actively looking for a job. The education matter can be interpreted in two ways – either moving from job seeking to education serves as a way of exclusion from NEET or one's involvement in additional training increases one's chances to attain employment of some sort. For instance, it could be a temporary contract typically issued for interns. On the contrary, an increase in the number of members of the household, the need to perform house chores, as well as residing in a location described as a "village" is likely to negatively affect the odds of a respondent securing employment.

Surprisingly, among the variables which could be attributed to being a characteristic of a minority, such as age, foreign origins, experiencing discrimination, as well as low income and low level of achieved education, only the latter was found significant. Ironically, not having a high school diploma positively contributed to the odds of finding a job. It could be linked to the profile of respondents who were reportedly residing in rural farm areas and joined the workforce early without the intention to continue education and instead focused on their career. Nonetheless, it is rather reassuring to not see clear discriminatory trends in the analysis result with respect to the minority groups which were the main focus of the study.

Certainly, the results have to be treated cautiously as the above-discussed limitations are quite significant. Namely, the R-squared coefficient has remained around 20% for all of the model iterations, meaning that a large portion of the relationships between variables cannot be explained solely by the considered values. Although it was expected at the beginning of the study and outlined in Chapter 2, a low R-square value still indicates potential issues with data fit and hence has to be accounted for when interpreting the results. The random forest

classification also showed an error percentage as high as 13, as well as the highest accuracy of 58%, yet is still away from being perfect.

To conclude, the literature analysis outlines the basis of the study by introducing the main challenges Italy faces concerning youth unemployment, such as the gender gap, hard barriers the way to immigrants' integration into the labor force, the divide between opportunities in the North and the South of the country, as well as generally unfavorable employment conditions like the small likelihood of obtaining a stable contract and livable pay in earlier stages of one's career.

What hampers the process of diminishing the youth unemployment rate apart from global economic and political events which continuously challenge the Italian economy, are the poor educational background and overall lack of motivation to pursue tertiary education together with ineffective and scarce policy-making and unfavorable welfare conditions. The statistical analysis confirms the higher chances of one getting a job at an older age and with better education background, while the results do not show clear correlation patterns between gender and ethnic profile and employment. The significant variables rather paint a profile of someone who is in an initially disadvantaged economic situation and is forced to sacrifice education attainment for supporting their household. Although one might speculate that some minorities find themselves in such conditions more often, as women are often expected to take care of house chores and babysitting, or families with immigrant backgrounds tend to reside in larger households located in rural areas for affordability reasons, there is no statistical evidence in the above-discussed results.

Bibliography

1. Azzolini, D., Schnell, P., & Palmer, J. (2012). Educational Achievement Gaps between Immigrant and Native Students in Two "New" Immigration Countries. The ANNALS Of The American Academy Of Political And Social Science, 643(1), 46-77. doi: 10.1177/0002716212441590
2. Bell, D., & Blanchflower, D. (2010). Youth Unemployment: Déjà Vu?. SSRN Electronic Journal. doi: 10.2139/ssrn.1545132
3. Bird, K. (2004). Analysis of Variance via Confidence Intervals [Ebook]. SAGE Publications Ltd. Retrieved from https://uk.sagepub.com/sites/default/files/upm-assets/9407_book_item_9407.pdf
4. Boeri, T., & Garibaldi, P. (2007). Two Tier Reforms of Employment Protection: A Honeymoon Effect?. The Economic Journal, 117(521), F357-F385. doi: 10.1111/j.1468-0297.2007.02060.x
5. Bradley, S., Migali, G., & Navarro Paniagua, M. (2020). Spatial variations and clustering in the rates of youth unemployment and NEET: A comparative analysis of Italy, Spain, and the UK. Journal Of Regional Science, 60(5), 1074-1107. doi: 10.1111/jors.12501
6. Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. - References - Scientific Research Publishing. (2001). Retrieved 7 September 2022, from https://www.scirp.org/reference/ReferencesPapers.aspx?ReferenceID=2597478
7. Cannari, L., Gambacorta, R., & D'Alessio, G. (2008). Capital Gains and Wealth Distribution in Italy. SSRN Electronic Journal. doi: 10.2139/ssrn.1182603
8. Cirillo, V., Fana, M., & Guarascio, D. (2017). Labour market reforms in Italy: evaluating the effects of the Jobs Act. Economia Politica, 34(2), 211-232. doi: 10.1007/s40888-017-0058-2
9. ESS. (2022). ESS Codebook, ESS9-2018 [Ebook]. Retrieved from https://stessrelpubprodwe.blob.core.windows.net/data/round9/survey/ESS9_appendix_a7_e03_1.pdf
10. Fernandez Macias, E. (2018). Labour market integration of migrants and their descendants. Eurofound.
11. Glossary | DataBank. Retrieved 22 August 2022, from https://databank.worldbank.org/metadataglossary/jobs/series/SL.UEM.1524.ZS#:~:text=Unemployment%2C%20youth%20total%20(%25%20of,International%20Labour%20Organization%2C%20ILOSTAT%20database.
12. Italy: employment by economic sector | Statista. (2022). Retrieved 22 August 2022, from https://www.statista.com/statistics/586899/employment-by-economic-sector-italy/
13. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2015). An introduction to statistical learning.
14. Lane, D. Chi Square Distribution. Retrieved 2 October 2022, from https://onlinestatbook.com/2/chi_square/distribution.html
15. Lavoro: grafici. (2022). Retrieved 26 August 2022, from https://www4.istat.it/it/giovani/lavoro/grafici

16. Liaw, A. (2002). Classification and Regression by randomForest. Retrieved 6 September 2022, from https://cran.r-project.org/doc/Rnews/
17. Lucidi, F., & Kleinknecht, A. (2009). Little innovation, many jobs: An econometric analysis of the Italian labour productivity crisis. Cambridge Journal Of Economics, 34(3), 525-546. doi: 10.1093/cje/bep011
18. Maitino, M. (2022). Employment effects of Reddito di cittadinanza, before and during the Covid-19 pandemic. IRPET, 2022(6). Retrieved from http://www.irpet.it/wp-content/uploads/2022/05/working-paper-6_2022-maggio-ravagli-et-al-1.pdf
19. Malgarini, M., Mancini, M., & Pacelli, L. (2013). Temporary hires and innovative investments. Applied Economics, 45(17), 2361-2370. doi: 10.1080/00036846.2012.663477
20. Montanari, M., Pinelli, D., & Torre, R. (2015). From tertiary education to work in Italy: a difficult transition. ECFIN Country Focus, 12(5).
21. Musolino, D. (2018). The North-South Divide in Italy: Reality or Perception?. European Spatial Research And Policy, 25(1), 29-53. doi: 10.18778/1231-1952.25.1.03
22. Pacifico, D., Browne, J., Fernandez, R., Immervoll, H., Neumann, D., & Thévenot, C. (2018). Faces of Joblessness in Italy: A People-Centred Perspective on Employment Barriers and Policies. SSRN Electronic Journal. doi: 10.2139/ssrn.3249882
23. Paolazzi, L. (2017). Le sfide della politica economica [Ebook]. Rome: CENTRO STUDI CONFINDUSTRIA. Retrieved from https://www.confindustria.it/wcm/connect/96f0b50b-1f13-4d75-8df1-1ba99a8f7c91/Scenari+Economici+30+-+web.pdf?MOD=AJPERES&CONVERT_TO=url&CACHEID=ROOTWORKSPACE-96f0b50b-1f13-4d75-8df1-1ba99a8f7c91-mtAIwj3
24. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models. Augmented Human Research, 5(1). doi: 10.1007/s41133-020-00032-0
25. Scaturro, R. (2021). Modern Slavery Made in Italy—Causes and Consequences of Labour Exploitation in the Italian Agricultural Sector. Journal Of Illicit Economies And Development, 3(2), 181-189. doi: 10.31389/jied.95
26. Unemployment statistics at regional level - Statistics Explained. (2021). Retrieved 22 August 2022, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Unemployment_statistics_at_regional_level#Regional_variations_in_youth_unemployment

Appendix

The code below is extracted using R Markdown.

Import data

```
library(openxlsx)
library(readxl)
library(pscl)

## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(carData)

ess = read_xls("final ess.xls")
```

Build target

To build target we use:

- uempla 1 unemployed and looking for a job 0 otherwise filter by uempli 1 unemployed and NOT looking for a job, 0 otherwise

```
# eliminate unemployed people not actively looking for a job
X = ess
X = X %>% filter(uempli == "0")

table(X$uempla)

##
##    0    1
## 1091  185

colSums(is.na(X))

##          essround           edition              idno             cntry
##                 0                 0                 0                 0
##           dweight           pspwght           pweight           aesfdrk
##                 0                 0                 0                 0
##           blgetmg          cntbrthc          ctzshipc           dscrage
##                 0               611               611                 0
##           dscrdsb           dscretn           dscrgnd           dscrgrp
##                 0                 0                 0                 0
##           dscrnap           dscrntn           dscrrce           dscrsex
##                 0                 0                 0                 0
##             happy            health           hlthhmp          livecnta
##                 0                 0                 0                 0
```

```
##            lnghom1            lnghom2             rlgdgr             rlgdnm
##                  0                  0                  0                  0
##              hhmmb               gndr            rshipa2            rshipa3
##                  0                  0                  0                  0
##               agea            anctry1            domicil              dsbld
##                  0                  0                  0                  0
##              edctn            edulvlb           edulvlfb           edulvlmb
##                  0                  0                  0                  0
##           edulvlpb             eduyrs            emplrel              hswrk
##                  0                  0                  0                  0
##             isco08           lvgptnea           maritalb            mnactic
##                  0                  0                  0                  0
##            pdjobev              pdwrk             pdwrkp            tporgwk
##                  0                  0                  0                  0
##            uemp12m             uemp3m             uempla             uempli
##                  0                  0                  0                  0
##              wkhct            wrkac6m            wrkctra             nacer2
##                  0                  0                  0                  0
##              yrbrn new_vars.hincsrca new_vars.hinctnta
##                  0                  0                  0
```

```r
#eliminate cuz too many are missing
X = subset(X, select = - c(cntbrthc, ctzshipc))

# eliminate variables used to build the target / similar to target
X = subset(X, select = - c(dscrdsb, happy,  health, wrkctra, wkhct, uemp12m, uemp3m, uem
pli, livecnta, rshipa2, rshipa3, anctry1, edulvlpb, yrbrn))

# eliminate non useful / redundant variables
X = subset(X, select = - c(essround, edition, idno, cntry, dweight, pspwght, pweight, ln
ghom2, isco08, nacer2, rlgdnm, emplrel, pdjobev, pdwrk, tporgwk, wrkac6m))

# make sure binary variables include 0, 1
X$gndr = X$gndr - 1

# eliminate subjects who identify as disabled as there are only two
X = X %>% filter(dsbld == 0)
X = subset(X, select = - c(dsbld))
dim(X)
```

```
## [1] 1271   30
```

```r
# check target
table(X$uempla)
```

```
##
##    0    1
## 1086  185
```

