SAN: Default Probability Prediction

Valeriia Klimova, Tigran Oganesian, Jan Pikman, Juraj Žilt

December 2023

Contents

1 Introduction

This document presents our semestral project for the Statistical Data Analysis course. In the current chapter, we introduce our research question and dataset, we also give a quick overview of our work plan, risks, and limitations. The second chapter provides background for our dataset, which is important for every data analysis, and introduces the concept of peer-to-peer lending. The next chapter introduces some of the existing works which solved the same task with the same dataset. The fourth chapter covers initial predictor selection, data preprocessing, and analysis. The fifth chapter describes feature engineering and statistical models, which are then used to solve our task. The following chapter presents our results and interprets learned models. The final chapter describes the individual contributions of the authors.

1.1 Question and Data

The main goal of our assignment is to predict the probability of default when taking out a loan. Information about the borrower will be analyzed, including personal characteristics such as age, education, marital status, credit history, or work experience.

The dataset for the loan research is sourced from Bondora[1], a financial support company catering to individuals overlooked by traditional banks. This dataset can be quickly summarized as follows:

- Number of predictors: 112
- Number of samples: 344 021
- Dataset timespan: 2009.02.28 - 2023.11.28
- NaN values percentage: 28.65%

Because the Bandora dataset has already been used for default prediction several times before we will describe some of these approaches and if possible we will compare their results with ours.

1.2 Work Plan

In the following two subsections, we ask instrumental questions for our default prediction task. Later, we propose approaches how to answer them.

1.2.1 Specific Instrumental Questions

1. What does the dataset look like?
2. Are there any patterns to be found in the data that would allow for the categorization of borrowers?
3. Are there linear relationships between the probability of default and other features?
4. Are there any features that are insignificant and can be excluded?
5. Are there any other advanced modeling techniques viable for customer default prediction?
6. Do the results meet expectations?

1.2.2 Answering IQs

1. Using visualizations, we will try to depict information about borrowers and see the underlying patterns. The main goal will be to track the relationship between the probability of default and other features.
2. During the exploratory analysis pairwise relationship between the inspected predictor and the customer default will be examined.
3. It is useful to try to build a linear model using the techniques learned in the SAN course. We will use a cross-validation to evaluate performance and find an appropriate model.
4. We can use forward or backward stepwise selection to choose an optimal model. We can also try using, for example, the lasso method, which eliminates unnecessary features, if its

[1] `https://www.bondora.com/en/public-reports#dataset-file-format`

assumptions will be met. Throughout the experiments, collinearity and its impacts will be considered.

5. We can try using dimensionality reduction and clustering of borrowers into groups to predict default.
6. We will compare our results to results presented by others.

1.2.3 Risks and Limitations

- *Is there enough properly sampled data? Doesn't the dataset contain sampling bias?*

  One of the characteristics of statistical models is their ability to provide valuable modeling even with a small dataset in case of proper sampling. The data chosen for the assignment contain 334 thousand samples which should be enough and therefore the only problem can be a sampling bias. Sampling bias cannot be easily detectable, but overcoming it can be done by using the model just on the population where it was sampled.

- *How collinearity can affect a statistical model?*

  During the preprocessing phase, we will identify a situation where predictor variables are highly correlated. The presence of collinearity can adversely affect the performance and interpretability of statistical models.

- *What to do about the influence of confounding?*

  The occurrence of confounding will depend on the model design. It cannot be avoided but should be kept in mind.

1.3 Analysis Tools

In this project, we utilize Python as the primary programming language due to its extensive support for data analysis. The following key libraries have been instrumental in conducting various statistical analyses and generating insightful visualizations:

- *NumPy* library has been employed for numerical operations and array manipulations.
- *Pandas* library has been used for data manipulation and exploration of the dataset.
- *Matplotlib* and *Seaborn* visualization libraries have played a crucial role in creating clear and informative plots, aiding in the interpretation of statistical findings.
- *SciPy* library has been employed for statistical tests.
- *Scikit-learn* has been employed to implement predictive models and evaluate their performance.

Additionally, Jupyter Notebooks were used as the interactive computing environment, providing a collaborative and transparent platform for code development, analysis, and documentation.

2 Data background

Bondora is one of the European pioneers in peer-to-peer (P2P) lending, allowing small investors to fund the needs of customers currently in Estonia, Spain, Finland, and the Netherlands. It used to offer its services to Slovak citizens. Unfortunately, that initiative was signed off due to their high default rate.

2.1 Peer-to-peer lending industry

A known alternative in the Czech Republic is the Zonky platform. Such platforms often serve as an opportunity for people with worse credit scores to receive loans, but often for higher (in some cases loan-sharking) interest. Therefore, such platforms should not be seen as altruistic places but rather as regular businesses trying to fulfill the need for money in all categories of society with a mind to their own prosperity.

There are two reasons why P2P businesses are able to function in such risky fields of entrepreneurship:

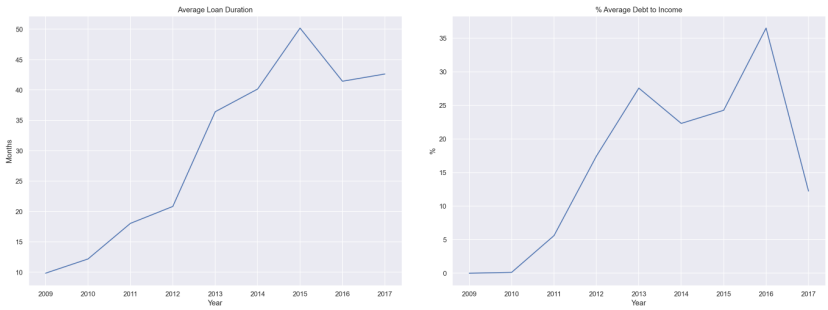

Figure 1: Change in average loan characteristics [2]

1. They do not have physical stores and do not provide personalized customer services with a vast number of employees, so their prose costs are lower than those of known big banks.
2. They do not use extensive amounts of expensive, trustworthy data as the banks do. Instead, they use cheaply available big data offered by many companies, which they can process by modern methods of machine learning to decide whether or not to provide the loan and if the loan is offered, at what interest rate.

   Such methods cannot be used in the banks because the banks need to be able to explain why the request for the loan was declined to overcome the state or organizational regulations of possible biases and unfairness. Therefore, machine learning is also a reason for multiple artificial intelligence enthusiasts to invest since they believe the decision whether or not to offer a loan is much more precise than the one used by the banks, influenced often by people's experience.

2.2 General overview and difficulties to consider

Important to mention is that the official open dataset contains data only about the applicants who received the loan. For that reason, it is not possible to predict the will to give or not to give the loan to the customers and compare it with their model and business rules. It is possible only to indicate the default probability of the customer, which can be seen as a potential risk classifier.

In reality, the Bondora's average investor has an internal rate of return (IRR) of -4% per annum [1], which suggests that the decision engines in Bondora are no correct.

The IRR strongly depends on the origin of the borrower. So far, the Estonian borrowers have an average IRR of 2.93%, Finland's -7.6%, Slovak's -15%, and Spanish's -22.6%. [1]

Two hypotheses come to mind when seeing such controversial differences in IRR .

1. Bondora does not see the lack of data about foreign customers as a barrier to lending money even when it is causing them negative income. They likely see it as an opportunity to spread across Europe, resulting in more significant income later. Negative IRR can be in short term compared to advertisement costs and initial investments.

   This line of thought is supported by the fact that Slovakia was eliminated, even though it did not have the worst IRR out of the countries where Bondora was operating. Most likely, they did not see further potential in Slovakia, and the expansion there was stopped.

2. Each country has its P2P companies, so when you decide to apply for a loan abroad, it means you are in need of money, and it implies alarming financial conditions.

Therefore, we will keep a close eye on the origins of the borrowers in our models so as not to cause a significant bias towards them.

Lastly, it is essential to mention that Bondora is a young and dynamically changing platform, resulting in significant changes in the chosen approach towards its customers. See Figure 1.

In the picture, we can see the characteristics of the average accepted loan by Bondora throughout the years. In the left image is the graph representing the average loan duration in months, and the right depicts the average debt to income in percentage. They might have transformed from

Rating	Number of loans	Average loan amount (EUR)	Standard deviation of loan amount	Average interest	Standard deviation of interest rate	Average expected return	Number of defaulted loans	Average default rate
AA	2,686	1,390	1,452	11.50%	4.50%	9.58%	255	9%
A	5,381	1,575	1,661	13.56%	4.68%	10.46%	566	11%
B	12,986	2,003	1,969	16.15%	3.91%	10.82%	1,740	13%
C	17,332	2,481	2,304	21.81%	3.94%	12.27%	3,073	18%
D	18,079	2,761	2,305	28.51%	3.98%	13.50%	4,623	26%
E	17,624	2,879	2,298	35.14%	4.19%	14.37%	5,727	32%
F	19,701	3,367	2,259	53.17%	11.30%	17.87%	9,710	49%
HR	13,799	1,750	1,516	77.24%	50.90%	15.52%	8,113	59%
Total	107,588	2,531	2,201	36.62%	27.72%	13.93%	33,807	31%

Figure 2: Statistics about the risk categories at Bondora dataset [1]

company borrowing short-term loans covering small expenditures with high interest to more mid-term loans with lower interest rates, allowing users with higher `DebtToIncome` to also receive money.[2]

3 Other Approaches to analyzing Bondora loan datasets

Due to the fact that some analyses of the Bondora Loan dataset already exist, one of the first steps of the exploration was to check other approaches and critically reflect on their decisions. The following subchapters discuss a few possible strategies and insights on the dataset.

