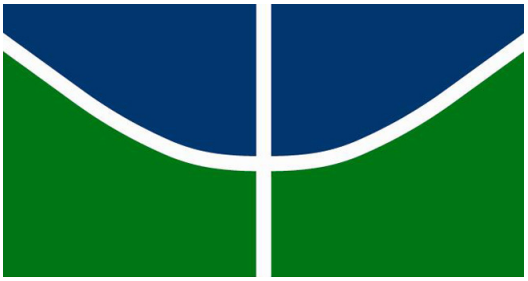


Cayan Atreio Portela Bárcena Saavedra

Essays in Machine Learning Applications in Credit Risk

Brasília - DF

May, 2023

Cayan Atreio Portela Bárcena Saavedra
Essays in Machine Learning Applications in Credit Risk/ Cayan Atreio Portela Bárcena Saavedra. – Brasília - DF, May, 2023-
84 p. : il. (some colored.) ; 30 cm.

Advisor: Dr. Herbert Kimura

Thesis (Ph.D.) – University of Brasilia
School of Economics, Business and Accounting (FACE)
Graduate Program in Management , May, 2023.

1. Machine Learning 2. Survival Analysis 3. Competing Risk 4. Machine Learning Fairness I. Advisor: Dr. Herbert Kimura II. University of Brasilia III. School of Economics, Business and Accounting (FACE) IV. Essays in Machine Learning Applications in Credit Risk

CDU 02:141:005.7

Cayan Atreio Portela Bárcena Saavedra

Essays in Machine Learning Applications in Credit Risk

Thesis submitted to the Graduate Program in Business Administration at University of Brasilia as partial fulfillment of the requirements for attainment of Ph.D. degree in Business Administration, with major in Finance and Quantitative Methods.

The examining committee, as identified below, approves this dissertation:

Dr. Herbert Kimura
Advisor

Dra. Juliana Betini Fachini Gomes
University of Brasília

Dr. Fabiano Guasti Lima
University of São Paulo

Dr. Leonardo Fernando Cruz Basso
Mackenzie Presbyterian University

Brasília - DF
May, 2023

To my mom, the most intelligent person I've ever known.

Abstract

This dissertation explores applications of machine learning models in credit risk. Statistical and machine learning techniques are investigated, seeking to develop alternative methods in the credit risk modeling pipeline, aiming at comply with standards and regulations. We develop three papers in this dissertation, analyzing different aspects of credit risk using machine learning. In the first paper, Algorithmic Credit Analysis and the use of Discriminatory Variables, concerning machine learning fairness and the use of sensitive variables. In the second paper, Lifetime Probability of Default with Survival Analysis and Ensemble Methods, application of survival analysis models for the entire time maturity of a credit operation. Finally, in the third paper, Credit Risk Assessment with Machine Learning and Competing Risk Survival Analysis Models, an adaptation in competing risks subdistribution hazards. In the three applications, different machine learning models are explored, and the results are discussed, aiming to contribute to the credit risk literature.

Keywords: Machine Learning; Survival Analysis; Competing Risk; Machine Learning Fairness.

List of Figures

List of Tables

List of abbreviations and acronyms

AI	*Artificial Intelligence*
ML	*Machine Learning*
IFRS 9	*International Financial Reporting Standards 9*
ECL	*Expected Credit Loss*
PD	*Probability of Default*
EAD	*Exposure at Default*
LGD	*Loss Given Default*
LR	*Logistic Regression*
DT	*Decision Tree*
RF	*Random Forest*
GBM	*Gradient Boosting Machine*
SVM	*Support Vector Machine*
SA	*Survival Analysis*
Cox PH	*Cox Proportional Hazards*
CWGBSA	*Componentwise Gradient Boosting Survival Analysis*
GBSA	*Gradient Boosting Survival Analysis*
RSF	*Random Survival Forests*

Contents

1 Introduction

1.1 Contextualization

The dynamic changes of contemporary world keeps getting fostered by technology evolution. Exponential advances in computing process power made feasible the process and management of large amounts of data. Choices and interactions that we make every day (even small and unnoticed ones) can generated information that can be used in further analysis. In addition, the fields of applications range from banking and securities, healthcare, education, communication, public policy and so on.

The availableness and use of such information led to what we now know as Artificial Intelligence (AI). It can be defined as the use of data to train computers that simulate human behavior in some aspects, e.g., face recognition, talk and interactions, learning, decision making etc (XU; LU; LI, 2021). AI permeates a multidisciplinary field of knowledge and has achieved exceptional results over the past few decades in many applications, such as speech and image recognition, natural language processing and intelligent systems (DUAN et al., 2009).

One of the enabling drivers of AI is machine learning (ZHANG; LU, 2021) which can be defined as the use of an algorithm that improves its performance by learning from new data (NILSSON, 2014). The use and development of machine learning by top companies has been increasing. According to a McKinsey survey *The State of AI in 2021* (CHUI M.; SUKHAREVSKY, 2021) AI adoption keeps rising, specially at companies headquartered in emerging economies. Common applications founded includes service operations, product and service development, marketing and sales and risk management. The survey also states that companies who are getting more profitable results from AI takes advantage of cloud technologies and advanced practices, such as Machine-Learning Operations (MLOps). The adoption of such practices can enhance usability in a specific service leading to growth lift, e.g., by using near-real-time predictions to customers. In this way, the use of data by industry is in a mature status, but constantly rising.

Concerning financial industry, one of the main use of such tools relates do credit risk management. Credit risk is considered to be the risk of a counterparty not honoring its financial obligation in accordance to agreed terms of a credit operation, representing potential loss. Therefore, banks and financial institutions need to manage the credit risk underlying their portfolio and individual transactions (SUPERVISION; SETTLEMENTS, 2000). The sub-prime mortgages credit crunch which started in 2006, for instance, have underlined a major impact of credit risk and credit risk management on the wellbeing and

profitability of business (BROWN; MOLES, 2014). In this way, lending decisions must be made cautiously and in a structured manner in order to avoid significant losses.

In such manner, the assessment of credit risk management plays a vital role on firm financial health. This assessment aims to measure three important characteristics on credit operations, such as: the exposure amount, the likelihood of repayment and the recovery rate. Consequently, decision on whether to grant the loan or not has to be made upon estimates on these three events. The approach on tackling this sort of problem may vary from expert judgment, relationship models or quantitative methods, with the latter being widely used with statistical methodologies and machine learning models. This work aims to reflect some of the use of machine learning for risk management, more specifically, in estimating probability of default, i.e., the parameter that reflects the likelihood of a borrower debt repayment.

When using machine learning models for estimating probability of default, two points of attention may arise from the perspective of a financial institution: i) machine learning fairness and ii) International Financial Reporting Standard 9, IFRS 9 (International Accounting Standards Board, 2014) compliance. Firstly, machine learning fairness is of utmost importance as it can be conceptualized as equality of opportunity, ensuring non-discrimination in automated decision-making processes (BAROCAS; HARDT; NARAYANAN, 2017). When algorithms are used to estimate default probabilities, it is crucial to ensure that the models are not biased or unfair towards specific individuals or groups. Fairness in machine learning algorithms can help prevent potential ethical issues, protect against legal challenges, and promote transparency and trustworthiness in the financial industry. Secondly, adherence to IFRS 9 (International Accounting Standards Board, 2014) is essential for financial institutions as it provides a standardized framework for recognizing and measuring credit impairments. IFRS 9 requires financial institutions to incorporate forward-looking information and consider expected credit losses when estimating default probabilities. By following IFRS 9 guidelines, institutions can enhance their risk assessment and provisioning processes, leading to more accurate and reliable estimates of expected credit losses.

In this way, this dissertation explores applications on machine learning models to estimate probability of default, addressing the points described above. The first paper, Algorithmic Credit Analysis and the use of Discriminatory Variables, addresses the challenge in machine learning fairness by comparing different pipeline possibilities that do not include sensitive variables. The second paper focus on IFRS 9, applying survival models to estimate probability of default, since this approach provides an estimate throughout the duration of the credit operation, thus adhering to regulation. The third paper, explores alternative on survival models in credit risk, by considering two competing risk events: default and early payment.

Chapter 2, uses a dataset of overdraft transactions to estimate probability of default. The studie analyze classification performance metrics when variables such as gender, education, and age are withdrawn from the models. Different classification models and pipeline configurations, with distinct balancing techniques and longitudinal analysis, are compared, seeking alternatives to the use of sensitive information while keeping a good model.

Chapter 3, shows an application of machine learning methods to risk management. Specifically, it investigates alternative approaches in order to conforms to regulation requirements stated by International Financial Reporting Standards 9 (IFRS 9) issued by the International Accounting Standards Board (IASB), which calls for estimates of a lifetime expected credit losses rather than a single point estimation for the specific time period. This can be achieved by using Survival Analysis since it consider the time until an event occurs, with decreasing survival probabilities over time. Therefore, it can provide estimations for every discrete period of time, in this context, monthly estimates. The methods are applied to a dataset consisting of refinancing operations, representing borrowers who has already defaulted on previously operations. Operations consisted on 36-month and 60-month time maturitie and different models were developed for each type. Classic Survival Analysis and Machine learning models applied to this context are compared, such as Cox Proportional Hazards (unpenalized and penalized), Survival Trees, Random Survival Forests, Gradient Boosting with regression trees, and Gradient Boosting with component-wise least squares.

Chapter 4 derives from chapter 3, but with a different approach to the survival function. Regarding the 36-month operations from the same data set of refinancing operations described in the second article, this paper aims to approach competing risks modelling, by considering default and early payment as primary and secondary risks, respectively. Since these are mutually exclusive events, reflecting their nature during estimation can bring valuable insights and results. More specifically, the paper consider subdistribution hazards in a component-wise gradient boosting model, comparing its results with other cause-specific survival models.

Therefore, this dissertation aims to contribute to the credit risk literature focused on the estimation of probability of default. Specifically, it address complex situations which calls for strategies that ensures unbiased decision-making and preventing potential ethical issues, and also complying with regulatory requirements, such as IFRS 9, while enabling accurate estimation of default probabilities and enhancing risk management practices.

2 Algorithmic Credit Analysis and the use of Discriminatory Variables

2.1 Introduction

Machine learning algorithms are being applied in various sectors. Although machine learning can be useful for society, it also imposes challenges, as results of models can reinforce bias or prejudice. In this context, there are initiatives to define regulations and guidelines for a fair use of machine learning. For instance, the European Union's General Data Protection Regulation (GDPR) establishes restrictions on automated individual decision-making, including profiling. The Settlements (FSI Insights on policy implementation, 2021) addresses regulatory expectations on the financial sector, suggesting the prevention of bias in AI models as a guidance. In addition Singapore (Thematic Review, 2022) argues that financial institutions should consider specifying a list of protected attributes and their proxies used on a ML model. In particular, the document states that the use of protected attributes and their proxies as input factors for AIDA-driven (Artificial Intelligence and Data Analysis) decisions should be justified.

Specifically in the financial industry, financial intermediation improves capital allocation by providing a more fluid flow between the participants of the financial system. Regarding this function, loan operations play a key role. The importance of lending has established credit scoring as one of the most successful real-world applications of statistics and operations research (CROOK; EDELMAN; THOMAS, 2007).

Credit scoring enables a prospective assessment of the risk of a loan operation. In particular, it helps the lender to better discriminate between applicants that are more likely to repay a loan from those who are less likely to meet initial agreed terms in a credit transaction. This evaluation mitigates credit risk (i.e., the risk of a potential loss due to the counterparty failing to fulfill its obligations), which represents the main risk that most banking institutions face (APOSTOLIK et al., 2009).

Although credit models typically relied on statistical foundations, more recently, machine learning (ML) scoring models have been applied in the analysis of loan applications (KOZODOI; JACOB; LESSMANN, 2022). Financial institutions have increasingly used ML models to support decision-making in credit risk (CROOK; EDELMAN; THOMAS, 2007).

However, especially for credit scoring, other aspects regarding automated decision-making have been questioned, such as political concerns on civil rights and the dangers of a

possible detracting effect over historically disadvantaged groups. Kauffman e Wang (2001) discuss points raised by the Executive Office of the President of The US, which emphasizes credit scoring as a critical element in the lending business, with a significant impact on society. Additionally, the European Commission establishes guidelines that highlight the need for regular and systemic monitoring of the sector (KAUFFMAN; WANG, 2001).

In this context, credit lending decisions based on historical data can provide biased estimates regarding attributes, e.g., age, race and gender (KALLUS; MAO; ZHOU, 2022). Even when clear discriminatory attributes are not allowed into supervised models, discrimination still can occur, if a variable can be related to some specific demographic characteristic. For instance, Kline, Rose e Walters (2022) identify race discrimination in hiring processes, performing an experiment with job applications with Black-American and White-American names.

