ERASMUS UNIVERSITY ROTTERDAM

ECONOMETRICS & MANAGEMENT SCIENCE

MASTER'S THESIS IN QUANTITATIVE FINANCE

Default Determinants in Peer-to-Peer Lending

Author

Łukasz FILAR

432650

Supervisor

Dr. Mikhail ZHELONKIN

Co-reader

Dr. Andrea NAGHI

May 1, 2020

Abstract

This paper examines performance of machine learning algorithms using data from LendingClub – the biggest peer-to-peer lending platform in the United States of America. The focus is on the regularization methods which allow to obtain accurate predictions and create a sparse subset of explanatory variables. Relaxed lasso is the overall winner as it combines high classification power with good calibration. Moreover, it consistently outperforms the logistic regression which is the market standard. I have found that screening borrowers can be significantly improved by looking into their other characteristic besides the credit grade. Moreover, default probabilities highly depend on the interest rates assigned by LendingClub and lenders should seek borrowers that have low debt to income ratio and possess bankcards with high limit.

Keywords: Classification, Machine learning, Lasso, Credit scoring, Peer-to-peer lending

Contents

1 Introduction

Peer-to-peer lending (P2P), as defined by the British Peer-to-Peer Finance Association, "enables people who have money to put it to work for competitive returns through lending to other individuals or businesses online". P2P lending is concentrated on the online platforms that facilitate the contact between borrowers and loaners. This way there is no traditional intermediary, that matches interested parties, and such arrangement benefits both sides in the following ways. Lenders earn interest, which frequently exceeds the amount of interest that can be obtained by traditional means (e.g. saving accounts or bonds), while borrowers obtain funds that may not have been available from the traditional intermediaries, like banks, due to their high credit criteria. Moreover, the rates associated with the P2P loans are usually lower. The main drawback of the P2P lending is the information asymmetry between the lenders and the borrowers as the lenders know less about the borrowers' ability and willingness to repay the loan than the borrowers themselves. This issue is frequently referred to as the information asymmetry while situation where lenders cannot properly distinguish between the borrowers with different credit risk levels leads to an adverse selection (Akerlof 1970). In order to balance the information asymmetry P2P lending companies grade all borrowers such that the lenders can assess one's ability to repay the loan. Moreover, some P2P companies provide much more detailed information about each borrower which facilitates the lending decision. The focus of this study are the individuals exchanging loans using LendingClub which is a P2P lending company founded in 2006 in the United States. It enables individuals to issue loans raging from \$1000 to \$40000 as well as provides internal credit score and detailed information about each borrower.

Credit ratings and thorough analysis of borrowers' personal information allow one to make more informed decisions and avoid a credit loss. The latter is affected by the following quantities: the *probability of default*, the *exposure at default*, and the *loss given default* which can be described as follows. Probability of default measures the risk of losses over a fixed time horizon, exposure at default is the principal at stake adjusted by the value of the interest payments, and loss given default provides information on the fraction of exposure lost in the event of default. In this study I focus on the probability of default, however, all three quantities depend on each other and together account for a comprehensive risk analysis. Probability of default can be transformed into credit score that gives a brief information on one's creditworthiness. Since scores are assumed to be monotonic, they automatically preserve rankings such that a borrower with a higher score is more credible than the one with a lower score. Probability of default and credit score can be used interchangeably.

Probability of default can be obtained by using the historical data, estimated from the prices of credit default swaps and bonds, or just gathered from external ratings agencies. I use the historical information since they allow to explore a vast area of statistical and machine learning methods. The most popular method in estimating the probabilities of default is the logistic regression which became a market standard and is frequently used as a benchmark for other approaches. In this paper I also use the logistic regression as one of the benchmarks for the regularization methods which are the main focus of this study. Regularization imposes the shrinkage on the coefficients which introduces additional bias but reduces the variance, and if the latter exceeds the former, the overall performance improves. Apart from the logistic regression, the regularization methods are also compared with the machine learning algorithms, namely support vector machine and random forest. All approaches are used to obtain the predictions of the probability of default and evaluated in terms of discrimination and calibration. I also assess the most popular evaluation methods to choose the one that is the most informative in distinguishing between the performance of different models. To obtain the results bootstrap aggregating, also known as bagging, is used since it improves the performance and stability of the employed methods (Breiman 1996).

Shrinkage imposed by some regularization methods reduces the coefficient values to zero thereby creating a subset of explanatory variables. It is worthwhile to examine these predictors as it could lead to better understanding of the relations between them and their impact on the probability of default.

The study showed that relaxed lasso is the best performing model as it combines both high classification power with good calibration. Ridge regression comes as the second best approach and is inferior to relaxed lasso due to worse calibration even though it is the most accurate classifier among all methods. Elastic-net is the third best performing method as it combines the discrimination power of ridge regression with good calibration. Concerning the benchmarks, logistic regression is one of the worst models as it is poorly calibrated and has mediocre classification accuracy. Since most methods outperform logistic regression its use as a market benchmark should be reviewed. I propose elastic-net as a challenger due to its good overall performance and ability of producing a sparse set of predictors. Support vector machine and random forest do not outperform relaxed lasso and ridge regression and are frequently inferior to all regularization methods with an exception of adaptive lasso. They, however, outperform logistic regression mainly due to much better calibration. Concerning the above conclusions it is evident that calibration plays a big role in the model evaluation. Therefore, the above statements come from the Brier score which accounts for both discrimination and calibration of the models. Area under the curve, which is the market standard, accounts only for the former and may lead to biased conclusions. The results of this analysis can be carried over towards any individual or institution that issues loans to consumers. Moreover, regularization methods can be used in the case of lending to the companies as well or in any other study focused on performing forecasts.

Borrowers' default probability heavily depends on the interest rate that was assigned to them by the LendingClub. Moreover, high values of debt to income ratio and high number of mortgage accounts lead to an increase in the probability of default. Solvency of the borrowers is improved by possession of bankcards, especially when their limit is high. Remarkably the number of credit related accounts do not necessarily impact the default probability unless such inquiries were made in the recent time.

The rest of this paper is organized as follows. In the following section I review the literature on the estimation of probability of default in the framework of P2P lending. In Section 3 the data used in this study is introduced alongside a short description of required manipulations. Regularization methods alongside with their competitors are introduced in Section 4. Moreover this section includes the algorithms required to obtain the results and the evaluation methods. Section 5 introduces the results obtained from fitting the models to empirical data. Section 6 summarizes the results and the most important corollaries. Moreover, it includes a short description of the variables retained by the best performing model. Appendix includes the descriptive statistics of the variables and the robustness check of the models.

2 Literature review

In finance, logistic regression was used to predict the direction of the stock markets (Leung et al. 2000), predict corporate bankruptcies (Martin 1977), or in the credit risk modeling. Regarding the last one, in his empirical study, Wiginton (1980) showed that it outperforms the linear discriminant analysis (LDA) and criticized LDA as an inadequate method, in this framework, due to an assumption of normally distributed predictors. Concerning the application of nonparametric approaches, Makowski (1985) introduced the classification trees while the support vector machines were first applied in the works of Härdle et al. (2005) and Härdle et al. (2007). In the latter, authors present the superior performance of the support vector machines over logistic regression. West (2000) and Lessmann et al. (2015) are two extensive works on the comparison of different approaches in the credit scoring. In the first study logistic regression and decision trees are confronted with neural network models and, even though, different variations of those methods outperform decision trees, which are the worst model in total, they are inferior to logistic regression which is the overall winner. Second study involves a comprehensive survey of 41 approaches including logistic regression, ridge regression, random forests, and support vector machines among others. Concerning the methods that are applied in this paper, random forest outperformed other approaches and authors suggest to use it as a benchmark for other methods instead of logistic regression, which comes as a second best model. Support vector machines are next, regardless of the kernel applied during the estimation, while the ridge regression is the worst approach. Most of the works on predicting the probability of default feature the logistic regression even though it has been frequently criticized or better performing methods have been introduced. Nevertheless concerning the tendencies in recent developments more and more research focuses on the application of machine learning algorithms rather than extensions of the regression methods. Concerning the regularization methods Chen and Xiang (2017) used group lasso in credit scoring and, as mentioned before, Lessmann et al. (2015) applied ridge regression in their survey study. Up to my best knowledge these are the only credible applications of regularization methods in the estimation of probability of default. I intend to complement the existing research with the application of the most popular regularization methods and compare their performance with the logistic regression, support vector machines and random forest.

Usually credit risk modeling is analyzed in the context of banking since it is the most traditional loan issuer. However, in the recent years credit scoring is getting an increasing attention in case of the P2P lending since the whole phenomenon is becoming more and more popular. Emekter et al. (2015) analyze credit risk and loan performance focusing on the selection of borrowers' characteristics. In the process they employ non-parametric test and logistic regression with stepwise selection method. Zhang et al. (2016) take a very different approach where they build a credit scoring model using social media data. Iyer et al. (2015) analyze the non-standard variables like profile photo and show that lenders that use other characteristic, besides the credit score, improve their overall performance by picking non-defaulting borrowers. Even though the interest in credit scoring in P2P lending is increasing there is still not much research done in that area. Therefore, I also contribute to the literature by examining how different approaches perform in such framework. As mentioned before, the requirements to loan money in P2P lending are lower than in the case of banks which leads to a nosier data sets, with many outliers and missing observations, which considerably obstructs the analysis.

3 Data

LendingClub's consumer loan data is used to carry out the empirical part of this study. The data is spanned from 2012 until 2016 and consists of 493855 loans (391040 fully paid and 102815 defaulted or charged-off). During this period LendingClub issued many more loans, however, some of them are still being repaid or late and it is not possible to say whether a particular loaner will default or not. Each borrower is characterized by a total of 111 variables.

Some variables are not relevant for the subsequent analysis due to the following reasons. LendingClub's ID number or loan title are added for administering purposes and do not add any relevant information to the research. Payments received to date or grade and sub-grade are transformations of other variables and would just add additional noise to the analysis. Some variables are relevant only for a particular group of loaners (e.g. defaulters) and do not contribute information to the whole sample. Moreover, influence of variables like zip-code, LendingClub's profile or application code is not a subject of this paper. Variables with more than 60% of missing values are also discarded since they would not contribute significantly to the research and imputing values in such a high proportion of cases would introduce additional bias. If the proportion of missing values is lower the median values are imputed. Variables with variance close to zero are also omitted since they would not add much additional information. In the end the total number of variables is 45 and they are shortly described in Table 1. Table 10 contains the descriptive statistics of the quantitative variables, and Table 11 contains the distribution of the levels of the qualitative variables. This table also contains the levels of the categorical variables. Both of them can be found in the Appendix. Analyzing the descriptive statistics it is easy to see that the following variables: annual income, revolving bankcards, average balance, revolving balance, months since recent account, current balance, credit limit, balance excluding mortgage, installment limit, and revolving credit limit exhibit high or extremely high kurtosis. In order to dampen it, a log transformation is used.

Data from 2012 and 2013 is used to estimate the models, while data from 2014 is used to validate them. In order to check how robust the models are, years 2015 and 2016 are used for the robustness check.

Table 1: List of predictor variables with a description

Variables	**Description**
Interest rate	Interest rate on the loan
Loan purpose	Category provided by the borrower (4 levels)
Loan amount	The listed amount of the loan
Term	Number of payments on the loan expressed in months (2 levels)
Listing status	The initial listing status of the loan (2 levels)
Annual income	Self-reported annual income provided by the borrower
Housing status	Home ownership status (3 levels)
Employment length	Employment length in years (3 levels)
Revolving bankcards	Total open to buy on revolving bankcards
Balance to credit limit ratio	Ratio of total current balance to credit limit for all bankcard accounts
Verification status	Indicates whether the income was verified (3 levels)
Inquiries in the last 6 months	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
Average balance	Average current balance of all accounts
Revolving balance	Total credit revolving balance
Current accounts	The number of credit lines currently in the borrower's credit file
Total accounts	The total number of credit lines in the borrower's credit file
Number of trades opened in the last 24 months	Number of trades opened in the last 2 years
Total number of mortgage accounts	The number of mortgage accounts
Months since oldest installment account	The number of months since oldest bank installment account opened
Months since oldest revolving account	The number of months since oldest revolving account opened
Months since recent revolving account	The number of months since the most recent revolving account opened
Months since recent account	The number of months since the most recent account opened
Months since last new bankcard	The number of months since most recent bankcard account opened
Months since recent inquiry	The number of months since the most recent inquiry
Active bankcard accounts	The number of currently active bankcard accounts
Total number of active revolving trades	The number of currently active revolving trades
Satisfactory bankcard accounts	The number of satisfactory bankcard accounts
Bankcard accounts	The number of bankcard accounts
Total number of installment accounts	The number of installment accounts
Open revolving accounts	The number of open revolving accounts
Revolving accounts	The number of revolving accounts
Revolving trades	The number of revolving trades with positive balance
Satisfactory accounts	The number of satisfactory accounts
Number of accounts opened in the last 12 months	The number of accounts opened in past 12 months
Percent of never delinquent trades	The percent of trades never delinquent
Current balance on all accounts	Total current balance of all accounts
Credit limit	Total credit limit
Installments	Monthly payment owed by a borrower
Bankcards above 75% of limit	Percentage of all bankcard accounts above 75% of the limit
Balance excluding mortgage	Total credit balance excluding mortgage
Bankcard limit	Total bankcard credit limit
Installment limit	Total installment credit limit
Revolving utility	The amount of credit the borrower is using relative to all available
Revolving credit limit	Total revolving credit limit
Debt to income ratio	Borrower's total monthly debt payments divided by the total debt obligations

4 Methodology

In order to estimate the probabilities of default I apply the binomial logistic regression model. It is an adequate method since it is a direct probability model that yields probabilities without a need of further transformations which is the case for many classifiers, like support vector machine or random forest, that are going to be introduced later in this section. Moreover, logistic regression allows for an easy and convenient addition of different regularizations that are also described afterwards.

McCullagh and Nelder (1989) note that logistic regression belongs to the family of the generalized linear models (GLMs) that extend the framework of standard linear regression by allowing the linear models to be related to the response variable via a link function. In the case of logistic regression the link function is the so-called *logit* (or log-odds)

$$\log\left(\frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)}\right)=\beta_0+\beta^T x, \tag{1}$$

where $Y \in \{0,1\}$ is i.i.d binomially distributed response variable informing whether a loaner remained solvent ($Y=0$) or defaulted ($Y=1$) and X is a vector of p predictors. Moreover, for observation i, $i=1,...,N$, let x_{ij} denote the value of predictor x_j, $j=1,...,p$ and $x_i=(x_{i1},...,x_{ip})$. β_0 is an intercept and β is a vector of coefficients while log stands for a natural logarithm.

Logistic regression allows for an easy interpretation of predictors and their impact, as an increase in jth predictor, by one unit, changes the log-odds by β_j assuming all other predictors are constant. Using (1) it is easy to obtain an expression for the conditional probability for ith loaner

$$p(x_i)=\Pr(y_i=1|x_i)=\frac{\exp(\beta_0+x_i^T\beta)}{1+\exp(\beta_0+x_i^T\beta)}. \tag{2}$$

Using the conditional probabilities, as presented in (2), is especially convenient since it yields values between 0 and 1 which directly translates into probabilities of default. In order to discuss the impact of the predictors in more detail, from (2) it can be seen that $1-p(x_i)=(1+\exp(\beta_0+x_i^T\beta))^{-1}$ so $p(x_i)$ is monotone in each predictor with regard to the sign of its coefficient. Concerning the interpretation of coefficient's magnitude, a straight line drawn tangent to the curve at any particular value, describes the instantaneous rate of change in $p(x_i)$ at that point

$$\frac{\partial p(x_i)}{\partial x_j}=\beta_j\frac{\exp(\beta_0+x_i^T\beta)}{(1+\exp(\beta_0+x_i^T\beta))^2}=\beta_j p(x_i)(1-p(x_i)).$$

Therefore, the effect of the jth predictor on the probability $p(x_i)$ depends on the coefficient β_j and the value of the probability. As log-odds may not be a very intuitive measure, taking an exponent of (1) gives the odds ratio

$$\frac{p(x_i)}{1-p(x_i)} = \exp(\beta_0 + x_i^T\beta). \tag{3}$$

From (3) it follows that the odds ratio is an exponential function of the predictors. The odds multiply by e^{β_j} per unit increase in x_j, keeping all other predictors constant. In order to give some more intuition on the impact of predictors on the response variable, one can note that regardless of the value of x_j, if the corresponding coefficient β_j is positive, then increasing x_i is associated with an increment of $p(x_i)$. Inverse relation follows for the coefficients that are negative. If the coefficient β_j is negative, then increasing x_i is associated with a decline of $p(x_i)$.