Getting rid of redundant variables

```r
# handle missing values
# create auxiliary dataframe
X_nonan = X
X_nonan = X_nonan %>% filter((lnghom1 != "777" & lnghom1 != "999" & lnghom1 != "UND"))
X_nonan = X_nonan %>% filter((rlgdgr != "77" | rlgdgr != "88"))
X_nonan = X_nonan %>% filter((maritalb != "77"))
X_nonan = X_nonan %>% filter((edulvlfb != "7777" | edulvlfb != "8888"))
X_nonan = X_nonan %>% filter((edulvlmb != "7777" | edulvlmb != "8888"))
X_nonan = X_nonan %>% filter((eduyrs != "88"))
X_nonan = X_nonan %>% filter((new_vars.hinctnta != "77" & new_vars.hinctnta != "88"))

# check if deleting the missing values led to a significant loss of observations
dim(X_nonan)
```

```
## [1] 607  30

table(X$uempla, X$edctn)

##
##       0   1
##   0 578 508
##   1 184   1

# remove missing values from main dataframe
X = X_nonan
# factorize
to_factorize = names(X) [! names(X) %in% c("agea", "eduyrs")]
X[,to_factorize] = lapply(X[,to_factorize],as.factor)

# print a summary view
summary(X)

##  aesfdrk blgetmg dscrage dscretn dscrgnd dscrgrp dscrnap dscrntn dscrrce
##  1:131   1: 30   0:603   0:602   0:604   1: 32   0: 32   0:593   0:602
##  2:328   2:570   1:  4   1:  5   1:  3   2:565   1:575   1: 14   1:  5
##  3:107   7:  1                           7:  3
##  4: 38   8:  6                           8:  7
##  7:  1
##  8:  2
##
##  dscrsex hlthhmp    lnghom1        rlgdgr        hhmmb     gndr
##  0:604   1:  3   ITA    :553   7      : 90   3      :200   0:323
##  1:  3   2: 11   RUM    : 14   5      : 85   4      :192   1:284
##          3:593   QAA    :  8   6      : 83   2      : 69
##                  ALB    :  6   0      : 71   5      : 68
##                  ARA    :  6   8      : 64   1      : 49
##                  GER    :  5   4      : 46   6      : 17
##                  (Other): 15   (Other):168   (Other): 12
##       agea       domicil edctn      edulvlb       edulvlfb      edulvlmb
##  Min.   :15.00   1: 58   0:414   213    :171   213    :242   213    :229
##  1st Qu.:21.00   2: 36   1:193   313    :164   323    : 83   313    : 99
##  Median :25.00   3:210           323    :121   113    : 81   113    : 82
##  Mean   :25.33   4:285           620    : 41   313    : 64   323    : 69
##  3rd Qu.:30.00   5: 18           321    : 38   720    : 40   720    : 34
##  Max.   :34.00                   720    : 33   321    : 39   321    : 31
##                                  (Other): 39   (Other): 58   (Other): 63
##      eduyrs      hswrk   lvgptnea maritalb mnactic  pdwrkp  uempla
##  Min.   : 3.00   0:578   1: 62    1: 97    1 :288   0:508   0:511
##  1st Qu.:11.00   1: 29   2:502    2:  8    2 :192   1: 99   1: 96
##  Median :13.00           6: 41    3:  3    3 : 96
##  Mean   :13.15           7:  1    4:  2    6 :  1
##  3rd Qu.:15.00           8:  1    5:  1    8 : 27
##  Max.   :24.00                    6:496    9 :  2
##                                            77:  1
##  new_vars.hincsrca new_vars.hinctnta
##  1      :446       4      : 95
##  2      : 65       3      : 91
##  4      : 34       7      : 75
##  3      : 18       5      : 68
##  8      : 16       2      : 67
##  5      : 10       8      : 60
##  (Other): 18       (Other):151
```

