Digital Finance (2024) 6:501–542
https://doi.org/10.1007/s42521-024-00114-3

ORIGINAL ARTICLE

Improving credit risk assessment in P2P lending with explainable machine learning survival analysis

Gero Friedrich Bone-Winkel[1] · Felix Reichenbach[2]

Received: 4 March 2024 / Accepted: 17 May 2024 / Published online: 12 June 2024
© The Author(s) 2024

Abstract

Recent research using explainable machine learning survival analysis demonstrated its ability to identify new risk factors in the medical field. In this study, we adapted this methodology to credit risk assessment. We used a comprehensive dataset from the Estonian P2P lending platform Bondora, consisting of over 350,000 loans and 112 features with a loan volume of 915 million euros. First, we applied classical (linear) and machine learning (extreme gradient-boosted) Cox models to estimate the risk of these loans and then risk-rated them using risk stratification. For each rating category we calculated default rates, rates of return, and plotted Kaplan–Meier curves. These performance criteria revealed that the boosted Cox model outperformed both the classical Cox model and the platform's rating. For instance, the boosted model's highest rating category had an annual excess return of 18% and a lower default rate compared to the platform's best rating. Second, we explained the machine learning model's output using Shapley Additive Explanations. This analysis revealed novel nonlinear relationships (e.g., higher risk for borrowers over age 55) and interaction effects (e.g., between age and housing situation) that provide promising avenues for future research. The machine-learning model also found feature contributions aligning with existing research, such as lower default risk associated with older borrowers, females, individuals with mortgages, or those with higher education. Overall, our results reveal that explainable machine learning survival analysis excels at risk rating, profit scoring, and risk factor analysis, facilitating more precise and transparent credit risk assessments.

Keywords P2P lending · Explainable AI · Cox model · Credit risk · SHAP · Survival analysis

JEL Classification G10 · G21 · G32 · G33 · G51

✉ Felix Reichenbach
f.reichenbach@tu-berlin.de

[1] Technische Universität Berlin, Straße des 17. Juni 135, 16023 Berlin, Germany

[2] Chair of Finance and Investment, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

1 Introduction

Driven by technological advancements, Peer-to-Peer (P2P) lending meets a growing demand for alternative financing and personal loans (Suryono et al., 2019). These loans are typically unsecured (De Roure et al., 2016) and lenders bear the risk of default. Most major P2P lending platforms set interest rates based on their internal rating models. On these platforms, accurate risk assessment is crucial: If the platform overestimates the risk of a loan, the borrower might be able to get a cheaper loan elsewhere. If the risk is underestimated, lenders might not be able to recover their investment in case of default and might not invest via the platform again. However, the lack of credit history and collateral makes it difficult to assess the creditworthiness of borrowers using traditional credit risk assessment methods (Bavoso, 2020).

Machine Learning (ML) is a promising technology that could prove key to more accurate risk evaluation (Zhang et al., 2015). While many studies have been conducted on the topic of credit risk assessment using ML, most focus on predicting binary default, rather than default timing. Moreover, ML is often seen as a "black box"—powerful at risk-classification, but not well suited for analyzing the economic impact of individual risk factors. Accordingly, this (perceived) trade-off between explanatory and predictive performance limits the usefulness of ML methods in risk management (Van Liebergen, 2017). However, recent developments in explainable ML demonstrate that the two are not necessarily at odds, with ML survival analysis methods identifying both known and novel risk factors for breast cancer survival (Liu et al., 2023; Moncada-Torres et al., 2021).

Our study applies this promising methodology to credit risk assessment in P2P lending. We use a dataset from the Estonian platform Bondora consisting of over 350,000 loans and 112 features with a volume of €915 million at the end of 2023. Using classical and extreme gradient-boosted Cox models, we predict the risk of P2P loans. Subsequently, we assign risk ratings using risk stratification. We show that the ratings based on the ML model (i.e., the boosted Cox model) significantly outperform both Bondora's risk rating and the rating based on the classical model (i.e., the linear Cox model).[1] We also discuss the practical implications for risk screening, setting fairer interest rates and the potential profit opportunities for investors when using these models.

Then, we open the "black box" of ML by using Shapley Additive Explanations (SHAP) to explain and compare the models (Lundberg & Lee, 2017). Here, we examine the differences between the classical and ML Cox model and evaluate how the ML model can be used to identify risk factors in P2P lending. Our analysis uncovers risk factors that align with those identified in prior studies, while also unveiling novel relationships with nonlinear and interaction effects.

[1] While it could be argued that "classical" (linear) Cox models are also a form of ML, we follow Ley et al. (2022) in distinguishing between the two: Classical models are defined by the user (top-down), whereas ML models are defined by the algorithm (bottom-up and driven by the data). Appendix B provides a more detailed discussion of this issue.

In summary, we demonstrate that with superior predictive and explanatory performance, explainable ML survival analysis is not only a useful tool for credit scoring but also for the examination of credit risk factors. The study is structured as follows. Section 2 presents the theory, reviews the existing research on risk assessment methodologies, and outlines the contribution and objectives of our study. Section 3 then introduces the datasets and our methodical approach.[2] Section 4 presents the model's performance and explanations, which we interpret and discuss in Sect. 5, followed by Sect. 6 that concludes the study.

2 Theoretical background and literature

In the following subsections, we briefly review the literature on credit risk modeling in P2P lending and explainable ML. Furthermore, we identify the research gap and discuss the objectives and contribution of this study.

2.1 Credit risk modeling in P2P lending

As outlined in the introduction, accurate risk assessment is essential for platforms, borrowers, and lenders. However, a review by Suryono et al. (2019) underscores that risk assessment poses a major challenge in P2P lending due to large information asymmetry, lack of credit history, gender discrimination, and low loan success rates. The review further identifies ML methods and big data as potential solutions to these issues.

Investigating credit risk is a well-established field of research and has been studied extensively using binary classification. Binary classification aims to predict whether a borrower will default. It has been studied widely on P2P lending datasets using statistical (e.g., Emekter et al., 2015; Serrano-Cinca et al., 2015), ML (e.g., Jiang et al., 2018; Xu et al., 2021; Zhou et al., 2019), and rarely using explainable ML methods (e.g., Ariza-Garzón et al., 2020; Bussmann et al., 2021). In practice, like most studies, banks typically use binary classification to calculate the probability of default for credit scoring (Dömötör et al., 2023). While their exact method is not disclosed, Bondora's ratings are based on expected loss, which also takes into account the likelihood of recovery after default (Bondora, 2023c).[3]

Survival analysis offers an alternative approach that has some advantages over traditional classification methods: First, survival analysis takes into account the time duration until a loan defaults. This is not done when analyzing the default status alone. However, time to default plays a crucial role in return on investment and

[2] Appendix B offers a more detailed exposition of the models, performance metrics, risk rating, and explainable ML, serving as an accessible guide for practitioners and researchers keen on applying this innovative methodology.

[3] These ratings are based in part on sensitive data that lenders legally cannot access. This includes prior loan applications, and information from credit bureaus, population registries, banks, and tax authorities (Bondora, 2023c). By encoding information unavailable to lenders in ratings, Bondora could reduce uncertainty between lenders and borrowers. However, if the ratings do not accurately reflect a loan's risk profile, it can lead to mispricing through too high or too low interest rates.

expected loss: With later defaults, the exposure at default and thus investment at risk is lower. This is particularly important in the case of fixed-rate loans, where the outstanding payments for investors are much higher when the loan defaults on the first payment as opposed to the last payment. Second, survival analysis also enables including loans in the training dataset that have not yet reached maturity through censoring. This can be a significant advantage, especially for long maturities, as it allows for the use of more recent data. A third advantage is that researchers can use survival functions to examine the influence of characteristics on solvency over time, which may provide insights into possible underlying causes. This is promising in practice, e.g., when evaluating loans in secondary markets, borrower characteristics may have a different effect after a certain period of time.

While statistical survival analysis models like linear Cox regression are used in some studies on P2P lending datasets (e.g., Emekter et al., 2015; Serrano-Cinca et al., 2015), few ML-based survival analysis studies exist (Suárez-Ramírez et al., 2022; Tan et al., 2019) and none of these use explainable ML methods. Nevertheless, when proposing a novel ML survival-analysis method (Bai et al., 2022) demonstrate that ML-based survival analysis can outperform statistical survival analysis methods in default classification.

2.2 Explainable machine learning and contribution

The scarcity of ML survival analysis studies in recent literature may be caused by seemingly conflicting goals of their methodologies: While popular survival analysis methods like linear Cox regression are commonly used for their interpretability in risk factor assessment (Emekter et al., 2015; Reichenbach & Walther, 2021; Serrano-Cinca et al., 2015), ML methods focus on predictive accuracy and are more difficult to interpret. Even the most precise predictions from a ML Cox model may not be very useful if they are not interpretable.

Recent research addresses this issue with the model agnostic and scalable explainable ML method SHAP (Lundberg and Lee, 2017; Lundberg et al., 2019, 2020; Mitchell et al., 2022). This method enables the explainability of ML Cox models, unveiling much more complex nonlinear relationships and interaction effects than the linear models could capture. This combination of methods yielded breakthrough results in clinical research, identifying both clinically confirmed and potentially novel risk indicators for breast cancer survival (Liu et al., 2023; Moncada-Torres et al., 2021) using explainable ML survival analysis.

To the best of our knowledge, explainable ML survival analysis has not yet been applied to P2P lending. This study aims to address this research gap by presenting explainable ML-based survival analysis as a useful tool for both credit scoring (i.e., classification using risk ratings) and the (inferential) analysis of different risk factors (e.g., debt-to-income, age, education) in P2P lending. Thus, we seek to contribute to both the literature on credit scoring techniques and the analysis of risk factors of individual borrowers.

Concerning the Bondora dataset used to test the methods, previous studies found room for improvement in credit risk scoring (e.g., Dömötör et al., 2023; Lyócsa et al., 2022; Teply & Polena, 2020), and signs for asset mispricing on the secondary market (Caglayan et al., 2020). However, as discussed above, the data has not been extensively studied using interpretable ML and survival analysis methods.

3 Data and methods

In this section, we briefly present the data and methods used in this study. For readers not familiar with survival analysis, ML, or SHAP, we recommend reading our more detailed introduction to these methods in Appendix B, where we introduce classical (Cox) models, their predictions and how these can be generalized using ML methods. We also explain how SHAP values are used to explain ML predictions.

The methodology of our study consists of 6 major steps: (1) preprocessing, (2) sampling, (3) training of the models, (4) rating assignment, (5) performance measurement and (6) analysis of risk factors. These are illustrated by Fig. 1 and are discussed in the following subsections after the presentation of the datasets.

3.1 Datasets

We use two datasets provided by Bondora (2023a) for our analysis. These are updated daily and were last retrieved on January 3, 2024. The loan dataset contains data on all loans originated on the Bondora platform, including over 350,000 loans from Estonia, Finland, and Spain. Its 112 features provide details on the borrowers' demographics, financials, and borrowing history, as well as risk ratings calculated by Bondora and information on the loan terms and outcome (see Appendix A for a full table with Bondora's definitions). With loans ranging from 2009 until the end of 2023, the total cash volume of all loans sums up to 915 million Euros. The loans in the dataset have varying maturities, the most common being 5 and 3 years. Our study focuses on 3-year-loans as the duration is short enough, in light of the limited data timespan available, to allow for the splitting of training, validation, and test data in a temporal order. The fixed loan duration ensures comparability between borrowers when analyzing the risk factors using survival analysis (see Appendix B.1). The following data and plots address this subset.

As seen in Fig. 2 the majority of loans were originated after 2017. The overall risk structure of the portfolio appears to have shifted towards less risky loans after 2019. This may be explained by Bondora's shift to hands-off[4] investment from 2016 to 2020 (Bondora, 2016). The lower amount of loans since 2020 may be due to the COVID-19 pandemic and a focus on higher-quality lenders.

[4] Today, Bondora exclusively offers investments into a single platform-managed portfolio at a fixed annual rate of return.