Logistic regression models are usually estimated by maximizing the binomial log-likelihood, or equivalently by minimizing the negative binomial log-likelihood. The log-likelihood can be written as

$$\ell(\beta_0,\beta) = \frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + x_i^T\beta) - \log(1+\exp(\beta_0 + x_i^T\beta)). \tag{4}$$

4.1 Regularization methods

Estimating a model with the whole set of independent variables $X = (x_1, ..., x_p)$ may result in overfitting and high standard errors. Therefore in order to obtain accurate estimates and a sparse set of independent variables $\mathcal{M} = (x_1, ..., x_k)$ (where $1 \leq k \leq p$ and $\mathcal{M} \subseteq \{1, ..., p\}$) logistic regression is performed alongside different regularizations. Such methods take into account all of the potential predictors, however, only a subset of them remains in the model. According to Hastie et al. (2009) applying regularizations usually improves prediction accuracy, trading off decreased variance for increased bias. The first regularization method is called lasso which stands for least absolute shrinkage and selection operator. It was introduced by Tibshirani (1996) and in terms of logistic regression framework, it aims to minimize the negative log likelihood function while forcing the sum of the absolute value of the regression coefficients to be less than a fixed value t

$$\hat{\beta}^{lasso} = \underset{\beta_0,\beta}{\operatorname{argmin}}\big\{-\ell(\beta_0,\beta)\big\}, \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq t,$$

where t is a free parameter controlling the amount of regularization on the regression coefficients. Alternatively lasso can be expressed in the so-called Lagrangian form

$$\hat{\beta}^{lasso} = \underset{\beta_0,\beta}{\operatorname{argmin}}\big\{ -\ell(\beta_0,\beta) + \lambda||\beta||_1\big\}, \tag{5}$$

where λ is a non-negative penalty factor and $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1 norm of β. Due to the nature of L_1 constraint, sufficiently large penalty parameter λ will force some of the coefficients to be exactly equal to zero. This way lasso simultaneously performs subset selection and parameter estimation. There are actually two reasons for introducing the penalty term λ: prediction accuracy and interpretability. Compared to maximum likelihood estimates (MLE), lasso sacrifices the (asymptotical) unbiasedness of its estimators for the sake of their lower variance. It is a well know fact that standard MLE tend to have large variance (especially in case of many predictor variables) which negatively affects their prediction accuracy. Interpretation, in terms of model complexity, also becomes clearer since lasso allows to obtain a subset of variables that exhibits the strongest influence towards the target variable. Even though lasso is a convenient approach it is also troubled by its deficiencies. If predictors are highly correlated lasso fails to distinguish irrelevant predictors from the true ones and drops them arbitrarily (Zhao and Yu 2006). Moreover, it is inferior in terms of prediction accuracy to ridge regression (Hoerl and Kennard 1970) that uses all potential predictors. The latter was empirically observed by Tibshirani (1996) for the case where the number of observations N is bigger than the number of variables p. Lasso shortcomings continue for high dimensional data ($p > N$), where lasso always chooses at most N variables due to the nature of convex optimization. Moreover, as mentioned before, lasso shrinks the non-zero coefficients towards zero, relative to maximum likelihood fit, and hence introduces an additional bias. This issue can be related to the work of Fan and Li (2001) and Meinshausen and Bühlmann (2006) who point out that lasso does not have the oracle properties which refers to model's consistency in parameter estimation and variable selection. In order to circumvent these shortcomings a few extensions were developed.

One of the methods that corrects for an additional bias, introduced by shrinking the nonzero coefficients, is the relaxed lasso introduced by Meinshausen (2007). This is a two stage approach where first we perform lasso on the whole set of data X and then perform

$$\hat{\beta}^{relasso} = \underset{\beta_0,\beta}{\operatorname{argmin}} - \left[\frac{1}{N} \sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T \{\beta \cdot \mathbf{1}_{\mathcal{M}}\}) - \log(1 + e^{(\beta_0 + x_i^T \{\beta \cdot_{\mathcal{M}}\})}) \right] + \phi\lambda ||\beta||_1, \quad (6)$$

on the subset $\mathcal{M}$. $\mathbf{1}_{\mathcal{M}}$ is the indicator function on the particular set of variables

$$\{\beta \cdot \mathbf{1}_{\mathcal{M}}\} = \begin{cases} 0, & k \notin \mathcal{M}, \\ \beta_k, & k \in \mathcal{M}. \end{cases}$$

Relaxation parameter ϕ controls the shrinkage intensity of the coefficients. Depending on the value of ϕ there are three different scenarios. For $\phi = 1$, the relaxed lasso becomes just regular lasso and for $\phi = 0$ it corresponds to a two stage procedure where first, lasso is performed only to choose the variables and then, the model is estimated by means of MLE using only those variables. This idea was first introduced by Efron et al. (2004) and is a special case of relaxed lasso which is also going to be implemented in this study. In subsequent sections I refer to this approach as the *two step method*. For $\phi < 0$, the shrinkage of coefficients in the second stage is lower compared to lasso estimation. Meinshausen (2007) claims that for the cross-validated penalty parameters relaxed lasso estimates have oracle properties.

Another extension of the lasso is the adaptive lasso introduced by Zou (2006)

$$\hat{\beta}^{adalasso} = \underset{\beta_0,\beta}{\operatorname{argmin}} \big\{ -\ell(\beta_0, \beta) + \lambda w ||\beta||_1 \big\}, \quad (7)$$

where w is a vector of known weights. In order to estimate the weights first MLE or ridge regression are conducted using (8) on the whole data set to obtain the coefficients. Next the weights are obtained by taking an inverse of those coefficients. In this paper the weights are obtained using the ridge regression. Similarly to the relaxed lasso adaptive lasso yields unbiased estimates and has oracle properties.

The final extension of lasso, considered in this study, is the elastic-net introduced by Zou and Hastie (2005). Elastic-net is a compromise between the lasso and the ridge regression. Introduced by Hoerl and Kennard (1970), ridge regression is very similar to the lasso since it also shrinks the coefficients, however, it does not put any of them to zero. It is due to the fact that ridge regression uses the L_2 constraint. Application of ridge estimators in the logistic regression framework was introduced by Le Cessie and Van Houwelingen (1992). Log likelihood function of the ridge regression is very similar to the lasso one (5)

$$\hat{\beta}^{ridge} = \underset{\beta_0,\beta}{\operatorname{argmin}}\big\{-\ell(\beta_0,\beta) + \lambda||\beta||_2/2\big\}. \tag{8}$$

Ridge regression is also going to be implemented in this study since it is known for its high prediction accuracy. Lasso penalty, alongside with the ridge penalty, constitute for the elastic-net regularization

$$\hat{\beta}^{en} = \underset{\beta_0,\beta}{\operatorname{argmin}}\big\{-\ell(\beta_0,\beta) + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]\big\}, \tag{9}$$

where $\alpha \in [0,1]$ is a parameter that controls the amount of lasso and ridge shrinkage. It is easy to see that for extreme values of α elastic-net becomes either ridge or lasso regression. The main reason for introducing the elastic-net is its improved forecasting accuracy (Zou and Hastie 2005) compared to the lasso.

MLE of the logistic regression without regularizations are obtained by the means of the Newton's algorithm which boils down to the iteratively reweighted least squares. Using the current estimates $(\widetilde{\beta_0}, \widetilde{\beta})$ a second order Taylor expansion is formed about them, which leads to the quadratic objective function (10). The updates are obtained by minimizing the quadratic objective function. To solve a problem featuring regularization Friedman et al. (2010) introduced the cyclic coordinate descent algorithm. In order to compute an entire path of solutons, for each λ, an outer loop computes the quadratic approximation ℓ_Q about the current parameters $(\tilde{\beta_0}, \tilde{\beta})$. Then coordinate descent algorithm solves the penalized weighted least-squares problem denoted in (12). Each inner coordinate descent loop continues until the maximum change in the objective function is less than a very small threshold times null deviance. The threshold in this paper is 1e−7. Then the next step is to decrease the value of λ and repeat all three loops until convergence. The cyclic coordinate descent is outlined in Algorithm 1.

Algorithm 1 Cyclic Coordinate Descent

outer loop Diminish the λ value.

middle loop Update the quadratic approximation of the log-likelihood function about the current estimates $(\widetilde{\beta}_0, \widetilde{\beta})$:

$$\ell_Q(\beta_0, \beta) = -\frac{1}{2N}\sum_{i=1}^{N} w_i(z_i - \beta_0 - x_i^T\beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})^2, \tag{10}$$

where

$$z_i = +x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1 - \tilde{x}_i)} \quad \text{is the current working response} \tag{11a}$$

$$w_i = \tilde{p}(x_i)(1 - \tilde{p}(x_i)) \quad \text{is the weight,} \tag{11b}$$

where $\widetilde{p}(x_i) = p(x_i; \widetilde{\beta}_0, \widetilde{\beta})$ is introduced to ease the notation and $C(\tilde{\beta}_0, \tilde{\beta})$ is the constant.

inner loop Use the coordinate descent algorithm to solve the penalized weighted least-squares problem:

$$\underset{\beta_0, \beta}{\operatorname{argmin}}\Big\{ -\ell_Q(\beta_0, \beta) + \lambda P_\alpha(\beta)\Big\}. \tag{12}$$

Since cyclic coordinate algorithm yields a sequence of models to choose from, some procedure is needed to choose the one deemed suitable for further analysis. For regularization models the penalty parameter λ is always estimated using 10-fold cross validation. The optimal λ is chosen using the fraction of deviance explained:

$$D_\lambda^2 = \frac{Dev_{null} - Dev_\lambda}{Dev_{null}}. \tag{13}$$

Fraction of deviance explained corresponds directly to the fraction of variance explained in linear regression R^2 while deviance Dev_λ is analogous to the residual sum of squares and is defined as minus two times the difference between log likelihoods of a model with a penalty parameter λ and the *saturated* model that includes a separate parameter for each observation Dev_{null} stands for the null deviance obtained from the intercept-only model. This study employs the largest value of λ that is within one standard error of the λ that gives the smallest cross-validation error. This way the most parsimonious model is chosen such that its error is not higher than one standard error above the error of the best model. Such choice is motivated by the sparsity of a model - model with the lowest mean cross-validated error retains more variables and fits the data better which may lead to overfitting. In other words choosing the one standard error λ leads to the simplest model with an accuracy comparable to the best model. Elastic-net and relaxed lasso employ additional parameters α and ϕ respectively. Similarly to the case of λ, 10-fold cross validation is used to estimate those parameters but in this case the chosen values are the ones that yield smallest cross-validation error.

4.2 Support vector machine

As mentioned before, support vector machine (SVM) is one of the machine learning algorithms used as a competitor for the regularization methods. Therefore, in this work, SVM is not analyzed in so much detail as the shrinkage methods but used as a benchmark to show empirically that regularization methods can achieve higher level of prediction accuracy.

Hastie et al. (2009) present SVM as a non-probabilistic classifier capable of performing both linear and non-linear classification using the separating hyperplane. In other words, given the training data, the support vector machine outputs an optimal hyperplane which discriminates the observations into two subspaces. Unlike in the case of logistic regression the binary response variable is denoted as $y \in \{-1, 1\}$ which is a more convenient approach concerning the formal expressions. Support vector machine solves the following optimization problem

$$\begin{aligned} &\underset{\beta_0,\beta,\xi}{\text{argmin}} && ||\beta|| \\ &\text{subject to} && y_i(\beta_0 + x_i^T\beta) \geq W(1-\xi), \forall i, \\ & && \xi_i \geq 0, \sum \xi_i \leq C, \end{aligned} \tag{14}$$

where W is the width of the margin. ξ is a *slack* variable informing where the observations are located relative to the margins and hyperplane (if $\xi_i = 0$ then the ith observation is on the correct side of the margin and hence the hyperplane). C is a non-negative tuning parameter or *budget* for the amount that the margin can be violated by the n observations (if $C = 0$ then there is no budget for violations and therefore $\xi = 0$ must hold).

SVM is a generalization of support vector classifier (SVC) that employs a linear hyperplane that optimizes the bias-variance trade-off. SVM follows similar idea however it generalizes to non-linear cases using the so-called kernel trick that maps the predictors into high-dimensional feature spaces (Boser et al. 1992). However, the separation does not need to be perfect as some observations may be on the incorrect side of the hyperplane. Such treatment allows for a greater robustness against the individual observations and provides a better classification of most of the observations. This *soft margin* incarnation of support vector machines was formulated by Cortes and Vapnik (1995).

It turns out that the solution[1] to SVC involves only the *inner products* of the observations. The inner product of two observations $x_i, x_{i'}$ is denoted by

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij}, x_{i'j}. \tag{15}$$

The inner product can be generalized using the *kernel* K. Kernel is a function that quentifies the similarity of two observations. The linear kernel is denoted as following

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}. \tag{16}$$

The above equation would return the SVC because it is linear in the features. In short, linear kernel quantifies the similarity of a pair of observations using Pearson correlation. Another popular kernel is polynomial kernel

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x'_{ij})^d, \tag{17}$$

where d ia a positive integer. Polynomial kernel boils down to fitting a support vector classifier in a higher-dimensional space involving polynomials of d degree, rather than in the original feature space. This way polynomial kernel allows for a more flexible decision boundary and since it is not linear the resulting classifier is known as an SVM. Another popular option is the radial kernel:

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2), \tag{18}$$

where γ is a positive constant. Given a test observation $x^* = (x_1^*, ..., x_p^*)^T$ is far away from a training observation x (in terms of Euclidean distance), then $\sum_{j=1}^{p} (x_j^* - x_{ij})^2$ will be large, which results in (18) being small. Hence, training observation that are far from x^*

[1] It is a quadratic programming solution that can be obtained using e.g. Lagrange multipliers. It will not be presented in this work but is available in Hastie et al. (2009).

will not considerably affect the predicted class label for x^*. Each of the above kernels will be tested on the empirical data such that the best fitting one is chosen in the subsequent analysis.

4.3 Random forest

The second competitor for the regularization methods is the random forest that was introduced by Breiman (2001). In this approach first multiple decision trees are grew and then, in case of classification problem, the mode of the classes of the individual trees is calculated. Before describing the random forest it is useful to first introduce the decision trees and the bootstrap aggregating which are two building blocks of the random forests.

Decision trees divide the predictor space into m regions, each denoted as R_m, using the recursive binary splitting which begins at point where all the observations belong to a single region and then it splits the predictor space by introducing two new branches. The split is done such that the classification error is minimized. Hastie et al. (2009) use entropy-like measure and try reducing the amount of entropy to obtain the best branch split. They also suggest using Gini index

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{19}$$

which measures the total variance across the K classes, where $\hat{p}_{mk}$ is the proportion of training observations in the mth region that are from the kth class. High variance negatively affects the classification accuracy so low values of Gini index produce better classifications. Once there is the first split, one or both of previously created regions can be further divided into two more regions. This process is continued until some minimum node size is reached. This approach leads to a large tree T_0 which may overfit the data and result in a poor classification performance. In order to prevent it, the tree is pruned to obtain a subtree T such that $T \subset T_0$ using cost-complexity pruning. This approach considers a sequence of trees indexed by a nonnegative tuning parameter δ that governs the trade-off between the complexity of a tree and its goodness of fit to the data. For each δ one wants to find the subtree T_δ that minimizes

$$C_\delta(T) = \sum_{m=1}^{|T|} N_m G + \delta|T|, \tag{20}$$

where N_m is the number of observations belonging to region R_m and $|T|$ is the number of terminal nodes in T. The large values of δ yield smaller trees while the small values produce bigger trees. In order to choose a value of δ that minimizes "average error" we use 10-fold cross validation. James et al. (2015) note that (20) is similar to the formulation of lasso (5) since both parameters δ and λ control the complexity of the models produced by those approaches.