3.1 Peer-to-peer lending: Legal loan sharking or altruistic investment? Analyzing platform investments from a credit risk perspective

The study comes from [1] and was inducted by the Department of Finance at Corvinus University of Budapest. The paper discusses the main P2P platform's functionalities and why it exists in the market from a financial perspective. The goal of the paper was to see whether using extensive alternative data of multiple predictors is measurably better than using a classic "banking-like" information dataset.

The number of predictors were reduced from 112 to 12 without further explanation. The remaining predictors were analyzed more closely, and two (`Country` and `CreditScoreAsEquifaxRisk`) were discussed thoroughly.

Firstly, common transformations were done, such as using natural logarithm on variables with high variance (such as income) or reducing the number of categories in some categorical predictors [3]. Later, to overcome the issue of biased data based on country, they transformed the categorical `Country` to continuous values by applying the Weight of Evidence (WOE) transformation.

$$WOE = ln(\frac{PortionOfGoods}{PortionOfBads}) \tag{1}$$

Secondly, an analysis of risk categories was shown, showing that risk categories assigned by Bondora categories describe the default classes correctly. See Figure 2.

In the end, six predictors [4] were inserted into the logistic regression after elimination of correlation within 12 previously selected features. There were two conclusions of the work:

1. No improvement in using additional data collected by Bondora over classic "banking-like" data was detected.
2. The threshold for accepting the loan applications should be higher than it is currently, since the model's risk categories are correct.

[2] It is just a hypothesis and the real reason for drastic changes was not found.

[3] E.g., `HomeOwnershipType` had ten categories originally, and later, they ended up with two - 1 meant owner and 0 was used otherwise

[4] Country, IncomeTotal, LiaiblitiesTotal, NoOfPreviousLoansBeforeLoan, HomeownershipType

3.2 Risk Prediction for Loan Applications By Machine Learning Algorithms

In data science and machine learning, diverse methodologies and approaches are employed to unravel insights from datasets. In this section, we explore two Kaggle[5] approaches from contributors, each providing a unique lens into dataset analysis and one bachelor thesis testing nine different machine learning methods. Rohan Chitale's[3] methodology extensively utilizes all 100 columns without specific feature selection, employing a Decision Tree Classification[6] model. In contrast, the second approach, presented by Levent Ozdemir[4], adopts a more targeted strategy, working with a total of 43 features, replacing remaining null values with the median, and employing the XGBClassifier[7] method.

3.2.1 Rohan Chitale

- **Data Preprocessing**
  - *Feature Utilization*

    In a notable feature utilization strategy, Chitale includes all available columns, even those with all `NaN` values.
  - *Handling NaN Values*

    `NaN` values in features are replaced with the median, which is a robust measure of central tendency. If the dataset contains outliers, using the median can provide a more representative measure of the "typical" value.
  - *Data Transformation*

    Chitale's approach to data transformation involves the use of `Z-score` normalization on the selected columns. It ensures that features are on a comparable scale, preventing any particular feature from dominating the analysis due to its magnitude.
- **Model Selection**
  - *Model*

    For classification, author opts for the `Decision Tree Classifier`. Decision Tree is a non-linear model that makes decisions based on a tree-like model of decisions and their possible consequences. This model can handle both numerical and categorical data.
  - *Dependent Variable*

    The dependent variable in this analysis is the `Status` column, representing the state of loans as `Late`, `Repaid`, or `Current`. Notably, records with a `Status` of `Current` are excluded from the analysis. This decision is grounded in the uncertainty surrounding the repayment status of current loans, making it a prudent choice to focus on loans that have either been repaid or are late.
- **Results**

  Given the exceptionally high `AUC` value of 99% obtained in Rohan Chitale's approach, it raises the possibility that certain parameters used in the model are strongly correlated with the dependent variable.
- **Revision**

  During both the training and prediction phases, it's observed that features such as `ActiveLateCategory`, `DefaultDate`, `PrincipalOverdueBySchedule`, `RecoveryStage`, were included. These features appear to be highly correlated with the `Status` and may not be known during the initiation or early stages of a loan. This could lead the model to learn the correlation of `Status` with data that is produced by such status, potentially compromising the model's generalizability and being leading to overfit.

[5] https://www.kaggle.com/
[6] https://scikit-learn.org/stable/modules/tree.html
[7] https://xgboost.readthedocs.io/en/stable/

3.2.2 Levent Ozdemir

- **Data Preprocessing**
  - *Feature Utilization*

    As part of the data preprocessing steps, author adopts a feature selection strategy by dropping columns with a significant number of missing values. Specifically, features with more than 50% missing values are excluded from the analysis. This results in a streamlined dataset, utilizing only 43 columns for further analysis.

  - *Handling NaN Values*

    To address missing values in the remaining dataset, Levent replaces `NaN` values with the median. This imputation strategy ensures that missing values are filled with a representative measure.

  - *Data Transformation*

    The author employs the `ColumnTransformer` from the `sklearn` library to perform specific transformations on different types of columns. Categorical columns transform using the `OneHotEncoder` technique, which is well-suited for converting categorical variables into a binary matrix representation. Numerical columns are transformed using `RobustScaler`. This scaling technique is robust to outliers, ensuring that the presence of extreme values does not unduly influence the transformation process.

- **Model Selection**
  - *Model*

    In terms of classification models, author opts for the `XGBoost Classifier` (XGBClassifier). It is an ensemble learning method based on the concept of boosting, which builds a series of weak learners (typically decision trees) sequentially, each one correcting the errors of its predecessor.

  - *Dependent Variable*

    Unlike the first approach, Ozdemir retains records with a `Status` column value of `Current` in the analysis. This decision is notable as it includes ongoing loans in the predictive model. The rationale behind this choice might involve considering the potential predictive value of including current loans in the analysis.

- **Results**

  The model's performance is assessed using the Area Under the Curve (AUC) metric, with a noteworthy `AUC` score of 99%.

- **Revision**

  Similar to the feedback for Rohan Chitale, it's worth examining the temporal aspects of the features used in his analysis. Some features may not be known during the initiation or early stages of a loan, and their inclusion could affect the model's real-world applicability. The classifier heavily relies on the `PrincipalBalance` feature to make decisions. This observation aligns with expectations, considering that a repaid loan typically has a zero `PrincipalBalance`.Therefore, it is a mistake to use this aspect in the analysis.

3.2.3 Sina Ansari Fard

This bachelor thesis [5] offers a closer look at nine machine learning methodologies. Unfortunately, after reading the paper, we concluded that the approach and data engineering were flawed. The values presented with 95% prediction accuracy should not be considered correct and easily achievable, as one might assume after reading [5]. Following the subchapter, the concerns will be explained more closely.

At the beginning of the thesis, irrelevant columns such as name and `LoanId` were removed, as well as columns with `NaN` values. Later, categorical values were assigned numeric values, and the first concern came with the `Status` column, where all active loans of status Late were assigned to one category, and already paid loans or the ones that were currently active without delay were in the second group. There were two main problems with such approach:

1. The case that the loan is paid does not mean it was not delayed or required additional time to extort money. So seeing it as "positive" predictor can be misleading for a model.
2. Currently active loans were included in so-called "safe" loans without considering the length of the account's existence. It means applications currently active for a few months were considered safe even though we know nothing about their future. In reality, most delays do not happen in the first months. It can be expected, that many loans currently assigned to a safe group can in the next model trained end up in the second group easily.

Later, the columns were also reduced based on the correlogram, resulting in 13 columns.

After that, the data were inserted into all the methods, and it was said that results will be analyzed not just with accuracy but also with precision, recall, and F1 score. The best performing methods were Gradient boosting and Random forest, both achieving an astonishing 95% accuracy. Unfortunately, the data contained columns such as interest rate or probability of default (which are highly correlated, but also both are columns added by Bondora engines and so the researcher of the thesis was using data generated from someone else's model, resulting probably in high dependenci on those predictors).