First attempt - keep all categories for all the variables

```
# fit the model
mod = glm(X$uempla ~ ., family = binomial(link = "logit"), data = X)
```

```
## Warning: glm.fit: algorithm did not converge

# print summary views
summary(mod)

##
## Call:
## glm(formula = X$uempla ~ ., family = binomial(link = "logit"),
##     data = X)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06   2.409e-06
##
## Coefficients: (3 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.657e+01  5.090e+05   0.000    1.000
## aesfdrk2            -2.569e-11  4.258e+04   0.000    1.000
## aesfdrk3            -3.770e-11  5.620e+04   0.000    1.000
## aesfdrk4            -1.148e-12  7.599e+04   0.000    1.000
## aesfdrk7             9.512e-11  3.729e+05   0.000    1.000
## aesfdrk8             3.895e-10  2.700e+05   0.000    1.000
## blgetmg2            -1.148e-11  1.265e+05   0.000    1.000
## blgetmg7             2.498e-10  5.550e+05   0.000    1.000
## blgetmg8            -5.861e-11  2.177e+05   0.000    1.000
## dscrage1            -1.168e-10  2.424e+05   0.000    1.000
## dscretn1             3.190e-11  2.405e+05   0.000    1.000
## dscrgnd1            -4.914e-11  3.256e+05   0.000    1.000
## dscrgrp2            -5.193e-11  1.272e+05   0.000    1.000
## dscrgrp7            -2.383e-10  2.664e+05   0.000    1.000
## dscrgrp8            -4.249e-11  2.250e+05   0.000    1.000
## dscrnap1                    NA         NA      NA       NA
## dscrntn1            -6.247e-11  2.095e+05   0.000    1.000
## dscrrce1             1.117e-10  3.016e+05   0.000    1.000
## dscrsex1            -4.646e-11  2.842e+05   0.000    1.000
## hlthhmp2            -1.981e-11  3.609e+05   0.000    1.000
## hlthhmp3            -1.206e-10  3.272e+05   0.000    1.000
## lnghom1ARA           6.667e-11  2.816e+05   0.000    1.000
## lnghom1BEN          -2.003e-10  7.358e+05   0.000    1.000
## lnghom1FRE           7.298e-11  4.142e+05   0.000    1.000
## lnghom1GER           7.368e-11  2.563e+05   0.000    1.000
## lnghom1HIN          -1.129e-10  4.815e+05   0.000    1.000
## lnghom1ITA           3.492e-11  1.910e+05   0.000    1.000
## lnghom1QAA           1.089e-10  2.397e+05   0.000    1.000
## lnghom1RUM           1.184e-10  2.149e+05   0.000    1.000
## lnghom1SCN          -3.432e-10  3.007e+05   0.000    1.000
## lnghom1SPA           1.208e-10  2.757e+05   0.000    1.000
## lnghom1SRD          -1.082e-10  5.793e+05   0.000    1.000
## lnghom1ZGH           3.099e-11  3.374e+05   0.000    1.000
## rlgdgr1              4.509e-11  8.747e+04   0.000    1.000
## rlgdgr2              4.923e-11  7.779e+04   0.000    1.000
## rlgdgr3              2.030e-11  7.673e+04   0.000    1.000
## rlgdgr4              2.313e-11  7.736e+04   0.000    1.000
## rlgdgr5              2.561e-11  6.632e+04   0.000    1.000
## rlgdgr6              2.011e-11  6.689e+04   0.000    1.000
## rlgdgr7              5.062e-11  6.536e+04   0.000    1.000
## rlgdgr8              2.949e-11  7.413e+04   0.000    1.000
## rlgdgr9              5.451e-11  1.099e+05   0.000    1.000
## rlgdgr10             7.983e-11  1.015e+05   0.000    1.000
## rlgdgr77             8.856e-11  2.974e+05   0.000    1.000
## rlgdgr88            -4.548e-11  2.256e+05   0.000    1.000
## hhmmb2              -6.488e-11  8.400e+04   0.000    1.000
## hhmmb3              -1.891e-11  7.303e+04   0.000    1.000
```

```
## hhmmb4              -3.344e-12  7.526e+04   0.000    1.000
## hhmmb5              -1.947e-11  8.353e+04   0.000    1.000
## hhmmb6               3.989e-11  1.196e+05   0.000    1.000
## hhmmb7              -7.548e-11  2.042e+05   0.000    1.000
## hhmmb8               6.602e-12  3.250e+05   0.000    1.000
## hhmmb77              1.652e-11  2.169e+05   0.000    1.000
## gndr1                1.574e-11  3.768e+04   0.000    1.000
## agea                 2.416e-14  4.997e+03   0.000    1.000
## domicil2            -1.967e-11  8.496e+04   0.000    1.000
## domicil3            -2.183e-11  6.020e+04   0.000    1.000
## domicil4            -4.442e-11  5.939e+04   0.000    1.000
## domicil5            -1.512e-10  1.110e+05   0.000    1.000
## edctn1              -2.788e-09  3.850e+05   0.000    1.000
## edulvlb213          -2.112e-10  2.040e+05   0.000    1.000
## edulvlb222           3.778e-11  4.529e+05   0.000    1.000
## edulvlb229          -2.775e-10  2.265e+05   0.000    1.000
## edulvlb313          -2.587e-10  2.130e+05   0.000    1.000
## edulvlb321          -2.439e-10  2.153e+05   0.000    1.000
## edulvlb323          -2.666e-10  2.113e+05   0.000    1.000
## edulvlb421          -1.491e-10  3.176e+05   0.000    1.000
## edulvlb520          -2.074e-10  2.670e+05   0.000    1.000
## edulvlb620          -2.664e-10  2.274e+05   0.000    1.000
## edulvlb710          -2.993e-10  2.721e+05   0.000    1.000
## edulvlb720          -2.704e-10  2.362e+05   0.000    1.000
## edulvlb800          -4.039e-10  4.684e+05   0.000    1.000
## edulvlb5555         -2.421e-10  4.407e+05   0.000    1.000
## edulvlb7777         -1.458e-10  4.235e+05   0.000    1.000
## edulvlfb113          3.367e-11  1.978e+05   0.000    1.000
## edulvlfb213          9.765e-11  1.988e+05   0.000    1.000
## edulvlfb222          6.897e-11  2.445e+05   0.000    1.000
## edulvlfb229          4.443e-11  2.501e+05   0.000    1.000
## edulvlfb313          1.160e-10  2.044e+05   0.000    1.000
## edulvlfb321          7.709e-11  2.100e+05   0.000    1.000
## edulvlfb323          7.425e-11  2.028e+05   0.000    1.000
## edulvlfb421         -6.762e-11  4.850e+05   0.000    1.000
## edulvlfb510         -3.130e-11  4.874e+05   0.000    1.000
## edulvlfb520          1.266e-10  2.715e+05   0.000    1.000
## edulvlfb620          5.737e-12  2.649e+05   0.000    1.000
## edulvlfb710          1.414e-10  3.389e+05   0.000    1.000
## edulvlfb720          8.324e-11  2.106e+05   0.000    1.000
## edulvlfb800          3.526e-10  3.042e+05   0.000    1.000
## edulvlfb5555        -7.005e-11  4.958e+05   0.000    1.000
## edulvlfb7777         1.114e-11  3.972e+05   0.000    1.000
## edulvlfb8888         1.300e-10  2.311e+05   0.000    1.000
## edulvlmb113          1.626e-10  1.416e+05   0.000    1.000
## edulvlmb213          1.998e-10  1.444e+05   0.000    1.000
## edulvlmb222          2.141e-10  2.535e+05   0.000    1.000
## edulvlmb229          4.281e-11  2.124e+05   0.000    1.000
## edulvlmb313          1.916e-10  1.498e+05   0.000    1.000
## edulvlmb321          2.099e-10  1.605e+05   0.000    1.000
## edulvlmb323          2.098e-10  1.527e+05   0.000    1.000
## edulvlmb421          2.145e-10  3.186e+05   0.000    1.000
## edulvlmb520          2.048e-10  2.422e+05   0.000    1.000
## edulvlmb620          1.674e-10  1.890e+05   0.000    1.000
## edulvlmb720          2.022e-10  1.626e+05   0.000    1.000
## edulvlmb800          3.300e-10  2.501e+05   0.000    1.000
## edulvlmb5555        -4.624e-06  4.916e+05   0.000    1.000
## edulvlmb7777         2.209e-10  2.735e+05   0.000    1.000
## edulvlmb8888        -2.470e-11  2.330e+05   0.000    1.000
## eduyrs               2.376e-12  9.500e+03   0.000    1.000
## hswrk1              -5.174e-10  3.257e+05   0.000    1.000
## lvgptnea2           -3.188e-11  6.076e+04   0.000    1.000
## lvgptnea6           -6.051e-12  1.113e+05   0.000    1.000
```

```
## lvgptnea7           -3.682e-11  4.194e+05   0.000    1.000
## lvgptnea8            5.456e-11  3.881e+05   0.000    1.000
## maritalb2            6.078e-11  1.752e+05   0.000    1.000
## maritalb3           -4.521e-10  2.644e+05   0.000    1.000
## maritalb4           -9.407e-11  3.324e+05   0.000    1.000
## maritalb5            2.614e-11  4.689e+05   0.000    1.000
## maritalb6           -6.914e-11  7.820e+04   0.000    1.000
## mnactic2             2.763e-09  3.916e+05   0.000    1.000
## mnactic3             5.313e+01  5.219e+04   0.001    0.999
## mnactic6                    NA         NA      NA       NA
## mnactic8             4.933e-10  3.343e+05   0.000    1.000
## mnactic9             3.965e-11  2.680e+05   0.000    1.000
## mnactic77           -2.150e-10  5.401e+05   0.000    1.000
## pdwrkp1              9.501e-13  8.589e+04   0.000    1.000
## new_vars.hincsrca2  -9.570e-12  5.412e+04   0.000    1.000
## new_vars.hincsrca3  -3.821e-11  9.840e+04   0.000    1.000
## new_vars.hincsrca4  -8.443e-11  7.567e+04   0.000    1.000
## new_vars.hincsrca5   6.043e-10  1.336e+05   0.000    1.000
## new_vars.hincsrca6          NA         NA      NA       NA
## new_vars.hincsrca7   3.543e-11  1.968e+05   0.000    1.000
## new_vars.hincsrca8   2.804e-10  1.115e+05   0.000    1.000
## new_vars.hincsrca77 -2.078e-13  3.006e+05   0.000    1.000
## new_vars.hincsrca88  5.407e-11  1.409e+05   0.000    1.000
## new_vars.hinctnta2  -1.724e-10  8.196e+04   0.000    1.000
## new_vars.hinctnta3  -1.783e-10  7.784e+04   0.000    1.000
## new_vars.hinctnta4  -1.596e-10  7.873e+04   0.000    1.000
## new_vars.hinctnta5  -1.435e-10  8.308e+04   0.000    1.000
## new_vars.hinctnta6  -1.096e-10  8.969e+04   0.000    1.000
## new_vars.hinctnta7  -1.477e-10  8.409e+04   0.000    1.000
## new_vars.hinctnta8  -1.462e-10  8.994e+04   0.000    1.000
## new_vars.hinctnta9  -1.407e-10  1.038e+05   0.000    1.000
## new_vars.hinctnta10 -1.686e-10  1.113e+05   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.3003e+02  on 606  degrees of freedom
## Residual deviance: 3.5216e-09  on 468  degrees of freedom
## AIC: 278
##
## Number of Fisher Scoring iterations: 25

# print summary views
summary.glm(mod)

##
## Call:
## glm(formula = X$uempla ~ ., family = binomial(link = "logit"),
##     data = X)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06   2.409e-06
##
## Coefficients: (3 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.657e+01  5.090e+05   0.000    1.000
## aesfdrk2            -2.569e-11  4.258e+04   0.000    1.000
## aesfdrk3            -3.770e-11  5.620e+04   0.000    1.000
## aesfdrk4            -1.148e-12  7.599e+04   0.000    1.000
## aesfdrk7             9.512e-11  3.729e+05   0.000    1.000
## aesfdrk8             3.895e-10  2.700e+05   0.000    1.000
## blgetmg2            -1.148e-11  1.265e+05   0.000    1.000
## blgetmg7             2.498e-10  5.550e+05   0.000    1.000
```

```
## blgetmg8            -5.861e-11  2.177e+05   0.000    1.000
## dscrage1            -1.168e-10  2.424e+05   0.000    1.000
## dscretn1             3.190e-11  2.405e+05   0.000    1.000
## dscrgnd1            -4.914e-11  3.256e+05   0.000    1.000
## dscrgrp2            -5.193e-11  1.272e+05   0.000    1.000
## dscrgrp7            -2.383e-10  2.664e+05   0.000    1.000
## dscrgrp8            -4.249e-11  2.250e+05   0.000    1.000
## dscrnap1                    NA         NA      NA       NA
## dscrntn1            -6.247e-11  2.095e+05   0.000    1.000
## dscrrce1             1.117e-10  3.016e+05   0.000    1.000
## dscrsex1            -4.646e-11  2.842e+05   0.000    1.000
## hlthhmp2            -1.981e-11  3.609e+05   0.000    1.000
## hlthhmp3            -1.206e-10  3.272e+05   0.000    1.000
## lnghom1ARA           6.667e-11  2.816e+05   0.000    1.000
## lnghom1BEN          -2.003e-10  7.358e+05   0.000    1.000
## lnghom1FRE           7.298e-11  4.142e+05   0.000    1.000
## lnghom1GER           7.368e-11  2.563e+05   0.000    1.000
## lnghom1HIN          -1.129e-10  4.815e+05   0.000    1.000
## lnghom1ITA           3.492e-11  1.910e+05   0.000    1.000
## lnghom1QAA           1.089e-10  2.397e+05   0.000    1.000
## lnghom1RUM           1.184e-10  2.149e+05   0.000    1.000
## lnghom1SCN          -3.432e-10  3.007e+05   0.000    1.000
## lnghom1SPA           1.208e-10  2.757e+05   0.000    1.000
## lnghom1SRD          -1.082e-10  5.793e+05   0.000    1.000
## lnghom1ZGH           3.099e-11  3.374e+05   0.000    1.000
## rlgdgr1              4.509e-11  8.747e+04   0.000    1.000
## rlgdgr2              4.923e-11  7.779e+04   0.000    1.000
## rlgdgr3              2.030e-11  7.673e+04   0.000    1.000
## rlgdgr4              2.313e-11  7.736e+04   0.000    1.000
## rlgdgr5              2.561e-11  6.632e+04   0.000    1.000
## rlgdgr6              2.011e-11  6.689e+04   0.000    1.000
## rlgdgr7              5.062e-11  6.536e+04   0.000    1.000
## rlgdgr8              2.949e-11  7.413e+04   0.000    1.000
## rlgdgr9              5.451e-11  1.099e+05   0.000    1.000
## rlgdgr10             7.983e-11  1.015e+05   0.000    1.000
## rlgdgr77             8.856e-11  2.974e+05   0.000    1.000
## rlgdgr88            -4.548e-11  2.256e+05   0.000    1.000
## hhmmb2              -6.488e-11  8.400e+04   0.000    1.000
## hhmmb3              -1.891e-11  7.303e+04   0.000    1.000
## hhmmb4              -3.344e-12  7.526e+04   0.000    1.000
## hhmmb5              -1.947e-11  8.353e+04   0.000    1.000
## hhmmb6               3.989e-11  1.196e+05   0.000    1.000
## hhmmb7              -7.548e-11  2.042e+05   0.000    1.000
## hhmmb8               6.602e-12  3.250e+05   0.000    1.000
## hhmmb77              1.652e-11  2.169e+05   0.000    1.000
## gndr1                1.574e-11  3.768e+04   0.000    1.000
## agea                 2.416e-14  4.997e+03   0.000    1.000
## domicil2            -1.967e-11  8.496e+04   0.000    1.000
## domicil3            -2.183e-11  6.020e+04   0.000    1.000
## domicil4            -4.442e-11  5.939e+04   0.000    1.000
## domicil5            -1.512e-10  1.110e+05   0.000    1.000
## edctn1              -2.788e-09  3.850e+05   0.000    1.000
## edulvlb213          -2.112e-10  2.040e+05   0.000    1.000
## edulvlb222           3.778e-11  4.529e+05   0.000    1.000
## edulvlb229          -2.775e-10  2.265e+05   0.000    1.000
## edulvlb313          -2.587e-10  2.130e+05   0.000    1.000
## edulvlb321          -2.439e-10  2.153e+05   0.000    1.000
## edulvlb323          -2.666e-10  2.113e+05   0.000    1.000
## edulvlb421          -1.491e-10  3.176e+05   0.000    1.000
## edulvlb520          -2.074e-10  2.670e+05   0.000    1.000
## edulvlb620          -2.664e-10  2.274e+05   0.000    1.000
## edulvlb710          -2.993e-10  2.721e+05   0.000    1.000
## edulvlb720          -2.704e-10  2.362e+05   0.000    1.000
```

```
## edulvlb800          -4.039e-10  4.684e+05   0.000    1.000
## edulvlb5555         -2.421e-10  4.407e+05   0.000    1.000
## edulvlb7777         -1.458e-10  4.235e+05   0.000    1.000
## edulvlfb113          3.367e-11  1.978e+05   0.000    1.000
## edulvlfb213          9.765e-11  1.988e+05   0.000    1.000
## edulvlfb222          6.897e-11  2.445e+05   0.000    1.000
## edulvlfb229          4.443e-11  2.501e+05   0.000    1.000
## edulvlfb313          1.160e-10  2.044e+05   0.000    1.000
## edulvlfb321          7.709e-11  2.100e+05   0.000    1.000
## edulvlfb323          7.425e-11  2.028e+05   0.000    1.000
## edulvlfb421         -6.762e-11  4.850e+05   0.000    1.000
## edulvlfb510         -3.130e-11  4.874e+05   0.000    1.000
## edulvlfb520          1.266e-10  2.715e+05   0.000    1.000
## edulvlfb620          5.737e-12  2.649e+05   0.000    1.000
## edulvlfb710          1.414e-10  3.389e+05   0.000    1.000
## edulvlfb720          8.324e-11  2.106e+05   0.000    1.000
## edulvlfb800          3.526e-10  3.042e+05   0.000    1.000
## edulvlfb5555        -7.005e-11  4.958e+05   0.000    1.000
## edulvlfb7777         1.114e-11  3.972e+05   0.000    1.000
## edulvlfb8888         1.300e-10  2.311e+05   0.000    1.000
## edulvlmb113          1.626e-10  1.416e+05   0.000    1.000
## edulvlmb213          1.998e-10  1.444e+05   0.000    1.000
## edulvlmb222          2.141e-10  2.535e+05   0.000    1.000
## edulvlmb229          4.281e-11  2.124e+05   0.000    1.000
## edulvlmb313          1.916e-10  1.498e+05   0.000    1.000
## edulvlmb321          2.099e-10  1.605e+05   0.000    1.000
## edulvlmb323          2.098e-10  1.527e+05   0.000    1.000
## edulvlmb421          2.145e-10  3.186e+05   0.000    1.000
## edulvlmb520          2.048e-10  2.422e+05   0.000    1.000
## edulvlmb620          1.674e-10  1.890e+05   0.000    1.000
## edulvlmb720          2.022e-10  1.626e+05   0.000    1.000
## edulvlmb800          3.300e-10  2.501e+05   0.000    1.000
## edulvlmb5555        -4.624e-06  4.916e+05   0.000    1.000
## edulvlmb7777         2.209e-10  2.735e+05   0.000    1.000
## edulvlmb8888        -2.470e-11  2.330e+05   0.000    1.000
## eduyrs               2.376e-12  9.500e+03   0.000    1.000
## hswrk1              -5.174e-10  3.257e+05   0.000    1.000
## lvgptnea2           -3.188e-11  6.076e+04   0.000    1.000
## lvgptnea6           -6.051e-12  1.113e+05   0.000    1.000
## lvgptnea7           -3.682e-11  4.194e+05   0.000    1.000
## lvgptnea8            5.456e-11  3.881e+05   0.000    1.000
## maritalb2            6.078e-11  1.752e+05   0.000    1.000
## maritalb3           -4.521e-10  2.644e+05   0.000    1.000
## maritalb4           -9.407e-11  3.324e+05   0.000    1.000
## maritalb5            2.614e-11  4.689e+05   0.000    1.000
## maritalb6           -6.914e-11  7.820e+04   0.000    1.000
## mnactic2             2.763e-09  3.916e+05   0.000    1.000
## mnactic3             5.313e+01  5.219e+04   0.001    0.999
## mnactic6                    NA         NA      NA       NA
## mnactic8             4.933e-10  3.343e+05   0.000    1.000
## mnactic9             3.965e-11  2.680e+05   0.000    1.000
## mnactic77           -2.150e-10  5.401e+05   0.000    1.000
## pdwrkp1              9.501e-13  8.589e+04   0.000    1.000
## new_vars.hincsrca2  -9.570e-12  5.412e+04   0.000    1.000
## new_vars.hincsrca3  -3.821e-11  9.840e+04   0.000    1.000
## new_vars.hincsrca4  -8.443e-11  7.567e+04   0.000    1.000
## new_vars.hincsrca5   6.043e-10  1.336e+05   0.000    1.000
## new_vars.hincsrca6          NA         NA      NA       NA
## new_vars.hincsrca7   3.543e-11  1.968e+05   0.000    1.000
## new_vars.hincsrca8   2.804e-10  1.115e+05   0.000    1.000
## new_vars.hincsrca77 -2.078e-13  3.006e+05   0.000    1.000
## new_vars.hincsrca88  5.407e-11  1.409e+05   0.000    1.000
## new_vars.hinctnta2  -1.724e-10  8.196e+04   0.000    1.000
```

```
## new_vars.hinctnta3  -1.783e-10  7.784e+04   0.000    1.000
## new_vars.hinctnta4  -1.596e-10  7.873e+04   0.000    1.000
## new_vars.hinctnta5  -1.435e-10  8.308e+04   0.000    1.000
## new_vars.hinctnta6  -1.096e-10  8.969e+04   0.000    1.000
## new_vars.hinctnta7  -1.477e-10  8.409e+04   0.000    1.000
## new_vars.hinctnta8  -1.462e-10  8.994e+04   0.000    1.000
## new_vars.hinctnta9  -1.407e-10  1.038e+05   0.000    1.000
## new_vars.hinctnta10 -1.686e-10  1.113e+05   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.3003e+02  on 606  degrees of freedom
## Residual deviance: 3.5216e-09  on 468  degrees of freedom
## AIC: 278
##
## Number of Fisher Scoring iterations: 25

# print r squared and adjusted r squared
pR2(mod)

## fitting null model for pseudo-r2

##           llh       llhNull            G2      McFadden          r2ML
## -1.760725e-09 -2.650147e+02  5.300294e+02  1.000000e+00  5.823849e-01
##          r2CU
##  1.000000e+00

library(pscl)
library(car)

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

# run ANOVA
Anova(mod)

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Analysis of Deviance Table (Type II tests)
##
## Response: X$uempla
##                   LR Chisq Df Pr(>Chisq)
## aesfdrk               0.00  5          1
## blgetmg               0.00  3          1
```

```
## dscrage               0.00  1          1
## dscretn               0.00  1          1
## dscrgnd               0.00  1          1
## dscrgrp               0.00  2          1
## dscrnap                     0
## dscrntn               0.00  1          1
## dscrrce               0.00  1          1
## dscrsex               0.00  1          1
## hlthhmp               0.00  2          1
## lnghom1               0.00 11          1
## rlgdgr                0.00 12          1
## hhmmb                 0.00  8          1
## gndr                  0.00  1          1
## agea                  0.00  1          1
## domicil               0.00  4          1
## edctn                 0.00  1          1
## edulvlb               0.00 14          1
## edulvlfb              0.00 17          1
## edulvlmb              0.00 15          1
## eduyrs                0.00  1          1
## hswrk                 0.00  1          1
## lvgptnea              0.00  4          1
## maritalb              0.00  4          1
## mnactic             228.18  5     <2e-16 ***
## pdwrkp                0.00  1          1
## new_vars.hincsrca     0.00  8          1
## new_vars.hinctnta     0.00  9          1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Second attempt - break categories into groups in order to have less dummies