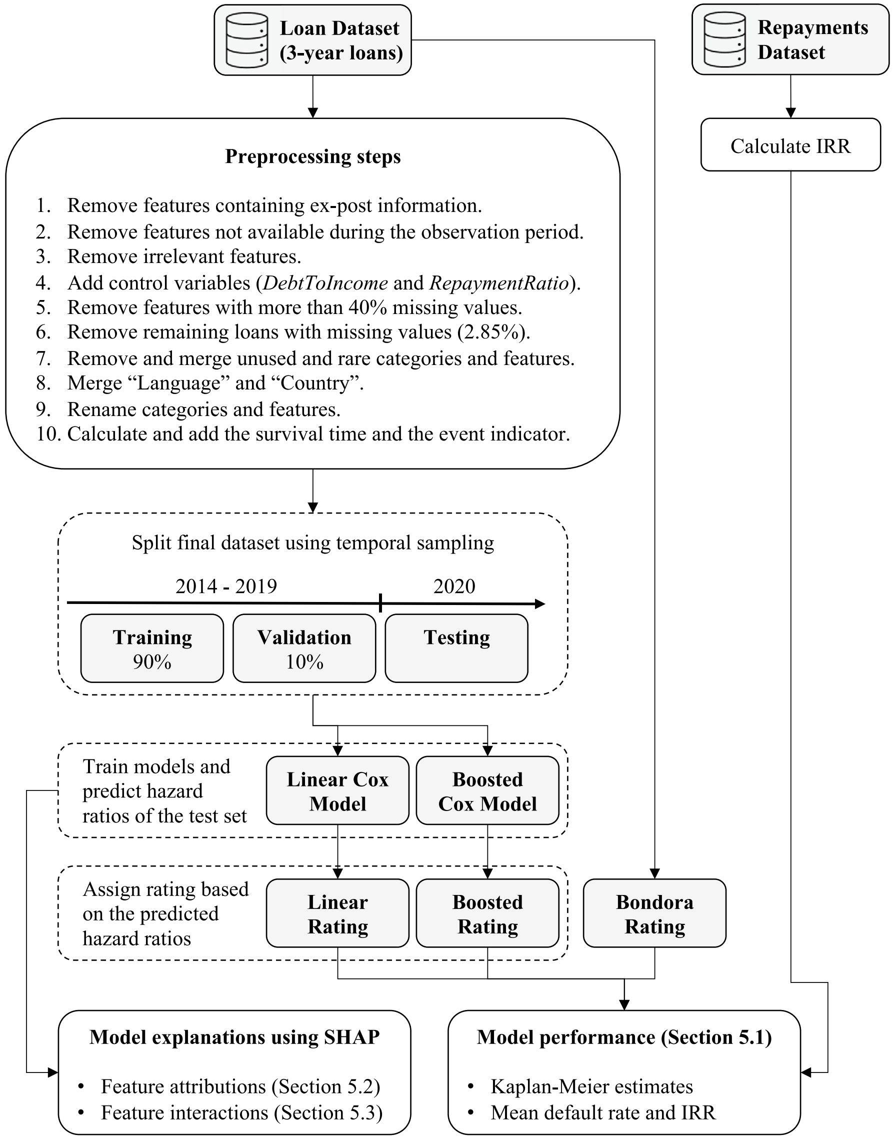

Loan Dataset (3-year loans)
Repayments Dataset
Calculate IRR
Preprocessing steps
1. Remove features containing ex-post information.
2. Remove features not available during the observation period.
3. Remove irrelevant features.
4. Add control variables (DebtToIncome and RepaymentRatio).
5. Remove features with more than 40% missing values.
6. Remove remaining loans with missing values (2.85%).
7. Remove and merge unused and rare categories and features.
8. Merge "Language" and "Country".
9. Rename categories and features.
10. Calculate and add the survival time and the event indicator.
Split final dataset using temporal sampling
2014 - 2019
2020
Training 90%
Validation 10%
Testing
Train models and predict hazard ratios of the test set
Linear Cox Model
Boosted Cox Model
Assign rating based on the predicted hazard ratios
Linear Rating
Boosted Rating
Bondora Rating
Model explanations using SHAP
• Feature attributions (Section 5.2)
• Feature interactions (Section 5.3)
Model performance (Section 5.1)
• Kaplan-Meier estimates
• Mean default rate and IRR

Fig. 1 Overview of the main preprocessing, training and evaluation steps

In addition, we use Bondora's repayments dataset (Bondora, 2023a), which contains all payments received by investors (over 6.2 million in total) to calculate the loans' internal rate of return (IRR). In contrast to the return on investment, the IRR takes into account the time of payment and therefore indicates

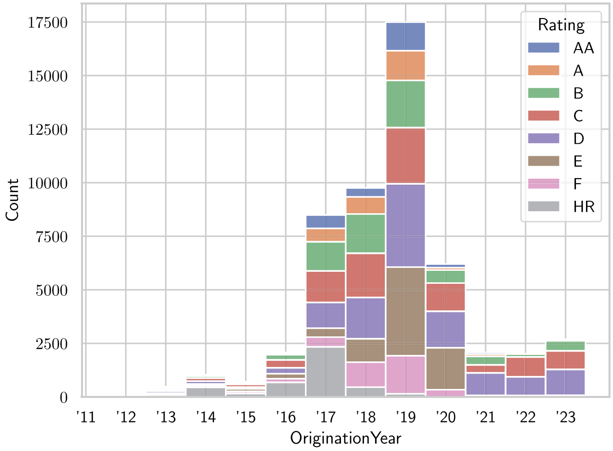


Fig. 2 Number of 3-year-loans by origination year and risk rating

the actual rate of return realized by the investors. The results of this analysis are presented in Sect. 4.1.

3.2 Preprocessing

As seen in Fig. 1, we preprocessed the data for model training in 10 steps. To achieve better comparability, we train both models on the same datasets. As such, these steps follow the stricter requirements of the linear model by excluding missing values. Benefiting the interpretability of both models, we reduce high dimensionality, co-dependencies and sparsity of the data, leading to simpler models and thus simpler SHAP explanations.

Next, we present the preprocessing steps in more detail: First, we removed 37 features not available at the time of the auction to avoid target leakage.[5] This includes data about the loan status, secondary market, and debt collection process. To identify these features, we consulted the Bondora auction Application Programming Interface (API) documentation (Bondora, 2023b) and the Bondora website (Bondora, 2023a). We then removed another 16 features not available due to data protection laws after June 1, 2017 (e.g., private information like marital status and employment position, but also financials like debt-to-income (DTI)). Furthermore to avoid overfitting, we dropped features not relevant to the analysis, like the loan ID, loan number, username, and features about the exact timing of the listing like the payment day, listing time, weekday, and month.

As a fourth step, we modeled some features found relevant in prior literature that are not present in the data and added control variables: *DebtToIncomeModeled* measures the ratio between monthly income and monthly liabilities plus loan

[5] Target leakage occurs when data not available in the real world is used to train a model (Kaufman et al., 2011).

payments.[6] *RepaymentRatio* measures the previous repayment amounts divided by the previous loan amounts.[7]

Next, we removed features with a high percentage of missing values: *DateOfBirth*, *City*, and *County* were missing for all borrowers, likely because they were retracted from the public data. Additionally, we removed the country-specific credit scores from external rating agencies due to missing data (between 41 and 91%). This left us with a few features with less than 2% missing. For these, we removed data points with any missing values. This affected 2.85% of the remaining borrowers. For categorical features, we removed unused and extremely rare categories (less than 0.1% of the borrowers), i.e., the only remaining homeless and 41 Slovakian borrowers. Note that this step may introduce bias to the benefit of improved comparability between the models and could be skipped for the boosted model.

Then, we merged the two categories “income unverified” and “income unverified, cross-referenced by phone” due to a very low number of loans in the latter category (less than 1%). As a significant portion of Estonian borrowers is Russian-speaking, this population was separated from the other Estonian borrowers as “EE_Ru”. For other country codes, the language spoken was more homogenous, and other languages were too rare to draw any conclusions. After that, the features “Language” and “Country” were removed from the dataset and replaced by the new feature “Country_Lang”.

We renamed categorical features encoded as numbers (as seen in Appendix A) to strings to improve readability in the plots and remove false ordinality. Additionally, we renamed *NewCreditCustomer* to *NewBondoraCustomer*, as this reflects the meaning of the feature better, and corrected spelling mistakes. Furthermore, we use one-hot encoding for the categorical variables when fitting the linear models. One-hot encoding creates a new column for every categorical value (e.g., *Rating_AA, Rating_A*, … , *Rating_HR*). The new columns contain a 1 if the loan has the corresponding rating and a 0 otherwise. To avoid multicollinearity, we dropped the first category of each feature in these models.

Finally, we added two variables needed for survival analysis. That is the survival time (*SurvivalTime*) and an event indicator (*Defaulted*). In our analysis of the Bondora dataset, we define survival time and the default indicator as follows. Survival time is the time in days between the loan origination and the first of the following events: default (defaulted), end of the loan term (not defaulted), and end of the observation period (not defaulted), i.e., the split date for training and validation sets or the date of the report for the testing set. In other words, we investigate how long a subject can observably meet the loan’s conditions without default. Crucially,

[6] Loan payments appear not to be included in Bondora’s monthly liabilities variable. This calculated ratio appears to deviate from the DTI ratio calculated by Bondora for data prior to 2017. This deviation is likely caused by the exclusion of some liabilities and income sources from the dataset for data protection reasons. Unfortunately, the exact calculation of the DTI ratio, income, and liabilities are not disclosed by Bondora.

[7] However, some loans had no data on previous repayments. This could be due to the loans being still active or the data being unavailable. On top, some loans had no previous loans. For these variables, the ratio was not calculable and was set to 0. To control for this, we added variables for unknown previous repayments, namely *NanEarlyRepayment* and *NanRepaymentHistory*.

in the case of early repayment, the loan is marked as censored at the complete loan duration of 36 months as well.[8]

3.3 Sampling

Then, we partitioned the data into training, validation, and test sets using temporal sampling. Temporal sampling ensures that the test data chronologically follows the training data. Both training and validation data are censored at the split date. This implies that, regardless of subsequent knowledge about a loan's default post-split, such information is disregarded since it wasn't available at the time of the split. This methodology mirrors a more authentic scenario, reflecting the real-world decision-making context where investors and platforms can only consider information from previously issued loans for risk assessment. Hence, we opted for temporal over random sampling.

The split date of January 1, 2020 was set based on the distribution of loans shown in Fig. 2, aiming to leave sufficient loans for training and enough completed loans in the test set in order to calculate the IRR correctly. The training set contains 90% of the loans between 2014 and the split date, while the validation set contains the remaining 10%.[9] The test set consists of all loans originated in 2020.

We trained the models on the training set, optimized the hyperparameters on the validation set, and evaluated them on the test set. We did not use the test set for any other purpose than evaluation. In contrast, we calculated the SHAP values on the training set to reveal potential overfitting.[10]

3.4 Training

We used the python package *lifelines* (Davidson-Pilon, 2023) to train the linear Cox models. For the boosted Cox models, we used the GPU-accelerated XGBoost package (Chen & Guestrin, 2016). Refer to Appendix B for an in-depth introduction to Extreme Gradient Boosting (XGBoost) survival analysis.

We tuned the hyperparameters of the XGBoost model on the validation set using a Tree-structured Parzen Estimator hyper-parameter optimizer (Bergstra et al., 2011) implemented in the python package *Optuna* (Akiba et al., 2019).

[8] If early repayments had shorter censored survival times than on-time repayments, on-time repayments would be considered less risky than early repayments by the model as their known survival is certainly longer. However, from a financial risk perspective, early repayment could be considered superior to on-time repayment, as it resolves the uncertainty of repayment and releases the bound capital for reinvestment. That is why we set the survival time of early repayments to the entire loan duration.

[9] These were sampled randomly from the pre-split timeframe.

[10] Using the test set could obscure potential overfitting: For example, the boosted model might have learned that all borrowers aged 19 in the training set defaulted and thus assign a spurious risk to all borrowers aged 19. If no borrowers aged 19 existed in the test set, this correlation would be invisible in the test set.

3.5 Rating assignment

We assigned risk ratings to the loans to evaluate the models' predictive performance and compare them to Bondora's rating. For these ratings we used the predicted Hazard Ratios (HR) of the linear and boosted Cox models. To achieve similar sized rating groups, we determined risk intervals that partition the training set predictions into equally-sized groups. Every interval then represents a rating. This approach is similar to the one demonstrated by (Bai et al., 2022), however, they use the test set risk predictions to define the risk intervals. As we use temporal sampling to simulate a realistic application scenario, we need to base the intervals on the training dataset.

3.6 Model and rating performance measures

We measure both direct model performance and the more indirect performance of the model-derived ratings. While including both, we focus on the latter in results Sect. 4.1, as argued below.

As a direct performance measurement of the survival models, we use the concordance index (*c*-index) (Harrell et al., 1982; Uno et al., 2011). It measures the rank-correlation of the predicted survival times with the observed survival times, or in other words, the probability that a randomly selected pair of loans is ranked correctly by their survival time. However, the c-index is likely to be less relevant to investors and platforms, for whom the distinction between early and late defaults is more important than the exact order of survival times.[11]

We therefore focus on the rating's ability to distinguish between good and bad loans. Hence, the linear and boosted ratings (derived from the linear and boosted models, respectively) and the Bondora rating are evaluated on the test set by comparing the default rates of the resulting risk groups. We also calculate the average IRR for each risk rating to assess whether using the models to select loans would be profitable compared to using Bondora's rating. The IRR is defined as the interest rate at which the present value of future cash flows equals the amount of the original investment (Dudley, 1972). Thus, as the name suggests, it can be interpreted as the investor's annual rate of return (assuming the cash flows can be reinvested at the same rate). We use the xirr function of the "tvm" package to calculate the annual effective return of all completed loans (Truppia, 2023).

Additionally, we estimate the survival functions of the risk groups using the Kaplan–Meier estimator from the package *Lifelines* (Davidson-Pilon, 2023). The Kaplan–Meier estimator is a non-parametric estimator of the survival function in the face of censoring (Kaplan & Meier, 1958). The resulting survival function is a step function that decreases with every loan default and expresses the empirical probability of survival beyond each time step.

[11] For an in-depth discussion of Harrell's *c*-index and the Inverse Probability of Censoring Weighting (IPCW) *c*-index, refer to Appendix C.

3.7 Analysis of risk factors using SHAP

Finally, we are interested in the practicability of using the models for an analysis of individual risk factors in P2P lending. For the linear Cox model, this is straightforward, as the coefficients of the Cox model directly indicate whether a feature increases (coefficient > 0) or decreases (< 0) the default risk. The trained boosted models, however, are up to 10 layers deep and consist of up to 2594 trees. This makes them very difficult to interpret. However, the *XGBoost* package includes a method to explain the predictions of the model using SHAP values, which are explained in more detail in Appendix B.

Thus, to be able to compare the predictions of the linear and the boosted Cox models, we calculated the SHAP values for both models. For the linear model, SHAP values equal the product of it's coefficients with the centered values of the variables (see Appendix B for a theoretical explanation). For the boosted model, we used the SHAP implementation in the *XGBoost* python package to estimate the values, including the SHAP interaction effects as developed by Lundberg et al. (2019). The resulting model explanations can be found in Sect. 4.2. In total, the final GPU-leveraged training of each model and calculating explanations took around four minutes.[12]

4 Results

The results are divided into three subsections. First, we present the performance results of our two models and the ratings (derived from the model predictions). Second, we examine the most influential features (risk factors) and display the differences between the two models using SHAP feature dependence plots. Third, we present specific feature interaction effects.