Disparities in credit access from individuals who belong to groups with sensitive attributes can be detrimental from ethical and social perspectives. Therefore, a non-discriminatory credit scoring (e.g., for student loans) is an essential tool to democratize socioeconomic opportunities (DE-ARTEAGA; FEUERRIEGEL; SAAR-TSECHANSKY, 2022). Also, legal frameworks, regulations and standards (MAKHLOUF; ZHIOUA; PALAMIDESSI, 2021) bring attention to adjustments that financial institutions may have to perform.

As few studies have assessed fairness in credit models (e.g. Zetten, Ramackers e Hoos (2022)), we contribute to the discussion on fairness artificial intelligence (AI) and discriminatory bias, by investigating the impact of the use of explanatory variables (gender, level of education, and age) in the prediction performance of credit scoring models, using a dataset of loans from overdraft checking accounts.

We apply several machine learning models for predicting default probability on an imbalanced dataset regarding overdraft loans operations of a large Brazilian bank. Different pipeline configurations are analyzed and compared, such as split-validation methods, resampling techniques and different subset of features, by removing some attributes that might be considered sensitive features. Resulting metrics are analyzed to evaluate the impact of pre-processing steps in a pipeline for the predictive modeling. More specifically, we aim at identifying whether sensitive information, which could be discriminatory and impact fairness of machine learning models, can impact prediction accuracy.

The paper is structured as follows. In the next section, we briefly discuss related studies. Then, we describe the data and methods used in the research. We discuss the main results and finally, we present the conclusion, highlighting implications of the paper.

2.2 Theoretical background

Bias in algorithms may reflect poor and unrepresentative datasets, inadequate models, erratic human behavior, resulting in unfair outcomes to individuals Akter et al. (2022). According to Ashok et al. (2022), AI biased by design may divert its original purpose of serving mankind.

In fact, Giffen, Herhausen e Fahse (2022) argue that, when machine learning helps decision-making, bias derived by superficial algorithm designs and human stereotypes may generate discriminatory resolutions, with several negative impacts, in various dimensions such as financial, social, and reputational.

Makhlouf, Zhioua e Palamidessi (2021) contend that in application processes automated screening systems may assign a lower probability of acceptance of applicants whose documents have misspelling and grammatical errors. However, this feature of automated systems may hinder the acceptance of non-native English speakers since these language mistakes would be more common in individuals from specific races and birthplaces. In this context, using machine learning techniques to analyze applicants may weaken diversity and reinforce bias and discrimination.

Particularly to credit risk analysis, AI can exacerbate bias and prejudice, making it harder to some groups of individuals get access to loans. This lack of access may preclude social and economic mobility. Kumar, Hines e Dickerson (2022) point to the phenom of credit invisibility and observational bias. Credit invisibility arises when a model is trained based on historical data that predominantly belongs to a certain demographic group, resulting in disparate predictions among other groups (KUMAR; HINES; DICKERSON, 2022).

In addition, observational bias relates to external factors that change natural conditions and affect loan repayment distribution (e.g. increase in default due to pandemic conditions), but may not generalize well for future applications (KUMAR; HINES; DICKERSON, 2022).

Jagtiani e Lemieux (2019) discuss the use of alternative data on credit scores. Alternative data refers to information that usually is not available in consumers credit files (KUMAR; HINES; DICKERSON, 2022), such as posts and interactions in social networks or inquiries in search engines. Jagtiani e Lemieux (2019) argue that the use of such data has improved credit score of borrowers who have fewer or inaccurate credit records (based on FICO scores), allowing them to get lower-priced credit.

Although alternative sources for data collection may leverage business decision, Uejio e Bureau (2021) emphasize the need for caution, since it may be difficult or impossible to ensure data quality information in social media or online searches. Additionally, the very characteristics of interacting in social networks may become proxies for discri-

minatory attributes. Nevertheless, "less alternative" information, such as cash-flow data, can provide significant predictive power, specially in situations where traditional credit history is not available (FINREGLAB, 2019).

The importance of evaluating the absence of discrimination on automated tasks is fostering the relevance of the concept on fairness in machine learning. Fair ML's objective is to make sure that model predictions adhere to statistical fairness standards (KOZODOI; JACOB; LESSMANN, 2022). In this context, fair ML refers to preventing discrimination against sensitive groups in the decisions made by machine learning models.

Some measures of statistical fairness have been proposed for a better assessment on machine learning predictions. More specifically, Barocas, Hardt e Narayanan (2017) define statistical non-discrimination criteria as statistical formulations that consider random variables and the decision surface, aiming at an absence of discrimination. The authors define such criteria as properties of the joint distribution of sensitive features, target variables and the estimated predicted scores. Three main fairness categories are suggested: independence, separation and sufficiency (BAROCAS; HARDT; NARAYANAN, 2017).

Considering loan applications, and the inherent missclassification costs, Kozodoi, Jacob e Lessmann (2022) suggest separation as a more desirable criterion to evaluate fairness ML by financial institutions. Separation compares the false positive rate and the false negative rate across groups, requiring that both rates are equal regarding sensitive attributes.

According to Makhlouf, Zhioua e Palamidessi (2021), fairness in ML can be categorized into two dimensions: (i) the task and (ii) the type of learning. The former involves two tasks concerning fairness-aware ML: (i) discrimination discovery, which focuses on measuring bias in datasets and model predictions, and (ii) discrimination removal, which prevent discrimination in pre-processing (e.g. manipulating datasets), in-processing (e.g. model adjusting) and post-processing steps (e.g. modifying predictions). The latter separate learning types, such as regression, classification and reinforcement learning.

Legal frameworks and the existence of regulations also highlight an increasing concern in the area (MAKHLOUF; ZHIOUA; PALAMIDESSI, 2021). For instance, Barocas e Selbst (2016) characterize two frameworks: (i) disparate treatment and (ii) disparate impacts. A decision is deemed unfair under the disparate treatment framework if it directly or indirectly uses the person's sensitive attribute information. For disparate impact, a decision is unfair if leads to disproportional outcomes for individuals, according to their sensitive attributes. In this context, financial institutions should be aware of possible changes that their risk scoring machine learning pipeline should face in order to adhere to regulations and standards.

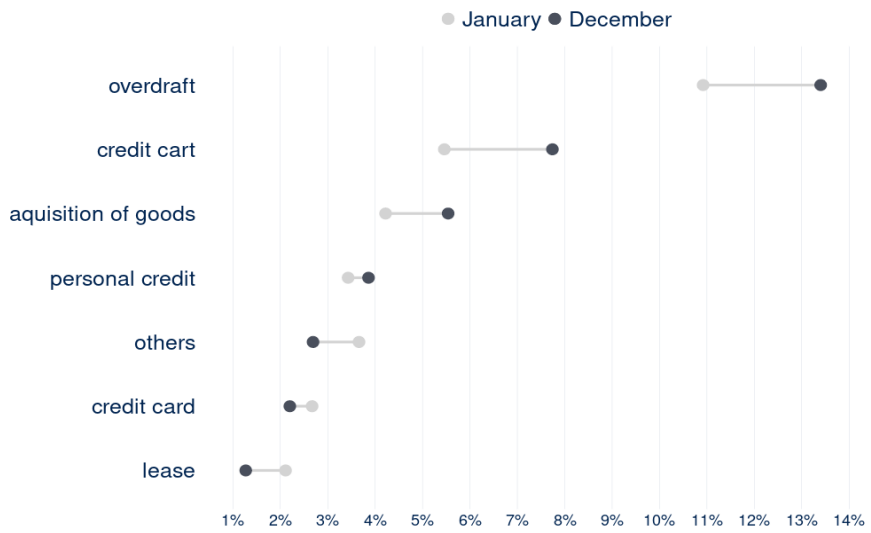


Figure 1 – Range of the monthly interest from 2022 from different loan types (Brazilian Central Bank).

2.3 Data and methods

2.3.1 Dataset

We use a dataset of a large Brazilian financial institution that operates both in the retail and wholesale markets. The dataset contains information regarding overdraft transactions, one of the most common type of personal loans in Brazil. These transactions mix a financial service associated with a checking account and a revolving pre-approved line of credit, without collateral. This loan is usually directed to emergency cash needs and has very high interest rates. Figure 1 shows, using data from the Brazilian Central Bank, the range of the monthly interest rate for a variety of loans for individuals charged by the main financial institutions as of March, 2022. Overdraft accounts have the higher interest rate, with an annual median interest rate of approximately 150%.

Following the Brazilian General Law for Data Protection that took effect in September 2020, the data of the overdraft accounts was anonymized. Some variables were modified, for instance, multiplied by a factor, to avoid disclosure of strategic information of the bank. Therefore, although the descriptive statistics do not represent the real numbers of the bank, the models reflect the relevance of the variables and the prediction performance.

The dataset initially included 168,800 observations (automatic pre-approved loans for overdraft accounts) and 30 columns (attributes of the clients). The study uses the same original dataset as in Santos, Saavedra e Kimura (2023). However due filtering criteria and different scope of the study, the final data for train and test are very different. We

removed redundant columns and columns with more that 60% missing data. Next, we aggregated some features in order to mitigate the percentage of non-available information and extract a random sample of 50% of observations.

This process resulted in a final database of 44,612 observations and 8 columns (7 features and the target variable). The features were: (i) start date of operation, (ii) genre, (iii) level of education, (iv) age, (v) average balance in savings account over the last 2 months, (vi) declared gross income and (vii) the time with declared income. The target variable is default, which is considered in this study as the delinquency defined as operations that exceed 90 days late in any month, during the first 12 months of duration.

We performed a series of analysis to identify how demographic variables can influence prediction results of the credit models. First, we use all available variables to assess performance of the complete models. We apply a variety of ML techniques: k-Nearest Neighbors, Decision Tree, Random Forest and Gradient Boosting. The Logistic Regression is set as the baseline model.

Second, we use a stepwise procedure to sequentially remove variables that could bias the model towards some specific group of borrowers (e.g., genre, age and level of education). These partial models are then compared with the complete models. We assess how the absence of some demographic variables, which could lead to bias or prejudice, impact the forecast of good and bad borrowers.

Additionally, we considered two strategies regarding longitudinal information: (i) completing shuffling observations in the train dataset and (ii) considering the month of the operation in a cumulative time-dependent window (e.g., using information from previous months to predict the next one).

In the first approach, we do not take in consideration differences in grants of overdraft accounts that occur in different months. Therefore, the analysis of granting a loan would not depend on a specific economic situation of the country or strategy performed by the bank to attract new clients in a given month.

In the second approach, the focus is on the specific month of the grant of the pre-approved loan. Since the bank can have different strategies and efforts to entice clients, the risk profile of the pre-approved overdraft accounts may have fluctuations among different months. Using this procedure, we aim to analyze differences in performance metrics considering separate models for each time window.

The train and the test data were splitted in 2/3 and 1/3 of final dataset, respectively. For the strategy considering longitudinal information, two splits were made: (i) leaving out operations in the last month (test data) and completely shuffling remaining operations (training data) and (ii) leaving out operations in the m-th month (test data) and completely shuffling operations on all $m-1$ previous months (training data).

Before fitting each final model, we search for the best hyperparameters considering a 3-fold validation in the train dataset. To compare results, we use different classification methods to predict defaulters and non-defaulters in these overdraft transactions.

2.3.2 Methods

We compare outcomes of classification of default and non-default prediction from the traditional Logistic Regression with outcomes of more recent machine learning techniques: k-Nearest Neighbors, Decision Trees, Random Forest and Gradient Boosting.

Logistic Regression (LR) is the most used classification technique among supervised models. Often defined as the benchmark model in classification problems, LR belongs to the class of generalized linear models (NELDER; WEDDERBURN, 1972) that covers probability distributions related to the exponential family.

In LR, a systematic component is built as a linear combination that carries information about features. The logit function relates the linear predictor to the expectation of the target variable, which is assumed to have a Bernoulli distribution. The model is given by:

$$P(Y = 1|X) = \log\left(\frac{p}{1-p}\right) = \sum_{r=1}^{p} x_{ir}\beta_r \tag{2.1}$$

where Y is the binary dependent variable, X is the vector of independent variables, and $\sum_{r=1}^{p} x_{ir}\beta_r$ is the linear predictor. The coefficients β_r are estimated from the training data using maximum likelihood estimation, equivalent to an iterative least squares process reweighted on an adjusted variable.

k-Nearest-Neighbor Classifiers (kNN) categorizes a given observation considering its k nearest neighbors. Neighbors are chosen by similarity on the feature space using Euclidean distance $d_{(i)} = ||x_{(i)} - x_0||$ (HASTIE et al., 2009). Therefore, given a query point x_0 and its k nearest neighbors $x_{(r)}, r = 1, ..., k$, a new observation is classified using majority vote among the k neighbors.

