LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Engineering Science
Industrial Engineering and Management

Vitaly Zaytsev

SELECTION AND EVALUATION OF RELEVANT PREDICTORS FOR CREDIT SCORING IN PEER-TO-PEER LENDING WITH RANDOM FOREST BASED METHODS

Supervisor: Professor Mikael Collan

Examiners: Professor Mikael Collan
Post-Doctoral Researcher Mariia Kozlova

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Industrial Engineering and Management
Vitaly Zaytsev

Selection and evaluation of relevant predictors for credit scoring in peer-to-peer lending with random forest based methods

Master's Thesis 2020

97 pages, 19 figures, 3 tables, 8 appendices
Examiners: Professor Mikael Collan
Post-Doctoral Researcher Mariia Kozlova
Keywords: Credit scoring, feature selection, machine learning, peer-to-peer lending, random forest, variable importance

This study focuses on identifying the important predictor variables for credit scoring using an empirical case of peer-to-peer (P2P) lending data. It contributes to other existing papers that also studied the relevance of predictors in P2P lending by incorporating a larger collection of datasets coming from two European and two U.S.-based P2P platforms and providing an aggregated overview of variables consistently found as important. The variable selection is performed with wrapper methods based on underlying random forest classification models, which predict the loan outcome either as defaulted or repaid. Since the underlying model was able to achieve a reasonable prediction accuracy, it became possible to further analyze the relevance of variables that contributed to making accurate predictions by comparing their respective relative importance metrics. The robustness of findings with respect to identified variables is further improved by combining the results from several methods. In addition to the actual variable selection results and extracted importance metrics, the most important variables are further assessed with regards to their estimated marginal effect on loan outcome. The results are verified for their conformity with other similar papers, and their implications are analyzed from different aspects relevant for investors and those in charge of operating P2P lending platforms.

ACKNOWLEDGEMENTS

I am very grateful to my supervisor, Mikael Collan, for providing valuable feedback and guiding through the overall thesis writing process. Those efforts encouraged me to focus on covering the relevant aspects of applying machine learning methodology in the context of credit scoring and also to ensure the smooth outlining of all contents. In addition, I would like to thank Mariia Kozlova for clarifying some questions with respect to credit scoring practices and providing helpful ideas on approaching the chosen topic in the initial stage of the writing process.

Lappeenranta, June 2020

Vitaly Zaytsev

TABLE OF CONTENTS

1 INTRODUCTION 4

1.1 BACKGROUND 4

1.2 RESEARCH GOALS 7

1.3 RESEARCH METHODS 9

1.4 STRUCTURE OF THE THESIS 10

2 FEATURE SELECTION 11

2.1 CORE PRINCIPLES AND MOTIVATIONS FOR PERFORMING FEATURE SELECTION 11

2.2 SPECIFICS OF FEATURE SELECTION WITH RANDOM FORESTS 13

2.3 STEPWISE SELECTION METHODS BASED ON VARIABLE IMPORTANCE 15

2.4 STATISTICAL HYPOTHESIS TESTING METHODS BASED ON VARIABLE IMPORTANCE 17

2.5 GENERIC BACKWARD SELECTION WRAPPER BASED ON PREDICTION ACCURACY 21

3 DECISION TREE BASED CLASSIFIERS 24

3.1 FUNDAMENTALS AND ADVANTAGES OF DECISION TREE BASED CLASSIFIERS 24

3.2 RANDOM FORESTS AS AN IMPROVEMENT OVER DECISION TREE CLASSIFIERS 27

3.3 INTERPRETATION OF RANDOM FORESTS 28

4 LITERATURE REVIEW 32

4.1 NOTES ON LITERATURE SEARCH PROCESS 32

4.2 RANDOM FORESTS AS A TOOL FOR PREDICTIONS 33

4.3 FEATURE SELECTION WITH RANDOM FORESTS 34

4.4 RELEVANT PREDICTORS FOR CREDIT SCORING IN P2P LENDING 37

5 DETAILS ON DATASETS, PRE-PROCESSING AND IMPLEMENTATION 39

5.1 GENERAL OVERVIEW OF USED DATASETS 39

5.2 BONDORA 40

5.3 LENDINGCLUB 42

5.4 PROSPER 44

5.5 ZOPA 45

5.6 SPECIFICS OF FEATURE SELECTION PROCEDURE AND CLASSIFIER CONFIGURATION 45

6 RESULTS OF VARIABLE SELECTION PROCESS 50

6.1 General overview of classification accuracy and variable selection results 50

6.2 Bondora 51

6.3 LendingClub 54

6.4 Prosper 58

6.5 Zopa 61

7 CONCLUSIONS 63

7.1 Summary of the results and conclusions 63

7.2 Limitations 67

7.3 Suggestions for further research 70

REFERENCES 72

APPENDICES 76

APPENDIX 1. Variable importance and selection results by four methods (Bondora data) 76

APPENDIX 2. Variable importance and selection results by four methods (LendingClub data) 79

APPENDIX 3. Variable importance and selection results by four methods (Prosper data) 83

APPENDIX 4. Variable importance and selection results by four methods (Zopa data) 86

APPENDIX 5. Data dictionary for Bondora dataset 87

APPENDIX 6. Data dictionary for LendingClub dataset 89

APPENDIX 7. Data dictionary for Prosper dataset 94

APPENDIX 8. Data dictionary for Zopa dataset 97

LIST OF SYMBOLS AND ABBREVIATIONS

CV	Cross-validation
GBS	Greedy backward selection
OOB	Out-of-bag
P2P	Peer-to-peer

1 INTRODUCTION

1.1 Background

Credit scoring is a task that is common at many financial institutions, which deal with a problem of evaluating the suitability of potential borrowers in order to qualify for a loan. This study considers the problem of approaching the credit scoring process in the context of peer-to-peer (P2P) lending. The P2P services are known as online platforms that allow individual persons to apply for loans financed by other individuals. On such platforms, the individual borrowers and lenders can connect with each other for funding the loan. Then, a portion of the monetary transaction is taken as a fee by the platform operator. These services are largely automated and use modern information technologies. They greatly facilitate the process of screening the applicants, handling transactions, and servicing the loans.

P2P lending has been considered as an increasingly popular service that emerged as an online credit alternative to the intermediated financial services industry. They operate largely on the basis of trust between people that have only met in cyberspace. Based on information shared by prospective borrowers, the lenders can decide whether to invest in a particular loan. (Galloway, 2009)

There are several major reasons that contributed to the success of P2P lending platforms. First of all, unlike traditional financial institutions, such platforms have lower overhead costs and risk exposure, which allow keeping servicing fees on a relatively low level. As a result, they are able to offer a remarkably high return on investment for the lenders. In addition, the borrowers that are typically considered as undesirable by traditional financial institutions have been granted much better access to loan financing. Finally, P2P lending is often associated with increased transparency, flexibility, and a more convenient service for customers. (Milne and Parboteeah, 2016)

Nowadays, many P2P lending platforms publish data on issued loans, which can be valuable for potential investors. Some of those only publish high-level aggregated statistics on loan performance, while others provide the detailed lists of all previously issued loans in an anonymized format. This data provides great opportunities for developing appropriate investment strategies and researching P2P lending processes in general. A number of such

papers focused on P2P lending are considered in greater detail later in this thesis. As an example, the recent paper by Croux et al. (2020) investigated the factors connected with the risk of default in the dataset from the popular U.S.-based P2P platform LendingClub and reported the following key facts illustrating the loans issued during the years 2007-2018:

- Average loan amount: $14 370 (overall range from $500 to $40 000)
- Average loan maturity: 41,8 months (possible terms include 3 and 5 years)
- Average annual interest rate: 13,37% (overall range between 5,31% and 30,99%)
- Loan default rate: 20%.

For the loans issued in the U.S., the most popular consumer loan purposes include debt consolidation, credit card repayment, and home improvement. On the other hand, the small business owners also sometimes apply for online loans, and their stated purposes typically include buying inventory, acquiring new business equipment, and working capital. The business loans are issued for a rather short term and are also quite small compared to those provided by traditional financial institutions. (U.S. Department of the Treasury, 2016)

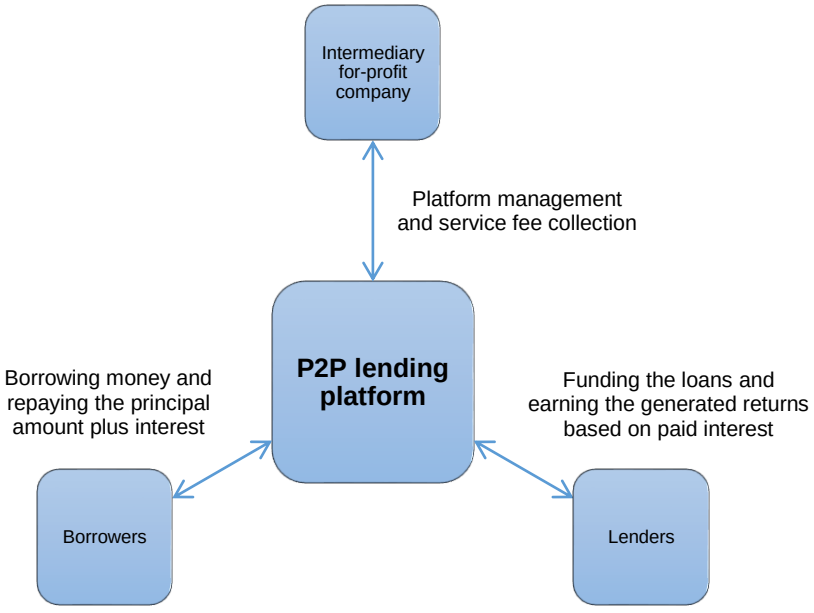

Figure 1. The summarized operation principles of common P2P lending platforms.

In the same way as for lenders at other financial organizations, the crucial aspect of financing a loan is to assess the potential risks for a specific loan application and determine the creditworthiness of a borrower. Traditionally, this task has been referred to as credit scoring, which is represented by a set of methodologies for determining borrower's risk for the consistent and objective decision-making based on empirically derived information (Siddiqi, 2005).

Selecting an appropriate statistical analysis framework that aggregates the information on the available predictors for producing an estimate of risk is often not enough. The other major issue encountered when approaching the credit scoring procedure is the identification of the best subset of the available predictors to be used in the assessment of creditworthiness. The background information on the borrower is typically comprised of a set of various indicators, such as financial and demographical. With today's wealth of data and modern information systems, it is possible to collect hundreds or even thousands of available predictor variables that have the potential to be good estimators for a dependent variable in a given classification setting. However, it is unlikely that all of them would be equally important for achieving a high prediction accuracy. This issue can be tackled with feature selection, a methodology for selecting a subset of features that appear to be the most representative for a particular application case or dataset. This process assumes retaining classification accuracy close to the highest possible level with a restricted number of variables used in the model. The terms "feature" and "variable" are used interchangeably in this thesis, along with some other similar terms (e.g. "factors", "predictors", and "determinants").

In the specific case of P2P lending, knowledge on important predictors of financial risks is valuable not only for the lenders that aim to implement credit scoring for aiding their loan evaluation process, but also for those who establish and operate such platforms. Reducing the amount of information (which sometimes turns out to be irrelevant) required for loan applications is likely to encourage more people to use the lending platform services. Moreover, the acquired information on relative importance of variables and the specific impact on the risk of default can help policymakers and researchers at various financial and governmental institutions. As an example, they would be able to understand better why

certain categories of borrowers carry an increased risk of becoming delinquent and what can be done in order to help such individuals to get better access to financial aid.

Despite the strong advantages of performing feature selection, this task is not always straightforward. A wide range of different feature selection methods is available nowadays, and it becomes necessary to evaluate their suitability and efficiency in a systematic manner. Some of those methods are based entirely on measuring certain characteristics of the original data, while the others are only concerned with the improvement in prediction accuracy provided by features. In the same way as for choosing the optimal classifier, the exact choice of a feature selection method is an application-dependent task. It can be affected both by problems in data and the definition of a final goal, where the focus might shift from improving classification accuracy to gathering the insights or conducting exploratory data analysis. It might appear rather easy to identify if a specific type of feature selection methods is undoubtedly more suited for a concrete task, but then the choice from several highly similar methods can become a notable issue.

1.2 Research goals

The aim of this study is to apply and demonstrate an approach of automated feature selection methods to obtain the information on relevant predictors using an empirical case of P2P lending data acquired from several publicly available datasets. It takes a closer look at a number of feature selection methods that are argued to be especially well suited when the final goal is to produce the most accurate predictions with a specific machine learning method belonging to the category of decision tree based classification models. The study attempts to establish additional evidence of certain indicators being especially useful for estimating the risk of borrower's default in the context of credit scoring for P2P lending. The importance of individual variables available in studied datasets is discussed in detail with respect to their contribution to accurate predictions of loan outcome and their estimated impact on the probability of default. The obtained results are expected to provide an improved understanding of factors potentially responsible for the loan defaults and their actual impact on risk of default. Such information could appear to be especially useful for P2P lending platform operators and investors interested in further improvement of their data collection and predictive model optimization.

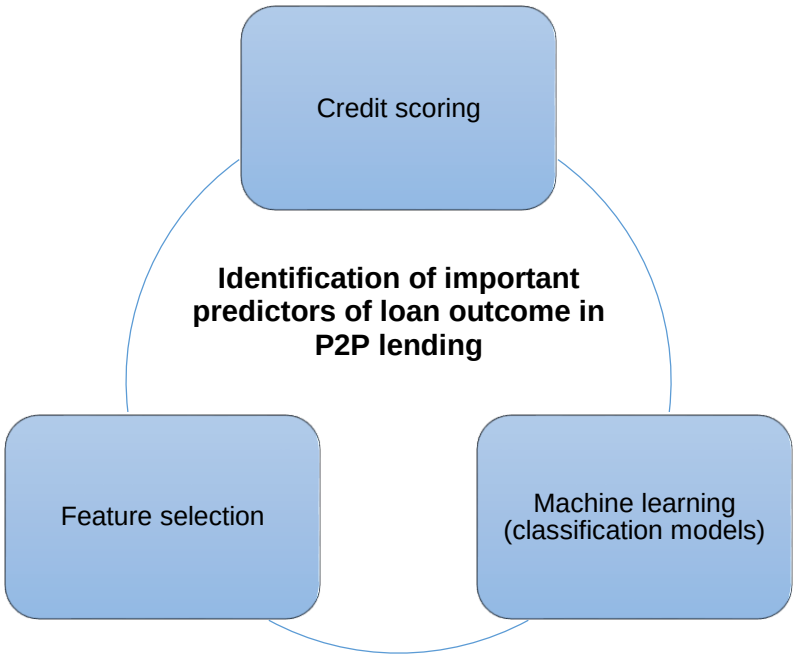

Figure 2. The interconnection of various research areas in an empirical study conducted in this thesis.

The concrete research questions can be defined as follows:

1. Which variables tend to be the most and least important for estimating the risk of default for loan applicants in the context of the P2P lending classification task?

2. What is the direction of influence on the probability of default among numerical variables and separate levels in categorical variables?

3. Are there notable discrepancies among subsets of variables estimated as important predictors when comparing results across several P2P lending datasets?

4. Does the information on variable importance acquired through feature selection correspond to the results from previous studies employing more traditional statistical approaches for determining variable significance?

1.3 Research methods

This study considers the problem of selecting the relevant predictor variables in a classification setting. The classification task here is to predict whether a specific borrower is likely to default on their financial obligations based on available information that describes the actual loan and borrower, including a set of financial and demographical variables collected through the loan application process.

In order to make this study more focused, the decision has been made to concentrate on the implementation of automated feature selection methods (wrappers) used in conjunction with decision tree ensemble classifier (random forest) in an empirical case study.

The features which appeared to be important for making accurate classification and their impact on the risk of default are discussed from the viewpoint of their relevance for credit scoring using built-in importance metrics of random forests and results of feature selection. Some additional methods that enable deeper interpretation of the model and specific impact of predictor variables on the target variable, such as partial dependence plots, are also used in this work.

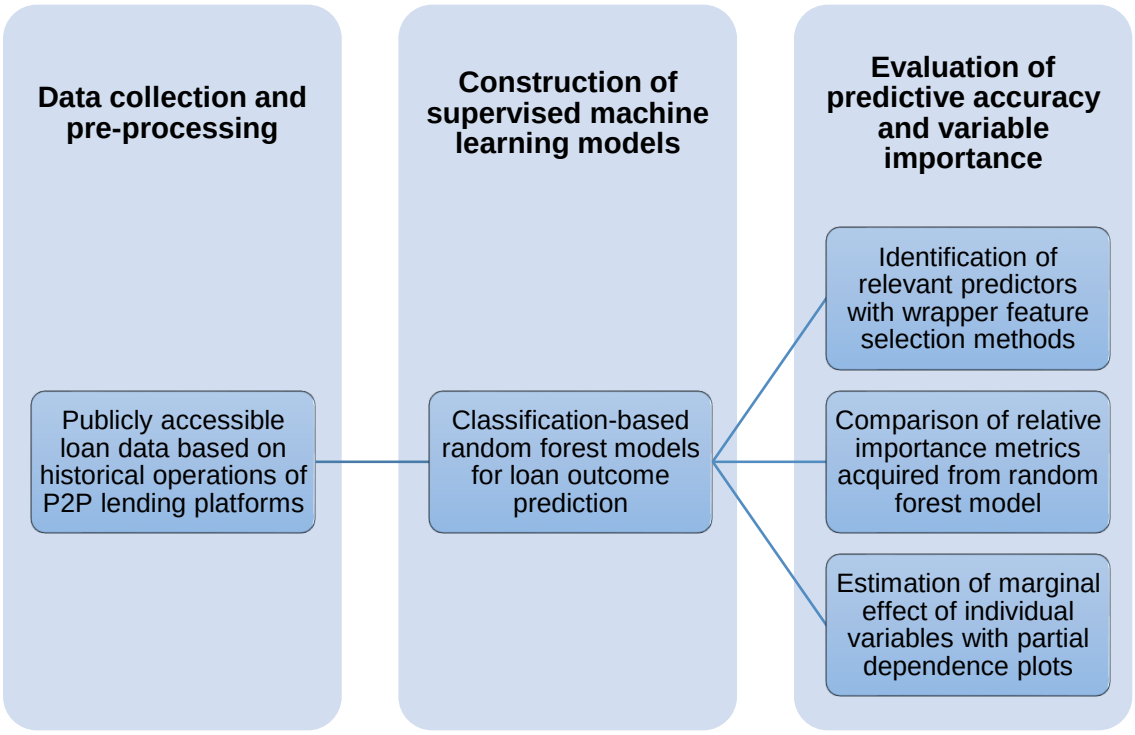

Figure 3. The summarized steps of research process implemented in this thesis.

For the empirical study part, the decision has been made to use four different datasets comprised of loan application historical records from the popular P2P lending platforms. Those are publicly available datasets based on historical operations of these P2P lending services that were accumulated over the years and made accessible on the web. The records contain anonymized information on the cases when a loan was granted to a certain individual. Those consist of detailed loan characteristics (such as amount, interest rate, term, purpose, and status) and a large number of indicator variables linked to the applicant (financial and demographical). The ranking of variables is then aggregated over all datasets, and the final results are also compared with findings from other papers. It should be noted that the terms "samples", "records", and "cases" will have a similar meaning that corresponds to the loan applications stored in the studied datasets.

1.4 Structure of the thesis

Section 2 describes the principles and motivations for utilizing the feature selection methods in general and also specifics of those used in this study. Section 3 provides the theoretical foundations of decision trees and random forest classifiers while also discussing their advantages and interpretation capabilities. Section 4 is focused on literature review related to the practical application of random forests, feature selection, and previously identified important predictors for credit scoring in P2P lending. Section 5 provides an in-depth description of datasets used for the empirical part and some additional notes of the feature selection and evaluation procedure. Section 6 presents the results acquired from feature selection methods as well as more specific information on the actual impact of most relevant predictors on the risk of default. Section 7 aggregates the results acquired from the previous section while providing the final conclusions and limitations of the study, as well as suggestions for further research.