4.1 Model and rating performance

In both Harrell's *c*-index and the more conservative IPCW *c*-index, the ML models performed slightly better than the linear models (0.674 vs 0.659).[13]

The loans in the test set (i.e., loans that originated in 2020) were rated using hazard ratio thresholds calculated on the training predictions (i.e., loans that originated between 2014 and 2019), as described in the methods and data sections. We use the same number of ratings as Bondora in the testing set for comparability (i.e., “AA” to “F”). Bondora rated no loans in the “High Risk” category since 2020.

The following plots show the Kaplan–Meier estimates of the survival probability over time of the risk groups for the different models. The shades around the functions are the 99.9% confidence bands of the empirical survival function estimates (Figs. 3, 4 and 5).

[12] Calculated on a 16 GB, AMD 5600x CPU with an Nvidia RTX3080 GPU.

[13] Please refer to Appendix C for an in-depth discussion on the *c*-index and its limited practical use.

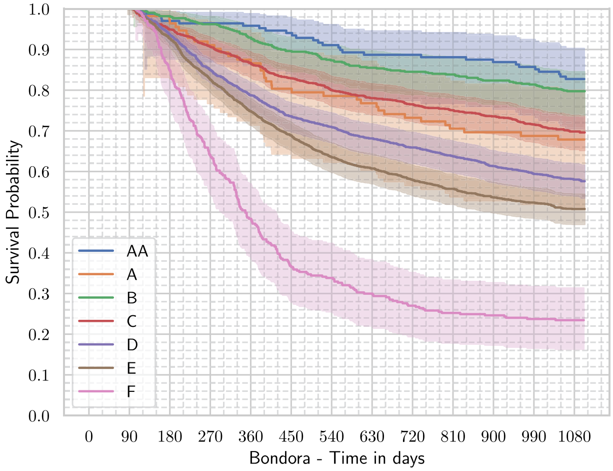


Fig. 3 Kaplan–Meier survival probability estimate of the test set for ratings based on Bondora's ratings. The step function shows the probability of survival of a loan in its respective risk group from the beginning of the loan term to a given point in time. The shaded areas surrounding the functions represent the 99.9% confidence intervals of the Kaplan–Meier survival function estimates

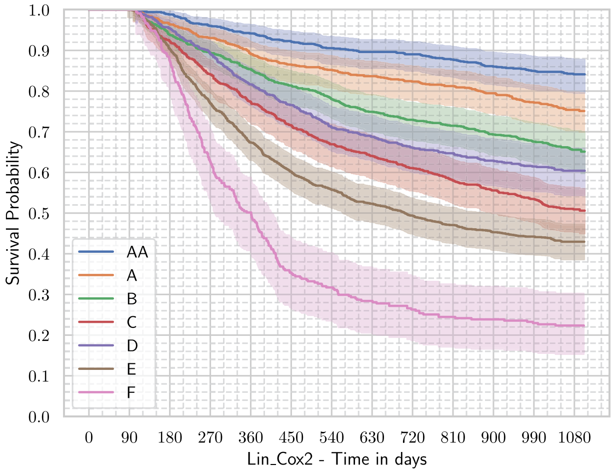


Fig. 4 Kaplan–Meier survival probability estimate of the test set for ratings based on the linear Cox model. The step function shows the probability of survival of a loan in its respective risk group from the beginning of the loan term to a given point in time. The shaded areas surrounding the functions represent the 99.9% confidence intervals of the Kaplan–Meier survival function estimates

The ratings based on the boosted model (Fig. 5) have the smallest confidence intervals and largest margins between risk categories, followed by the linear ratings (Fig. 4) and the Bondora ratings (Fig. 3). In the case of the Bondora ratings, multiple survival estimates cross, with the "A" group having a lower survival probability than the "B" and "C" group at the end of the distribution. For the ratings based on the linear Cox model, loans in group "D" end up having a higher probability of survival

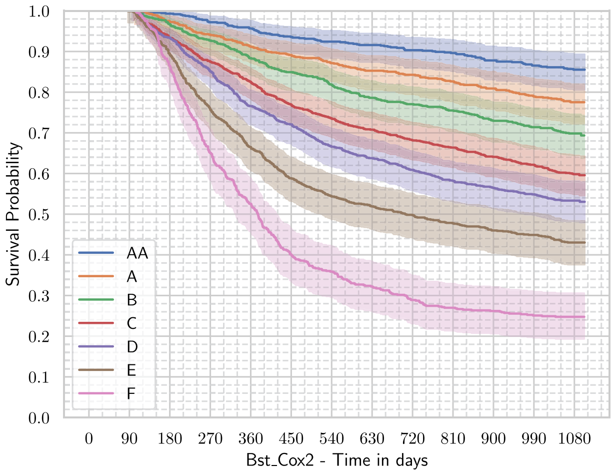


Fig. 5 Kaplan–Meier survival probability estimate of the test set for ratings based on the boosted Cox model. The step function shows the probability of survival of a loan in its respective risk group from the beginning of the loan term to a given point in time. The shaded areas surrounding the functions represent the 99.9% confidence intervals of the Kaplan–Meier survival function estimates

than loans in group “C”. For the boosted ratings, on the other hand, the groups are approximately proportional to each other and the survival curves show considerable margins between them after 36 months. These time-to-default estimates are mathematically tied with the default rates (the mortality rate at the end of the observation period equals the default rate) and thus IRR, which we present in the table below.

Table 1 shows the size of each group, the mean of the applied amount, IRR,[14] interest, and default rate for the Bondora and boosted Cox ratings. Despite the more than four times larger group size of the “AA” group of the boosted ratings, it has a lower default rate than Bondora’s (14.45% vs 17.26%) and significantly higher IRR (15.63% vs – 3.20%). Similarly, the ML model rated twice as many as F, at a higher loss (– 40.69% vs – 36.26%), again indicating better differentiation. In addition, the interest rates are not as strongly correlated with the boosted Cox ratings as they are with the Bondora ratings. For example, the interest rate in the top group of the boosted model is almost three times higher (27.77%) than the top group of the Bondora ratings (9.52%). Overall, the boosted Cox ratings have a more ordinal relationship with default rates and IRR. This is in line with the Kaplan–Meier estimates above (with B having a lower default rate (20.26%) than A (32.14%) for Bondora). Additionally, the boosted model has a significantly wider spread for the IRR between the rating groups (Boosted model: 15.63% to – 40.69% vs Bondora: 3.91% to – 36.26%).

[14] The IRR can only be meaningfully determined for loans that are already closed, which is why only these are included in the table. When determining the loan status, we follow Dömötör et al. (2023) for the sake of comparability and, in addition to the loans marked as closed by Bondora, also take into account those for which no payment has been made for at least one year. This approach is very conservative, as the debt collection process could lead to further payments being made at a later date, which would increase the realized return.

Table 1 Mean values grouped by Bondora and boosted Cox ratings

Rating	Count	$\overline{\text{Amount}}$	$\overline{\text{Interest}}$ (%)	$\overline{\text{Defaults}}$ (%)	$\overline{\text{IRR}}$ (%)
Bondora ratings					
AA	168	3310.78	9.52	17.26	– 3.20
A	112	2603.55	12.79	32.14	– 7.33
B	607	2497.39	17.78	20.26	2.75
C	1316	2542.40	25.14	30.24	3.32
D	1704	2278.13	35.80	42.08	3.91
E	1958	2871.94	40.23	49.18	– 8.60
F	337	2573.88	65.10	76.56	– 36.26
Boosted ratings					
AA: [0, 0.16]	713	2076.35	27.77	14.45	15.63
A: (0.16, 0.24]	801	2202.34	28.97	22.47	12.23
B: (0.24, 0.35]	902	2338.30	29.87	30.27	7.19
C: (0.35, 0.53]	1136	2563.83	32.11	40.23	1.35
D: (0.53, 0.84]	1109	2863.70	35.02	46.71	– 5.37
E: (0.84, 1.45]	911	3096.99	37.49	56.97	– 16.82
F: (1.45, inf)	630	2886.87	46.34	75.24	– 40.69

The table includes the number of loans, the mean amount, interest rate, default rates and IRR for each group. Loan status for the default rate and IRR is based on the approach of Dömötör et al. (2023). For boosted Cox ratings, the Hazard Ratio (HR) thresholds are shown in parentheses. The results for the linear ratings, which exhibits intermediate performance, are reported in Appendix D

4.2 Model explanations: SHAP feature attributions

This section presents the SHAP explanations of the Cox models using scatter plots for the most influential features. In these plots, each point consists of a borrowers' feature value in the training set (x-axis), paired with the SHAP estimate of how that value impacted the models prediction (y-axis). E.g., for age, a point (52, – 0.2) would indicate that SHAP estimated a 0.2 reduced predicted logarithmic HR based on that borrowers age of 52.

Note that these estimates are additive, meaning that adding all SHAP estimates of one borrower across all features yields that borrowers predicted logarithmic HR. As the models predict the logarithmic HR of the borrowers, these plots allow us to interpret the estimated economic impact of a borrowers feature value directly: In non-logarithmic terms, this reduction by 0.2 equals multiplying the borrowers default hazard by $82\% (\approx e^{-0.2})$. For detailed explanations of Cox models and SHAP explanations, refer to Appendices B.2 and B.5.

We color-coded the linear SHAP values by the statistical significance of their corresponding Cox coefficient (full table in Appendix E),[15] as indicated by the color

[15] We set the coloring as significant because when this variable is true, all other coefficients are multiplied by 0, and thus its impact is always 0.

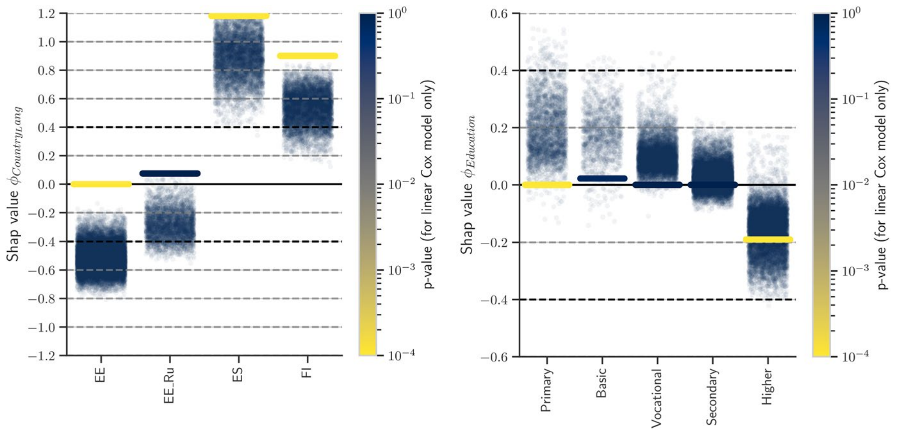


Fig. 6 SHAP values for the features *Country_Lang* (left) and *Education* (right). The blue dots represent the predictions of the boosted model, while the lines are the coefficients of the linear model, with their statistical significance indicated by the color bar on the right (colour figure online)

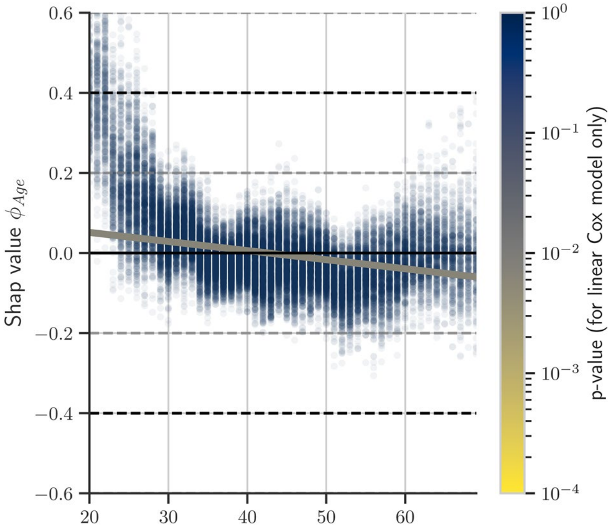


Fig. 7 SHAP values for the feature *Age*. The blue dots represent the predictions of the boosted model, while the slope of the line is determined by the coefficient of the linear model. Its statistical significance is indicated by the color bar on the right (colour figure online)

bars on the right. The vertical dispersion of the boosted SHAP values is explained by feature interaction effects. We will dissect these for selected features later.

Figures 6 and 7 show the SHAP values for the boosted and linear Cox model for three selected features from the demographics category, *Country_Lang*, *Education Age*. The full dependence plots for all 21 features in the categories demographics, financials, and borrowing history are presented in Appendix F.

Country. The most impactful demographic feature is the country and language of the borrower. For this feature, the Shapley values of the linear and boosted Cox model appear to be quite similar in magnitude and direction, except for the centering difference. While the boosted model is centered on the average contribution of all feature values, the linear model is centered on the first feature value dropped in one hot encoding (i.e., *EE*). This results in slightly higher values for the linear model. The linear model's SHAP values are equal to the coefficients of the linear Cox regression (see Appendix E).

Both models show that Estonian-speaking Estonians have the lowest risk, followed by Russian-speaking Estonians and Finnish borrowers. Spanish borrowers exhibit the highest default risk (HR $* e^{1.2} \approx 3.22$).

Education. One of the more intuitive features in the demographics category is the educational level of the borrower. While the boosted model finds a monotonically decreasing risk with higher education, the linear model only finds a significant impact on risk for the highest educational level (HR $* e^{-0.2} \approx 0.82$).