In reality, AUC, precision, recall, and F1 were shown, but nothing was said about them. Even more, the conclusion was short and did not mention any reasonable statement based on the achieved results. Overall, the thesis was poorly written. No equations or closer explanations of the methods used were done, and it was not mentioned why some methods would work better than others. Therefore, this work will not be compared with ours.

3.2.4 Summary

As we can see, other methodologies focus more on machine learning techniques than the statistical methods like lasso or ridge linear regression used in introductory courses of statistical learning. No lasso/ridge or dimension reduction was used to eliminate parameters. Instead, human perception and insights with simple statistics were applied. It allows us to test more technical approaches and explore possible outcomes of these techniques.

4 Exploratory Data Analysis

In this section, we detail the initial exploration and preprocessing of the Bandora dataset that consists of 344 021 rows and 112 columns. Firstly, we will describe the principles according to which we are selecting columns for further use. Then the rows are inspected w.r.t remaining columns for unknown or missing values. Finally, we provide an overview of the values present in a preprocessed dataset while pointing out some observations which we consider interesting.

4.1 Data Preprocessing

The first thing we noticed during the initial data inspection is that four columns, respectively `DateOfBirth`, `County`, `City`, and `EmploymentPosition`, are completely filled with `NaN` values hence these columns were not included in our dataset. Furthermore, no predictor explicitly states whether the loan defaulted or not and thus we had to create the binary `Default` column by combining the `DefaultDate` and `Status` columns. The `DefaultDate` column stores the date when the loan has defaulted if the loan has not defaulted its entry is `NaN`. This information by itself could be used to create the `Default` predictor but we have to keep in mind that our dataset also includes currently active loans which might still become defaulted. Therefore, we also use the `Status` column which has three possible values: `Late`, `Current`, and `Repaid`, to filter out rows representing loans that have not yet defaulted (status is `Late` or `Current`). This reduces the number of rows to 215 648, out of which 102 917 has been successfully repaid and 112 731 has defaulted.

The remaining columns were further filtered according to their meaning which is explained in the table on the Bandora website[8]. Some of the columns were omitted because their description was too technical and it would be difficult to interpret their true meaning. But most of the removed columns were deleted because their entry values could not be known before the loan was issued and therefore such columns cannot be used for our task of default prediction. The number of remaining features after this step is 22 and they are detailed in Appendix A.

[8] `https://www.bondora.com/en/public-reports#dataset-file-format`

Column	NaN count	NaN percentage	Undefined count	Undefined percentage
Age	0	0.00	0	0.00
AmountOfPreviousLoansBeforeLoan	8	0.00	0	0.00
AppliedAmount	0	0.00	0	0.00
Country	0	0.00	0	0.00
DebtToIncome	50	0.02	0	0.00
Education	50	0.02	2 476	1.15
EmploymentDurationCurrentEmployer	3 222	1.49	0	0.00
EmploymentStatus	202	0.09	**179 888**	**83.42**
ExistingLiabilities	0	0.00	0	0.00
FreeCash	50	0.02	0	0.00
Gender	45	0.02	0	0.00
HomeOwnershipType	1 657	0.08	3	0.00
IncomeTotal	0	0.00	0	0.00
LiabilitiesTotal	0	0.00	0	0.00
LoanDuration	0	0.00	0	0.00
MaritalStatus	50	0.02	**179 864**	**83.41**
NewCreditCustomer	0	0.00	0	0.00
NoOfPreviousLoansBeforeLoan	8	0.00	0	0.00
NrOfDependants	**180 824**	**83.85**	0	0.00
OccupationArea	91	0.04	**179 918**	**83.43**
UseOfLoan	0	0.00	**179 856**	**83.40**
WorkExperience	**179 922**	**83.43**	0	0.00

Table 1: Counts and proportions of NaN and unknown values for each column

Next, we inspected the remaining dataset of 215 648 rows and 22 columns for NaN and unknown values. The overview of this inspection is presented in Table 1 from which can be seen that the six predictors, namely: EmploymentStatus, MaritalStatus, NrOfDependants, OccupationArea, UseOfLoan, and WorkExperience, are mostly empty. Since, we consider these columns to be interesting for further analysis and the methods that we are going to use have relatively low data demand we have decided to remove rows that have missing values in these columns. This further reduces our dataset to 34 745 rows where 15 198 loans were successfully repaid and 19 547 loans resulted in default. The remaining missing cells of numerical predictors were filled by the median of the corresponding column. The same was done with the mode for categorical columns.

4.2 Summary Statistics and Observations

After the preprocessing, we computed simple statistical properties for each of the selected columns, which are shown in Table 2. Discrete and continuous columns include minimum, mean, standard deviation, and maximum, while nominal and ordinal columns display only mode. It can be seen that the standard deviation of many predictors is very large. This fact combined with enormous maximum values suggests that the dataset might contain outliers with high values in columns.

We have also analyzed the direct relationship between the individual columns and Default. This analysis was done using histograms or possibly bar graphs, depending on the data type of the analyzed column. We compared the shape of histograms (or bar graphs) produced by repaid and defaulted loans. Because this analysis was done for each predictor, in this text we provide only one example which deals with Age column that is shown in Figure 3. From this figure, we see that borrowers aged between 23 and 35 years are more likely to successfully repay the loan than otherwise while the older borrowers are slightly more likely to default. Following this analysis for each column, we can state that there is no "magical" predictor that would by itself be able to almost perfectly predict the occurrence of default.

Interesting to note is that the original dataset included loans dated from 2009-02-28 to 2023-11-30 but after the removal of rows with missing entries for the above-mentioned six predictors only the loans from 2009-02-28 to 2017-06-02 remained. This coincides with the General Data Protection Regulation (GDPR) which was approved on 2016-04-14 and implemented on 2018-05-25. We speculate that the GDPR is the reason for the lack of data entries which goes from 2017-06-02 onward.

Column	Type	MIN	MEAN ± SD MODE	MAX
Age	discrete	18	37.85 ± 11.37	77
AmountOfPreviousLoansBeforeLoan	continuous	0.00	1 190.85 ± 2 467.87	30 000.00
AppliedAmount	continuous	31.96	2 875.02 ± 2520.76	10 630.00
Country	nominal	-	Estonia	-
DebtToIncome	continuous	0.00	28.25 ± 19.36	198.02
Education	ordinal	-	Secondary	-
EmploymentDurationCurrentEmployer	ordinal	-	MoreThan5Years	-
EmploymentStatus	nominal	-	FullyEmployed	-
ExistingLiabilities	discrete	0	4.43 ± 3.47	36
FreeCash	continuous	-2 332.00	452.95 ± 1 300.08	158 748.64
Gender	nominal	-	Male	-
HomeOwnershipType	nominal	-	Owner	-
IncomeTotal	continuous	0.00	1 920.84 ± 3 589.46	228 550.00
LiabilitiesTotal	continuous	0.00	813.75 ± 1 299.91	172 510
LoanDuration	discrete	1	41.16 ± 19.02	60
MaritalStatus	nominal	-	Single	-
NewCreditCustomer	nominal	-	True	-
NoOfPreviousLoansBeforeLoan	discrete	0	0.72 ± 1.52	24
NrOfDependants	ordinal	-	0	-
OccupationArea	nominal	-	Other	-
UseOfLoan	nominal	-	Other	-
WorkExperience	ordinal	-	15To25Years	-

Table 2: Simple statistical overview of selected columns displaying minimum, mean with standard deviation, and maximum. For discrete and nominal columns only mode is provided

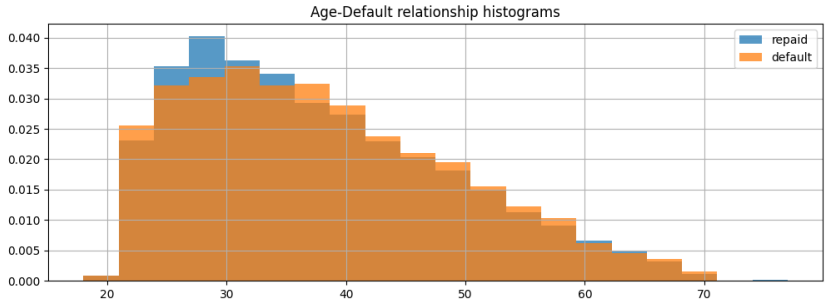

Figure 3: Superimposition of histograms based on the age predictor for both the repaid and defaulted loans.

5 Model Construction

5.1 Feature Engineering

In order to build an efficient and interpretable prediction model, an important step in our analysis is the selection of independent parameters. In this section, we use two approaches that take into account the nature of our data: the correlation matrix for numerical attributes and the Chi-square test for categorical attributes. These methods aim to identify relevant traits while mitigating multicollinearity problems and detecting statistically significant associations between categorical predictors and the target variable. The resulting subset of attributes derived from a thorough examination of numerical and categorical attributes is the basis for subsequent stages of model development.

