LUISS Università Guido Carli

Department of Business and Management

Bachelor's degree in Management and Computer Science
Course of Finance and Financial Technologies

Demographic discrimination in the peer-to-peer lending market: the Bondora case

Supervisor:
Prof. Michela Altieri

Candidate:
Alida Brizzante

Academic Year 2021/2022

Abstract

Peer-to-peer lending is expanding worldwide, exploiting the technology evolution and offering a valid financial alternative with respect to traditional banks. Even if the process for obtaining a loan appears to be easier than ever, there may be a chance for demographic discrimination to arise (Klein et al., 2021). Demographic discrimination in the peer-to-peer lending market may be present whenever a group of borrowers, identified on the basis of demographic characteristics, has an at least equally good ability of repayment but it is still associated with disadvantageous loan conditions due to some demographic characteristics (for example gender, race, education, income, marriage, and others).

Using data on P2P loans of an Estonian platform, I perform classification on the default probability and then a regression on the interest rate. Combining the results, discrimination is tested for some demographic variables of *Gender*, *Education*, and *Married*. The outcomes obtained employing traditional methods are cautiously discussed under a machine learning approach.

The results show that there are grounds for discrimination on the basis of gender and the marital status of the borrower. More precisely: (i) female borrowers appear to be disadvantaged with respect to their male peers; (ii) discrimination against those not being married or not willing to disclose it is present. However, there is no discrimination evidence for the education level of the borrower, even if higher education is associated with a lower default likelihood.

Table of Contents

List of Tables

List of Figures

Introduction

The peer-to-peer lending market is expanding at incredible pace all around the world. It allows people, i.e. peers, to request and receive loans interacting directly through online platforms, in a disintermediated way (Milne and Parboteeah, 2016).

This new financial peer-to-peer model can increase the accessibility of financial services, with an important focus on the portion of the population being underserved by traditional banks. However, this benefit may be opposed to the existence of peer-to-peer lending market inefficiencies, leading to potential discrimination (Klein et al., 2021).

The purpose of this research is to analyze the data coming from an Estonian P2P lending platform, in order to understand whether demographic discrimination is present, due to an incorrect evaluation of the borrower's demographic characteristics. Using traditional and machine learning methods, it is possible to test whether a group of borrowers, identified on the basis of demographic characteristics, is subject to unfavorable loan conditions even when the ability to repay the debt is at least equally good with respect to the opposed group.

The main analysis consists of a classification of the loan default, using a logit and probit analysis, and a regression of the loan interest rate, using a multivariate linear regression. The demographic variables considered as independent variables are the following: *Married, Education* and *Gender*. The results of the models constructed are then combined to yield some evidence about demographic discrimination. In other words, this allows spotting whether disadvantageous loan conditions are faced by demographically different categories of borrowers. The results show that female borrowers appear to be disadvantaged with respect to males and that those not being married or not willing to disclose it are discriminated as well.

This paper contains two chapters. In the first chapter, the theoretical background is explained, introducing the functioning of peer-to-peer lending and the main features making it a valid alternative to traditional banking. Going on, the criticalities concerning

peer-to-peer lending are introduced, focusing on the demographic biases that may be present in such platforms. Last, the most relevant recent research is reported and two hypotheses are formulated to investigate discrimination.

The second chapter introduces the loan dataset analyzed and dives deep into the methodology employed to test the hypotheses. The results stemming from the conducted analysis are presented and compared to the prior research. Discrimination is then tested coherently with what was previously explained. Last, a machine learning approach is explained and used to test the robustness of the findings.

Concluding, this research provides an analysis of potential demographic discrimination in the peer-to-peer lending market, with a focus on the European platform Bondora. The discrimination outcomes will be presented at the end of the second chapter.

Chapter 1
Theory and Background

1.1 P2P Lending

Peer-to-Peer (P2P) lending, otherwise called social lending or crowdlending, is a lending model that uses a double-sided platform to connect individual borrowers with a crowd of individual lenders (Ribeiro-Navarrete et al., 2021). This alternative source of credit, contrary to what normally happens with traditional banks, allows the two parties to interact without the need of a central authority, in a disintermediated fashion (Milne and Parboteeah, 2016). In this way, borrowers and lenders can communicate through the platform and conclude the deal directly, without referring to a traditional financial institution.

The peer-to-peer lending models were introduced for the first time in 2005 in Europe, with the launch of Zopa. The first two US platforms were inaugurated the following year: Prosper and Lending Club. However, the initial expansion started right after the 2008 global financial crisis, due to the strict regulations enacted with which traditional financial institutions had to comply (Atz and Bhoalt, 2016). Apart from the financial crisis, the spread of the Internet and the development of new technologies certainly facilitated its adoption, leveraging on the network effect that characterizes the platforms (Havrylchyk et al., 2017; Milne and Parboteeah, 2016).

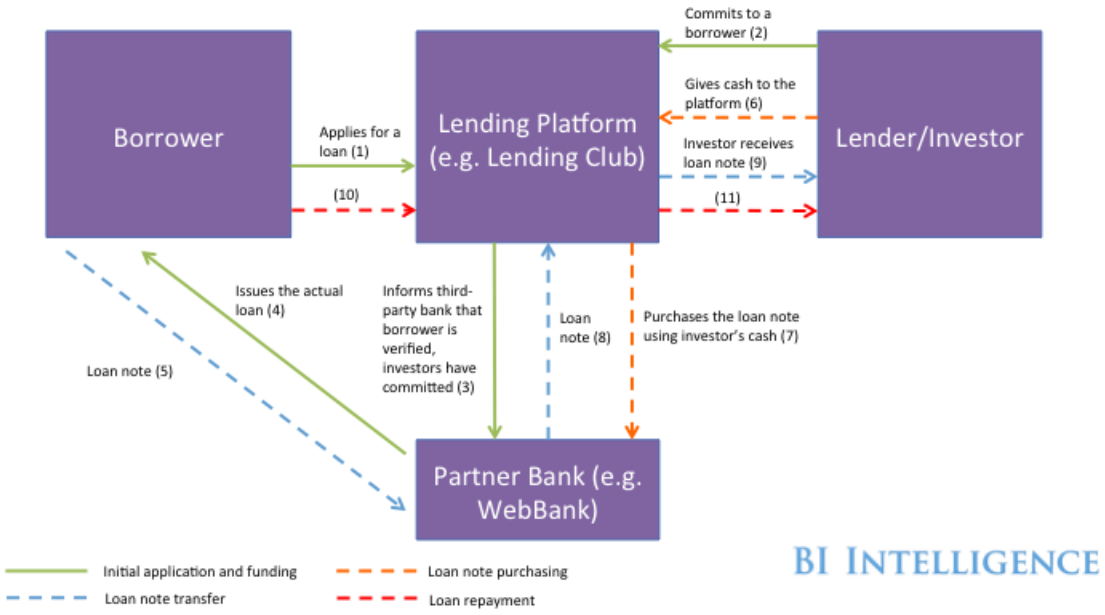
During the following years, many competitors entered the market and the global crowdfunding market[1] grew from $0.5 billion in 2011 to $290 billion in 2016 (Rau, 2019). Ziegler et al. (2018) provided additional evidence of its importance, finding that 17% of

[1] There are different types of crowdfunding: reward and donation based, debt based (P2P lending) and equity based. Thus, the value is comprehensive of all the types.

alternative market share in Europe was held by P2P lending. Furthermore, in 2021 the global peer-to-peer lending market reached a value of $112.9 billion[2].

The mechanism behind peer-to-peer lending platforms is briefly described here. First, the borrower applies online for the loan, providing the information required by the platform. At the same time, investors can notify their willingness and availability to provide credit to borrowers. The platform has the fundamental function of being an intermediary, creating the link between borrowers and lenders and verifying the information provided by the borrower. Together with this, it is important to underline that another third party enters into the mechanism: a parent bank. The partner bank has the role of issuing the actual loan to the borrower, usually net to an origination fee that the platform will receive.

Peer-to-Peer Lending Mechanism



*This is a simplified graphic showing how a loan is processed through a peer-to-peer marketplace – revenue sources such as fees are not included

Source: https://www.businessinsider.com/peer-to-peer-lending-how-digital-lending-marketplaces-are-disrupting-the-predominant-banking-model-2015-5?r=US&IR=T

[2] Peer to Peer (P2P) Lending Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2022-2027, IMARC Group.

Apart from the mechanism, peer-to-peer lending has the following characterizing elements, making it revolutionary concerning traditional banks. First, peer-to-peer lending platforms are constructed in a way that the consequences of information asymmetries are reduced, providing availability of both soft and hard financial information, that will also contribute to the assessment of the creditworthiness (Iyer et al., 2016; Cummins et al., 2019). Second, the platforms do not bear the credit risk which, contrary to what happens with traditional banks, is on the investor side. Last, peer-to-peer lending has less strict requirements to access credit, with respect to traditional banks (Serrano-Cinca et al., 2015).

1.2 Underbanked Phenomenon

The companies offering alternative financial services, including peer-to-peer lending, are expanding unevenly among European counties and the selection of the markets to serve is not done by chance. This can be partially explained by correlated phenomena involving the financial sector and occurring in a non-standardized manner. One of the most important, even if somehow still overlooked, is certainly the underbanked phenomena.

In this section, the underbanked problem will be defined, so that to explore why a high underbanked presence may translate into more clients willing to rely on alternative financial services (like those provided by Fintech companies like Bondora) rather than on the traditional banks.

The term “underbanked” refers to those who already possess bank accounts but are patrons of alternative financial services providers at the same time (Xu, 2019). This portion of the population may appear negligible but, looking at the Global Findex Database, the reality is completely different: the 12% of account holders worldwide can be classified to be

"underbanked". Furthermore, it has to be considered that all these people not having full access to financial services are bringing costs to the whole economy[3].

In recent years, many researchers have struggled to build a profile of the underbanked category (Xu, 2019) or, more in general, of alternative financial service users (Despard et al., 2015; Goodstein and Rhine, 2017). What they have found out is the following: the underbanked are characterized, on average, by lower income, lower education level, income volatility, unemployment, and as being part of a minority, for example racial (Xu, 2019; Despard et al., 2015; Goodstein and Rhine, 2017). However, they have also discovered that the willingness to choose alternative financial services stems from an informed decision, deriving from the fact that for them is easier to do so instead of qualifying for traditional bank services.

Additionally, for what concerns the factors influencing the underbanked's use of alternative financial services, the researchers' opinions converged into three main causes. The first one is not affordable bank fees, either for the unclear structure or for the high cost, that become unsustainable if combined with already limited income (Guy Birken, 2020). This also includes the fact that alternative solutions often come with innovative and easy-to-use tools to manage the financing that, on the other side, banks are lacking. Second, most underbanked people do not have trust in banks and this lack of trust is usually derived from past experience and perceptions, experienced by the underbanked (Guy Birken, 2020; Xu, 2019). Last, banks are not providing satisfactory and targeted solutions for this category, not properly covering their needs. This is exacerbated by the high bank credit denial rate, which makes the underbanked feeling discouraged by banks (Xu, 2019).

[3] Citibank estimated that a wealth of 30$ billion of potential savings could be generated reducing the cash usage of the underbanked and unbanked in developed countries of the world. This is because the underbanked and unbanked, instead of using traditional financial services, are more prone to rely on cash or alternative financial services.

1.3 Biases in Fintech Lending

In the previous section, the discussion focused on the fact that P2P lending opens the possibility to receive loans to underserved categories of borrowers, praising the inclusive characters of crowdlending fintech platforms. However, the devotion to making financial services accessible to everyone may be opposed to the existence of P2P lending market inefficiencies, allowing discrimination to unveil (Klein et al., 2021).

In general, discrimination occurs whenever the terms of a transaction are affected by the personal characteristic of the participants that are not relevant to the transaction (Blanchflower, Levine and Zimmerman, 2003). In the credit market, discrimination means that different groups of the borrowers, with non-significative differences in the ability to repay their loans, are offered different loan terms (for example interest rate) or are subject to differences in loan's funding success (Ladd, 1998; Turner and Skidmore, 1999).

As many researchers have previously reported, the reality is that P2P lending platforms are still far from a perfect and unbiased allocation of credit. The following Lending Club events of 2016 are a representative example of the potential flaws existing in this kind of platform: the company was reported to have improperly sold $22 million of loans to an institutional investor that did not meet the investor's standards (The New York Times, 2016).

Another possible type of bias may be related to the inability of the investor to perform a correct evaluation of the borrower requiring the loan. This is especially noticeable when considering the demographic characteristics of the borrowers that, when the lender is called to choose the allocation of its investment, are not considered correctly. To support this statement, many previous empirical researches reveal that demographically distinct borrowers appear to encounter significantly different loan terms (Nickerson, 2022).

Thus, how should demographic discrimination be recognized when looking at P2P lending data? The answer is rather intuitive. Discrimination is present whenever the group having

disadvantageous demographic[4] characteristic experiences either higher interest rates, which translates into higher return for investors to compensate for the higher risk or a higher rejection rate, even if no financially concrete explanation exists (for instance, even when the default rate of that group is lower than the other).

The cause of this evidence of discrimination in P2P lending markets is certainly not unique. However, some researchers have tried to contribute with their interpretation and a possible explanation of the phenomenon.

A sociological interpretation is given by Harkness (2016) and Pager and Shepherd (2008), stating that the cultural stereotypes and the well-rooted prejudices of the society towards minorities are having a deep influence on the decisional journey of the investors. In other words, the lender becomes more prone to model a taste-based or prejudice-based funding assessment, preferring demographic characteristics over financial histories.

Another contribution is presented by Kim (2020), recognizing in the lack of professional and cumulative experience is one possible driver of irrational decisions when dealing with P2P lending. Last, visualizing such demographic information when making the funding choice may be misleading for the inexpert investor, nudging his perception and choice.

1.4 Literature Review

In recent years, the existence of discrimination based on demographic characteristics in P2P lending platforms has been studied extensively around the world. However, the evidence that has been generated is far from being generalizable, due to the number of differences existing among the various platforms.

Pope and Syndor (2011) analyzed a fintech lending platform in the US and shows that the market discriminates according to age, disfavoring older people. The same holds for male

[4] For example race, gender, education level, homeownership type, marital status or age.

borrowers, finding that the probability for women to obtain a loan is higher than for men. Turning to the race, Pope and Syndor (2011) provided evidence of black borrowers being less likely to receive funding. However, they also recognize that this is not taste-based discrimination.