Age. The linear model identified age as a significant risk factor to the $p = 10^{-2}$ level (blue line), and found that the risk decreased monotonically with age. For example, a borrower aged 60 ($\phi_{\text{age}} \approx -0.04$) would have around 92% the risk of a borrower aged 20 ($\phi_{\text{age}} \approx 0.04$) with HR $= e^{-0.04}/e^{0.04} \approx 0.92$. By definition, the slope of the line corresponds to the coefficient of the linear Cox regression (– 0.002) and has its x-intercept at approximately 40 years, which is the average age of borrowers. In the boosted model the risk of default also generally decreases with age. However, young borrowers between 20 and 25 and borrowers older than 60 have a higher risk than the linear model would predict.

While these SHAP dependence plots visualize the individual feature's contributions, some features may interact with other features in the model, leading to vertical dispersion in the plots (e.g., for *Country_Lang* and *Education*). These can be investigated further using SHAP interaction values, as shown in the following section.

4.3 Model explanations: SHAP feature interactions

The SHAP interaction effects estimate how much the SHAP value of a feature changes when combined with another feature. Subtracting all interaction effects from the total effects results in the more focused main effect of a feature. This main effect is often more interpretable than the total effect, as it is not influenced by interaction effects and thus shows less vertical dispersion.

In the plots below, we present the main effect and the two most important interaction effects (which are our two examples from the last section, *Country_Lang* and *Age*) for the categorical feature with the most vertical dispersion, *HomeOwnershipType*. For the interaction effects, the dots in the scatter plots are colored by the value of the interacting variable. Here, red indicates a higher and blue indicates a lower value. The coloring is illustrated with the color bars on the right.

Compared to the total SHAP value (see Appendix F), the main effect for the homeownership type in Fig. 8 shows less vertical dispersion.

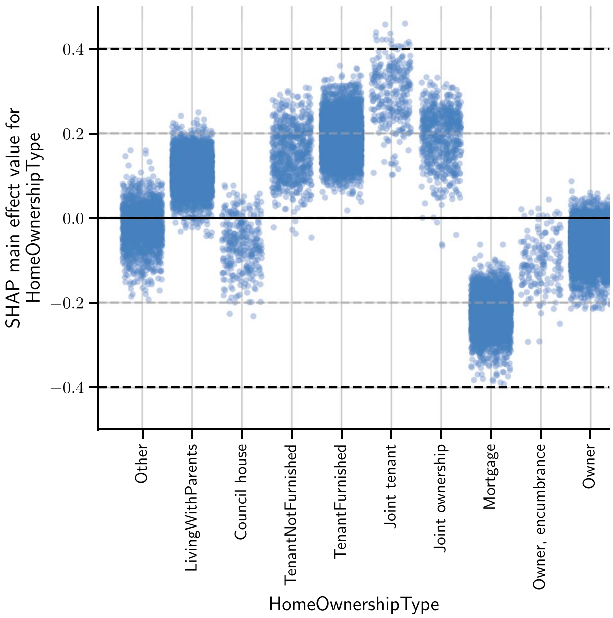


Fig. 8 SHAP values for the main effect of *HomeOwnershipType*. The main effect is calculated by subtracting all interaction effects from the total effect (for *HomeOwnershipType*, see Appendix F)

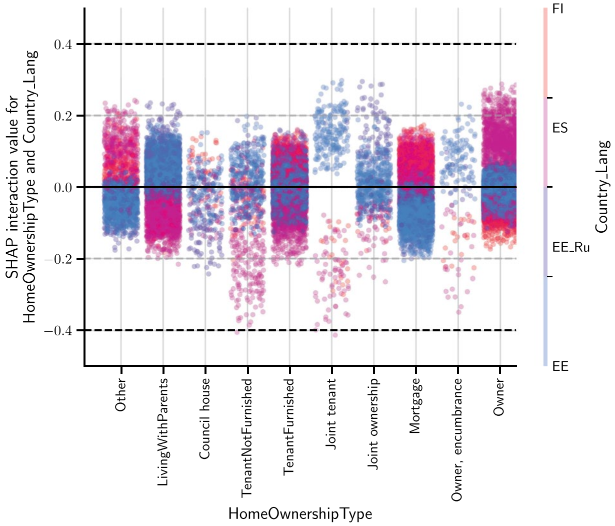


Fig. 9 SHAP values for the interaction effect of *HomeOwnershipType* and *Country_Lang*. The dots are colored by the value of *Country_Lang* (color bar on the right) (colour figure online)

Looking at the interaction with the country (and language for Estonians), as seen in Fig. 9, borrowers living with their parents or in furnished apartments are deemed

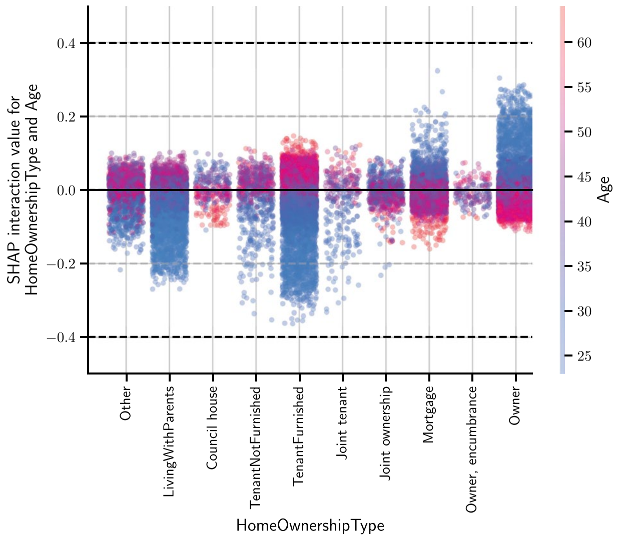


Fig. 10 SHAP values for the interaction effect of *HomeOwnershipType* and *Age*. The dots are colored by the value of *Age* (color bar on the right) (colour figure online)

less risky in Finland and Spain, while the opposite is true for Estonians. The inverse relationship is found for the other categories, except for owners. Here, Spanish borrowers are deemed more risky, followed by Estonians and Finns.

Figure 10 shows the interaction with age. The risk decreases for older borrowers most notably in the categories *Mortgage* and *Owner*, while increasing for *LivingWithParents* and *TenantFurnished.*

In summary, these interaction effects reveal more interpretable feature contributions on credit risk for less clear effects such as *HomeOwnershipType*. These interaction effects are derived from the boosted model without any manual modeling.

5 Discussion

The principal goal of this study was to evaluate the utility and performance of explainable ML survival analysis within the P2P lending sector, with a particular focus on its potential to improve risk assessment and risk indicator analysis compared to linear models.

5.1 Model performance

Our results show that the ML model outperformed the linear model. This applies to both Harrell's *c*-index and the more conservative IPCW *c*-index. While the differences appear slight for the *c*-indices, the following results on practical rating performance demonstrate that the boosted model performed significantly better at risk assessment. For an in-depth analysis of the *c*-index results, including their limited significance in practice and further robustness checks, see Appendix C.

The boosted model achieved the best risk differentiation across the rating groups and vastly outperformed Bondora's rating. The Kaplan–Meier survival estimates revealed that the boosted model was the only model with non-crossing survival curves ordered correctly by risk ranking (Fig. 5). Here, it had the largest margins between groups, the smallest confidence bands and also the largest span between the highest and lowest risk groups, indicating a superior ability to discriminate loans by risk. When looking at the tabular results including default- and IRR rate, compared to Bondora, the boosted model showed a more ordinal, even distribution and wider spread of the IRR between the rating groups from 15.63 to − 40.69% for the boosted-, and 3.91 to − 36.26% for Bondora's ratings.

Investing an equal amount in boosted "AA" rated loans would have outperformed the Bondora-rated "AA" loans by almost 19% p.a. (IRR of 15.63% compared to − 3.20%). Moreover, the boosted "AA" rated loans were less risky, at a default rate of 14.45%, in contrast to 17.26% for the Bondora "AA" group. These results were especially striking when considering the risk-group sizes. The boosted model assigned more than four times as many loans as Bondora to the "AA" rating (713 vs 168). This favorable risk-return profile could be further enhanced by targeting mispriced loans (predicted low risk and high interest rates). Future research could investigate the model's potential for identifying mispriced loans on the secondary market.

On the opposite spectrum, ML survival analysis ratings could aid in the screening of high-risk borrowers, addressing a major challenge for P2P platforms. The boosted model rated almost twice as many borrowers as F, returning − 40.69% annually compared to Bondora's F group at − 36.26%. Screening out these high-risk borrowers would benefit investors, platforms and lower-risk borrowers alike, as argued in the theory section.

In addition, boosted survival analysis can enable fairer interest rates on P2P platforms. Multiple crossing survival curves and poor predictiveness of time-to-default indicated that the Bondora ratings may not be well-calibrated. This could cause unfair interest rates, as detailed in Sect. 2. If the interest rates on loans were fairly pricing risk, they should compensate for it. However, we found a different picture, e.g., the average interest rate of the boosted "AA" group (27.77%) was almost three times that of Bondora's "AA"-group (9.52%), despite the boosted-"AA" group's lower default rate. Moreover, we observed a low correlation between default risk and interest rates. The rating-performance differences may be attributed to the fact that Bondora's ratings encode expected loss rather than expected time-to-default. However, expected loss is linked to expected time-to-default, as argued in the theory section (Sect. 2). Furthermore, the boosted ratings outperformed Bondora's ratings regardless of the used performance measures (i.e., *c*-index, differentiation based on Kaplan–Meier curves, mean IRR and mean default rate). Hence, the interest rates may be set inaccurately and borrowers should get quotes from multiple platforms to promote fairer pricing.

While the boosted model outperformed the linear model, the linear model exhibited decent risk-rating performance, except for "C" having a significantly higher default rate (49.17%) than D (39.58%, Appendix D). We will address the model differences in the next section.

In summary, the application of boosted survival analysis to rate loans originated in 2020, utilizing historical performance data available to an investor by the end of 2019, demonstrated effective results. Our findings suggest that, within our dataset, boosted ML survival analysis models emerged as a promising tool for improving credit risk assessment, supporting investors' decision-making, advancing risk screening and promoting fairer interest rates.

5.2 Model explanations

Another objective of this study was to investigate whether explainable ML survival analysis can be beneficial to risk indicator analysis in P2P lending. With SHAP explanation values, we explained and compared both linear and boosted Cox models using feature dependence plots (Sect. 4.2).

Overall, the boosted model extracted at least as much information from the data as the linear model. In our data, where the linear model identified significant effects, the boosted model consistently found similar effects (see Appendix F for a full display of all feature contributions). Unsurprisingly, these similar findings are mostly in line with prior research. To name a few: Borrowing risk reduced with age (e.g., Albert & Duffy, 2012; Kurnianingsih et al., 2015), male gender indicated higher risk (e.g., Lin et al., 2017), while having a mortgage reduced risk (e.g., Serrano-Cinca et al., 2015).

Furthermore, the boosted model uncovered more detailed and new relationships between borrower information and default risk. For example, all grades of education were well differentiated by the boosted model, with higher education decreasing risk. This is in line with prior research (e.g., Chen et al., 2018; Lin et al., 2017). In contrast, the linear model only found a significant impact on risk for the highest educational level, and no significance for the intermediate levels. The same applied for *homeownershiptype* (Fig. 12b), where the boosted model clearly distinguished *LivingWithParents* (higher risk), and *Owner* (lower risk) in addition to the categories identified significant by the linear model (*Mortgage*, lower risk and *TenantFurnished*, higher risk). While the higher risk for borrowers living with their parents makes intuitive sense (indicates lower assets), the lower risk for homeowners is supported by prior research (e.g., Serrano-Cinca et al., 2015). Furthermore, some variables impacted the boosted model where the linear model did not find significance at all. The relationship with DTI (Fig. 13b) found by the boosted model is a known factor in risk prediction (Emekter et al., 2015; Serrano-Cinca et al., 2015) and was not identified by the linear model.

In addition to the more detailed linear effects, the boosted model found both known and novel non-linear and interaction effects exploratively. For example, credit risk decreased sharply for younger borrowers with age, flattening out until the age of 55 years and then increasing slightly again. As mentioned, research confirms this general trend of the models, as risk aversion increases with higher age (e.g., Albert & Duffy, 2012; Kurnianingsih et al., 2015). Additionally, evidence points towards a nonlinear and domain-specific relationship between risk-taking and age, with increased risk-taking in early adulthood (e.g., Rolison et al., 2014; Willoughby et al., 2021), which was also picked up by the boosted model. On top, the newly quantified risk increase for borrowers at 55 years may be explained by lower income due to retirement, increased medical costs and reduced life expectancy. Analyzing interaction effects also unveiled novel risk indicators and helped us dissect the vertical dispersion of primary effects. For *HomeOwnershipType*, we found that tenants living with their parents or in a furnished apartment were deemed less risky when younger (Fig. 10), which intuitively makes sense. The interaction with the country (Fig. 9) indicates that the model may be able to account for systemic and cultural differences. To the best of our knowledge, these particular interaction effects have not been identified exploratively in prior research. Especially in the European market, analyzing cultural and systemic differences (e.g., due to different retirement systems) might be valuable for credit risk assessment.

Finally, one advantage of the classical model is that it can test for statistical significance. However, in the cases where the ML-SHAP values were less dense, the linear models often found the feature value to be less or even insignificant (white coloring). This indicates that SHAP feature importances (average absolute SHAP values) may be consistent with statistical significance, which was also found by Bussmann et al. (2021).

Overall, the boosted model not only found the same significant risk factors as the linear model but also found more detailed and even new relationships between risk factors and default risk. This included nonlinear and interaction effects that are not quantifiable in linear models in the same explorative way.[16] This is in line with prior research on this methodology in oncology (Moncada-Torres et al., 2021), supporting the argument that explainable machine learning survival analysis can reveal known and potentially novel risk factors in risk research.