5.1.1 Numerical Features

To identify relationships between numerical features, a Pearson correlation matrix was used. It provides insights into the linear relationships between pairs of numerical variables. Specifically, the correlation coefficient was calculated for each pair of numerical features, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation.

Features exhibiting strong correlations may offer redundant information, and as part of the feature selection process, consideration was given to removing one of the features from highly correlated pairs.

Before undertaking feature selection, an initial correlation matrix, which is shown in Figure 4, was computed for all numerical features in the dataset. This initial correlation matrix served as a baseline reference.

An examination of the initial correlation matrix unveiled notable relationships among features. Some instances of strong correlation were identified, including:

- `NewCreditCustomer` and `NoOfPreviousLoansBeforeLoan` and `AmountOfPreviousLoansBeforeLoan`:

  A decision was made to exclude `NoOfPreviousLoansBeforeLoan` and `AmountOfPreviousLoansBeforeLoan` in favor of retaining only the informative `NewCreditCustomer`. `NewCreditCustomer` is the representative feature, given its significance in capturing relevant information related to the customer's credit history.

- `ExistingLiabilities` and `DebtToIncome`:

  It was also agreed to prioritize `DebtToIncome` over `ExistingLiabilities` and consequently exclude the latter. This strategic choice stems from the understanding that `DebtToIncome` encapsulates essential information related to the borrower's financial situation, offering a comprehensive indicator of the ability to manage existing debts in relation to income.

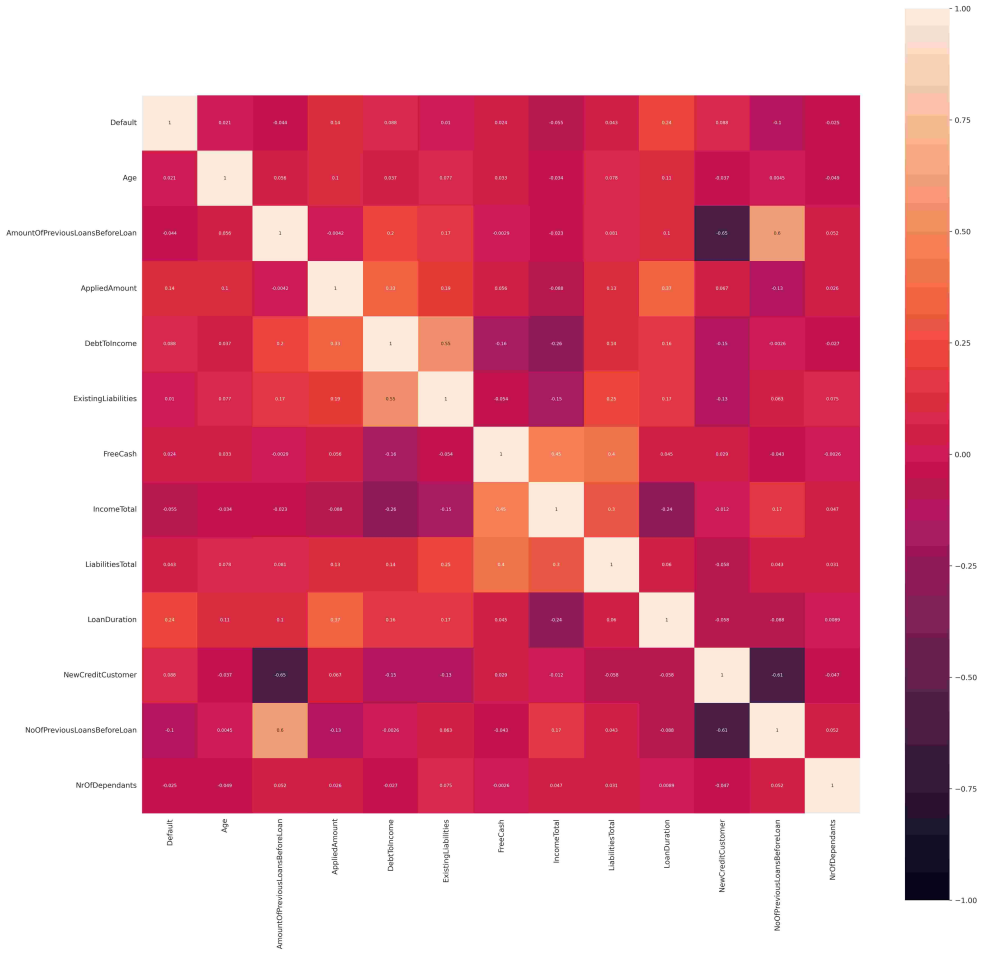

Figure 4: Correlation matrix before feature selection

After removing the columns named above, the resultant correlation matrix looks as in Figure 5.

In the end, based on the Table 2 logarithm was applied to the `Age`, `AppliedAmount`, `DebtToIncome`, `IncomeTotal`, `LiabilitiesTotal` and also `FreeCash`. Plus one was applied to each feature be-

fore the logarithmic transformation [9] apart from `FreeCash`, which contained also negative values. With `FreeCash`, we did not add just one, but also the absolute value of the minimum of the feature `FreeCash`. The logarithmic approach helped us to reduce the impact of outliers and feature variance, which was, in some cases, higher than the mean.

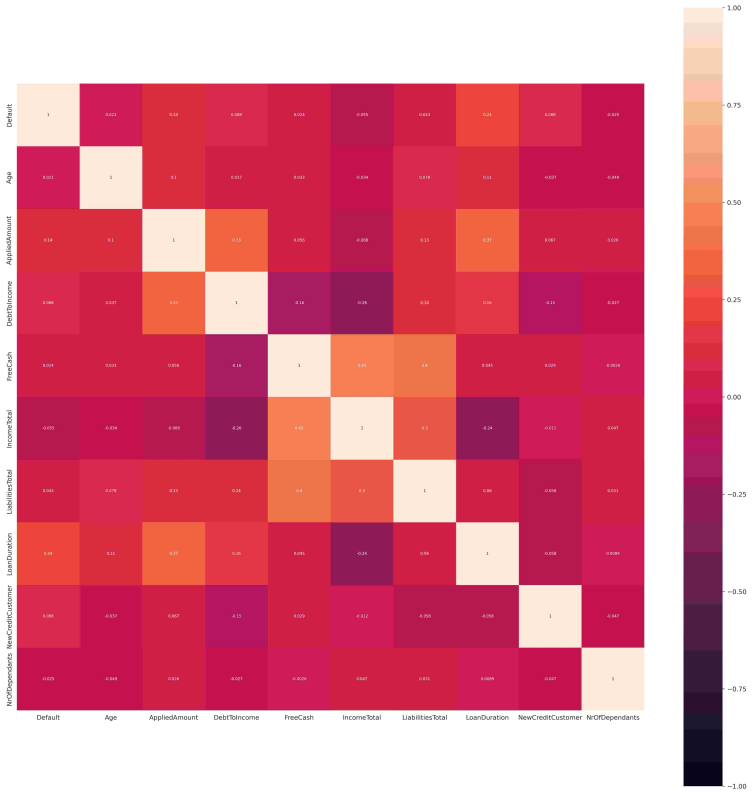

Figure 5: Correlation matrix after feature selection

5.1.2 Categorical Features

To evaluate the association between pairs of categorical features, we will use the Cramér's V coefficient that is normalized based on the number of observations, making it less dependent on the sample size. In contrast, the Chi-squared test statistic is influenced by the sample size, and larger datasets may lead to statistically significant results even for small effect sizes.

With a large number of categories in a column, there is an increased risk of finding statistically significant associations purely by chance (Type I errors). Reducing categories helps control this risk and ensures that identified associations are more likely to be genuine. These reductions, based on preprocessing and merging by meaning, affected the following columns:

- `Education`

  This variable was streamlined to four broad categories: *Secondary*, *Higher*, *Basic*, and *Vocational*. The decision to combine the *Primary* and *Basic* categories was driven by the need for a simpler representation of educational attainment to facilitate clarity and interpretability of subsequent analyses.

- `EmploymentStatus`

  The revised column now comprises three overarching categories: *Unemployed*, *Self-Employed*, and *Employed*. In our modified representation, we have condensed the categories *Partially Employed* and *Fully employed* into a unified *Employed* category.

  Additionally, the *Retiree* category has been merged with *Unemployed* to form a singular category, recognizing the relatively smaller proportion of retirees within the dataset.

- `HomeOwnershipType`

[9] to overcome undefined zero and minus infinity issueas

We have consolidated related categories to form three overarching groups: *Owner*, *Tenant*, and *Other*. The category *Joint Ownership* has been seamlessly merged with *Owner*.

The categories *Tenant Pre-Furnished Property*, *Tenant Unfurnished Property*, and *Mortgage* have been merged into a singular *Tenant* category. This consolidation is driven by a recognition of the commonality in housing arrangements for individuals falling under these classifications. The remaining categories with relatively smaller representation have been encompassed within the comprehensive *Other* category.