2 FEATURE SELECTION

2.1 Core principles and motivations for performing feature selection

Feature selection attempts to identify a subset of the original features to maintain a similar prediction accuracy (or even improve it) compared to the original features. When dealing with high-dimensional data, it is often the case that not all variables are actually contributing to improving a model prediction accuracy, so the inclusion of irrelevant variables might lead to unnecessary complexity of the predictive model. Limiting the number of variables provides significant benefits from the model perspective, such as reducing computational complexity, avoiding the curse of dimensionality, improving prediction performance, and making model more interpretable (Guyon and Elisseeff, 2003). However, there are also some less obvious reasons why one would consider feature selection, which might be closely related to the actual application case. Those can include, for instance, savings in costs and labor efforts associated with collection of data or performing further research on variable importance to establish potential causality between predictor and target variables in the studied phenomena. The automated methods for performing feature selection can be especially useful if there is a lack of domain knowledge for manual selection of features.

There are numerous existing feature selection methods that have different properties and their specific advantages. According to Guyon and Elisseeff (2003), it is possible to identify the following main categories of such methods:

1. Wrapper methods
2. Filter methods
3. Embedded methods.

Wrapper methods measure the performance of a machine learning model as a proxy for identifying a subset of variables useful for predictions. They allow utilizing the underlying machine learning model as a "black box" for identifying the subset of relevant features. The famous examples of such methods include forward and backward selection, which iteratively add or remove variables based on how they contributed to improving the goodness-of-fit statistics (measured on train or validation datasets). Such approaches are often considered "greedy", as they typically require multiple runs of machine learning algorithms. On the

other hand, they can identify the features that achieve the maximum prediction accuracy while taking into consideration the combined power of all available predictors.

Filter methods are used independently and before applying any machine learning algorithms for measuring certain properties of training data (e.g. generic class separability measures or measures of association between individual predictor variables and target variable). They are very fast to compute and can be used as a pre-processing step for basic identification of relevant features before applying any classification or regression algorithms. However, they are often not able to consider the possible interaction effects between variables, which could potentially improve the predictive accuracy. In addition, they might impose restrictions on the possible types of predictor and target variables to be used with a specific method. Some common examples of such methods for ranking the features include mutual information criteria, statistical hypothesis tests, and correlation coefficients.

Finally, embedded methods are an inherent part of the specific machine learning algorithms and applied during the actual model building process. They are somewhat similar to wrapper methods in the sense that they are linked with predictive model classification accuracy, thus sharing the aforementioned advantages of wrappers. Unlike wrapper methods, they are more efficient as they do not require a separate validation set for predictive accuracy assessment or model retraining. Instead, they typically perform the internal feature selection on training data as a way to achieve the optimization of a certain objective function that attempts to obtain the best goodness-of-fit metrics while penalizing the inclusion of irrelevant variables into the model. The most prominent example is regularization by shrinking the coefficients of irrelevant variables in a predictive model towards zero (e.g. lasso or ridge regression). The decision trees also perform internal feature selection by considering various predictors at each node split and prioritizing predictors resulting in lowest node impurity.

Although there are numerous feature selection methods of various types available nowadays, there were certain reasons to favor wrapper methods in this work:

1) Due to the classification accuracy (or similar metric) being used as a proxy for evaluating the feature relevance, the whole focus in the task of feature selection is shifted from measuring generic correlation scores and class separability metrics

(which sometimes impose unrealistic statistical assumptions on the underlying distribution in data) to evaluating their direct impact on achieving best possible predictive accuracy.

2) Since the methods used in this work inherit the key properties of the underlying classification model, it becomes possible to take into account various aspects considered difficult for other approaches to feature selection, such as a non-linear relationship between a predictor and target variable, efficient handling of both quantitative and qualitative variables, and also interaction effects (combined power) of several predictors.

The core principles behind the employed feature selection methods can be described as follows: use the dataset containing all predictors, fit a model with those predictors, measure the importance of individual predictors with respect to model classification capabilities and decide on selecting the top-performing predictors. The final conclusion regarding the importance of the different variables is made by comparing the results acquired from several closely-related datasets of the same application area. Despite that specific collection of available variables differs from one dataset to another, it is still possible to observe if there are some similar patterns present in all of them.

Obviously, such a wrapper-based feature selection methodology will only make sense in case if the base classifier is able to achieve a reasonable prediction accuracy. In addition, it is essential to note that the resulting subset of variables is highly dependent on the underlying classification method. As an example, the variables that have non-linear and interaction effects, which were selected with random forests, might appear to be useless in a logistic regression context.

2.2 Specifics of feature selection with random forests

Although there is an embedded feature selection process that occurs during the construction of decision tree classifier, the feature selection methods applied in this study ultimately rely on prediction accuracy or variable importance metrics acquired from a complete classifier for aiding further decision-making on important variables. Therefore, they can be regarded as wrapper methods. The main idea behind them is to measure the changes in prediction

accuracy or importance metrics after altering predictors or target variable in a certain way. They avoid an exhaustive search of best subset of features that is considered as a truly superior but also an unrealistic approach with a large number of predictors by applying certain heuristics for identifying relevant variables via stepwise selection procedures or statistical tests over a large number of fitted random forests.

In principle, one could simply use the variable importance metrics to manually select a pre-determined fraction of top-performing predictors based on information from single random forest fit, but there is a number of caveats associated with that (Janitza et al., 2018). The most prominent one is that there is no straightforward method to check if variables with small (non-negative) importance are truly relevant or simply ended up there by chance. Having a large set of variables with low importance might seem to overcomplicate the model, but their combined predictive power provides a good reason to identify them all. In addition, the importance metrics alone cannot be used to distinguish the redundant variables in those that were indicated as important, but this is a lot more difficult problem than simply finding all relevant predictors.

In order to be considered redundant from the perspective of used feature selection methods, the variables do not necessarily need to have a perfect correlation or multicollinearity, which is generally undesirable with any kind of predictive model. Instead, the redundancy here means that removing all but one of such highly correlated variables would only lead to a slightly suboptimal model from the prediction error perspective. However, the authors of the feature selection methods considered in this thesis do not provide a more concrete definition of redundancy, so in terms of this thesis, the redundancy will simply mean a high level of correlation that makes adding several correlated variables unjustified with respect to the supposed costs versus benefits perspective. Alternatively, it could be explained as follows: the amount of predictive error increase after removal of additional redundant (but important) variables is of a similar magnitude as the one resulting from the removal of non-correlated weakly important variables.

Even if one would be determined to use random forests for feature selection, there is still a problem of selecting the best performing approach specific to that type of classifier. Over the years, multiple strategies have been proposed by various authors for conducting variable

selection process with random forests, which attempt to solve the problems illustrated above with regards to identifying the subset of all relevant variables in a systematic way, such as backward selection based on variable importance metrics (Díaz-Uriarte and Alvarez de Andrés, 2006), comparison of the importance coming from original variables with their randomly shuffled versions added into the same model (Kursa and Rudnicki, 2010) and detection of statistically significant difference in importance metrics after permutation of the target variable (Altmann et al., 2010). Several of the most widely applied methods have been selected for use in this thesis. Although there are various metrics available for deciding on important variables, all these methods utilize the permutation importance coming from random forest models. The details of the methods used in this thesis are discussed in the following subsections (allocated according to the categories of methods).

2.3 Stepwise selection methods based on variable importance

The general idea of stepwise selection is not specific to random forests. A similar approach is widely applied e.g. in linear regression setting by iteratively removing the least significant variables according to certain criteria. However, the methods described below utilize the metrics of importance that are found specifically in random forest models.

One of the earliest methods for using random forests as a tool for selecting features has been suggested by Díaz-Uriarte and Alvarez de Andrés (2006) and implemented in R package "varSelRF". In their paper, they introduced a backward elimination approach that involves removing a fraction of variables iteratively based on the least permutation importance so that the best performing model can be selected. They note that this method is more suitable for identifying a smaller subset of variables excluding redundant predictors that provide a negligible increase in predictive accuracy.

The detailed procedure for this method can be explained as follows:

1. Fit the full random forest model containing all available predictors.

2. Discard the certain number of predictors that have the smallest unscaled permutation importance (the proposed ratio is 0,2 of all predictors at each iteration of the procedure).

3. Re-fit the model and continue performing the previous step until no more variables can be removed.

4. Select the variables from a model that resulted in the lowest error on unseen out-of-bag (OOB) samples.

This method was not used in this thesis, since the available R package utilizes a different implementation of random forests (package "randomForest") than the one used by other methods (package "ranger"), so the comparison of results could appear inappropriate between those. In addition, the generic wrapper backward selection introduced in subsection 2.5 seemed to be a very similar method that also achieves the goal of selecting the non-redundant variables.

Another somewhat similar method introduced by Genuer et al. (2010) that is implemented in R package "VSURF" requires a more complex two-step procedure, where the first step involves a cutoff-based selection of top-performing variables that exceed the minimum observed importance threshold and the other consists of a stepwise ascending variable introduction.

The method applies the following algorithm for selecting the features in the first preliminary ranking step:

1. Fit random forest models with a full set of predictors repeatedly (the suggested number is 50) and average the resulting unscaled permutation variable importance.

2. Rank the variables according to averaged importance in descending order.

3. Discard the variables with the smallest importance according to certain heuristics (explained below).

4. Order the remaining variables in decreasing order of importance.

The actual rule for deciding on the cutoff value is somewhat sophisticated. The authors claim that the variability of variable importance is larger for relevant variables than for useless ones. Therefore, they identify the minimum expected threshold of variable importance for separating useless variables from others by fitting a pruned CART tree model (where the standard deviation of variable importance is predicted by the acquired rank of variables) and finding the minimum prediction value of standard deviation. Finally, only the variables with variable importance exceeding this threshold are retained for the next steps. The authors argue that this preliminary step is rather conservative and leaves more variables than necessary for the next step.

The second step for the actual variable selection involves the following:

1. Build the random forest models starting with one variable and calculate OOB error by averaging results from repeated runs (the suggested number is 25).

2. Iteratively add new variables according to the importance ranking until all of the previously identified variables are used.

3. Select the variables from the model with fewest predictors that resulted in the OOB error rate within one standard deviation from the smallest observed OOB error.

There is also a third optional step that aims to further reduce the set of predictors in order to remove redundant ones, but it was not considered for use in this study.

2.4 Statistical hypothesis testing methods based on variable importance

In contrast to the methods from the previous category, the methods presented in this subsection are focused on identifying all of the relevant predictors regardless of their possible redundancy.

Kursa and Rudnicki (2010) suggested a notably different method (corresponds to R package "Boruta") that applies a significance testing procedure over variable importance metrics. The key principle of the method consists of adding "shadow" variables to the random forest model by taking the original variables and randomly shuffling their values. The authors of

the method argue that non-zero importance of a "shadow" variable can only be attributed to the random fluctuations, which arise from random shuffling and stochastic nature of the random forest classifier. Therefore, the importance can be verified by applying a statistical significance test that compares the importance of the original variable and highest achieved importance of an artificially created variable. However, many random forests need to be grown in order to acquire statistically valid results.

In order to apply this method, the following procedure needs to be performed:

1. Add the copies ("shadow" versions) of the original predictor variables into a model with all predictors. If there were only a few predictor variables in the first place, this step ensures that at least five copies were added into model.

2. Randomly shuffle the values in the added copies of variables to eliminate their association with the target variable.

3. Run a random forest with an extended set of predictors and collect the scaled permutation importance metrics.

4. Identify the highest obtained importance among "shadow" variables.

5. For all variables with the pending decision on their importance, run a two-sided test of equality with the importance value identified in the previous step.

6. Declare variables that have significantly higher importance than the one identified in step 4 as important and those with significantly lower importance as unimportant.

7. Remove the shadow variables from the model and repeat the whole procedure again until a decision is made on all variables or if the pre-determined limit of random forest runs is exceeded (default number is 100).

Another method selected for use in this study was developed by Altmann et al. (2010), and unlike the previously mentioned method, it applies permutation on the target variable instead

of predictors. This process is repeated many times so that it becomes possible to obtain a distribution of nullified importance measures. Then, the p-values are calculated to estimate the probability of observing the importance equal to or higher than the originally reported importance based on acquired null distribution. The final decision is made by selecting important variables according to the acquired p-values and chosen significance level.

A more detailed description of the process behind selecting features with this method is given below:

1. Permute the values of the target variable and run random forests using all available variables to predict the permuted versions of target variables (50 to 100 permutations are recommended).

2. Collect the chosen variable importance metrics from each run of random forest in the previous step as a vector of nullified importance measures.

3. Fit a probability distribution to the acquired importance metrics according to the specified form of distribution (Gaussian, lognormal, or gamma) or by detecting an appropriate form with Kolmogorov-Smirnov tests (parameters of distribution estimated with maximum likelihood method).

4. Calculate p-values denoting the probability of observing the importance metrics equal or higher to the original (true) variable importance metrics with the given distribution. In case if automatic tests were not able to identify a distribution similar to the three mentioned ones, run a non-parametric estimation of p-values by determining the fraction of nullified importance metrics that were more extreme than the true importance.

5. Decide on important variables based on the chosen significance level (0,05 as a standard choice).

The authors note that this method is efficient with a large number of variables, as the number of permutations could be a lot less than the number of used variables. In addition, such a method could use any of the available importance metrics coming from a predictive model.

An interesting approach has also been recently demonstrated by Janitza et al. (2018) and represents another way of determining important variables (R package "vita"). It assumes that the importance of truly irrelevant predictors varies randomly around zero. However, as mentioned earlier in this thesis, it is impossible to distinguish important variables with very low importance scores from noise variables. Therefore, this method attempts to mirror the negative part of the distribution in order to acquire its unobserved part. The robustness of the method is further improved by calculating the importance using holdout samples instead of OOB samples. Finally, the p-values can be derived for identifying whether a certain variable corresponds to the null distribution of irrelevant variables.

The feature selection with this method can be performed as follows:

1. Divide dataset into two parts of equal size.

2. Create a random forest based on all variables with each set and compute the hold-out variable importance metrics (i.e. not the same OOB importance available by default).

3. Approximate the null distribution for the hold-out variable importance metrics based on observed non-positive variable importance values by creating an empirical cumulative distribution function comprised of three parts (the negative importance scores, importance scores of zero, and negative importance scores multiplied by -1).

4. Calculate p-values corresponding to the original variable importance metrics with respect to the null distribution identified in the previous step.

5. Decide on important variables based on the chosen significance level.

Unfortunately, this method cannot always be applied as it is not possible to reliably estimate the distribution, unless it is known beforehand that the dataset contains a very large number of truly irrelevant variables, so it was not used along with other methods in this work.

2.5 Generic backward selection wrapper based on prediction accuracy

The last presented method out of those used in this thesis is not attributed to any particular author, and since it does not use the random forest importance metrics at all, it did not exactly fall into the previously mentioned categories of random forest based feature selection methods. Instead of variable importance metrics, it relies on plain classification error rates acquired from the random forest model. Therefore, it is a generic wrapper method that could be used in principle with any classifier. However, it still makes use of OOB error that is naturally produced while building random forest and, therefore, does not require having a separate validation set. The heuristics behind this approach can be described as follows:

1. Fit a random forest using a complete set of all available predictors and record the initial baseline OOB error rate.

2. Attempt to remove each predictor (one at a time), re-fit random forest, and record the new error estimates of the reduced models.

3. Find the model that resulted in the lowest classification error and remove the variable associated with that model completely from further consideration.

4. Continue removing variables according to steps 2-3 until the error rate of a reduced model starts to be considerably larger than the one of a full model. The actual stopping criterion is fulfilled when the error of a reduced model becomes larger than the maximum observed error rate of a model with all predictors plus one standard deviation (acquired from distribution of error rates from 10 preliminary runs of full random forest model). The threshold was chosen in a way to account for the possible fluctuations of error rates between different runs of random forest and allow the exclusion of some weakly important variables (that provide negligible improvement in predictive accuracy) before stopping the process.

5. Finally, the last removed variable that resulted in stopping criteria becoming fulfilled is returned back to the model, and the resulting set of variables is considered as important.

This method is further referred to as greedy backward selection. It could appear to be computationally expensive when compared to others, although it is not a fully exhaustive search. On the other hand, in contrast to other feature selection methods presented above, it avoids dealing with variable importance. The main rationale for including this method into analysis was that unlike the other three methods, it allows discovering a smaller set of non-redundant variables. Although it is crucial to learn about all the relevant variables when one investigates the reasons of defaults, from the practical perspective, it could be wiser to use the model with the smallest set of predictors to avoid dealing with a complex model. Despite that it was possible to run a backward selection method using variable importance (e.g. as in method "varSelRF"), it seemed reasonable to complement the other available methods with the one that avoids dealing with variable importance. The more elaborate explanation for such a decision is given below.

There is evidence that variables, which are highly relevant but also highly correlated with each other, might occupy the top-ranking position in the variable importance list (Archer and Kimes, 2008). This means that highly important but redundant variables would be removed from the model always approximately at the end of the backward elimination procedure. At the same time, the increase in the error resulting from the removal of some of these variables could be of similar magnitude as for the non-redundant variables with weak importance or even smaller (e.g. for perfectly correlated variables), despite having higher importance according to model. Although such variables would naturally get very high importance in a full model, if one would attempt to remove some of them, their remaining correlated versions could simply compensate for the effect that was provided by the removed ones. Ultimately, the redundant variables would have smaller chances of ending up in the final reduced model, which might be a desirable option from the variable selection perspective. With the described wrapper method, all variables would be considered for removal with an equal chance at each iteration of variable elimination procedure with respect to the improvement they contribute when compared to an updated model with a reduced set

of variables. This might bring an additional perspective on the variable impact comparing to the importance metrics calculated within a single model.

A similar and less computationally expensive approach could be the reversed version of the abovementioned algorithm, which tries new variables for adding to a model over removing the existing ones. However, the problem with such an approach is that the model cannot take advantage of being initially exposed to all available interactions between predictors, so it becomes a lot more difficult to find such interactions that could potentially boost the prediction performance.

3 DECISION TREE BASED CLASSIFIERS

3.1 Fundamentals and advantages of decision tree based classifiers

Decision trees are some of the widely used machine learning models that are powerful yet simple to understand and implement. Although the single decision trees are often criticized for their weak predictive accuracy, the ensembles of trees can overcome their major drawbacks. Random forests were favored for use in this work as an underlying model for feature selection since they are generally regarded as some of the top-performing machine learning methods currently available. Hastie et al. (2009) consider decision trees as close to the ideal "off-the-shelf" machine learning methods due to the following benefits:

- Relatively fast construction and ease of interpretation.
- Natural ability to handle the mixtures of quantitative and qualitative predictors, without any transformation required (also being invariant to strictly monotone transformations of individual predictors).
- Robustness to outliers and resistance to the inclusion of irrelevant predictors.
- Very little tuning required for achieving good prediction accuracy (when building ensembles of trees).

Another important aspect is that many statistical software packages nowadays provide readily available algorithms for straightforward implementation of decision tree and random forest models. Ultimately, it becomes possible to streamline the experimentation process and focus on the most crucial part of this work, which is the estimation of predictor variable relevance.

The main principle used for obtaining a decision tree model is recursive partitioning of input variable space. This process results in a set of rules being derived from training data, which resemble a tree structure and can be used for classifying new samples. CART models (introduced by Breiman et al. in 1984) became a widely popular method for classification and regression tasks. Despite the simplicity of logic behind building the trees, they are very powerful models that can deal with a wide spectrum of classification problems. Their ability to capture interactions amongst predictors as well as non-linear relationships between target and predictor variables puts them on par with some of the most advanced machine learning

methods available today, such as artificial neural networks. Unlike the traditional statistical methods like logistic regression, decision trees are non-parametric models that do not require making assumptions about underlying distributions in data and naturally able to identify the most suitable representation of the relationship between predictor and target variables.

The basic procedure for obtaining a decision tree in a classification task can be described as follows:

1. Select one of the available predictor variables X_j in training data to create a tree node.

2. Pick a such threshold value s_j that divides the input space for that variable into two regions, which correspond to values below and above the threshold.

3. Try all possible values s_j until the best possible threshold is found (the one that results in lowest node impurity).