5.3 Limitations and implications

Further research is required to test these methods more widely. While we were able to achieve exceptional results on the Bondora dataset, future studies should test the reliability of the models by applying them to different datasets and split dates. For the latter, we suggest a rolling window approach, as this would allow the adaptability and validity of the models to be tested. Here, the US Lending Club dataset could be suitable. The lending club dataset is widely used in prior research and has a much larger number of completed loans ranging over a longer period. The Bondora dataset did not allow for this at the time of writing due to the limited available observation period combined with the need for completed loans to calculate the IRR. However, it can be revisited in the future, as the dataset grows and more loans are completed. For our purposes, to assess risk rating ability and risk-factor analysis the dataset we

[16] While interaction effects can be quantified in linear Cox regression, they need to be carefully modeled by hand based on domain knowledge and prior research. This can be very resource-consuming, especially with many features.

used is well-suited. Especially for the latter, as seen in the explanations, the dataset provided valuable hypotheses into cultural differences important for risk assessment in the European P2P market (this insight would not be possible on the US Lending Club dataset). Furthermore, the good performance of the ML models on the test set despite the changed risk profile (changes in rating distribution in Fig. 2) indicated that the models are likely robust and can be used for future loans. As discussed in more detail in Appendix C, we also applied random sampling and rated 5-year loans, which resulted in even higher performance for the ML models and qualitatively similar results for the SHAP values. Additionally, simply removing borrowers with missing and sparse values from the dataset can introduce bias to the model. This is especially problematic if the missingness is systematic and should be investigated to avoid bias against demographics (e.g., in our case, the excluded homeless and Slovakian borrower). While our decision to compare the models lead to following the more restricted preprocessing steps of the linear model, the boosted model can be trained (and predict) on incomplete and sparse datasets in future use.

Furthermore, not all of the explained relationships, as shown in Appendix F, were straightforward. Some required further investigation (e.g., using interaction effects), and some contradicted intuition. For example, we found a risk decrease with larger liabilities and an increase with larger income. We suspect that this is caused by collinearities, e.g., with DTI.[17] This indicates that explainable ML findings need to be validated, as Shapley values explain the model, and not the data directly. If the model finds complex mathematical relationships that obscure the true underlying risk factors, explanations may be incapable of identifying meaningful risk indicators that are intuitive, straightforward, and supported by theory. Developing a model with useful explanations may require some trial and error—especially through regularization and adequate data prepossessing, including feature selection and modeling of variables likely relevant in reality. Nevertheless, spurious relations are not only found in ML models but also in linear models. To overcome this issue, a hypothesis-based approach is commonly used with linear models. For explainable ML methods matching quantitative with qualitative theory is equally important, although its explorative approach allows deriving hypotheses directly, which should be verified with existing theory.

6 Conclusion

In this study, we used classical and ML survival analysis to predict default risk in P2P lending using a European dataset with over 350,000 loans. We compared the performance of the models in rank correlation, classification, and credit risk rating. We then opened the ML models' black box using SHAP to explain the performance differences and identify credit risk indicators.

[17] While liabilities do not contain the applied amount, the DTI ratio does. At a fixed DTI ratio, larger liabilities may indicate a lower relative loan amount to a borrower's financial situation. Additionally, DTI may typically decrease with higher incomes, and thus, high DTI at larger incomes may indicate higher risk.

Our results demonstrated that ML survival analysis performs exceptionally well in credit scoring, significantly outperforming the platform's risk ratings. For investors, our ratings revealed a profit opportunity through targeting high-interest loans with low estimated risk. On the platform side, more accurate risk assessments could promote fairer pricing and improve the screening process, ultimately reducing overall risk and increasing portfolio performance.

Using SHAP, we were able to explain the models decision making and discover both novel and known credit risk factors. This yielded compelling and intuitive hypotheses, establishing promising avenues for future research.

Altogether, the methodology's exceptional performance results, combined with its meaningful explanations, confirmed its ability to improve credit-risk ratings through more accurate while transparent credit risk assessments. With analogous findings in oncology that validate explainable ML survival analysis' ability to generate knowledge, we are confident that this approach can further the understanding of time-to-event data across various domains. This methodology could spark a new wave of survival analysis research, including the reexamination of studies previously conducted with linear survival analysis methods.

Appendices

A Bondora dataset feature descriptions

The following table shows an excerpt of the features presented in the loan dataset, grouped by categories. The features were selected based on public availability and relevance for the analysis. Bondora excluded 16 features due to data protection regulations starting on June 1, 2017. The excluded features include information about family status, employment, loan usage, and financial details like sources of income, free cash, and debt-to-income (DTI) ratio (Table 2).

B Detailed explanation of the methodology

This appendix explains the survival analysis models used in this study in more detail and demonstrates how classical models can be generalized to allow for more complex models using Machine Learning (ML). Finally, we introduce the methods used to evaluate and explain the models.

B.1 Survival analysis

Statistical survival analysis investigates the timing of an event of interest. This event can be anything that occurs over time, like death, failure of a machine, or credit default.

Table 2 Data features of the Bondora loan dataset

Group	Name	Description
Application and contract details	AppliedAmount	The amount borrower applied for originally
	ContractEndDate	The date when the loan contract ended
	FirstPaymentDate	First payment date according to initial loan schedule
	LoanDate	Date when the loan was issued
Bondora rating details	ExpectedLoss	Expected Loss calculated by the current Rating model
	ExpectedReturn	Expected Return calculated by the current Rating model
	HomeOwnershipType	0 Homeless 1 Owner 2 Living with parents 3 Tenant, pre-furnished property 4 Tenant, unfurnished property 5 Council house 6 Joint tenant 7 Joint ownership 8 Mortgage 9 Owner with encumbrance 10 Other
	LossGivenDefault	Gives the percentage of outstanding exposure at the time of default that an investor is likely to lose if a loan actually defaults. This means the proportion of funds lost for the investor after all expected recovery and accounting for the time value ...
	ProbabilityOfDefault	Probability of Default refers to a loan's probability of default within a one-year horizon.
	Rating	Bondora Rating issued by the Rating model
Borrower demographics	Age	The age of the borrower when signing the loan application
	Country	Residency of the borrower
	Education	1 Primary education 2 Basic education 3 Vocational education 4 Secondary education 5 Higher education
	EmploymentStatus	1 Unemployed 2 Partially employed 3 Fully employed 4 Self-employed 5 Entrepreneur 6 Retiree
	Gender	0 Male 1 Woman 2 Undefined
	LanguageCode	1 Estonian 2 English 3 Russian 4 Finnish 5 German 6 Spanish 9 Slovakian
	NewBondoraCustomer (renamed from NewCreditCustomer)	Did the customer have a prior credit history in Bondora 0 Customer had at least 3 months of credit history in Bondora 1 No prior credit history in Bondora
	VerificationType	Method used for loan application data verification 0 Not set 1 Income unverified 2 Income unverified, cross-referenced by phone 3 Income verified 4 Income and expenses verified
Credit and borrowing history	AmountOfPreviousLoansBeforeLoan	Value of previous loans
	NoOfPreviousLoansBeforeLoan	Number of previous loans

Table 2 (continued)

Group	Name	Description
	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
Debt servicing costs	InterestAndPenaltyDebtServicingCost	Service cost related to the recovery of the debt based on the interest and penalties of the investment
	PrincipalDebtServicingCost	Service cost related to the recovery of the debt based on the principal of the investment
Financials	ExistingLiabilities	Borrower’s number of existing liabilities
	IncomeTotal	Borrower’s total income
	LiabilitiesTotal	Total monthly liabilities
Loan information	Amount	Amount the borrower received on the Primary Market. This is the principal balance of your purchase from Secondary Market
	Interest	Maximum interest rate accepted in the loan application
	LoanDuration	Current loan duration in months
	MonthlyPayment	Estimated amount the borrower has to pay every month
Payment and default details	DefaultDate	The date when a loan went into defaulted state and the collection process was started
	EAD1	Exposure at default, outstanding principal at default
	EAD2	Exposure at default, loan amount less all payments prior to default
	InterestAndPenaltyPaymentsMade	Note owner received loan transfers earned interest, penalties total amount
	InterestAndPenaltyWriteOffs	Interest that was written off on the investment
	PrincipalPaymentsMade	Note owner received loan transfers principal amount
	PrincipalWriteOffs	Principal that was written off on the investment
Privacy excluded	DebtToIncome	Ratio of borrower’s monthly gross income that goes toward paying loans
	FreeCash	Discretionary income after monthly liabilities
	NrOfDependants	Number of children or other dependants

Table 2 (continued)

Group	Name	Description
	OccupationArea	1 Other 2 Mining 3 Processing 4 Energy 5 Utilities 6 Construction 7 Retail and wholesale 8 Transport and warehousing 9 Hospitality and catering 10 Info and telecom 11 Finance and insurance 12 Real-estate 13 Research 14 Administrative 15 Civil service ...
	WorkExperience	Borrower’s overall work experience in years

With survival analysis methods, the influence of explanatory variables on survival time can be assessed (Kleinbaum & Klein, 2012).

In survival analysis, survival time T is a random variable that measures the time between the start of the observation period and the event of interest (e.g., time to default). The survival function $S(t)$ yields the probability that this event occurs after time t:

$$S(t) = P(T > t). \tag{1}$$

As such, if we estimate $S(t)$ for the event credit default, we can calculate the probability that default does not occur before time t.

We might be interested in the risk of default after a certain survival time, for example when buying an existing loan on the secondary market. This risk could be calculated based on the hazard function $h(t)$. The hazard function $h(t)$ yields the instantaneous risk that the event of interest occurs at exactly $t,$ given that it has not occurred yet:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2}$$

This function is directly related to the more intuitive survival function $S(t)$. That is:

$$h(t) = \frac{\mathrm{d}S(t)/\mathrm{d}t}{S(t)}. \tag{3}$$

By estimating the hazard function $h(t),$ the survival function $S(t)$ can be calculated and vice versa.

In survival analysis, some subjects do not experience the event of interest during the observation period, such as outstanding loans in our case. These cases are considered censored. While we do not know the exact survival time of these subjects, we know that they survived at least as long as they were observed. Survival analysis methods allow us to extract this information from censored loans.

B.2 Cox proportional hazards model

Various models exist to estimate the survival function $S(t)$ and the hazard function $h(t)$. In our analysis, we focus on the widely used Cox model (Cox, 1972). This is a semi-parametric model that estimates the Hazard Ratio (HR) of subjects to evaluate the effect of covariates on the survival time. In the Cox model, the hazard function $h(t)$ can be formulated as follows:

$$h(t \mid \mathbf{x}_i) = h_0(t) \cdot \exp(\langle \mathbf{w}, \mathbf{x}_i - \bar{\mathbf{x}} \rangle). \tag{4}$$

Here, $h_0(t)$ is the baseline hazard function, $\mathbf{w}$ is a vector of regression coefficients, $\mathbf{x}_i$ is a vector of features for a subject $i,$ and $\bar{\mathbf{x}}$ is the mean of all features.[18]

[18] The centering of the features using $\mathbf{x}$ is optional and does not change the HRs of any two subjects i and $j,$ as seen in Eq. 5. However, centering the features increases the numerical stability of the models and allows for a more direct comparison in the plots used in the results section.

To calculate the regression coefficients $\mathbf{w}$, the parameters $\mathbf{w}$ are estimated by maximizing the partial likelihood of the model based on the observed data (Kleinbaum & Klein, 2012). In this approach, the baseline hazard function $h_0(t)$ cancels out and does not need to be estimated. With an unknown baseline hazard function the Cox model can be used to estimate the HRs of any two subjects i and j, that is:

$$HR = \frac{h(t \mid \mathbf{x}_i)}{h(t \mid \mathbf{x}_j)} = \exp(\langle \mathbf{w}, \mathbf{x}_i - \bar{\mathbf{x}} \rangle - \langle \mathbf{w}, \mathbf{x}_j - \bar{\mathbf{x}} \rangle) = \exp(\langle \mathbf{w}, \mathbf{x}_i - \mathbf{x}_j \rangle). \tag{5}$$

The constant HR measures the relative risk of experiencing the event of interest at any point t in the observation period for two subjects (or groups) i, j. The linear scaling factors $\mathbf{w}$ of the HR offer easy interpretability of the features' impact on hazard and thus survival. With these coefficients, we can estimate the effect of a feature on the HR: For example, a coefficient of $w_i = -0.05$ for the i-th feature *age* means that the HR decreases with a factor of $\exp(-0.05) \approx 0.9512$ for every increasing year of age compared to the baseline hazard. That is in our example, a subject that is one year older has a 4.88% lower chance of defaulting at any time than a subject with the same features that is 1 year younger.

Another benefit of the Cox model is that it is semi-parametric, meaning that we do not need to make assumptions about the shape of the hazard function by investigating the HR instead.

However, these HRs are independent of time, leading to the eponymous proportional hazards assumption of the Cox model. This assumption limits the model's flexibility, and may not hold in practice. For example, when some borrowers are more at risk for early default while others are more at risk for late default, their true hazard functions would look quite different and not necessarily proportional to each other. Furthermore, the log-linear scaling factor $\langle \mathbf{w}, \mathbf{x} \rangle$ is unable to model non-linear- and feature interaction effects. These properties limit the Cox model's ability to model complex relationships between features and survival time. Finally, as we investigate HRs and not the hazard function directly, the Cox model cannot predict survival times directly.[19]

An alternative approach to the widely used semi-parametric Cox model is the fully parametric Accelerated Failure Time (AFT) AFT model that estimates the survival time T directly.[20] However, both models are quite similar in their mathematical limitations due to the linearity of the scaling factor $\langle \mathbf{w}, \mathbf{x} \rangle$. Thus, both models are incapable of modeling time-dependent scaling factors, non-linearity in the coefficients, and interaction effects.

[19] However, the baseline hazard function $h_0(t)$ can be estimated, for example using the Breslow estimator (Breslow, 1972).