- `MaritalStatus`

  We have simplified the marital statuses into two comprehensive groups: *Cohabitant* and *Single*. Notably, the category *Married* has been seamlessly merged with *Cohabitant*, recognizing the shared characteristics and relational dynamics between these two statuses. The categories *Widow* and *Divorced* have been merged into the *Single* category.

- `OccupationArea`

  We have grouped related occupational areas into four overarching categories: *Industry*, *Commerce*, *Service*, and *Other*. The categories *Mining*, *Processing*, *Energy*, *Utilities*, *Construction*, and *Agriculture, forestry, and fishing* have been seamlessly merged into the comprehensive *Industry* group, acknowledging the commonalities in these sectors.

  Similarly, *Retail and wholesale* and *Hospitality and catering* have been thoughtfully combined into the more encompassing *Commerce* group, recognizing the shared characteristics of these occupational domains.

  The diverse occupational areas encompassing *Transport and warehousing*, *Info and telecom*, *Finance and insurance*, *Real estate*, *Research*, *Administrative*, *Civil service and military*, *Education*, and *Healthcare and social help* have been consolidated into the *Service* group. This combination from smaller sample sizes in specific subgroups while maintaining key information on occupational diversity in the service sector.

  Finally, the category *Art and entertainment* has been merged into the overarching *Other* category, recognizing its unique characteristics within a broader context.

- `UseOfLoan`

  The loan utilization purposes have been grouped into three comprehensive categories: *Financial*, *Personal*, and *Business*.

  Specifically, the categories *Loan consolidation*, *Real estate*, *Home improvement*, and *Business* have been seamlessly merged into the more encompassing *Financial* category, reflecting shared financial objectives within these loan utilization purposes.

  Simultaneously, the diverse loan purposes encompassing *Education*, *Travel*, *Vehicle*, *Other*, and *Health* have been thoughtfully regrouped into the *Personal* category. This consolidation captures the individual and personal nature of these loan utilization objectives.

  Finally, the remaining categories have been streamlined into the more encompassing *Business* category.

- `WorkExperience`

  The work experience categories have been restructured into three overarching groups: *MoreThan15Years*, *5To15Years*, and *LessThan5Years*.

  The categories *15To25Years* and *MoreThan25Years* have been seamlessly merged into the *MoreThan15Years* category, reflecting the overarching group of individuals with extensive work experience.

  Simultaneously, the *5To10Years* and *10To15Years* categories have been thoughtfully regrouped into the *5To15Years* category, capturing a range that encompasses mid-level work experience.

  Finally, the categories *2To5Years* and *LessThan2Years* have been streamlined into the *LessThan5Years* category, recognizing the shared characteristics of individuals with relatively shorter work experience.

- `EmploymentDurationCurrentEmployer`

  The employment duration categories have been restructured into three overarching groups: *MoreThan5Years*, *LessThan3Years*, and *3To5Years*.

  Simultaneously, the categories *TrialPeriod*, *UpTo1Year*, *UpTo2Years* and *UpTo3Years* have been regrouped into the *LessThan3Years* category, reflecting the overarching group of individuals with less than three years of current employment.

  Finally, the categories *UpTo4Years* and *UpTo5Years* have been seamlessly merged into the *3To5Years* category, capturing individuals with a current employment duration ranging from three to five years.

- `Gender`

  The Bondora dataset contains three gender categories *Male*, *Female*, and 1909 out of the remaining 34K rows were *Undefined*. Since there are just two genders and there was no description of why someone is undefined, it was decided to replace the missing values by random sampling with gender probabilities based on the overall dataset.

By analyzing pairs of columns with reduced number of categories, the coefficients depicted in Figure 6 were computed. The results show that no significant relationships were found, so no columns were removed after this phase of feature engineering.

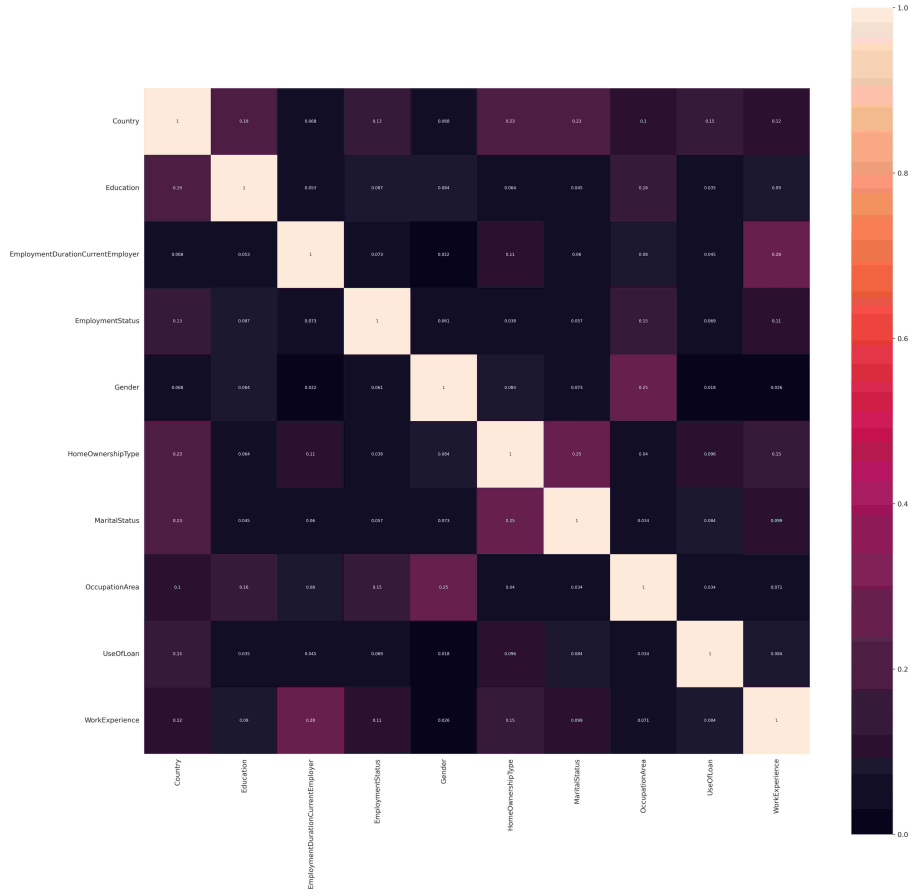

Figure 6: Cramér's V coefficients

Later on, all categorical features were transformed into dummy variables for further modeling since multiple models can not work with categorical values. Dummy transformation adds n-1 new features into a model, where n equals to a number of categories, and the original feature is removed from the dataset. It resulted in 10 new features, so the dataset has 29 features at the current stage.

The dummy transformation is not always necessary. For example, previously mentioned `WorkExperience` can be replaced with integers or ordinal numbers, indicating that a growing number of years can increase/decrease(based on the assigned coefficient) the probability of the predicted value. Based on

preprocessing, we have seen that the likelihood of default fluctuates across the categories. Therefore, the dummy transformation was preferred instead of using some polynomial transformations on assigned values to particular categories, which would decrease the interpretability of the dataset.

The previous correlation matrix and Cramer's V coefficients were computed separately upon numerical and categorical features. After adding the dummy variables, correlation can also be checked between categorical and numerical features. See Figure 7. Based on the results, it was dedicated to drop the newly created dummy variable `MoreThan15Years` since it was highly correlated with `Age`. Therefore, the final dataset has 28 features.

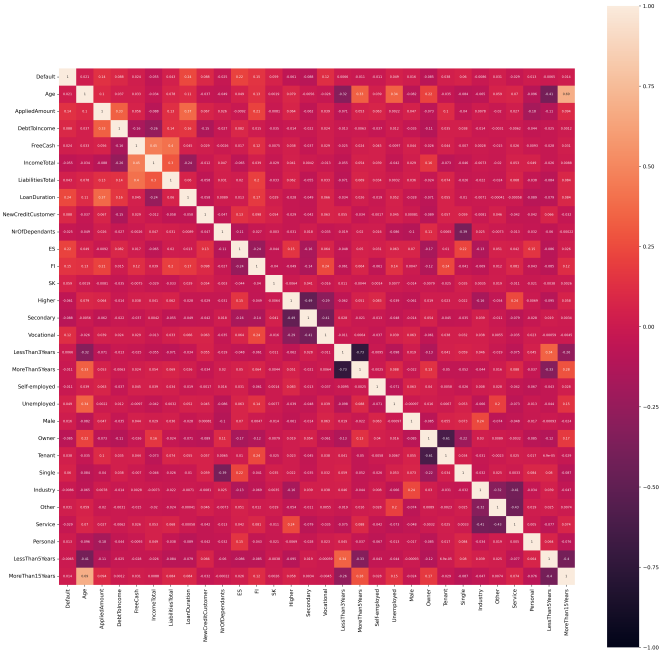

Figure 7: Correlation between all features

It is interesting to point out that there are squares on the diagonal of 7. The squares are correlated but do not need to be excluded from the dataset since those rows/columns are newly created dummy variables, and when modeling, each customer will have for sure maximally[10] just one value.