4. Repeat steps 1 to 3 for finding the best predictor variable to use in the current node (the one that results in lowest node impurity).

5. Continue growing the tree by forming new nodes with smaller subregions from the existing ones when it leads to a decrease in node impurity (steps 1 to 4).

6. The tree forming is finished when all terminal nodes contain samples of the same class (this can also be restricted by some stopping criterion, such as the minimum number of samples in terminal nodes).

For the impurity measure function, the common choice is Gini index, which is defined by

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

where $\hat{p}_{mk}$ denotes the proportion of samples in node m that belong to class k. This measure is minimized when the proportion $\hat{p}_{mk}$ is close to either zero or one, coresponding to the pure nodes occupied by a single class.

The prediction process for new samples with a finished tree is rather straightforward. It is done simply by selecting the most common class in the terminal node, which corresponds to the predictor values in a given sample.

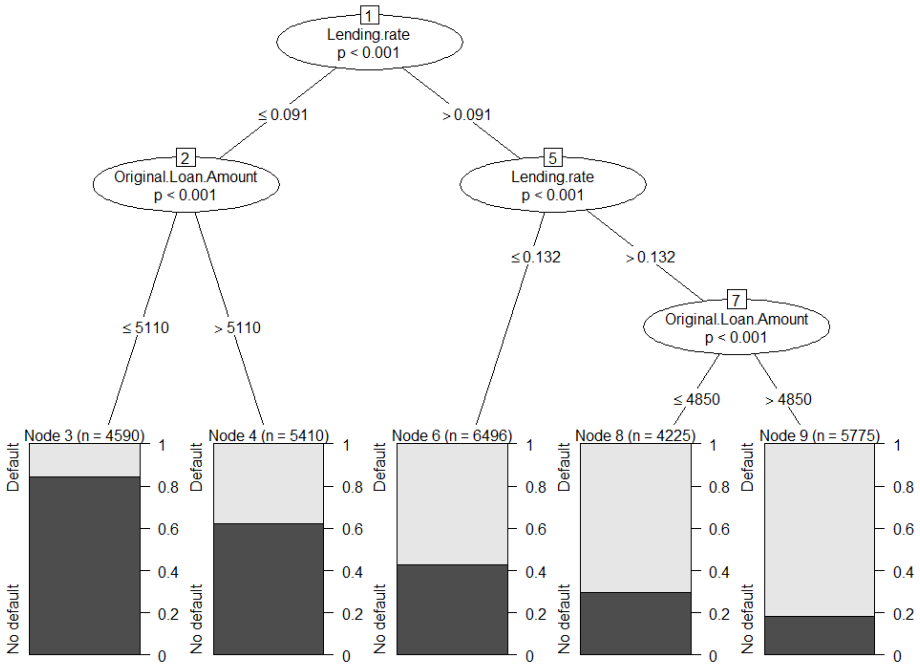

Figure 4. A simplified single classification tree structure for predicting the loan outcome in the Zopa dataset (created with R package "party"). Although acquired from a different implementation of random forests (conditional inference trees) than the one utilized in this thesis, this graph illustrates the common logic behind the generic decision tree algorithm. The variables identified as best for obtaining a split at each node (associated with the lowest impurity) are provided in the ellipses corresponding to the position of each node in the overall tree structure. The chosen thresholds for performing a split are located below the ellipses. The decision at each terminal node is made according to the majority class.

After constructing a decision tree, it is possible to extract its structure and see exactly how certain combinations of variable values in the input lead to a conclusion on the class label. The possibility of such model representation provides great opportunities for its interpretation and might be useful for getting a better understanding of the studied phenomena.

Despite all the mentioned strengths, the basic implementation of a decision tree algorithm has certain drawbacks, which lead to noticeable issues when applied in practice. One of those is that the decision tree is naturally prone to overfitting when grown to maximal depth (so that terminal nodes contain only a few training samples). Although this can be somewhat mitigated by limiting tree complexity via an appropriate cross-validation procedure, the resulting accuracy of single decision trees is likely to be worse when compared to other properly optimized classification methods. In addition, such trees are very unstable (small changes in training data might lead to a very different tree structure and predictions), which is a rather undesirable behavior.

3.2 Random forests as an improvement over decision tree classifiers

Random forest is an extension of the same idea introduced by decision trees that allows to overcome its major shortcomings while maintaining the similar benefits. It is represented by an ensemble of single decision tree models that can vote on class prediction, resulting in a selection of most popular class (Breiman, 2001).

Random forests provide several significant improvements when compared to single decision tree. The most prominent one is reduced overfitting, which results from using bagging combined with random sample and feature selection while creating tree splits.

The core principle behind random forests is bootstrap aggregation (bagging), which requires building a large number of separate decision trees. Normally, the classification error is steadily reduced with the addition of new trees, but starting from few hundreds of trees, the improvement in prediction error tends to be negligible. When the predictions from individual trees are averaged, it is possible to obtain a significant decrease in variance of a classification procedure, which results in higher prediction accuracy. Each tree in an ensemble is only allowed to use a random fraction (approximately 2/3) of training samples (selected with replacement), and is not restricted in its maximal depth, which would be necessary to do for a single tree model in order to reduce overfitting.

Further de-correlation of individual trees in the ensemble is obtained by restricting the number of predictor variables to be considered for a split in each new node of a tree. For a

classification case, this results in randomly choosing the m number of predictors out of p total predictors such that $m \approx \sqrt{p}$.

The abovementioned parameters for random forests are normally regarded as default options that tend to perform the best regardless of the specific dataset and were initially recommended by their inventors. However, it is certainly possible to use other approaches and perform a tuning procedure with the validation dataset. For instance, it is possible to consider all available variables at each split without restricting their number or perform the selection of training samples without replacement. In a similar way as for the single decision trees, it is also possible to manually specify other parameters that limit the complexity of each tree in a forest, but those are not typically used in the random forest setting.

A valuable property of random forests is their ability to provide an approximation of the unbiased accuracy estimate (referred to as OOB error). It is acquired by running predictions on unseen out-of-bag samples with trees, where such samples were not present during their construction process, thus eliminating the need for a separate cross-validation procedure.

3.3 Interpretation of random forests

Unfortunately, the interpretation of a single tree structure does not make much sense in the model with a large number of trees. However, random forests provide another good opportunity for a model interpretation, which is variable importance measures. These will act as a foundation for feature selection methods discussed and implemented in this work. There are two widely used importance measures in decision tree algorithms, which provide relative importance for each predictor variable used in a model. Those are called Gini importance and permutation importance.

Gini importance is directly linked to the Gini index mentioned previously and is calculated from training data by averaging the decrease in node impurity from all splits done with a specific variable. The obvious problem with this approach is that it is based on training samples, which results in overly optimistic importance measures. On the other hand, the permutation importance uses OOB samples to calculate the average decrease in prediction accuracy after randomly shuffling all the values of that variable among OOB samples. The

idea behind this approach is that if a certain variable has little or no predictive power, then such a permutation procedure would not impact the classification accuracy. It is assumed to be a more robust indicator of variable importance from the viewpoint of model predictive capabilities. Another good side of this method is that it can be considered a model-agnostic way for assessing variable importance and, in principle, can be applied for use with other machine learning methods besides decision trees (a separate test dataset would be needed in such case). Some researchers have pointed out that permutation importance should be preferred over the standard version of Gini importance as a generally more unbiased metric (Strobl et al., 2007). Therefore, the feature selection methods based on this metric have been favored for use in this work.

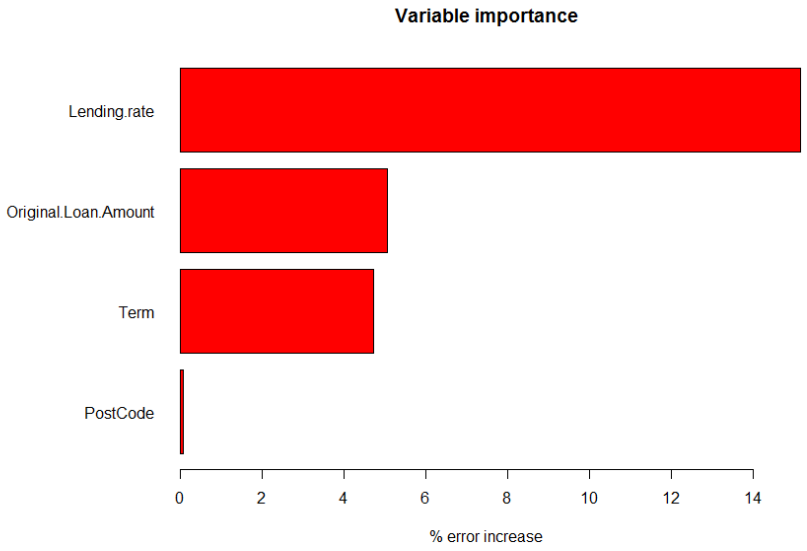

Figure 5. An example of variable importance metrics acquired from the random forest model based on the Zopa dataset. The percentage of error increase provides the relative measure of how individual variables in a random forest model (comprised of all available predictor variables) contribute to making accurate predictions by measuring the decrease in overall prediction accuracy in OOB samples before and after applying random permutation on predictor variable values in those samples.

The last note here regarding permutation importance is that there are two existing versions of permutation importance (unscaled and scaled). The latter is different from the unscaled version in the matter that the average decrease in prediction accuracy over trees in an

ensemble is normalized by its standard error. The idea behind this approach is that it allows to account for differences in importance metrics averaged over a large number of trees. On the other hand, it does not properly account for the effect of various sample sizes, so it has been found to be a non-robust metric of importance with undesirable statistical properties (Strobl and Zeileis, 2008). Therefore, the unscaled version was used in this thesis to report the resulting variable importance.

As it been mentioned previously, one of the points of criticism aimed at random forests is that they are essentially ”black box” models that do not provide as good interpretation opportunities as single trees. Although this is already mitigated with variable importance metrics, there exist some strategies for deriving additional insights from random forests. One of the most popular ones corresponds to the construction of partial dependence plots, which were initially suggested by Friedman (2001) for gradient boosting. The great feature of such plots is that they allow observing the marginal effect of any chosen predictor variable on the target variable. In the case of classification, the y-axis of such plot represents the log of the fraction of votes for the class of interest, while the x-axis corresponds to various category levels in case of categorical variables or available range of numerical variable according to training data.

In principle, generation of such plots can be viewed as a way to represent the change in the target variable as a function of its predictor by accounting for the average effect of all other predictors. More formally, the plotted function is represented by following equation:

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}), \tag{2}$$

where x is the variable for which the marginal effect is being sought and x_{iC} is the rest of variables in data. The equation can be further expanded as follows:

$$f(x) = \log p_k(x) - \frac{1}{K}\sum_{j=1}^{K} \log p_j(x), \tag{3}$$

with K representing the number of classes, k corresponding to the class for which the partial plot is built and p_j corresponding to the proportion of class j.

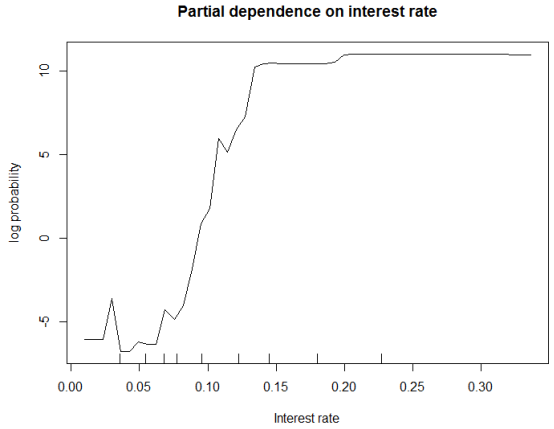

Figure 6. An example of a partial dependence plot that is based on the Zopa dataset (created with "randomForest" package). This is a graphical illustration of how the changes in the underlying predictor variable affect the risk of default (the larger values on the y-axis correspond to an increased probability of default). The predictor variable is displayed in its original scale. The tick marks on top of the x-axis correspond to deciles in the variable used for building the plot (outliers removed prior to plot generation).

Although it is technically possible to extract a single tree structure to visualize it in a similar manner as in the subsection 3.1, in case of the random forest, it would make little sense, as the predictive accuracy is acquired by growing very complex trees combined with decision averaging. Hence, it becomes impossible to extract any meaningful insights without artificially restricting its depth. This task becomes even more problematic when considering tens and hundreds of variables in the same tree at once.

4 LITERATURE REVIEW

4.1 Notes on literature search process

This thesis made use of various sources of information on covered topics (39 in total), including academic papers (32), books (3), and reports (4). The majority of those were found through a web search engine (e.g. freely accessible e-books and reports from the web pages of the corresponding research organizations), publication databases (including Elsevier ScienceDirect, SpringerLink, EBSCO, and JSTOR), and also in references of some papers. The papers which were initially found via keyword search were briefly evaluated with regards to their relevance to the topics and issues considered in this thesis (based on their abstracts and also taking into account the citation count), so only a fraction of the most relevant ones were further studied in detail and used in the thesis.

The following diagram outlines the keywords that were used most frequently (either separately or in combination) through the literature search process:

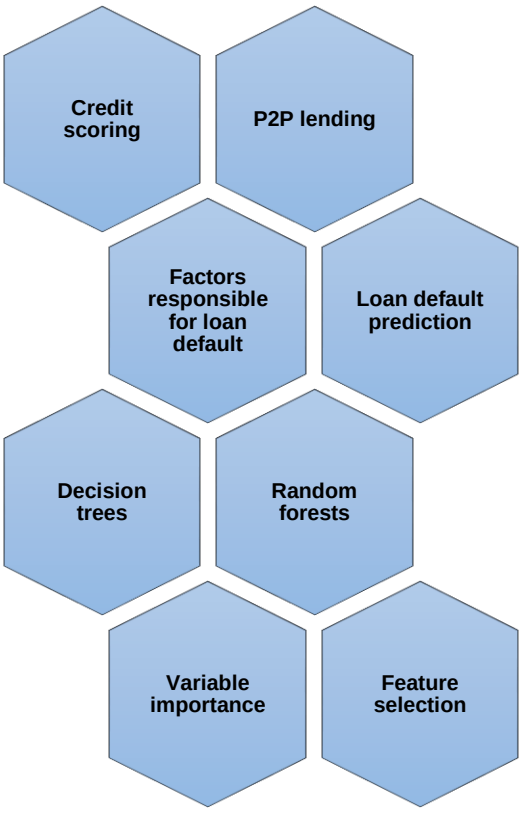

Figure 7. The main keywords used during the literature search process.

The four main topics of papers, which were referenced in this study, can be categorized as follows:

1. Credit scoring practices and methods
2. Prior studies of important predictors in P2P lending
3. Specifics of decision trees and random forests
4. Feature selection methods in general and their specialized modifications used with random forests.

The following subsections are focused on the actual literature review with respect to practical application of random forests and the associated feature selection methods, as well as previously identified relevant predictors for credit scoring in P2P lending. They cover only a fraction of all referenced papers, while the rest are mentioned in the other parts of this thesis.

4.2 Random forests as a tool for predictions

Although this thesis is focused on the identification of relevant variables, it would make sense first to provide a general overview of random forest classifiers in regard to their suitability for the given task.

According to Hastie et al. (2009), the random forest is a remarkably well-performing machine learning method, which is highly praised by various authors due to its prediction accuracy and interpretability. As per credit risk modeling analytical report by Deloitte Risk Advisory (2018), "Random forest can be used to save time on data management steps and to identify the most important characteristics in a dataset". Hapfelmeier and Ulm (2013) indicate that random forests are frequently applied in research fields such as social, econometric, and clinical studies as they achieve a high prediction accuracy and are able to identify informative variables.

Hand and Henley (1997) state in their paper on credit scoring methods that decision trees are considered as one of the possible methods in the development of scoring cards along with discriminant analysis and logistic regression, but they also stress that even in the context

of credit scoring it is difficult to know beforehand what type of classifier is the most suitable in a particular scenario.

There is a large number of papers, which attempt to compare the performance of various machine learning methods, including random forests. Although it is unlikely that there exists a single best method, there has been some evidence that random forests tend to achieve better predictive accuracy in general and also when compared to the classical methods used in credit scoring. Such examples are presented below.

A highly comprehensive study by Fernández-Delgado et al. (2014), which compared a wide variety of classifiers with respect to their performance on a very large number of datasets, concludes that prediction accuracy of random forests tends to be among the best. According to a large-scale experiment by Couronné et al. (2018) that involved 243 datasets from various scientific fields, the standard implementation of random forests with default hyperparameters has been found to outperform the logistic regression in a majority of cases. The latter is a more classical statistical learning technique that has been regarded as a standard tool in most financial applications (Siddiqi, 2005). A more focused study by Malekipirbazari and Aksakalli (2015) that considered the credit scoring with machine learning methods for P2P lending revealed that random forests outperform more basic machine learning methods, such as logistic regression, with respect to predictive accuracy.

4.3 Feature selection with random forests

As per Hand and Henley (1997), the use of wrapper methods (e.g. stepwise selection) for identifying the important variables is considered as one of the standard approaches in credit scoring, along with relying on expert opinion and using methods focused on the relevance of individual variables, such as information value.

In a similar way as for classifiers, it is unlikely that there is a single best-performing feature selection method, so it could be beneficial to test a variety of methods in order to choose the one most suitable for the studied data. Despite that such an extensive comparison was outside the scope of this thesis, an attempt has been made to establish justification for using specifically selected methods, while also identifying their possible shortcomings based on existing literature.

Liu and Schumann (2005) compared a number of different feature selection methods specifically in the context of credit scoring and concluded that wrapper-based methods performed among the best, also noting that automated data mining procedures for feature selection are likely to become more frequently used over time. According to Guyon (2003), wrapper feature selection methods are especially notable as being universally applicable and simple methods to use, although being criticized for the "brute force" nature of their approach. Xue et al. (2015) note that wrapper methods are often argued to result in more accurate classifications, at the cost of being more computationally expensive. When speaking of random forests specifically, the use of their variable importance metrics has been shown to provide very good results on par with more traditional methods like univariate statistical tests and information-theoretic methods (Bommert et al., 2020). The variable importance measures were found to be a suitable tool for reducing the number of predictors while maintaining predictive accuracy similar to one of a full model (Liaw and Wiener, 2002). In addition, a simulation study by Archer and Kimes (2008) that considered the effectiveness of random forest variable importance metrics in identifying the true predictor variables found that those metrics are indeed able to provide accurate measures of discriminative ability of individual predictors.

Despite all the previously mentioned benefits of random forests, some researchers have identified certain issues with them, which can also affect the feature selection process. Those are elaborated in greater detail below.

The Gini index, which is used through the construction of CART trees and for calculating the Gini importance, can supposedly introduce bias during the construction process of random forests and also when performing feature selection with them. It seems that while considering variables for a split when adding new nodes, the trees favor continuous variables and categorical variables with a large number of levels. Moreover, the bootstrap sampling procedure used with sample replacement can act as another source of bias also when using permutation importance, which by itself appears to be more a robust indicator comparing to Gini importance. As a consequence, when considering metrics aggregated over all trees, this might result in misleading variable importance scores for both Gini and permutation importance. (Strobl et al., 2007)

Another important issue when considering the task of finding the relevant predictors is the correlation among several features. Strobl et al. (2008) conclude in their study that the tree construction algorithm and variable importance metrics in the standard implementation of random forests tend to overestimate the relevance of correlated predictors. Similarly, Archer and Kimes (2008) report that for the case of most important predictors, their correlates were also appearing close to the top positions of importance ranking. On the other hand, results from the paper by Genuer et al. (2010) indicate that the importance of several correlated predictors tends to get distributed among them.

A viable solution for penalizing correlated predictors of low influence was suggested by Strobl et al. (2008), which requires applying a conditional permutation importance procedure in combination with unbiased conditional inference trees. Those take a somewhat different approach for constructing a more robust version of the random forest using the methods described by Hothorn et al. (2006). Conditional inference trees also seem to treat features of the varying scale of measurement or number of categories in a more reasonable way, although this requires way larger amounts of computational power comparing to traditional random forests (Strobl et al., 2007). As a result, given the relatively large size of studied datasets, this limitation did not allow to apply the potentially better modification of random forests in the current study, although there is no strong evidence that it offers any practical advantages in real-world applications. In fact, the large-scale classifier comparison experiment by Fernández-Delgado et al. (2014) found such a modification to actually perform worse than the standard implementation of random forests, which further justifies disregarding the conditional inference approach.