```
# build alternative dataframe
X2=X

## lnghom1
# the person speaks Italian at home or not
X2["lnghom1ITA"] = as.factor(ifelse(X2["lnghom1"] == "ITA", 1, 0))
# the person speaks Southern Italian at home or not
X2["lnghom1DIAL"] = as.factor(ifelse(is.na(match(X2$lnghom1, c("NAP", "SCN", "SRD", "QAA
"))), 0, 1))
# the person speaks a European language which is not Italian at home or not
X2["lnghom1EURO"] = as.factor(ifelse(is.na(match(X2$lnghom1, c("ALB", "ENG", "FIL", "FRE
", "GER", "RUM", "SPA"))), 0, 1))
# eliminate original variable
X2 = subset(X2, select = - c(lnghom1))

## edulvlb
X2["edulvlbLOW"] = as.factor(ifelse(is.na(match(X2$edulvlb, c("0", "113", "213", "222"))
), 0, 1))
X2["edulvlbHIGH"] = as.factor(ifelse(is.na(match(X2$edulvlb, c("510", "520", "620", "720
", "800"))), 0, 1))
# eliminate original variable
X2 = subset(X2, select = - c(edulvlb))

## edulvlfb
X2["edulvlfbLOW"] = as.factor(ifelse(is.na(match(X2$edulvlfb, c("0", "113", "213", "222"
))), 0, 1))
X2["edulvlfbIGH"] = as.factor(ifelse(is.na(match(X2$edulvlfb, c("510", "520", "620", "72
0", "800"))), 0, 1))
# eliminate original variable
X2 = subset(X2, select = - c(edulvlfb))

## edulvlmb
```

```r
X2["edulvlmbLOW"] = as.factor(ifelse(is.na(match(X2$edulvlmb, c("0", "113", "213", "222"
))), 0, 1))
X2["edulvlmbIGH"] = as.factor(ifelse(is.na(match(X2$edulvlmb, c("510", "520", "620", "72
0", "800"))), 0, 1))
# eliminate original variable
X2 = subset(X2, select = - c(edulvlmb))

# run second model
mod2 = glm(X2$uempla ~ ., family = binomial(link = "logit"), data = X2)

## Warning: glm.fit: algorithm did not converge

summary(mod2)

##
## Call:
## glm(formula = X2$uempla ~ ., family = binomial(link = "logit"),
##     data = X2)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06   2.409e-06
##
## Coefficients: (2 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.657e+01  4.101e+05   0.000    1.000
## aesfdrk2            -3.236e-11  4.048e+04   0.000    1.000
## aesfdrk3             1.086e-09  5.301e+04   0.000    1.000
## aesfdrk4             3.366e-10  7.369e+04   0.000    1.000
## aesfdrk7            -3.365e-09  3.687e+05   0.000    1.000
## aesfdrk8            -1.208e-10  2.658e+05   0.000    1.000
## blgetmg2            -7.336e-10  1.030e+05   0.000    1.000
## blgetmg7             5.738e-10  3.819e+05   0.000    1.000
## blgetmg8            -2.346e-09  2.047e+05   0.000    1.000
## dscrage1             8.419e-10  2.199e+05   0.000    1.000
## dscretn1             2.966e-10  2.262e+05   0.000    1.000
## dscrgnd1             1.964e-09  2.630e+05   0.000    1.000
## dscrgrp2            -3.197e-10  1.163e+05   0.000    1.000
## dscrgrp7            -6.326e-10  2.530e+05   0.000    1.000
## dscrgrp8             1.629e-10  2.104e+05   0.000    1.000
## dscrnap1                    NA         NA      NA       NA
## dscrntn1            -2.344e-09  1.757e+05   0.000    1.000
## dscrrce1             4.308e-10  2.542e+05   0.000    1.000
## dscrsex1            -2.083e-09  2.530e+05   0.000    1.000
## hlthhmp2            -2.826e-10  3.290e+05   0.000    1.000
## hlthhmp3            -1.469e-09  3.041e+05   0.000    1.000
## rlgdgr1              1.342e-09  8.332e+04   0.000    1.000
## rlgdgr2              1.162e-09  7.405e+04   0.000    1.000
## rlgdgr3             -1.197e-09  7.380e+04   0.000    1.000
## rlgdgr4              1.935e-09  7.398e+04   0.000    1.000
## rlgdgr5              1.399e-09  6.307e+04   0.000    1.000
## rlgdgr6              1.910e-09  6.272e+04   0.000    1.000
## rlgdgr7              1.797e-09  6.191e+04   0.000    1.000
## rlgdgr8              2.927e-09  7.103e+04   0.000    1.000
## rlgdgr9              5.843e-10  1.021e+05   0.000    1.000
## rlgdgr10             2.949e-09  9.587e+04   0.000    1.000
## rlgdgr77             1.875e-09  2.646e+05   0.000    1.000
## rlgdgr88             1.582e-09  2.211e+05   0.000    1.000
## hhmmb2              -2.622e-09  8.009e+04   0.000    1.000
## hhmmb3              -3.268e-09  6.906e+04   0.000    1.000
## hhmmb4              -3.068e-09  7.155e+04   0.000    1.000
## hhmmb5              -3.400e-09  7.943e+04   0.000    1.000
## hhmmb6              -2.597e-09  1.124e+05   0.000    1.000
```