Decision Trees (DT) divide the feature space into simpler regions (JAMES et al., 2013). A Decision Tree can be represented as a binary tree, where each internal node represents a decision based on a specific feature, and each leaf node represents a class label. The method recursively split the data into two regions, modeling the mean of the dependent variable Y in each region (HASTIE et al., 2009).

The region is chosen considering a split-point to achieve the best fit, measured by the information gain, taking into account the feature provided. Information gain is a measure of the reduction in entropy, or impurity, achieved by splitting the data based on a specific feature, in a particular threshold or region. In this study, we consider two

measures of entropy, Gini and Cross-Entropy. The entropy for each tree-based models were chosen in a grid search combined with other parameters. Therefore, each model selected the entropy measure that was in the hyperparemeter set that led to the best performance during cross validation. Let $H(\cdot)$ be the loss function, Gini and Cross-entropy impurity measures are respectively given by equations 2.2 and 2.3:

$$H_g(Q_m) = \sum_k \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2.2}$$

$$H_{ce}(Q_m) = -\sum_k \hat{p}_{mk}\log(\hat{p}_{mk}) \tag{2.3}$$

where Q_m represents the data at node m with n_m samples and $\hat{p}_{mk} = \frac{1}{n_m}\sum_{y_i \in Q_m} I(y_i = k)$ is the proportion of observations in node m that belongs to class k.

Random Forest (RF) constitutes an ensemble method which is built upon a combination of multiple decision tree predictors (BREIMAN, 2001). Each tree is trained on a random subset of the training data and a random subset of the input features, resulting in a group of *de-correlated* trees (HASTIE et al., 2009). RF uses a bagging algorithm (BREIMAN, 1996), resulting in variance reduction by decreasing the correlation between trees (HASTIE et al., 2009). The process can be illustrated as in Algorithm 2.1 (HASTIE et al., 2009):

Algorithm 2.1 *Random Forest*

1.5

1. For $b = 1$ to B:
   a) Draw a bootstrap sample of size N from the training data.
   b) Grow a decision tree T_b to the selected data into its terminal nodes.
2. Output of the ensemble of trees: T_b

In Random Forests, B subsets are drawn from bootstrap samples for building each b tree and the final prediction is made by averaging them (equation 2.4):

$$\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{2.4}$$

where $T_b(x)$ is the prediction of the b-th tree for a new point x.

Boosting is an ensemble method introduced by Freund e Schapire (1997). In the "AdaBoost.M1."Freund e Schapire (1997), one the most popular boosting algorithm, the

model is built in sequential steps, with different weights on observations given by the prediction errors in the previous steps.

The idea is to sequentially apply weak classifiers to repeated modified versions of the data Hastie et al. (2009). The final predictions are then estimated by a weighted majority vote, described as follows (HASTIE et al., 2009):

$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{2.5}$$

where α_m is the weight of the contribution of each weak classifier G_m. Applying weights in each m step results in a modification of the data set, where observations are individually reweighted according to misclassifications on step $m-1$.

Since the dataset is unbalanced, as it is usual in credit analysis, we perform two balancing techniques to train the models: (i) Synthetic Minority Oversampling Technique and (ii) Bagging for Imbalanced Data.

Synthetic Minority Oversampling Technique (SMOTE) is one of the most popular over-sampling methods for dealing with imbalanced datasets. Proposed by Chawla et al. (2002), it over-samples the minority class with synthetic observations. A new observation x_{new} is generated by computing the difference on the feature space under consideration, as:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \tag{2.6}$$

where x_{zi} is one of the k neareast-neighbors from x_i and λ is a random number between 0 and 1.

In addition, we use an algorithm based on Bagging for Imbalanced Data. Bagging Classifier is an ensemble of models fitted on a random redistribution of the training set (MACLIN; OPITZ, 1997). Bagging for Imbalanced Data provides, for each classifier, a balanced subsample dataset from the training set. Therefore, it builds multiple base learners, trained on balanced observations, aggregating their predictions (HIDO; KASHIMA; TAKAHASHI, 2009). Final classification is made by averaging predicted probabilities. The framework is an ensemble-based meta-learning algorithm (HIDO; KASHIMA; TAKAHASHI, 2009), since it can be applied for general classifiers.

2.4 Results

In this section, we compare evaluation metrics that arise from different scenarios of algorithm setup. Considering the loan application context, some metrics are especially compelling. For instance, Recall reflects the proportion of default operations correctly

Table 1 – Total operations per month

Operation start month	Frequency	Relative Freq.	Default rate
May	2956	6.60 %	5.68 %
June	5260	12.6 %	5.76 %
July	7805	18.7 %	5.52 %
August	9033	21.7 %	7.66 %
September	9646	23.2 %	7.09 %
October	9912	23.8 %	5.99 %

identified by a given model. From a financial institution perspective, an increase in Recall would mean that a granted loan is less likely to default. In contrast, Precision reflects the fraction of actual default operations among the ones predicted as default by the model. There is a trade-off between these metrics where an increase in one may, very often, be accompanied by a decrease in the other. Therefore (among other metrics), we also compute the area under the receiver-operating characteristic curve (AUC), which reflects the probability that a model predicts a higher risk score for a randomly chosen default operation than for a randomly non-default operation selected.

We also show results for Specificity, F1-Score, Geometric mean, and Index Balanced Accuracy. Specificity indicates the proportion of good borrowers with a low probability of default, thus identified as good borrowers. Geometric mean aims to maximize accuracy for each class while keeping those accuracies balanced, reaching high values if both Precision and Recall are high and in equilibrium, giving the same weight for both scenarios (KUBAT; MATWIN et al., 1997). In a more elaborated fashion, F1-score computes the harmonic mean from Precision and Recall. The Index Balanced Accuracy measure combines accuracy across both classes in a weighted manner, favoring the most important class (GARCÍA; MOLLINEDA; SÁNCHEZ, 2009), in this case, non-performing operations.

From an initial analysis, the number of observations increases in the more recent months. The last month contains almost a quarter of the total observations. This result may reflect different strategies of the bank to grant clients access to pre-approved loans.

Default rate sharply increase on August (7.66%) and September (7.09%), but then returns to 5.99% in October, a level close to the first three months. As an 1% of additional default risk can be significant, we identify that the analysis of loans by different granting approval dates can be relevant. The difference on rates and total number of observations per month highlight the caution needed for longitudinal assessments. Strategies of the bank in granting these pre-approved loans within the months can reflective more conservative or aggressive approaches to win over the customer.

Qualitative variables considered sensitive information presented different default by level (Table 2). Gender shows a 1% lower default rate for female customers. Loans granted to customers with lower registered educational levels present higher default rates. For instance, customers up to elementary school have twice the default rate of those

with higher education. The quantitative, sensitive variable Age, showed close values with mean (and standard deviation) of 34.4 (±10) for non-default and 33.8 (±10.4) for default customers.

Table 2 – Default rate by gender and educational level

Sensible Feature	Level	Frequency	Relative Freq.	Default rate
Gender	Male	24402	54.7 %	6.9 %
	Female	20210	45.3 %	5.9 %
Educational Level	Elementary school	1950	4.37 %	9.9 %
	Middle school	30414	68.17 %	7.2 %
	Higher education	11245	25.21 %	4.2 %
	Graduate school	1003	2.25 %	2.9 %

We consider five classification models: Logistic Regression, KNN, Decision Tree, Random Forest, and Gradient Boosting, with two strategies to split observations into training/test data: (i) an utterly random split and (ii) a time-dependency split; and with two approaches for balancing the training dataset: (i) balanced bagging estimators and (ii) SMOTE.

We run each configuration with four different set of input features, by sequentially removing sensitive variables in the following way: (None) without removing any features; (G) removing Gender; (GE) removing Gender and Education; and (GEA) removing Gender, Education, and Age. All approaches combined result in 80 different models. The following metrics are computed: AUC, Recall, Precision, Specificity, F1-score, Geometric Mean Score, and Index Balanced Accuracy. Evaluation metrics from 40 models trained on a random split dataset are presented in Table 3, whereas Table 4 displays metrics from 40 models trained considering a time-window dependency.

When comparing AUC of each classifier within the same setup, ensemble methods present better metrics. With random training/test split, Random Forest has the highest values in 6 out of 8 setups and Gradient Boosting in 1 out of 8 (while being second best in other 6). When considering time-dependency split, Random Forest outperforms 4 out of 8 (while being second in other 2) and Gradient Boosting in 1 out of 8 scenarios (while being second in other 4).

The ROC curve and respective AUC values are displayed considering this longitudinal view (varying classifiers within each strategy configuration): without time-dependency with balanced bagging estimators (Figure 2) and SMOTE oversapling (Figure 3); and considering time-dependency split with balanced bagging estimators (Figure 4) and SMOTE oversampling (Figure 5). It is worth noting the significant drop in performance presented by k-NN when trained on synthetic observations generated by SMOTE. Since these synthetic observations are also generated by k-NN, it may lead to an overfitting cycle.

Figure 6 depicts (a) Recall and (b) Precision for models trained on a shuffled dataset and considering a sequential removal of features. With balanced bagging estimators

Table 3 – Results for models trained considering a random split

Strategy	Modelo	Features Removed	AUC	Recall	Precision	Specificity	F1-Score	Geo	IBA
Bagging Classifier	Logistic Regression	None	0.6302	0.5865	0.0932	0.5998	0.1609	0.5931	0.3514
		Gender	0.6316	0.5907	0.0943	0.6022	0.1627	0.5964	0.3553
		Gender, Education	0.6172	0.5440	0.0891	0.6097	0.1531	0.5759	0.3295
		Gender, Education, Age	0.6169	0.5627	0.0904	0.6027	0.1557	0.5823	0.3378
	KNN	None	0.6131	0.5699	0.0871	0.5809	0.1511	0.5754	0.3307
		Gender	0.6216	0.5948	0.0905	0.5806	0.1571	0.5877	0.3458
		Gender, Education	0.6111	0.6145	0.0886	0.5566	0.1549	0.5848	0.3440
		Gender, Education, Age	0.6124	0.6228	0.0872	0.5425	0.1529	0.5813	0.3406
	Decision Tree	None	0.6459	0.6321	0.0957	0.5809	0.1662	0.6060	0.3691
		Gender	0.6464	0.6425	0.0944	0.5676	0.1646	0.6039	0.3674
		Gender, Education	0.6416	0.6456	0.0934	0.5604	0.1632	0.6015	0.3649
		Gender, Education, Age	0.6348	0.6570	0.0917	0.5437	0.1610	0.5976	0.3612
	Random Forest	None	0.6502	0.6591	0.0959	0.5641	0.1674	0.6097	0.3753
		Gender	0.6488	0.6466	0.0960	0.5729	0.1672	0.6087	0.3732
		Gender, Education	0.6400	0.6580	0.0934	0.5521	0.1636	0.6027	0.3671
		Gender, Education, Age	0.6402	0.6819	0.0918	0.5268	0.1618	0.5993	0.3648
	Gradient Boosting	None	0.6496	0.6435	0.0961	0.5755	0.1673	0.6086	0.3729
		Gender	0.6476	0.6456	0.0968	0.5777	0.1684	0.6107	0.3755
		Gender, Education	0.6412	0.6756	0.0926	0.5358	0.1629	0.6017	0.3671
		Gender, Education, Age	0.6400	0.6974	0.0912	0.5126	0.1613	0.5979	0.3641
SMOTE	Logistic Regression	None	0.6280	0.5979	0.0901	0.5764	0.1566	0.5871	0.3454
		Gender	0.6268	0.6010	0.0914	0.5808	0.1586	0.5908	0.3498
		Gender, Education	0.6138	0.6155	0.0872	0.5479	0.1527	0.5807	0.3395
		Gender, Education, Age	0.6154	0.6124	0.0875	0.5522	0.1532	0.5816	0.3402
	KNN	None	0.5725	0.4860	0.0815	0.6156	0.1395	0.5470	0.2953
		Gender	0.5685	0.4788	0.0812	0.6198	0.1388	0.5447	0.2926
		Gender, Education	0.5638	0.4642	0.0788	0.6195	0.1348	0.5363	0.2831
		Gender, Education, Age	0.5636	0.4705	0.0820	0.6306	0.1397	0.5447	0.2919
	Decision Tree	None	0.6306	0.6187	0.0969	0.5953	0.1675	0.6069	0.3692
		Gender	0.6266	0.6135	0.0968	0.5985	0.1672	0.6060	0.3677
		Gender, Education	0.6140	0.6062	0.0949	0.5946	0.1642	0.6004	0.3609
		Gender, Education, Age	0.6220	0.6653	0.0889	0.5219	0.1569	0.5893	0.3522
	Random Forest	None	0.6414	0.6446	0.0938	0.5634	0.1638	0.6026	0.3661
		Gender	0.6254	0.6663	0.0861	0.5040	0.1525	0.5795	0.3413
		Gender, Education	0.6335	0.6280	0.0931	0.5707	0.1621	0.5987	0.3604
		Gender, Education, Age	0.6375	0.6135	0.0957	0.5933	0.1655	0.6033	0.3647
	Gradient Boosting	None	0.6241	0.6166	0.0968	0.5966	0.1674	0.6065	0.3686
		Gender	0.6446	0.6062	0.0959	0.5989	0.1655	0.6025	0.3633
		Gender, Education	0.6331	0.6922	0.0912	0.5163	0.1612	0.5978	0.3637
		Gender, Education, Age	0.6371	0.6953	0.0904	0.5092	0.1600	0.5950	0.3607