5.2 Statistical Methods

In this section, we outline the statistical techniques selected for the analysis, drawing upon methodologies acquired from the Statistical Data Analysis (SAN) course. The chosen techniques encompass a diverse range, each tailored to address specific characteristics and patterns within the dataset.

Our goal is to develop a predictive model that can efficiently estimate the probability of default, offering valuable insights into the potential risks associated with a particular set of characteristics.

The binomial distribution, inherent to logistic regression, aligns seamlessly with our objective. Because logistic regression is designed for binary classification tasks, where outcomes are characterized by two distinct possibilities, such as default or non-default in our current context.

[10]Since dummy transformation makes n-1 features, in case of the nth type of customer all three predictors will have value zero.

5.2.1 Logistic Regression

To analyze the probability of default, we began by applying logistic regression, incorporating all available features into the model. However, in order to account for nonlinear relationships between the features and the target variable, we experimented with applying polynomial transformations. This allowed us to capture more complex interactions between variables, thereby enhancing the model's ability to predict the probability of default. The results of this stage of analysis will be valuable for a more accurate prediction and understanding of the factors influencing credit portfolio risks.

5.2.2 Feature Selection

Two popular methods for feature selection are forward and backward stepwise regression, which iteratively add or remove features based on their contribution to the model. These techniques are particularly useful when dealing with datasets containing numerous potential predictors.

- *Forward Stepwise Selection*

  Forward stepwise selection begins with an empty model and iteratively adds the most significant predictor variable at each step. The process continues until a predefined criterion, such as a statistical test or information criterion, indicates that further additions do not significantly improve the model's fit.

- *Backward Stepwise Selection*

  Contrary to forward stepwise selection, backward stepwise selection begins with a model containing all available predictors and iteratively removes the least significant ones.

Among the feature selection methods, we opted for forward stepwise selection. This method is particularly valuable for its ability to efficiently navigate through the feature space, making it suitable for large datasets with numerous potential predictors.

Following the feature selection step, we proceeded to apply logistic regression. This allowed us to assess the model's performance with the subset of selected features. Subsequently, we revisited logistic regression, including the application of polynomial transformations to capture any non-linear patterns among the variables.

The combination of feature selection and the application of logistic regression with and without transformations forms a comprehensive approach, balancing model simplicity and predictive power. This iterative process ensures that our final model is both interpretable and capable of capturing the nuanced relationships influencing the likelihood of default.

5.2.3 Shrinkage Methods

In our pursuit of refining the predictive model for default probability, we also explored the application of shrinkage methods, including Lasso and Ridge regression. However, despite the potential benefits of these regularization techniques in handling multicollinearity and reducing overfitting, their application did not yield improvements in model performance.

One possible reason for this outcome is that the dataset may not have exhibited severe multicollinearity issues, which are situations where predictor variables are highly correlated. Shrinkage methods like Lasso and Ridge are particularly effective in mitigating multicollinearity, and their impact is more pronounced when this issue is prevalent. In cases where multicollinearity is not a significant concern, the regularization provided by these methods might not substantially enhance the model's predictive capabilities.

5.2.4 Support Vectors Machine

In our pursuit of optimizing the predictive model for default probability, we also experimented with the application of Support Vector Machine (SVM). Despite its effectiveness in handling complex relationships and non-linearities, the use of SVM did not result in a discernibly improved model in our specific context.

Several factors could contribute to this outcome. SVMs may thrive in scenarios where the decision boundary is intricate and not easily captured by linear models. However, our dataset might not exhibit the complexities that make SVM particularly advantageous. Additionally, SVM performance is influenced by the choice of hyperparameters and the kernel function. Suboptimal selection of

these parameters may hinder the model's ability to effectively capture the underlying patterns in the data.

While SVM is a powerful algorithm, its success often depends on the characteristics of the dataset and careful tuning of hyperparameters. In our case, the iterative nature of our modeling approach allows us to explore a variety of methods, acknowledging that the effectiveness of each algorithm can vary depending on the unique features of the data at hand.

5.2.5 Linear and Quadratic Discriminant Analysis

Discriminant analysis is employed when the outcome variable is categorical. Linear Discriminant Analysis (LDA) identifies linear combinations of predictors that best differentiate between categories by assuming that the categories have a normal distribution and equal variance. This results in a linear decision boundary which gives LDA its name. Quadratic Discriminant Analysis (QDA) is almost the same as LDA but the variance is considered to be unique for each class which increases the number of learned parameters and results in a quadratic decision boundary. The parameter increase also results in higher data demand compared to the LDA method.

Thanks to the size of our dataset the LDA and even the QDA method can be applied. On the other hand, the real-life application of either method would be problematic because the model has to be completely recomputed every time new samples are added which is not ideal for use in the banking sector.

6 Results

6.1 Main Findings

In Table 3, it is possible to see achieved results by using various statistical methods on the same pre-engineered dataset. The accuracies and AUC values did not differ much and did not receive astonishing results. Still, the dataset holds data about people and their finances, two highly problematic areas to study and predict. Especially in the P2P bubble, customers often come with difficult situations and backgrounds. Even for two almost identical persons with the same education level, age, origin, and employment status, the financial habits differ significantly. Therefore, it is not possible to expect higher accuracy rates. More in 6.2.4

After applying logarithms and dummy predictors, the accuracy has improved just slightly, but the main impact of the feature engineering was the variance reduction, which fell under 1%.

Used model	Accuracy	AUC
Logistic regression	67.73 ± 0.83	0.7420
Logistic regression - Lasso	67.80	72.92
Logistic regression - NoCountry	64.08	68.86
Linear regression - Ridge	68.75	0.7419
Linear regression - Lasso	68.38	0.7377
Logistic regression - Forward Stepwise Selection	67.94 ± 0.83	0.7274
Logistic regression - Backward Stepwise Selection	64.08 ± 0.7	0.6886
Linear Discriminant Analysis	69.03 ± 0.0748	0.7331 ± 0.0834
Quadratic Discriminant Analysis	66.33 ± 0.0640	0.7175 ± 0.0697
Radial Basis Function Support Vector Machine	64.04	0.7485

Table 3: Comparison of results produced by different models. Both the accuracies (%) and the AUC values are presented as means and standard deviations of values produced by 10-fold cross-validation

6.2 Interpretation

The following sections will present the results of all approaches more closely. For each approach, 10-fold cross-validation was used to determine the expected accuracy in real-life applications. Since the data has more than 34K of rows, k-fold was not necessary (variance under 1%) and regular 80/20 train-test set would be enough. But due to initially various datasets, it was decided to use 10-fold cross-validation as a "more robust metric."

6.2.1 Logistic Regression

Basic logistic regression with all features already provides results identical to the models after reduction thanks to optimizers built in the modern libraries. Unfortunately, multiple p-values are way above the 5% threshold. Therefore, we will focus on discussing the possible solutions and approaches to feature reduction while keeping the accuracy rate as high as possible in the following chapters. Furthermore, by decreasing the number of features, the interpretability of the model will also be increased. Result 67.73% upon unreduced dataset will be seen as a baseline for further modeling.

6.2.2 Lasso and Ridge

With Lasso and Ridge, it is necessary to cross-validate alphas, penalizing the number of predictors in use. We have decided to test smaller alphas to use still the variety of predictors, which is crucial for P2P decision schemas. The Table 4 is ordered based on the assigned coefficients(just Lasso the orders are diferent for Lasso and Ridge) to the predictors. The biggest influence have education in both methods [11] as well as origin of the applicant[12].

Used model	Lasso	Ridge
Higher	−0.0395	−0.0625
Secondary	−0.0206	−0.0428
Service	−0.0079	−0.0120
MoreThan5Years	−0.0039	−0.0125
Other	0.0	0.0015
Industry	0.0	0.0069
Single	0.0	0.0033
Tenant	0.0	−0.0099
Owner	0.0	−0.0034
Male	0.0	−0.0060
Unemployed	0.0	−0.0015
Self-employed	0.0	0.0028
LessThan3Years	0.0	−0.0000
Vocational	0.0	−0.0122
Age	0.0	−0.0027
NrOfDependants	0.0	0.0048
LiabilitiesTotal	0.0	−0.0212
FreeCash	0.0	−0.0080
DebtToIncome	0.0	0.0120
Personal	0.0	−0.0005
LessThan5Years	0.0031	0.0106
IncomeTotal	0.0032	0.0223
NewCreditCustomer	0.0119	0.0163
AppliedAmount	0.0139	0.0241
SK	0.0250	0.0365
FI	0.0699	0.0851
LoanDuration	0.0847	0.0963
ES	0.1150	0.1322

Table 4: Comparison of results produced by Lasso and Ridge shrinkage methods.