Next, Santoso et al. (2018) studied how borrower characteristics impact the loan rate and the loan default with regression analysis, using three Indonesian platforms. They found out that women tend to receive a higher interest rate with respect to men and that they are associated with a higher probability of default. For what concerns the education level, they discovered that higher education is associated with a lower probability of loan default. Furthermore, both marital status and age have a significant impact on the default, but the sign of their influence is different among the three platforms.

Evidence from Chinese P2P lending platforms is provided by both Chen et al. (2019) and Chen et al. (2020). First, Chen et al. (2019) used logistic regression and marginal effects to look at how the variables affect the loan default and funding success. The discoveries are the following: (i) females are less likely than males to default, making them better clients in P2P lending; (ii) the higher the education level, the higher the funding success of the loan; and (iii) marriage is a sign of creditworthiness, increasing the funding success of women with respect to single women.

Second, Chen et al. (2020) employed a logit and probit regression and revelated that only the education factor is correctly evaluated by the market. There is evidence of taste-base discrimination against young people and women and the married status increases the probability of obtaining funds.

To conclude, Kim (2020) used data from three Korean lending platforms to study the lender preferences and borrower repayment performance with linear regression. For what concerns the age, it came out that investors discriminate according to the borrower's age, even if the effect is minimal. Turning to gender, there was no gender discrimination, meaning not only that it does not affect the repayment ability, but also that gender does not influence lenders' decisions.

1.4 Research Question

The multitude of researchers, investigating the existence of demographic discrimination in P2P lending markets has not yet converged on a unique finding. The results change according to the geographical area served by the platform and its internal functioning, related to the technology and regulatory framework.

This paper aims at contributing to this literature by analyzing the funded loans of the European P2P lending platform Bondora. The recent research dealing with EU companies is scarce, leaving room for additional analysis of the demographic discrimination in the European peer-to-peer lending market.

The demographic variables considered for the analysis are the following: *Gender*, *Education* and *Married*. First, the gender of the borrower is relevant because some studies claim that females are more risk-averse when dealing with financial decisions (Byrnes et al., 1999), reducing their inability to repay the loan and, as a consequence, their default (Chen et al., 2019).

Past research finds that a higher education level is associated with high income, lowering the default risk (Santoso et al., 2018; Chen and Huang, 2016; Dorfleitner et al., 2016). I do not expect different results from the peer-to-peer lending default modeling outcomes.

Third, the married status of the borrower is recognized to have a double-folded perception. On one side, being married may be a sign of financial stability, reducing the risk and the default probability (Chen et al., 2019; Santoso et al, 2018). On the other side, it is also true that being married could mean having more financial constraints, having to look after the family and increasing the default (Santoso et al, 2018).

The first part of the subsequent analysis will focus on a study of the default behavior of the demographically different groups. The main goal here is to investigate whether the demographic characteristics (*Gender*, *Married*, and *Education*) of the borrower have an

association with the default likelihood. Based on previous evidence there is a mix of contrasting opinions. However, the first hypothesis can be formulated as follows.

H1: The demographic characteristics of borrowers do influence the defaulting behavior. More precisely, the probability of default is lower for women (-), for borrowers having higher education (-), and for married borrowers (-).

The focus of the second part of the analysis is on the borrowers' interest rates and investors' expected returns. The analysis may let emerge, as prior research reported, some differences in the characteristics of the loan that only a specific group receives. However, the hypothesis is grounded on good faith, meaning that no demographic assessment should be present when deciding the interest rates.

H2: the interest rates of the borrowers are not influenced by any demographic variable. Thus, the three demographic variables should bring no significant results as predictors of the interest rate.

Finally, combining the results of the two hypotheses, there may be traces of demographic discrimination whenever one of the groups is associated with a better repayment performance, meaning lower default probability, but still, it experiences higher interest rates. More precisely, higher interest rates can be mapped to a higher perceived risk, making the investor willing to require a higher compensation (i.e. return). However, this higher interest may not be justified if it is determined by non-financial characteristics and it may be even more irrational when the repayment ability is superior.

Chapter 2
The Bondora Case

2.1 Bondora P2P Platform

Bondora is an Estonian platform offering financial services, including P2P unsecured personal loans to borrowers residing in Estonia, Finland, and Spain[5]. The company was founded in 2009. Since that year, it has been able to grow exponentially, becoming one of the European leaders in the sector and positioning itself in the 8th position in the European P2P lending market ranking, with a 3%[6] of market share. According to its loan statistics, Bondora can boast over 980 000 borrowers, with an average loan amount of €2696, a 52 months average loan duration, and a 20.9% average loan interest[7]. Furthermore, the loan procedure that Bondora offers is 100% digitalized, making (almost) everything automated and effortless, for both borrowers and lenders.

The main innovative attribute of Bondora is that of offering a digital environment in which borrowers and investors are directly connected, making it much easier and faster to obtain financing. Not only this but Bondora can also be considered a network enabling cross-border credit operations, which connect investors living either in the EEA or Norway with borrowers living in either Estonia, Finland, or Spain.

The choice of the countries from which people can ask for personal loans reflects a strategic analysis. In fact, in some of these countries, there is a high underbanked rate[8], allowing Bondora to offer a higher rate, as well as more financing opportunities to people that could

[5] The analyzed dataset also includes loans generated in Slovakia. However, Bondora is no longer operating in this credit market.
[6] April 2022, European Crowdfunding Investments, Peer-to-Peer & Online Lending Statistics, P2PMarketData
[7] Updated to the 21st March 2022 and referring to time frame of 14 years of business activity.
[8] Ventura L, "World's Most Unbanked Countries 2021", Global Finance.

not otherwise obtain credit. In addition, it is also worth considering that each country is characterized by a highly complex regulatory framework, that would require an in-depth analysis before expanding the business in it.

Even if the segment targeted by Bondora is primarily the one considered underserved by the banking sector, the platform has in place several procedures and cross-checks to verify the creditworthiness of the borrowers and the truthfulness of the information provided by them.

The loans offered by the platform, which are also those analyzed afterward, have the following characteristics: a principal amount comprised between €100 and €10000, a repayment period of 3 to 60 months, and an annual percentage rate (APR) from 25.11% and 50.85%, depending on the individual circumstances of the borrower. Bondora's business is regulated by the Estonian Financial Supervisory Authority.

From the investor side, the company proposes flexible investment opportunities (with a minimum amount of €1) giving the possibility to pick the program that better fits your investing strategy and risk propension, also thanks to the exploitation of diversification. There are two types of investments[9] involving P2P lending, primarily differing by whether the borrower would bear the whole liability or not. In case of default, the platform claims to be leaving nothing to the case, clearly outlining all parties' obligations.

From the borrower side, the person is left free to disclose, on a voluntary basis, optional information which may improve the creditworthiness, may attract more investors, and may give the chance to be entailed a lower interest rate. As already mentioned above, no collateral will be asked to obtain the loan, making the whole process smooth and fast.

Another key feature of the platform is the transparency: all non-personal data are made available to the public, and most importantly to the potential investors, reducing the information asymmetry that may usually exist between borrowers and lenders. This is

[9] Direct versus indirect investment structure. Indirect structure may be preferred by lenders which are more risk averse.

combined with a very strict verification phase, during which the submitted data is analyzed and cross-referenced with third parties' registries[10].

Once information is collected and verified, Bondora assigns a risk rating to the loan application. The company uses a proprietary scoring system, which reflects the expected loss[11] of the loan by classifying it into eight possible ratings from AA (best) to HR (worse) and, according to their model, no demographic variables are considered in the evaluation.

According to Bondora, the transparency of data offered by the platform together with the rating system should help the lender in making an informed and fair decision.

2.2 Data and Descriptive Statistics

The data employed for the subsequent analysis is retrieved from a publicly available and daily updated dataset that the P2P lending platform Bondora provides on its website under the "Public Reports" section[12]. The raw dataset consists of a collection of loan features, which are not covered by the data protection laws, having a total of 217692 observations and 112 variables each. One observation represents one loan and there can be multiple loans for the same borrower.

The loans are both defaulted and non-defaulted and refer to a timeframe starting from 1st March 2009 up to the 17th February 2022. The different variables available belong to different categories and are of several types, including loan-specific information, as well as demographic characteristics of the borrowers. The default information can be easily retrieved, depending on whether there is a default date for the loan taken into consideration.

[10] This includes local credit bureau, local county court judgment database, population registry, property registry and behavioral data by data vendors, social network and server logs.
[11] Expected loss = Probability of Default * Loss Given Default * Exposure at Default
[12] Link: https://www.bondora.com/it/public-reports#shared-legend

Before starting the analysis, the dataset has been cleaned and pre-processed in order to remove not so useful variables, either being obsolete or having too many missing values. The general data cleaning is described in Figure 1.

2.2.1 Variables Description and Feature Engineering

The number of loans considered after the data cleaning is 200378 and for each of the loans, there is a total of 30 variables. Please refer to Table 1 for a brief description of each variable, together with its data type.

Most of the features are directly provided by Bondora. However, few of them are less straightforward, requiring more attention since they have been obtained performing some feature engineering operations. These variables are: *SATO*, *DScore*, *LoanToIncome,* and *Married*.

First, *SATO* stands for "spread at origination" and is defined as the difference between a loan's interest rate and the average interest rate of loans originated in the same calendar quarter (Fuster, Goldsmith-Pinkham, Ramadorai and Walther, 2021). This variable is extremely useful since it allows to look at how much the interest of a loan differs with respect to the average interest value in the same period. In other words, the higher the SATO, the higher the spread and the higher the return required by investors to accept that loan.

Second, the *DScore* (Chen, Huang, and Ye, 2019) is a measure of how much information the borrower decided to disclose to the lender, presumably trying to increase his/her creditworthiness. It ranges from 0 to 9 and it is built by assigning a "point" each time the borrower discloses one of the following pieces of information: use of the loan, education, marital status, number of dependents, employment status, employment duration of current employer, work experience, work occupation area, and homeownership type.

The rationale behind this variable is to keep track of the amount of optional information disclosed by the borrower while removing the considered features with too many NAs.

Third, the *LoanToIncome*[13] (Chen, Gu, Lui, and Tse, 2020) feature combines the applied loan amount with the total income of the borrower. This will be useful to look in a very interpretable way at the eventual differences between the considered groups of borrowers.

Last, the *Married* variable simply states whether a borrower decided to declare to be married or not. Notice that the not-married case may reflect one of the following two cases: (i) the borrower is not married but has another marital status; or (ii) the borrower decided not to disclose the marital status information on Bondora[14].

2.2.2 Descriptive Statistics

A comprehensive table with the main descriptive statistics for all the single variables is provided in Table 2. However, considering the research question and the willingness to investigate eventual differences between groups of borrowers having different *Gender* or *Education*, it is worth focusing on some descriptive statistics directly considering the different levels of these categorical variables. For this reason, in Table 3.a and 3.b there are reported the mean, standard deviation, median, mix, and max for a subset of relevant variables, differentiating on *Gender*, *Married,* and *Education* values.

For what concerns the *Gender* (Table 3.a), more than half of borrowers on the Bondora platform are of male gender (62.63%). Looking at the statistics, the male gender is characterized, on average, by higher total income, higher interest rates and higher expected loss, with respect to female borrowers. Furthermore, it is interesting to highlight the fact

[13] Only keep loans having a LoanToIncome ration smaller than 20, as the authors suggest because considered to be outliers.
[14] The marital status is optional for the borrower. Thus, the proportion of missing values is very high.

that women are more prone to disclose information (higher *DScore* mean)[15] and they are also requesting for higher loan amount compared to their income, w.r.t. men.

Turning to the *SATO* variable, we can see a substantial difference between males and females: the *SATO* for males has a positive mean, whereas the *SATO* for females has a negative mean. In other words, it seems that women on the Bondora platform have been able to access loans having an interest that is lower than the quarter's average[16]. On the contrary, the same is not true for males.

For what concerns the *Education* (Table 3.b), one would expect, also based on previous literature, to encounter the following relations: the higher the educational level, the lower the expected loss, the higher the income, the better the rating and the lower the expected loss. However, from the statistics there aren't major differences between the different groups and the general expectation is not even always respected (for instance, the lower expected loss on average is associated to the secondary education and not the high one).

Something relevant to highlight is the unusual behavior of the *Basic* level, which is the only group of borrowers going in opposite direction with respect to the general trend. This is reflected by the following findings: (i) the ratings in this group are much worse concerning the average, even if the *Basic* is not even the lower educational level; (ii) the default proportion is almost doubled (60.8%) w.r.t. the overall (33.4%); and (iii) the disclosed information is much more with a median of 9.

Before moving to the methodology, it is useful to outline a deducted profile of the average defaulting borrower in the Bondora platform. This can be done by looking at Table 4, in which the descriptive statistics are reported on the default vs non-default distinction. The defaulting borrower usually has: a lower income, higher *LoanToIncome* ratio, more disclosed information and positive *SATO*. At the same time, Bondora seems to be forecasting rather well the higher probability of default, with higher *ExpectedLoss* and

[15] This has been further tested with a glm with DScore as dependent variable. The result obtained confirms the fact that females disclose higher amount of information then men (coefficient of 3.029e-01 statistically significant).
[16] Recall that the SATO is given by the interest minus the average interest of loans issued in the same quarter.

worse *Rating* associated to the *Default* group. Last, it is interesting to highlight that the portion of males in the defaulting category is much higher w.r.t. the non-defaulting group. Note that in the Figures there are some useful plots, clearly visualizing the discoveries commented along this section.

2.3 Methodology

The main goal is to investigate whether some non-voluntary discriminations and differences in the default inclination exist among groups, distinguished on the basis of demographic characteristics. I now present my empirical strategy.

2.3.1 Logistic Regression

For the analysis of the dataset, both *logit* and *probit* methods have been tried but the results obtained are so similar that only the logit version is presented here.

Logistic Regression (or *logit*) is a classification method for modelling the probability of a categorical binary outcome (i.e. *0-1* or *NoDefault-Default*) given a set of independent variables, either categorical or numerical or both.

The logit method belongs to the family of Generalized Linear Models[17], using as link function the log of odds ratio. The link function takes the following shape:

$$\ln\left(\frac{p}{1-p}\right)$$

where the odd ratio is defined to be $\frac{p}{1-p}$. This means that the equation modelled using the logistic regression, considering x_i to be the independent variables and β_i the corresponding coefficient, has the following formula:

[17] Using a glm() function in R, with family = "binomial" and link function "logit".