Table 4 – Results for models trained considering a time-window dependency

Strategy	Classifier	Features Removed	AUC	Recall	Precision	Specificity	F1-Score	Geo	IBA
Bagging Classifier	Logistic Regression	None	0.6334	0.6296	0.0835	0.5596	0.1475	0.5936	0.3548
		Gender	0.6347	0.6481	0.0858	0.5599	0.1516	0.6024	0.3661
		Gender, Education	0.6197	0.6279	0.0820	0.5519	0.1451	0.5887	0.3492
		Gender, Education, Age	0.6195	0.6364	0.0826	0.5495	0.1462	0.5913	0.3527
	KNN	None	0.6319	0.6515	0.0819	0.5347	0.1456	0.5902	0.3524
		Gender	0.6345	0.6700	0.0838	0.5327	0.1489	0.5975	0.3618
		Gender, Education	0.6251	0.6684	0.0800	0.5099	0.1429	0.5838	0.3462
		Gender, Education, Age	0.6219	0.6582	0.0796	0.5145	0.1419	0.5819	0.3435
	Decision Tree	None	0.6457	0.6380	0.0856	0.5653	0.1509	0.6005	0.3633
		Gender	0.6448	0.6263	0.0863	0.5771	0.1516	0.6012	0.3632
		Gender, Educaion	0.6422	0.6768	0.0828	0.5222	0.1476	0.5945	0.3589
		Gender, Education, Age	0.6374	0.6414	0.0850	0.5601	0.1502	0.5994	0.3622
	Random Forest	None	0.6493	0.6633	0.0838	0.5379	0.1488	0.5973	0.3613
		Gender	0.6489	0.6700	0.0857	0.5444	0.1520	0.6040	0.3694
		Gender, Education	0.6398	0.7222	0.0819	0.4842	0.1472	0.5914	0.3580
		Gender, Education, Age	0.6339	0.7340	0.0807	0.4671	0.1454	0.5855	0.3520
	Gradient Boosting	None	0.6451	0.6717	0.0842	0.5340	0.1496	0.5989	0.3636
		Gender	0.6454	0.6818	0.0838	0.5249	0.1493	0.5982	0.3635
		Gender, Education	0.6401	0.6987	0.0837	0.5127	0.1495	0.5985	0.3648
		Gender, Education, Age	0.6366	0.6886	0.0846	0.5249	0.1506	0.6012	0.3673
SMOTE	Logistic Regression	None	0.6300	0.6902	0.0844	0.5230	0.1505	0.6008	0.3670
		Gender	0.6301	0.6835	0.0851	0.5313	0.1513	0.6026	0.3687
		Gender, Education	0.6185	0.6330	0.0808	0.5408	0.1433	0.5851	0.3455
		Gender, Education, Age	0.6195	0.6380	0.0814	0.5412	0.1444	0.5876	0.3487
	KNN	None	0.5590	0.4630	0.0696	0.6056	0.1210	0.5295	0.2764
		Gender	0.5653	0.4747	0.0718	0.6087	0.1247	0.5376	0.2851
		Gender, Education	0.5675	0.4764	0.0726	0.6121	0.1260	0.5400	0.2877
		Gender, Education, Age	0.5765	0.4916	0.0742	0.6091	0.1290	0.5472	0.2959
	Decision Tree	None	0.6212	0.6145	0.0863	0.5854	0.1514	0.5998	0.3608
		Gender	0.6182	0.5589	0.0907	0.6427	0.1561	0.5994	0.3562
		Gender, Education	0.6170	0.6549	0.0824	0.5350	0.1464	0.5919	0.3546
		Gender, Education, Age	0.6189	0.7121	0.0794	0.4737	0.1429	0.5808	0.3454
	Random Forest	None	0.6380	0.6633	0.0842	0.5401	0.1494	0.5986	0.3627
		Gender	0.6306	0.6582	0.0829	0.5356	0.1472	0.5938	0.3569
		Gender, Education	0.6214	0.6953	0.0811	0.4981	0.1453	0.5885	0.3531
		Gender, Education, Age	0.6287	0.7020	0.0810	0.4926	0.1453	0.5881	0.3531
	Gradient Boosting	None	0.6069	0.6380	0.0834	0.5532	0.1476	0.5941	0.3560
		Gender	0.6393	0.5993	0.0872	0.6002	0.1523	0.5998	0.3597
		Gender, Education	0.6207	0.7104	0.0806	0.4833	0.1447	0.5859	0.3511
		Gender, Education, Age	0.6206	0.7172	0.0804	0.4772	0.1446	0.5850	0.3505

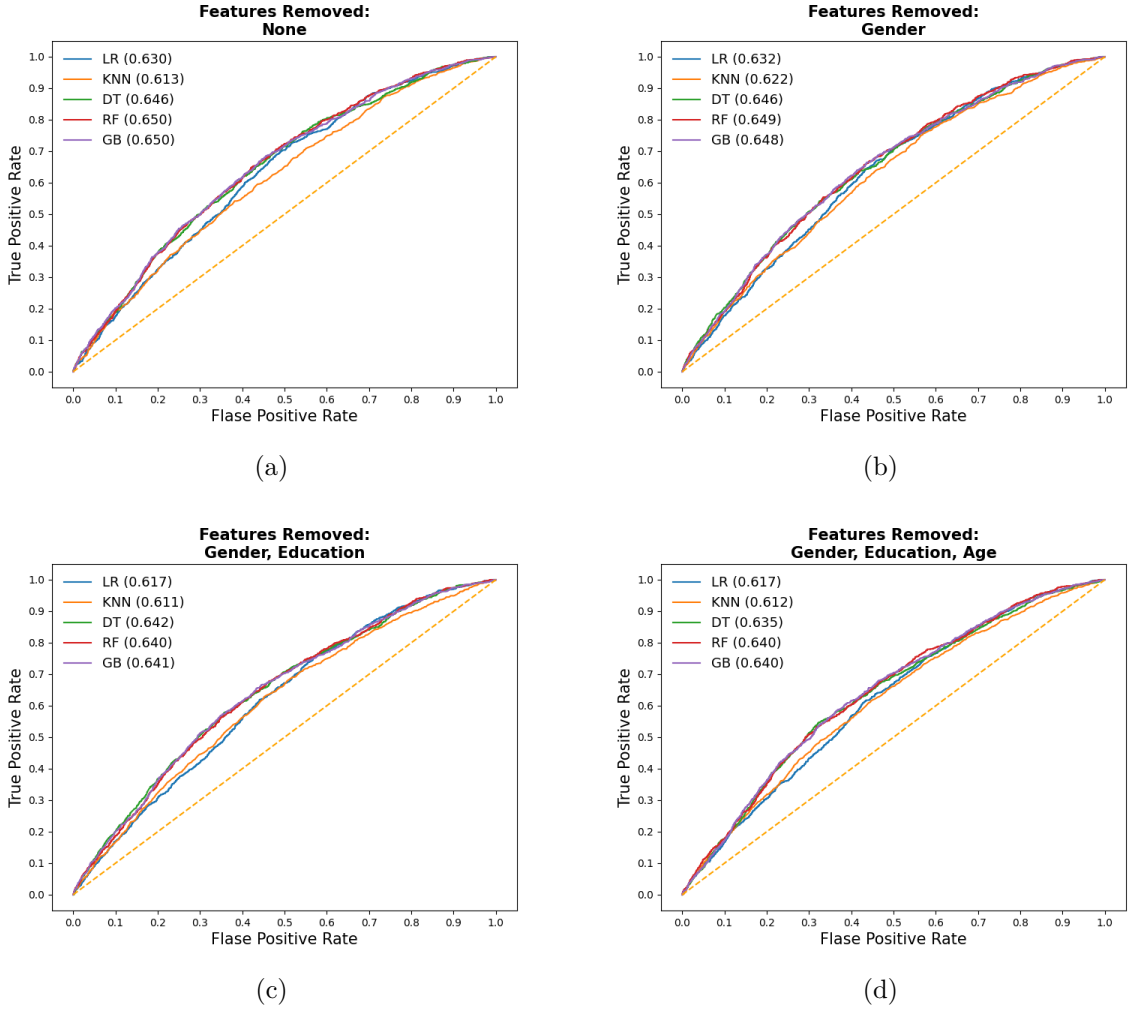


Figure 2 – ROC Curve and AUC for balanced bagging classifiers trained on a random split dataset and removing the following features: (a) None (b) Gender (c) Gender and Education (d) Gender, Education and Age

(left), ensemble methods outperform all models in both Recall and Precision. As sensible features are removed, an increase in Recall can be observed (Figure 6a, left), which leads to a decrease in Precision (Figure 6b, left). Therefore, for this class of models, the removal of sensitive features indicates a more restrictive threshold for granting operations. With SMOTE (Figure 6, right) a higher variability can be observed, with no clear pattern. Gradient boosting shows a significant increase in Recall (and decrease in precision) when education is removed.

With time-split models better Recall are achieved by ensemble models (Figure 7a). Considering Precision (Figure 7b), Decision Trees perform well, but also shows a significant decrease when educational level are removed for syntethic observations (Figure 7b, right). Since the data presented a greater disparity in default rates among educational level, oversampling the minority class with synthetic observations can emphasize its importance during training.

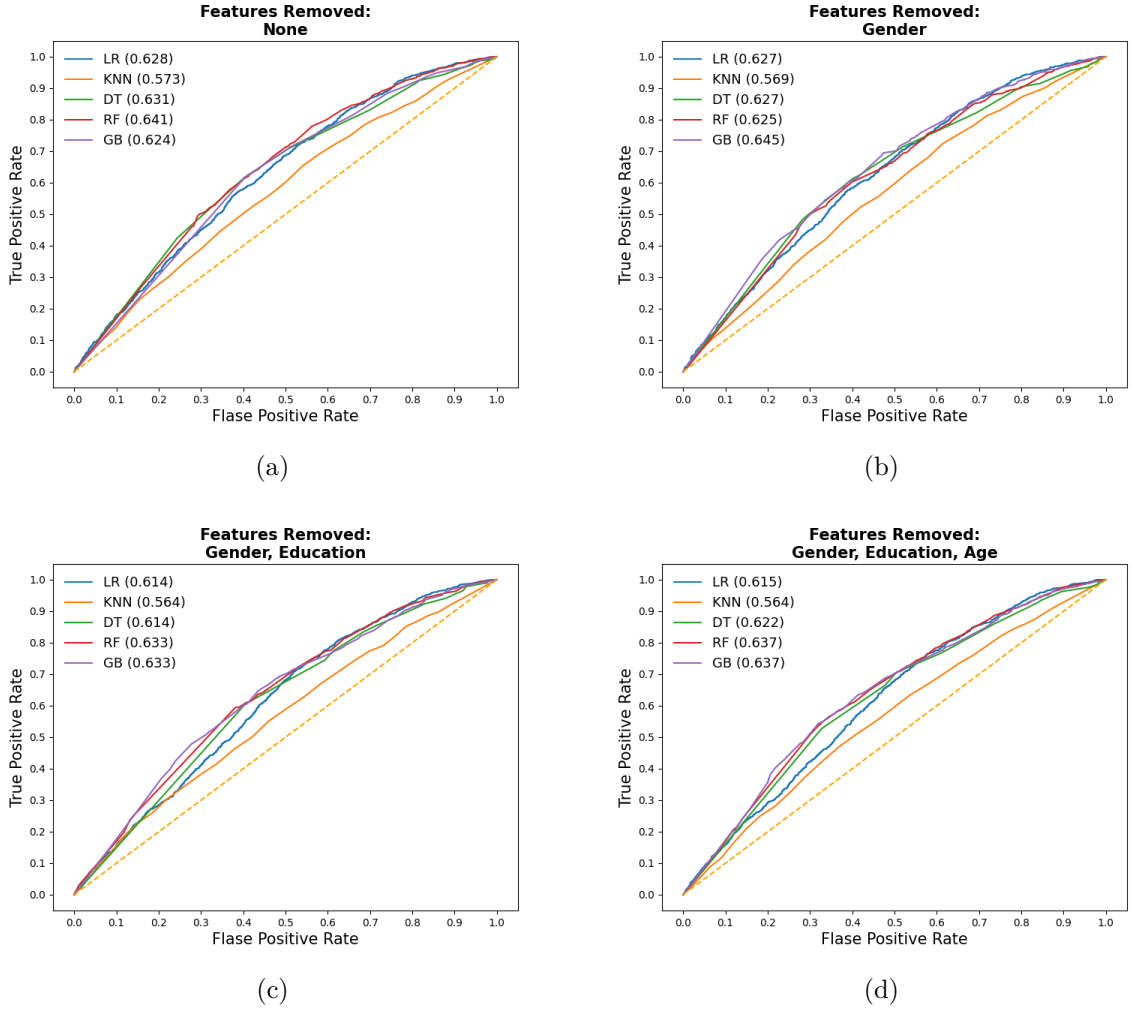


Figure 3 – ROC Curve and AUC for smote classifiers trained on random split dataset and removing the following features: (a) None (b) Gender (c) Gender and Education (d) Gender, Education and Age

In general, better AUC values are achieved by Balanced Bagging Classifier estimators rather than SMOTE. However, Balanced Bagging estimators shows a consistent decrease in AUC as sensible features are withdrawn. In contrast, SMOTE results presents a more variable behavior due to removal of features. In some cases, it even shows an increase in performance with ensemble models. In this sense, a pre-defined strategy could help choosing which pipeline to use, for example knowing before hand which (if any) sensitive feature are strictly required to be removed.