```
## hhmmb7              -4.246e-09  1.833e+05   0.000    1.000
## hhmmb8              -3.171e-09  3.097e+05   0.000    1.000
## hhmmb77             -4.657e-09  2.089e+05   0.000    1.000
## gndr1               -3.628e-10  3.540e+04   0.000    1.000
## agea                -1.886e-11  4.625e+03   0.000    1.000
## domicil2             1.714e-09  8.095e+04   0.000    1.000
## domicil3             1.194e-09  5.773e+04   0.000    1.000
## domicil4             1.364e-09  5.681e+04   0.000    1.000
## domicil5            -1.575e-09  1.056e+05   0.000    1.000
## edctn1              -9.412e-09  3.786e+05   0.000    1.000
## eduyrs              -4.191e-11  8.124e+03   0.000    1.000
## hswrk1               3.049e-08  3.093e+05   0.000    1.000
## lvgptnea2           -4.365e-10  5.634e+04   0.000    1.000
## lvgptnea6            2.109e-09  1.038e+05   0.000    1.000
## lvgptnea7            2.711e-09  4.079e+05   0.000    1.000
## lvgptnea8           -2.515e-10  3.817e+05   0.000    1.000
## maritalb2           -6.403e-09  1.614e+05   0.000    1.000
## maritalb3            1.062e-07  2.534e+05   0.000    1.000
## maritalb4           -5.090e-09  3.207e+05   0.000    1.000
## maritalb5            4.417e-09  4.129e+05   0.000    1.000
## maritalb6           -2.437e-09  7.304e+04   0.000    1.000
## mnactic2             1.068e-08  3.858e+05   0.000    1.000
## mnactic3             5.313e+01  5.029e+04   0.001    0.999
## mnactic6                    NA         NA      NA       NA
## mnactic8            -2.975e-08  3.192e+05   0.000    1.000
## mnactic9             2.453e-09  2.653e+05   0.000    1.000
## mnactic77            5.608e-09  5.007e+05   0.000    1.000
## pdwrkp1             -2.324e-09  7.948e+04   0.000    1.000
## new_vars.hincsrca2   5.769e-10  5.205e+04   0.000    1.000
## new_vars.hincsrca3   1.378e-09  9.368e+04   0.000    1.000
## new_vars.hincsrca4  -5.479e-09  7.312e+04   0.000    1.000
## new_vars.hincsrca5   3.520e-08  1.270e+05   0.000    1.000
## new_vars.hincsrca6   1.357e-08  5.111e+05   0.000    1.000
## new_vars.hincsrca7   1.209e-09  1.900e+05   0.000    1.000
## new_vars.hincsrca8   1.680e-09  1.038e+05   0.000    1.000
## new_vars.hincsrca77  2.292e-09  2.439e+05   0.000    1.000
## new_vars.hincsrca88 -5.678e-10  1.281e+05   0.000    1.000
## new_vars.hinctnta2   3.071e-09  7.738e+04   0.000    1.000
## new_vars.hinctnta3   1.125e-09  7.389e+04   0.000    1.000
## new_vars.hinctnta4   2.642e-09  7.475e+04   0.000    1.000
## new_vars.hinctnta5   3.310e-09  7.916e+04   0.000    1.000
## new_vars.hinctnta6   1.561e-09  8.545e+04   0.000    1.000
## new_vars.hinctnta7   3.332e-09  8.050e+04   0.000    1.000
## new_vars.hinctnta8   2.984e-09  8.492e+04   0.000    1.000
## new_vars.hinctnta9   4.494e-09  9.941e+04   0.000    1.000
## new_vars.hinctnta10  3.245e-09  1.073e+05   0.000    1.000
## lnghom1ITA1          9.293e-10  1.643e+05   0.000    1.000
## lnghom1DIAL1        -1.492e-08  1.994e+05   0.000    1.000
## lnghom1EURO1        -2.121e-10  1.687e+05   0.000    1.000
## edulvlbLOW1         -7.386e-10  4.773e+04   0.000    1.000
## edulvlbHIGH1        -2.025e-10  5.763e+04   0.000    1.000
## edulvlfbLOW1        -5.239e-10  3.893e+04   0.000    1.000
## edulvlfbIGH1         7.172e-11  6.320e+04   0.000    1.000
## edulvlmbLOW1         2.179e-09  3.966e+04   0.000    1.000
## edulvlmbIGH1         8.657e-10  6.345e+04   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.3003e+02  on 606  degrees of freedom
## Residual deviance: 3.5216e-09  on 516  degrees of freedom
## AIC: 182
##
## Number of Fisher Scoring iterations: 25
```

```
pR2(mod2)

## fitting null model for pseudo-r2

##           llh       llhNull            G2      McFadden          r2ML
## -1.760725e-09 -2.650147e+02  5.300294e+02  1.000000e+00  5.823849e-01
##          r2CU
##  1.000000e+00

Anova(mod2)
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: algorithm did not converge

## Analysis of Deviance Table (Type II tests)
##
## Response: X2$uempla
##                   LR Chisq Df Pr(>Chisq)
## aesfdrk               0.00  5          1
## blgetmg               0.00  3          1
## dscrage               0.00  1          1
## dscretn               0.00  1          1
## dscrgnd               0.00  1          1
## dscrgrp               0.00  2          1
## dscrnap                     0
## dscrntn               0.00  1          1
## dscrrce               0.00  1          1
## dscrsex               0.00  1          1
## hlthhmp               0.00  2          1
## rlgdgr                0.00 12          1
## hhmmb                 0.00  8          1
## gndr                  0.00  1          1
## agea                  0.00  1          1
## domicil               0.00  4          1
## edctn                 0.00  1          1
## eduyrs                0.00  1          1
```

```
## hswrk                 0.00  1          1
## lvgptnea              0.00  4          1
## maritalb              0.00  4          1
## mnactic             264.73  5     <2e-16 ***
## pdwrkp                0.00  1          1
## new_vars.hincsrca     0.00  9          1
## new_vars.hinctnta     0.00  9          1
## lnghom1ITA            0.00  1          1
## lnghom1DIAL           0.00  1          1
## lnghom1EURO           0.00  1          1
## edulvlbLOW            0.00  1          1
## edulvlbHIGH           0.00  1          1
## edulvlfbLOW           0.00  1          1
## edulvlfbIGH           0.00  1          1
## edulvlmbLOW           0.00  1          1
## edulvlmbIGH           0.00  1          1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#removing income and main activity to see if they are too similar to target

X3 = subset(X2, select = - c(new_vars.hincsrca, new_vars.hinctnta, mnactic, pdwrkp))
mod3 = glm(X3$uempla ~ ., family = binomial(link = "logit"), data = X3)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

typeof(mod3)

## [1] "list"

summary(mod3)

##
## Call:
## glm(formula = X3$uempla ~ ., family = binomial(link = "logit"),
##     data = X3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8263  -0.5913  -0.1468  -0.0295   4.1383
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -29.06415 3936.67450  -0.007  0.99411
## aesfdrk2       -0.25218    0.36316  -0.694  0.48743
## aesfdrk3        0.46234    0.48439   0.954  0.33984
## aesfdrk4       -0.15017    0.67154  -0.224  0.82305
## aesfdrk7      -15.60500 6522.63866  -0.002  0.99809
## aesfdrk8        0.74298    1.57691   0.471  0.63753
## blgetmg2        0.13638    0.99127   0.138  0.89057
## blgetmg7      -17.17335 6522.63871  -0.003  0.99790
## blgetmg8        0.94983    1.69235   0.561  0.57463
## dscrage1      -16.76082 2649.82851  -0.006  0.99495
## dscretn1      -14.50837 2559.46582  -0.006  0.99548
## dscrgnd1      -16.06238 2784.07123  -0.006  0.99540
## dscrgrp2       -0.18916    1.00742  -0.188  0.85106
## dscrgrp7        0.96861    1.84578   0.525  0.59974
## dscrgrp8        0.11002    1.82212   0.060  0.95185
## dscrnap1             NA         NA      NA       NA
## dscrntn1        0.31043    1.81660   0.171  0.86432
## dscrrce1       -1.34205 2769.35108   0.000  0.99961
## dscrsex1      -13.49580 3223.89573  -0.004  0.99666
## hlthhmp2       12.89886 3509.29433   0.004  0.99707
## hlthhmp3       12.17920 3509.29417   0.003  0.99723
```