The major drawback of the decision trees is the high variance which considerably deteriorates their accuracy. Moreover decision trees are not robust and even a small change in the data can substantially change the shape of the tree. To overcome these deficiencies one can aggregate decision trees in order to improve their predictive performance. One of the methods that allows for an analysis of an ensemble of decision trees is bootstrap aggregating, which is frequently referred to as bagging. It was introduced by Breiman (1996) and is a very general method which is also employed in the design of this study. To perform bagging first I obtain B bootstrap samples from the training set, fit the model on the bth bootstrapped sample to obtain $\hat{f}^b(x)$, and then average all the predictions

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x).$$

By averaging the predictions bagging reduces the variance and hence improves the prediction accuracy of employed model. Therefore it suits well to the framework of decision trees that yield unstable and highly variable predictions.

Decision trees alongside bagging constitute for the random forests which differ from the bagged trees by building a collection of de-correlated trees that are subsequently averaged. Hastie et al. (2009) present an Algorithm 2 that summarizes the process of growing the random forest.

Algorithm 2 Random forest algorithm

```
1. Draw $B$ bootstrap samples from the training data and for each sample $b$ do as follows.
2. Grow a decision tree $T_b$ by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is achieved.
    2.1. Choose $m$ variables at random from $p$ variables
    2.2. Pick the best split-point among $m$
    2.3. Split the node into two further nodes
3. Output the ensemble of trees $\{T_b\}^B$
4. Let $\hat{C}_b(x)$ be the class prediction for the $b$th decision tree. Then $\hat{C}^B(x) = \text{majority vote}\{\hat{C}_b(x)\}$
```

SVM and random forest are both non-probabilistic classifiers, so in order to evaluate their outcomes, and compare them to the ones obtained from the logistic regression models, I need to obtain the probabilities. Platt (2000) introduced a method that basically performs a logistic regression on the output of the algorithms with respect to the true class label. This way the results given by the SVM and random forest are turned into the probabilities.

4.4 Model evaluation

Medema et al. (2009) analyze different aspects of the default models evaluation and state that nowadays they are being validated by determining their discrimination and calibration abilities. Discrimination deals with a correct separation between two classes while calibration concerns the statistical consistency between the distributional forecasts and the observations.

Given the estimated probabilities one needs to pick a cut-off value that discriminates the loaners into defaulters and non-defaulters depending on their probability of default. In other words this cut-off value allows one to migrate from a probabilistic forecast to a point forecast. There is no simple rule on how to pick this value and the approach chosen in this study is to choose a cut-off that minimizes the sum of the error frequencies which corresponds to maximization of the sum of *sensitivity* and *specificity*. *Sensitivity* refers to the ratio of correctly identified cases (true positives) to the sum of correctly identified and incorrectly rejected ones (false negatives). In other words it measures the proportion of correctly identified defaulters (true positive rate). *Specificity* measures the ratio of correctly rejected cases (true negatives) to the sum of correctly rejected cases and incorrectly identified ones (false positives). In other words it measures the proportion of correctly identified non-defaulters (true negative rate) (James et al. 2015). A convenient way to present these relations is a confusion matrix presented in Table 2:

Table 2: Possible outcomes of loaners classification

		Predicted class	
		Default	Solvent
True class	Default	True positive (TP)	False negative (FN)
	Solvent	False positive (FP)	True negative (TN)

Sensitivity and specificity constitute for the most popular discrimination measure - the receiver operating characteristics curve that is frequently abbreviated as the ROC curve. It shows the performance of a classifier over all cut-off values and its overall performance is measured by the area under the ROC curve (AUC). In order to create the ROC curve we plot the *sensitivity* against 1− *specificity*. Even though AUC allows to assess the discriminatory power of a classifier without choosing a subjective cut-off, it is not an adequate measure for classifiers comparisons. Lobo et al. (2007) point out that AUC does not take into account the calibration of the model, treats true positive rate and false positive rate equivalently, or operates in the irrelevant prediction ranges.

Unlike AUC, Brier score (Brier 1950) does not require a cut off value and directly uses the estimated probabilities. It is denoted as an average of the squared difference between the probability of default, estimated by the model $\hat{p}_i$, and the observed outcome y_i (Wilks 2006):

$$B = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2. \tag{21}$$

Due to Murphy (1973) Brier score can be decomposed into three terms: uncertainty, reliability, and resolution. Uncertainty component measures the uncertainty inherent in the event occurrence like going default or staying solvent. In other words it informs how likely it is that a particular event is being forecasted. The possible values of uncertainty lie within the interval from 0 for events that either occur all the time or never to 0.25 for events that have a frequency of 0.5. The latter is the case for this study since always an even amount of defaulters and non-defaulters is sampled from the population. Reliability component measures the distance between the true and predicted probabilities. The perfectly plausible forecast has a score equal to 0. This holds when all events occur with their respective estimated probabilities over the successive observations. Resolution component reflects how well the forecasts predict the occurance of an event with a forecast probability that is very different from its frequency. Analytically, those three measures can be denoted as

$$\begin{gathered} B = \mathit{Reliability} - \mathit{Resolution} + \mathit{Uncertainty} = \\ \frac{1}{N}\sum_{k=1}^{K} n_k(\hat{p}_k - y_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(y_k - c)^2 + c(1-c), \end{gathered} \tag{22}$$

where k is some forecast-event pair, n_k denotes the number of forecasts belonging to the same probability category and c is the relative frequency. As mentioned before $c = 0.25$ in this research.

In order to facilitate the analysis reliability, resolution and uncertainty can be expressed in terms of two components - calibration, which is equal to reliability, and refinement that is equal to the difference between uncertainty, and resolution. Refinement component measures the extent to which each forecast-event pair, assessed with the same probability, is uniform in exhibiting occurrence or no occurrence of particular event. Similarly to AUC, refinement can be used as a measure of classification accuracy. Analytically it can be denoted as

$$\begin{gathered} B = \mathit{Calibration} + Refinement = \\ \frac{1}{N}\sum_{k=1}^{K} n_k(\hat{p}_k - y_k)^2 + \frac{1}{N}\sum_{k=1}^{K} n_k(y_k(1 - y_k)). \end{gathered} \tag{23}$$

Values of Brier score close to 0 inform about high accuracy of the model. Therefore I aim for both calibration and refinement to be as low as possible. Since the inherent uncertainty in the forecasts is 0.25, which is not affected by the choice of model, the values of Brier score that are below 0.25 indicate a skill in the forecasts.

Another measure that is useful in determining the calibration of the models is a calibration plot (curve) also known as reliability diagram. It relates the estimated probabilities (x-axis) with the actual outcomes (y-axis). The perfect calibration is denoted as a 45° line and outcome above this line informs about *underconfidence* of the model while outcomes below the line informs about *overconfidence* of the model. In other words this line has an intercept equal to 0 and slope equal to 1. The closer the slope and intercept of a particular model is to those values the better it is calibrated.

5 Empirical results

Models are estimated using the data spanned from 2012 until 2013 with a total of 173840 loaners (145875 non-defaulters and 27965 defaulters). Initially the model is evaluated on the sample from 2014, where there are 137003 loaners (105806 non-defaulters and 31197 defaulters). In order to check how robust the models are, they are evaluated on the samples from 2015 with 139547 loaners (102991 non-defaulters and 36556 defaulters) and 2016 with 43465 loaners (36368 non-defaulters and 7097 defaulters). This way it is easier to determine whether it is necessary to update the models each year or it is sufficient to estimate them once. Algorithm 3 is employed to obtain and evaluate the probabilities of default:

Algorithm 3 Research design

1. Sample 1000 defaulters and non-defaulters from the training and validation samples.
2. Estimate the models and make predictions using the methods outlined in Sections 4.1 and 4.2.
3. Validate the models using methods outlined in Section 4.4.
4. Repeat steps $1 - 3$ 1000 times.
5. Calculate a mean of the validation measures.

The choice of the sample size is motivated by the fact that qualitative variables have levels, and a considerable number of occurrences (for each level) is required for a sufficient representation. Moreover, such sample size allows to carry out the computations in a reasonable amount of time. Each sample is created such that the number of defaulters and non-defaulters is equal, even though, the number of non-defaulters is much higher. Haibo and Garcia (2009) point out that imbalanced data tends to significantly deteriorate the overall performance of different approaches. As mentioned before the research design employs bagging since it allows to obtain stable results and decrease the variance of the estimates (Breiman 1996).

Since most of the quantitative predictors are expressed in different units they are standardized before the model estimation such that

$$\sum_{i=1}^{N} x_{ij} = 0 \qquad \text{and} \qquad \sum_{i=1}^{N} x_{ij}^2 = 1.$$

This way predictors have mean equal to 0 and variance equal to 1. Standardization allows to avoid the bias towards the predictors with large coefficients and allows for a fair penalization of the parameters. If the predictors would not be standardized their coefficients would differ significantly in their magnitude and therefore contribute differently towards the penalty factor. Standardized predictors are expressed on the same scale and hence allow for a fair shrinkage of the coefficients.

In Section 4.2 it was mentioned that SVM can be tuned using three different kernels: linear, polynomial, and radial. SVMs, with all three different kernels, are fitted according to the Algorithm 3 in order to choose the leading one that is employed in the subsequent analysis. The results are presented in Table 3.

Table 3: Evaluation of support vector machines tuned with three different kernels: linear, polynomial and radial in years 2012-2013 and validated in year 2014

Kernel	AUC (95% conf. int.)	Brier score (95% conf. int.)
Linear	0.6891 (0.6663 – 0.7122)	0.2232 (0.2228 – 0.2236)
Polynomial	0.5890 (0.5643 – 0.6140)	0.2471 (0.2465 – 0.2478)
Radial	0.6831 (0.6599 – 0.7062)	0.2252 (0.2248 – 0.2257)

The SVMs are evaluated using the AUC and the Brier score. SVC obtains the best results, both in terms of AUC and Brier score and hence is going to be used in the subsequent analysis. Radial kernel performs a bit worse compared to the SVC, while polynomial kernel is inferior to both mentioned kernels.

The comparison starts with an evaluation of discriminatory power of different models using AUC alongside its constituents, sensitivity and specificity. The results of the evaluation in year 2014 are presented in Table 4.

Table 4: Area under the curve (with 95% confidence intervals), sensitivity and specificity of the models estimated in years 2012-2013 and validated in year 2014

Model	AUC (95% conf. int.)	Sensitivity	Specificity
Logistic regression	0.6913 (0.6683 – 0.7142)	0.6670	0.6245
Lasso	0.6958 (0.6730 – 0.7186)	0.6920	0.6037
Ridge	0.7024 (0.6798 – 0.7251)	0.6741	0.6321
Elastic-net	0.6968 (0.6788 – 0.7241)	0.6944	0.6025
Relaxed lasso	0.6996 (0.6771 – 0.7224)	0.7056	0.6031
Lasso with logistic regression	0.6971 (0.6748 – 0.7203)	0.6879	0.6111
Adaptive lasso	0.6807 (0.6577 – 0.7040)	0.6874	0.5871
Support vector classifier	0.6891 (0.6663 – 0.7122)	0.6630	0.6251
Random forest	0.6901 (0.6672 – 0.7131)	0.6708	0.6166

All methods obtain similar results and their AUC oscillates around 0.7 with the ridge regression being the only approach that exceeds this value. It is followed by the relaxed lasso and the two step approach that combines lasso with logistic regression. Elastic-net and lasso are the next best performing methods with the former benefiting from the good classification accuracy of the ridge regression. Logistic regression is next, right before two machine learning algorithms which are one of the worst performing models, with SVC performing better than random forest. Adaptive lasso is the worst approach overall. With an exception of adaptive lasso I note that all regularization methods outperform their benchmarks when evaluated under the AUC. Concerning its constituents all models yield better results for sensitivity which means that they do better in distinguishing defaulters than non-defaulters. Generally, a method yielding high sensitivity also gives low specificity. Relaxed lasso yields the highest sensitivity but at the same time its specificity is one of the lowest. Elastic-net is the second best model in terms of sensitivity, however, it performs poorly in terms of specificity, being the second worst approach. Lasso is the third best approach in terms of sensitivity but obtains only a moderate result in terms of specificity. Two step approach obtains a similar result to adaptive lasso and those models are placed fourth and fifth in terms of sensitivity. Sixth best approach is ridge regression which is the best approach considering the specificity. Random forest is one of the poorest performing models in terms of sensitivity and obtains only a moderate result in terms of specificity. Logistic regression and SVC are the two worst models, however, they obtain high results in terms of specificity. It is worth repeating that specificity and sensitivity are calculated using the subjective cut-off value and it might be the case that using a different cut-off would give very different results. Still, concerning both the AUC and sensitivity with specificity we can sum up by stating that ridge regression and relaxed lasso are the leading methods.

In Section 4.4 it was pointed out that AUC has many flaws and may be misleading measure in the model evaluation. Therefore models are also evaluated using the Brier score which assesses both the discriminatory power and the calibration of the models. The results for an evaluation in 2014 are presented in Table 5.

Table 5: Brier score decomposition of the models estimated in years 2012-2013 and validated in year 2014

Model	Brier score (95% conf. int.)	Calibration	Refinement
Logistic regression	0.2246 (0.2242 – 0.2251)	0.0032	0.2214
Lasso	0.2220 (0.2215 – 0.2224)	0.0021	0.2199
Ridge	0.2209 (0.2204 – 0.2214)	0.0030	0.2179
Elastic-net	0.2219 (0.2214 – 0.2223)	0.0023	0.2196
Relaxed lasso	0.2197 (0.2193 – 0.2221)	0.0012	0.2187
Lasso with logistic regression	0.2211 (0.2206 – 0.2215)	0.0016	0.2195
Adaptive lasso	0.2258 (0.2254 – 0.2263)	0.0010	0.2248
Support vector classifier	0.2232 (0.2228 – 0.2236)	0.0011	0.2221
Random forest	0.2237 (0.2233 – 0.2241)	0.0019	0.2218

Analogous to the AUC analysis, all the methods obtain similar Brier scores. As it was mentioned before, a successful Brier score requires a good performance both in terms of classification and calibration. Since the latter is involved now, the corollaries differ from what was pointed out in terms of AUC. Analyzing Table 5 I note that relaxed lasso is the best model under the Brier score. It combines both high calibration and refinement. Second best model is the ridge regression which maintains high classification accuracy but, at the same time, it is poorly calibrated. It can be seen by means of Figure 1 where ridge regression is very distant from the perfect calibration line. Therefore, even though it yields accurate predictions, its performance is deteriorated by poor calibration.