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_i * x_i$$

which is also equivalent to:

$$p = \frac{e^{\beta_0+\beta_1*x_1+\cdots+\beta_i*x_i}}{1 + e^{\beta_0+\beta_1*x_1+\cdots+\beta_i*x_i}}$$

The application of a logistic regression method to the Bondora dataset is structured are follows. The variables are divided into three groups: (1) loan specific variables; (2) demographic variables; and (3) credit related variables.

A logit model is constructed for each of the aforementioned groups, including a set of control variables equal for all of them. A summary of the variables classification used to construct the models is reported below.

Logit models' Variables

Here you can find the classification of the variables used to construct the three logit models. At the bottom there are the control variables, included in all the models.

Loan Specific	Demographic	Credit Related
Interest *LoanDuration* *Amount* *AppliedAmount* *Verified* *LoanToIncome* *ExpectedReturn*	*EducationLevel* *Gender* *Age* *Married* *HomeOwnershipType* *IncomeTotal*	*NewCreditCustomer* *NoOfPreviousLoansBeforeLoan* *DebtToIncome* *ExpectedLoss* *Rating* *ExistingLiabilities* *SATO*
Control Variables		
LoanDuration, Interest, Amount, LoanToIncome, Rating and DScore.		

2.3.2 Linear Regression

A Multiple Linear Regression is a statistical technique that uses a set of numerical or categorical independent variables, also called explanatory, to make predictions of the values of a numerical continuous variable Y, called the depended variable. It is an extension of Linear Regression, using multiple variables as predictors and, as such, it is based on the following assumptions: (i) linearity of the relationship between independent and dependent variables; (ii) homoscedasticity; (iii) independence of observations; and (iv) normality.

The multiple linear regression allows to model a linear relationship between the predictors and the response variable, using the following formula:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \epsilon$$

In which Y is the response variable, X is the set of independent variables, β_0 is the intercept and β_i is the set of coefficient associated to each predictor. An error term ϵ is also present.

In addition, each predictor may be associated in a positive (i.e. the higher the predictor's value, the higher the response's value), negative or flat way to the response variable, depending on the sign of the corresponding estimated coefficient. Thus, the coefficients can be easily interpreted to look at how strong the independent-dependent variables relationship is.

Furthermore, the most widely used metric to assess the goodness of fit of the multiple linear regression model is the R^2 statistics. The R-squared represents the proportion of response's variance that is explained by the predictors employed in the model. The formula is the following:

$$R^2 = 1 - \frac{RSS}{TSS}$$

With RSS being the sum of squares of the residuals and TSS being the total sum of squares.

A multiple linear regression will be employed to test the second hypothesis. More precisely, the model uses Interest as response and the demographic variables as independent variables. The control variables are also included. Last, before the model construction, multicollinearity has been checked and the numeric variables have been standardized.

MLR Models

The multiple linear regression model having Interest as dependent variable. Here, the response, independent and control variables are reported. The interactions between the demographic variables are included as predictors.

Response variable	Predictors	Control variables
Interest	*Education* *Verified* *Gender* *Married* *Gender*Married* *Education*Married* *Education*Gender*	*LoanDuration* *Interest* *Amount* *LoanToIncome* *Rating* *Dscore*

2.3.3 Random Forests

The logit and the multiple linear regression are both useful methods to assess and evaluate the two hypotheses. However, they are grounded on too stringent assumption (such as the linearity of the relationship). For this reason, a machine learning approach is also

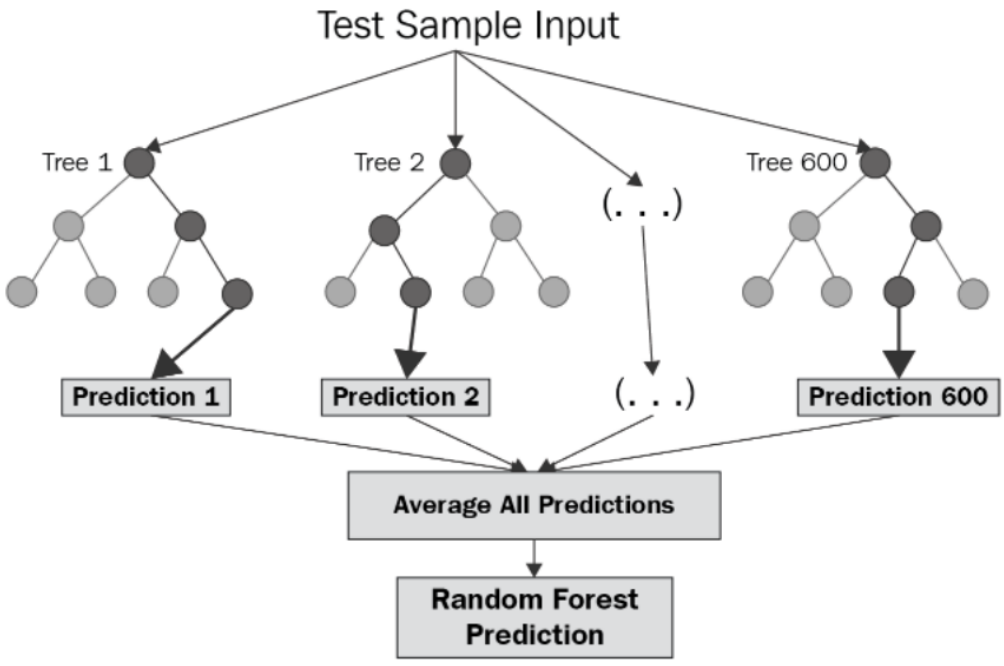
employed, with the main purpose of investigating whether the demographic variables are important predictors even once the assumptions are relaxed.

Random Forests is a supervised Machine Learning algorithm using an ensemble learning method for both classification and regression problems. The Random Forest algorithm is a bagging technique that builds many decisions trees to take the average of the several trees' predictions.

The various trees are trained on different portions of the same training set, allowing to experience a reduction in the variance and, at the same time, boosting both the accuracy and robustness of the model. Below a clear visualization of the function of the algorithm is reported.

Random Forests Structure

The Random Forest prediction is the average of the outcomes coming from a multitude of decision trees. In other words, the predictions of the single trees are aggregated.



Source: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Random Forests are also useful to rank the importance of variables of a classification or regression model. In the importance matrix each predictor will be associated with a metric, tracking its importance in the model.

One importance metric is the Mean Decrease Accuracy, showing how much the model accuracy decreases when dropping that specific variable. In other words, the higher the value of mean decrease accuracy, the higher the importance of the variable in the model.

2.3.4 Marginal Effects

The Logistic Regression, as a generalized linear model, involves a non-linear transformation (i.e. logarithmic) which makes the coefficients not directly interpretable.

For this reason, it is worth considering the marginal effects, which are a useful way to describe the average effect of changes in explanatory variables on the change in the probability of outcomes in logistic regression (Norton, Dowd and Maciejewski, 2019).

The marginal effects, in other words, communicate the rate at which the dependent variable (*Default*) changes at a given point and with respect to one dimension, keeping all the other variables constant. In particular, the interpretation of the result will be based on the Average Marginal Effects (AMEs), which returns a single summary, computing the average, for each of the variables in the model.

Going back to the Bondora dataset analysis, the marginal effects will turn out to be particularly useful for assessing how much the unitary changes in the *Education*, *Married* and *Gender* variables have influence on the *Default* likelihood.

2.3.5 Interactions

An interaction arise when the relation between one predictor (X) and the response variable (Y) depends on the value of another independent variable, for instance Z (Fisher, 1926). In other words, the effects of the independent variables on the dependent variable are moderated by the presence and the values taken by Z, called moderator variable.

The presence of interaction terms has important implications for statistical models, such as linear regression and logistic regression, and the interaction is integrated in the equation as the product of two or more independent variable ($X * Z$).

Some interactions are added in the models, between different pairs of demographic variables (for example, *Married*Gender*). Additionally, the interactions terms are plotted to better visualize the impact of the moderator on the dependent-independent variables relationship.

2.4 Results

The results of the analyses and methods used to test the hypotheses are reported and commented below. First, you can find a table containing the two initial hypothesis.

H1	*The demographic characteristics of borrowers do influence the defaulting behavior. More precisely, the probability of default is lower for women (-), for borrowers having higher education (-), and for married borrowers (-).*
H2	*The interest rates of the borrowers are not influenced by any demographic variable. Thus, the three demographic variables should bring no significant results as predictors of the interest rate.*

2.4.1 First Hypothesis

The first hypothesis investigated the association between demographic predictors and the *Default* as dependent variable. The Table 5 reports the estimated coefficients, the significance and the AIC for each model constructed and, since the results are sufficiently similar across the models, only the *LOGIT* model will be discussed below. Furthermore, the AIC is a relative measure of goodness of fit, useful to compare the performance of the models and it works in the following way: the lower the AIC, the better the model. In this case, the *LOGIT* model is the second best (AIC = 220868.727), only after the Logit with interactions (AIC = 220764.265).

All the three demographic variables of interest resulted to be statistically significant in the prediction of *Default*: *Education* and *Gender* having a negative estimated coefficient and *Married* having a positive one. Furthermore, it is interesting to notice that also the interaction between *Gender*Married* appeared to be significant (see Figure 7). From this results, we can deduce that the demographic characteristics are actually being evaluated by the model but the sign of the association with *Default* will be more evident studying the odd ratios and the marginal effects.

The estimated coefficients displayed are the log odds for the logistic regression. In order to make them more interpretable, it is useful to transform and report the corresponding Odd Ratio (Table 6). The odd ratio would have the following interpretation: when the odd ratio is greater than 1, then a higher value of the predictor increases the likelihood of default. Referring to the model and interpreting the odd ratios of the demographic variables it results that the likelihood of default is lower for female borrowers (0.5528) and higher education borrowers (0.9564). On the contrary, being married results in an increased likelihood of default (1.1443). The last result is consistent with the results obtained by Santoso et al. (2018) in some of the Indonesian platforms.

This is consistent with the Average Marginal Effect[18] (Table 7) obtained for the variables. For *Gender*, changing the value from Male to Female leads to a lower default probability for a total decrease of -0.1097. For *Education*, the AME is negative (-0.0083) once again showing that higher education level leads to lower default. Last, changing the information about the marital status from not given or not married (0) to married (1) translates into an increase of 0.0250 in the default probability.

2.4.2 Second Hypothesis

The second hypothesis is verified only if none of the demographic variables is considered to be significant in the prediction of the *Interest*.

The results reported in Table 8 imply the following considerations. First, for what concerns *Education*, the coefficient is negative (-0.003) but the p-value is too large, making it a non-significant predictor for the interest rate. Second, *Gender* is a significant variable in predicting *Interest* and the Female level is associated with a positive coefficient (0.089), indicating that, according to the linear regression, female borrowers are on average associated with higher interest rates concerning male borrowers[19]. Last, also *Married* is statistically significant with a negative coefficient meaning that being married decreases the interest rate faced by borrowers.

Furthermore, the interactions between all the pairs of the demographic variables have been included in the model. The interesting result is that both *Education*Gender* and *Education*Married* synergies are statistically relevant. First, the *Education*Gender* (Figure 8) shows that depending on the level of the *Gender*, the *Education* is going to have a different effect on *Interest*. More precisely, when the borrower is a female, a unitary

[18] The AMEs reported here are computed examining the partial effects, meaning the contribution of each variable on the outcome scale, conditional on the other variables involved in the link function transformation of the linear predictor (R Documentation).

[19] The Gender variables is included into the model as a factor having two levels: 0 (Male) and 1 (Female). The coefficients for both levels are retrievable in the following way: Gender1 (Female) = Intercept + Gender1 and Gender0 (Male) = Intercept.

change of *Education* will affect with a higher magnitude the interest rate[20], with respect to when the borrower is a male.

Last, the interaction *Education*Married* (Figure 9) is a clear representation of a cross-over interaction. The moderator *Married* allows the sign of Education's coefficient to change from negative (-0.003 when Married = 0) to positive (+0.034 when Married = 1). In other words, the higher the education of married borrowers, the higher the interest faced. The contrary is true for borrowers that did not disclose being married.

The results suggest that the first hypothesis is verified only for two out of the three demographic variables: both *Gender* and *Education* are relevant in predicting the *Default* and both female borrowers and high education borrowers reduced the likelihood of *Default*. For what concerns *Married*, the hypothesis is only partially verified: it is true that *Married* has a statistically significant influence on the defaulting behavior but, contrary to the assumption, married borrowers increase the probability of default, underling a positive relationship between *Married* and *Default*.

Turning to the second hypothesis, the interest rates faced by the borrowers resulted in being affected in a significant way by demographic characteristics such as *Married* and *Gender*. Thus, the second hypothesis is verified only for the *Education* level of the borrower, which does not bring significant results in the prediction of *Interest*. However, it is also true that all the synergies including Education are significant in the model. For this reason, even for Education the second hypothesis is only partially verified. Here I report a table containing a summary of the two hypotheses' results.

[20] Steepest line. Furthermore, the regression line for Male (Gender = 0) is Interest = -2.403 -0.003Education while for Gender = 1 is Interest = -2.305 – 0.033Education.

	H1	H2
Gender	Verified. "Female" reduces Default (-)	Not verified. Significant and higher interest for females (+)
Education	Verified. "High education" reduces Default (-)	Partially verified. Not significant unless interactions are considered.
Married	Partially verified. Being married increases Default (+)	Not verified. Significant and lower interest for married (-)

Demographic discrimination in the P2P lending market may be present whenever a group of borrowers, identified on the basis of demographic characteristics, has an at least equally good ability of repayment (i.e. same or lower default probability, with respect to the other group) but it is still associated with disadvantageous loan conditions (i.e. higher interest rates). Below there is the discussion for each demographic variable.

The first variable analyzed is *Gender*. On one side, female borrowers in the Bondora P2P lending platform have a lower default rate. On the other side, the average interest rate to which female borrowers are subject is higher than with respect to the males. In other words, the loan conditions faced by females are worse than those of males. Combining the results, there are traces of discrimination based on the gender of the borrower, penalizing women. This is consistent with what was discovered by Chen et al. (2020) in China.