[20] While the AFT model is capable of predicting survival times and the Cox model is not, the Cox model does not make assumptions about the distribution. This is especially important in our case, as the distribution of survival times is unknown. However, if the proportional hazards assumption does not hold, the AFT model could be more appropriate when given the correct distribution. As a robustness check, we repeated the main analysis using an AFT model with qualitatively similar results (see Appendix C).

B.3 Machine learning survival analysis

The limitation of linearity and lack of feature interaction effects in the classical models can be overcome by replacing $\exp(\langle \mathbf{w}, \mathbf{x} \rangle)$ with an arbitrary function $\mathcal{T}: \mathbb{R}^n \to \mathbb{R}$, where n is the number of features. One way to approximate the unknown function $\mathcal{T}$ is to use ML methods.

In supervised learning, a subfield of ML, we aim to learn a mapping from input to output of a known dataset (training set) to predict the output of new data points (test set). For example, we could be interested in predicting the number of months a loan is serviced based on features like age, income and existing debts. Here, the input would be the features and the output would be the time to default on the loan. We can solve this learning problem as an optimization problem by using a function that measures how well a model performs (loss function). For example, a loss function could measure the difference between the predicted and the actual survival time of a loan. By minimizing this difference, we can find the best[21] model for predicting the time until a loan defaults.

One could argue that the survival models presented above are a form of ML. After all, they maximize the partial likelihood (or its negative, if used as a loss function), which itself measures how well the model predicts the relative survival times of the data. What sets apart ML from classical methods, is their approach to generating models. In statistics, the model and its mathematical properties are defined by the user, whereas in ML, the model is created by the algorithm itself (Ley et al., 2022).

As mentioned above, the Cox model can be generalized by replacing the scaling factor $\exp(\langle \mathbf{w}, \mathbf{x} \rangle)$ with an arbitrary function that maps from any data point $\mathbf{x}$ to a real number $\mathcal{T}: \mathbf{x} \to \mathbb{R}$. This results in the Cox model with a generalized scaling factor:

$$h(t \mid \mathbf{x}) = h_0(t) \cdot \mathcal{T}(\mathbf{x}). \tag{6}$$

With $\mathcal{T}$ being an arbitrary function, ML methods can approximate the unknown function $\mathcal{T}$ without linear restrictions. With this more general model, feature interaction effects and non-linear, non-monotonic relationships between the features and the survival time can be modeled.

However, machine-learning Cox models are still somewhat constrained. Due to the unknown baseline hazard $h_0(t)$, the model is not entirely defined by the algorithm. This leads to the aforementioned proportional hazards assumption. Interestingly, the optimization algorithm may use the same scoring function as the classical models by maximizing the partial likelihood of the training data. This is how they are optimized in this study using Extreme Gradient Boosting (XGBoost), the method presented below.

[21] Finding the global minimum is a major challenge in ML research, thus the resulting model is not certainly the best possible model.

B.4 Extreme gradient boosting

One method to approximate the unknown function $\mathcal{T}$ is to use XGBoost (Chen & Guestrin, 2016). XGBoost is a machine-learning technique for regression and classification problems that performs particularly well on tabular data. While deep learning models have achieved breakthroughs in image and text processing, they are usually outperformed by gradient boosting on tabular data and harder to train and optimize than XGBoost (Shwartz-Ziv & Armon, 2021). XGBoost is essentially an (extreme) optimization of gradient-boosted trees. Thus, we first briefly explain the key concept of gradient-boosted trees and how they can be used to approximate the unknown function $\mathcal{T}$.

Gradient boosting is an ensemble method that combines multiple weak learners to create a strong learner. A weak learner is a model that performs at least slightly better than random guessing. The weak learners used in tree-based gradient boosting are regression trees that consist of nodes and leaves. The nodes contain a split criterion, e.g., whether a feature is above or below a certain value. Then the data is passed through the tree until it reaches a leaf node, and the output variable of that leaf node is returned. This output variable is usually the average of the training set ending up in that leaf.

The splits are chosen to minimize the loss function. Using the example from the previous subsection, this could be the difference between the predicted and the actual time to default on loans. To find an optimal split for a feature, the algorithm searches for the splitting value that results in the most homogenous groups regarding their target variables (time to default). In the case of personal loans, the regression tree would look for the best split and may find that borrowers with a monthly income below a certain threshold (X) tend to struggle to service their loan. Thus, the regression tree would split the data into two groups, borrowers with income below X and above X. This continues recursively for the two groups until the algorithm finds no more meaningful splits. For new predictions, the model would return the predicted survival time of the leaf node the data point ends up in (which is the average from the training data in each leaf).

This example uses only one feature to build the regression tree. In reality, the regression tree would use all available features to build the tree, and for every split would choose the feature with the best-split point. For example, the optimal split points for borrowers with low income could be the country of residence as the cost of living and income level might differ significantly. This would introduce an interaction effect between the income and the country of the borrower. In this process, regression trees grow in complexity quickly. For example, adding independent binary features may double the tree size. Instead of doubling the leaves with every binary feature, we could just add additional trees with two leaves each. This is one of the key benefits of ensemble learning methods. Ensemble learning methods combine multiple weak learners to create one strong learner. For example, *Random Forests* simply train multiple trees at once and combine their outputs, e.g., by calculating their mean or voting in the case of classification.

In boosting, however, trees are trained sequentially. Each additional tree is trained on the summed residuals of the previous trees. The residuals are the difference between the predicted and the actual value. This way, each tree tries to correct the error of the previous trees. The trees can then be combined using simple (weighted) addition.

The *XGBoost* package used in this study is an implementation of a boosted regression tree. *XGBoost* uses ML principles like regularization to prevent overfitting, and achieves breakthrough performance in speed and memory usage using ("extreme") optimization methods (Chen & Guestrin, 2016).

B.5 Model explanation using SHAP

Even though ML methods are powerful tools for prediction, they are often considered black boxes (Van Liebergen, 2017). For example, the resulting XGBoost models used in this analysis consist of up to 2594 trees with a maximum depth of up to 10 node layers. With over a million nodes ($2594 * 2^9$), it is difficult to gauge how the model makes its predictions from the model parameters. Instead, we can explain the model by looking at the feature contributions to the model output. That is the difference in the prediction of the model with and without that feature.

This is quite simple in the linear case. For example, the feature contributions of a feature x_j in the model $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is simply the product of its coefficient β_j and its value x_j, minus the expected value of this product:

$$\phi_j(x) = \beta_j x_j - E(\beta_j x_j). \tag{7}$$

For more complex ML models like gradient boosting, this is not as straightforward.

In our example, the effect of income on the time to default could be dependent on other variables, like the country of the borrower. These interaction effects are often present in ML models, where the impact of a feature on model output may depend on the values of other features. So how can we consider these interaction effects when estimating the contribution of each feature to the model output?

One solution to this problem stems from cooperative game theory. Shapley values (Shapley, 1952) are a unique solution to the problem of fairly distributing the gain (e.g., pay-off) of a cooperative game among its players. In a cooperative game, the interactions between the players are important. One player who works well in a team may improve the productivity of the whole team more than a player who struggles working in a team. Shapley's values take these effects into account by comparing the achievements of the whole team with and without each player. Every player is then assigned a Shapley value that represents their marginal contribution to the outcome (e.g., the estimated additional pay-off that the team achieved with, compared to without the player). Thus, adding up all Shapley values yields the total gain of the game. We can use this solution for our problem of explaining ML predictions by distributing the prediction contributions (pay-off) optimally among the features (players). See (Lundberg et al., 2019) for details on this estimation in the XGBoost case.

Shapley Additive Explanations (SHAP) can estimate pair-wise feature contributions as well, yielding the interaction effects of any feature pair. By removing these interaction values from the SHAP value of a feature, we can isolate the main effect. The main effect is often easier to interpret, while the interaction effects can help understand how a feature interacts with other features in a model. We will use this approach for two features in the results section.

To explain the overall decision-making of a model and not just single predictions, we can calculate the SHAP values of a large sample and analyze the global relationship of feature values and model output (Lundberg et al., 2020). In our study, we use SHAP dependence plots to visualize our models.

While other methods (e.g., *LIME*, Ribeiro et al., 2016) can also be used to explain ML models, SHAP has several properties desirable for this analysis: Most importantly, the SHAP framework is model agnostic. This means that it can be applied to explain and compare the predictions of any model, independent of the model's inner workings. Additionally, it is the only method that can satisfy three desirable explanation attributes: local accuracy (single explanations capture the difference between expected model output and the prediction), missingness (missing features get an attribution of zero), and consistency (if a changed model increases the impact of a feature, its SHAP value will not decrease). Providing this unique solution to these criteria, Lundberg and Lee's SHAP framework unified six popular additive model explanation methods, including *LIME*, at the time of its publication. Furthermore, SHAP is well-optimized for some ML models. The XGBoost implementation uses graphical processing units to estimate SHAP values (Lundberg et al., 2019). This provides a significant speedup compared to other methods and makes it feasible to calculate SHAP values for large datasets (Lundberg et al., 2020) like the Bondora dataset used in this study.

C Model performance: *c*-index

This appendix takes a closer look at the performance of the models using the concordance index (c-index). We begin by discussing its calculation and meaning.

Harrell's c-index (Harrell et al., 1982) measures the rank correlation between predicted risk scores and the actual survival times. It is defined as the ratio of concordant pairs to comparable pairs. Two observations are comparable when we can determine whether one of the observations has a shorter survival time than the other. A pair of comparable observations is concordant when the observation with the shorter survival time also has a higher predicted risk score. In other words, the c-index is the probability that the model correctly predicts the survival order of two randomly chosen loans. A c-index of 0.5 would indicate model randomness, while a c-index of 1.0 would indicate perfectly right and a c-index of 0.0 perfectly wrong predictions.

While the c-index is a popular metric for evaluating survival models, it has some limitations. With larger numbers of censored observations, the c-index becomes too optimistic (Uno et al., 2011). To account for this issue, Uno et al. (2011) propose the Inverse Probability of Censoring Weighting (IPCW) c-index that takes the censoring distribution of the data into account. As we have active, censored loans in our dataset, we additionally use this estimator to evaluate the models.

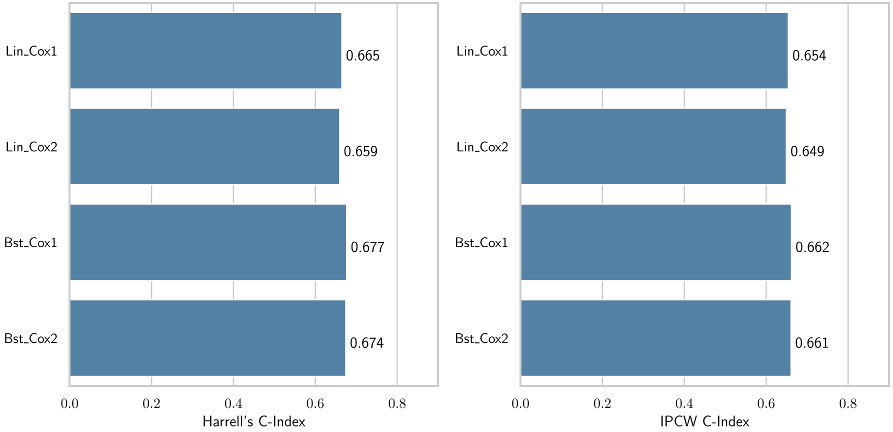


Fig. 11 Harrell's *c*-index (left) and the IPCW *c*-index (right) of the models. "Lin_" prefixes linear models, while "Bst_" prefixes the boosted models. Additionally, a "1" denotes models that do not use the Bondora rating variables for their prediction (which is true for all models discussed in the main text). A "2" denotes models that also use the Bondora rating. We use the python package *lifelines* (Davidson-Pilon, 2023) to estimate the *c*-indices

Figure 11 presents Harrell's *c*-index and the more conservative censoring independent IPCW *c*-index for all models. The results indicate that the boosted models perform better than the linear models. In addition, the inclusion of the Bondora rating only slightly improves performance. This suggests that the ratings do not encode much additional information useful to the models. Furthermore, the IPCW *c*-index is slightly lower than the regular *c*-index for all models. This is expected, as the IPCW *c*-index is a more conservative measure when a large number of subjects is censored.

In general, it is worth noting that although the c-index for all models is significantly higher than 0.5 (i.e. not random), it is still relatively low (< 0.7). The reasons for this are twofold. First, while sampling the test after the training data provides a more realistic setting, the portfolios strongly fluctuating risk profile and an overall shift toward less risky loans may be challenging to account for the models. As a robustness check, we reran the models with random sampling, 5-year loans and monthly time periods, which resulted in a *c*-index above 0.75. Second, by definition, the *c*-index is highly sensitive to slight differences in survival time. While in practice, a difference in loan survival time of one day likely does not matter, it can lower the *c*-index if the loan risks are not ranked in the same order. This again illustrates that the importance of these indicators should not be overstated. In our case, the metrics in the main analysis (Sect. 4.1) are likely to be more relevant to stakeholders.

As a further robustness check, we tested the performance of ML AFT models instead of Cox models. Notably, both performed very similarly. However, we did observe that the AFT model was more difficult to tune as two additional parameters regarding the survival distribution needed to be set. This slowed optimization convergence significantly. The similar performance supports the idea that the models are similarly restricted because they cannot model time-varying effects. As the proportional hazards assumption of the Cox model may be violated, further research could investigate whether less restrictive models could perform better in this context. For example, instead of setting the survival time of early repayment to the loan duration, repayments could be predicted and modeled as a competing risk. Other classical models that performed well on loan data (Dirick et al., 2017) could be implemented and tested as ML models, and advanced ML-models like boosted survival trees (Bai et al., 2022) may yield improved results.