Several authors have attempted to compare the random forest based feature selection methods using a large number of simulated and real datasets, including all methods applied in this thesis (except greedy backward selection). For instance, the paper by Speiser et al. (2019) claims that the method “VSURF” tends to outperform the other methods, which was also in line with some other preceding studies. On the other hand, the study by Degenhardt et al. (2019) reports that methods “Boruta” and “vita” appear to be the most powerful methods, although they did not include method “VSURF” in their comparison. It is unclear if results from such papers could be universally applicable, which was the reason to utilize different methods in this thesis.

4.4 Relevant predictors for credit scoring in P2P lending

Despite that credit scoring is applied in various contexts, it is likely that predictors relevant for bank loans would be different from those suitable for P2P lending risk assessment, as the characteristics of typical borrowers using such platforms seem to be unique (Jiang et al., 2018). The rest of this subsection is focused on analyzing the information on relevant predictors specifically in the P2P lending context.

A number of papers have previously considered a similar task of identifying important predictors in P2P lending. Their major limitation is that they are mostly based on the same source of data (LendingClub dataset). In addition, they do not seem to consider all the available variables and instead limit the number of the initially considered predictors to the most promising ones. Some of the notable conclusions of those papers are presented below.

Serrano-Cinca et al. (2015) studied LendingClub data from years 2008-2014 and applied several approaches for identifying important predictors (survival analysis, statistical hypothesis testing, and logistic regression). They came to a conclusion that some of the important predictors include loan grade (assigned by LendingClub based on their own estimators of risk), interest rate, loan purpose, and a number of borrower characteristics (income, housing situation, credit history, and amount of debt), also noting that interest rate is highly correlated with loan grade. In addition, they conclude that the length of employment and loan amount are not statistically significant variables.

Another paper by Emekter et al. (2015) also considered the same data but from the different period (years 2007-2012). The same methods, as in the previously mentioned study, were used for evaluating the variable importance. Similarly, it concludes that loan grade and interest rate are the most relevant predictors for determining the risk of default. In addition, there were some other variables identified as important (loan term, FICO score, and revolving line utilization).

A more comprehensive study by Croux et al. (2020) covers nearly all variables available in LendingClub dataset. Although they also use logistic regression, they note that it becomes somewhat unreliable when attempting to include such a large number of variables, so they

additionally verify its results with regularized regression (lasso). Interestingly, they decided to complement the original data with some other relevant information from external sources corresponding to loans in the dataset. This information included the borrower's occupation category, average taxable income specific to borrower's location, and several macroeconomic variables indicating the overall market performance. They conclude that the most important variables include credit score, homeownership, borrower's job category, loan maturity, and purpose. Other important variables include the number of collections over last year (excluding medical expenses), whether the applicant applied alone, debt-to-income ratio, number of delinquencies over the past two years, number of credit inquiries, derogatory public records, and bankruptcies. Some macroeconomic variables also seem to have an impact on risk of default (return on the Russell 2000 Index and GDP growth). In addition, they stress that even in the presence of a large number of variables, the loan grades assigned by platform appear to be significant, while having a low correlation with credit score coming from the external agencies. Altogether, approximately 58 variables have been identified as significant out of 100 initially considered variables over the first run of regression, and a similar number was achieved after additional verification with the lasso. Finally, the authors examine whether the fact of defaulted loans being overrepresented in data from recent years (which occurs after discarding current loans) creates some sort of sampling bias, which might result in less reliable conclusions on variable importance. By conducting certain robustness checks, they conclude that there is no apparent selection bias.

Although not being a part of the current study, the unconventional sources of information on loan applicants are also actively researched as a potential way of predicting loan outcome. As an example, Berg et al. (2018) argue that information obtained from digital footprints of consumer loan applicants (i.e. information that people leave online simply by accessing or registering on a website) can be as useful as traditional credit scores for predicting the risk of default. Duarte et al. (2012) claim that borrowers with a trustworthy appearance based on their photographs tend to have better credit scores and lower default occurrences. Finally, Dorfleitner et al. (2015) conclude that text-based soft information extracted from loan applications, such as spelling mistakes and mentioning of positive emotion, only affects the probability of getting the loan funded but not its outcome, also noting that interest rate appears to be the most influential factor contributing to the risk of default.

5 DETAILS ON DATASETS, PRE-PROCESSING AND IMPLEMENTATION

5.1 General overview of used datasets

All datasets used in this study were acquired from the publicly available data provided by P2P lending platforms. They contain historical information on funded loans and associated indicator variables related to borrowers. The currently active loans were excluded from all datasets, since their outcome is unknown yet. Nearly all of the originally available variables were used in further variable selection process regardless if they are seemingly redundant or irrelevant, except those that contain identification data that is unique for each loan or information that became known through the actual loan repayment process (this does not apply for the past loans). Such an approach became possible as the applied variable selection methods can identify all variables relevant for loan outcome regardless of their correlation with each other.

The table below contains the summary information on all datasets used in feature selection and evaluation procedure. The indicated number of variables includes the loan outcome used as a target variable in the classification model. All of the displayed numbers correspond to datasets that were acquired after completing all pre-processing steps but before performing any other steps related to the further feature selection and evaluation procedure outlined subsection 5.6 (e.g. dataset splitting and balancing the distribution of defaulted versus non-defaulted cases in training sets).

Table 1. Summary of P2P lending datasets used in this study.

Dataset name	Area of operation	Issue years of loans	Number of variables	Number of samples	Percentage of defaults
Bondora	Estonia, Finland and Spain (mainly)	2009-2017	47	27 071	69,22%
LendingClub	USA	2012-2018	105	54 026	21,72%
Prosper	USA	2007-2014	50	44 144	30,31%
Zopa	UK	2005-2020	5	476 777	12,59%

The actual pre-processing steps and identified issues related to each dataset are covered in greater detail in the following subsections.

The important remark that needs to be done with respect to results outlined further in this thesis is that variables with the similar meaning across datasets but a somewhat different definition (corresponding to each P2P lending platform) still are presented under the unified notation. As an example, the variables originally named as “lending rate”, “borrower rate”, and “interest rate” across different datasets are presented using the common generic label (“interest rate”). Although they might be calculated in a somewhat different manner by each P2P lending platform, for the purposes of this thesis, it seemed appropriate to treat them as approximately equal for the ease of comparison and aggregation. However, the reader can clarify the original notation and meaning of each variable used across all datasets in the corresponding appendices of this thesis.

In addition, there were few occasions when it was not possible to find a publicly available unambiguous description of a certain variable (nearly all of those cases correspond to variables in the Zopa dataset). Since it seemed wasteful to exclude such variables from the analysis, the most likely definitions were assumed based on the originally existing labels in datasets. Therefore, the provided descriptions might not match the up-to-date definitions of the variables completely according to the P2P lending platforms.

5.2 Bondora

The initially downloaded dataset contained 144 430 loan application cases from the years 2009-2020. The first issue that became evident after the quick exploration of the dataset is that essentially all records starting from the year 2017 contain missing values over a large number of potentially important columns. It is unknown whether platform operators stopped collecting this information or if something else caused this issue. Since applying imputation for such a large number of samples would certainly be unrealistic, the most natural choice seemed to remove loans originated during those years, although they comprised a substantial fraction of data. There was also another small group of old loans that were financed for the currently unsupported purposes. That group also had a lot of missing or invalid values, so it was removed as well. At the end of this step, there still was a small fraction of missing values in a certain numerical variable (monthly payment amount), and since there was no available information on the proper calculation of those values in case of this specific dataset, those cases were omitted completely.

As a next step, the binary class column (indicating the occurrence of default event) was created based on a column that contained the date of default for defaulted loans and empty value for non-defaulted loans. Despite there already was a column of loan status ("Current", "Repaid", or "Late"), the information in that column seemed to be inadequate for establishing "good" and "bad" loans for the following reasons:

1. Loans with "Repaid" status were partially comprised of defaulted loans or loans that had gone overdue for a very long period of time, while loans with "Late" status had some cases of non-defaulted loans that had never been overdue (or only for a very short period).

2. Some non-defaulted loans have been overdue long enough to be considered as "bad" loans.

Due to these reasons, the decision has been made to categorize loans into "bad" and "good" classes with the following scheme: loans should be considered as defaulted if they have an entry indicating default or if they have been more than 30 days overdue, otherwise they should be regarded as non-defaulted.

Once the loans with "Current" status were filtered out, the proportion of loans considered as defaulted in remaining data became extremely high (69,22%). While considering these numbers, one needs to remember that the general goal in the currently used methodology is to discover the "bad" loans even if some of those do not meet the strict definition of default from the perspective of the P2P lending platform. Nevertheless, even if one would disregard the information on overdue periods and simply categorize loans based on the occurrence of entry indicating default, they would still comprise a majority of all samples. Interestingly, this cannot be completely attributed to the recent loans with long maturity periods that resulted in early default, since even after filtering out all loans after 2014, the defaulted loans were making up an unusually large proportion (approximately half of all samples).

Certain columns were converted into a more appropriate format. For instance, there were two columns denoting the original amount of requested loan and amount that was actually issued. Instead of the original amount, a new column was created, which indicates the

difference between the applied and originally requested amount. In addition, since there were several credit score columns and each borrower had only one of those scores available (corresponding to the country of residence), they were converted into a single column for ease of further interpretation.

Even though the majority of entries with missing values were removed during the initial pre-processing step, a small number of those were identified later in categorical columns as invalid values (i.e. those that do not correspond to any of the values available in the official data dictionary). For all of those cases, the imputation procedure was carried out. Since only the categorical variables had missing values at that point, and the number of missing cases was quite small relative to dataset size, the most frequent category level was assigned. This operation was performed with respect to the borrower's country.

Finally, for categorical variables with a very high number of levels (those were the city, county, and employment position), only 50 most frequent levels were kept, while the rest were aggregated as a single category level labeled "Other". This was done mainly for two reasons:

1. There is a technical limitation of the maximum number of levels in each categorical variable imposed by the original implementation of the random forest algorithm in R, which was used for generating the partial dependence plots.

2. It is unlikely that category levels that were present only in a few cases in the whole dataset can provide reliable information for the model (those could include e.g. an unusually rare employment position or very small city with few applicants).

5.3 LendingClub

The original dataset contained 2 260 701 loans from years 2007-2018 with the following states: "current", "in grace period", "late" (possible range of 15-30 or 31-120 days), "charged off", "default", "fully repaid", and "not meeting credit policy". Due to the absence of reliable information on overdue periods in fully repaid loans, all of them were regarded as "good". The ones with status indicated as "current", "late (15-30 days)", "in grace period", and "not meeting credit policy" were excluded from further analysis due to uncertainty associated

with those. The rest were aggregated to represent two classes of loans as follows: those with status “repaid” were considered as non-defaulted loans, while those with status “late (31-120 days)”, “charged off”, and “default” were considered as defaulted loans.

During the next pre-processing step, some columns were modified in order to become more meaningful for further classification. This included the conversion of columns containing only month and year related to a certain event (such as the opening of borrower’s first credit line) to the number of days passed from that event until the loan issue date.

While preparing this dataset, there were certain issues associated with the treatment of missing values over a large number of variables, especially numerical ones. It became necessary to develop a flexible strategy for dealing with such data. The number of missing values seemed to vary significantly across different variables: it could be as small as only a few missing records per variable, but in most cases, it was close to 100 000 or even 1 000 000 missing records. Moreover, it became obvious that some variables have a large number of missing values simply due to their nature (such as the number of months since last borrower’s delinquency), while for others, the reason is unknown.

The typical way to deal with variables where missing cases make a majority is simply to remove such variables, but then there is no way to check if they appeared to be important, at least for the non-missing cases. Since random forests can model complex non-linear dependencies in numerical variables efficiently, a reasonable approach seemed to manually impute the “dummy” values in place of missing records so that they can be reliably treated by random forest as a different category of samples. Therefore, the following strategy was applied:

1. When the amount of missing values per variable exceeds 200 000, replace missing values in that variable with the fixed negative number (-100) that is outside of the normal range for such variables (where lowest possible value is normally 0).

2. If the number of missing values does not exceed 200 000, omit those cases completely.

In a similar way as for the Bondora dataset, there were few interesting categorical variables with a too large number of category levels (zip code and employment title). Those were reduced to represent only 50 most frequently occurring levels, while others were aggregated as a single category level "Other".

After filtering out irrelevant loans and those with missing values, the remaining ones corresponded to a very large number of cases (approximately one million). Due to limited computational resources, it would not be possible to acquire feature selection results in a reasonable time. Therefore, it became necessary to reduce the dataset size to a small fraction of the original one. It was acquired by randomly selecting 5% of samples, maintaining the approximately same ratio between defaulted and non-defaulted loans.

5.4 Prosper

Prosper dataset contained loan statuses, which were quite similar to those previously covered for LendingClub data. The total number of loan application cases was 113 937 for loans originated in the years 2007-2014. The loans which were overdue for more than 30 days, charged off, and defaulted ones were considered as belonging to the single class of defaulted loans, while completed loans were considered as non-defaulted. The rest of the loans with statuses "Current" and "Cancelled" were rejected.

Few variables were modified in order to become more suitable for further analysis. For instance, the date of opening the borrower's first credit line was replaced by the corresponding number of days before the loan issue date. The variables indicating the old and updated versions of loan grade (with only one of those available depending on origination year) were combined in a single column. Finally, the borrowers' occupation categories were aggregated to represent only 50 most frequently occurring occupations.

With respect to the treatment of missing values, a hybrid strategy was followed. For the variables where missing values were making a majority of cases (information on previous loans funded on the same platform), those values were simply replaced with zeros. On the other hand, there were several thousands of entries that seemed to have missing values across several important variables, so those entries were removed completely.

5.5 Zopa

The dataset coming from Zopa platform was considerably smaller than the others. Although it contained 734 079 samples (years 2005-2020), after the removal of unnecessary variables, only four predictors were left for further analysis: loan amount, interest rate, loan term, and postal code of the borrower. Despite such a small number of variables available, they were anticipated to be quite relevant for classification task and expected to provide additional verification of findings from other datasets.

There was very little pre-processing required. Since there were only three possible loan states ("Current", "Completed", "Late", and "Defaulted") and no exact information on overdue periods, it seemed reasonable to eliminate current loans and consider late loans also as defaulted.

With the exception of some blank entries in the postal code variable, there were no missing values, while the blank entries would simply be considered as another categorical variable level. The postal codes were aggregated in the same manner as before to include separate category levels only for 50 most frequently occurring ones.

Since there were only a few variables in the first place, it became possible to use a larger number of samples compared to other datasets without compromising on computational times.

5.6 Specifics of feature selection procedure and classifier configuration

The empirical study part was implemented with R software and packages corresponding to the feature selection methods chosen for this study, with settings mostly left at their defaults suggested by authors (the utilized methods and corresponding R packages are summarized in the table on page 46).

Table 2. Summary of used feature selection methods and corresponding R packages.

Feature selection method	**Feature selection package**	**Random forest package**	**Additional notes on feature selection procedure**
Altmann et al. (2010)	Available as an internal part of the corresponding random forest package	"ranger"	Cutoff for important features based on p-values less than 0,01
Kursa and Rudnicki (2010)	"Boruta"	"ranger"	Cutoff for important features based on p-values less than 0,01. Tentative variables considered as non-important
Genuer et al. (2010)	"VSURF"	"ranger"	Variables from second stage ("interpretation") considered as important
Greedy backward selection (generic wrapper)	n/a	"ranger"	Implemented directly with R script as a wrapper using "ranger" random forest package

Similarly, there was no extensive tuning performed for random forest and mostly default parameters were used. Although the chosen number of trees was somewhat smaller than the default number (500), it was not possible to use more trees in the case of this thesis due to long computational times associated with repeated runs of random forests in feature selection procedure. During the experimentation stage, the different numbers were tried and 500 trees appeared to provide only slightly better prediction accuracy than 200 trees and slightly more stable variable importance metrics. Ultimately, since the primary goal of this thesis is relative variable importance comparison rather the maximization of predictive accuracy, it seemed reasonable to proceed with a smaller number of trees. At the same time, it is possible that some of the weakly important variables would not be identified successfully due to a decrease in the detection power of feature selection methods, but the improvement they would provide could be almost unnoticeable in the practical applications.

The actually used random forest parameters are given in the following list:

- Number of trees: 200
- Number of variables tried at each split: square root of the number of variables in the model
- Maximal tree depth: unlimited
- Minimal node size: 1
- Handling of levels in unordered categorical variables: according to their proportion falling in second class.

A problem that became evident during the experimentation process is that random forests seem to focus on the majority class during the tree building process. As a result, very high accuracy was achieved for the class of non-defaulted loans and very low for the defaulted loans. Since it is arguably better to identify most of the riskier loans at the cost of increased false negatives (i.e. "good" loans classified as "bad"), the majority class was artificially downsampled so that parts of the dataset used for feature selection and training of model contained the exactly same number of samples from each class. However, the original ratio of defaulted versus non-defaulted loans was preserved in testing folds of cross-validation procedure to obtain an unbiased accuracy estimate, which served as the main metric for comparing various sets of selected variables. Table 1 on page 39 contains the original fraction of defaulted versus non-defaulted cases, where the number of samples is comprised of all samples used for feature selection and cross-validation error estimation process (before performing any downsampling).

The general procedure that was followed for performing feature selection could be described as follows:

1. Randomly split datasets acquired after the pre-processing stage into two parts of equal size while preserving the original ratio of two classes in each part. The first part is used only for the feature selection process, while the second is used for cross-validation procedure.

2. Downsample the majority class in the first part in a way that ensures an equal ratio of defaulted and non-defaulted cases.

3. Run the feature selection methods using the first part of data to identify the corresponding subsets of important features with respect to each method.

4. Run the cross-validation procedure on the second part of data as follows: create ten stratified folds (i.e. preserving the original ratio of both classes) and use 90% of data for training and 10% for testing at each iteration to acquire the ten estimates of cross-validation accuracy across all folds. The majority class is downsampled only in the training fold (similarly as in step 2).

5. Repeat the cross-validation procedure to compare the different subsets of variables (also including a model with all variables) identified in step 3 and calculate the mean 10-fold cross-validation accuracy for each case.

6. Make a final decision on important variables by using the following voting rule: the variable must be selected by at least 2 out of 4 methods in order to be considered important.

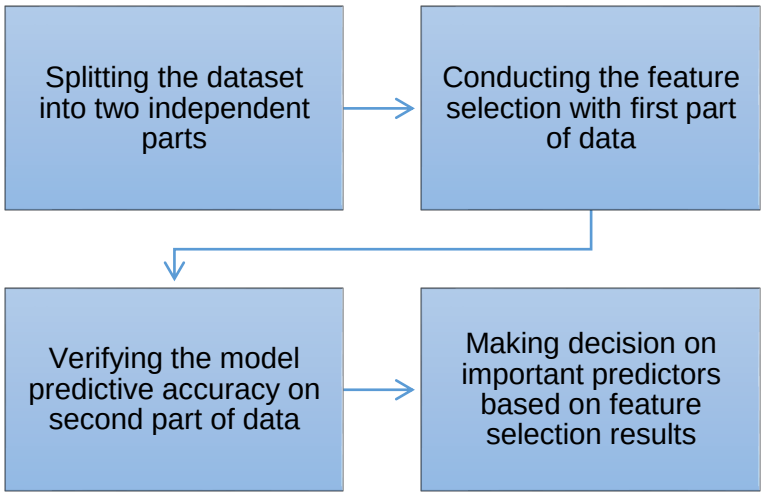

Figure 8. The main steps in applied feature selection procedure.

Since the role of GBS method is rather auxiliary (as it been pointed out in subsection 2.5) and also taking into account that it almost always selected variables which were already selected by at least two other methods, this rule could be regarded as a majority voting among the remaining three main methods, despite not meeting the exact definition of it. The justification for applying the voting rule in this context stems from the fact that using only a

single feature selection method for decision-making on importance could result in an increased false-positive discovery rate. It also allows ensuring that possibly redundant variables not selected by some of the used methods could still be identified as important.