```
## rlgdgr1        -0.23113    0.76050  -0.304  0.76119
## rlgdgr2        -1.15023    0.88301  -1.303  0.19270
## rlgdgr3        -0.10954    0.65459  -0.167  0.86710
## rlgdgr4        -0.15468    0.71261  -0.217  0.82816
## rlgdgr5         0.37318    0.54412   0.686  0.49282
## rlgdgr6         0.28165    0.55728   0.505  0.61328
## rlgdgr7         0.45436    0.55329   0.821  0.41153
## rlgdgr8         0.59890    0.62143   0.964  0.33517
## rlgdgr9        -0.60323    0.98171  -0.614  0.53890
## rlgdgr10        1.50056    0.75741   1.981  0.04757 *
## rlgdgr77      -15.07185 4075.05212  -0.004  0.99705
## rlgdgr88      -15.83129 3054.78220  -0.005  0.99587
## hhmmb2          1.36766    0.72627   1.883  0.05968 .
## hhmmb3          1.41749    0.64286   2.205  0.02745 *
## hhmmb4          0.98500    0.64460   1.528  0.12649
## hhmmb5          1.36665    0.69904   1.955  0.05058 .
## hhmmb6          2.28244    0.99895   2.285  0.02232 *
## hhmmb7          4.05552    1.76859   2.293  0.02184 *
## hhmmb8        -12.47988 4513.67738  -0.003  0.99779
## hhmmb77       -15.23027 2163.66352  -0.007  0.99438
## gndr1          -0.13394    0.31347  -0.427  0.66918
## agea           -0.12005    0.03948  -3.040  0.00236 **
## domicil2       -0.21586    0.98657  -0.219  0.82681
## domicil3        0.82684    0.57125   1.447  0.14778
## domicil4        1.03217    0.55257   1.868  0.06177 .
## domicil5        0.61468    0.95559   0.643  0.52006
## edctn1         -5.86514    1.13026  -5.189 2.11e-07 ***
## eduyrs          0.02101    0.07862   0.267  0.78930
## hswrk1         -3.75296    1.53596  -2.443  0.01455 *
## lvgptnea2      -0.54216    0.43983  -1.233  0.21770
## lvgptnea6      -1.85700    0.80092  -2.319  0.02042 *
## lvgptnea7     -14.94755 6522.63900  -0.002  0.99817
## lvgptnea8     -18.58078 6522.63872  -0.003  0.99773
## maritalb2     -15.20843 1934.86296  -0.008  0.99373
## maritalb3       2.70722    1.63189   1.659  0.09713 .
## maritalb4     -15.53720 4202.32929  -0.004  0.99705
## maritalb5     -18.58005 6522.63877  -0.003  0.99773
## maritalb6       0.09908    0.43795   0.226  0.82102
## lnghom1ITA1    16.90899 1783.88821   0.009  0.99244
## lnghom1DIAL1   16.51904 1783.88834   0.009  0.99261
## lnghom1EURO1   16.31282 1783.88827   0.009  0.99270
## edulvlbLOW1     1.09066    0.48396   2.254  0.02422 *
## edulvlbHIGH1    0.07350    0.56307   0.131  0.89615
## edulvlfbLOW1    0.17964    0.35396   0.508  0.61180
## edulvlfbIGH1   -0.02046    0.70438  -0.029  0.97683
## edulvlmbLOW1   -0.04019    0.34725  -0.116  0.90785
## edulvlmbIGH1    0.09038    0.78554   0.115  0.90840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.03  on 606  degrees of freedom
## Residual deviance: 361.84  on 540  degrees of freedom
## AIC: 495.84
##
## Number of Fisher Scoring iterations: 17

pR2(mod3)

## fitting null model for pseudo-r2
```

```
##          llh      llhNull           G2     McFadden         r2ML         r2CU
## -180.9175922 -265.0146892  168.1941940    0.3173299    0.2420144    0.4155575
```

```r
Anova(mod3)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: X3$uempla
##             LR Chisq Df Pr(>Chisq)
## aesfdrk        3.568  5  0.6131299
## blgetmg        0.801  3  0.8491700
## dscrage        1.179  1  0.2776437
## dscretn        0.147  1  0.7013949
## dscrgnd        0.637  1  0.4248669
## dscrgrp        0.576  2  0.7496066
## dscrnap                0
## dscrntn        0.029  1  0.8647371
## dscrrce        0.000  1  0.9999251
## dscrsex        0.050  1  0.8237442
## hlthhmp        0.478  2  0.7875832
## rlgdgr        12.468 12  0.4088723
## hhmmb         12.578  8  0.1272281
## gndr           0.183  1  0.6689282
```

```
## agea           9.659  1  0.0018845 **
## domicil        5.704  4  0.2223431
## edctn         78.035  1  < 2.2e-16 ***
## eduyrs         0.071  1  0.7899213
## hswrk         11.963  1  0.0005426 ***
## lvgptnea       6.766  4  0.1487723
## maritalb       5.311  5  0.3791376
## lnghom1ITA     2.428  1  0.1192098
## lnghom1DIAL    1.671  1  0.1960624
## lnghom1EURO    1.556  1  0.2122363
## edulvlbLOW     5.107  1  0.0238264 *
## edulvlbHIGH    0.017  1  0.8962043
## edulvlfbLOW    0.259  1  0.6107882
## edulvlfbIGH    0.001  1  0.9768136
## edulvlmbLOW    0.013  1  0.9078822
## edulvlmbIGH    0.013  1  0.9088036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#reducing the variables
X4 = subset(X3, select = - c(aesfdrk, blgetmg, hlthhmp, dscrage, dscretn, dscrgnd, dscrg
rp, dscrntn, dscrrce, dscrsex))
mod4 = glm(X4$uempla ~ ., family = binomial(link = "logit"), data = X4)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(mod4)

##
## Call:
## glm(formula = X4$uempla ~ ., family = binomial(link = "logit"),
##     data = X4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6281  -0.5812  -0.1521  -0.0415   4.0754
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.772e+01  1.628e+03  -0.011   0.9913
## dscrnap1      3.239e-01  7.167e-01   0.452   0.6513
## rlgdgr1      -2.255e-01  7.350e-01  -0.307   0.7590
## rlgdgr2      -1.142e+00  8.573e-01  -1.332   0.1828
## rlgdgr3      -9.047e-03  6.467e-01  -0.014   0.9888
## rlgdgr4      -2.605e-01  6.950e-01  -0.375   0.7078
## rlgdgr5       3.599e-01  5.309e-01   0.678   0.4978
## rlgdgr6       3.952e-01  5.413e-01   0.730   0.4654
## rlgdgr7       4.587e-01  5.418e-01   0.847   0.3971
## rlgdgr8       6.847e-01  6.074e-01   1.127   0.2596
## rlgdgr9      -5.683e-01  9.283e-01  -0.612   0.5404
## rlgdgr10      1.511e+00  7.370e-01   2.050   0.0404 *
## rlgdgr77     -1.537e+01  4.079e+03  -0.004   0.9970
## rlgdgr88     -1.614e+01  3.055e+03  -0.005   0.9958
## hhmmb2        1.227e+00  7.080e-01   1.734   0.0830 .
## hhmmb3        1.386e+00  6.164e-01   2.249   0.0245 *
## hhmmb4        9.095e-01  6.282e-01   1.448   0.1477
## hhmmb5        1.269e+00  6.798e-01   1.867   0.0618 .
## hhmmb6        2.237e+00  9.475e-01   2.361   0.0182 *
## hhmmb7        3.843e+00  1.720e+00   2.234   0.0255 *
## hhmmb8       -1.203e+01  4.611e+03  -0.003   0.9979
## hhmmb77      -1.535e+01  2.264e+03  -0.007   0.9946
## gndr1        -1.191e-01  2.857e-01  -0.417   0.6768
## agea         -1.102e-01  3.808e-02  -2.895   0.0038 **
```

```
## domicil2     -5.321e-01  9.558e-01  -0.557   0.5777
## domicil3      6.564e-01  5.359e-01   1.225   0.2207
## domicil4      9.190e-01  5.147e-01   1.785   0.0742 .
## domicil5      6.261e-01  8.868e-01   0.706   0.4801
## edctn1       -5.674e+00  1.113e+00  -5.097 3.45e-07 ***
## eduyrs        1.247e-02  7.825e-02   0.159   0.8734
## hswrk1       -3.737e+00  1.530e+00  -2.442   0.0146 *
## lvgptnea2    -5.554e-01  4.287e-01  -1.296   0.1951
## lvgptnea6    -1.786e+00  7.836e-01  -2.279   0.0227 *
## lvgptnea7    -1.416e+01  6.523e+03  -0.002   0.9983
## lvgptnea8    -1.848e+01  6.523e+03  -0.003   0.9977
## maritalb2    -1.553e+01  1.938e+03  -0.008   0.9936
## maritalb3     2.552e+00  1.559e+00   1.637   0.1016
## maritalb4    -1.535e+01  4.268e+03  -0.004   0.9971
## maritalb5    -1.796e+01  6.523e+03  -0.003   0.9978
## maritalb6     6.146e-02  4.247e-01   0.145   0.8849
## lnghom1ITA1   1.739e+01  1.628e+03   0.011   0.9915
## lnghom1DIAL1  1.694e+01  1.628e+03   0.010   0.9917
## lnghom1EURO1  1.680e+01  1.628e+03   0.010   0.9918
## edulvlbLOW1   1.024e+00  4.679e-01   2.189   0.0286 *
## edulvlbHIGH1  1.738e-01  5.485e-01   0.317   0.7514
## edulvlfbLOW1  7.075e-02  3.380e-01   0.209   0.8342
## edulvlfbIGH1  8.752e-02  7.058e-01   0.124   0.9013
## edulvlmbLOW1  4.815e-02  3.367e-01   0.143   0.8863
## edulvlmbIGH1 -7.992e-02  7.668e-01  -0.104   0.9170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 530.03  on 606  degrees of freedom
## Residual deviance: 369.72  on 558  degrees of freedom
## AIC: 467.72
##
## Number of Fisher Scoring iterations: 17

pR2(mod4)

## fitting null model for pseudo-r2

##          llh      llhNull           G2     McFadden         r2ML         r2CU
## -184.8605659 -265.0146892  160.3082465    0.3024516    0.2321027    0.3985383