Moving to *Married*, the group including the borrowers that decided to disclose the information of being married has a higher default rate, with respect to those that either did not disclose the information or that are not married. This is consistent with the findings of Santoso et al. (2018) but the opposite of what Chen et al. (2019) discovered, suggesting that the contrasting results may be due to exogenous variables specific of the geographical area considered. The married borrowers, however, are also associated with lower interest rates, not evaluating in a correct way the risk associated with the higher default. This means

that, based on the framework employed, discrimination against those not being married or not willing to disclose it is present.

Last, no discrimination evidence is present for the *Education* variable. In fact, even if higher education reduces the default probability, as previously discovered also by Santoso et al. (2018), there are no disadvantageous loan terms for those having a higher education.

2.4.4 Random Forest Check

The methods employed up to now for testing the hypotheses rely on many stringent assumptions. The flexibility could be improved using a machine learning approach, such as Random Forests, that is able to better capture the relationship between the dependent and the independent variables. A machine learning approach would, on one side, considerably increase the predicting power of the model but, on the other side, the interpretability is decreased, requiring more tests.

Fuster et al. (2017) adopted a slightly different approach, focusing their research on whether a change in the technology (from traditional to machine learning models) can influence the default distribution across categories of borrowers. The results they obtained revealed that, in the US mortgage market, improving the technology leads to an increased disparity across different borrowers' groups, with respect to traditional technology.

For this reason, while keeping valid the discrimination results obtained so far, the last step of this study is comparing the prediction power of the traditional models (Logistic Regression and Linear Regression) to the one of Random Forests. For this purpose, the following procedure is employed: (i) the dataset is split into training and test set[21]; (ii) for both Default and Interest, traditional models and machine learning models are trained and then tested; (iii) the predictors are the same of the models constructed in the previous

[21] Training set containing 70%, randomly sampled, of the observation and the test set containing the remaining 30%.

section; (iv) prediction accuracy is evaluated for all models in the test set and results are reported below.

Models Comparison

The comparison of the random forests versus the logit and linear regression. For the classification of Default, the random forests and logit have been compared using the AUC score. For the Interest regression, the random forests and linear regression have been compared using the R squared. The reported results refer to the test set.

	Default (AUC)	**Interest (R^2)**
Random Forests	0.823	0.894
Logit or Linear Regression	0.743	0.735

It is clear that the best performing technology, in both Default classification and Interest regression, is the Random Forests one, with an Area Under the Curve equal to 0.823 (for the Default) and an R-squared of 0.894 (for the Interest). The former (AUC) indicates how good the model is for correctly classifying the two classes. The latter (R-squared) tells the goodness of the model in explaining the dependent variable's variability.

Last, the Random Forests method allows checking the importance of the single variables considered as predictors. For this study, the relevance of this feature is looking at how important the demographic variables are considered by the machine learning model. Table 9 and Table 10 show that the three least important variables are the demographic ones of *Gender*, *Education*, and *Married*.

The result can be interpreted in the following way: the main accuracy of the model is not derived from the demographic characteristics of the borrower.

2.5 Conclusion

The main purpose of this research is to yield some evidence about the existence of implicit demographic discrimination in the peer-to-peer lending market. Discrimination is tested by contrasting the default likelihood with the loan terms obtained, across groups of borrowers distinguished on demographic characteristics.

The results, stemming from both traditional and machine learning techniques, suggest that some kind of demographic discrimination exists when it comes to the gender and the marital status of the borrower. More precisely, the analysis suggests that there is a tendency to disfavor female borrowers and non-married borrowers.

These findings contribute to the existing literature without, however, providing a generalizable judgment concerning the peer-to-peer lending demographic discrimination. In fact, even if the results are consistent with what was previously discovered by other researchers (Chen et al., 2020; Chen et al, 2019; Santoso et al., 2018), there are still a lot of discrepancies leading to different results (Pope and Syndor, 2011; Kim, 2020), depending on the platform and geographical area considered.

The research could be expanded to other demographic variables, replicating the methodology employed. For instance, the age of the borrower would be particularly interesting to test, considering that Kim (2020) and Pope and Syndor (2011) found the existence of age discrimination in the peer-to-peer lending market. Additional curiosity on the borrower's age stems also from the variable importance's results of the Random Forests model (Table 9), showing considerable importance associated with *Age* in predicting the *Default*, even greater than the other demographic variables.

Concluding, additional studies on the topic are needed, especially with more focus on the machine learning approach. Even if there are clear signs of demographic discrimination, their magnitude requires more attention for the following reason: on one side, prior research provides evidence of the fact that the introduction of a machine learning approach enlarged the distributional gap found with traditional methods; on the other side, the

Random Forests model reveals that the demographic variables are contributing less than the others.

References

Atz U., Bholat D., "Peer-to-peer lending and financial innovation in the United Kingdom", 7 May 2016, Bank of England Working Paper.

Baxter G., Rengarajan S., "The march towards digital money: bringing the underbanked in from the cold", March 2017, Citi, Imperial College London.

Byrnes J.P., Miller D.C., Schafer W.D., "Gender differences in risk taking: A meta-analysis", 1999, Psychological Bulletin.

Chajure A., "Random Forest Regression", 29 June 2019, The Startup.

Chen S., Gu Y., Lui Q. and Tse Y., "How do lenders evaluate borrowers in peer-to-peer lending in China?", 2020, International Review of Economics and Finance.

Chen X., Huang B. and Ye D., "The gender gap in peer-to-peer lending: evidence from the people's republic of China", July 2019, Asian Development Bank Institute.

Coleman A., "Why Business Is Booming In The Baltics", 20 September 2015, Forbes.

Cummins, K. M., Diep, S. A., and Brown, S. A., "Alcohol expectancies moderate the association between school connectedness and alcohol consumption", 2019, Journal of School Health.

Demirguc-Kunt A. and Klapper L., "Measuring Financial Inclusion: The Global Findex Database", April 2012, The World Bank, Policy Research Working Paper.

Dorfleitner G., Priberny C., Schuster S., Stoiber J., Weber M., De Castro I. and Kammler J., "Text related soft information in peer-to-peer lending – Evidence from two leading European platforms", 2016, Journal of Banking & Finance.

Faridi O., "European Lender Bondora has Now Returned €64M to Investors with €571M Invested", 28 March 2022, Crowdfund Insider.

Fisher R.A., "The Arrangement of Field Experiments", 1926, Journal of the Ministry of Agriculture of Great Britain.

Fuster A., Goldsmith-Pinkham P. S., Ramadorai T., and Walther A., "Predictably Unequal? The Effects of Machine Learning on Credit Markets", November 2017, Journal of Finance.

Guy Birken E., "The costs of being unbanked or underbanked", 28 July 2020, Forbes Advisor.

Han S., "On the Economics of Discrimination in Credit Market", October 2001, United States Federal Reserve Board.

Harkness S. K., "Discrimination in Lending Markets: Status and the Intersection of Gender and Race", 2016, Social Psychology Quarterly, American Sociological Association.

Havrylchyk O., Mariotto C., Rahim T., Verdier M., "What drives the expansion of the peer-to-peer lending?", 2017, European Banking Authority.

European Commission, Directorate-General for Energy, "Impact Assessment Report accompanying the Proposal for a Directive of the European Parliament and of the Council on energy efficiency", 17 July 2021, Commission Staff Working Document, EU Commission.

Iyer R., Khwaya Ijaz A., Luttmer E., Shue K., "Screening Peers softly: Inferring the Quality of Small Borrowers", 12 August 2015, Journal of Management Science.

Kim D., "Sexism and Ageism in P2P Lending Market: Evidence from Korea", 05 May 2020, Journal of Asian Finance.

Klein G., Shtudiner Z. and Zwilling M., "Why do peer-to-peer (P2P) lending platforms fail? The gap between P2P lenders' preferences and the platforms' intentions", 25 May 2021, Electronic Commerce Research.

Ladd H. F., "Evidence on Discrimination in Mortgage Lending", 1998, Journal of Economic Perspectives.

Milne A., Parboteeah P., "The Business Models and Economics of Peer-to-Peer Lending", May 2016, European Credit Research Institute.

Morgenson G., "Lending Club, a Story Stock That Skimped on the Details", 13 May 2016, The New York Times

Nickerson D., "Credit Risk, Regulatory Costs and Lending Discrimination in Efficient Residential Mortgage Markets", 12 April 2022, Journal of Risk and Financial Management.

Norton EC, Dowd BE, Maciejewski ML., "Marginal Effects—Quantifying the Effect of Changes in Risk Factors in Logistic Regression Models", 2019, JAMA.

Pope D., Sydnor J, “What’s in a picture?: Evidence of discrimination from Prosper.com”, January 2011, Journal of Human Resources.

Ribeiro-Navarrete S., Pineiro-Chousa J., Lopez-Cabarcos M. A., Palacios-Marqués D., “Crowdlending: mapping the core literature and research frontiers”, 10 November 2021, Review of Managerial Science.

Santoso W., Trinugroho I., Risfandy T. and Prabowo M., “What determine loan rate and default status in financial technology online direct lending? Evidence from Indonesia”, March 2018, Otoritas Jasa Keuangan.

Savarese C., “Crowdfunding and P2P lending: which opportunities for Microfinance?”, May 2015, European Microfinance Network.

Say N., “Bondora Review: P2P Loans Investment Platform”, 9 June 2021, Money Check.

Serrano-Cinca C., Gutiérrez-Nieto B., Lopez-Palacios L., “Determinants of Default in P2P Lending”, 2015, PLoS ONE.

Turner M. A. and Skidmore F, “Mortgage Lending Discrimination: A Review of Existing Evidence”, 1999, The Urban Institute.

Xu X., “The underbanked phenomena”, 2019, Journal of Financial Economic Policy.

Ziegler, T., Shneor, R., Garvey, K., Wenzlaff, K., Yerolemou, N., Rui, H., Zhang, B., “Expanding Horizons: The 3rd European Alternative Finance Industry Report”, 2018, Cambridge Center for Alternative Finance.

Tables

Table 1
Variables Description

This table lists the variables used in the different models. For each variable a short description is present, together with the type of variable and the eventual possible levels if it is a categorical variable. The IncomeTotal, Amount, AppliedAmount and LiabilitiesTotal variables are expressed in Euro. The Income and LiabilitiesTotal variables refer to the monthly amount.

Variable	Type	Description
Age	Numerical, discrete	The age of the borrower at origination.
Gender	Numerical	The gender of the borrower. The levels are: Male and Female.
Education	Categorical	Education level of the borrower at origination. The levels are: Primary, Basic, Secondary and Higher (including also Vocational). The levels are mapped with integer numbers between 1-4.
HomeOwnership Type	Categorical	Home ownership type of the borrower at origination. The levels are: Tenant (including pre furnished, unfurnished and joint tenant), Owner (including joint ownership), Living with Parents, Mortgage and Other (including owner with encumbrance, homeless, council house and others).
Married	Numerical	Married variable is equal to "1" if the borrower declared to be married and "0" otherwise. Notice that the "0" includes also the not given case.

IncomeTotal	Numerical, continuous	The total income of the borrower (monthly basis). It is the sum of the income coming from employment, pension, paternity leave, child support, social support, alimony payments and others.
NewCreditCustomer	Logical	Whether or not the borrower has had prior credit history with Bondora. The value is "1" if the customer is new and "0" if the customer had at least 3 months of credit history.
NoOfPreviousLoansBeforeLoan	Numerical, discrete	Number of previous loans.
Year	Numerical	The year of origination of the loan.
AppliedAmount	Numerical, continuous	The amount of the loan the borrower applied for at origination.
Amount	Numerical, continuous	The amount received by the borrower on the Primary Market.
Interest	Numerical, continuous	Maximum interest rate accepted in the loan application.
LoanDuration	Numerical	The current loan duration in months.
Verified	Categorical	Whether the information associated with the loan has been verified or not. The levels are: 0 for unverified and 1 for verified.
DebtToIncome	Numerical, continuous	Ratio of borrower's monthly gross income that goes toward paying loans.

SATO	Numerical, continuous	Spread at origination. The difference between a loan's interest rate and the average interest rate of loans originated in the same calendar quarter (Fuster, Goldsmith-Pinkham, Ramadorai and Walther, 2021).
LoanToIncome	Numerical, continuous	The ratio between the applied amount variable and the total income of the borrower.
ExpectedLoss	Numerical, continuous	Expected loss computed according to the rating model of Bondora. It is the product of the probability of default, loss given default and exposure at default.
ExpectedReturn	Numerical, continuous	Expected Return calculated by the current Rating model.
Rating	Numerical, discrete (1-8)	The rating of Bondora issued by the rating model. The levels are: AA (best), A, B, C, D, E, F, HR (worse). For the model construction, the rating is an integer number between 1-8.
ExistingLiabilities	Numerical, continuous	The number of existing liabilities of the borrower.
LiabilitiesTotal	Numerical, continuous	The total monthly liabilities.
DScore	Numerical, discrete (0-9)	DScore stands for "Disclosure Score" and it is a measure of the amount of information disclosed by the borrower. The maximum value is 9, which corresponds to borrowers disclosing all optional information.
Default	Categorical	Whether or not the loan went into defaulted state and collection process was started.

Table 2.a
Summary Statistics

The table is a summary statistics for all the variables considered in the analysis. For each numerical variable there is the mean (1), standard deviation (2), minimum (3), maximum (4) and standard error (5).

	Mean	**Std. dev.**	**Min**	**Max**	**SE**
	(1)	(2)	(3)	(4)	(5)
Age	40.47	12.44	18	75	0.03
IncomeTotal	1498.97	793.89	500	3394	1.77
NewCreditCustomer	0.54	0.5	0	1	0
NoOfPreviousLoansBeforeLoan	1.64	2.56	0	27	0.01
Amount	2544.64	1921.28	530	6911	4.29
AppliedAmount	2669.31	2055.59	530	7442	4.59
Interest	28.87	13.51	11.4	59.73	0.03
LoanDuration	49.01	16.61	1	120	0.04
Verified	0.74	0.44	0	1	0
DebtToIncome	3.97	10.27	0	37.37	0.02
ExpectedLoss	0.12	0.1	0	1.01	0
ExpectedReturn	0.12	0.06	-0.8	0.69	0
ExistingLiabilities	2.91	3.15	0	40	0.01
LiabilitiesTotal	364.75	363.77	0	1283	0.81
DScore	3.95	2.18	3	9	0
SATO	-0.7	12.42	-20.51	24.41	0.03
LoanToIncome	2.08	1.78	0.29	6.78	0
Married	0.05	0.21	0	1	0

Table 2.b
Summary Statistics

Furthermore, for the categorical variables there is the absolute frequency (2), the relative frequency (3) of each level and the cumulative percentage (4).