Overall, the performance results support the findings of prior literature that ML models can perform at least as well as linear survival analysis models (Moncada-Torres et al., 2021) while showing good results when used as classification models (Bai et al., 2022).

D Model performance: linear rating model

See Table 3.

E Coefficients of the linear Cox models

The following table presents the results of the linear Cox model. Stars indicate the statistical significance of the coefficients. The table shows the coefficients of the two linear models (with and without Bondora ratings) and the standard errors in parentheses (Table 4).

Table 3 Mean values grouped by linear Cox ratings

Rating	Count	$\overline{\text{Amount}}$	$\overline{\text{Interest}}$ (%)	$\overline{\text{Defaults}}$ (%)	$\overline{\text{IRR}}$ (%)
AA: (− inf, 0.55]	858	2005.78	28.20	15.85	13.23
A: (0.55, 0.65]	906	2364.26	28.77	24.72	10.55
B: (0.65, 0.75]	1020	2589.93	28.90	34.51	6.45
C: (0.75, 0.97]	907	2518.62	30.73	49.17	− 0.69
D: (0.97, 1.51]	768	2848.00	37.55	39.58	− 3.25
E: (1.51, 2.1]	1412	2961.59	39.38	57.01	− 18.95
F: (2.1, inf]	331	2791.44	50.01	77.64	− 47.04

The table includes the number of loans, the mean amount, interest rate, default rates and IRR for each group. Loan status for the default rate and IRR is based on the approach of Dömötör et al. (2023). The HR thresholds are shown in parentheses

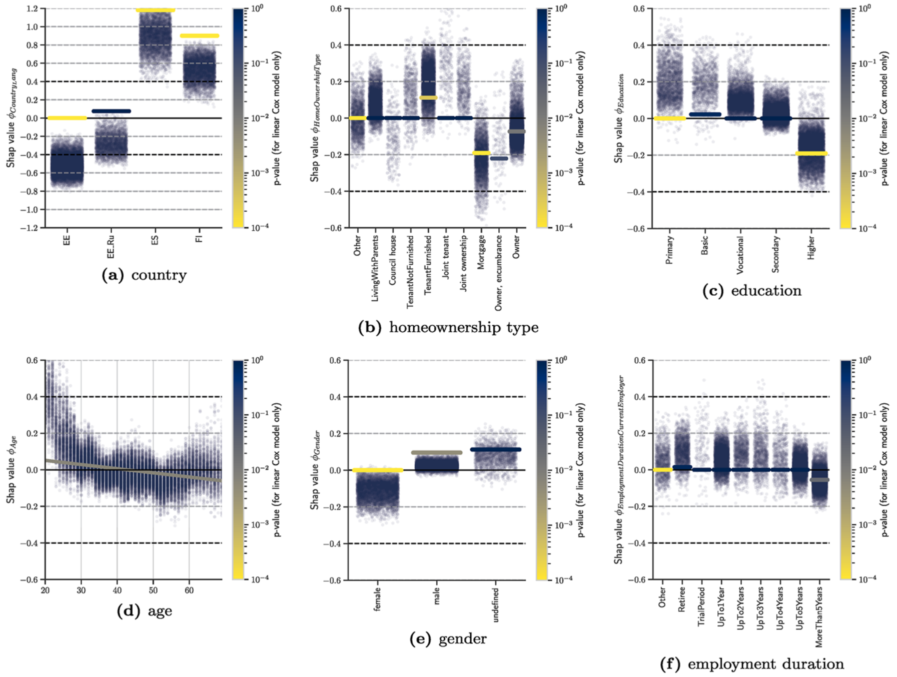
Table 4 Cox proportional hazard model coefficients

Parameters	Cox_1	Cox_2
IncomeTotal	− 0.0 (0.0)	− 0.0 (0.0)
DebtToIncomeModeled	0.0 (0.0005)	0.0 (0.0)
LiabilitiesTotal	0.0 (0.0)	− 0.0 (0.0)
ExistingLiabilities	0.0106* (0.0033)	− 0.0 (0.0)
AppliedAmount	0.0*** (0.0)	0.0 (0.0)
AmountOfPreviousLoansBeforeLoan	0.0 (0.0)	− 0.0 (0.0)
PreviousRepaymentsBeforeLoan	0.0* (0.0)	0.0 (0.0)
RepaymentRatio	− 0.7338*** (0.0533)	− 0.5419*** (0.0395)
PreviousEarlyRepaymentsBefoleLoan	0.0* (0.0)	0.0 (0.0)
NoOfPreviousLoansBeforeLoan	− 0.0353*** (0.0075)	− 0.0285*** (0.0058)
PreviousEarlyRepaymentsCountBeforeLoan	− 0.1977*** (0.0403)	− 0.06* (0.0262)
Country_Lang_EE_Ru	0.1866*** (0.0332)	0.0696* (0.0343)
Country_Lang_ES	0.9406*** (0.0416)	1.1699*** (0.0276)
Age	− 0.0007 (0.0009)	− 0.002* (0.0009)
Country_Lang_FI	0.7053*** (0.0331)	0.888*** (0.0269)
Gender_male	0.1497*** (0.0252)	0.0757* (0.0243)
Gender_undefined	0.1895*** (0.0424)	0.0938* (0.0425)
Education_Basic	0.0205 (0.0497)	0.0478 (0.0434)
Education_Vocational	− 0.1239** (0.0343)	− 0.0 (0.0)
Education_Secondary	− 0.0875* (0.0318)	− 0.0 (0.0)
Education_Higher	− 0.3069*** (0.0328)	− 0.2082*** (0.0207)
EmploymentDurationCurrentEmployer_Retiree	0.1291* (0.0543)	0.0157 (0.0424)
EmploymentDurationCurrentEmployer_TrialPeriod	0.0355 (0.1262)	0.0 (0.0003)
EmploymentDurationCurrentEmployer_UpTo1Year	0.0171 (0.0457)	0.0 (0.0001)
EmploymentDurationCurrentEmployer_UpTo2Years	0.0161 (0.0586)	− 0.0 (0.0001)
EmploymentDurationCurrentEmployer_UpTo3Years	0.0116 (0.062)	− 0.0 (0.0001)
EmploymentDurationCurrentEmployer_UpTo4Years	− 0.0346 (0.0716)	− 0.0 (0.0002)
EmploymentDurationCurrentEmployer_UpTo5Years	0.0293 (0.0447)	0.0 (0.0001)
EmploymentDurationCurrentEmployer_MoreThan5Years	− 0.0484 (0.0431)	− 0.0624* (0.0204)
HomeOwnershipType_LivingWithParents	− 0.014 (0.0409)	0.0 (0.0001)
HomeOwnershipType_Council house	− 0.1425 (0.1014)	− 0.0 (0.0003)
HomeOwnershipType_TenantNotFurnished	− 0.0821 (0.0645)	− 0.0 (0.0002)
HomeOwnershipType_TenantFurnished	0.0868* (0.0384)	0.1076*** (0.026)
HomeOwnershipType_Joint tenant	0.0502 (0.0877)	0.0 (0.0003)
HomeOwnershipType_Joint ownership	0.0673 (0.0684)	0.0 (0.0002)
HomeOwnershipType_Mortgage	− 0.28*** (0.043)	− 0.1727*** (0.0319)
HomeOwnershipType_Owner, encumbrance	− 0.476** (0.135)	− 0.2549* (0.115)
HomeOwnershipType_Owner	− 0.1273** (0.0373)	− 0.0794* (0.0247)
VerificationType_income	0.0376 (0.0454)	− 0.0204 (0.0404)
VerificationType_income and expenses	− 0.0269 (0.0206)	0.0 (0.0)
NewBondoraCustomer_True	− 0.0246 (0.0279)	0.076* (0.0268)
NanRepaymentHistory_True	− 0.2317*** (0.0273)	− 0.1922*** (0.027)

Table 4 (continued)

Parameters	Cox_1	Cox_2
NanEarlyRepayment_True	– 0.3377*** (0.028)	– 0.2455*** (0.0246)
Interest	0.0021* (0.0007)	
ExpectedLoss	0.0211 (0.2302)	
ExpectedReturn	– 0.5082* (0.2032)	
ProbabilityOfDefault	1.2587*** (0.115)	
LossGivenDefault	– 0.3897*** (0.0475)	
Rating_A	0.0235 (0.0577)	
Rating_B	– 0.0065 (0.0478)	
Rating_C	0.0792 (0.0457)	
Rating_D	0.276*** (0.046)	
Rating_E	0.2142*** (0.0502)	
Rating_F	0.2208*** (0.0567)	
Rating_HR	0.0201 (0.0755)	
Observations	31517	31517
c_cph train	0.711838	0.701238
c_cph test	0.665275	0.659625
LR	5973.63661	4449.765007
LR_p	0.0	0.0

$*p < 0.01; **p < 0.001; ***p < 0.0001$



(a) country

(b) homeownership type

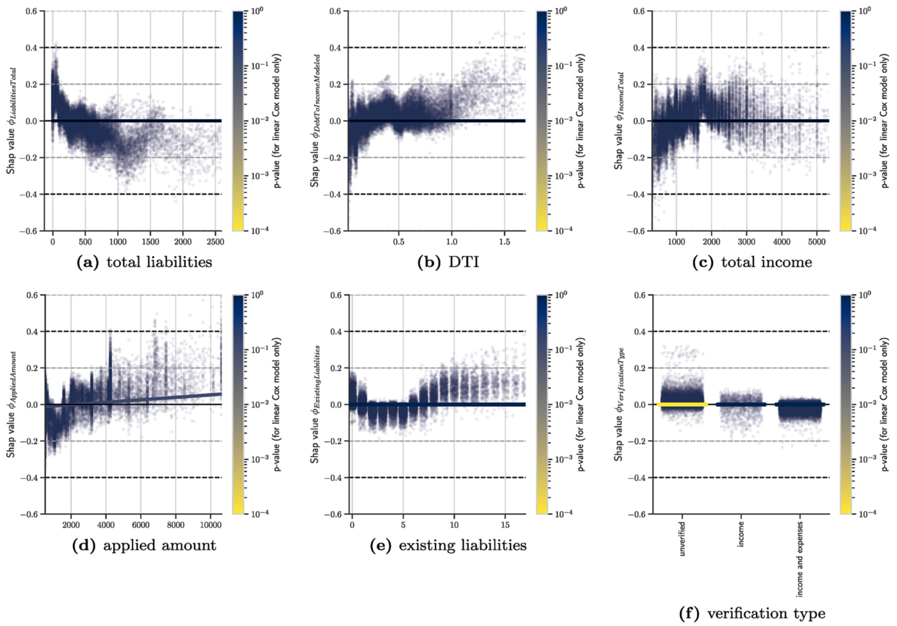
(c) education

(d) age

(e) gender

(f) employment duration

Fig. 12 SHAP values—demographics



(a) total liabilities

(b) DTI

(c) total income

(d) applied amount

(e) existing liabilities

(f) verification type

Fig. 13 SHAP values—financials

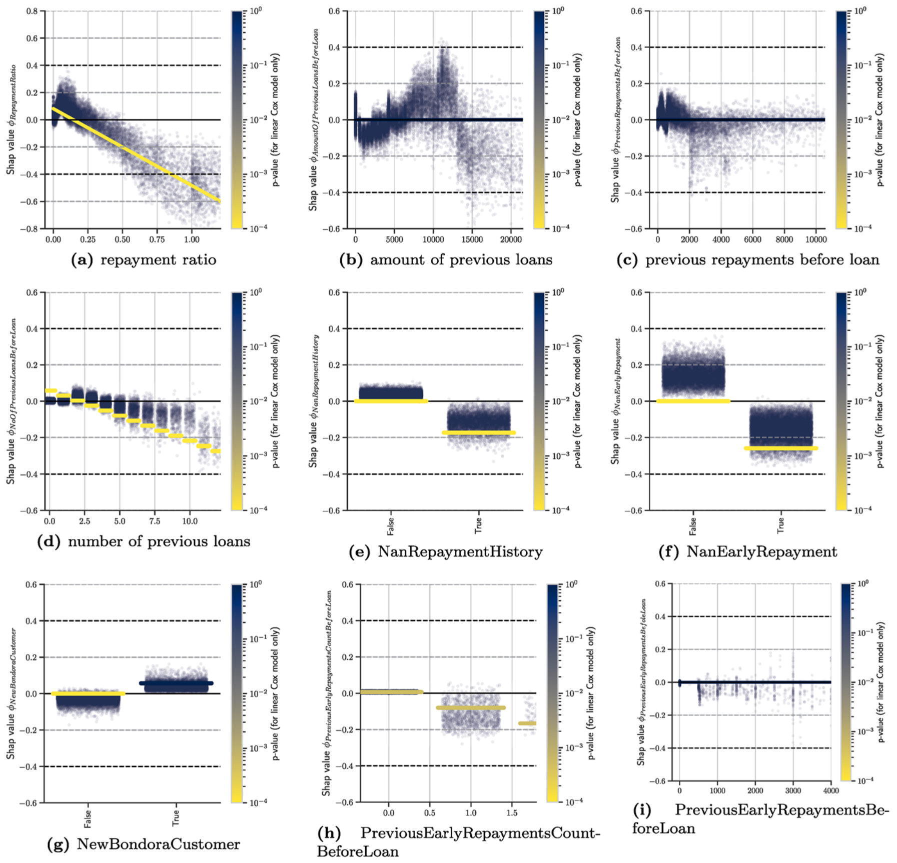
F Dependence plots for all features

In this appendix, the dependence plots for all features of the categories demographics, financials, and borrowing history are presented. Within categories, the plots are ordered by the magnitude of their mean absolute SHAP value (i.e., feature importance).