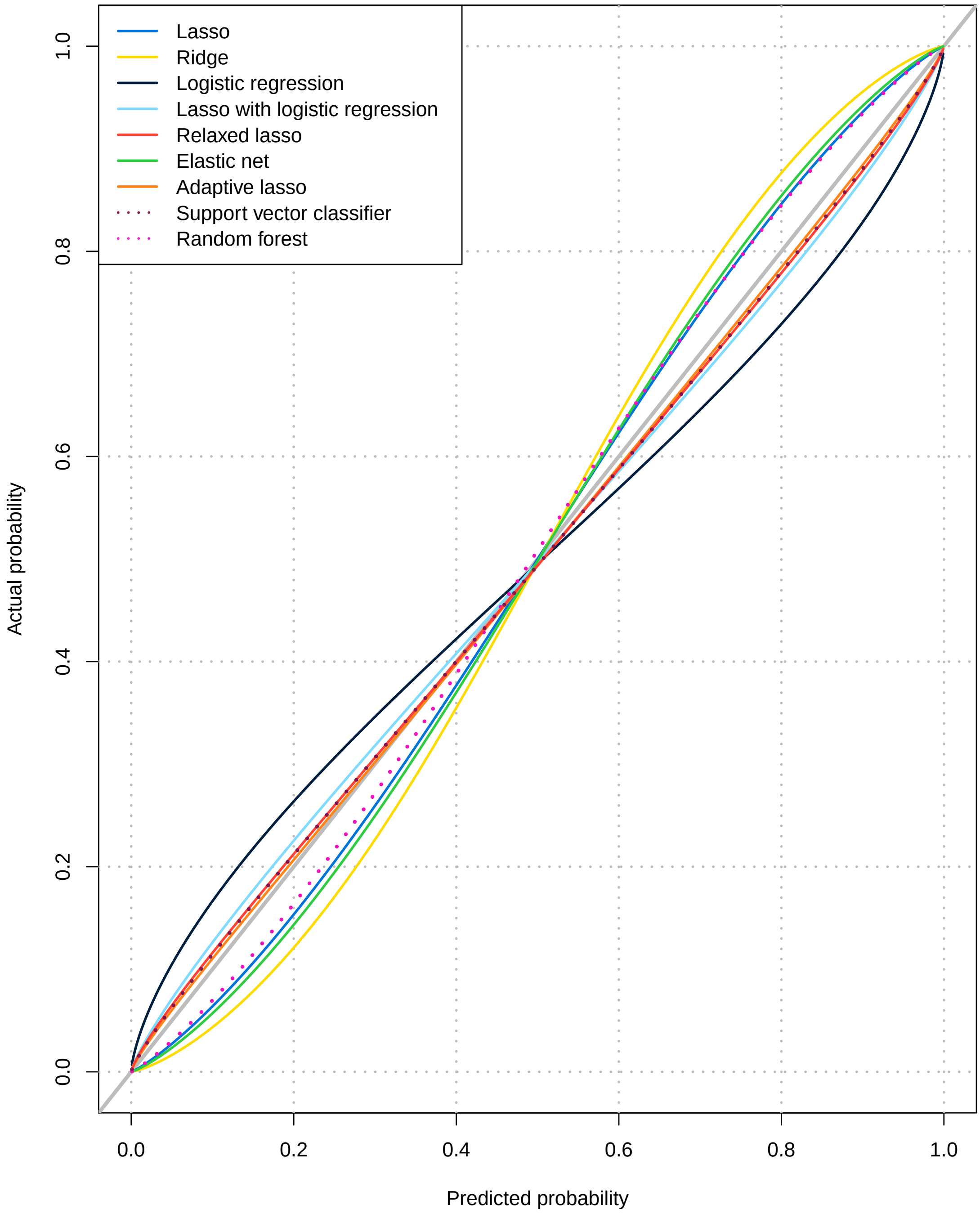

Figure 1: Calibration plot of the models estimated in years 2012-2013 and validated in year 2014

The two stage procedure is the third best performing model with a score slightly higher than the ridge regression. Even though it yields worse results in terms of classification it is well calibrated and improves upon both lasso and logistic regression applied individually.

Concerning the former one can see that debiasing the coefficients considerably improves the calibration by looking at the calibration plot. The curve of the two stage method is much closer to the 45° line compared to the curve of logistic regression. Similarly to the analysis under AUC, elastic-net and lasso are the next best performing methods, however, the difference between them is smaller as elastic-net inherits poor calibration of ridge regression. Logistic regression is inferior to both SVC and random forest as it is much worse calibrated. Unlike in the case of AUC, SVC outperforms random forest as, alongside adaptive lasso, it's the best calibrated model. Even though adaptive lasso is well calibrated it is the worst model just as in the case of evaluation under the AUC.

Concerning the results of AUC and Brier score I sum up by naming relaxed lasso the overall leader. It has high discriminatory power and is well calibrated. Moreover, it is clear that AUC and Brier score do not yield the same results but concerning the flaws of AUC, the Brier score is the decisive measure to assess the models.

To check whether the above conclusions hold in general, I perform the robustness check for the two subsequent years. The results for year 2015 are presented in Tables 22, 23 and Figure 8 which can be found in the Appendix. In terms of AUC, the results for 2015 are similar to the ones for 2014. Ridge regression and relaxed lasso are two best performing models and the biggest difference compared to the prior analysis is a poor performance of two step approach which is no longer one of the best performing models as it dropped to sixth place. It is inferior to elastic-net and lasso, which come respectively as the third and fourth best performing methods, and random forest, which also obtains slightly better classification accuracy than logistic regression. The two bottom places remain unchanged as SVC and adaptive lasso obtain the lowest AUC values. Concerning the sensitivity and specificity, again, models perform better in distinguishing defaulters than non-defaulters as they all yield better results for sensitivity than specificity. Logistic regression, lasso and two step approach obtain the highest values for sensitivity but at the same time their specificity results are poor. Adaptive lasso obtains similar results in terms of sensitivity but it yields the lowest specificity out of all nine approaches. Elastic-net obtains average results in both measures. Ridge regression and relaxed lasso obtain the highest values in terms of specificity with the former obtaining second lowest result in terms of sensitivity. Random forest produces high specificity, however, it also yields the lowest sensitivity. In terms of Brier score, relaxed lasso is the leading model again with ridge regression coming as the close second best. In 2015 the difference between them is smaller than in 2014. As in the previous year, relaxed lasso overcomes ridge regression due to a poor calibration of the latter. Elastic-net is the third best performing model as both ridge and lasso perform well which automatically elevates this approach. Lasso comes very close as the fourth best model obtaining almost identical result as the elastic-net. The difference between them lays more in the classification power, which is better in case of the elastic-net, since it inherits the discrimination accuracy of the ridge regression. The next two model are random forest and SVC. The latter still performs very well in terms of calibration but its classification accuracy is second lowest among all the approaches. Random forest improved in terms of classification and maintained similar level of calibration compared to the previous year. As mentioned before, the discriminatory power of the two step approach considerably deteriorated which is also reflected in the poor performance under Brier score, where it comes as the third worst model. Logistic regression maintains its poor performance as it comes second worst with both weak classification accuracy and calibration. As in 2014, adaptive lasso is the worst method overall.

Tables 24, 25, and Figure 9 introduce the evaluation measures for the models validated in 2016. Similarly to the results for year 2015, they can all be found in the Appendix. In general, the discriminatory power of all methods has deteriorated compared to the previous years. For the third consecutive year ridge regression is the leading classifier, and relaxed lasso is the second best under AUC. Surprisingly, random forest is the third best method as it was performing moderately in 2014 and 2015. Elastic-net continues its good performance and obtains almost the same result as lasso. Unexpectedly, logistic regression model performs better than the two step approach which is the second worst performing model. SVC is again dominated by the logistic regression and improves only upon the two step approach and the adaptive lasso which is, once more, the worst model overall. Similarly to the previous years, all models obtain higher results for sensitivity than specificity. Random forest, elastic-net and adaptive lasso obtain the highest sensitivity values while the ridge regression yields the best result in terms of specificity which is only slightly lower than its sensitivity. Lasso and relaxed lasso are next with the latter also obtaining the third highest specificity that is slightly higher for logistic regression which is the second best approach in terms of specificity. Still its sensitivity is on moderate level as ridge regression, two step approach and SVC obtain the lowest values. Concerning the performance under the Brier score there are no changes in terms of the best performing models as relaxed lasso and ridge regression obtain the lowest results. In 2016, however, they obtain identical results which origins from a much better calibration of the ridge regression compared to the previous years. As in terms of AUC, random forest is the third best performing method combining high discriminatory power with good calibration. Lasso is the next best method, obtaining the best calibration among all the methods. Elastic-net follows lasso closely as its calibration is a little worse due to the involvement of ridge regression. SVC manages to beat the logistic regression and two step approach, while adaptive lasso is the worse performing model for the third consecutive year.

Apart from the ridge regression all regularization methods allow to obtain a subset of variables from the original sample. Identification of such variables may be helpful in the process of distinguishing defaulters from non-defaulters. The leading method in this study, the relaxed lasso, on average, retains 21 out of 45 variables which are presented in the Table 7 alongside with their frequency. In the beginning of this paper it was noted that lenders may be able to improve their decision making process by screening borrowers using other features than only their credit grade. To check whether the predictors improve the discrimination of the borrowers two logistic regression models were estimated. In the first one the credit grade is the single predictor:

$$\log\left(\frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)}\right) = \gamma_0 + \gamma_1 C, \tag{24}$$

where C is the credit grade assigned by the LendingClub. It takes the values from A to G, where grade A is assigned to the best borrowers and grade G to the worst ones. The second regression involves credit grade alongside all the predictors kept on average by the relaxed lasso:

$$\log\left(\frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)}\right) = \gamma_0 + \gamma_1 C + \gamma_2 x_1 + ... + \gamma_{k+1} x_k. \tag{25}$$

Table 6: Comparison of the predictive accuracy between model using LendingClub credit grade and model using LendingClub credit grade with predictors

Model	Measure	2014	2015	2016
Credit grade	AUC	0.5014	0.5026	0.4961
	Brier score	0.2690	0.2691	0.2695
Credit grade with predictors	AUC	0.6900	0.6920	0.6827
	Brier score	0.2236	0.2238	0.2242

Table 7: Frequency of variables chosen by relaxed lasso in years 2012-2013 (in %)

Variables	Frequency
Annual income	**100.0**
Interest rate	**100.0**
Term	**100.0**
Debt to income ratio	**91.2**
Number of trades opened in the last 24 months	**87.5**
Loan amount	**72.8**
Months since last new bankcard	**69.1**
Revolving bankcards	**66.9**
Number of accounts opened in the last 12 months	**63.2**
Loan purpose	**57.4**
Inquiries in the last 6 months	**53.7**
Revolving trades	**52.9**
Bankcard limit	**51.5**
Total number of active revolving trades	**50.0**
Housing status	**49.3**
Months since recent account	**43.4**
Percent of never delinquent trades	**41.2**
Total number of installment accounts	**40.4**
Total number of mortgage accounts	**40.4**
Revolving utility	**40.4**
Average balance	**39.0**
Months since recent revolving account	36.0
Credit limit	35.3
Bankcards above 75% of limit	34.6
Months since oldest revolving account	34.6
Months since recent inquiry	34.6
Total accounts	34.6
Verification status	34.5
Listing status	31.6
Installments	30.9
Employment length	25.0
Months since oldest installment account	25.0
Installment limit	22.1
Satisfactory bankcard accounts	21.3
Active bankcard accounts	19.9
Balance to credit limit ratio	19.1
Revolving balance	19.1
Bankcard accounts	17.6
Balance excluding mortgage	15.4
Current balance on all accounts	15.4
Revolving credit limit	14.7
Revolving accounts	11.8
Open revolving accounts	8.1
Current accounts	7.4
Satisfactory accounts	7.4

Table 6 shows that model using credit grade and predictors, chosen by relaxed lasso, results in considerably better predictive accuracy then model using credit grade only. The latter performs closely to a random guess indicating that lenders should also look into other characteristics of the potential borrowers.

Furthermore, the relations between particular predictors and their impact on the probabilities of default are checked. The following analysis was performed using relaxed lasso estimated on the sub-sample equal to 2000 observations from years $2012-2013$ and evaluated in year 2014. To ensure the representativeness of the sub-samples, the descriptive statistics were calculated (Tables 18, 19, 20, and 21) and compared with the corresponding metrics for the whole samples (Tables 10, 11, 12, and 13). Since the results are comparable across years the following part of this paper is based on the results from 2014 only.

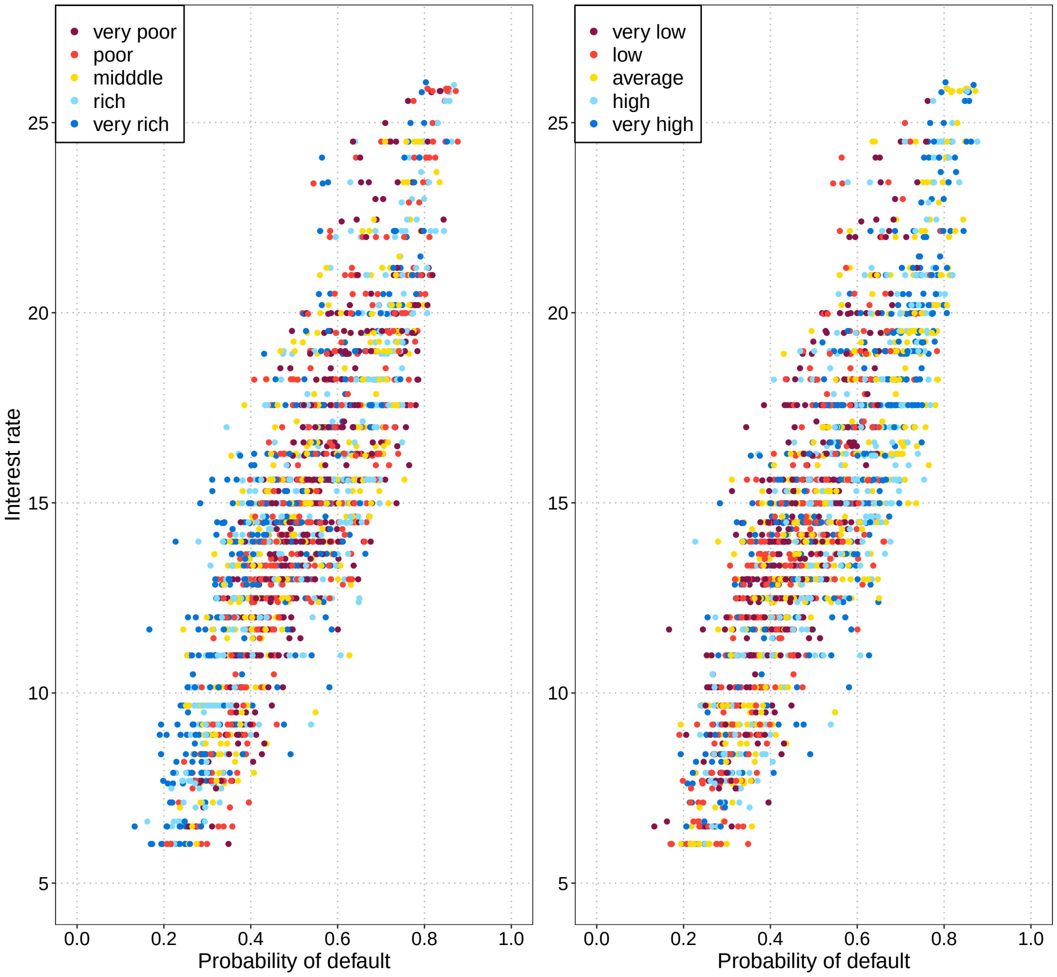

Figure 2: Relation between interest rate and probability of default. Left: Annual income of the borrowers split into five groups. Right: Loan amount split into five groups

Looking at Figure 2, there is a positive relation between the probability of default and the interest rates as the individuals with high interest-bearing loans are usually viewed as doubtful borrowers whose high probability of insolvency must be compensated with high interest rates. LendingClub sets the interest rates based on the following factors: (i) the LendingClub base rate, set to 5.05%, which is the starting point for each loan, (ii) risk and volatility adjustment depending on the FICO Score, (iii) loan amount limits determined by the FICO Score, and (iv) other risk modifiers like length of credit history. Interest rates can also be altered depending on the macroeconomic conditions, supply and demand on the platform, and the default and charge off rates.

Borrowers' credibility can also be assessed by checking their income. To assess whether different income levels affect the default probabilities, the annual income is split into five distinct groups that indicate whether a borrower is very poor, poor, in the middle, rich or very rich. Looking at the left plot of Figure 2 it is hard to notice a considerable differences between the probability of default and the annual income. One may argue that borrowers with higher annual income are characterized by both low probability of default and low interest rates, however, this relation is not very clear. The same treatment is applied to the loan amount which, similarly to annual income, does not show a clear relation with the probability of default as can be seen on the right plot of Figure 2. The reason is that the borrowers do not necessarily borrow in accordance to their salaries. To account for the relation between the income and loan amount the same analysis is performed with debt to income ratio. Looking at Figure 3 one can see that borrowers with high values of debt to income ratio tend to have higher probability of default and interest rates. Therefore, the probability of default does not necessarily depend on annual income or loan amount unless these values are analysed with respect to each other.