Anova(mod4)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table (Type II tests)
##
## Response: X4$uempla
```

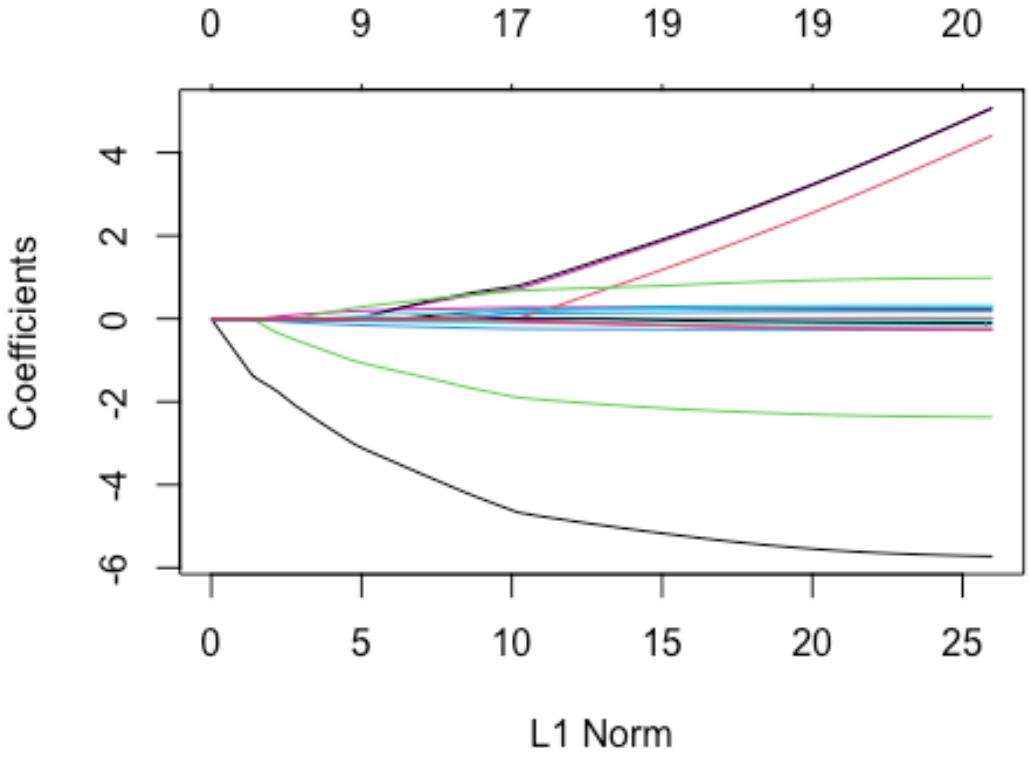
```
##             LR Chisq Df Pr(>Chisq)
## dscrnap        0.213  1  0.6444502
## rlgdgr        14.240 12  0.2856351
## hhmmb         13.110  8  0.1081295
## gndr           0.174  1  0.6764391
## agea           8.715  1  0.0031555 **
## domicil        6.515  4  0.1638683
## edctn         76.865  1  < 2.2e-16 ***
## eduyrs         0.025  1  0.8736340
## hswrk         12.026  1  0.0005246 ***
## lvgptnea       6.470  4  0.1666744
## maritalb       5.052  5  0.4095737
## lnghom1ITA     4.527  1  0.0333663 *
## lnghom1DIAL    2.944  1  0.0861841 .
## lnghom1EURO    2.802  1  0.0941649 .
## edulvlbLOW     4.805  1  0.0283849 *
## edulvlbHIGH    0.100  1  0.7517339
## edulvlfbLOW    0.044  1  0.8340451
## edulvlfbIGH    0.015  1  0.9016370
## edulvlmbLOW    0.020  1  0.8862428
## edulvlmbIGH    0.011  1  0.9166339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#convergence reached

```
# trying to identify the coefficients
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.1

set.seed(123)
X5 = subset(X4, select = - c(uempla))
X5 = as.matrix(X5)
y = X4$uempla
lassomod = glmnet(X5, y, alpha = 1, family = "binomial")
print(lassomod)

##
## Call:  glmnet(x = X5, y = y, family = "binomial", alpha = 1)
##
##    Df  %Dev   Lambda
## 1   0  0.00 0.104400
## 2   1  1.62 0.095170
## 3   1  3.03 0.086710
## 4   1  4.26 0.079010
## 5   1  5.33 0.071990
## 6   1  6.27 0.065600
## 7   1  7.10 0.059770
## 8   1  7.83 0.054460
## 9   1  8.47 0.049620
## 10  1  9.04 0.045210
## 11  1  9.55 0.041200
## 12  3 10.04 0.037540
## 13  4 10.89 0.034200
## 14  4 11.83 0.031160
## 15  5 12.79 0.028390
## 16  5 13.95 0.025870
## 17  5 14.94 0.023570
## 18  6 15.94 0.021480
## 19  6 16.82 0.019570
## 20  7 17.60 0.017830
```

```
## 21  7 18.27 0.016250
## 22  7 18.85 0.014810
## 23  8 19.38 0.013490
## 24  9 19.89 0.012290
## 25  9 20.36 0.011200
## 26  9 20.75 0.010200
## 27 10 21.10 0.009298
## 28 10 21.41 0.008472
## 29 10 21.67 0.007719
## 30 11 21.91 0.007034
## 31 11 22.13 0.006409
## 32 11 22.31 0.005839
## 33 12 22.47 0.005321
## 34 12 22.62 0.004848
## 35 14 22.76 0.004417
## 36 15 22.88 0.004025
## 37 15 22.98 0.003667
## 38 16 23.08 0.003342
## 39 17 23.16 0.003045
## 40 18 23.28 0.002774
## 41 18 23.38 0.002528
## 42 18 23.48 0.002303
## 43 18 23.56 0.002099
## 44 18 23.64 0.001912
## 45 19 23.70 0.001742
## 46 19 23.76 0.001587
## 47 19 23.81 0.001446
## 48 19 23.86 0.001318
## 49 20 23.90 0.001201
## 50 20 23.93 0.001094
## 51 20 23.97 0.000997
## 52 20 23.99 0.000908
## 53 20 24.02 0.000828
## 54 20 24.04 0.000754
## 55 20 24.06 0.000687
## 56 20 24.07 0.000626
## 57 20 24.09 0.000571
## 58 19 24.10 0.000520
## 59 19 24.11 0.000474
## 60 19 24.12 0.000432
## 61 19 24.13 0.000393
## 62 19 24.14 0.000358
## 63 19 24.14 0.000327
## 64 19 24.15 0.000298
## 65 19 24.15 0.000271
## 66 19 24.16 0.000247
## 67 19 24.16 0.000225
## 68 20 24.17 0.000205
## 69 20 24.17 0.000187
## 70 20 24.17 0.000170
## 71 20 24.18 0.000155
## 72 20 24.18 0.000141
## 73 20 24.18 0.000129
## 74 20 24.18 0.000117
## 75 20 24.19 0.000107
## 76 20 24.19 0.000097
## 77 20 24.19 0.000089
## 78 20 24.19 0.000081
## 79 20 24.19 0.000074
## 80 20 24.19 0.000067
## 81 20 24.19 0.000061
## 82 20 24.19 0.000056
```

```
plot(lassomod)
```




```
#to see if education was too similar to being unemployed
library(expss)


chisq.test(X6$uempla, X6$gndr, correct=FALSE)


        Pearson's Chi-squared test


data:  X6$uempla and X6$gndr

X-squared = 0.42259, df = 1, p-value = 0.5156


library(maditr)
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
## The following object is masked from 'package:expss':
##
##     vars
```

```
library(ggplot2)
#table(X3$uempla, X3$edctn)
#table(X3$uempla)
cross_cases(X3, uempla, edctn)
```

edctn
0
1
uempla
0
319
192
1
95
1
#Total cases
414
193

#education is not the same as uempla, continue searching

```
#split between employed and unemployed

X6 = data.frame(X5)
X6$uempla = y

#checking for missing values once again
colSums(is.na(X6))
```

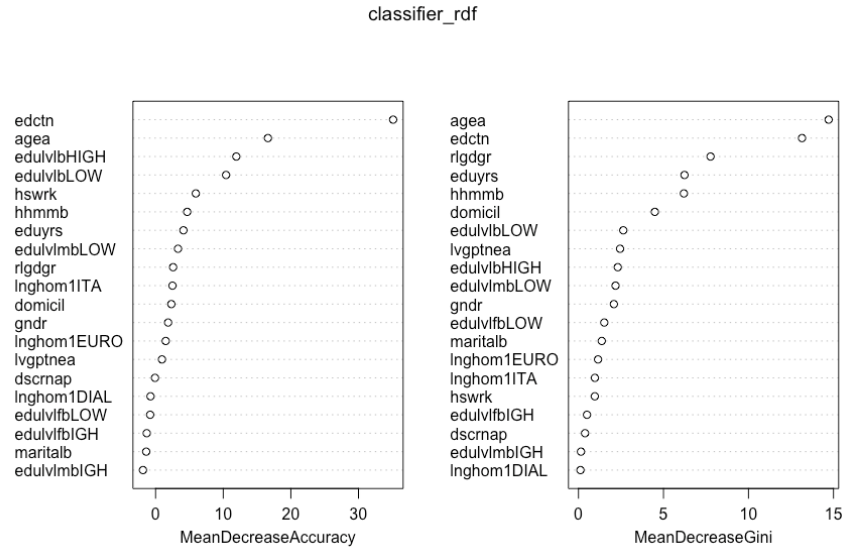
```
##     dscrnap      rlgdgr       hhmmb        gndr        agea     domicil
##           0           0           0           0           0           0
##       edctn      eduyrs       hswrk    lvgptnea    maritalb  lnghom1ITA
##           0           0           0           0           0           0
## lnghom1DIAL lnghom1EURO  edulvlbLOW edulvlbHIGH edulvlfbLOW edulvlfbIGH
##           0           0           0           0           0           0
## edulvlmbLOW edulvlmbIGH      uempla
##           0           0           0
```

```
X_emp = X6 %>% filter(uempla == "0")
X_unemp = X6 %>% filter(uempla == "1")
```

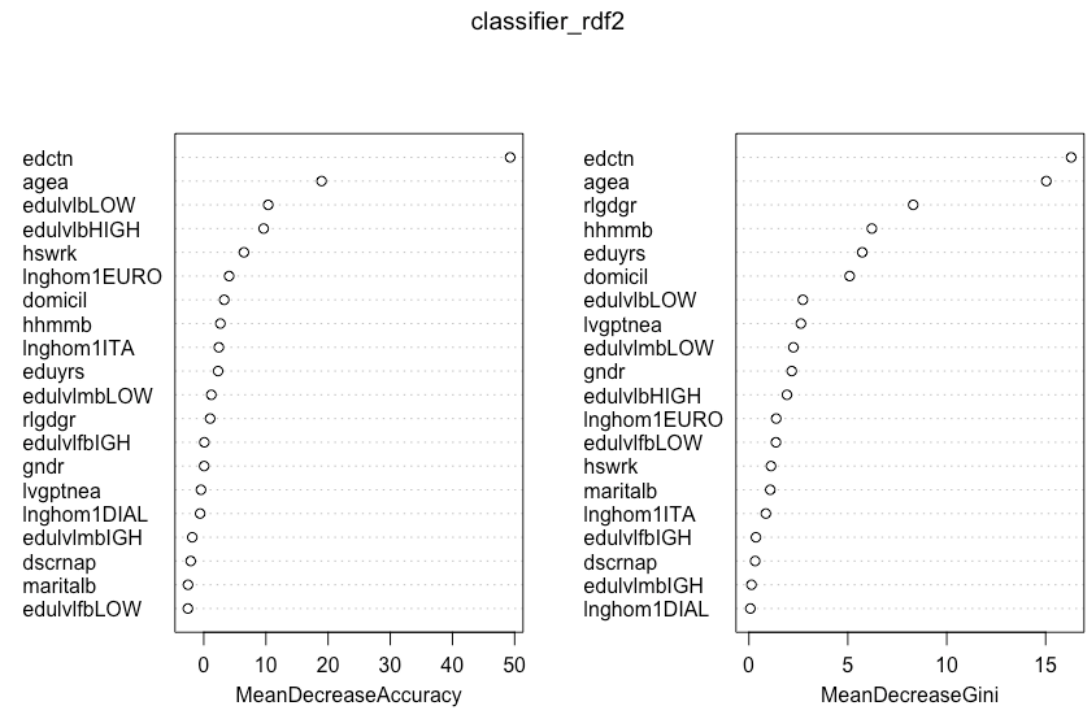
```
train_size = floor(0.8 * nrow(X_unemp))

# Randomly sample instance numbers for the train set
train_instances = sample(seq_len(nrow(X_unemp)), size = train_size)
# Build the train and test sets
train_emp = X_emp[train_instances, ]
test_emp = X_emp[-train_instances, ]

train_unemp = X_unemp[train_instances, ]
test_unemp = X_unemp[-train_instances, ]

train = rbind(train_emp, train_unemp)
test = rbind(test_emp, test_unemp)
```