Although prediction accuracy is calculated for several sets of features via cross-validation procedure, it is not directly used in the decision-making process for variable identification, but rather to give an idea of how classification accuracy based on various subsets of features of different sizes diverges from the model with all predictors. The most important predictive accuracy metric is arguably the one that is calculated from the full model, as it provides a generalized understanding of whether the underlying classification model is good enough to make any further inference on variable importance.

The results of the variable selection are further complemented with the actual importance metrics derived from the model built on the whole set of variables along with partial dependence plots for some of the most important variables (focused on the class of defaulted loans). The relative level of importance across several variables is compared by using the actual variable importance metrics (specifically, unscaled permutation importance based on OOB samples) from a model containing all available predictors.

Since some of the studied datasets were comprised from a very large number of variables, the detailed results outlined in the following section are only provided for the most noteworthy variables. The full list of variable selection results and the corresponding importance metrics can be found in the appendices.

6 RESULTS OF VARIABLE SELECTION PROCESS

6.1 General overview of classification accuracy and variable selection results

As it been previously mentioned, the interpretation of a random forest prediction model would only make sense if the underlying model is good enough for making predictions. Since the achieved prediction accuracy was reasonably high for all datasets, it becomes possible to consider the acquired variable selection and importance results as quite reliable and proceed with a more detailed analysis. Although the overall accuracy does not seem to be exceptionally high, the majority of defaulted loans are successfully identified. This supports the general idea that it is better to miss some of the potentially "good" loans while minimizing the investment into "bad" ones. Note that with respect to classification accuracy sensitivity (true positive rate) and specificity (true negative rate), the default as an outcome of the loan is considered here as a true positive class, while the absence of default corresponds to a true negative class.

Table 3. The prediction accuracy of baseline random forest classifiers including all available variables for each dataset.

Dataset	Overall 10-fold CV mean accuracy	Sensitivity (defaulted loans)	Specificity (non-defaulted loans)	Variables estimated to be important
Bondora	69,87%	70,83%	67,43%	27/46
LendingClub	64,25%	67,29%	63,36%	32/104
Prosper	66,67%	72,23%	64,33%	39/49
Zopa	70,79%	75,51%	70,11%	3/4

The number of times a variable was selected by different methods decreased almost steadily with the corresponding decrease of permutation importance metrics. The general idea for justifying the application of feature selection has been proved to be valid, as only part of all available variables were found to be important. For most of the variables, the rate of predictive accuracy improvement per adding a single variable was of a similar magnitude or even lower than the accuracy fluctuations arising from different runs of the same random forest model (the difference between two runs could reach approximately 0,5%). Still, their

combined power seemed to accumulate with large sets of variables. Overall, this means that only a fraction of variables was necessary to achieve a similar predictive accuracy as in the full model.

6.2 Bondora

The model with a full set of predictors was able to achieve the 69,87% overall accuracy and 70,83% accuracy for the class of defaulted loans. The smallest model with approximately the same accuracy (69,28% and 70,25%, respectively) had only 23 out of 46 predictor variables. A slightly larger set of variables (27) have been identified as important according to the voting rule based on feature selection methods. Unlike other datasets, the importance here decreased more gradually from the most important variables towards the least important ones.

The variables with the strongest importance included information on borrower's origin (location of residence and language), previous loan repayments, loan grade (denoting borrower's risk score), and the loan-specific indicators (monthly payment, interest rate, and term). The least important ones corresponded mostly to demographic variables (age and education), verification of income, and monthly payment day. Finally, the majority of variables in the middle of importance range were comprised of those describing borrower's existing liabilities, financial situation, and occupation, while also including the loan amount.

The partial plots on pages 52-53 illustrate how the probability of default is being affected by loan term, interest rate, amount, borrower's total income, monthly payment amount, previous loan repayments, number of existing liabilities, and loan grade. It is important to remember that the original data was comprised of borrowers residing in three different countries, which should be taken into account while interpreting the variables indicating the effect of borrower's location and income on prediction of default. The first notable observation here is that longer loan term and higher amount seem to be associated with an increased risk of default. The most likely explanation could be that it is naturally more probable to run into financial problems over an extended period of time. The higher interest rates are also associated with the increase in occurrence of default, but since the platforms initially assign the higher interest rates to riskier borrowers (Emekter et al., 2015), it is not possible to imply direct causation here.

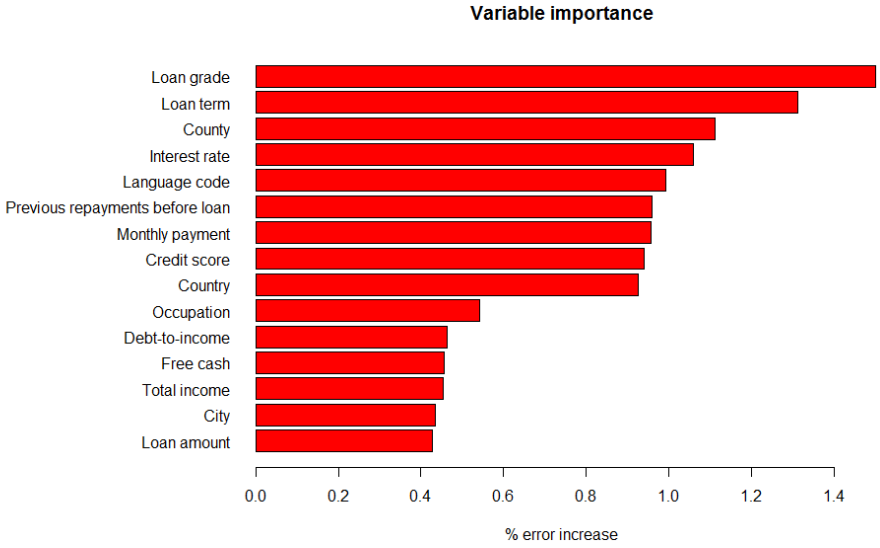

Figure 9. Variable importance for the top 15 variables in the Bondora data. The percentage of error increase provides the relative measure of how individual variables in a random forest model (comprised of all available predictor variables) contribute to making accurate predictions by measuring the decrease in overall prediction accuracy in OOB samples before and after applying random permutation on predictor variable values in those samples.

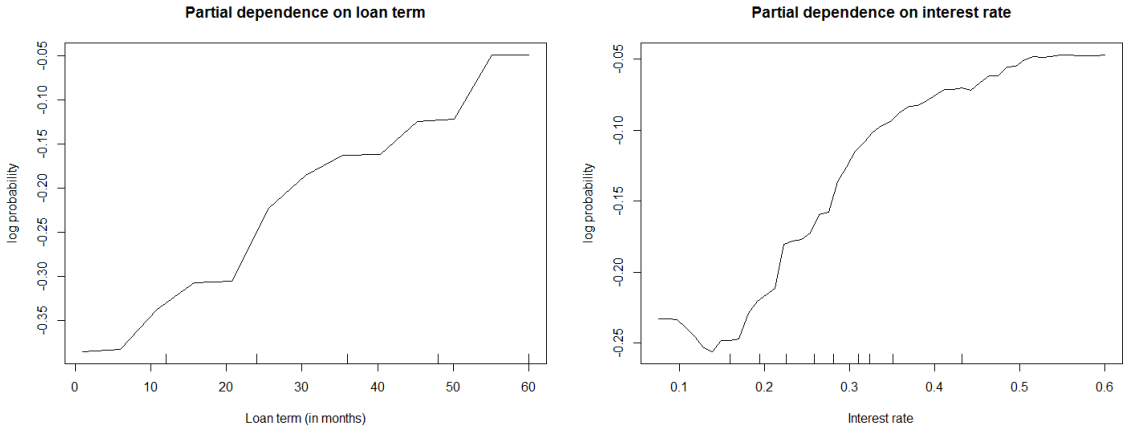

Figure 10. Partial dependence plots for several important variables in the Bondora data. These plots provide a graphical illustration of how the changes in the underlying predictor variables affect the risk of default (the larger values on the y-axis correspond to an increased probability of default). The predictor variables are displayed in their original scale. The tick marks on top of the x-axis correspond to deciles in the variable used for building the plot (outliers removed prior to plot generation).

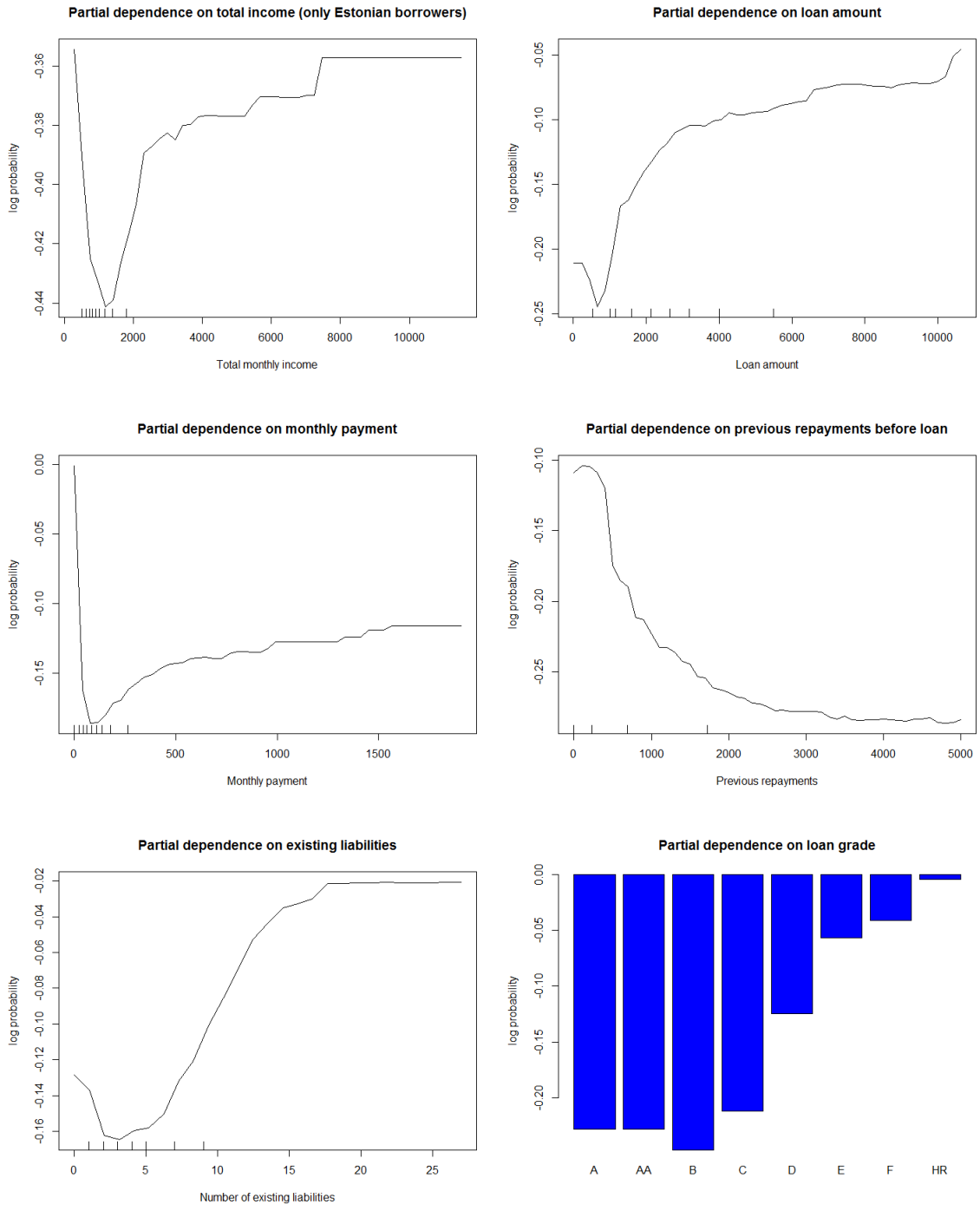

Figure 11. Partial dependence plots for several important variables in the Bondora data (continuation).

According to the U-shaped association between the amount of income and risk of default, the latter is very high when a borrower has an income close to zero, but then it steadily decreases with the increase in income, which is quite expected. However, when the income starts to be unusually high (starting approximately from the 80th percentile of the income distribution), the risk of default starts to increase rapidly. A possible explanation for this unusual behavior could be the misreported income data provided by the borrower.

The graph illustrating the dependence of the probability of default on monthly installments is difficult for interpretation, since the monthly paid amount is dependent on multiple underlying factors. The lowest monthly paid amounts appear to correspond to a very high risk of default (probably linked to longer loan duration). With the increase in paid amount, the risk quickly decreases until some point and then starts to increase again.

Previous repayments on loans seem to be negatively correlated with the risk of default. Basically, this means that borrowers who were able to repay their past loans successfully are also likely to repay the future loans, with higher amounts resulting in a decreased risk.

A somewhat unusual association can be seen between the number of borrower's existing liabilities and the risk of default. According to the partial dependence plot, having no liabilities at all or a very large amount of liabilities correspond to a high risk of default, while having few liabilities indicates a lower risk. The most likely explanation goes as follows: the borrowers without existing liabilities might have never taken a loan, and so they are comprised of a large number of potential defaulters, while those who have a very large number of liabilities might have a difficult time to fulfill them all (if they ever intend to do so).

Finally, the partial dependence plot with respect to Bondora custom rating (loan grade) displays the expected pattern: the borrowers with higher risk profile tend to default more often than those with a lower risk profile.

6.3 LendingClub

The estimated overall model accuracy using all variables in the LendingClub dataset appeared to be 64,25%, with 67,29% of defaulted loans correctly identified. Reducing the

number of predictors to 54 out of 104 (as suggested by one of the methods) resulted only in slightly worse accuracy (64,05% and 66,62%, respectively).

According to the voting rule results, 32 variables were found to be important. The importance of the four top-ranked variables (loan term, interest rate, grade, and subgrade) is notably higher than for the subsequent ones.

In a similar manner as for Bondora data, the loan-specific indicators were contributing the most (monthly payment, interest rate, loan term, and amount), along with variable describing borrower's risk profile (loan grade, subgrade, and credit score) and financial situation (credit limit, bank account balance, total income, and debt-to-income ratio). The demographic variables related to borrower were found to be rather weakly important (occupation and location of residence). A large number of less frequently used borrower's financial indicators related to various types of banking accounts and credit limits were found in the middle of importance range.

The partial plots on page 56-57 show the marginal effect of some important variables on the risk of default (loan term, interest rate, amount, borrower's income, credit limit, current balance, credit score, and loan grade). In the same way as in the Bondora dataset, the longer loan term, higher interest rate, and higher amount are associated with an increased risk of default. The dependence on borrower's current balance and income is very similar and likely to have a similar explanation as in the previously mentioned case for the Bondora dataset, although the increase of risk is not that drastic for the unusually high income levels.

According to the plot illustrating the effect of different credit limits on the risk of default, the higher credit limits are associated with a reduction in that risk. The simple explanation is that for a borrower, it is easier to repay the loan when they have a larger available credit on their bank card.

Lastly, the graph illustrating dependence on credit score seems to correspond with common knowledge: the higher risk profile as per score assigned by a consumer reporting agency is associated with an increased probability of default. The same logic applies to the internal loan grade metrics calculated by the lending platform.

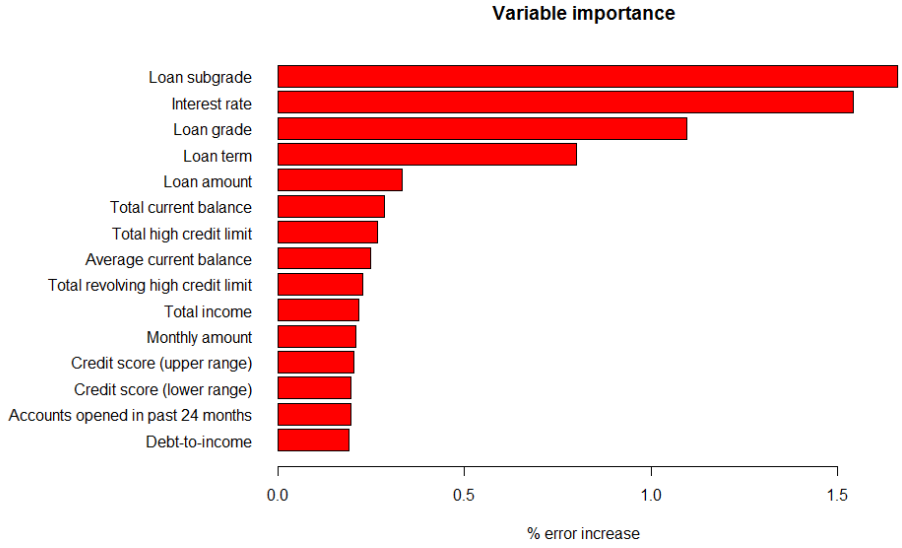

Figure 12. Variable importance for the top 15 variables in the LendingClub data. The percentage of error increase provides the relative measure of how individual variables in a random forest model (comprised of all available predictor variables) contribute to making accurate predictions by measuring the decrease in overall prediction accuracy in OOB samples before and after applying random permutation on predictor variable values in those samples.

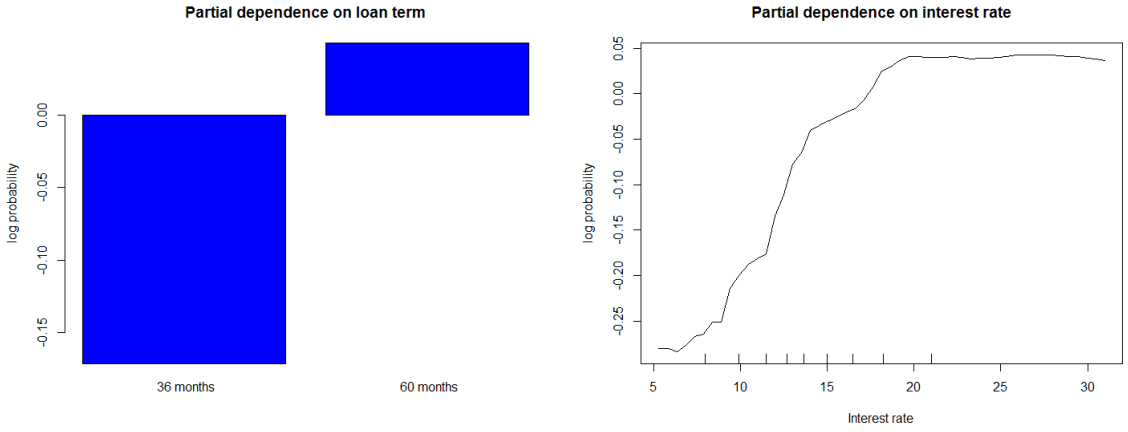

Figure 13. Partial dependence plots for several important variables in the LendingClub data. These plots provide a graphical illustration of how the changes in the underlying predictor variables affect the risk of default (the larger values on the y-axis correspond to an increased probability of default). The predictor variables are displayed in their original scale. The tick marks on top of the x-axis correspond to deciles in the variable used for building the plot (outliers removed prior to plot generation).