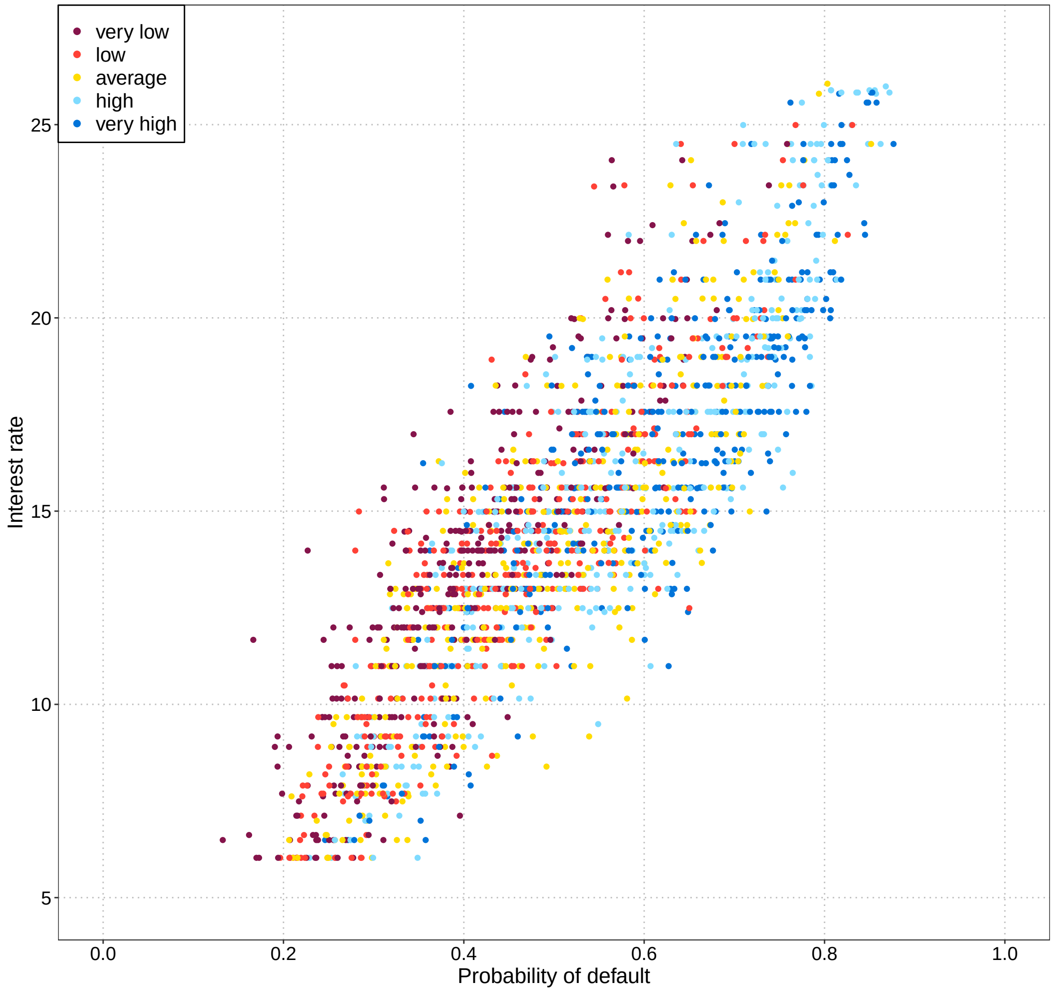

Figure 3: Relation between interest rate and probability of default with respect to the debt to income ratio

LendingClub clients can receive either 36- or 60-month loans. 60-month loans are associated with higher interest for two reasons. First, since the lenders freeze their funds for a prolonged period of time they expect a liquidity premium. Moreover, high interest rates set on loans act like a protection from an increase in interest rates set by the central banks which are more likely to change in a longer period of time. The second reason comes directly from the right plot of Figure 4 where one can see that lending for 60 months is related to considerably higher probability of default. Lending for a long time periods is more risky as the total payout obligation is higher due to the interest accrued over time.

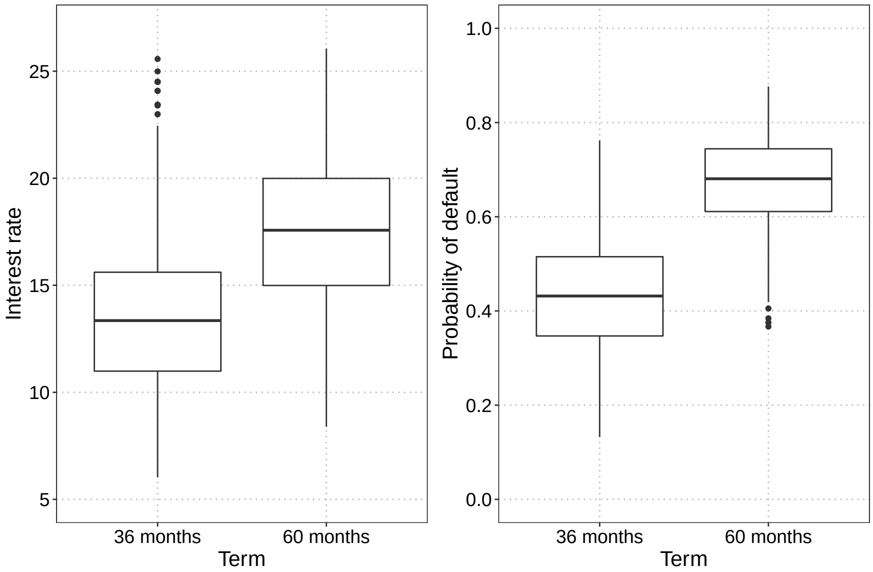

Figure 4: Relation between term and interest rate (left) and probability of default (right)

Borrowers frequently have more financial activity while applying for a loan. Figure 5 shows that taking any additional financial activity leads, on average, to higher probability of default, however for 5 or more accounts opened in the last 12 months the probability of default decreases. Regardless of the number of recently started trades or opened accounts the probability of default frequently spans from low values of default probability (around 0.2) to high values (around 0.8). On one hand, opening additional accounts does not necessarily have to lead to an increase in the default probability. Borrowers always need to meet certain criteria while applying for the next loan and it can be the case that some borrowers have a stable financial situation or a collateral. On the other hand, as most of the LendingClub borrowers need to consolidate already existing debt (Figure 7) peer-to-peer lending may just be the way to obtain the required funds since the requirements are not as stringent as in case of the traditional loan issuers. However, taking a loan to repay already existing loans can lead to infinite loop of insolvency which can considerably damage borrowers' creditworthiness. Moreover, as indicated in Table 9 most of the LendingClub borrowers do not own a house so they cannot increase they credibility with a collateral.

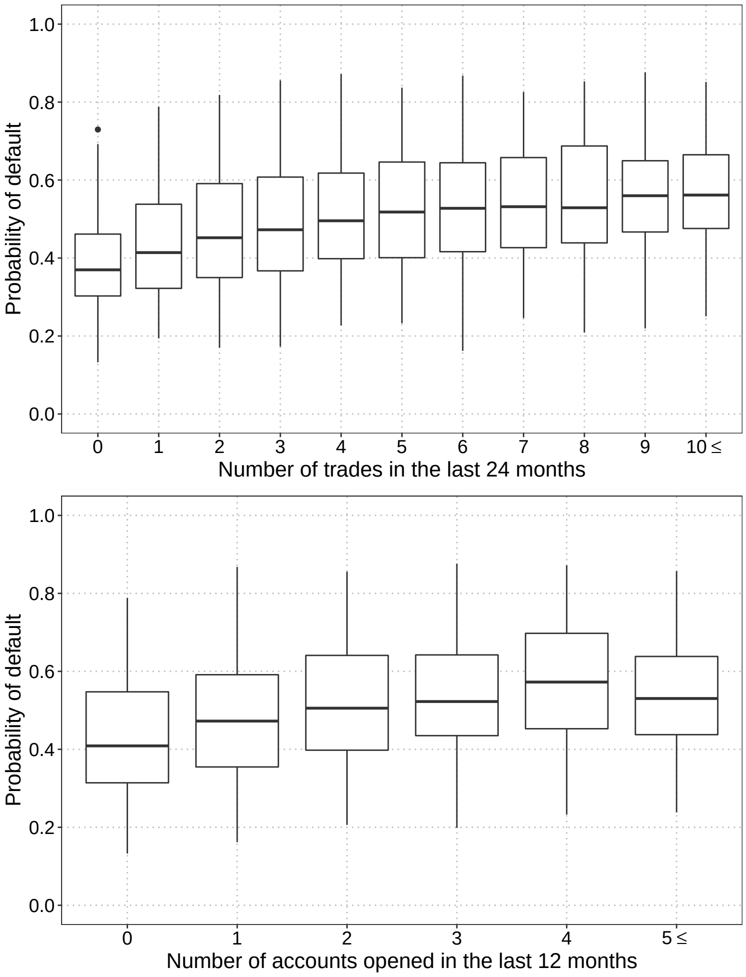

Figure 5: Relation between the probability of default and recent financial activity of the borrowers

Figure 6 shows that the relation between default probability and the total number of revolving, installment, and mortgage accounts is not very clear. It is worth mentioning that total number of active revolving trades and total number of revolving trades are very highly correlated and relaxed lasso never picked both of them in a single model. For both revolving trades and installment accounts the probability of default fluctuates between different numbers and, on average, remains on the same level. For mortgage accounts the default probability decreases at first, however, for 5 or more mortgages it increases. To further explore this issue I list variables that are potentially related with the number of the mortgage accounts in Table 8.

Table 8: Average values of selected predictors per number of the mortgage accounts

Total number of mortgage accounts	Probability of default	Interest rate	Debt to income ratio	Bankcard limit	Number of accounts opened in the last 12 months
0	53.01%	15.11%	19.21%	14602	4.67
1	50.89%	14.52%	18.99%	18382	4.84
2	49.63%	14.51%	18.24%	19409	4.65
3	49.51%	14.48%	17.74%	20098	4.88
4	44.99%	13.13%	16.95%	21090	4.47
Above 5	48.13%	14.29%	16.92%	23731	5.32

Debt to income ratio gradually decreases with the number of mortgage accounts, however, the difference between borrowers who have 4 and 5 or more such accounts is very small. There is no clear relation between the number of recent accounts and the number of mortgage accounts as, for instance, borrowers with 4 mortgages have, on average, less recent accounts than borrowers with lower number of mortgages. However, for borrowers with 5 mortgage accounts the number of recent accounts is the highest.[2] The bankcard limit increases gradually with the number of mortgage accounts.

Table 9: Frequency of loan term and housing status per number of the mortgage accounts

Total number of mortgage accounts	Term		Housing status		
	36 months	60 months	Mortgage	Own	Rent
0	75.12%	24.88%	0.00%	17.04%	82.96%
1	66.15%	33.85%	68.32%	8.70%	22.98%
2	64.98%	35.02%	69.26%	8.56%	22.18%
3	62.00%	38.00%	75.50%	5.50%	19.00%
4	70.32%	29.68%	80.65%	5.16%	14.19%
Above 5	58.72%	41.28%	78.90%	5.50%	15.60%

For the borrowers with 5 or more mortgages more than 40% of the accounts are long term. As it was indicated before 60-month loans are more risky than 36-month ones. Regarding the housing status the borrowers with 5 or more mortgages do not differ considerably from the borrowers with 3 or 4 mortgages. Borrowers without mortgage accounts have a completely different housing situation as most of them rent a place and more than 17% own a flat or a house.

Summarizing borrowers with 5 or more mortgages have comparable debt to income ratio with borrowers that have 4 mortgages, high number of recently opened accounts and

[2] Number of recent accounts do not necessarily reflect the number of mortgage accounts as the Spearman's correlation between them is very low ($\rho = 0.0125$).

many long term loans. Therefore, their riskiness may come from the way they are conducting their finances as they seem to allocate funds in different directions for prolonged time periods. Some borrowers may treat buying real estates as an investment and even if they are very credible at some point their creditworthiness is put in doubt which requires to cushion potential risk by higher interest rates that lead to higher probability of default. It is much harder to justify why the recently inquired accounts lead to a higher probability of default but already existing obligations do not. One potential explanation is that more recent activity can have a larger impact on borrowers' creditworthiness compared to already established ones as the latter are already accounted for by the borrowers in their financial plans while the burden of new inquiries may eventually become overwhelming.

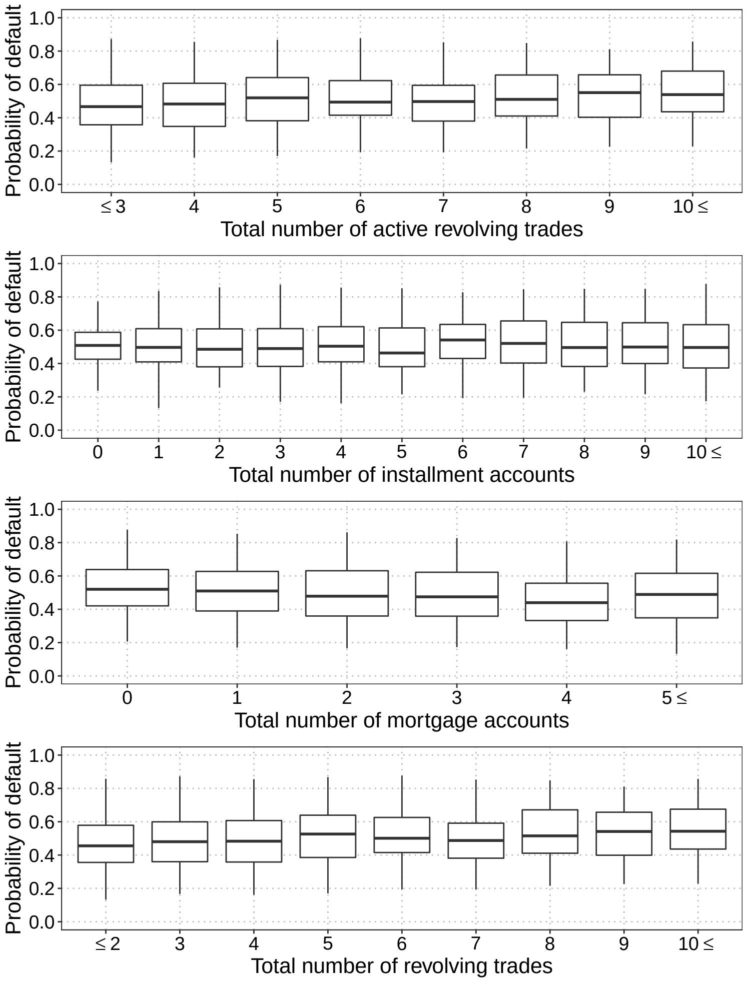

Figure 6: Relation between the probability of default and current financial activities of the borrowers

Bankcards are one of the most common and flexible lines of credit that one can use. Since they are issued by banks the requirements to obtain one are more stringent compared to the ones set by LendingClub. Therefore, the two bankcard related variables: total limit and total open to buy on revolving bankcards decrease one's default probability. Individuals with high card limits tend to be more reliable in terms of loan issuing as they have more funds which can be used to cover for repaying installments or the whole loan. Higher available revolving credit limit means that an individual is more credible which makes him a safer borrower. Looking at Figure 7 one can see that the default probability associated with credit cards is the lowest which confirms the above statements.