	Level	**Frequency**	**Percent**	**Cumulative %**
	(1)	(2)	(3)	(4)
Verified	1	148893	74.3	74.3
	0	51485	25.7	100
Rating	AA	8213	4.1	4.1
	A	9553	4.8	8.9
	B	28935	14.4	23.3
	C	41530	20.7	44
	D	46230	23.1	67.1
	E	34682	17.3	84.4
	F	19723	9.8	94.2
	HR	11512	5.7	100
Default	0	133532	66.6	66.6
	1	66846	33.4	100

Table 3.a
Descriptive Statistics for Gender

This table reports summary statistics for the different levels of the *Gender* variable (Male and Female). It reports the mean, standard deviation, median, minimum and maximum for the numerical variables and the absolute and relative frequency for the categorical variables. The *Income* and *AppliedAmount* variables are expressed in Euro. The *IncomeTotal* variable refers to the monthly amount. The *DScore* is the amount of disclosed information on a voluntary basis (ranging from 0 to 9). The *SATO* is the spread at origination, i.e. the difference between the loan's interest and the average interest in the same financial quarter. The "*t-Test*" column contains the means difference and the t-test value (for categorical variables a Chi-squared test of independence is used).

	Male (N=125494)	Female (N=74884)	t-Test	p-value
IncomeTotal				
Mean (SD)	1580 (819)	1360 (730)	220.314	<2e-16
Median [Min, Max]	1370 [502, 3400]	1200 [502, 3400]	(62.386)	
AppliedAmount				
Mean (SD)	2660 (2060)	2690 (2040)	-36.03	0.000141
Median [Min, Max]	2130 [530, 7440]	2130 [530, 7440]	(-3.807)	
SATO				
Mean (SD)	-0.810 (13.1)	-0.515 (11.1)	-0.295	7.41e-08
Median [Min, Max]	-2.54 [-20.5, 24.4]	-1.78 [-20.5, 24.4]	(-5.381)	
Interest				
Mean (SD)	29.4 (14.1)	28.0 (12.4)	1.404	<2e-16
Median [Min, Max]	26.0 [11.5, 59.7]	25.0 [11.5, 59.7]	(23.301)	
LoanToIncome				
Mean (SD)	1.95 (1.70)	2.30 (1.89)	-0.342	<2e-16
Median [Min, Max]	1.38 [0.277, 6.75]	1.67 [0.277, 6.75]	(-40.728)	
ExpectedLoss				
Mean (SD)	0.123 (0.101)	0.109 (0.0863)	0.014	<2e-16
Median [Min, Max]	0.0923 [0.000848, 0.972]	0.0900 [0.000724, 1.01]	(31.968)	
Rating				
A	6326 (5.0%)	3227 (4.3%)	(1663.627)	<2e-16
AA	4733 (3.8%)	3480 (4.6%)		
B	17793 (14.2%)	11142 (14.9%)		
C	25218 (20.1%)	16312 (21.8%)		
D	27832 (22.2%)	18398 (24.6%)		
E	20996 (16.7%)	13686 (18.3%)		
F	14199 (11.3%)	5524 (7.4%)		
HR	8397 (6.7%)	3115 (4.2%)		
DScore				
Mean (SD)	3.85 (2.08)	4.11 (2.32)	-0.262	<2e-16
Median [Min, Max]	3.00 [3.00, 9.00]	3.00 [3.00, 9.00]	(-25.337)	
DEFAULT				
0	78574 (62.6%)	54958 (73.4%)	(2450.638)	<2e-16
1	46920 (37.4%)	19926 (26.6%)		

Table 3.b
Descriptive Statistics for Education

This table reports summary statistics for the different levels of the Education variable (Primary, Basic, Vocational, Secondary, High). Please refer to Table 3.a for variable-specific description.

	Primary (N=24284)	Basic (N=5874)	Secondary (N=70429)	High (N=99791)
IncomeTotal				
Mean (SD)	1330 (710)	1120 (630)	1260 (674)	1730 (831)
Median [Min, Max]	1110 [502, 3400]	918 [502, 3400]	1090 [502, 3400]	1600 [502, 3400]
AppliedAmount				
Mean (SD)	2610 (2040)	2430 (2040)	2460 (2060)	2840 (2040)
Median [Min, Max]	2130 [530, 7440]	1910 [530, 7440]	1870 [530, 7440]	2340 [530, 7440]
SATO				
Mean (SD)	1.78 (12.1)	-1.77 (11.2)	-2.73 (12.3)	0.195 (12.4)
Median [Min, Max]	-0.0819 [-20.5, 24.4]	-1.69 [-20.5, 24.4]	-3.94 [-20.5, 24.4]	-1.49 [-20.5, 24.4]
Interest				
Mean (SD)	30.5 (13.3)	30.1 (11.5)	27.3 (12.5)	29.5 (14.2)
Median [Min, Max]	29.0 [11.5, 59.7]	29.0 [11.5, 59.7]	24.9 [11.5, 59.7]	25.2 [11.5, 59.7]
LoanToIncome				
Mean (SD)	2.22 (1.85)	2.47 (1.95)	2.22 (1.88)	1.93 (1.66)
Median [Min, Max]	1.63 [0.277, 6.75]	1.80 [0.277, 6.75]	1.52 [0.277, 6.75]	1.41 [0.277, 6.75]
ExpectedLoss				
Mean (SD)	0.116 (0.0766)	0.157 (0.128)	0.107 (0.0943)	0.124 (0.0987)
Median [Min, Max]	0.0983 [0.000724, 0.746]	0.120 [0.00721, 0.885]	0.0880 [0.000848, 1.01]	0.0900 [0.000848, 0.999]
Rating				
A	753 (3.1%)	134 (2.3%)	4704 (6.7%)	3962 (4.0%)
AA	736 (3.0%)	28 (0.5%)	3446 (4.9%)	4003 (4.0%)
B	3186 (13.1%)	700 (11.9%)	12505 (17.8%)	12544 (12.6%)
C	4868 (20.0%)	1146 (19.5%)	15503 (22.0%)	20013 (20.1%)
D	5958 (24.5%)	1196 (20.4%)	15262 (21.7%)	23814 (23.9%)
E	5581 (23.0%)	1115 (19.0%)	10814 (15.4%)	17172 (17.2%)
F	2417 (10.0%)	652 (11.1%)	4833 (6.9%)	11821 (11.8%)
HR	785 (3.2%)	903 (15.4%)	3362 (4.8%)	6462 (6.5%)
DScore				
Mean (SD)	3.07 (0.650)	7.28 (2.68)	4.09 (2.31)	3.86 (2.10)
Median [Min, Max]	3.00 [3.00, 9.00]	9.00 [3.00, 9.00]	3.00 [3.00, 9.00]	3.00 [3.00, 9.00]
DEFAULT				
0	16570 (68.2%)	2303 (39.2%)	47271 (67.1%)	67388 (67.5%)
1	7714 (31.8%)	3571 (60.8%)	23158 (32.9%)	32403 (32.5%)

Table 4
Descriptive Statistics for Default

This table reports summary statistics for the default vs non-default borrowers. Please refer to Table 3.a for variable-specific description. The "*t-Test*" column contains the means difference and the t-test value (for categorical variables a Chi-squared test of independence is used).

	0 (N=133532)	1 (N=66846)	t-Test	p-value
IncomeTotal				
Mean (SD)	1510 (808)	1480 (766)	24.984	1.43e-11
Median [Min, Max]	1290 [502, 3400]	1300 [502, 3400]	(6.756)	
AppliedAmount				
Mean (SD)	2570 (2020)	2870 (2100)	-297.312	<2e-16
Median [Min, Max]	2080 [530, 7440]	2230 [530, 7440]	(-30.208)	
SATO				
Mean (SD)	-2.34 (11.1)	2.58 (14.1)	-4.918	<2e-16
Median [Min, Max]	-3.28 [-20.5, 24.4]	1.54 [-20.5, 24.4]	(-78.645)	
Interest				
Mean (SD)	25.7 (11.6)	35.2 (14.8)	-9.446	<2e-16
Median [Min, Max]	21.1 [11.5, 59.7]	33.6 [11.5, 59.7]	(-144.27)	
DebtToIncome				
Mean (SD)	2.73 (8.66)	6.43 (12.5)	-3.706	<2e-16
Median [Min, Max]	0 [0, 37.2]	0 [0, 37.2]	(-69.015)	
LoanToIncome				
Mean (SD)	2.01 (1.77)	2.23 (1.79)	-0.223	<2e-16
Median [Min, Max]	1.38 [0.277, 6.75]	1.66 [0.277, 6.75]	(-26.382)	
ExpectedLoss				
Mean (SD)	0.0938 (0.0673)	0.166 (0.123)	-0.072	<2e-16
Median [Min, Max]	0.0855 [0.000724, 0.837]	0.144 [0.000859, 1.01]	(-141.465)	
DScore				
Mean (SD)	3.64 (1.85)	4.55 (2.62)	-0.907	<2e-16
Median [Min, Max]	3.00 [3.00, 9.00]	3.00 [3.00, 9.00]	(-80.108)	
Rating				
A	7482 (5.6%)	2071 (3.1%)	(27758.554)	<2e-16
AA	6882 (5.2%)	1331 (2.0%)		
B	23675 (17.7%)	5260 (7.9%)		
C	32516 (24.4%)	9014 (13.5%)		
D	34015 (25.5%)	12215 (18.3%)		
E	18984 (14.2%)	15698 (23.5%)		
F	7128 (5.3%)	12595 (18.8%)		
HR	2850 (2.1%)	8662 (13.0%)		
Gender				
Male	78574 (58.8%)	46920 (70.2%)	(2450.638)	<2e-16
Female	54958 (41.2%)	19926 (29.8%)		
Education				
Mean (SD)	3.55 (1.22)	3.50 (1.22)	0.055	<2e-16
Median [Min, Max]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	(9.444)	

Table 5
Results of the Logit models

The results of the logistic regression models and the probit model are reported here. For each model, there are the estimated coefficients of each predictor, the significance level based on the p-value, and the standard error in parenthesis. At the bottom of the table, there is the AIC value, useful for model comparison. The models from (3) to (5) include, among the demographic variables of interest, only the one present in the name. The model (6) is equal to the (1) but for the demographic variables' interactions included.

	Logit	**Probit**	**Logit Education**	**Logit Gender**	**Logit Married**	**Logit Interactions**
	(1)	(2)	(3)	(4)	(5)	(6)
Education	-0.045 ***	-0.027 ***	-0.047 ***			-0.039 ***
	(0.005)	(0.003)	(0.005)			(0.007)
Gender	-0.593 ***	-0.364 ***		-0.576 ***		-0.550 ***
	(0.011)	(0.007)		(0.011)		(0.037)
Age	0.010 ***	0.006 ***				0.010 ***
	(0.000)	(0.000)				(0.000)
Married	0.135 ***	0.079 ***			0.175 ***	-0.278 **
	(0.027)	(0.016)			(0.026)	(0.108)
IncomeTotal	-0.000 ***	-0.000 ***	-0.000 ***	-0.000 ***	-0.000 ***	-0.000 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
DScore	0.168 ***	0.102 ***	0.161 ***	0.170 ***	0.151 ***	0.169 ***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
LoanDuration	-0.001 ***	-0.001 ***	-0.003 ***	-0.001 ***	-0.003 ***	-0.001 ***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Interest	0.034 ***	0.021 ***	0.033 ***	0.034 ***	0.033 ***	0.034 ***
	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Amount	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LoanToIncome	0.061 ***	0.037 ***	0.060 ***	0.065 ***	0.060 ***	0.060 ***
	(0.006)	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)
Rating	0.216 ***	0.116 ***	0.217 ***	0.209 ***	0.220 ***	0.216 ***
	(0.006)	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)
Gender:Married						0.464 ***
						(0.045)
Education:Gender						-0.022 *
						(0.011)
Education:Married						0.059
						(0.031)
Constant	-3.385 ***	-1.983 ***	-3.186 ***	-3.141 ***	-3.291 ***	-3.399 ***
	(0.036)	(0.021)	(0.032)	(0.028)	(0.029)	(0.038)
N	200378	200378	200378	200378	200378	200378
AIC	220868.727	220989.709	224188.314	221526.401	224223.410	220764.265

*** p < 0.001; ** p < 0.01; * p < 0.05.

Table 6
Odd Ratios for Logit model

The table reports each variable (predictor) of the "LOGIT" model with the corresponding odd ratio. The odd ratios are computed starting from the estimated coefficients reported in the previous table. More precisely, the odd ratio is equal to the exponential of the coefficient in the summary table. An odd ratio greater than 1 increases the likelihood of default for high values of the predictor. An odd ratio smaller than 1 decreases the likelihood of default for high values of the predictor.

Variable	Odd Ratio
Intercept	0.034
Education	0.956
Gender	0.553
Age	1.010
Married	1.144
IncomeTotal	1.000
DScore	1.183
LoanDuration	0.999
Interest	1.034
Amount	1.000
LoanToIncome	1.063
Rating	1.241

Table 7
Marginal Effects for Logit model

The R package margins() is used to compute the Average Marginal Effects (1) of the predictors of the "LOGIT" model. The results are reported here, together with the standard error (2), lower (5)and upper bound (6). The Average Marginal Effect reported here is computed examining the partial effects, meaning the contribution of each variable on the outcome scale, conditional on the other variables involved in the link function transformation of the linear predictor (R Documentation).