The Lasso excluded 16 predictors out of the original 28, and the Ridge kept 14 predictors under 0.01 coefficient, which then probably had a negligible effect on the final result. The coefficients of Ridge are slightly higher than the ones of Lasso. This is because keeping all the predictors in the models brings some noise into it, and therefore, main Ridge coefficients are forced to minimize the others by maximizing themselves.

Both methods have almost identical results, so further discussion about the preferred variance in the Bondora dataset does not make sense. Still, due to small colinearities between the predictors, Lasso might be a preferred approach in feature reduction.

[11] `Higher` and `Secondary`
[12] `SK`, `FI` and `ES`

6.2.3 Stepwise Selection

Forward stepwise selection was no different from the Lasso and Ridge regression in 6.2.2 and ended up with nine features. 8 out of which were from the top or the bottom of the Table 4 of the Lasso column (which is ordered based on assigned coefficients) and just `LoanDuration` differs. It suggests to trust in selected features. The truthfulness is also supported by p-values, which test H0 hypotheses of zero coefficients, all of which were rejected. See Figure 8.

Logit Regression Results

Dep. Variable:	Default	No. Observations:	27796
Model:	Logit	Df Residuals:	27787
Method:	MLE	Df Model:	8
Date:	Sun, 31 Dec 2023	Pseudo R-squ.:	0.1112
Time:	14:39:01	Log-Likelihood:	-16943.
converged:	True	LL-Null:	-19062.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
AppliedAmount	-0.2202	0.013	-17.009	0.000	-0.246	-0.195
LoanDuration	0.0228	0.001	29.552	0.000	0.021	0.024
ES	1.5099	0.037	40.497	0.000	1.437	1.583
FI	1.0232	0.038	26.658	0.000	0.948	1.098
SK	2.4504	0.243	10.064	0.000	1.973	2.928
Higher	-0.6552	0.036	-18.121	0.000	-0.726	-0.584
Secondary	-0.4147	0.031	-13.366	0.000	-0.476	-0.354
Unemployed	-0.1518	0.065	-2.345	0.019	-0.279	-0.025
Service	-0.1934	0.028	-6.793	0.000	-0.249	-0.138

Figure 8: P-values of remaining features after the forward stepwise selection.

The shrunk features were inserted into the logistic regression, and the results were 68%, identical to the ones achieved by logistic regression with all features.

On the other hand, backward selection differs more from the previously selected features. `NrOfDependants` and `Age` were selected resulting in 9 features. But still, the majority of predictors remained, and after the introduction into the model, an accuracy of 64.1% was achieved again with low p-values. See Figure 9.

Logit Regression Results

Dep. Variable:	Default	No. Observations:	27796
Model:	Logit	Df Residuals:	27787
Method:	MLE	Df Model:	8
Date:	Sun, 31 Dec 2023	Pseudo R-squ.:	0.08198
Time:	22:48:07	Log-Likelihood:	-17499.
converged:	True	LL-Null:	-19062.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-0.0311	0.033	-0.930	0.353	-0.097	0.034
IncomeTotal	-0.1106	0.033	-3.337	0.001	-0.176	-0.046
LiabilitiesTotal	0.0988	0.021	4.705	0.000	0.058	0.140
NrOfDependants	0.0340	0.013	2.652	0.008	0.009	0.059
ES	1.5485	0.037	42.213	0.000	1.477	1.620
FI	1.2067	0.038	31.735	0.000	1.132	1.281
SK	2.5211	0.242	10.406	0.000	2.046	2.996
Higher	-0.4664	0.030	-15.459	0.000	-0.526	-0.407
Other	0.0705	0.030	2.347	0.019	0.012	0.129

Figure 9: P-values of remaining features after the backward stepwise selection.

Since models have different predictors even after cross-validation and achieve almost identical results, two hypotheses are relevant:

1. There are multiple combinations of predictors similarly indicating the probability of default.
2. Few features that repeat over and over again in the "bucket of valuable predictors" have such an impact on the prediction that any other features are not relevant. More closely discussed in 6.2.5.

The first hypothesis is improbable. Therefore, even though the p-values suggest using all the predictors, the coefficients were checked and discussed. Predictors related to countries of origin have much higher coefficients than any other predictors, especially the ones at which the forward and backward selection differs. There are again two reasons for this:

1. Data are highly biased towards countries of origin.
2. `Country` is a binary predictor 0/1, but `Age` can go up to 100, so after the multiplication, the weight will play a similar role in predicting the default.

More about it will be said in Subsection 6.2.5.

6.2.4 Dimension reduction by LDA and QDA

Two dimensional reduction techniques (LDA and QDA) were applied to the dataset to show the diversity of the dataset and the intersection between default and non-default groups to highlight the variety of the data. See Figure 10 created by other dimensionality reduction technique PCA. We can see a slight domination of single-colored sides of the image, but there is still a significant noise and overlapping throughout the dataset.

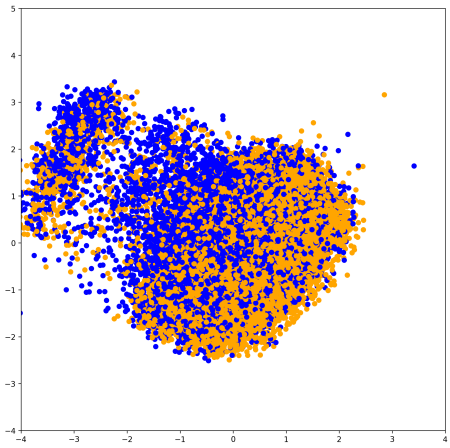

Figure 10: Dimension reduction by PCA to 2 dimensions

After the evaluation, we could see that the achieved accuracy reached almost 70%, but unfortunately also values around 65%. This is due to the enormous number of points around the borderline, which are often stacked on each other, and so even a small shift of the borderline can result in a notable change in the model. So, there is a difference to other approaches by the variance the results can have, while the other approaches offer stable metrics of the models.

So, not only the model needs to be always recomputed when predicting the default but even more, the results are not stable. Due to those facts PCA/LDA/QDA should not be used for modeling, max just for visualization purposes as we used it to highlight the variance in the data.

6.2.5 Logistic regression without countries

Some of the papers with different approaches mentioned in Section 3 already pointed out the high dependence of the models on `Country` predictors but were not doing anything about it. Therefore, we decided to exclude those dummy variables originally belonging to the `Country` variable. It left us with 25 features. Two main differences were detected:

1. The models with all approaches mentioned above with reduced dataset were not able to reach accuracy of the original dataset, but instead were circling around 63/64% each (shrinkage methods were slightly above 65%).

   In reality, a 2% difference is an impactful decision to make. Bondora has some prose costs and, like all the other companies, is living out of the margin. In the field of lending money, you have just a limited amount of money to offer. Therefore, each mistake is not costing you the money just in the form of not having the ability to receive them back, but even more, you cannot get the potential interest out of them. [13] So, each blunder hits twice, so there had to be discussions about the origins of the customers in Bondora headquarters.

2. Instead of reducing the computation time due to fewer predictors, the computation time is extended. We do not have an explanation for that, but it was just an empirical experience.

To return to the first point, shrinkage methods still preferred `Education` dummy variables, `Service` and `LoanDuration`, but the model replaced the origin of the applicant by `IncomeTotal` or/and `NewCreditCustomer`. Those new features make much more sense from the human perspective to

[13] In using money to make money, there are three options the investor can consider. The first is I will invest money and have a negative interest. So it is preferred not to do anything with them. The second is to keep the cash volatile and wait for a safe opportunity to use it but potentially lose it due to inflation. The third is to hope for positive interest. The P2P platforms have just the first and third options and nothing in between. And so they cannot say, after one mistake, I will have two positive trades. But rather, after one mistake, I will not have money to offer for the other two positive trades.

be relevant predictors for default rather than the applicant's origin. Therefore, even though it does not make economic sense to exclude `Country` from the dataset, based on moral perspective and overcoming bias towards some minorities, e.g., Slovaks, it should be discharged. Of course, it can be said that each predictor is somehow biased, otherwise it could not be used for prediction. But, models should be built by features with small biases, not few or even one strong classifier. Their combination will decide the final word in offering the loan to the customer. Therefore, we are not concluding:" `Country` should be excluded from all the models out there (even though Country is a too general predictor, that from a moral and human perspective it should be excluded), " because it can improve the accuracy. In reality, it should be discussed why such a general predictor, such as `Country` is, has an impactful word in the final decision. We already at the beginning pointed out that bias towards countries other than Estonia is due to the fact that Bondora is an Estonian company known by Estonians, and customers outside of Estonia are people with probably terrible financial backgrounds, taking the ability to apply for the money, on their own country, seeking for money somewhere else. So in the Bondora dataset, the "problem" was not in the `Country` but in potential customers from those country.