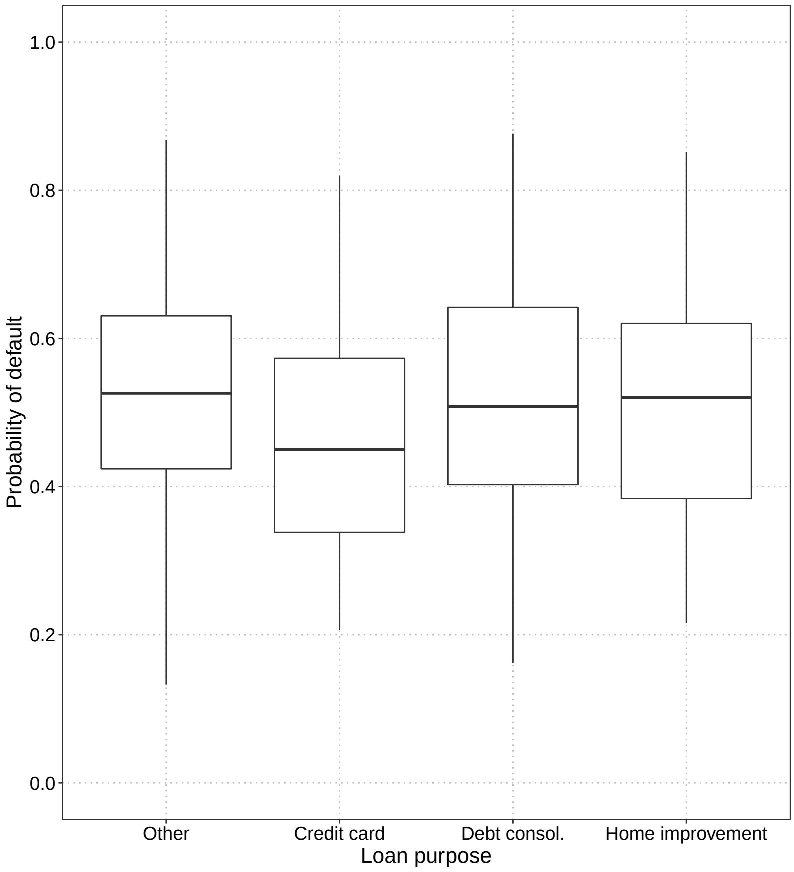

Figure 7: Relation between the probability of default and loan purposes

6 Discussion

The analysis of the results obtained in 2014 alongside with robustness check in years 2015 and 2016 allow to form some general conclusions. As it was mentioned before, Brier score is a more reliable evaluation method, compared to the AUC, since it accounts for both calibration and classification accuracy. Therefore, Brier score is a decisive evaluation method when it yields different results compared to AUC. In years 2014-2016 relaxed lasso and ridge regression constantly outperform other methods in terms of both AUC and Brier Score. However, since the relaxed lasso obtains the lowest Brier score in two out of three years, and is a joint best method alongside ridge regression in the last year, it is the overall leader. Hastie et al. (2017) note that relaxed lasso overcomes the lasso in their study and point out its high prediction accuracy. Ridge regression comes as the second best model and is the best classifier in general, however, it suffers from unsatisfactory calibration which deteriorates its performance. Ridge regression's high prediction accuracy is noted also in the work of Tibshirani (1996). Choosing the third best performing model is harder as there is not a method that always outperform the remaining models. Two step approach, elastic-net and random forest all have been third best performing methods in terms of both Brier score and AUC. Outside of 2015 when elastic-net is the third best performing method, it is fourth and fifth best in terms of Brier score and fourth in terms of AUC. Both two stage approach and random forest tend to perform moderately or poorly in other years which leads to a conclusion that elastic-net is the third best performing method. Up to certain extent its good performance is due to high classification accuracy of ridge regression but also due to better calibration compared to its constituent. Pointing out the fourth and next best performing models is much tougher and of limited use so more general remarks are presented.

Logistic regression is a market standard approach that is frequently employed as a benchmark for the other approaches. In this study it is usually outperformed by all methods with an exception of adaptive lasso. This is in line with some previous research as Härdle et al. (2007) note that SVM using radial kernel is better than logistic regression. On one hand logistic regression is a very straightforward approach that, theoretically speaking, should be outperformed by the more advanced methods so this result should not be surprising, however on the other hand a method that is a benchmark for other approaches should be harder to beat. At that point I agree with Lessmann et al. (2015)

who suggest using different approach as a benchmark but according to this study random forest, which they endorse, is not the right candidate. Even though it performs well in 2016 it is performing moderately at best in other years. As a competitive benchmark I suggest the elastic net which combines good calibration of lasso with high classification power of ridge regression. Another rationale behind choosing the elastic-net is that it is not a *black box* approach with uninterpretable results but a model in regression framework that yields estimates of a subset of variables from the original sample. Therefore it enables one to identify the most important predictors which may be helpful in determining the reasons behind someone's default. Relaxed lasso could also be used as a benchmark since it possesses all the virtues of the elastic-net and moreover it has oracle properties which is beneficial in terms of choosing the subset of variables. Another important issue that emerged is the debiasing. Both relaxed lasso and the two step approach debias the coefficients but do it in two different ways. Relaxed lasso uses *adaptive* debiasing that is suited for each case while the two step approach uses *hard* debiasing that simply derives from the design of the method. The results show that hard debiasing lead to inconsistent results as the two step approach performs quite well in one year but then very poor in the next year. In this paper machine learning algorithms are outperformed by most regularization methods which seems to be quite unlikely concerning the fact that the latter are rarely employed in the forecasting studies. In Lessmann et al. (2015) a total of 41 models are compared coming from years of research in credit scoring and the only regularization method concerned is the ridge regression while most approaches are different machine learning algorithms. Even though ridge regression performs poorly in their study, one should actually consider using other regularizations which might lead to an improved performance. Adaptive lasso is constantly the worst method both in terms of the AUC and the Brier score. It seems to be quite improbable concerning the performance of other regularizations, however, it is worth noting that it was introduced for the case of high-dimensional data, and since it is not the case for this research adaptive lasso is just not adequate for this framework.

Adding additional predictors considerably improves the prediction accuracy. Iyer et al. (2015) also point out that, in the P2P market, lenders using additional information outperform the credit scores assigned to the borrowers. Besides the financial variables they also examine non-standard information like borrowers' profile picture or text description of their loan purpose. Subset of variables, obtained from relaxed lasso, allows to better understand relations between predictors and probability of default. Stiglitz and Weiss (1981) note that higher interest rates lead to higher probability of default as borrowers experience considerable difficulties with repaying high installments which got confirmed in this work. On the other hand there was no visible relation between the level of income and default probabilities. This relation was also studied in literature with different results. Albanesi et al. (2017) point out a positive relation between income and default probability while Beer et al. (2018) point out that income does not account for much of the variability in the credit scores, especially when other variables are taken into account. High debt to income ratio increases the probability of default which is an intuitive outcome as borrowers use too much of their funds to repay the loan. This variable was also used by Emekter et al. (2015).

Apart from the results presented above there are a couple of things that further research could take into account. Considering that the differences between the employed approaches are small it would be beneficial to test the regularization methods on different data sets, as in Lessmann et al. (2015). Moreover, by complementing this research one could check the performance of the regularization methods compared to wider range of models and validation methods. As it was mentioned before adaptive lasso and relaxed lasso have oracle properties. In order to further examine this property one could check whether those methods retain a consistent subset of predictors with consistent parameter estimates, as well as, verify whether relaxed lasso manages to distinguish between highly correlated predictors. Further extensions of the presented framework could include an examination of combined approaches, similar to the two step approach, as in the works of Zhang X. et al. (2016). Combination of stronger regularization methods with a good performing classifier could lead to a method that is superior, compared to the approaches outlined in this work.

7 Appendix

7.1 Descriptive statistics

Table 10: Descriptive statistics of quantitative and ordinal variables in 2012-2013

Variables	Minimum	Maximum	Mean	Median	St.dev	Skewness	Kurtosis
Loan amount	1000.00	35000.00	13844.58	12000.00	7976.16	0.79	0.06
Installments	21.62	1408.13	437.51	387.76	245.02	0.97	0.89
Annual income	4800.00	7141778.00	71713.53	61360.00	52288.65	31.71	3452.00
Debt to income ratio	0.00	34.99	16.91	16.60	7.59	0.14	-0.66
Inquiries in the last 6 months	0.00	8.00	0.81	0.00	1.03	1.43	2.20
Current accounts	0.00	62.00	10.95	10.00	4.60	0.98	1.94
Revolving balance	0.00	2568995.00	15970.95	12092.00	19269.58	28.09	2642.40
Total accounts	2.00	105.00	24.44	23.00	11.11	0.77	0.57
Current balance on all accounts	0.00	8000078.00	125614.36	75804.00	140063.78	4.05	85.84
Revolving credit limit	0.00	9999999.00	28401.15	22500.00	35516.11	131.50	35876.23
Number of trades opened in the last 24 months	0.00	40.00	3.95	4.00	2.61	1.15	2.89
Average balance	0.00	958084.00	12618.80	7393.00	15251.92	5.86	162.59
Revolving bankcards	0.00	497445.00	8012.69	3500.00	12935.09	4.37	39.36
Balance to credit limit ratio	0.00	339.60	66.77	71.80	25.59	-0.73	-0.18
Months since oldest installment account	0.00	649.00	124.72	127.00	45.94	0.35	2.62
Months since oldest revolving account	5.00	760.00	174.06	159.00	81.11	1.34	2.72
Months since recent revolving account	0.00	228.00	13.17	9.00	14.78	3.06	13.56
Months since recent account	0.00	211.00	8.44	6.00	8.93	3.81	23.33
Total number of mortgage accounts	0.00	31.00	1.74	1.00	2.15	1.66	4.20
Months since last new bankcard	0.00	554.00	24.83	15.00	28.58	3.02	17.89
Months since recent inquiry	0.00	24.00	6.81	6.00	5.42	1.03	0.62
Active bankcard accounts	0.00	30.00	3.62	3.00	1.92	1.36	3.78
Total number of active revolving trades	0.00	37.00	5.54	5.00	2.68	1.34	3.52
Satisfactory bankcard accounts	0.00	35.00	4.59	4.00	2.38	1.27	3.53
Bankcard accounts	0.00	65.00	8.86	8.00	4.46	1.35	3.45
Total number of installment accounts	0.00	66.00	7.42	6.00	6.08	2.00	6.04
Open revolving accounts	0.00	58.00	7.90	7.00	3.59	1.35	3.64
Revolving accounts	0.00	94.00	14.79	14.00	6.82	1.26	2.92
Revolving trades	0.00	37.00	5.55	5.00	2.68	1.34	3.51
Satisfactory accounts	0.00	62.00	10.94	10.00	4.41	1.09	2.46
Number of accounts opened in the last 12 months	0.00	25.00	1.83	2.00	1.40	1.31	5.15
Percent of never delinquent trades	15.00	100.00	96.09	100.00	7.03	-2.62	9.13
Bankcards above 75% of limit	0.00	100.00	53.02	50.00	33.37	-0.09	-1.10
Credit limit	0.00	9999999.00	153443.82	103419.00	155285.11	5.17	166.02
Balance excluding mortgage	0.00	2644442.00	41814.42	32323.00	39191.00	6.39	200.76
Bankcard limit	0.00	522210.00	19729.64	14400.00	18465.05	2.76	15.54
Installment limit	0.00	1214546.00	32621.27	25343.00	33778.08	3.42	30.53
Interest rate	6.00	26.06	14.02	13.98	4.37	0.26	-0.46
Revolving utility	0.00	140.40	58.04	60.10	23.09	-0.37	-0.56

Table 11: Distribution of qualitative variables in 2012-2013

Variables	Level	Fraction (in %)
Loan purpose	Debt consolidation	58.69
	Credit card	23.12
	Other	12.74
	Home improvement	5.45
Term	36 months	82.72
	60 months	17.28
Listing status	Fractional	78.89
	Whole	21.11
Housing status	Mortgage	50.64
	Rent	41.08
	Own	8.27
Employment length	0-3 years	33.87
	4-9 years	34.38
	10+ years	31.75
Verification status	Verified	44.20
	Not verified	33.79
	Source verified	22.00

Table 12: Descriptive statistics of quantitative and ordinal variables in 2014

Variables	Minimum	Maximum	Mean	Median	St.dev	Skewness	Kurtosis
Loan amount	1000.00	35000.00	14454.18	12250.00	8341.76	0.77	-0.11
Installments	30.42	1409.99	440.86	380.98	248.88	0.99	0.74
Annual income	3000.00	7500000.00	74772.56	65000.00	57669.72	38.31	4354.42
Debt to income ratio	0.00	39.99	17.66	17.24	7.94	0.21	-0.55
Inquiries in the last 6 months	0.00	6.00	0.83	0.00	1.08	1.56	2.65
Current accounts	0.00	84.00	11.64	11.00	5.22	1.25	3.16
Revolving balance	0.00	1298783.00	15669.98	10990.00	20469.74	12.55	401.99
Total accounts	2.00	150.00	26.45	25.00	12.03	0.82	1.06
Current balance on all accounts	0.00	3796811.00	139332.71	81740.00	152151.56	2.67	19.48
Revolving credit limit	0.00	1508600.00	29949.94	22100.00	30400.04	7.14	155.25
Number of trades opened in the last 24 months	0.00	53.00	4.65	4.00	2.92	1.14	3.21
Average balance	0.00	497484.00	13441.09	7523.00	16011.59	3.75	37.62
Revolving bankcards	0.00	232482.00	8478.89	3682.00	13328.05	3.79	22.51
Balance to credit limit ratio	0.00	255.20	63.30	67.10	26.76	-0.49	-0.66
Months since oldest installment account	0.00	490.00	127.32	130.00	50.17	0.35	1.91
Months since oldest revolving account	4.00	842.00	181.35	165.00	90.55	1.07	1.59
Months since recent revolving account	0.00	372.00	12.25	7.00	15.02	3.40	20.26
Months since recent account	0.00	194.00	7.48	5.00	8.02	3.98	28.37
Total number of mortgage accounts	0.00	34.00	1.89	1.00	2.20	1.62	4.90
Months since last new bankcard	0.00	533.00	22.79	13.00	28.39	3.28	18.95
Months since recent inquiry	0.00	25.00	6.53	5.00	5.58	1.06	0.49
Active bankcard accounts	0.00	24.00	3.60	3.00	2.10	1.25	3.08
Total number of active revolving trades	0.00	36.00	5.67	5.00	3.06	1.32	3.23
Satisfactory bankcard accounts	0.00	35.00	4.63	4.00	2.72	1.48	4.53
Bankcard accounts	0.00	61.00	8.69	8.00	4.90	1.24	2.80
Total number of installment accounts	0.00	97.00	8.75	7.00	7.36	1.90	5.77
Open revolving accounts	0.00	62.00	8.24	7.00	4.27	1.32	3.36
Revolving accounts	2.00	105.00	15.52	14.00	8.14	1.17	2.31
Revolving trades	0.00	37.00	5.62	5.00	3.04	1.32	3.28
Satisfactory accounts	0.00	84.00	11.58	11.00	5.22	1.25	3.16
Number of accounts opened in the last 12 months	0.00	26.00	2.13	2.00	1.65	1.26	4.32
Percent of never delinquent trades	16.70	100.00	94.36	97.70	8.29	-2.07	5.33
Bankcards above 75% of limit	0.00	100.00	49.24	50.00	34.82	0.06	-1.23
Credit limit	0.00	9999999.00	169191.26	110520.00	170612.04	3.99	100.40
Balance excluding mortgage	0.00	1359090.00	47742.82	36221.00	45077.56	3.83	36.17
Bankcard limit	0.00	560800.00	19461.27	13300.00	19696.04	2.74	16.03
Installment limit	0.00	1164167.00	39787.24	30091.00	40951.01	3.01	22.02
Interest rate	6.00	26.06	13.87	13.65	4.38	0.41	-0.09
Revolving utility	0.00	892.30	54.25	54.60	23.33	0.25	11.64