	AME	**SE**	**z**	**p**	**lower**	**upper**
	(1)	(2)	(3)	(4)	(5)	(6)
Age	0.002	0.000	23.757	0.000	0.002	0.002
Amount	0.000	0.000	0.812	0.416	0.000	0.000
Dscore	0.031	0.001	63.088	0.000	0.030	0.032
Education	-0.008	0.001	-8.295	0.000	-0.010	-0.006
Gender	-0.110	0.002	-54.079	0.000	-0.114	-0.106
IncomeTotal	0.000	0.000	-13.196	0.000	0.000	0.000
Interest	0.006	0.000	47.121	0.000	0.006	0.007
LoanDuration	0.000	0.000	-3.751	0.000	0.000	0.000
LoanToIncome	0.011	0.001	9.584	0.000	0.009	0.014
Married	0.025	0.005	5.075	0.000	0.015	0.035
Rating	0.040	0.001	35.134	0.000	0.038	0.042

Table 8
Results of Multiple Linear Regression

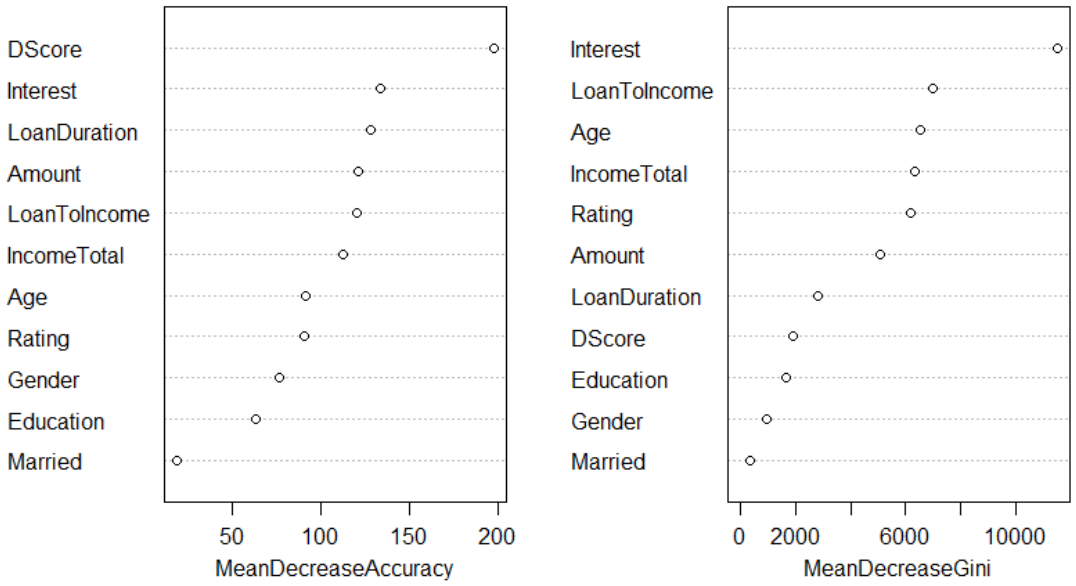
The results of the multiple linear regression model are reported here[22]. The model called "INTEREST" uses as dependent variable the Interest of the loan. There are the estimated coefficients of each predictor, the significance level based on the p-value and the standard error in parenthesis. At the bottom of the table the R squared and AIC are reported.

	Interest Regression
	(1)
LoanDuration	-0.025 ***
	(0.001)
Amount	-0.000 ***
	(0.000)
LoanToIncome	0.222 ***
	(0.014)
Rating	6.904 ***
	(0.009)
DScore	-0.644 ***
	(0.009)
Education	-0.037
	(0.021)
Verified1	-0.282 ***
	(0.036)
Gender1	1.324 ***
	(0.111)
Married1	-0.755 *
	(0.352)
Gender1:Married1	0.101
	(0.149)
Education:Gender1	-0.399 ***
	(0.033)
Education:Married1	0.498 ***
	(0.100)
Constant	0.322 **
	(0.100)
N	200378
R2	0.731
AIC	1348579.482
*** p < 0.001; ** p < 0.01; * p < 0.05.	

[22] The numerical variables have been standardized before the model construction.

Table 9
Variables Importance Random Forest: Default

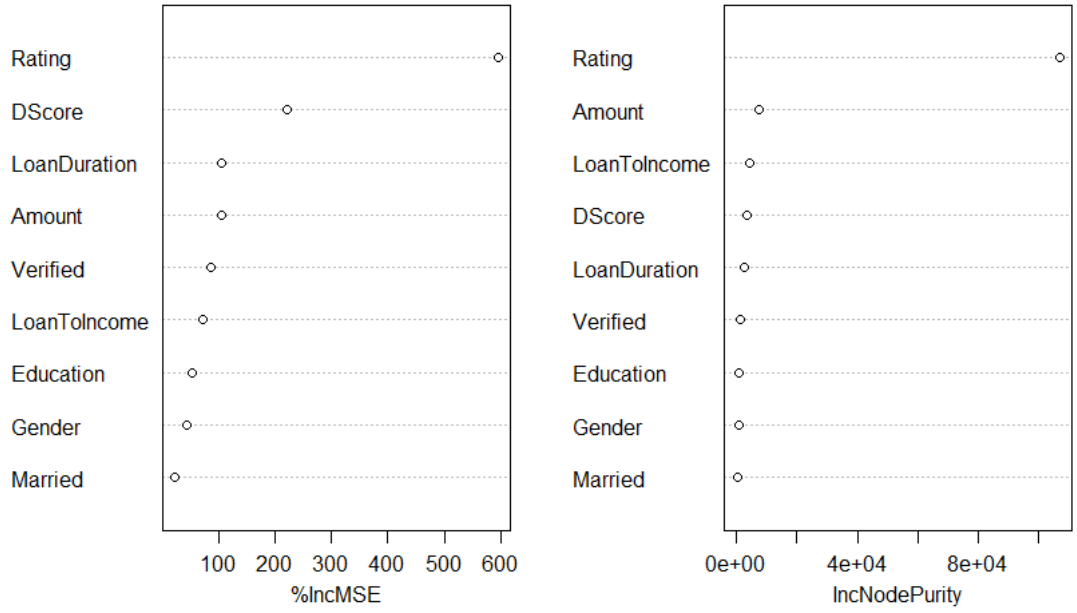
Importance of the variables used as predictors in the Random Forest classification model with *Default* as dependent variable. The variables are displayed in order of importance according to two different indexes: the Mean Decrease Accuracy (i.e. how much the accuracy of the model drops not considering that variable) and the Mean Decrease Gini (using the Gini Impurity index).



	MeanDecreaseAccuracy	MeanDecreaseGini
Education	63.269	1668.089
Gender	76.097	981.459
Age	90.857	6530.664
Married	18.271	360.933
IncomeTotal	112.264	6347.431
DScore	197.667	1890.671
LoanDuration	127.888	2818.298
Interest	133.436	11499.325
Amount	121.115	5075.183
LoanToIncome	120.210	6957.302
Rating	90.025	6170.452

Table 10
Variables Importance Random Forest: Interest

Importance of the variables used as predictors in the Random Forest regression model with *Interest* as dependent variable. The variables are displayed in order of importance according to two different indexes: the %IncMSE (i.e. the increase in MSE when the variable is randomly permuted) and IncNodePurity (i.e. increase in node purity).



	%IncMSE	IncNodePurity
LoanDuration	0.033	2555.742
Amount	0.121	7414.081
LoanToIncome	0.045	4224.642
Rating	1.651	106955.699
DScore	0.084	3301.850
Education	0.009	908.176
Verified	0.023	1097.222
Gender	0.006	526.860
Married	0.004	398.760

Figures

Figure 1.a
General Data Cleaning

Filtering on the age has been applied, to keep only loans issued to borrowers having at least 18 years, as this is the age limit imposed by the platform. In addition, the genders considered for the analysis are only "Male" and "Female", dropping the "Undefined" case, which is characterized by extreme values that make unwillingly high the likelihood of biased results. Then, LoanToIncome can be only greater than 20 and all the observations having Education equal to -1 are dropped (-1 indicates unknown or not specified).

Furthermore, *IncomeTotal*, *AppliedAmount*, *Amount*, *Interest*, *DebtToIncome*, *SATO*, *LoanToIncome*, and *LiabilitiesTotal* have been winsorized[23] using the 90% quantile.

Step of Data Cleaning	Number of Observations
Before Data Cleaning	217692
Age ≥ 18	217639
NA	214954
Gender (exclude Undefined)	202489
LoanToIncome ≤ 20	202223
Education ≠ -1	200378

[23] Winsorisation is useful to replace extreme values with less extreme ones. In this case, the 5% smallest and the 5% largest values are replaced by less extreme values.

Figure 1.b
General Data Cleaning

For what concerns the variables subset selection, the choice of the features to keep or remove is made according to the following two criteria: (i) on the basis of the proportion of missing values; and (ii) keeping in consideration the literature review reported before. The number of missing values for each variable before and after subset selection.

The first plot shows the number of missing values for the 20 variables having the highest percentage of missing.

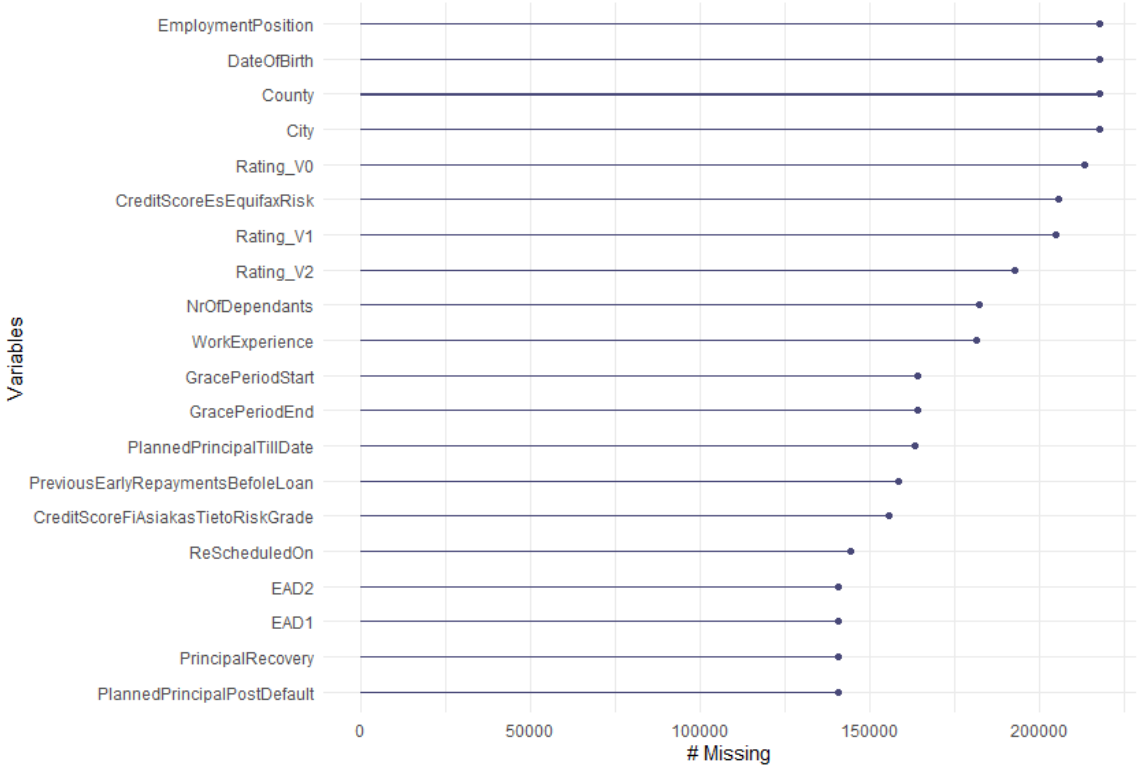


Figure 1.c
General Data Cleaning

The second plot shows the number of missing values for the variables selected after the subset selection. These variables are also those included in the models.

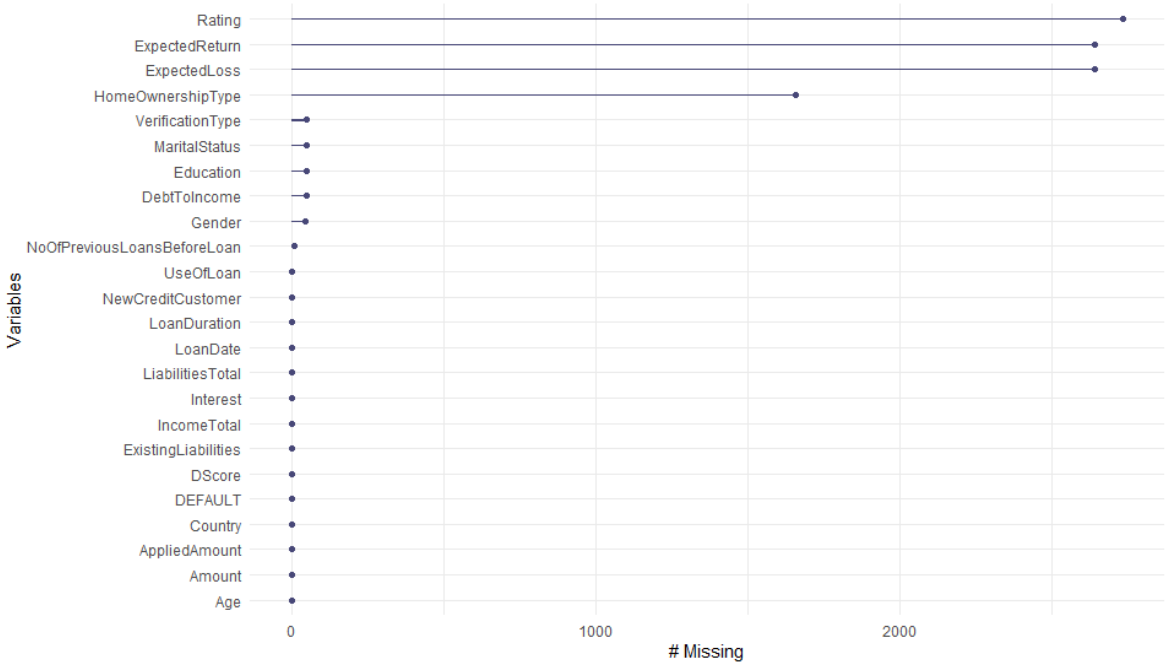
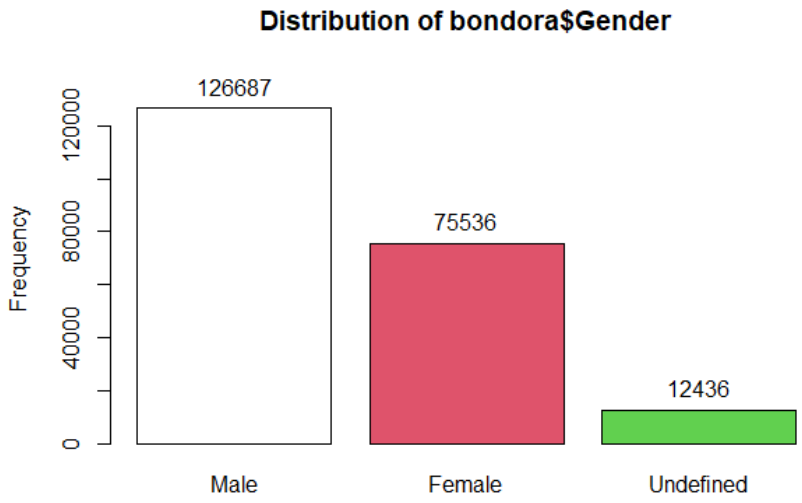


Figure 2
Frequency Plots

The two plots represent the distribution of the levels of Gender and Education. The numbers reported are the absolute frequencies for each group.