Demographics Overall the Shapley values of the linear and boosted Cox model appear to be quite similar in magnitude and direction for the demographic feature contributions (Fig. 12), except for centering differences (e.g., for country). Additionally, the boosted model shows vertical dispersion, the largest overall for the home-ownership type indicating interaction effects. The faint visibility in some categories for the boosted model (e.g., TrailPeriod) is caused by the small count present in the training set. In these cases, the linear models found the feature to be less or even insignificant (blue coloring). The boosted model found a nonlinearity for the feature interaction with age.

Financials For the financials, the differences between the linear and boosted models are more apparent. The boosted model found nonlinear relationships where the linear models found no significant relationship at all (Fig. 13a–c). Existing liabilities were found to decrease risk until 3–5 liabilities, and then increase risk slightly (Fig. 13e). The verification type (Fig. 13f) had a small impact on the risk, with a slight decrease for verified borrowers.

Borrowing history Regarding the borrowing history, the models again find similar relationships (Fig. 14a, d, f, g). However, again there are two features not found



(a) repayment ratio **(b)** amount of previous loans **(c)** previous repayments before loan

(d) number of previous loans **(e)** NanRepaymentHistory **(f)** NanEarlyRepayment

(g) NewBondoraCustomer **(h)** PreviousEarlyRepaymentsCountBeforeLoan **(i)** PreviousEarlyRepaymentsBeforeLoan

Fig. 14 SHAP values—borrowing history

very significant in the statistical model but quite important in the boosted model (Fig. 14b and c).

Author contributions The initial idea and concept of the study was developed by GBW under the guidance of FR. GBW conducted the initial literature research, analyzed the data, built the prediction and rating models, calculated the Shapley values, and prepared all figures using Python and CUDA. GBW wrote the first draft of the paper as part of a thesis, which FR supervised. Subsequently, FR contributed to the further development of the methodology, in particular regarding financial aspects, and the final draft. Both authors reviewed and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability All datasets used in this study were obtained from publicly available data sources (see references in Section 3).

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Akiba, T., et al. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '19: The 25th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 2623–2631). ACM. ISBN:978-1-4503-6201-6. https://doi.org/10.1145/3292500.3330701 (visited on 07/18/2023).

Albert, S. M., & Duffy, J. (2012). Differences in risk aversion between young and older adults. Neuroscience and Neuroeconomics. ISSN:2230-3561. https://doi.org/10.2147/NAN.S27184. PMID:24319671. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3852157/ (visited on 07/26/2023).

Ariza-Garzón, M. J., et al. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access, 8*, 64873–64890. ISSN:2169-3536. https://doi.org/10.1109/ACCESS.2020.2984412

Bai, M., Zheng,Y., & Shen, Y. (2022). Gradient boosting survival tree with applications in credit scoring. *Journal of the Operational Research Society, 73*(1), 39–55. ISSN:0160-5682, 1476-9360. https://doi.org/10.1080/01605682.2021.1919035 (visited on 07/10/2023).

Bavoso, V. (2020). The promise and perils of alternative market-based finance: The case of P2P lending in the UK. *Journal of Banking Regulation, 21*(4), 395–409. ISSN:1750-2071. https://doi.org/10.1057/s41261-019-00118-9 (visited on 06/04/2023).

Bergstra, J., et al. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (Vol. 24). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html (visited on 07/25/2023).

Bondora. (2016). *Primary market will be removed from the user interface on November 1st*. Bondora Blog. https://www.bondora.com/blog/primary-market-removed-from-user-interface-on-november-1st/ (visited on 06/23/2023).

Bondora. (2023a). *Bondora public reports*. Bondora.com. https://www.bondora.com/en/public-reports (visited on 06/04/2023).

Bondora. (2023b). *BondoraApi*. https://api.bondora.com/swagger/docs/v1 (visited on 05/25/2023).

Bondora. (2023c). *How are Bondora risk ratings calculated?* Bondora Support. https://support.bondora.com/en/how-are-bondora-risk-ratings-calculated (visited on 05/29/2023).

Breslow, D. R. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187–220. ISSN:0035-9246. JSTOR:2985181. https://www.jstor.org/stable/2985181 (visited on 05/22/2023).

Bussmann, N., et al. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*(1), 203–216. ISSN:1572-9974. https://doi.org/10.1007/s10614-020-10042-0 (visited on 07/29/2023).

Caglayan, M., et al. (2020). Asset mispricing in peer-to-peer loan secondary markets. *Journal of Corporate Finance, 65*, 101769. ISSN:0929-1199. https://doi.org/10.1016/j.jcorpfin.2020.101769. https://www.sciencedirect.com/science/article/pii/S0929119920302133 (visited on 07/31/2023).

Chen, J., Zhang, Y., & Yin, Z. (2018). Education premium in the online peer-to-peer lending marketplace: Evidence from the big data in China. *The Singapore Economic Review, 63*(01), 45–64. ISSN:0217-5908. https://doi.org/10.1142/S0217590818410023. https://www.worldscientific.com/doi/abs/10.1142/S0217590818410023 (visited on 02/17/2024).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16* (pp. 785–794). ACM. ISBN:978-1-4503-4232-2. https://doi.org/10.1145/2939672.2939785

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 187–220. ISSN:0035-9246. JSTOR:2985181. https://www.jstor.org/stable/2985181 (visited on 06/22/2023).

Davidson-Pilon, C. (2023). *Lifelines, survival analysis in Python*. Zenodo. https://doi.org/10.5281/zenodo.7883870. https://zenodo.org/record/7883870 (visited on 07/02/2023).

De Roure, C., Pelizzon, L., & Tasca, P. (2016). How does P2P lending fit into the consumer credit market? *Discussion Paper Deutsche Bundesbank No 30/2016*. https://www.bundesbank.de/resource/blob/704046/b53dc281b4666672e6d526a35e50fd50/mL/2016-08-12-dkp-30-data.pdf

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: A benchmark study. *Journal of the Operational Research Society, 68*(6), 652–665. ISSN:1476-9360. https://doi.org/10.1057/s41274-016-0128-9 (visited on 05/22/2023).

Dömötör, B., Illés, F., & Ölvedi, T. (2023). 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, 86*, 101801. ISSN:10424431. https://doi.org/10.1016/j.intfin.2023.101801. https://linkinghub.elsevier.com/retrieve/pii/S1042443123000690 (visited on 09/07/2023).

Dudley, C. L. (1972). A note on reinvestment assumptions in choosing between net present value and internal rate of return. *The Journal of Finance, 27*(4), 907–915.

Emekter, R., et al. (2015). Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics, 47*(1), 54–70. ISSN:00036846. https://doi.org/10.1080/00036846.2014.962222. https://search.ebscohost.com/login.aspx?direct=true &db=bth &AN=99283419 &site=ehost-live (visited on 06/29/2023).

Harrell Jr., F. E., et al. (1982). Evaluating the yield of medical tests. *JAMA, 247*(18), 2543–2546. ISSN:0098-7484. https://doi.org/10.1001/jama.1982.03320430047030 (visited on 06/20/2023).

Jiang, C., et al. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research, 266*(1), 511–529. ISSN:1572-9338. https://doi.org/10.1007/s10479-017-2668-z (visited on 07/24/2023).

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481.

Kaufman, S., Rosset, S., & Perlich, C. (2011). Leakage in data mining: Formulation, detection, and avoidance. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (Vol. 6, pp. 556–563). https://doi.org/10.1145/2020408.2020496

Kleinbaum, D. G., & Klein, M. (2012). Introduction to survival analysis. In D. G. Kleinbaum, & M. Klein (Eds.), *Survival analysis: A self-learning text. Statistics for biology and health* (pp. 1–54). Springer. ISBN:978-1-4419-6646-9. https://doi.org/10.1007/978-1-4419-6646-9_1(visited on 05/22/2023).

Kurnianingsih, Y. A., et al. (2015). Aging and loss decision making: In- creased risk aversion and decreased use of maximizing information, with correlated rationality and value maximization. *Frontiers in Human Neuroscience, 9*, 280. ISSN:1662-5161. https://doi.org/10.3389/fnhum.2015.00280. PMID:26029092. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4429571/ (visited on 07/26/2023).

Ley, C., et al. (2022). Machine learning and conventional statistics: Making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy, 30*(3), 753–757. ISSN:1433-7347. https://doi.org/10.1007/s00167-022-06896-6 (visited on 07/13/2023).

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: Evidence from a lending platform in China. *Applied Economics, 49*(35), 3538–3545. ISSN:0003-6846. https://doi.org/10.1080/00036846.2016.1262526 (visited on 07/30/2023).

Liu, X., et al. (2023). Combining machine learning with Cox models to identify predictors for incident post-menopausal breast cancer in the UK Biobank. *Scientific Reports, 13*(1), 9221. ISSN:2045-2322. https://doi.org/10.1038/s41598-023-36214-0. https://www.nature.com/articles/s41598-023-36214-0 (visited on 07/25/2023).

Lundberg, S. M., Erion, G., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. ISSN:2522-5839. https://doi.org/10.1038/s42256-019-0138-9. PMID:32607472. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7326367/ (visited on 07/22/2023).

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). Consistent Individualized feature attribution for tree ensembles. arXiv:1802.03888 (visited on 07/24/2023). Preprint.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 04/06/2023).

Lyócsa, S., et al. (2022). Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innovation, 8*(1), 32. ISSN:2199-4730. https://doi.org/10.1186/s40854-022-00338-5 (visited on 07/31/2023).

Mitchell, R., Frank, E., & Holmes, G. (2022). GPUTreeShap: Massively parallel exact calculation of SHAP scores for tree ensembles. arXiv:2010.13972 (visited on 07/24/2023). Preprint.

Moncada-Torres, A., et al. (2021). Explainable machine learning can out-perform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports, 11*(1), 6968. ISSN:2045-2322. https://doi.org/10.1038/s41598-021-86327-7. https://www.nature.com/articles/s41598-021-86327-7 (visited on 04/06/2023).

Reichenbach, F., & Walther, M. (2021). Signals in equity-based crowd-funding and risk of failure. *Financial Innovation, 7*(1), 54. ISSN:2199-4730. https://doi.org/10.1186/s40854-021-00270-0 (visited on 04/12/2023).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. arXiv:1602.04938 (visited on 07/22/2023). Preprint.

Rolison, J. J., et al. (2014). Risk-taking differences across the adult life span: A question of age and domain. *The Journals of Gerontology: Series B, 69*(6), 870–880. ISSN:1758-5368, 1079-5014. https://doi.org/10.1093/geronb/gbt081. https://academic.oup.com/psychsocgerontology/article-lookup/doi/10.1093/geronb/gbt081 (visited on 02/02/2024).

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLoS ONE, 10*(10), e0139427. ISSN:1932-6203. https://doi.org/10.1371/journal.pone.0139427. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139427 (visited on 06/29/2023).

Shapley, L. S. (1952). *A value for N-person games*. RAND Corporation. https://www.rand.org/pubs/papers/P295.html (visited on 06/18/2023).

Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need. arXiv:2106.03253 (visited on 07/13/2023). Preprint.

Suárez-Ramírez, C. D., Martínez, J.-C., & Loyola-González, O. (2022). A novel survival analysis-based approach for predicting behavioral probability of default. In: O. O. Vergara-Villegas, et al. (Eds.), *Pattern recognition. Lecture notes in computer science* (pp. 56–69). Springer International Publishing. ISBN:978-3-031-07750-0. https://doi.org/10.1007/978-3-031-07750-0_6

Suryono, R. R., Purwandari, B., & Budi, I. (2019). Peer to peer (P2P) lending problems and potential solutions: A systematic literature review. *Procedia Computer Science, 161*, 204–214. ISSN:18770509. https://doi.org/10.1016/j.procs.2019.11.116. https://linkinghub.elsevier.com/retrieve/pii/S1877050919318265 (visited on 06/01/2023).

Tan, F., et al. (2019). A deep learning approach to competing risks representation in peer-to-peer lending. *IEEE Transactions on Neural Networks and Learning Systems, 30*(5), 1565–1574. ISSN:2162-2388. https://doi.org/10.1109/TNNLS.2018.2870573.

Teply, P., & Polena, M. (2020). Best classification algorithms in peer- to-peer lending. *The North American Journal of Economics and Finance, 51*, 100904. ISSN:1062-9408. https://doi.org/10.1016/j.najef.2019.01.001. https://www.sciencedirect.com/science/article/pii/S1062940818302262 (visited on 07/31/2023).

Truppia, J. M. (2023). *Package 'tvm'—Time value of money functions*. https://cran.r-project.org/web/packages/tvm/tvm.pdf (visited on 08/30/2023).

Uno, H., et al. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine, 30*(10), 1105–1117. ISSN:1097-0258. https://doi.org/10.1002/sim.4154. PMID:21484848.

Van Liebergen, B. (2017). Machine learning: A revolution in risk management and compliance? *Journal of Financial Transformation, 45*, 60–67.

Willoughby, T., et al. (2021). Is adolescence a time of heightened risk taking? An overview of types of risk-taking behaviors across age groups. *Developmental Review, 61*, 100980. ISSN:02732297. https://doi.org/10.1016/j.dr.2021.100980. https://linkinghub.elsevier.com/retrieve/pii/S0273229721000356 (visited on 02/02/2024).

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: A machine learning methodology. *Scientific Reports, 11*(1), 18759. ISSN:2045-2322. https://doi.org/10.1038/s41598-021-98361-6. https://www.nature.com/articles/s41598-021-98361-6 (visited on 04/14/2023).

Zhang, S., et al. (2015). Value of big data to finance: Observations on an Internet credit service company in China. *Financial Innovation, 1*(1), 17. ISSN:2199-4730. https://doi.org/10.1186/s40854-015-0017-2 (visited on 06/04/2023).

Zhou, J., et al. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and Its Applications, 534*, 122370. ISSN:0378-4371. https://doi.org/10.1016/j.physa.2019.122370. https://www.sciencedirect.com/science/article/pii/S0378437119313652 (visited on 07/29/2023).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.