Table 13: Distribution of qualitative variables in 2014

Variables	Level	Fraction (in %)
Loan purpose	Debt consolidation	61.66
	Credit card	22.48
	Other	10.39
	Home improvement	5.47
Term	36 months	74.41
	60 months	25.59
Listing status	Fractional	49.62
	Whole	50.38
Housing status	Mortgage	50.74
	Rent	39.75
	Own	9.51
Employment length	0-3 years	35.97
	4-9 years	31.23
	10+ years	32.80
Verification status	Verified	29.26
	Not verified	30.68
	Source verified	40.06

Table 14: Descriptive statistics of quantitative and ordinal variables in 2015

Variables	Minimum	Maximum	Mean	Median	St.dev	Skewness	Kurtosis
Loan amount	1000.00	35000.00	14905.69	13000.00	8642.22	0.68	-0.31
Installments	30.54	1424.57	443.51	383.02	253.55	0.94	0.56
Annual income	0.00	8900060.00	76875.18	65000.00	72862.81	49.42	4929.10
Debt to income ratio	0.00	9999.00	18.84	18.22	28.11	320.70	113849.01
Inquiries in the last 6 months	0.00	6.00	0.66	0.00	0.94	1.65	2.95
Current accounts	1.00	90.00	11.95	11.00	5.59	1.25	2.88
Revolving balance	0.00	867528.00	16464.32	10950.00	23002.44	8.96	154.72
Total accounts	4.00	151.00	26.38	25.00	12.38	0.93	1.57
Current balance on all accounts	0.00	3726495.00	141944.36	81242.00	156235.16	2.72	18.65
Revolving credit limit	0.00	1035000.00	32648.87	23800.00	33517.53	5.31	61.97
Number of trades opened in the last 24 months	0.00	47.00	5.22	5.00	3.34	1.26	3.45
Average balance	0.00	395953.00	13379.61	7382.00	16040.74	3.63	29.43
Revolving bankcards	0.00	559912.00	9795.49	4518.00	14933.13	4.02	34.77
Balance to credit limit ratio	0.00	243.80	59.68	62.70	28.46	-0.34	-0.91
Months since oldest installment account	0.00	724.00	125.33	128.00	50.99	0.35	2.24
Months since oldest revolving account	4.00	775.00	179.87	163.00	91.64	1.07	1.65
Months since recent revolving account	0.00	315.00	11.89	7.00	15.04	3.65	21.41
Months since recent account	0.00	197.00	7.00	5.00	7.81	4.87	46.71
Total number of mortgage accounts	0.00	34.00	1.76	1.00	2.05	1.60	5.12
Months since last new bankcard	0.00	611.00	22.03	12.00	29.30	3.69	23.54
Months since recent inquiry	0.00	25.00	6.11	5.00	5.44	1.19	0.88
Active bankcard accounts	0.00	26.00	3.58	3.00	2.22	1.35	3.41
Total number of active revolving trades	0.00	44.00	5.64	5.00	3.35	1.55	4.57
Satisfactory bankcard accounts	0.00	63.00	4.76	4.00	3.03	1.78	7.39
Bankcard accounts	0.00	70.00	8.35	7.00	4.95	1.34	3.31
Total number of installment accounts	0.00	117.00	9.03	7.00	7.64	2.06	7.54
Open revolving accounts	0.00	83.00	8.43	8.00	4.64	1.40	3.76
Revolving accounts	2.00	102.00	15.31	14.00	8.47	1.26	2.69
Revolving trades	0.00	42.00	5.58	5.00	3.24	1.45	3.91
Satisfactory accounts	1.00	90.00	11.90	11.00	5.57	1.25	2.89
Number of accounts opened in the last 12 months	0.00	25.00	2.48	2.00	1.98	1.45	4.46
Percent of never delinquent trades	12.50	100.00	94.16	97.50	8.54	-2.10	5.70
Bankcards above 75% of limit	0.00	100.00	45.13	42.90	35.84	0.21	-1.26
Credit limit	2500.00	4214831.00	174383.07	112254.00	173574.64	2.63	17.34
Balance excluding mortgage	0.00	1684313.00	51449.57	39282.00	47739.77	3.69	33.25
Bankcard limit	0.00	684000.00	21129.64	14500.00	21478.89	2.95	21.44
Installment limit	0.00	2101913.00	43365.87	33322.00	43151.69	3.59	55.17
Interest rate	5.32	28.99	13.24	12.69	4.52	0.51	0.00
Revolving utility	0.00	153.70	51.37	51.40	24.47	-0.02	-0.79

Table 15: Distribution of qualitative variables in 2015

Variables	Level	Fraction (in %)
Loan purpose	Debt consolidation	61.95
	Credit card	20.55
	Other	11.24
	Home improvement	6.27
Term	36 months	70.60
	60 months	29.40
Listing status	Fractional	41.24
	Whole	58.76
Housing status	Mortgage	49.50
	Rent	39.69
	Own	10.81
Employment length	0-3 years	37.60
	4-9 years	29.61
	10+ years	32.79
Verification status	Verified	31.18
	Not verified	26.48
	Source verified	42.34

Table 16: Descriptive statistics of quantitative and ordinal variables in 2016

Variables	Minimum	Maximum	Mean	Median	St.dev	Skewness	Kurtosis
Loan amount	1000.00	40000.00	14819.83	12800.00	9250.85	0.69	-0.36
Installments	30.12	1536.95	453.88	383.94	284.21	0.94	0.41
Annual income	0.00	6500000.00	80880.73	69528.00	66177.67	26.03	2121.09
Debt to income ratio	0.00	9999.00	18.83	17.92	49.35	191.03	38491.52
Inquiries in the last 6 months	0.00	5.00	0.67	0.00	0.94	1.62	2.78
Current accounts	1.00	77.00	12.00	11.00	5.76	1.30	3.15
Revolving balance	0.00	1023940.00	16615.88	10644.00	25299.42	9.41	173.57
Total accounts	2.00	133.00	26.23	24.00	12.48	0.99	1.85
Current balance	0.00	2697239.00	155072.99	96780.00	168812.08	2.74	16.52
Revolving credit limit	0.00	1070650.00	35859.35	26300.00	37310.31	5.65	70.26
Number of trades opened in the last 24 months	0.00	42.00	5.53	5.00	3.56	1.35	3.89
Average balance	0.00	419840.00	14635.41	8572.00	17396.22	3.72	31.13
Revolving bankcards	0.00	454843.00	12101.00	6112.00	17278.27	3.93	31.08
Balance to credit limit ratio	0.00	189.80	54.00	54.90	29.56	-0.13	-1.08
Months since oldest installment account	1.00	506.00	124.89	129.00	50.98	0.27	1.85
Months since oldest revolving account	6.00	758.00	180.04	163.00	93.33	1.05	1.62
Months since recent revolving account	0.00	267.00	11.74	7.00	14.72	3.75	23.52
Months since recent account	0.00	166.00	6.73	5.00	7.28	4.65	42.61
Mortgage accounts	0.00	26.00	1.77	1.00	1.99	1.51	4.00
Months since last new bankcard	0.00	462.00	21.17	12.00	29.09	3.82	23.32
Months since recent inquiry	0.00	25.00	5.94	4.00	5.49	1.23	0.91
Active bankcard accounts	0.00	26.00	3.47	3.00	2.27	1.42	3.75
Active revolving trades	0.00	39.00	5.38	5.00	3.33	1.52	4.13
Satisfactory bankcard accounts	0.00	40.00	4.81	4.00	3.15	1.76	6.45
Bankcard accounts	0.00	68.00	8.13	7.00	4.94	1.42	3.98
Installment accounts	0.00	96.00	9.25	7.00	7.66	2.00	6.86
Open revolving accounts	0.00	73.00	8.41	7.00	4.78	1.46	4.23
Revolving accounts	2.00	103.00	14.95	13.00	8.47	1.33	3.21
Revolving trades	0.00	30.00	5.30	5.00	3.19	1.37	3.13
Satisfactory accounts	1.00	77.00	11.95	11.00	5.73	1.30	3.17
Number of accounts opened in the last 12 months	0.00	30.00	2.66	2.00	2.13	1.48	4.95
Percent of never delinquent trades	15.40	100.00	94.56	98.10	8.21	-2.27	7.01
Bankcards above 75% of limit	0.00	100.00	38.85	33.30	35.52	0.47	-1.09
Credit limit	2500.00	9999999.00	192399.17	132677.00	194325.83	5.51	165.06
Balance excluding mortgage	0.00	1548128.00	53439.82	40955.00	49688.97	3.80	39.67
Bankcard limit	0.00	474600.00	23309.83	16300.00	23054.89	2.74	14.96
Installment limit	0.00	2000000.00	45991.26	35313.00	45319.79	4.28	92.12
Interest rate	5.32	30.99	13.87	12.99	5.33	0.65	-0.07
Revolving utility	0.00	136.70	46.79	45.80	25.33	0.12	-0.83

Table 17: Distribution of qualitative variables in 2016

Variables	Level	Fraction (in %)
Loan purpose	Debt consolidation	56.62
	Credit card	18.14
	Other	16.86
	Home improvement	8.37
Term	36 months	74.51
	60 months	25.49
Listing status	Fractional	23.33
	Whole	76.67
Housing status	Mortgage	50.81
	Rent	35.65
	Own	13.53
Employment length	0-3 years	37.06
	4-9 years	28.15
	10+ years	34.79
Verification status	Verified	30.01
	Not verified	29.13
	Source verified	40.86

Table 18: Descriptive statistics of the quantitative and ordinal variables chosen by relaxed lasso in 2012 and 2013

Variables	Minimum	Maximum	Mean	Median	St.dev	Skewness	Kurtosis
Annual income	9765.00	600000.00	68839.53	60000.00	41651.28	3.21	21.85
Interest rate	6.03	25.89	14.95	14.47	4.42	0.17	-0.59
Debt to income ratio	0.00	34.91	17.77	17.58	7.70	0.05	-0.75
Number of trades opened in the last 24 months	0.00	20.00	4.16	4.00	2.65	1.04	1.95
Loan amount	1000.00	35000.00	14520.75	12662.50	8131.29	0.66	-0.19
Months since last new bankcard	0.00	227.00	22.95	15.00	26.22	2.74	11.19
Revolving bankcards	0.00	11.85	7.73	8.13	2.04	-1.63	4.03
Number of accounts opened in the last 12 months	0.00	12.00	1.94	2.00	1.46	1.30	4.05
Inquiries in the last 6 months	0.00	6.00	0.83	1.00	1.06	1.56	2.88
Revolving trades	0.00	21.00	5.68	5.00	2.71	1.20	2.28
Bankcard limit	0.00	165000.00	18711.43	14400.00	16588.57	2.40	9.66
Total number of active revolving trades	0.00	21.00	5.66	5.00	2.71	1.19	2.27
Months since recent account	1.00	94.00	7.92	8.38	0.87	3.88	22.92
Percent of never delinquent trades	52.40	100.00	96.21	100.00	6.82	-2.56	7.97
Installment accounts	0.00	39.00	7.30	6.00	5.67	1.67	3.78
Total number of mortgage accounts	0.00	14.00	1.60	1.00	2.03	1.56	2.58
Revolving utility	0.00	122.50	59.78	61.90	22.57	-0.44	-0.43
Average balance	0.00	11.89	8.87	8.91	1.05	-0.42	2.04

Table 19: Distribution of the qualitative variables chosen by relaxed lasso in 2012 and 2013

Variables	Level	Fraction (in %)
Loan purpose	Debt consolidation	58.45
	Credit card	21.45
	Other	14.65
	Home improvement	5.45
Term	36 months	74.85
	60 months	25.15
Housing status	Mortgage	48.75
	Rent	43.60
	Own	7.65

Table 20: Descriptive statistics of the quantitative and ordinal variables chosen by relaxed lasso in 2014

Variables	Minimum	Maximum	Mean	Median	St.dev	Skewness	Kurtosis
Annual income	11500.00	636000.00	72300.95	62500.00	43099.52	3.01	21.10
Interest rate	6.03	26.06	14.63	14.49	4.29	0.23	-0.20
Debt to income ratio	0.18	39.82	18.48	18.24	8.06	0.15	-0.61
Number of trades opened in the last 24 months	0.00	28.00	4.77	4.00	3.00	1.27	4.08
Loan amount	1000.00	35000.00	14665.00	12500.00	8508.59	0.75	-0.25
Months since last new bankcard	0.00	303.00	21.65	13.00	26.99	3.07	14.90
Revolving bankcards	0.00	11.84	7.82	8.15	1.90	-1.39	3.42
Number of accounts opened in the last 12 months	0.00	15.00	2.14	2.00	1.65	1.27	3.93
Inquiries in the last 6 months	0.00	6.00	0.86	1.00	1.09	1.55	2.73
Revolving trades	0.00	28.00	5.80	5.00	3.13	1.35	3.28
Bankcard limit	0.00	168500.00	17875.72	13000.00	17035.00	2.33	8.90
Total number of active revolving trades	0.00	28.00	5.83	5.00	3.15	1.36	3.31
Months since recent account	1.00	91.00	7.41	5.00	7.95	4.02	25.45
Percent of never delinquent trades	16.70	100.00	94.18	97.15	8.44	-2.37	8.91
Total number of installment accounts	0.00	67.00	8.68	7.00	7.56	2.07	6.56
Total number of mortgage accounts	0.00	23.00	1.71	1.00	2.13	1.76	6.25
Revolving utility	0.00	123.20	56.33	56.40	22.66	-0.12	-0.70
Average balance	5.61	12.14	8.83	8.74	1.12	0.02	-0.70

Table 21: Distribution of the qualitative variables chosen by relaxed lasso in 2014

Variables	Level	Fraction (in %)
Loan purpose	Debt consolidation	63.60
	Credit card	21.75
	Other	10.40
	Home improvement	4.25
Term	36 months	68.90
	60 months	31.10
Housing status	Mortgage	46.85
	Rent	44.15
	Own	9.00

7.2 Robustness check

Table 22: Area under the curve, sensitivity and specificity of the models estimated in years 2012-2013 and validated in 2015

	AUC (95% conf. int.)	Sensitivity	Specificity
Logistic regression	0.6908 (0.6678 – 0.7140)	0.6934	0.5963
Lasso	0.6951 (0.6723 – 0.7179)	0.6929	0.6003
Ridge	0.7041 (0.6826 – 0.7276)	0.6800	0.6301
Elastic-Net	0.6956 (0.6728 – 0.7184)	0.6907	0.6033
Relaxed lasso	0.7013 (0.6788 – 0.7241)	0.6856	0.6241
Lasso+Log. reg.	0.6919 (0.6690 – 0.7148)	0.6921	0.5964
Adaptive lasso	0.6767 (0.6536 – 0.6999)	0.6918	0.5755
Support vector classifier	0.6890 (0.6660 – 0.7120)	0.6874	0.5992
Random forest	0.6921 (0.6692 – 0.7150)	0.6720	0.6192

Table 23: Brier score decomposition of the models estimated in years 2012-2013 and validated in 2015

	Brier score (95% conf. int.)	Calibration	Refinement
Logistic regression	0.2249 (0.2245 – 0.2255)	0.0034	0.2215
Lasso	0.2220 (0.2216 – 0.2225)	0.0018	0.2202
Ridge	0.2200 (0.2195 – 0.2205)	0.0027	0.2173
Elastic-Net	0.2219 (0.2214 – 0.2223)	0.0019	0.2200
Relaxed lasso	0.2197 (0.2193 – 0.2201)	0.0015	0.2182
Lasso+Log. reg.	0.2234 (0.2229 – 0.2238)	0.0022	0.2212
Adaptive lasso	0.2275 (0.2271 – 0.2280)	0.0014	0.2261
Support vector classifier	0.2233 (0.2229 – 0.2238)	0.0012	0.2221
Random forest	0.2231 (0.2227 – 0.2235)	0.0020	0.2211