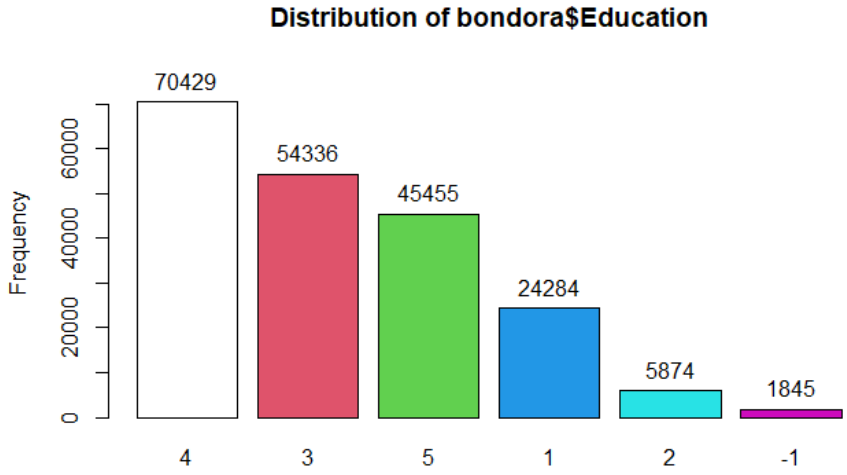


Figure 3.a
General Plots

Several different plots representing and visualize different features of the dataset. The plots include a title describing what they represent. For what concerns the country of the borrowers: EE = Estonia, ES = Spain, SK = Slovakia and FI = Finland.

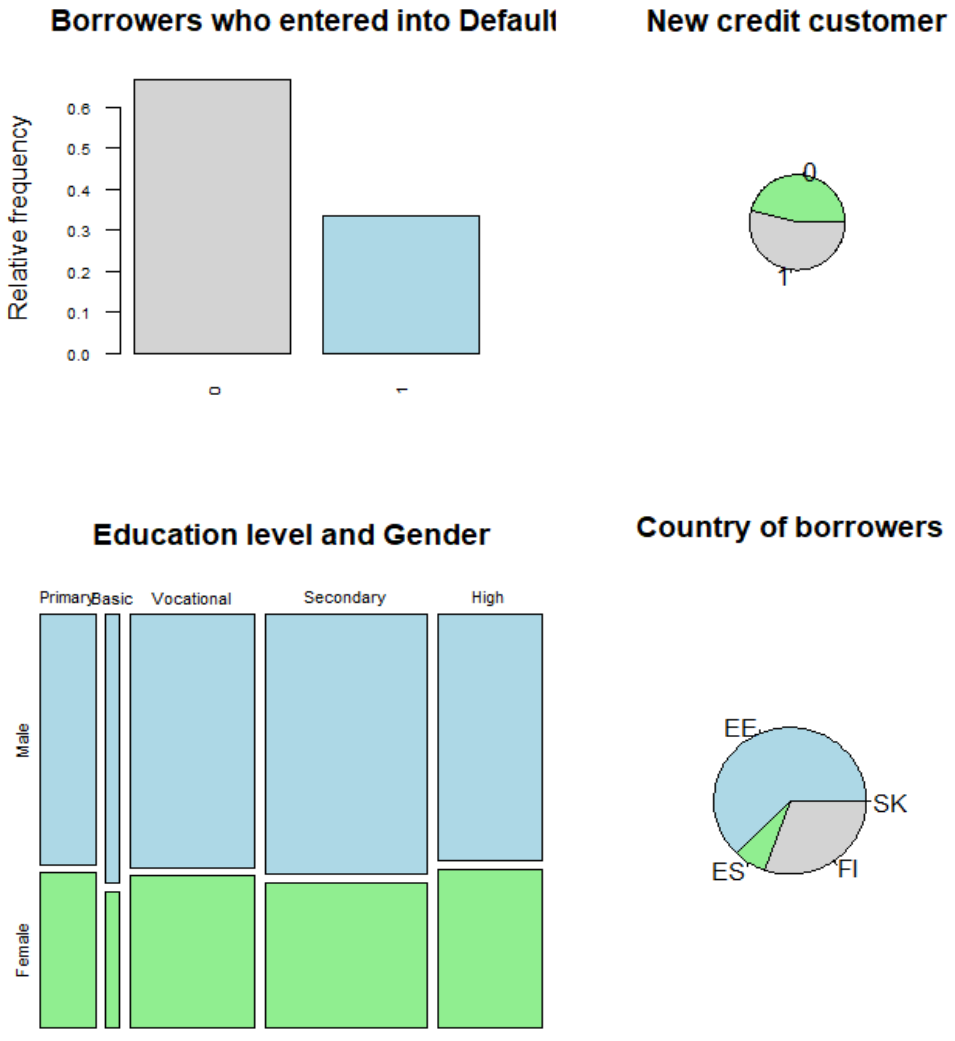


Figure 3.b
General Plots

Several different plots represent and visualize different features of the dataset. The plots include a title describing what they represent. For what concerns the use of loan: 0 Loan consolidation, 1 Real estate, 2 Home improvement, 3 Business, 4 Education, 5 Travel, 6 Vehicle, 7 Other and 8 Health.

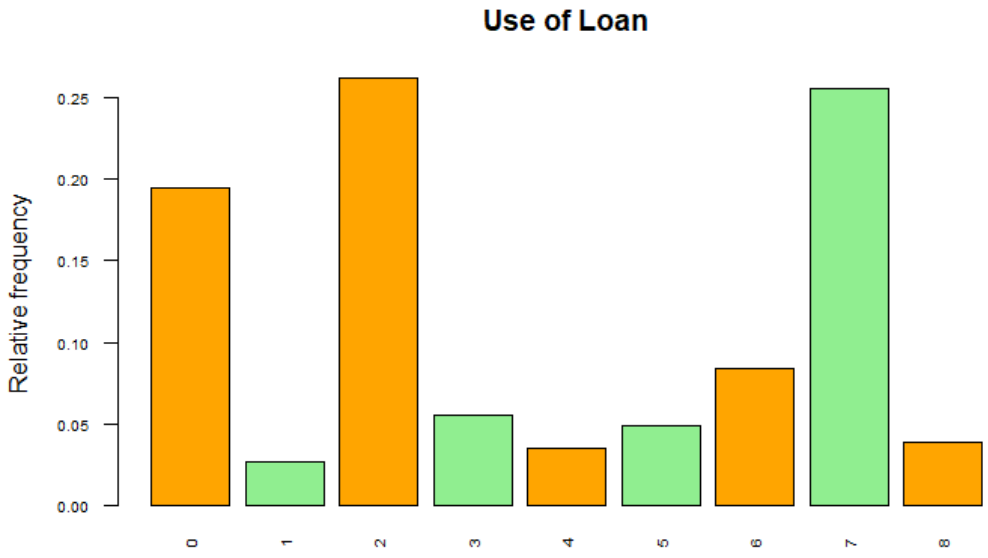
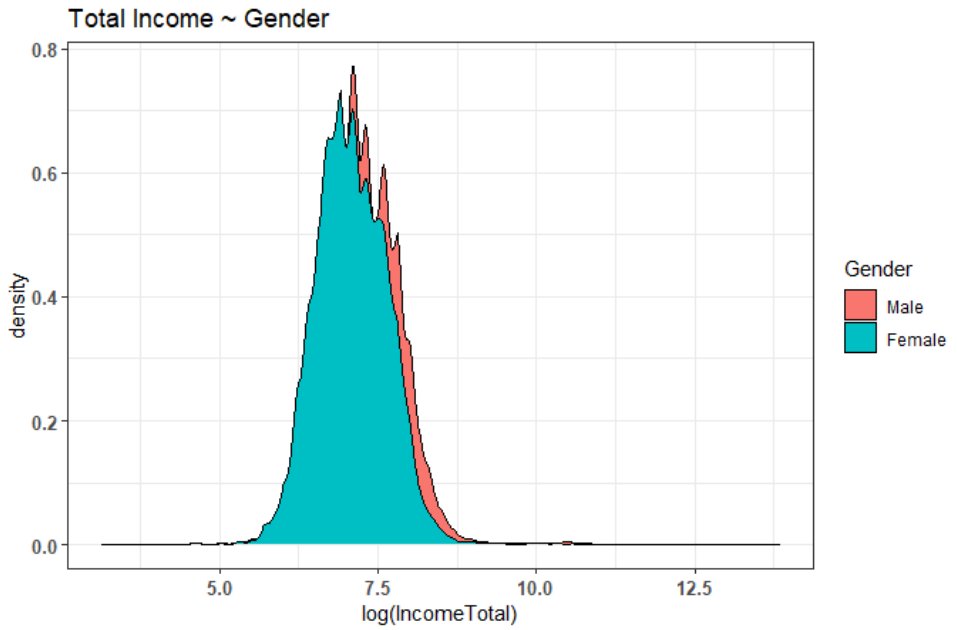


Figure 4
Gender Density Plots

The density plots for gender, showing with different colors the two gender groups (Male pink and Female blue). The density plots are present for Total Income and Amount to Income, both in log terms.



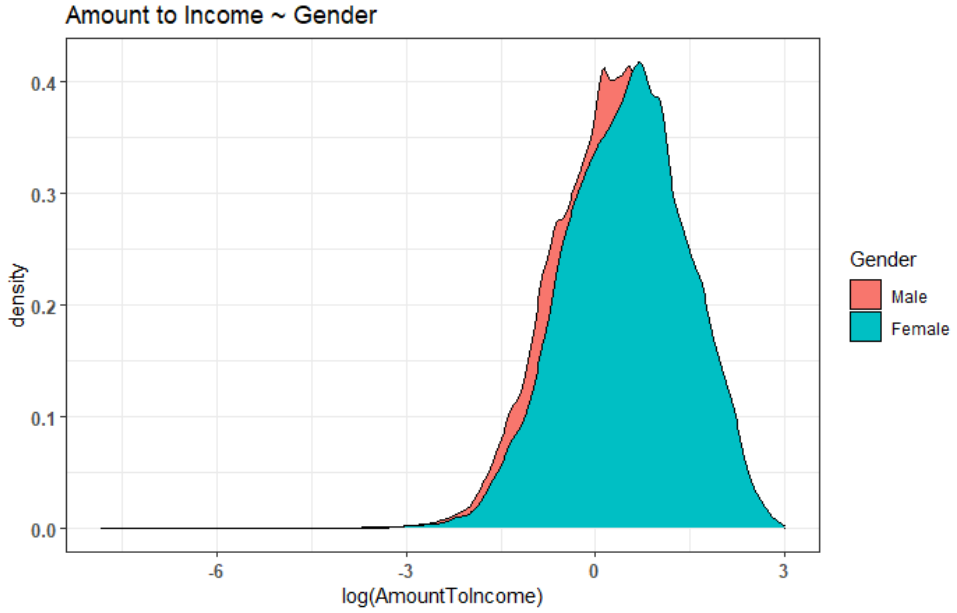


Figure 5
Education Box Plots

The box plots for education, visualizing a box plot for each level of the Education categorical variable. The box plots are present for Interest and Expected Loss.

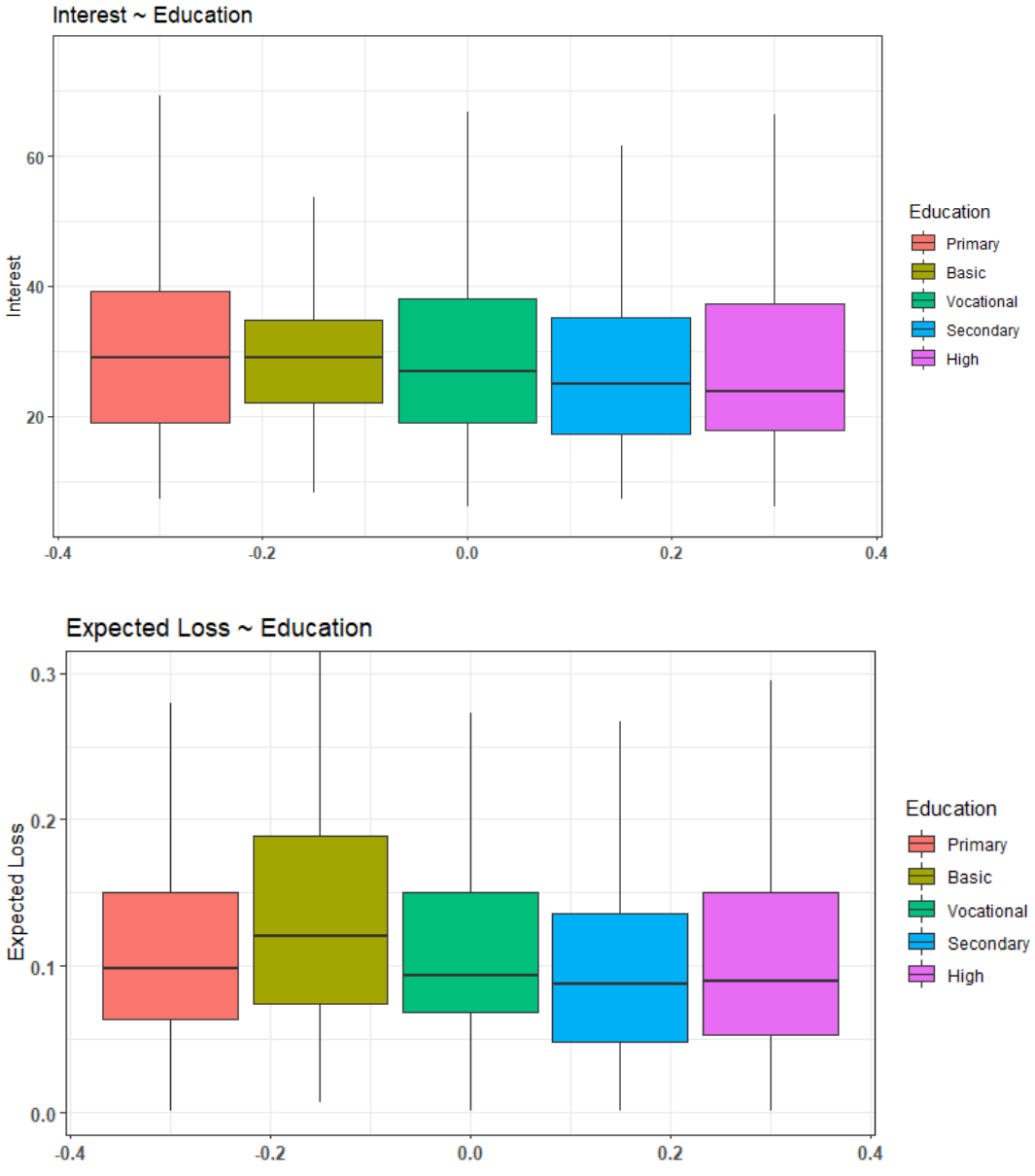


Figure 6
Correlation Matrix

The figure below represents the correlation matrix. For each pair of numerical variables, a visualization of the correlation value is reported. The colors are mapped in the following way: "red" = positive correlation (0-1), "blue" = negative correlation (-1-0). The correlation matrix shows that some pairs of variables are strongly correlated, either negatively or positively.