Maintaining all other factors constant, AUC values from different split strategies raise no significant difference. However, specific classifiers tend to perform better with a particular strategy. For instance, Gradient Boosting showed better performance considering a completely random split, whereas Logistic Regression perform better in a time-dependency training dataset. Therefore, this can vary greatly from a range of factors, such as, seasonal behavior, information available for using as input features and objective

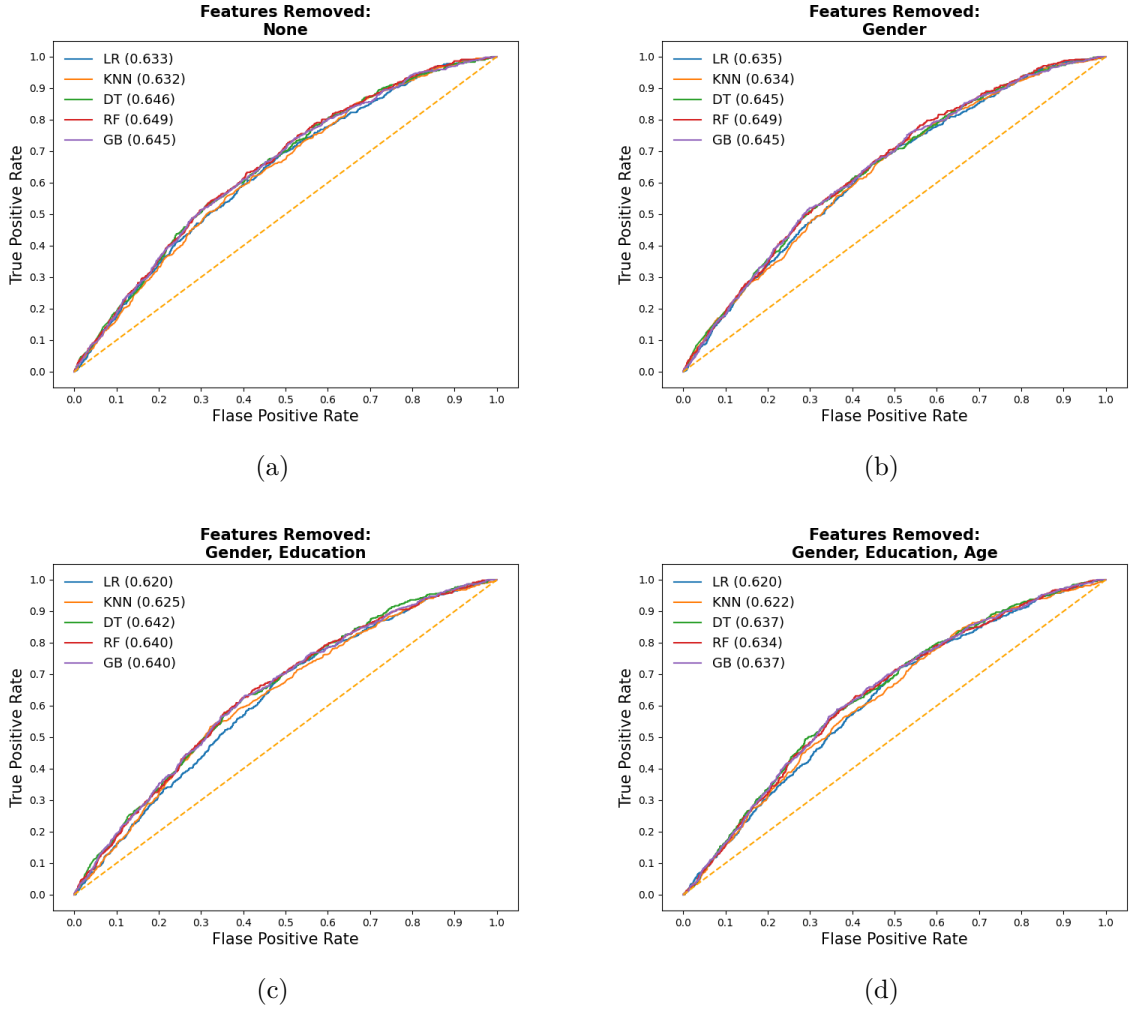


Figure 4 – ROC Curve and AUC for balanced bagging classifiers trained considering a time-window dependency and removing the following features: (a) None (b) Gender (c) Gender and Education (d) Gender, Education and Age

of the study.

All algorithms demonstrated a greater capacity to identify non-performing operations with time-dependency models, i.e., presented highest Recall values. This result could be lead because the training doesn't incorporate the plunge that happened in the default rate in the last month, and it gets more rigorous than models that considered that drop. In a scenario of more significant uncertainty, where banks do not know what to expect in subsequent periods, a model trained on previous operations may be more conservative, reflecting a stricter threshold for lending. Pipeline models must be aligned with the institution's business strategy, which aligns with the current conjuncture.

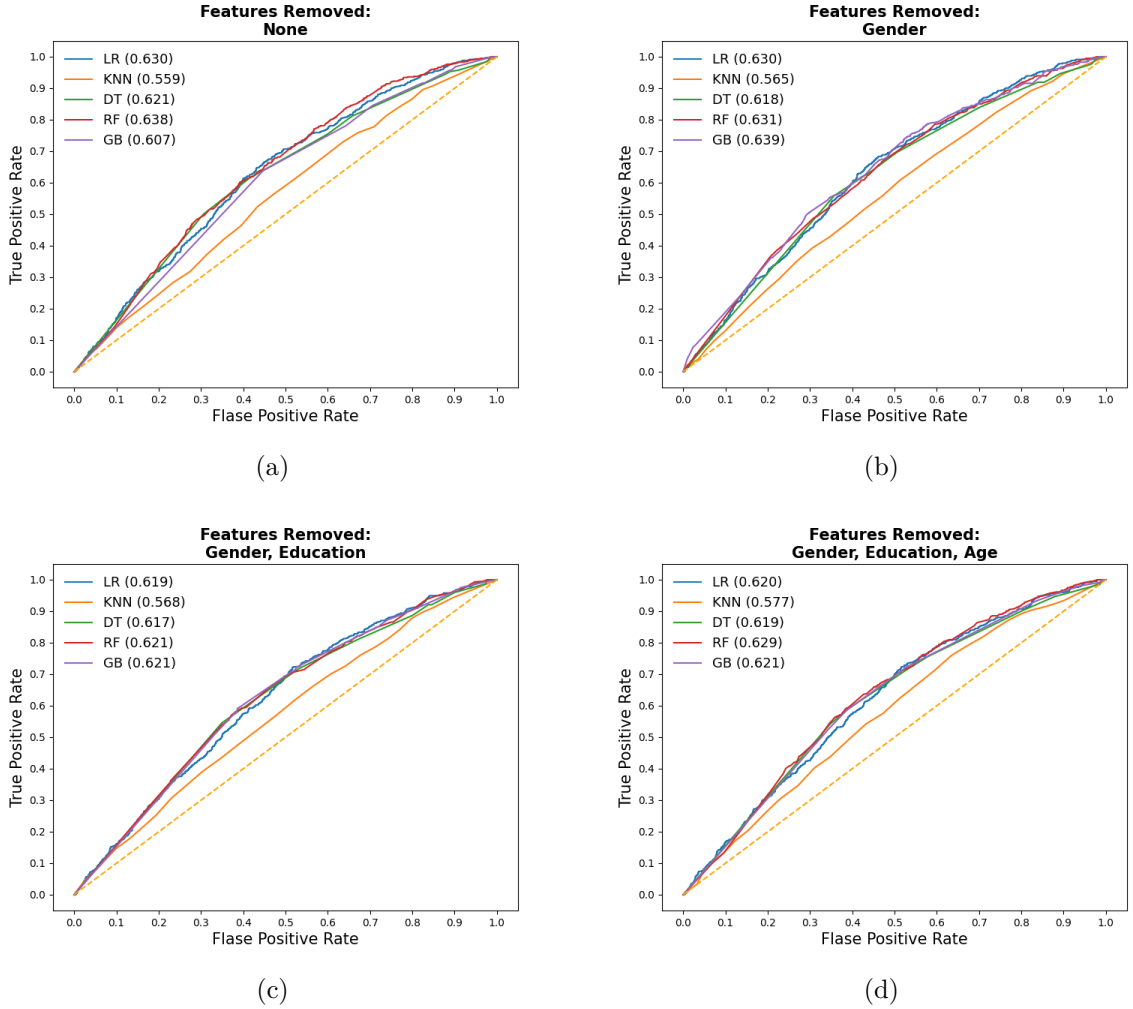
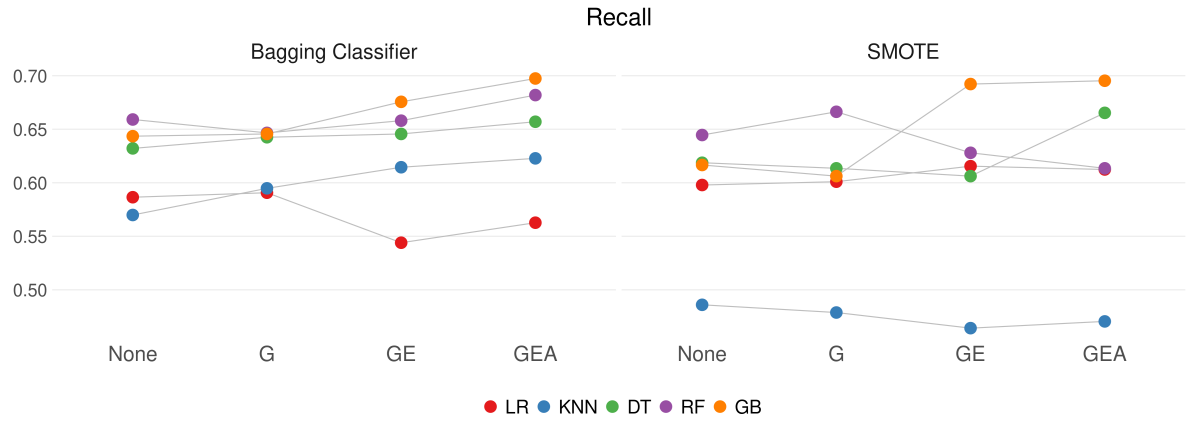


Figure 5 – ROC Curve and AUC for smote classifiers trained considering a time-window dependency and removing the following features: (a) None (b) Gender (c) Gender and Education (d) Gender, Education and Age

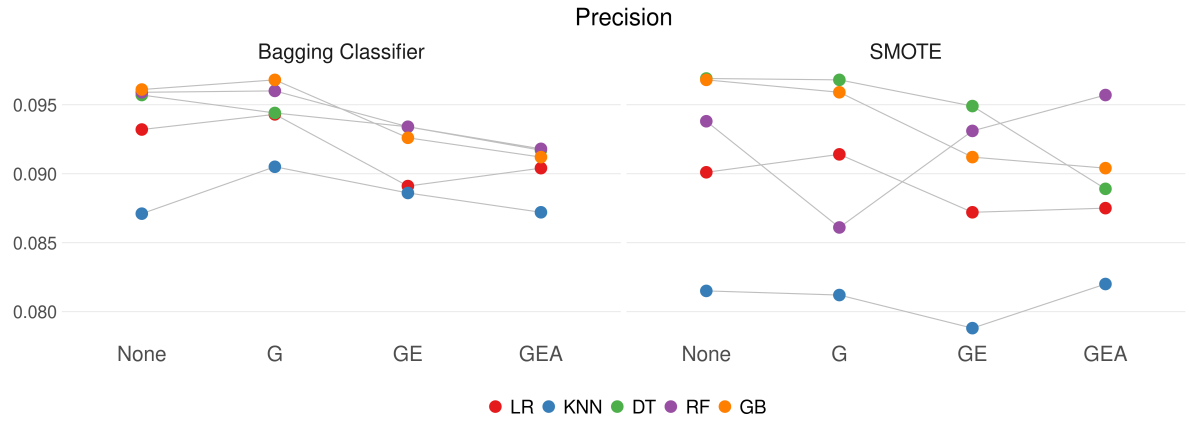
2.5 Conclusion

Decision in credit risk can be highly supported by machine learning automated models. These models should be optimized for known evaluation metrics, such as precision and recall, with the caution of avoiding historical bias rising from sensitive attributes.