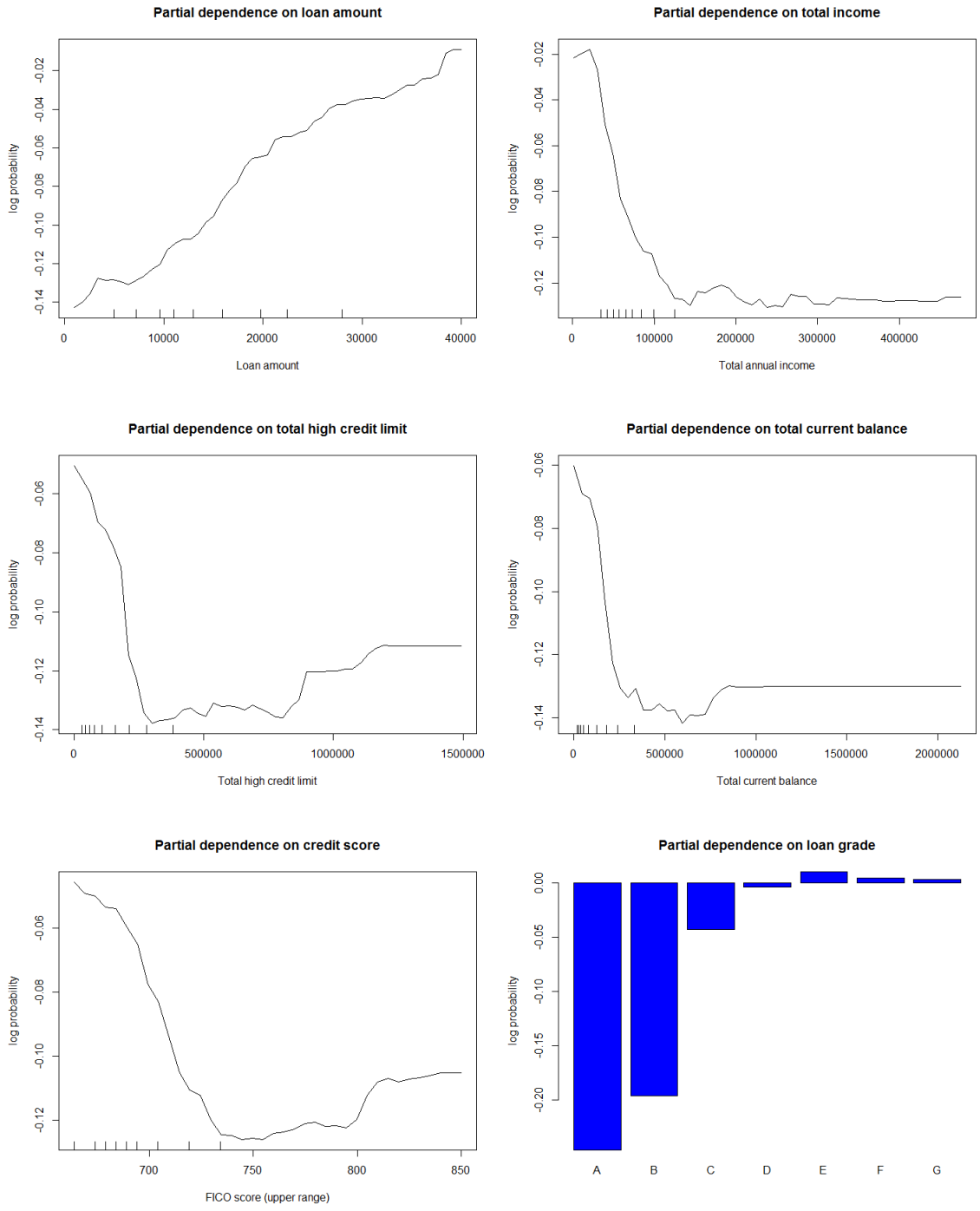

Figure 14. Partial dependence plots for several important variables in the LendingClub data (continuation).

6.4 Prosper

The model utilizing all of the available predictors from Prosper data resulted in the overall prediction accuracy of 66,67% and defaulted loans accuracy of 72,23%. The corresponding accuracy metrics of a smaller model (32 variables out of 49) were 66,21% and 71,59%. On the other hand, 39 variables appeared to be relevant for predicting loan outcome according to the voting decision. The variable denoting interest rate seems to be a lot more important comparing to other variables, while for the following ones, the importance decreases quite gradually.

In addition to interest rate, monthly payment, and loan amount, the most important variables included borrower's financial indicators (available bank card credit, credit utilization, debt-to-income ratio, and income), the variables related to risk profile (credit score and loan grade), and the number of credit inquiries. The borrower's occupation appeared to be considerably more important than in other datasets, while the loan amount appeared as less important. The weakly important variables and those with moderate importance consisted of information on previously taken loans and other less frequently used financial indicators. Except for the borrower's occupation area, the demographic indicators were weakly important (such as home ownership and employment status).

The partial plots on pages 59-60 illustrate the impact of several important variables on the risk of default (loan term, interest rate, amount, borrower's total income, number of credit inquiries, available bank card credit, FICO credit score, and Prosper custom risk score).

The interest rate, loan amount, and term are associated with the probability of default in a similar way as described previously. In addition, the higher number of credit inquiries is also associated with an increase in the probability of default. The more inquiries could be associated with a higher number of taken loans or other debt obligations, and having more such obligations makes it more difficult to repay them all.

The rest of the variables (credit limit, income, FICO credit score, and custom Prosper risk score) display similar straightforward patterns as previously described for corresponding variables in LendingClub data.

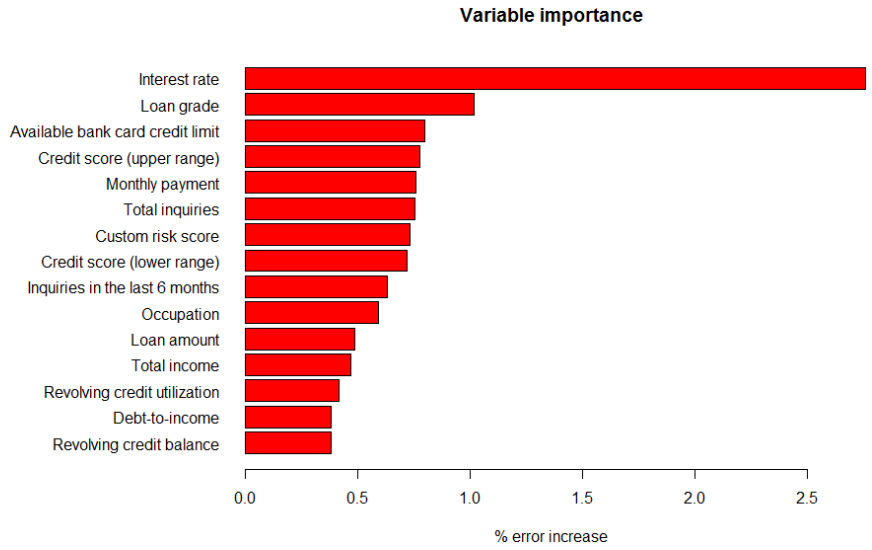

Figure 15. Variable importance for the top 15 variables in the Prosper data. The percentage of error increase provides the relative measure of how individual variables in a random forest model (comprised of all available predictor variables) contribute to making accurate predictions by measuring the decrease in overall prediction accuracy in OOB samples before and after applying random permutation on predictor variable values in those samples.

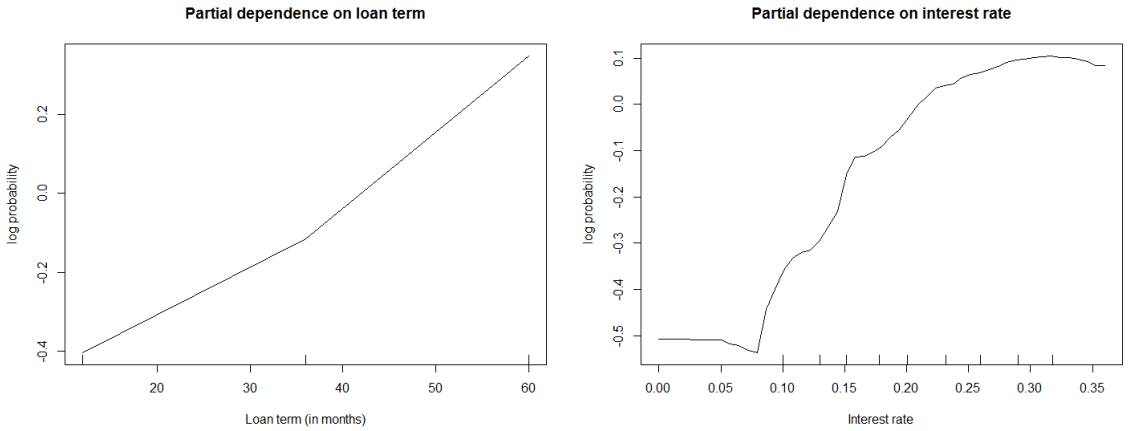

Figure 16. Partial dependence plots for several important variables in the Prosper data. These plots provide a graphical illustration of how the changes in the underlying predictor variables affect the risk of default (the larger values on the y-axis correspond to an increased probability of default). The predictor variables are displayed in their original scale. The tick marks on top of the x-axis correspond to deciles in the variable used for building the plot (outliers removed prior to plot generation).

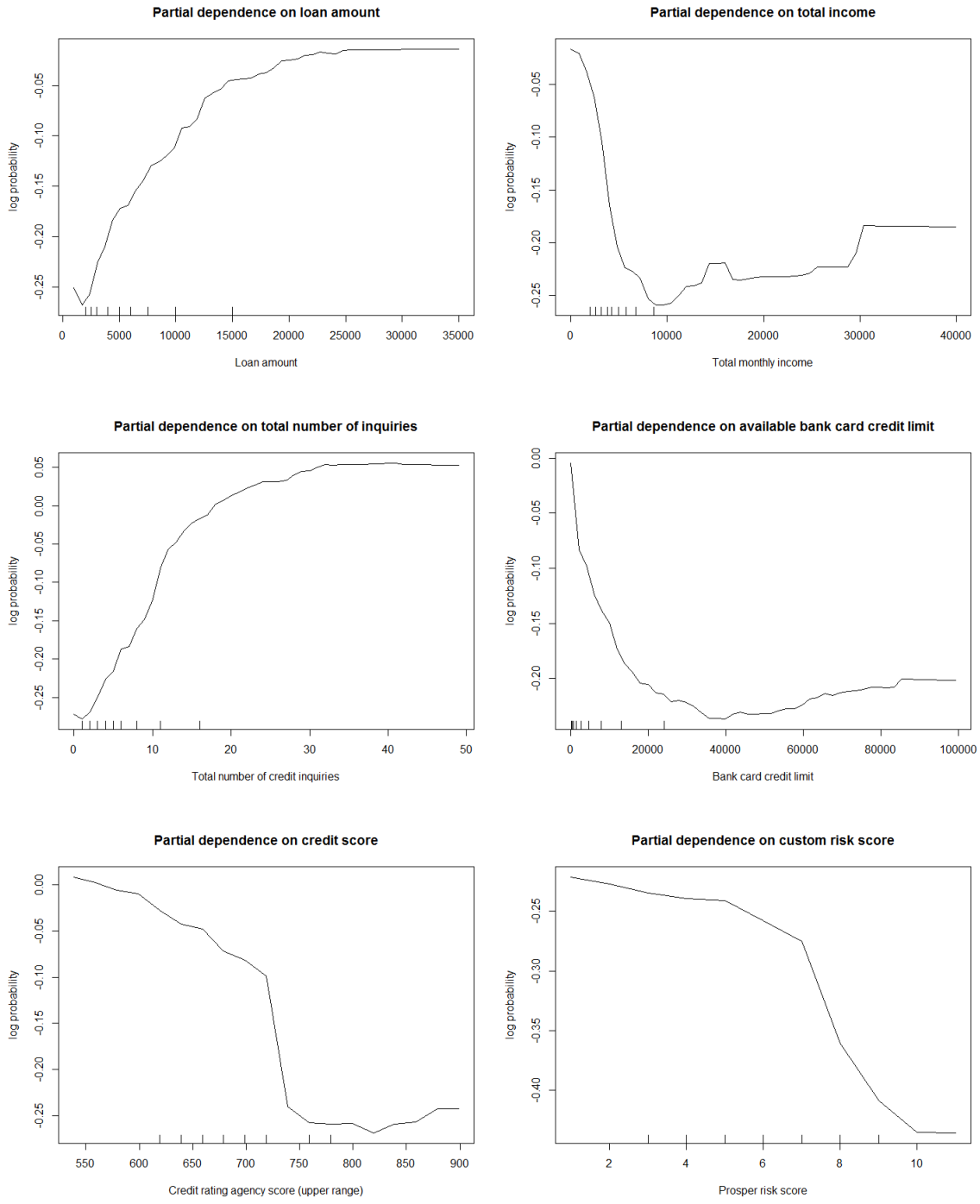

Figure 17. Partial dependence plots for several important variables in the Prosper data (continuation).

6.5 Zopa

The findings based on this dataset are strongly in line with the preceding ones. Three loan-based indicators were found to be highly important (loan term, interest rate, and amount). An interesting observation was that the full model with all variables including the non-selected fourth variable (denoting postal code) actually had worse predictive accuracy (70,79% overall and 75,51% for defaulted loans) comparing to model with three predictors (corresponds to 72,60% and 76,89%), which was probably a result of overfitting. It also appeared that the prediction accuracy was somewhat higher compared to models for predicting loan outcome in the other datasets despite an extremely small number of predictors in the model.

The partial dependence plots for all important variables in this dataset (loan amount, interest rate, and term) and provided on page 62. The loan amount, term, and interest rate are connected with the risk of default in the same way as previously described for other datasets.

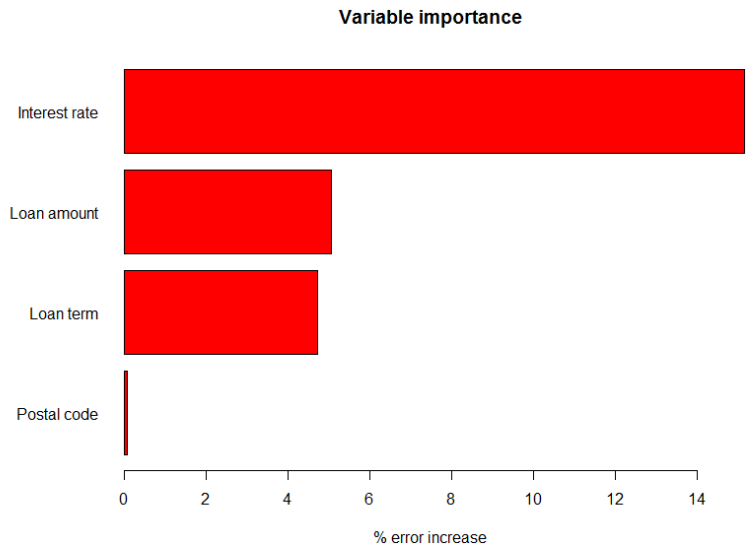

Figure 18. Variable importance for all variables in the Zopa data. The percentage of error increase provides the relative measure of how individual variables in a random forest model (comprised of all available predictor variables) contribute to making accurate predictions by measuring the decrease in overall prediction accuracy in OOB samples before and after applying random permutation on predictor variable values in those samples.

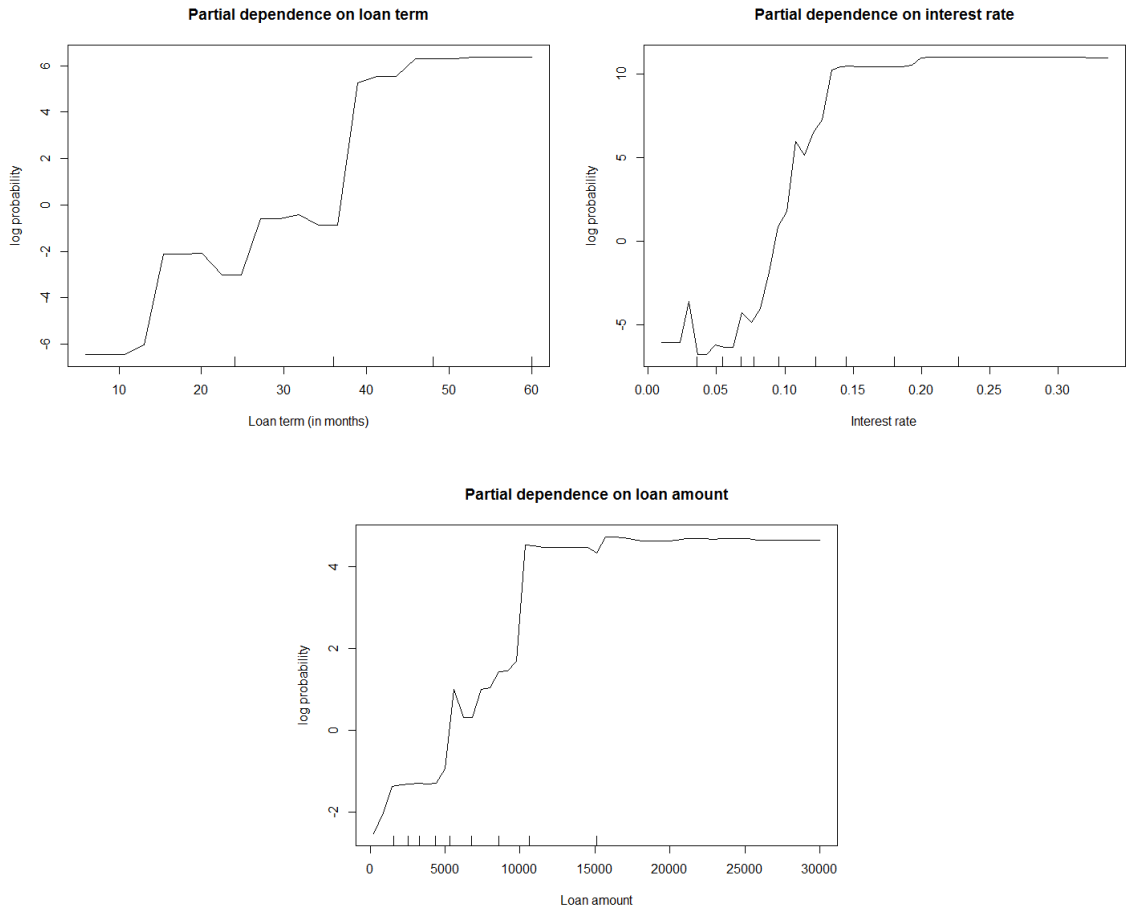

Figure 19. Partial dependence plots for all important variables in the Zopa data. These plots provide a graphical illustration of how the changes in the underlying predictor variables affect the risk of default (the larger values on the y-axis correspond to an increased probability of default). The predictor variables are displayed in their original scale. The tick marks on top of the x-axis correspond to deciles in the variable used for building the plot (outliers removed prior to plot generation).

7 CONCLUSIONS

7.1 Summary of the results and conclusions

This thesis applied random forest based feature selection methodology in order to acquire more comprehensive insights on relevant predictors for credit scoring in P2P lending. The study confirmed that many of the predictors available in the data collected by P2P lending platforms indeed are highly relevant for predicting the risk of loan defaults. On the other hand, a significant number of variables had no impact on predictions or appeared to be rather weakly important and provided only negligible improvement in prediction accuracy. This suggests that the collection and inclusion of such variables in predictive models should be reconsidered by the operators of P2P platforms as well as investors. The fact that baseline classification models achieved a reasonable predictive accuracy provided a justification to apply a detailed analysis of how various predictor variables can contribute to more accurate predictions. A focus was put on presenting the grounds for considering random forests as a promising tool for classification and identification of important features in general, while also taking note of possible caveats. In order to improve the reliability of findings, the decision-making process was aided by combining the results coming from several feature selection methods. The results were made more elaborate by incorporating several P2P lending datasets into the same study while also taking into consideration the insights obtained from other similar papers.

The rest of this subsection covers the acquired results by aggregating them across all studied datasets with respect to the initially set research questions.

1) *Which variables tend to be the most and least important for estimating the risk of default for loan applicants in the context of the P2P lending classification task?*

The variables with lowest permutation importance metrics, which were not found as important according to variable selection methods, appear to be unique for each dataset, so it is not possible to provide any generalized conclusions on those. On the other hand, the relative importance of variables, which were selected enough times to be deemed as important, displays some similar patterns across all datasets. Since the actual values of importance metrics are not suited for direct comparison across different datasets for various

reasons, the focus here is put on outlining the approximate position of those predictors in importance ranking list.

It seems unambiguous that main loan characteristics (interest rate, term, and amount) appear to be some of the most important variables and, in principle, provide a good foundation for achieving a reasonable predictive accuracy, which was seen with the Zopa dataset case that only had four variables in the first place. It did not have monthly payment amount in the available variables, but it appeared as highly important in other datasets. Although the loan grade and credit score are essentially "black boxes" calculated by using other indicators, they also appeared on the top of the importance list.