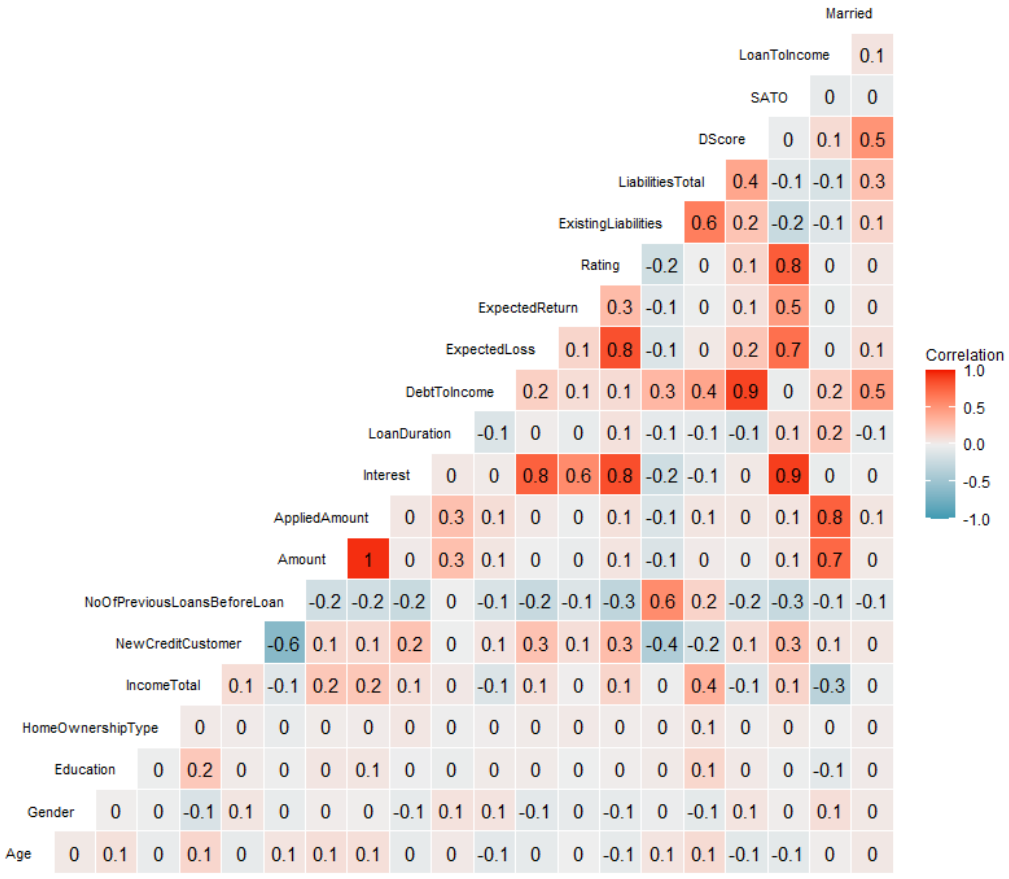


Figure 7
Interaction Gender*Married with Default as dependent

Visualization of the interaction effect between two categorical variables: Married and Gender with Gender as moderator. The interaction is included in the logistic regression model used to predict the Default and it resulted to be significant. For Gender = 0 (Male), being married is associated with a slightly lower Default probability. On the contrary, for Gender = 1 (Female), being married is associated with a higher Default probability with a considerably greater difference w.r.t. Married = 0. On average, males are associated with higher Default independently from the Married value.

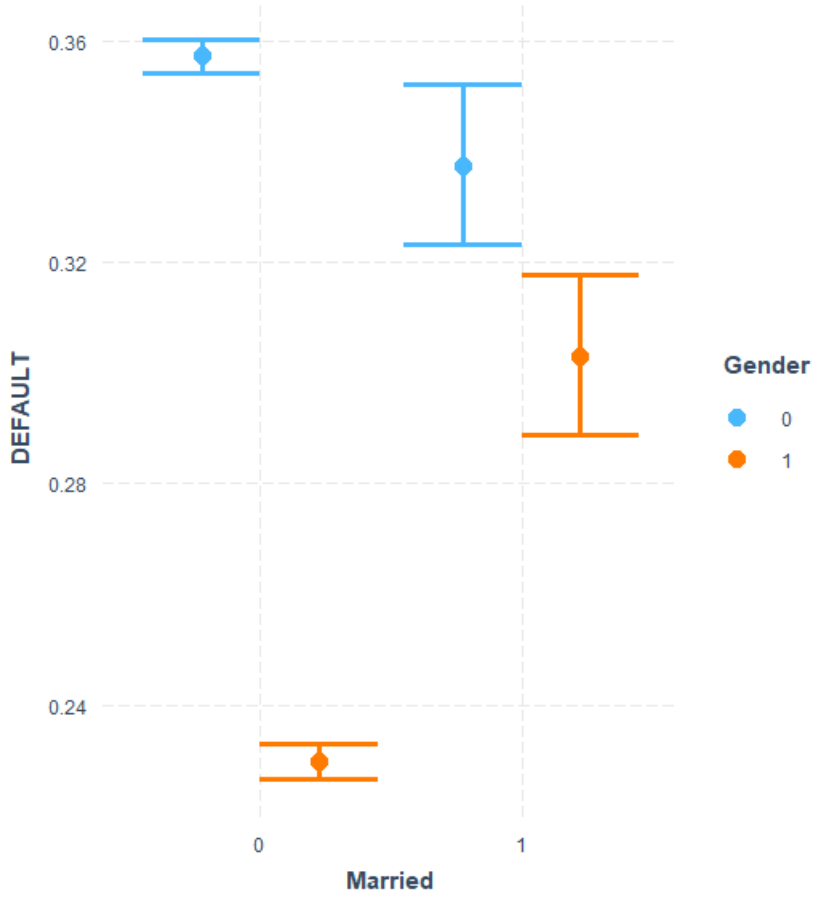


Figure 8
Interaction Education*Gender with Interest as dependent

Visualization of the interaction effect between Education and Gender with Gender as moderator. The interaction is included in the linear regression model used to predict the Interest and it resulted to be significant. It can be observed that for Gender = 1 the line has a negative steep slope. On the contrary, when Gender = 0 the line has a negative flatter slope.

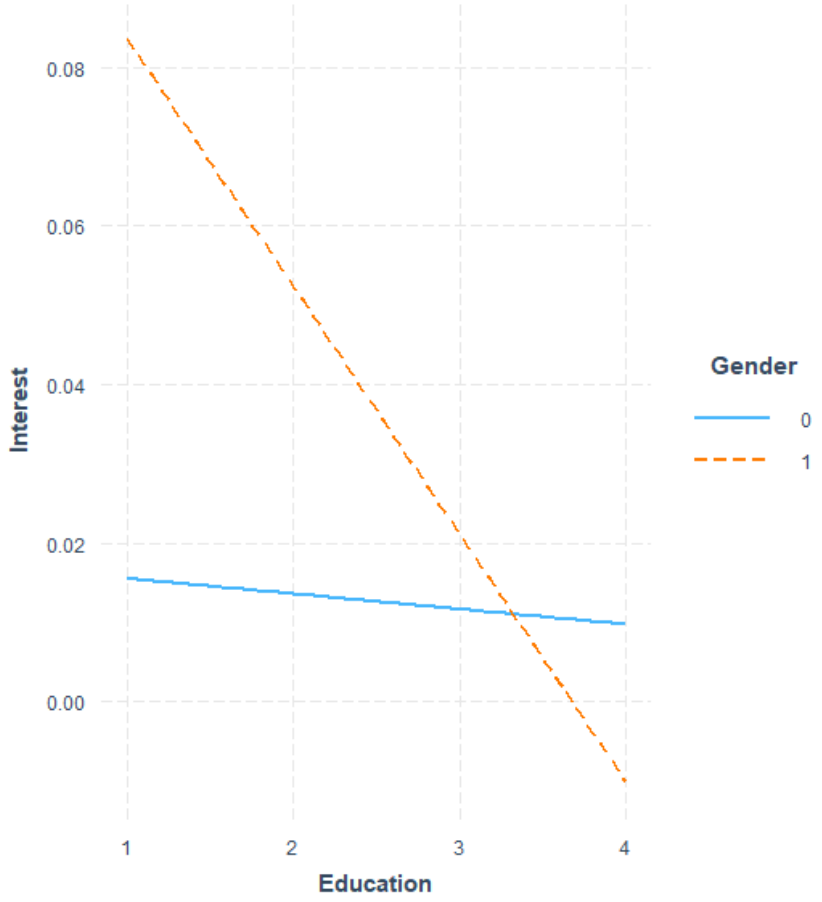


Figure 9
Interaction Education*Married with Interest as dependent

Visualization of the interaction effect between Education and Married with Married as moderator. The interaction is included in the linear regression model used to predict the Interest and it resulted to be significant. It can be observed that for Married = 1 the line has a positive and moderately steep slope. On the contrary, when Married = 0 the line has a negative slope and flat shape.

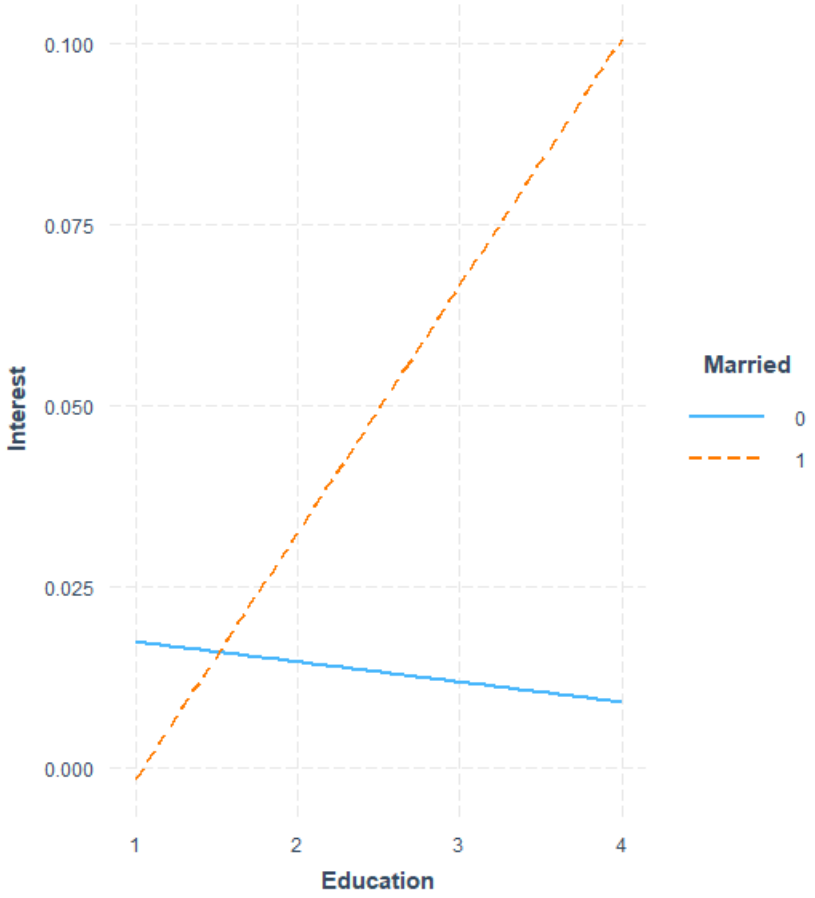
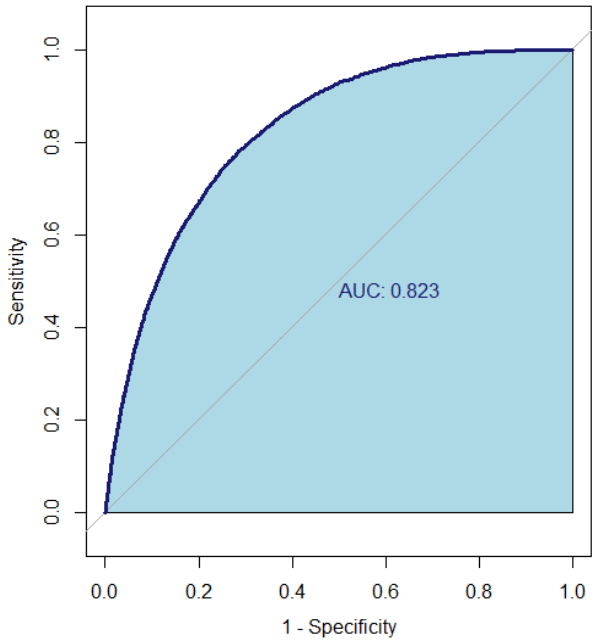


Figure 10
ROC Curves for Default

Representation of the ROC curves with the Area Under the Curve value for both logistic regression and random forest. The response variable is the Default.

Random Forest



Logit