Considering information on overdraft loans of a Brazilian financial institution, we illustrate several strategies for the trade-off of performance and bias of models by sequentially removing sensible information. We identified model set ups, such as Balanced Bagging ensemble estimators, that do not consider sensitive features and bring no significant performance decrease. In some cases, a higher proportion of identified defaulted transactions is achieved when these features are removed, satisfying mathematical properties of the modeling algorithm and avoiding historical bias as an input. Among the sensitive variables considered in this study, educational level results in more considerable



(a)



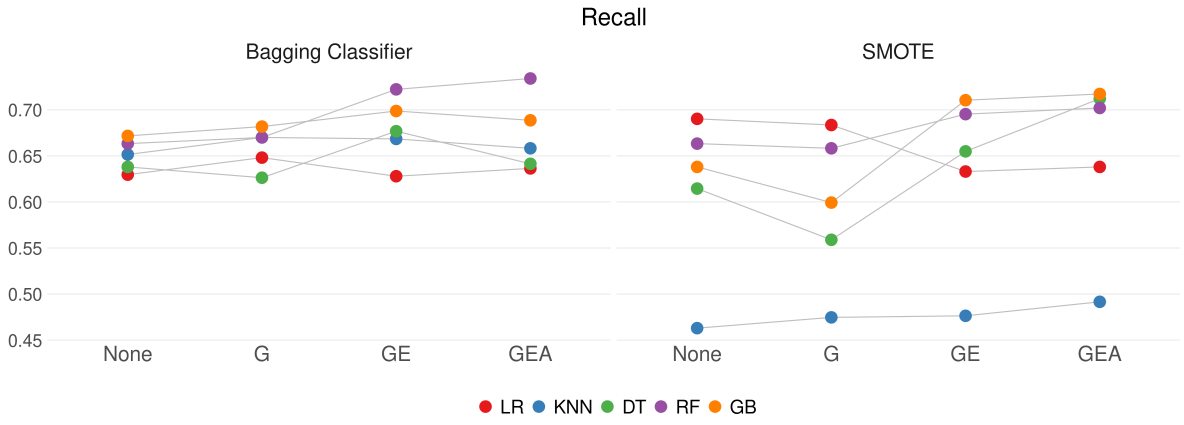
(b)

Figure 6 – Random split: (a) Recall and (b) Precision results.

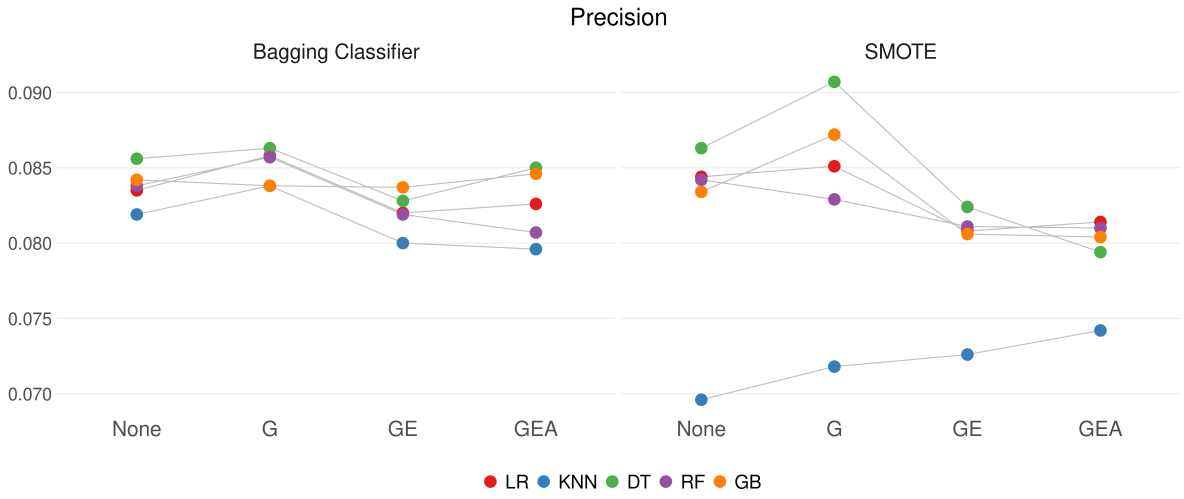
changes in performance metrics, which may due to its larger disparity on default rate among classes.

We analyze different quantitative models and balancing strategies, as well as different configurations of data. We try to assess whether removing discriminatory variables lead to distinct results when applied to different granting months, as bank strategies to win over clients could differ.

Financial regulation and market standards can directly impact financial institutions mechanisms of building predictive risk score models, taking into account the increasing concern on fairness of machine learning algorithms. As discussed, there are many initiatives regarding this issue, such as from those of the European Union's General Data Protection Regulation (GDPR), the Settlements (FSI Insights on policy implementation, 2021) and the Singapore (Thematic Review, 2022). Therefore, we contribute to the literature by providing possible strategies that do not necessarily reduce expected model performance, and at the same time refraining from using sensitive information.



(a)



(b)

Figure 7 – Time-dependency split: (a) Recall and (b) Precision results.

3 Lifetime Probability of Default with Survival Analysis and Ensemble Methods

3.1 Introduction

The recent spread of available data and the increase in computational processing power allow the development of more complex and robust predictive and prescriptive quantitative models in many fields. In the financial industry, models that demand more granular data and enhanced computing capability are becoming increasingly common. These tools help decision making in various fields such as customer segmentation, portfolio selection, derivatives pricing and risk management. However, banks and regulators still have concerns about the applicability of these quantitative and computational models, since although useful, may hinder potential and unknown systemic risks.

One common use of quantitative models in financial institutions relates to credit risk management, which is one of the major risks that financial institutions face. Credit risk is mainly associated with potential losses due to the possibility of a borrower defaulting. Credit scoring is one of the approaches that has been mainly used over the past years (THOMAS; CROOK; EDELMAN, 2017), which involves get an estimate/probability of the outcome. Traditionally, banks make use of statistical models to measure risk and make decisions about credit facilities. For instance, quantitative models are already used on decisions about (i) granting or rejecting a credit loan, (ii) defining limits of exposure to default risk, (iii) establishing interest rates of loans, (iv) calculating provisions necessary to cover expected losses, (v) setting equity capital to comply with regulatory requirements, etc.

More specifically, due to the frequent changes in banking regulations, there are several techniques to model the Probability of Default (PD) of a loan or borrower that were developed over the last years. Despite the overall guidance given by the Basel Committee and by the Central Banks of the countries, there is some flexibility for financial institutions to decide which methodology they will use to estimate PD for managerial and regulatory purposes.

Taking into account the broad academic literature on PD and the current regulation, one important research topic involves the study of credit risk in the context of the new guidelines from the International Financial Reporting Standards 9 (IFRS 9) issued by the International Accounting Standards Board (IASB). The IFRS 9 requires relevant changes in modelling PD, and therefore the topic of our paper is relevant both for academics and practitioners. In particular, new accounting standards call for an estimate

of a lifetime expected credit losses instead of only within a specific time period, i.e., 12 months, or when an impairment occurs.

In this study, we investigate an alternative approach to analyse the probability of default until the maturity of credit card refinancing transactions with classification models, exploring Ensemble Methods and Survival Analysis with tree base learners. This paper compares (i) the baseline Cox's Proportional Hazard (Cox PH) model, (ii) Cox Regression with elastic net regularization, (iii) Survival Trees, (iii) Random Survival Forest (RSF), (iv) Gradient Boosted Survival Trees and (v) Component Wise Gradient Boosted Survival Tree. In addition, the dataset consists of refinancing credit card loans, investigating a different type of borrower; the one that has already defaulted on the original loan.

This paper is structured as follows. In the next section we discuss the context and review some of the relevant related literature. We then describe the material and methods used in the study. Finally, we present our results and conclusions, highlighting the contribution of the paper the academic literature and to the practice of credit risk management, discussing limitations and challenges to study PD from a survival modelling perspective.

3.2 Contextualization

3.2.1 Credit Risk

By releasing the Basel II Accord, the Basel Committee on Banking Supervision (BCBS) presented an approach for financial institutions to measure the capital required to face credit risk losses through internally estimated risk parameters. From these parameters, financial institutions can calculate the value of the Expected Credit Loss (ECL), defined according to Equation 3.1 (BCBS, 2006).

$$\text{ECL} = \text{PD} \cdot \text{LGD} \cdot \text{EAD} \tag{3.1}$$

where the risk parameters are: (i) Probability of Default (PD), probability that a borrower will default on the agreed contract, (ii) Loss Given Default (LGD), percentage of the value of a loan that is lost when the borrower defaults, and (iii) Exposure at Default (EAD), the credit exposure, in monetary values, at the time of default.

Several techniques have been used to estimate credit risk parameters. However, the BCBS emphasizes the need for banks to monitor the effectiveness of their models in calculating the credit parameters PD, LGD, and EAD (BCBS, 2005). Although PD models were, in comparison with LGD and EAD models, more explored by academics and prac-

titioners, recent regulation, e.g., IFRS 9, imposed the need for significant enhancements for addressing the probability of a loan become delinquent.

3.2.2 IFRS9

Since 2018, a new accounting standard published by the International Accounting Standards (IASB), the IFRS 9, is in effect in a large number of countries (International Accounting Standards Board, 2014). This new standard changes the classification and measurement of financial assets and liabilities, impacting many elements of companies, such as income statement, credit risk calculation, data management, etc.

Specifically for credit risk, the main change brought by the new standard involves the Expected Credit Loss (ECL) measurement, which is based on the definition of three risk stages as a criterion for its calculation (International Accounting Standards Board, 2014):

1. For performing credit positions that do not significantly increase risk, the expected 12-month loss must be calculated;
2. For under performing credit positions that are classified, based on criteria defined by the institution, as having significant increases in risk, the expected loss must be calculated for the lifetime of the operation; and
3. For non-performing or defaulted assets, the expected loss is calculated for the entire lifetime of the credit transaction.

The inclusion of risk aggravation stages changes the estimation of the risk parameters of the ECL, including the PD. Rather than presenting an estimation for a limited time horizon, i.e., 12-month period, financial institutions should generate the lifetime probability of default of the credit exposure. A lifetime estimate of the probability of default is a challenging issue for financial institutions, due to the characteristics of their credit exposure, such as high volume and long term maturity.

3.2.3 Probability of Default

Although PD is used in the calculation of the ECL, banks can take advantage of default models for a variety of other purposes, such as (i) definition of credit limits, (ii) decision-making on a borrower's eligibility, (iii) definition of interest rates, (iv) calculation of credit provisions, and (v) calculation of regulatory and economic capital to cope with credit losses. Specific problems related to credit risk and PD modelling represent relevant challenges in the banking industry and can be investigated with the use of machine lear-

ning and data mining methods, which is associated with a process for discovering patterns in data (LEFEBVRE-ULRIKSON et al., 2016).

PD models aim at identifying the probability that a customer or a counterparty in a given transaction will not meet the clauses of a credit agreement. Banks build PD models using historical data on credit operations, personal characteristics and behaviour of their customers. In some models, macroeconomic variables are also used for credit risk modelling and can bring significant impact into the models as described by (DJEUNDJE; CROOK, 2018).