Another group of variables that seem to be highly relevant across all datasets corresponds to the financial indicators, consisting e.g. from information on borrower's income, available balance, and credit limits, as well as the debt-to-income ratio.

With the exception of the borrower's occupation (that is probably highly related to the information on income), the demographic variables were of little importance (such as age, home ownership, and education). This was also true for the borrower's location of residence (within a single country).

2) *What is the direction of influence on the probability of default among numerical variables and separate levels in categorical variables?*

Although the partial dependence plots were only created for a small portion of important predictors available in several datasets, they showed the presence of clear dependencies between underlying predictors and the actual changes in risk of default. Some of the common patterns are illustrated further below.

With respect to variables describing the loan, there is a strong positive correlation between loan interest rate, term, and amount with the probability of default. The variables denoting the borrower's risk profile (e.g. credit score and loan grade) displayed the uniform pattern of lower risk profiles corresponding to a decreased probability of default and higher risk profiles corresponding to an increased probability. In addition, the variables related to

available credit limit and borrower's income appeared to share the same association with the probability of default, with higher values generally corresponding to a decreased probability compared to lower values.

3) *Are there notable discrepancies among subsets of variables estimated as important predictors when comparing results across several P2P lending datasets?*

For the variables that are common for all datasets, the obtained importance metrics seem to remain quite consistent even across several platforms used in different countries, which is especially true for loan-specific indicators, such as interest rate, term, and amount. Nevertheless, some minor differences are certainly present, and the most notable ones are presented below.

With respect to variables related to the borrower's location of residence, those appeared on top importance ranking positions in the Bondora dataset, while for other datasets they were of little importance. However, this dataset contained information on borrowers from several countries, which could have different baseline default occurrence rates. This was confirmed after re-running the random forest model using smaller subsets of data split by country, so the importance of the borrower's location did not seem that high anymore.

Another notable dissimilarity was about the verification of the borrower's income, which seemed to be important only in the Bondora data. It had virtually no importance in the Prosper dataset, but the number of non-verified borrowers there amounted only to a tiny percentage of all borrowers, while for other datasets, the distribution between verified and non-verified ones was quite similar and more balanced.

Although the variables related to debt-to-income ratio and credit utilization are more or less equally important for all datasets, the variables describing borrower's previously taken and repaid loans were a lot more important in Bondora data than Prosper data, despite that fraction of those with previously taken loans was quite similar in both datasets.

For some reason, the borrower's occupation seems to have a notably higher position in Prosper and Bondora variable ranking compared to LendingClub. It should be mentioned

that in the Prosper dataset, the occupation was initially selected from one of the pre-determined categories, while in other datasets, the occupation was supplied directly as a free-form text by the borrower and then aggregated based on most frequently occurring ones during data pre-processing. As a result, the different written forms of the same employment title and possible typographical errors were not taken into account. On the other hand, the loan amount was considerably more important in LendingClub dataset when compared to others.

4) *Does the information on variable importance acquired through feature selection correspond to the results from previous studies employing more traditional statistical approaches for determining variable significance?*

Generally, the results with respect to variables found to be important agree with findings from other papers that studied the determinants of risk of default in P2P lending, although it is complicated to compare the actual level of importance directly with those papers, especially considering the differences in the used subsets of variables, the underlying methods, and metrics outlining the variable importance.

Similar to the conclusions by Serrano-Cinca et al. (2015), the important variables in the LendingClub dataset appear to include the loan grade, interest rate, borrower's income, credit history, and amount of debt, while excluding the length of employment. The results of this thesis do not confirm the importance of loan purpose, while loan amount, contrary to their findings, was identified as important. Although home ownership (identified by authors as important) was not selected by the minimum required number of methods, it still appeared in the selection results of one of the used methods.

The acquired results appear to be strongly in line with findings by Emekter et al. (2015), who mentioned the interest rate, loan term, and revolving line utilization as important predictors. In addition, the loan grade and FICO score also appeared as important.

The results of this thesis also seem to mostly agree with the paper by Croux et al. (2020), which found approximately 58 variables to be significant out of 100. Although this is a larger number than obtained from the voting rule, a similar number of variables was discovered by

one of the used feature selection methods (54), although this observation does not take into account the differences in chosen subsets of variables and their relative importance level. As the number of studied variables was very large, the comparison here is performed only for the variables which were explicitly mentioned in their conclusions. Similar to their results, the credit score, loan grade, term, and borrower's debt-to-income ratio appeared as important variables, along with job title, number of credit inquiries, and home ownership type (the latter selected only by one method). With regards to contradictions, the variables identified as important in this thesis do not confirm the importance of loan purpose and derogatory records, while the number of public bankruptcies appeared in the selection results of one of the methods. Finally, the number of delinquencies over past two years prior to the loan, number of collections over last year (excluding medical expenses) and variables linked to second loan applicant in joint applications also were not found as important, although some of those were present only in a tiny fraction of loan application cases studied in this thesis, which could diminish their chances of being properly identified as contributing to overall prediction error reduction.

7.2 Limitations

This study utilized the specific feature selection methods together with the random forest classifier in order to study the relevance of predictor variables. Although any other classifier could be used in a similar methodology for selecting features, it is possible that ranking of important variables would appear different, as different classifiers tend to capture different types of dependencies in data, which in turn would affect the variable importance rankings. Variables that can be beneficial in the case of random forest classifier could appear as useless when they are considered in another setting, e.g. with a classifier that cannot discover the interaction or non-linear effects among variables. Moreover, there are multiple existing approaches within random forest feature selection methodology, so it might happen that using other approaches compared to those applied in this work would lead to somewhat different results. Also, there was no extensive tuning performed for random forest classifier and chosen feature selection methods. Although the default configuration settings seem to work well, in principle, it could be possible to acquire more reliable results e.g. by increasing the number of trees in the ensemble or trying some other configurations of the used models. It is likely that some of the weakly important variables were not successfully detected with

only 200 trees in the random forest model, but the improvement in the accuracy that they could provide would appear practically negligible.

Although there is strong evidence from the existing literature on the potential of random forests for performing feature selection and numerous examples from various applied research fields, their use specifically for determining the significance of features is a relatively recent and not very established area unlike the traditional statistical significance testing and logistic regression. Also, taking into account that they belong to the category of "black box" models with a certain degree of instability in their construction process, there is unlikely to be a definite set of universally accepted methodologies in that matter. The methodologies used in this paper were acquired through an experimentation process, which can highly depend on specific classification models and datasets utilized in the study. The risk of false-positive discoveries was somewhat mitigated by incorporating the voting rule based on several feature selection methods. On the other hand, one could think that if a certain variable was selected at least once, this already provides a good reason to consider it as important. Nevertheless, the variables that have very small importance scores should be considered with caution, as the improvement in prediction error provided by them can be close to even smaller than one that can be attributed to internal fluctuations arising from different runs of same random forest model, although a large collection of such variables could accumulate a significant improvement in prediction accuracy.

The prediction accuracy rates of random forests can be somewhat unstable in the presence of unbalanced classes and are also subject to sudden changes when considering different sets of predictors. This was somewhat mitigated by artificially ensuring the equal ratio of defaulted and non-defaulted cases in all datasets, so the classifiers attempted to seek for the optimal tradeoff while predicting both classes as accurately as possible. However, depending on what strategy is considered as best by the potential investor, it could be reasonable to modify datasets further for achieving a better accuracy for predicting defaulted loans at the cost of making more mistakes in predicting non-defaulted loans. This could also impact the results of variable importance metrics and feature selection methods, as they do not take into account the class-specific accuracies of the model. Ultimately, a separate study would be needed to understand exactly the monetary tradeoffs for different class-specific accuracies in order to build valid investment strategies.

It is crucial to understand that the absolute values of importance metrics and marginal effects of variables in provided partial dependence plots should not be treated as ground truth. Such numbers are generally non-additive estimates of importance, subject to strong variations when considering various sets of features at once due to potential redundancy and interaction effects in predictors. As an example, the interest rate in the Zopa dataset is associated with a 15,16% reduction in error. However, in the other datasets where the amount of variables is considerably large, this number appears to be a lot smaller (although the interest rate is more important than most of the other variables). If one would attempt to artificially restrict the number of variables to use e.g. for predicting loan outcome in the LendingClub dataset, it is likely that the absolute reduction in error would be a lot higher than 1,54% in a model with the full set of available variables. Therefore, it is best to consider those values as relative measures in importance. Similarly, the partial dependence plot should be viewed only as an approximation of the relationship between the predictor and the target variable.

The strong caution is required when one attempts to establish the true (real-world) dependency between the available predictor and target variables (impact on the probability of default) by using variable importance metrics or selected subsets of features from feature selection methods. This applies to the interpretation of the results acquired in this work and also in general for such methodology, with more detailed explanations following below.

First of all, despite that P2P lending shares many aspects with traditional lending at other financial institutions, it is unclear if indicators found to be important in loans funded on P2P lending platforms would be equally important in any credit scoring task. Moreover, despite that selected predictors are very likely to be top-performing with currently studied datasets with respect to classification accuracy, this provides only limited evidence of their real-world predictive power when estimating the risk of default, especially for other types of loans than those found in P2P lending.

The data used in this study was collected by a third-party (P2P lending platforms), which makes it impossible to verify the quality of the original data collection process or dataset integrity. For some of the datasets, the major proportion of values in certain variables were missing or invalid. Although they were encoded appropriately during pre-processing, this

could result in potentially underestimated variable importance metrics and selection outcomes, also impacting the interpretability of findings.

The loan application records only contain the cases when the loan was granted, as a result of entry requirements set by platform for borrowers and the specific choice made by lenders. In order to conduct a completely unbiased study for determining a true underlying importance of variables with respect to risk of default, it would be necessary to let everyone who expressed an interest in taking a loan to actually get it. This is obviously an unrealistic scenario that could only be achieved in a controlled experiment and unlikely happen at a real financial institution (including P2P lending platforms) that applies various precautionary measures for minimizing their risks and avoiding unwanted borrowers.

Finally, in the same way as correlation should not be confused with causation, the indicators of variable importance with respect to classification accuracy cannot be regarded as a sole indicator of variable dependencies without ensuring the absence of spurious relationships in data, which point at visible association between variables but do not represent a true dependence or a root cause. As an example, the P2P lending platforms seem to assign higher interest rates to loans with riskier borrowers using their own risk estimators (Emekter et al., 2015). This does not allow to imply that high interest rates, in fact, are the actual cause of more frequent defaults, although they are strongly associated with the loan outcome.

7.3 Suggestions for further research

There are potentially many aspects that could provide additional insights into the importance of variables and classification accuracy in P2P lending, which were not covered in this study. The most prominent ones are further discussed in this subsection.

First of all, it would be interesting to see how different sampling strategies could affect the results. For instance, for samples ordered in time, it could make more sense to split the dataset into training and validation/testing parts so that one part always is followed by the other from a time perspective. Although the samples do not represent time series, it is unclear whether some sort of look-ahead bias is introduced due to random sampling procedures implemented in this work. On the other hand, the time-based splits could result in unequal ratios of defaulted versus non-defaulted cases due to variations occurring over the years,

which might be an unwanted scenario. In addition, it could be beneficial to see how more rigorous sampling procedures that aim to exclude e.g. periods of global economic instability in the years 2007-2008 and periods of policy changes (with respect to internal policies of P2P lending platforms) could affect variable selection results.

As mentioned earlier, the ratio between defaulted and non-defaulted cases affects the prediction trade-offs that were favored by the decision tree during its construction. Therefore, it is possible to try creating different classifiers that correspond e.g. to risk-averse or other strategies, while also seeking for variables that are useful for predicting the specific class either of defaulted or non-defaulted loans. Moreover, it could be beneficial to conduct a study that considers the whole problem of the decision-making process for the investor from the viewpoint of expected monetary gains and losses or probability estimates rather than the particular assignment of the loan into a certain class.

Finally, some extra effort could be put to verify the presence of interactions among the variables estimated as important, as it is crucial to know whether they would be useful only together or also in separation. In the context of random forests, this could be achieved e.g. by building three-dimensional partial dependence plots or extracting most frequently occurring rules that were created during the decision tree construction process in order to understand better how combinations of different factors influence the risk of default.

REFERENCES

1. Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10), pp. 1340-1347.
2. Archer, K. J., Kimes, R. V., 2008. Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, 52(4), pp. 2249-2260.
3. Berg, T., Burg, V., Gombović, A., Puri, M., 2020. On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies*, 33(7), pp. 2845-2897.
4. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics and Data Analysis*, 143.
5. Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp. 5-32.
6. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. *Classification and Regression Trees*. Chapman & Hall/CRC.
7. Couronné, R., Probst, P., Boulesteix, A.-L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19(270).
8. Croux, C., Jagtiani, J., Korivi, T., Vulanovic, M., 2020. Important factors determining Fintech loan default: Evidence from a LendingClub consumer platform. *Journal of Economic Behavior and Organization*, 173, pp. 270-296.
9. Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, 20(2), pp. 492-503.
10. Deloitte Risk Advisory, 2018. *Point of View: Using Random Forest for credit risk models* (analytical report). Accessed 5 June 2020. Available at: https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/financial-services/deloitte-nl-fsi-using-random-forest-for-credit-risk-models.pdf
11. Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3).

12. Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., Kammler, J., 2016. Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking and Finance*, 64, pp. 169-187.
13. Duarte, J., Siegel, S., Young, L., 2012. Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *The Review of Financial Studies*, 25(8), pp. 2455-2483.
14. Emekter, R., Tu, Y., Jirasakuldech, B. , Lu, M., 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1), pp. 54-70.
15. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15(1), pp. 3133-3181.
16. Friedman, J. H., 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), pp. 1189-1232.
17. Galloway, I., 2009. Peer-to-Peer Lending and Community Development Finance. *Community Investments*, 21(3), pp. 18-39.
18. Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2010. Variable selection using random forests. *Pattern Recognition Letters*, 31(14), pp. 2225-2236.
19. Guyon, I., Elisseeff, A., 2003. An introduction to Variable and Feature selection. *Journal of Machine Learning Research*, 3, pp. 1157-1182.
20. Hand, D. J., Henley, W. E., 1997. Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3), pp. 523-541.
21. Hapfelmeier, A., Ulm, K., 2013. A new variable selection approach using Random Forests. *Computational Statistics and Data Analysis*, 60, pp. 50-69.
22. Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer-Verlag.
23. Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), pp. 651-674.
24. Janitza, S., Celik, E., Boulesteix, A.-L., 2018. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 12, pp. 885-915.

25. Jiang, C., Wang, Z., Wang, R., Ding, Y., 2018. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 266, pp. 511-529.
26. Kursa, M. B., Rudnicki, W. R., 2010. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11).
27. Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. *The Newsletter of the R Project*, 2(3), pp. 18-22.
28. Liu, Y., Schumann, M., 2005. Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, 56, pp. 1099-1108.
29. Malekipirbazari, M., Aksakalli, V., 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), pp. 4621-4631.
30. Milne, A., Parboteeah, P., 2016. *The Business Models and Economics of Peer-to-Peer Lending* (ECRI Research Report No. 17). The European Credit Research Institute. Accessed 5 June 2020. Available at: https://www.ceps.eu/wp-content/uploads/2016/05/ECRI%20RR17%20P2P%20Lending.pdf
31. Serrano-Cinca, C., Gutiérrez-Nieto, B., López-Palacios, L., 2015. Determinants of Default in P2P Lending. *PLoS One*, 10(10).
32. Siddiqi, N., 2005. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, Inc.
33. Speiser, J. L., Miller, M. E., Tooze, J., Ip, E., 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134, pp. 93-101.
34. Strobl, C., Zeileis, A., 2008. *Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance* (Technical Report 17). Department of Statistics, University of Munich. Accessed 5 June 2020. Available at: https://epub.ub.uni-muenchen.de/2111/1/techreport.pdf
35. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. *BMC bioinformatics*, 9(307).
36. Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).

37. U.S. Department of the Treasury, 2016. *Opportunities and Challenges in Online Marketplace Lending* (white paper). Accessed 5 June 2020. Available at: https://www.treasury.gov/connect/blog/Documents/Opportunities%20and%20Challenges%20in%20Online%20Marketplace%20Lending%20vRevised.pdf
38. Wright, M. N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1).
39. Xue, B., Zhang, M., Browne, W. N., 2015. A Comprehensive Comparison on Evolutionary Feature Selection Approaches to Classification. *International Journal of Computational Intelligence and Applications*, 14(2).

APPENDICES

APPENDIX 1. Variable importance and selection results by four methods (Bondora data)

Variable name	Method (abbreviated)				selection count	OOB error increase on permutation
	Altmann	Boruta	VSURF	GBS		
Rating	+	+	+		3	1,50%
LoanDuration	+	+	+	+	4	1,31%
County	+	+	+		3	1,11%
Interest	+	+	+		3	1,06%
LanguageCode	+	+	+		3	0,99%
PreviousRepaymentsBeforeLoan	+	+	+		3	0,96%
MonthlyPayment	+	+	+	+	4	0,96%
CreditScoreComb	+	+	+	+	4	0,94%
Country	+	+	+		3	0,93%
EmploymentPosition	+	+	+	+	4	0,54%
DebtToIncome	+	+	+		3	0,46%
FreeCash	+	+	+	+	4	0,46%
IncomeTotal	+	+	+		3	0,45%
City	+	+	+	+	4	0,44%
Amount	+	+	+		3	0,43%
AmountOfPreviousLoansBeforeLoan	+	+	+	+	4	0,40%
ExistingLiabilities	+	+	+	+	4	0,35%
IncomeFromPrincipalEmployer	+	+	+		3	0,35%
LiabilitiesTotal	+	+	+		3	0,31%
Age	+	+	+		3	0,31%
MonthlyPaymentDay	+	+	+		3	0,30%
NoOfPreviousLoansBeforeLoan	+	+			2	0,25%
AppliedAmountDiff	+	+			2	0,22%

Education	+	+	+	+	4	0,22%
NewCreditCustomer	+	+			2	0,21%
OccupationArea	+				1	0,19%
VerificationType	+	+	+	+	4	0,19%
RefinanceLiabilities	+	+			2	0,18%
HomeOwnershipType	+				1	0,15%
UseOfLoan					0	0,08%
WorkExperience					0	0,07%
Gender	+				1	0,07%
IncomeFromPension	+				1	0,06%
MaritalStatus	+				1	0,06%
ApplicationSignedHour					0	0,03%
IncomeFromFamilyAllowance					0	0,03%
IncomeFromSocialWelfare				+	1	0,03%
EmploymentStatus					0	0,03%
NrOfDependants					0	0,03%
PreviousEarlyRepaymentsCountBeforeLoan					0	0,01%
PreviousEarlyRepaymentsBeforeLoan					0	0,01%
IncomeOther					0	0,01%
EmploymentDurationCurrentEmployer					0	0,01%
IncomeFromChildSupport					0	0,00%
IncomeFromLeavePay					0	-0,01%
ApplicationSignedWeekday					0	-0,02%
Overall 10-fold CV mean accuracy (full model accuracy: 69,87%)	69,55%	69,08%	69,28%	66,78%		
Sensitivity (full model sensitivity: 70,83%)	70,47%	70,23%	70,25%	67,71%		
Specificity (full model specificity: 67,43%)	67,31%	66,59%	66,97%	64,69%		

Total variables selected	32/46	27/46	23/46	11/46

APPENDIX 2. Variable importance and selection results by four methods (LendingClub data)