6.3 Comparing with Similar Studies

In reviewing the relevant literature, we looked at several papers using the same data set. However, after careful consideration, we concluded that the approaches used in these studies are flawed and not suitable for the task at hand of predicting default. Consequently, it may not be possible to make meaningful comparisons between our work and theirs, as their methodologies do not take into account certain nuances and subtleties.

6.4 Conclusion

In conclusion, the evaluation of different models in this study has shown that all tested models have a similar accuracy value of about 0.67 and an AUC of about 0.72. The use of dimensionality reduction techniques has shed light on the complex structure inherent in the data, making it difficult to clearly separate default and non-default cases by imposing a linear boundary.

Despite this complexity, the primary objective of this work - to determine the probability of default using linear models - was successfully achieved. The results, with an accuracy of 0.67, are validated within the limitations of the complex data structure.

However, it is worth noting a limitation of the study -the inability to analyse the same parameters for individuals who were not approved for a loan. As a result, the study lacks insights into the patterns that might be crucial indicators of likely default.

7 Contribution Statement

The collaborative effort of our team involved distinctive contributions from each member. Juraj Žilt played a pivotal role by presenting a selection of datasets, allowing us to choose one that aligned with our course-acquired knowledge. Valeriia Klimova's contribution was instrumental in preparing the Work Plan, providing a structured framework for our project. Jan Pikman significantly contributed to the project's foundation by undertaking the crucial task of data preprocessing and visualization. Both Juraj Žilt and Tigran Oganesian brought valuable insights by familiarizing themselves with related papers.

Moving forward, Valeriia Klimova and Tigran Oganesian took charge of handling features and multicollinearity. They processed both numerical and categorical parameters, eliminating certain variables and reducing the number of categories in categorical columns. Subsequently, Valeriia Klimova and Tigran Oganesian constructed initial linear models, while Juraj Žilt focused on optimizing their performance. Jan Pikman was responsible for the dimensionality reduction of the dataset.

In a collaborative discussion, all authors thoroughly reviewed and documented the outcomes, synthesizing our individual contributions into a cohesive and comprehensive project. The collective effort and varied expertise of each team member were essential in achieving the project's objectives and outcomes.

References

[1] B. Dömötör, F. Illés, and T. Ölvedi, "Peer-to-peer lending: Legal loan sharking or altruistic investment? analyzing platform investments from a credit risk perspective," *Journal of International Financial Markets, Institutions and Money*, vol. 86, p. 101801, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1042443123000690

[2] stuartlee165, "Bondora defaults," https://github.com/stuartlee165/Bondora_Defaults/tree/main, 2022.

[3] R. Chitale, "Bondora peer to peer lending," Oct 2021. [Online]. Available: https://www.kaggle.com/code/rohanc7/bondora-peer-to-peer-lending

[4] L. Ozdemir, "Default loans predictor," Jul 2021. [Online]. Available: https://www.kaggle.com/code/leventoz/default-loans-predictor-acc-99-99/notebook

[5] S. A. Fard, "Risk prediction for loan applications by machine learning algorithms," Bachelor Thesis, TALLINN UNIVERSITY OF TECHNOLOGY, 2023. [Online]. Available: https://digikogu.taltech.ee/et/Download/e7b8b0b3-b901-49ec-be82-58430ba8e58a

A Data Documentation

After data preprocessing and grouping according to the meaning, it was decided to keep the following features:

- General person characteristics
  - `Age`: The age of the borrower when signing the loan application.
  - `Gender`: The gender of the borrower.
    - `0.0`: Male
    - `1.0`: Female
    - `2.0`: Undefined
    - `NaN`: Undefined
  - `Country`: County of the borrower.
    - `EE`: Estonia
    - `FI`: Finland
    - `ES`: Spain
    - `SK`: Slovakia
    - `NL`: Netherlands
  - `Education`: Education of the borrower.
    - `1.0`: Primary education
    - `2.0`: Basic education
    - `3.0`: Vocational education
    - `4.0`: Secondary education
    - `5.0`: Higher education
    - `0.0`: Undefined
    - `-1.0`: Undefined
    - `NaN`: Undefined
- Household characteristics
  - `MaritalStatus`: Marital status of the borrower.
    - `1.0`: Married
    - `2.0`: Cohabitant
    - `3.0`: Single
    - `4.0`: Divorced
    - `5.0`: Widow
    - `0.0`: Undefined
    - `-1.0`: Undefined
    - `NaN`: Undefined
  - `NrOfDependants`: Number of children or other dependants.
  - `HomeOwnershipType`: Type of property owned by the borrower
    - `0.0`: Homeless
    - `1.0`: Owner
    - `2.0`: Living with parents
    - `3.0`: Tenant, pre-furnished property

        * `4.0`: Tenant, unfurnished property
        * `5.0`: Council house
        * `6.0`: Joint tenant
        * `7.0`: Joint ownership
        * `8.0`: Mortgage
        * `9.0`: Owner with encumbrance
        * `10.0`: Other
        * `-1.0`: Undefined
        * `NaN`: Undefined
- Person's liabilities
    - `ExistingLiabilities`: Borrower's number of existing liabilities.
    - `LiabilitiesTotal`: Total monthly liabilities.
- Employment characteristics
    - `EmploymentStatus`:
        * `1.0`: Unemployed
        * `2.0`: Partially employed
        * `3.0`: Fully employed
        * `4.0`: Self-employed
        * `5.0`: Entrepreneur
        * `6.0`: Retiree
        * `0.0`: Undefined
        * `-1.0`: Undefined
        * `NaN`: Undefined
    - `EmploymentDurationCurrentEmployer`: Employment time with the current employer.
    - `WorkExperience`: Borrower's overall work experience in years.
    - `OccupationArea`: Area in which the borrower is employed.
        * `1.0`: Other
        * `2.0`: Mining
        * `3.0`: Processing
        * `4.0`: Energy
        * `5.0`: Utilities
        * `6.0`: Construction
        * `7.0`: Retail and wholesale
        * `8.0`: Transport and warehousing
        * `9.0`: Hospitality and catering
        * `10.0`: Info and telecom
        * `11.0`: Finance and insurance
        * `12.0`: Real-estate
        * `13.0`: Research
        * `14.0`: Administrative
        * `15.0`: Civil service & military

        * `16.0`: Education
        * `17.0`: Healthcare and social help
        * `18.0`: Art and entertainment
        * `19.0`: Agriculture, forestry and fishing
        * `0.0`: Undefined
        * `-1.0`: Undefined
        * `NaN`: Undefined
- Person's income details
    - `IncomeTotal`: Borrower's total income.
    - `DebtToIncome`: Ratio of borrower's monthly gross income that goes toward paying loans.
    - `FreeCash`: Discretionary income after monthly liabilities.
- Credit history
    - `NewCreditCustomer`: Did the customer have prior credit history in Bondora
        * `False`: Customer had at least 3 months of credit history in Bondora
        * `True`: No prior credit history in Bondora
    - `NoOfPreviousLoansBeforeLoan`: Number of previous loans.
    - `AmountOfPreviousLoansBeforeLoan`: Value of previous loans.
- Loan application details
    - `Status`: The current status of the loan application.
        * `Current`: Loan application is current.
        * `Late`: Loan application is late.
        * `Repaid`: Loan application is repaid.
    - `AppliedAmount`: The amount borrower applied for originally.
    - `UseOfLoan`: Purpose for which the loan is applied for.
        * `0`: Loan consolidation
        * `1`: Real estate
        * `2`: Home improvement
        * `3`: Business
        * `4`: Education
        * `5`: Travel
        * `6`: Vehicle
        * `7`: Other
        * `8`: Health
        * `101`: Working capital financing
        * `102`: Purchase of machinery equipment
        * `103`: Renovation of real estate
        * `104`: Accounts receivable financing
        * `105`: Acquisition of means of transport
        * `106`: Construction finance
        * `107`: Acquisition of stocks
        * `108`: Acquisition of real estate

    - `109`: Guaranteeing obligation
    - `110`: Other business
    - `other in 1XX format`: Business loans that are not supported since October 2012
    - `-1`: Undefined
    - `NaN`: Undefined
  - `Interest`: Maximum interest rate accepted in the loan application.
- Loan details
  - `Amount`: Amount the borrower received on the Primary Market. This is the principal balance of your purchase from Secondary Market.
  - `MonthlyPayment`: Estimated amount the borrower has to pay every month.
  - `LoanDuration`: Current loan duration in months.
  - `ProbabilityOfDefault`: Probability of Default, refers to a loan's probability of default within one year horizon.
  - `DefaultDate`: The date when loan went into defaulted state and collection process was started.