Traditional statistical methods, such as logistic regression, discriminant analysis, and decision trees have been used to credit risk applications (HAND; HENLEY, 1997). However, more recently advances in computing processing power and artificial intelligence algorithms fostered other approaches to evaluate credit risk. For instance, (YEH; LIEN, 2009) compare different classification models on a Taiwan credit card dataset, comparing their predictive accuracy. The authors find that a higher coefficient of determination is produced by artificial neural networks.

In contrast, (NILOY; NAVID, 2018) concludes that Naive Bayes outperform Logistic Regression using credit card data to model probability of default. (LESSMANN et al., 2015) compares 41 classification models with 8 datasets and also found that ANN outperform several other individual classifiers, but recommend RF as benchmark for comparing new classification algorithms and suggests that outperforming LR can no longer be interpreted as a signal of methodological advancement.

To address limitations of traditional credit models, which usually focus on a probability of default within a given period, an alternative mechanism to estimate the lifetime PD is to apply Survival Analysis (SA). More particularly, SA takes into consideration the time until an event of interest occurs and its application in financial context has been growing (ANDREEVA, 2006; ANDREEVA; ANSELL; CROOK, 2007; DIRICK; CLAESKENS; BAESENS, 2017). In the context of credit operations, the event of interest can be the default. Therefore, SA is useful to analyze PD throughout the total term of the credit exposure. In the context of survival analysis being used to investigate credit risk, literature has shifted from the traditional model towards machine learning algorithms.

(NARAIN, 1992b) first applied survival analysis in credit risk management, fitting an accelerated life exponential model to a 24 months loan data. In addition, the author built a scorecard using multiple regression and concluded that supporting the score with estimated survival times could lead to a better credit-granting decision. (CHOPRA; BHILARE, 2018) found that logistic regression outperform Random Survival Forests in out-of-sample evaluation, considering the default for Small and Medium Enterprises (SMEs).

(FANTAZZINI; FIGINI, 2008) compares base decision tree classifiers with ensem-

ble methods. The study reveals that ensemble methods outperformed classical methodologies, highlighting the usefulness of gradient boosting. More recently, (XIA et al., 2021) proposes the SurvXGBoost algorithm and indicates that it outperform other benchmark models in terms of predictability and misclassification cost, considering the probability of default of a loan application. The dataset used was regarded to a major P2P lending platform in the US, between jan/2009 and Dec/2013.

Exemplifying the broad scope of studies using machine learning techniques, (BAI; ZHENG; SHEN, 2021) shows that gradient boosting survival tree outperforms other existing methods by C-index, KS and AUC. The study was conducted using the Lending Club loan dataset retrieved from Kaggle, with operations between 2007 and 2015.

Although RSF methods have been used mainly in medicine (BELLE et al., 2011; PARIZADEH et al., 2017; BALAZY et al., 2019), its applications in credit risk growing in a fast pace, once the results can provide useful information on the field (BREIMAN, 1984).

3.3 Methodology

3.3.1 Survival Analysis

Survival Analysis methods explore the time to an event in a given population and, in comparison with others traditional classification models, such as Discriminant Analysis and Logistic Regression, add the feature of assessing probability over time (BELLOTTI; CROOK, 2008).

In credit risk context, the probability of the borrower not manifesting the event of interest (default) longer than time t is quantified by the survival function:

$$S(t) = P(T \geq t) \tag{3.2}$$

Therefore, Survival Analysis can be used to investigate whether the borrower will default or not, as well as the change in the rate of occurrence of default until a given time t:

$$h(t) = \lim_{\delta_t \to 0} \frac{P(t \leq T < t + \delta_t | T \geq t)}{\delta_t} \tag{3.3}$$

where T is a random variable associated with the survival function, specified in equation 3.2 and $h(t)$ is the harzard function that quantifies the event rate at time t conditional on survival up to t.

Applying SA to credit, (BELLOTTI; CROOK, 2008) performed a study in which they used the Cox PH model to conduct a Survival Analysis studying a sample of credit card transactions provided by a UK bank. The technique follows a semi-parametric model, where the effect of the covariates is used to adjust an unknown probability distribution to a known probability distribution. The results showed that the Survival Analysis has a competitive performance in relation to the Logistic Regression, with the advantage of allowing the estimation of a lifetime probability of default.

Additionally, (DUROVIĆ, 2019) carried out a study on PD modelling, taking into consideration the implementation of the IFRS 9 guidelines. In the study, the authors concluded that the use of Survival Analysis can be relevant to estimate the PD, considering the mandatory estimation of the exposure during the whole lifetime period of the credit facility.

3.3.2 Cox Proportional-Hazards Model

The Cox Proportional-Hazards (Cox PH) model (COX, 1972) is one of the most traditional approaches based on time-to-event techniques. It assumes that the failure rates of two groups are constant, i.e., follow proportional functions.

Considering p covariates and a vector $x = (x_1, ..., x_p)$, the general form of the Cox Model is given by:

$$\lambda(t) = \lambda_0(t)\ g(X\beta)$$

where $g(.)$ is a non-negative function. This model is comprised by two components: a non-parametric $\lambda_0(t)$, which is not specified; and a parametric component which is usually used as:

$$g(X\beta) = \exp(X\beta) = \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$$

where β is a $p \times 1$ vector of parameters for each covariable. The paremeters estimation are given by Breslow's approximation to the log likelihood (KALBFLEISCH; PRENTICE, 1973), using Newton-Raphson method.

For penalized Cox Models, a penalty parameter λ is considered in the partial log-likelihood (VERWEIJ; HOUWELINGEN, 1994) function:

$$l_\lambda(\beta) = l(\beta) - \frac{1}{2}\ \lambda\ P(\beta)$$

where λ is a non-negative weight parameter and $P(\beta)$ is the penalty function. Zou and Hastie (ZOU; HASTIE, 2005) proposed elastic net penalty, defined as:

$$P_{\lambda,\alpha} = \sum_{j=1}^{p} \lambda \left(\alpha|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2 \right)$$

with $\lambda > 0$ anda $0 < \alpha \leq 1$ combining l_1 and l_2 norms. The use of such penalties leads to well-know regression models (such as Lasso, Ridge and Elastic-Net) applied to survival analysis context.

3.3.3 Tree-based Models

Tree-based models involve segmenting the covariates space into a number of simpler regions aiming at making a prediction for a given observation. These models provide an alternative to linear and additive models, regression problems, and linear logistic/additive logistic models for classification problems. In this context, tree-based models are suitable for classification and regression problems in which there is a set of explanatory variables and a single-response variable.

Decision Trees (DT) have many advantages when comparing to others traditional classification and regression models, such as (i) the easiness of the explanation of the relationship between independent and dependent variables, (ii) the logical process for creating the nodes that are more similar to the human decision-making process, (iii) the intuitive classification rules conveyed in figures that depict the decision trees, etc.

An important disadvantage of the Decision Tree models is that its accuracy tends to be lower than of other regression and classification models. Despite this disadvantage, there are different methods to optimise the predictive power of the tree-based models by aggregating many decision trees, through bagging, random forests, and boosting. The intuition of Decision Tree can be applied in a Survival Analysis context, allowing the investigation of PD models through the lifetime of the credit facility, as required by the regulation.

In this study, we investigate Survival Trees and Random Survival Trees, which are data mining techniques based on machine learning that are extensions of Decision Trees.

3.3.4 Survival Tree

Survival Tree (ST) is a tree-based method in which a splitting rule is used for grouping individuals or observations from their covariates. Each group is selected based on its survival behaviour (BOU-HAMAD et al., 2009).

ST has been applied in various areas. For instance, (COHN et al., 2009) used the survival tree technique to identify the risk factors in the diagnose of children with nephroblastoma. In the study, characteristics that influence the time to relapse, malignancy

or death of the patient were identified and a risk hierarchy was created that allows the indication of different treatments given certain characteristics.

In our work, we build the survival tree technique using a splitting criterion based on maximum likelihood, where the chosen cut-off point maximizes the observed log-likelihood function, described in equation 3.4 (BOU-HAMAD et al., 2009):

$$ll(j) = \sum_{t=1}^{K} \Big[n_{t1}(t) ln(\hat{\pi}(t)) + n_{t0}(t) ln(\hat{S}(t)) \Big] \tag{3.4}$$

where $n_{td}(j)$ is the number of individuals in node j with observed time $\tau_i = t$, $d = 0, 1$ stands for right-censored or true time-to-event observations, respectively; $\hat{\pi}(j)$ and $\hat{S}(j)$ are the maximum likelihood estimators of the parameters $\pi(j)$ and $S(j)$, which stands for the probability of events and survival probabilities, respectively.

3.3.5 Random Survival Forest

Random Survival Forest (RSF), proposed by (ISHWARAN et al., 2008a), uses ensemble methods on survival trees to obtain the ensemble cumulative hazard function (CHF). The RSF takes advantage of randomisation in two ways; (i) by randomly drawing B bootstrap samples for each tree, and at each node of a tree, and (ii) by randomly selecting a subset of variables to split. Therefore, the split node is chosen using the selected candidate variable that maximise the survival differences in the child nodes. (BELLINI, 2019) considered the Random Survival Forest an alternative for modeling PD lifetime, as well as other techniques of survival analysis and machine learning.

(ISHWARAN et al., 2008a) give an overview of the framework of RSF, described in the following steps:

1. Draw B bootstrap samples
2. Build a survival tree for each sample. Randomly select p variables at each node. The cut-off point maximizes the survival difference between the child nodes.
3. Grow the tree until the constraints are not violated.
4. Calculate the CHF for each single tree and average them to obtain the ensemble CHF.
5. With the out-of-sample data, compute the prediction error for the ensemble CHF.

3.3.6 Gradient Boosted Survival Trees

Gradient boosting (FRIEDMAN, 2001) stands for an alternative approach for the optimization problem. More specifically, it works in an additive manner, where models

are sequentially updated on former residuals. These models are usually referred to as base learners or weak learners, as they are often simple models. Hence, the overall form of the final model can be described as:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m g(\mathbf{x}; \theta_m)$$

where M is the number of base learners $g(.)$ with parameter θ_m and the overall model is given by a β_m-weighted sum. Following this it can be seen that different base learners, as well as different loss functions, grows into different models. Regarding to this framework, this study applies two different base learners (GBSA and CWGB) resulting in two different models.

Gradient Boosting Survival Analysis (GBSA) implements gradient boosting with regression tree base learner. Therefore, in each step, a regression tree is growing using cox partial likelihood as the loss function. The final estimate is a β_m-weighted sum of trees.

Component-Wise Gradient Boosting (BUEHLMANN, 2006) (CWGB) uses component-wise least squares as base learner. These weak learners will perform a simple regression using as covariable the one that minimizes the Cox proportional harzards partial likelihood. In this way, the overall model will also be a linear model, as it is a result of a linear combinations of m simple linear models.

3.3.7 Evaluation Metrics

The performance evaluation among applied models was conducted according to Concordance Index (C-index) and Integrated Brier Score (IBS). Analysing these two metrics can give a reasonable assessment on model performance The first one indicates how good the model discrimination is and the former can, in addition, bring insights on calibration, as it is evaluated over time periods.

The Concordance Index (HARRELL et al., 1982) performs a rank correlation between estimated risks and observed times, comparing pair observations. A comparable pair for two observations i, j is made if the sample with lower observed time has experienced the event. It is considered a concordant pair if the observation with lower survival time has the higher predicted risk score. For each $y_k < y_w$ where y_k represents a censored observation, the c-index is described as:

$$c\text{-index} = \frac{\sum \sum_{i<j} I(y_i < y_j) I(\hat{S_i}(t) > \hat{S_j}(t)) + I(y_j < y_i) I(\hat{S_j}(t) > \hat{S_i}(t))}{\sum \sum_{i<j} I(y_i < y_j) + I(y_j < y_i)}$$

On the other hand, it is also an interesting that the model shows a good calibration performance. In this sense, we expect that the risk predictions over time follows the

observed events behavior. Integrated Brier Score can provide an assessment for such aspect in addition of also being a measure of discrimination. For a given time period t, IBS can be defined as:

$$\text{IBS} = \frac{1}{n}\sum_{i=1}^{n}\int_{t_1}^{t_{max}} I(y_i \leq t \wedge \delta_i = 1)\frac{(0-\hat{\pi}(t|x_i))^2}{\hat{G}(y_i)} + I(y_i > t)\frac{(1-\hat{\pi}(t|x_i))^2}{\hat{G}(t)dw(t)}$$

where $w(t) = \frac{t}{t_{max}}$. In this way, IBS can be interpreted as how well a model is calibrated for its periods of time.