Variable name	Method (abbreviated)				selection count	OOB error increase on permutation
	Altmann	Boruta	VSURF	GBS		
sub_grade	+	+	+		3	1,66%
int_rate	+	+	+		3	1,54%
grade	+	+	+	+	4	1,10%
term	+	+	+	+	4	0,80%
funded_amnt	+	+	+	+	4	0,33%
tot_cur_bal	+	+	+		3	0,29%
tot_hi_cred_lim	+	+	+		3	0,27%
avg_cur_bal	+	+	+		3	0,25%
total_rev_hi_lim	+	+	+		3	0,23%
annual_inc	+	+	+	+	4	0,22%
installment	+	+	+		3	0,21%
fico_range_high	+	+	+		3	0,20%
fico_range_low	+	+	+	+	4	0,20%
acc_open_past_24mths	+	+	+	+	4	0,20%
dti	+	+	+		3	0,19%
bc_open_to_buy	+	+	+		3	0,18%
total_bc_limit	+	+	+		3	0,17%
all_util	+	+	+		3	0,16%
total_bal_il	+	+			2	0,14%
revol_util	+		+		2	0,14%
total_bal_ex_mort	+		+		2	0,14%
max_bal_bc	+	+	+		3	0,13%
bc_util	+		+		2	0,12%
num_rev_accts	+				1	0,11%
total_acc	+				1	0,11%
revol_bal	+				1	0,11%

emp_title	+	+	+		3	0,11%
num_op_rev_tl	+				1	0,11%
open_rv_24m	+	+	+		3	0,11%
total_il_high_credit_limit	+				1	0,10%
inq_last_12m	+	+			2	0,10%
num_rev_tl_bal_gt_0	+	+			2	0,09%
num_bc_tl	+				1	0,09%
percent_bc_gt_75	+				1	0,09%
mo_sin_rcnt_rev_tl_op	+				1	0,09%
num_actv_rev_tl	+	+			2	0,09%
addr_state	+		+	+	3	0,09%
mths_since_recent_bc					0	0,08%
mo_sin_old_rev_tl_op	+		+		2	0,08%
num_bc_sats					0	0,08%
open_il_24m	+				1	0,08%
num_sats	+				1	0,08%
open_rv_12m					0	0,08%
num_tl_op_past_12m	+				1	0,08%
open_acc	+				1	0,07%
open_act_il	+			+	2	0,07%
mths_since_rcnt_il	+	+	+		3	0,07%
num_actv_bc_tl	+				1	0,07%
mo_sin_rcnt_tl	+				1	0,07%
inq_fi	+				1	0,06%
mort_acc					0	0,06%
verification_status	+				1	0,06%
home_ownership	+				1	0,05%
num_il_tl	+				1	0,05%
zip_code					0	0,05%
open_il_12m					0	0,05%
open_acc_6m	+				1	0,05%

earliest_cr_line_length			0	0,04%
mo_sin_old_il_acct			0	0,04%
mths_since_last_major_derog			0	0,04%
mths_since_recent_inq			0	0,03%
total_cu_tl			0	0,03%
il_util	+		1	0,03%
tot_coll_amt			0	0,03%
mths_since_recent_bc_dlq			0	0,03%
inq_last_6mths	+		1	0,02%
delinq_2yrs			0	0,02%
purpose			0	0,01%
pub_rec_bankruptcies	+		1	0,01%
mths_since_last_record			0	0,01%
emp_length			0	0,01%
pct_tl_nvr_dlq			0	0,01%
pub_rec			0	0,01%
mths_since_recent_revol_delinq			0	0,01%
initial_list_status			0	0,01%
num_tl_90g_dpd_24m			0	0,00%
collections_12_mths_ex_med			0	0,00%
tax_liens			0	0,00%
num_accts_ever_120_pd			0	0,00%
mths_since_last_delinq			0	0,00%
sec_app_inq_last_6mths			0	0,00%
disbursement_method			0	0,00%
sec_app_mths_since_last_major_derog			0	0,00%
chargeoff_within_12_mths			0	0,00%
sec_app_mort_acc			0	0,00%
delinq_amnt			0	0,00%
sec_app_chargeoff_within_12_mths		+	1	0,00%
num_tl_30dpd			0	0,00%

application_type					0	0,00%
sec_app_open_acc					0	0,00%
sec_app_earliest_cr_line_length					0	0,00%
annual_inc_joint					0	0,00%
num_tl_120dpd_2m				+	1	0,00%
amnt_diff					0	0,00%
sec_app_fico_range_high					0	0,00%
revol_bal_joint					0	0,00%
sec_app_num_rev_accts					0	0,00%
acc_now_delinq					0	0,00%
sec_app_collections_12_mths_ex_med					0	0,00%
sec_app_fico_range_low					0	0,00%
verification_status_joint					0	0,00%
dti_joint					0	0,00%
sec_app_open_act_il					0	0,00%
sec_app_revol_util					0	0,00%
Overall 10-fold CV mean accuracy (full model accuracy: 64,25%)	64,05%	64,22%	64,25%	64,04%		
Sensitivity (full model sensitivity: 67,29%)	66,62%	66,07%	66,41%	63,82%		
Specificity (full model specificity: 63,36%)	63,37%	63,74%	63,60%	64,10%		
Total variables selected	54/104	26/104	27/104	10/104		

APPENDIX 3. Variable importance and selection results by four methods (Prosper data)

Variable name	Method (abbreviated)				selection count	OOB error increase on permutation
	Altmann	Boruta	VSURF	GBS		
BorrowerRate	+	+	+	+	4	2,76%
ProsperGradeComb	+	+	+		3	1,02%
AvailableBankcardCredit	+	+	+		3	0,80%
CreditScoreRangeUpper	+	+	+	+	4	0,78%
MonthlyLoanPayment	+	+	+	+	4	0,76%
TotalInquiries	+	+	+		3	0,76%
ProsperScore	+	+	+		3	0,73%
CreditScoreRangeLower	+	+	+		3	0,72%
InquiriesLast6Months	+	+	+	+	4	0,63%
Occupation	+	+	+	+	4	0,59%
LoanOriginalAmount	+	+	+		3	0,49%
StatedMonthlyIncome	+	+	+	+	4	0,47%
BankcardUtilization	+	+	+	+	4	0,42%
DebtToIncomeRatio	+	+	+		3	0,38%
RevolvingCreditBalance	+	+	+		3	0,38%
OpenRevolvingMonthlyPayment	+	+	+		3	0,37%
TradesNeverDelinquent..percentage.	+	+	+		3	0,37%
CurrentCreditLines	+	+	+		3	0,29%
CurrentDelinquencies	+	+	+		3	0,29%
AmountDelinquent	+	+	+		3	0,29%
Investors	+	+	+		3	0,28%
Term	+	+	+	+	4	0,27%
OpenCreditLines	+	+	+		3	0,27%
OpenRevolvingAccounts	+	+	+		3	0,21%

TotalTrades	+	+	+		3	0,20%
TotalCreditLinespast7years	+	+	+		3	0,20%
IncomeRange	+	+	+		3	0,19%
OnTimeProsperPayments	+	+		+	3	0,15%
DelinquenciesLast7Years	+	+	+		3	0,14%
TradesOpenedLast6Months	+	+	+	+	4	0,14%
FirstRecordedCreditLineLength		+	+		2	0,14%
ListingCategory..numeric.	+	+	+	+	4	0,13%
EmploymentStatus	+	+			2	0,13%
ProsperPrincipalBorrowed	+	+			2	0,11%
TotalProsperPaymentsBilled	+	+			2	0,11%
BorrowerState	+	+	+	+	4	0,09%
IsBorrowerHomeowner	+			+	2	0,06%
ProsperPrincipalOutstanding		+		+	2	0,05%
PublicRecordsLast10Years					0	0,05%
TotalProsperLoans	+	+			2	0,03%
EmploymentStatusDuration					0	0,03%
Recommendations				+	1	0,01%
InvestmentFromFriendsCount					0	0,01%
InvestmentFromFriendsAmount					0	0,00%
ProsperPaymentsLessThanOneMonthLate					0	0,00%
ProsperPaymentsOneMonthPlusLate					0	0,00%
FundedAmountDiff					0	0,00%
IncomeVerifiable					0	0,00%
PublicRecordsLast12Months					0	0,00%
Overall 10-fold CV mean accuracy (full model accuracy: 66,67%)	66,62%	66,68%	66,21%	65,51%		

Sensitivity **(full model sensitivity: 72,23%)**	71,63%	71,87%	71,59%	70,65%
Specificity **(full model specificity: 64,33%)**	64,44%	64,33%	63,87%	63,18%
Total variables selected	37/49	38/49	32/49	15/49

APPENDIX 4. Variable importance and selection results by four methods (Zopa data)

Variable name	Method (abbreviated)				selection count	OOB error increase on permutation
	Altmann	Boruta	VSURF	GBS		
Lending.rate	+	+	+	+	4	15,16%
Original.Loan.Amount	+	+	+		3	5,06%
Term	+	+	+	+	4	4,74%
PostCode		+			1	0,07%
Overall 10-fold CV mean accuracy (full model accuracy: 70,79%)		70,85%	72,60%	73,03%		
Sensitivity (full model sensitivity: 75,51%)	same as for third method	75,52%	76,89%	73,60%		
Specificity (full model specificity: 70,11%)		70,16%	71,98%	72,95%		
Total variables selected	3/4	4/4	3/4	2/4		

APPENDIX 5. Data dictionary for Bondora dataset

The table below contains the original and also partially supplemented labels and definitions of used variables according to the Bondora P2P lending platform.

Original variable name	Description
Age	The age of the borrower when signing the loan application
Amount	Amount the borrower received on the Primary Market (in euro)
AmountOfPreviousLoansBeforeLoan	Value of previous loans
ApplicationSignedHour	The hour of the day when the loan application was signed
ApplicationSignedWeekday	The day of the week when the loan application was signed
AppliedAmountDiff (calculated column)	The difference between the originally requested and funded loan amount
City	City of the borrower
Country	Residency of the borrower
County	County of the borrower
CreditScoreComb (calculated column)	The combined credit score column that is comprised of the borrower's credit score as provided by a consumer credit rating agency in the corresponding country
DebtToIncome	Ratio of borrower's monthly gross income that goes toward paying loans
Education	Educational level of the borrower
EmploymentDurationCurrentEmployer	Employment time with the current employer
EmploymentPosition	Employment position with the current employer
EmploymentStatus	Employment status of the borrower
ExistingLiabilities	Borrower's number of existing liabilities
FreeCash	Discretionary income after monthly liabilities
Gender	Gender of the borrower
HomeOwnershipType	The type of home ownership supplied by borrower
IncomeFromChildSupport	Borrower's income from alimony payments
IncomeFromFamilyAllowance	Borrower's income from child support

IncomeFromLeavePay	Borrower's income from paternity leave
IncomeFromPension	Borrower's income from pension
IncomeFromPrincipalEmployer	Borrower's income from its employer
IncomeFromSocialWelfare	Borrower's income from social support
IncomeOther	Borrower's income from other sources
IncomeTotal	Borrower's total income
Interest	Maximum interest rate accepted in the loan application
LanguageCode	The language that was used in the loan application by borrower
LiabilitiesTotal	Total monthly liabilities
Loan status (calculated column)	Indicates whether the loan is considered as defaulted (based on the original loan status and the information on overdue period)
LoanDuration	Current loan duration in months
MaritalStatus	Marital status of the borrower
MonthlyPayment	Estimated amount the borrower has to pay every month
MonthlyPaymentDay	The day of the month the loan payments are scheduled (the actual date is adjusted for weekends and bank holidays)
NewCreditCustomer	Indicates whether the customer has prior credit history in Bondora
NoOfPreviousLoansBeforeLoan	Number of previous loans
NrOfDependants	Number of children or other dependants
OccupationArea	Borrower's occupation area
PreviousEarlyRepaymentsBeforeLoan	How much was the early repayment amount before the loan
PreviousEarlyRepaymentsCountBeforeLoan	How many times the borrower had repaid early
PreviousRepaymentsBeforeLoan	How much the borrower had repaid before the loan
Rating	Bondora custom rating issued by the rating model
RefinanceLiabilities	The total amount of liabilities after refinancing
UseOfLoan	The category of loan purpose supplied by borrower
VerificationType	Method used for loan application data verification
WorkExperience	Borrower's overall work experience in years

APPENDIX 6. Data dictionary for LendingClub dataset

The table below contains the original and also partially supplemented labels and definitions of used variables according to the LendingClub P2P lending platform.

Original variable name	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent
acc_open_past_24mths	Number of trades opened in past 24 months
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
amnt_diff (calculated column)	The difference between the originally requested and funded loan amount
annual_inc	The self-reported annual income provided by the borrower during registration
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts
chargeoff_within_12_mths	Number of charge-offs within 12 months
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent
disbursement_method	The method by which the borrower receives their loan
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income

earliest_cr_line_length (calculated column)	Amount of time since opening of the earliest credit line at time of application (in days)
emp_length	Employment length in years
emp_title (calculated column)	The job title supplied by the borrower when applying for the loan (reduced to include only 50 most frequently supplied job titles)
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to
funded_amnt	The total amount committed to that loan at that point in time
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	Indicates whether the loan was initially listed in the whole or fractional market
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates
int_rate	Interest rate on the loan
loan_status (calculated column)	Indicates whether the loan is considered as defaulted (based on the original loan status and the information on overdue period)
max_bal_bc	Maximum current balance owed on all revolving accounts
mo_sin_old_il_acct	Months since oldest installment account opened
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
mo_sin_rcnt_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts
mths_since_last_delinq	The number of months since the borrower's last delinquency
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record
mths_since_rcnt_il	Months since most recent installment accounts opened
mths_since_recent_bc	Months since most recent bankcard account opened

mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_inq	Months since most recent inquiry
mths_since_recent_revol_delinq	Months since most recent revolving delinquency
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
open_acc	The number of open credit lines in the borrower's credit file
open_acc_6m	Number of open trades in last 6 months
open_act_il	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
pct_tl_nvr_dlq	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request

revol_bal	Total credit revolving balance
revol_bal_joint	Sum of revolving credit balance of the co-borrowers, net of duplicate balances
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
sec_app_chargeoff_within_12_mths	Number of charge-offs within last 12 months at time of application for the secondary applicant
sec_app_collections_12_mths_ex_med	Number of collections within last 12 months excluding medical collections at time of application for the secondary applicant
sec_app_earliest_cr_line_length (calculated column)	Amount of time since opening of the earliest credit line at time of application for the secondary applicant (in days)
sec_app_fico_range_high	FICO range (low) for the secondary applicant
sec_app_fico_range_low	FICO range (high) for the secondary applicant
sec_app_inq_last_6mths	Credit inquiries in the last 6 months at time of application for the secondary applicant
sec_app_mort_acc	Number of mortgage accounts at time of application for the secondary applicant
sec_app_mths_since_last_major_derog	Months since most recent 90-day or worse rating at time of application for the secondary applicant
sec_app_num_rev_accts	Number of revolving accounts at time of application for the secondary applicant
sec_app_open_acc	Number of open trades at time of application for the secondary applicant
sec_app_open_act_il	Number of currently active installment trades at time of application for the secondary applicant
sec_app_revol_util	Ratio of total current balance to high credit/credit limit for all revolving accounts
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
term	Loan duration (in months)
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently in the borrower's credit file
total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment accounts
total_bc_limit	Total bankcard high credit/credit limit

total_cu_tl	Number of finance trades
total_il_high_credit_limit	Total installment high credit/credit limit
total_rev_hi_lim	Total revolving high credit/credit limit
verification_status	Indicates if the borrower's income was verified by LC
verification_status_joint	Indicates if the co-borrowers' joint income was verified by LC
zip_code (calculated column)	The first 3 numbers of the zip code provided by the borrower in the loan application (reduced to include only 50 most frequently supplied zip codes)

APPENDIX 7. Data dictionary for Prosper dataset

The table below contains the original and also partially supplemented labels and definitions of used variables according to the Prosper P2P lending platform.

Original variable name	Description
AmountDelinquent	Dollars delinquent at the time the credit profile was pulled
AvailableBankcardCredit	The total available credit via bank card at the time the credit profile was pulled
BankcardUtilization	The percentage of available revolving credit that is utilized at the time the credit profile was pulled
BorrowerRate	The borrower's interest rate for this loan
BorrowerState	The two letter abbreviation of the state of the address of the borrower at the time the Listing was created
CreditScoreRangeLower	The lower value representing the range of the borrower's credit score as provided by a consumer credit rating agency
CreditScoreRangeUpper	The upper value representing the range of the borrower's credit score as provided by a consumer credit rating agency
CurrentCreditLines	Number of current credit lines at the time the credit profile was pulled
CurrentDelinquencies	Number of accounts delinquent at the time the credit profile was pulled
DebtToIncomeRatio	The debt to income ratio of the borrower at the time the credit profile was pulled
DelinquenciesLast7Years	Number of delinquencies in the past 7 years at the time the credit profile was pulled
EmploymentStatus	The employment status of the borrower at the time they posted the listing
EmploymentStatusDuration	The length in months of the employment status at the time the listing was created
FirstRecordedCreditLineLength (calculated column)	The amount of time since the first credit line was opened (in days)
FundedAmountDiff (calculated column)	The difference between the originally requested and funded loan amount
IncomeRange	The income range of the borrower at the time the listing was created
IncomeVerifiable	The borrower indicated they have the required documentation to support their income
InquiriesLast6Months	Number of inquiries in the past six months at the time the credit profile was pulled
InvestmentFromFriendsAmount	Dollar amount of investments that were made by friends

InvestmentFromFriendsCount	Number of friends that made an investment in the loan
Investors	The number of investors that funded the loan
IsBorrowerHomeowner	Specifies whether or not the member has is a verified Homeowner at the time the listing was created
ListingCategory	The category of the listing that the borrower selected when posting their listing
LoanOriginalAmount	The origination amount of the loan
LoanStatus (calculated column)	Indicates whether the loan is considered as defaulted (based on the original loan status and the information on overdue period)
MonthlyLoanPayment	The scheduled monthly loan payment
Occupation (calculated column)	The occupation selected by the borrower at the time they created the listing (reduced to include only 50 most frequently supplied occupations)
OnTimeProsperPayments	Number of on time payments the borrower had made on Prosper loans at the time they created this listing
OpenCreditLines	Number of open credit lines at the time the credit profile was pulled
OpenRevolvingAccounts	Number of open revolving accounts at the time the credit profile was pulled
OpenRevolvingMonthlyPayment	Monthly payment on revolving accounts at the time the credit profile was pulled
ProsperGradeComb (calculated column)	The combined loan grade that is comprised of previously existing and the newly created custom Prosper rating (CreditGrade and numerical ProsperRating)
ProsperPaymentsLessThanOneMonthLate	Number of payments the borrower made on Prosper loans that were less than one month late at the time they created this listing
ProsperPaymentsOneMonthPlusLate	Number of payments the borrower made on Prosper loans that were greater than one month late at the time they created this listing
ProsperPrincipalBorrowed	Total principal borrowed on Prosper loans at the time the listing was created
ProsperPrincipalOutstanding	Principal outstanding on Prosper loans at the time the listing was created
ProsperScore	A custom risk score built using historical Prosper data. Applicable for loans originated after July 2009
PublicRecordsLast10Years	Number of public records in the past 10 years at the time the credit profile was pulled
PublicRecordsLast12Months	Number of public records in the past 12 months at the time the credit profile was pulled
Recommendations	Number of recommendations the borrower had at the time the listing was created
RevolvingCreditBalance	Dollars of revolving credit at the time the credit profile was pulled
StatedMonthlyIncome	The monthly income the borrower stated at the time the listing was created

Term	The length of the loan expressed in months
TotalCreditLinespast7years	Number of credit lines in the past seven years at the time the credit profile was pulled
TotalInquiries	Total number of inquiries at the time the credit profile was pulled
TotalProsperLoans	Number of Prosper loans the borrower at the time they created this listing
TotalProsperPaymentsBilled	Number of on time payments the borrower made on Prosper loans at the time they created this listing
TotalTrades	Number of trade lines ever opened at the time the credit profile was pulled
TradesNeverDelinquent	Percentage of trades that have never been delinquent at the time the credit profile was pulled
TradesOpenedLast6Months	Number of trades opened in the last 6 months at the time the credit profile was pulled

APPENDIX 8. Data dictionary for Zopa dataset

The table below contains the original and also partially supplemented labels and definitions of used variables according to the Zopa P2P lending platform.

Original variable name	**Description**
Original.Loan.Amount	The funded loan amount
Term	The length of the loan (in months)
Lending.rate	The lender yield on the loan (interest rate)
Latest.Status (calculated column)	Indicates whether the loan is considered as defaulted (based on the original loan status)
PostCode (calculated column)	The borrower's postal code (reduced to include only 50 most frequently supplied postal codes)