Random Forest

```
classifier_rdf <- randomForest(uempla~., data =train, mtry=5,importance = TRUE, ntrees =
50)
classifier_rdf

##
## Call:
##  randomForest(formula = uempla ~ ., data = train, mtry = 5, importance = TRUE,      n
trees = 50)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 19.08%
## Confusion matrix:
##    0  1 class.error
## 0 57 19   0.2500000
## 1 10 66   0.1315789

classifier_pred_rdf <- predict(classifier_rdf, test, type = "class")
classifier_rdf_acc <-
mean(test$uempla == classifier_pred_rdf)
classifier_rdf_acc

## [1] 0.6043956

varImpPlot(classifier_rdf)
```



```
classifier_rdf2 <- randomForest(uempla~., data =train, mtry=10,importance = TRUE, ntrees
= 50)
classifier_rdf2

##
## Call:
##  randomForest(formula = uempla ~ ., data = train, mtry = 10, importance = TRUE,
ntrees = 50)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##         OOB estimate of  error rate: 21.05%
## Confusion matrix:
##    0  1 class.error
## 0 57 19   0.2500000
## 1 13 63   0.1710526

classifier_pred_rdf2 <- predict(classifier_rdf2, test, type = "class")
classifier_rdf_acc2 <-
mean(test$uempla == classifier_pred_rdf2)
classifier_rdf_acc2

## [1] 0.6263736

varImpPlot(classifier_rdf2)
```



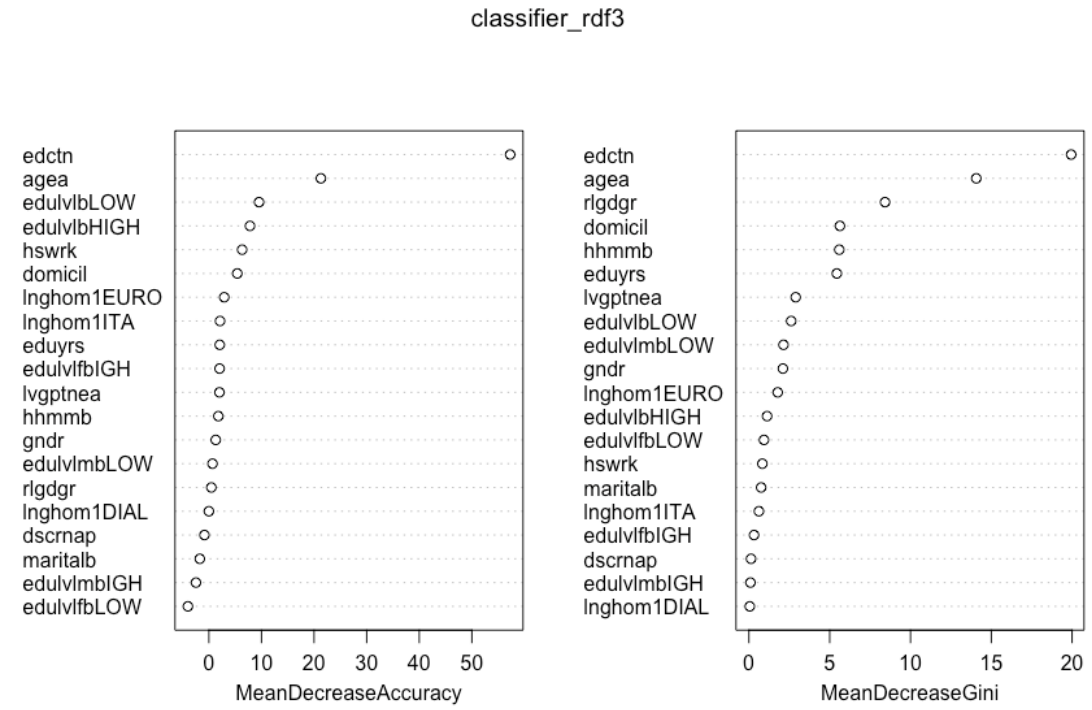

```
classifier_rdf3 <- randomForest(uempla~., data =train, mtry=19,importance = TRUE, ntrees
= 50)
classifier_rdf3

##
## Call:
##  randomForest(formula = uempla ~ ., data = train, mtry = 19, importance = TRUE,
ntrees = 50)
##                Type of random forest: classification
```

```
##                       Number of trees: 500
## No. of variables tried at each split: 19
##
##         OOB estimate of  error rate: 23.03%
## Confusion matrix:
##    0  1 class.error
## 0 57 19   0.2500000
## 1 16 60   0.2105263
```

```
classifier_pred_rdf3 <- predict(classifier_rdf3, test, type = "class")
classifier_rdf_acc3 <-
mean(test$uempla == classifier_pred_rdf3)
classifier_rdf_acc3
```

```
## [1] 0.6263736
```

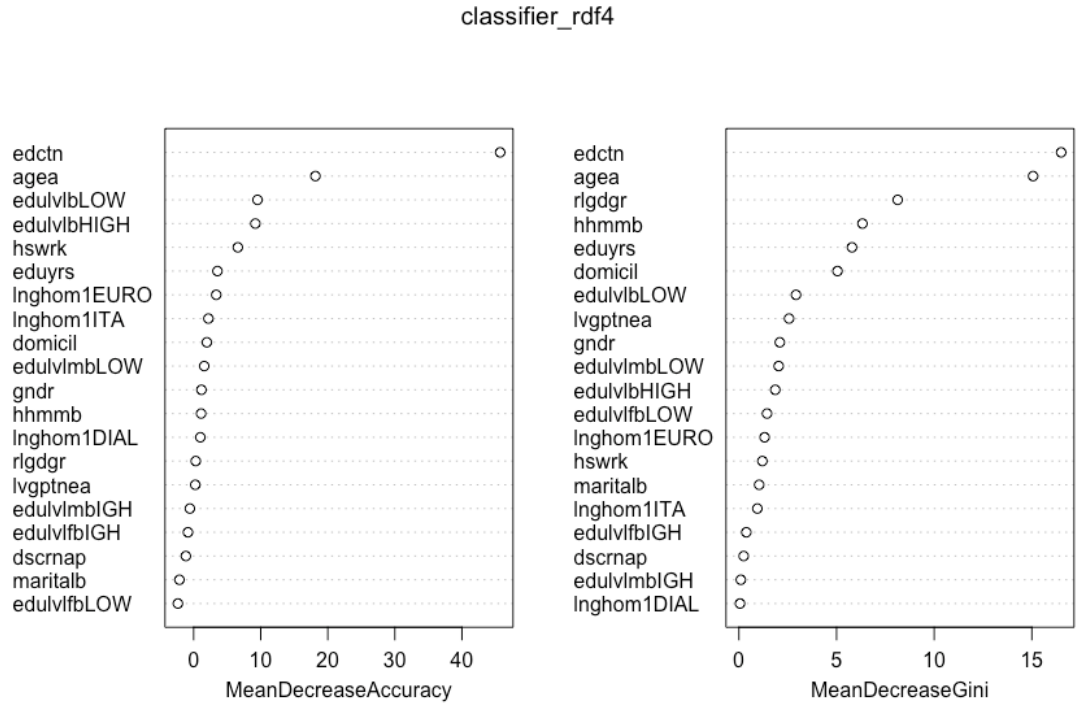
```
varImpPlot(classifier_rdf3)
```




```
classifier_rdf4 <- randomForest(uempla~., data =train, mtry=10,importance = TRUE, ntrees
= 100)
classifier_rdf4
```

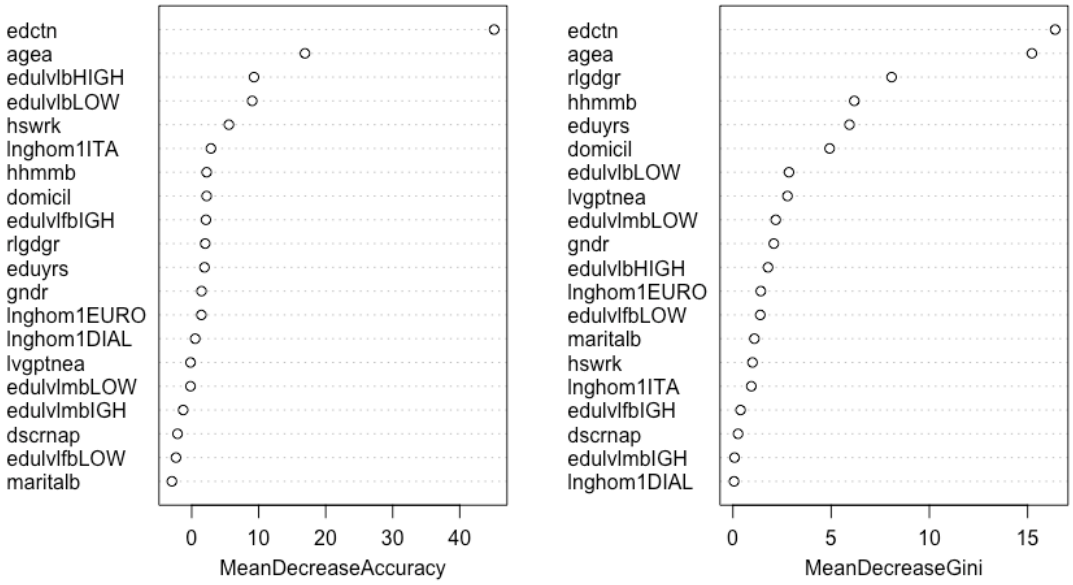
```
##
## Call:
##  randomForest(formula = uempla ~ ., data = train, mtry = 10, importance = TRUE,
ntrees = 100)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##         OOB estimate of  error rate: 19.74%
## Confusion matrix:
##    0  1 class.error
## 0 57 19   0.2500000
## 1 11 65   0.1447368
```

```
classifier_pred_rdf4 <- predict(classifier_rdf4, test, type = "class")
classifier_rdf_acc4 <-
mean(test$uempla == classifier_pred_rdf4)
classifier_rdf_acc4

## [1] 0.6263736

varImpPlot(classifier_rdf4)
```




```
classifier_rdf5 <- randomForest(uempla~., data =train, mtry=10,importance = TRUE, ntrees
= 500)
classifier_rdf5

##
## Call:
##  randomForest(formula = uempla ~ ., data = train, mtry = 10, importance = TRUE,
ntrees = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##         OOB estimate of  error rate: 21.71%
## Confusion matrix:
##    0  1 class.error
## 0 57 19   0.2500000
## 1 14 62   0.1842105

classifier_pred_rdf5 <- predict(classifier_rdf5, test, type = "class")
classifier_rdf_acc5 <-
mean(test$uempla == classifier_pred_rdf5)
classifier_rdf_acc5

## [1] 0.6263736

varImpPlot(classifier_rdf5)
```

classifier_rdf5



...