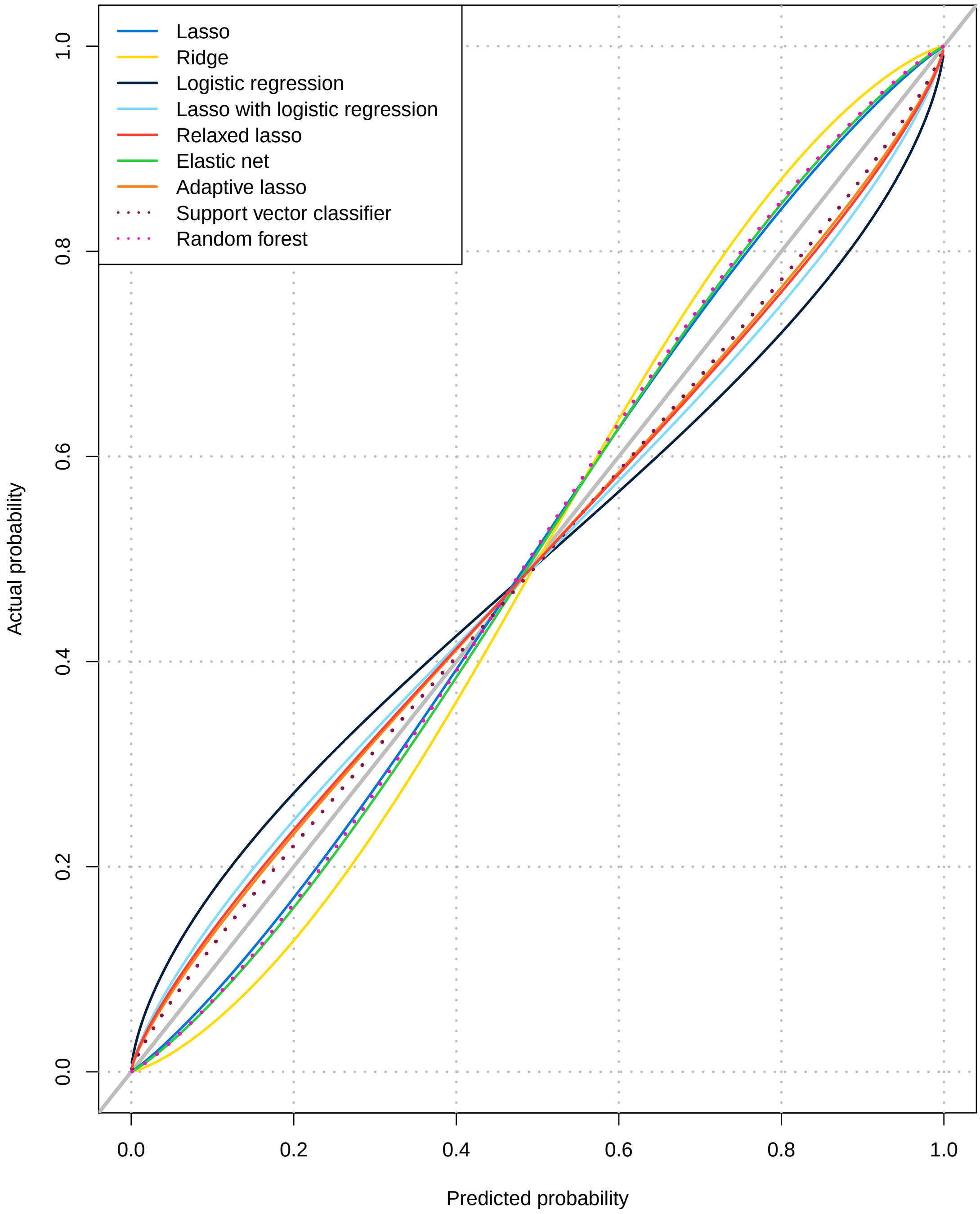

Figure 8: Calibration plot of the models estimated in years 2012-2013 and validated in 2015

Table 24: Area under the curve, sensitivity and specificity of the models estimated in years 2012-2013 and validated in 2016

	AUC (95% conf. int.)	Sensitivity	Specificity
Logistic regression	0.6760 (0.6526 – 0.6993)	0.6505	0.6207
Lasso	0.6774 (0.6541 – 0.7007)	0.6693	0.6037
Ridge	0.6940 (0.6711 – 0.7169)	0.6467	0.6458
Elastic-Net	0.6775 (0.6542 – 0.7007)	0.6726	0.5994
Relaxed lasso	0.6922 (0.6690 – 0.7148)	0.6689	0.6202
Lasso+Log. reg.	0.6726 (0.6492 – 0.6960)	0.6466	0.6179
Adaptive lasso	0.6531 (0.6293 – 0.6769)	0.6722	0.5689
Support vector classifier	0.6737 (0.6503 – 0.6971)	0.6496	0.6173
Random forest	0.6800 (0.6568 – 0.7033)	0.6854	0.5925

Table 25: Brier score decomposition of the models estimated in years 2012-2013 and validated in 2016

	Brier score (95% conf. int.)	Calibration	Refinement
Logistic regression	0.2281 (0.2277 – 0.2286)	0.0044	0.2237
Lasso	0.2250 (0.2245 – 0.2254)	0.0017	0.2233
Ridge	0.2204 (0.2199 – 0.2209)	0.0025	0.2179
Elastic-Net	0.2253 (0.2248 – 0.2258)	0.0020	0.2233
Relaxed lasso	0.2204 (0.2199 – 0.2208)	0.0020	0.2184
Lasso+Log. reg.	0.2290 (0.2286 – 0.2295)	0.0041	0.2249
Adaptive lasso	0.2346 (0.2340 – 0.2350)	0.0028	0.2318
Support vector classifier	0.2271 (0.2266 – 0.2276)	0.0026	0.2245
Random forest	0.2243 (0.2238 – 0.2248)	0.0019	0.2224

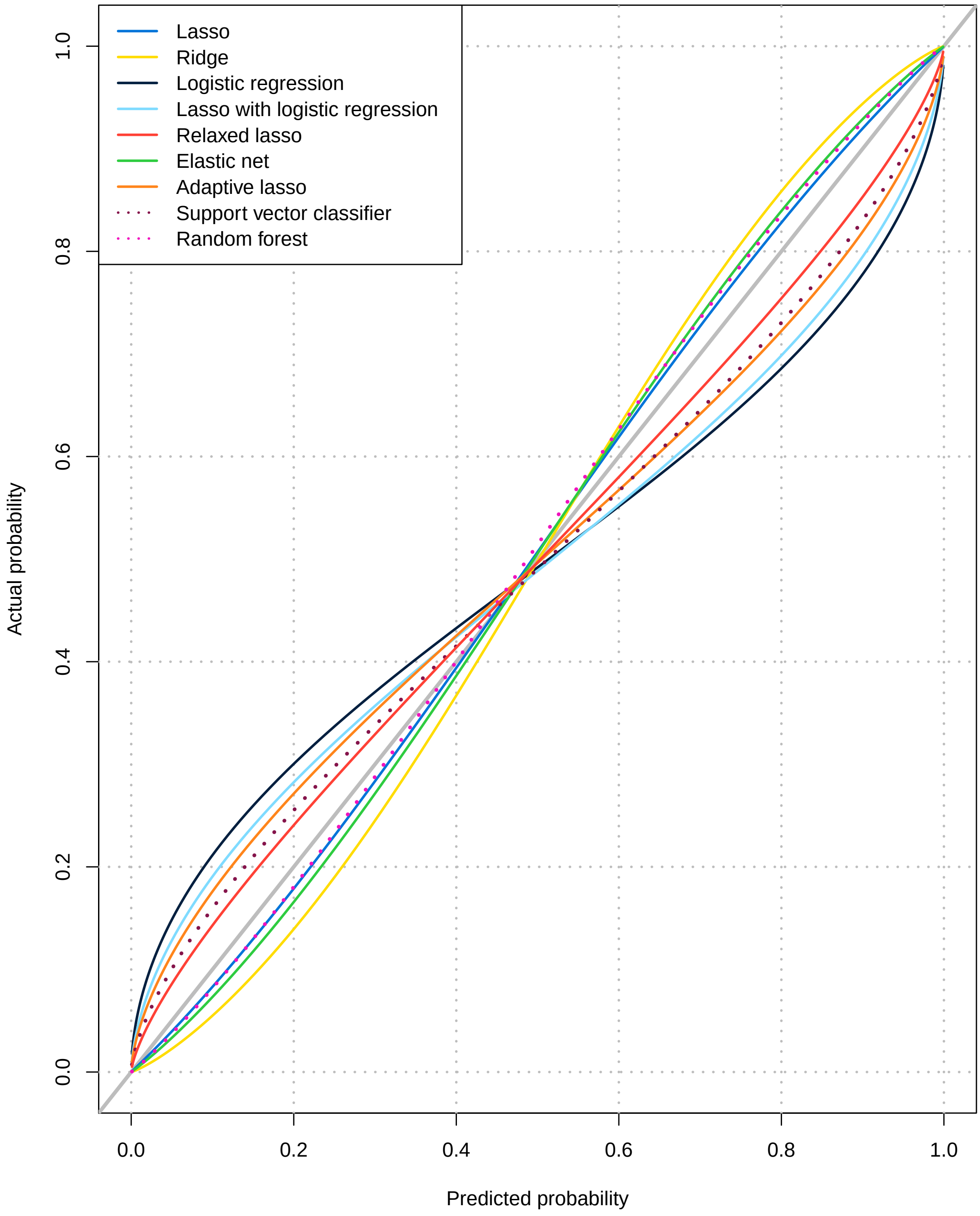

Figure 9: Calibration plot of the models estimated in years 2012-2013 and validated in 2016

References

Akerlof G. (1970) *The market for lemons: quality uncertainty and the market mechanism*, Quarterly Journal of Economics **84**, 488-500

Albanesi S., De Giorgi G. and Nosal J. (2017) *Credit growth and the financial crisis: a new narrative* CEPR Discussion Paper No. DP12230

Beer R., Ionescu F. and Li G. (2018) *Are income and credit scores highly correlated?*, FEDS Notes. Washington: Board of Governors of the Federal Reserve System

Boser B.E., Guyon I.M. and Vapnik, V.N. (1992) *A training algorithm for optimal margin classifiers*, Proceedings of the fifth annual workshop on computational learning theory, 144-152

Breiman L. (1996) *Bagging predictors*, Machine Learning **24**, 123-140

Breiman L. (2001) *Random forests*, Machine Learning **45**, 5-32

Brier G.W. (1950) *Verification of forecasts expressed in terms of probability*, Monthly Weather Review **78**, 1-3

Le Cessie S. and Van Houwelingen J.C. (1992) *Ridge estimators in logisitic regression*, Applied Statistics **41**(1), 191-201

Chen H. and Xiang Y. (2017) *The Study of Credit Scoring Model Based on Group Lasso*, Information Technology and Quantitative Management **122**, 677-684

Cortes C. and Vapnik V.N. (1995) *Support-vector networks*, Machine Learning **20**(3), 273-297

Durkin T.A. (2000) *Credit Cards: Use and Consumer Attitudes, 1970–2000*, Federal Reserve Bulletin, 623-634

Efron B., Hastie T., Johnstone I. and Tibshirani R. (2004) *Least angle regression*, Annals of Statistics **32**(2), 407-499

Emekter R., Yanbin T., Jirasakuldech B. and Lu M. (2015) *Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending*, Applied Economics **47**(1), 54-70

Fan J. and Li R. (2001) *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association **96**, 1348-1360

Friedman J., Hastie T. and Tibshirani R. (2010) *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Journal of Statistical Software **33**(1), 1-22

Haibo H. and Garcia E. A. (2009) *Learning from imbalanced data*, IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263-1284

Harrell F.E. (2015) *Regression modeling strategies: with applications to linear models, logistic and ordinal regression and survival analysis*, Springer Series in Statistics (Second edition)

Hastie T., Tibshirani R. and Friedman J. (2009) *The elements of statistical learning: data mining, inference and prediction*, Springer Series in Statistics (Second edition)

Hastie T., Tibshirani R. and Tibshirani R.J. (2017) *Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso*, ArXiv pre-prints

Hastie T., Tibshirani R and Wainwright M. (2015) *Statistics learning with sparisty: the lasso and generalizations*, Monographs on statistics and Applied Probability, CRC Press

Härdle W., Moro R.A. and Schäfer D. (2005) *Predicting bankruptcy with support vector machines*, Statistical Tools for Finance and Insurance, Springer Verlag

Härdle W., Moro R.A. and Schäfer D. (2007) *Estimating probabilities of default with support vector machines*, SFB **649** Discussion Paper 2007-035

Hoerl A.E. and Kennard R.W. (1970) *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics **12**(1), 55-67

Iyer R, Khwaja A. I., Luttmer E. F. P. and Shue K (2015) *Screening peers softly: inferring the quality of small borrowers*, Management Science **62**, 1554–1577

James G., Witten D., Hastie T. and Tibshirani R (2015) *An introduction to statistical learning with applications in R*, Springer Texts in Statistics

Lessmann S., Baesens B., Seow HV. and Thomas L.C. (2015) *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*, European Journal of Operational Research **247**, 124-136

Leung M., Daouk H. and Chen A (2000) *Forecasting stock indices: a comparison of classification and level estimation models*, International Journal of Forecasting **16**(2),

Lobo J.M., Jimenez-Valverde A. and Real R. (2007) *AUC: a misleading measure of the performance of predictive distribution models*, Global Ecology and Biogeaography **17**(2), 145-151

Makowski P. (1985) *Credit scoring branches out*, The Credit World **75**, 30-37

Martin D. (1977) *Early warning of bank failure: A logit regression approach*, Journal of Banking & Finance **1**, 249-276

McCullagh P. and Nelder J.A. (1989) *Generalized linear models*, Monographs on statistics and Applied Probability, Chapman and Hall (Second edition)

Medema L., Koning R.H. and Lensink R. (2009) *A practical approach to validating a PD model*, Journal of Banking & Finance **33**, 701-708

Meinshausen N. (2007) *Relaxed lasso*, Computational Statistics & Data Analysis **52**, 374-393

Meinshausen N. and Bühlmann P. (2006) *High-dimensional graphs and variable selection with the lasso*, The Annals of Statistics **34**(3), 1436-1462

Murphy A.H. (1973) *A new vector partition of the probability score*, Journal of Applied Meteorology **12**, 595-600

Peterson J. (2001) *The policy relevance of institutional economics*, Journal of Economic Issues **35**(1), 173-183

Platt J. (2000) *Probabilities for support vector machines*, In A. Smola, P. Bartlett, B Schölkopf, & D. Schuurmans (Eds.). *Advances in large margin classifiers.* Cambridge, MA: MIT Press

Stiglitz J.E. and Weiss A. (1981) *Credit Rationing in Markets with Imperfect Information*, The American Economic Review **71**(3), 393-410

Tibshirani R. (1996) *Regression shrinkage and selection via lasso*, Journal of the Royal Statistical Society, Series B (Methodological) **58**, 267-288

Wang G., Hao J., Ma J. and Jiang H. (2011) *A comparative assessment of ensemble learning for credit scoring*, Expert Systems with Applications **38**, 223-230

West D. (2000) *Neural network credit scoring models*, Computers & Operations Research **27**, 1131-1152

Wiginton J.C. (1980) *A note on the comparison of logit and discriminant models of consumer credit behavior*, Journal of Financial and Quantitative Analysis **15**(3), 757-770

Wilks D.S. (2006) *Statistical methods in the atmospheric sciences*, Academic Press (Second edition)

Zhang Y., Jia H., Diao Y., Hai M. and Li H. (2016) *Research on Credit Scoring by fusing social media information in Online Peer-to-Peer Lending* Information Technology and Quantitative Management **91**, 168-174

Zhang X., Wu Y., Wang L. and Li R. (2015) *Variable selection for support vector machines in moderately high dimensions*, Journal of the Royal Statistical Society, Series B (Methodological) **78**(1), 53-76

Zhao R. and Yu B. (2006) *On Model Selection Consistency of Lasso*, Journal of Machine Learning Research **7**, 2541-2563

Zou H. (2006) *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association **101**, 1418-1429

Zou H. and Hastie T. (2005) *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society, Series B (Methodological) **67**(2), 301-320