Models which presents best out-of-sample results are also compared in an out-of-time set of observations. In this case, evaluations are made considering a dynamic time-dependent AUC. The measure is extended to survival context by considering sensitivity (true positive rate) and specificity (true negative rate) as time-dependent measures. Considering $\hat{f}(x_i)$ as the i-th observation's risk score estimative and and ω_i as the inverse probability of censoring weights (IPCW), the dynamic time-dependent AUC, at time t, can be defined as:

$$\widehat{AUC}_{(t)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I(y_i \leq t)\omega_i I(\hat{f}(x_j) \leq \hat{f}(x_i))}{(\sum_{i=1}^{n} I(y_i > t))(\sum_{i=1}^{n} I(y_i \leq t)\omega_i)}$$

Therefore, the measure distinguish observations who fail by time $t_i \leq t$ from those who fail after time t_i ($t_i > t$), giving an intuition on calibration power by analyzing discrimination performance over discrete periods of time.

3.4 Variables and data

We used data from credit card refinancing operations of a US financial institution. Differently from studies that use traditional datasets for analysis of application scoring or loans already granted but still solvent, our research explores a different profile of borrowers. More specifically, we investigate potential default in refinancing borrowers who had already been delinquent in their credit card debt. Therefore, our study adds to the understanding of the credit risk phenomenon in the context of a different borrower profile.

Data spam from January 2012 to December 2018. Each observation in the database is a contract of credit card refinancing of a borrower with financial information. We developed different models by each time-maturities; 36-month and 60-month operations. For training and test, we used operations that started until 2014 for 36 month maturities, and operations that started until 2013 for 60 months. Hence, the out-of-time performance was evaluated on operations beginning on the following year for each model. The split rule applied to both models was the same; the data set was divided in 70% for training and 30% for testing.

The observed proportion of default was 12% in 36-month and 22% in 60-month operations. We validated the models in an out-of-sample and out-of-time dataset. The out-of-time data consists in operations contracted in 2015 (36-month) and 2014 (60-month), from January to December.

The dataset comprises observations with the following 15 attributes:

1. Loan amount: the listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
2. Interest rate: interest rate of the loan.
3. Installment: the monthly payment owed by the borrower.
4. Employment length: employment length in years, ranging from zero to ten, where zero means less than one year and ten means ten or more years.
5. Home ownership: the home ownership status provided by the borrower during registration or obtained from the credit report (rent, own, mortgage or other).
6. Annual income: the self-reported annual income provided by the borrower during registration.
7. Verification status: indicate if income was verified or not.
8. Dti: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
9. Total acc: The total number of credit lines currently in the borrower's credit file.
10. Public record of bankruptcies: number of public record bankruptcies.
11. Tax liens: number of tax liens.
12. Total limit: Total bankcard high credit/credit limit.
13. Earliest credit line: time since borrower's earliest reported credit line was opened.
14. Time to default (in months).
15. Status: binary variable with 1 (default) or 0 (non default).

Unlike the tradition models of PD, in which the relevant dependent variable is a binary variable related to the occurrence of default, the survival methods requires the observation of time to an event of interest, which in this case is the default. This variable

was build as the difference, in months, between the date of the contract of the credit card refinancing and the date of default.

For the survival analysis techniques, we relied on the following assumptions: (i) the event of interest in the survival analysis is the default, (ii) non-default operations are considered right censored, (iii) all trees have been pruned by the total amount of operations on the end nodes.

After cleaning up the original database, we build a data-set containing relevant information that allow the study of the probability of default, throughout time, of credit operations. The analysis were made using the scikit-learn survival framework.

3.4.1 Exploratory Data Analysis

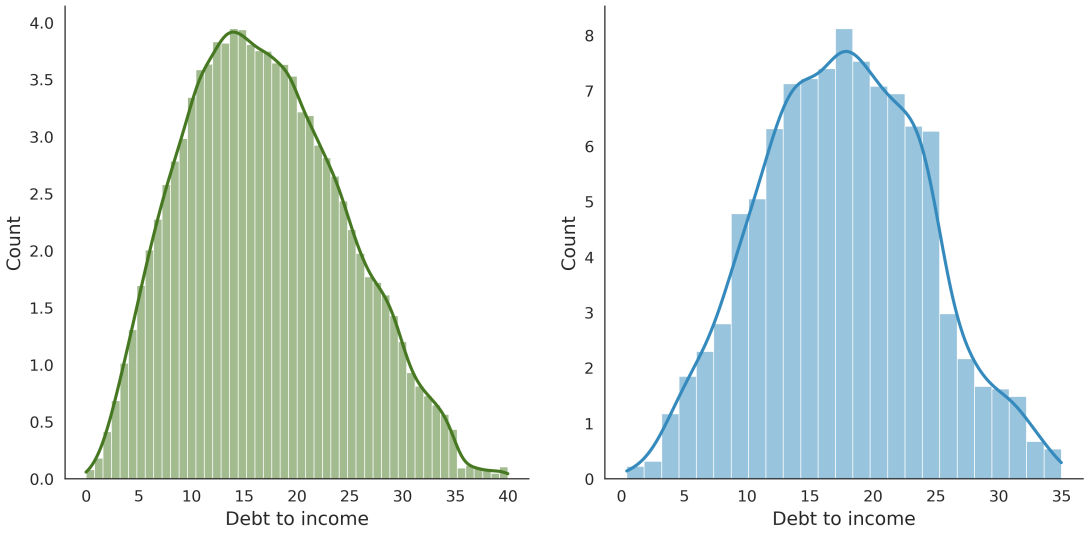
As the dataset consist of a specific type of operation (refinancing), we carried an exploratory data analysis in order to understand some common characteristics among facilities. We then compare features distributions among the two-time maturities described before. Some features shows clear differences that are inherent to operations time maturities, such as installment and loan amount. On the other hand, some features remain very similar, as they reflect common characteristics among borrowers, e.g., total number of accounts and earliest credit line. In addition, public record of bankrupticies and tax liens were dropped because of low variance.

Figure 1 shows that the ratio of debt to income have similar behavior, with most observations falling in between 10 and 25 for both type of loans. The time from the earliest credit line that the borrower reported (Fig. 2) also shows a homogeneous distribution among facilities, but slightly more right-skewed on 36-month. Since the number of 36-month operations is significantly larger, it is expected that some points presents a higher value.

The distribution of total number of credit lines in the borrower's credit file (Figure 3.) also present similar values on the majority of observations, apart from some extreme values (bigger than 65) that also shapes the 36-month distribution as right-skewed. It indicates that most customers have been carrying out credit operations for some time.

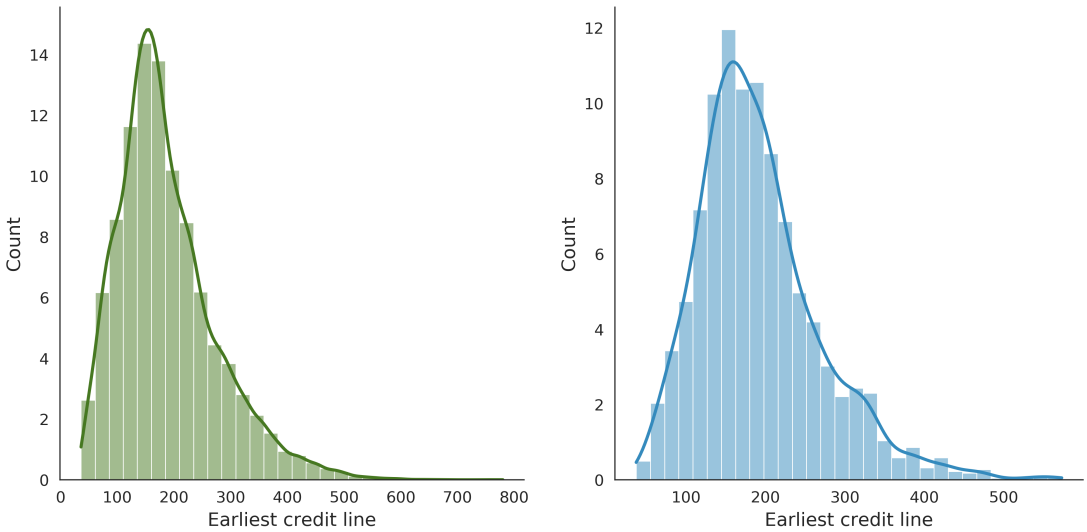
When contrasting the loan amount, 5 years operations seems to yield higher amounts. It's peak is around $20,000, greater than the peak observed in 3 years operations, which presents is slight shifted to the left, around $ 10,000 (Fig 3.). This suggests that, in general, higher value operations are contracted with a longer term for payment of installments, illustrating an expected behavior.

The distribution of installment, which stands for the monthly payment owed by the borrower, seems to present a clear separation, as the maximum value observed in 60-months operation is a value of 0.029 and only 114 (around 0,15%) observations have a



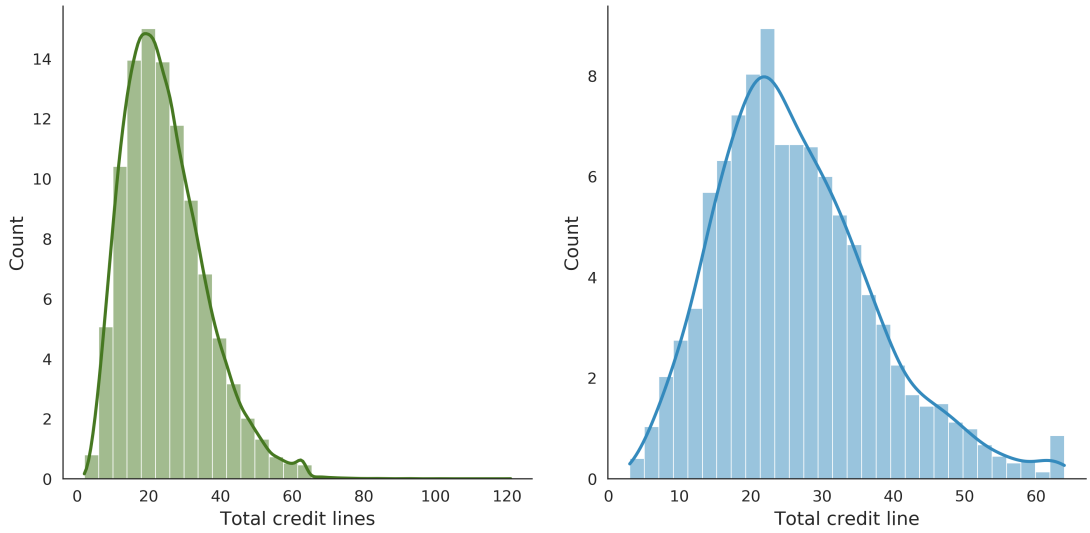
(a) 36 months (b) 60 months

Figure 8 – Time-Debt to income distribution per time maturity.



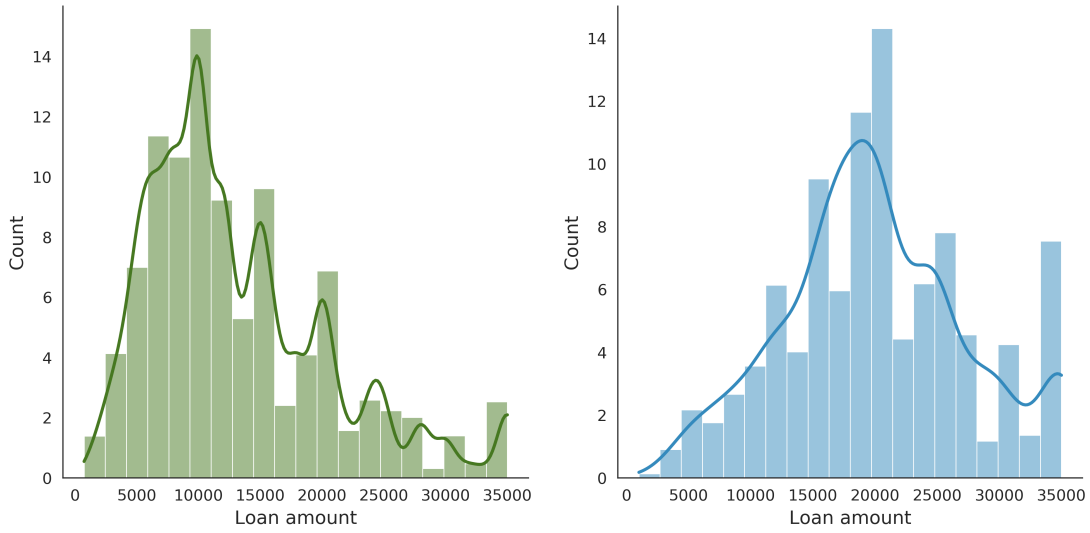
(a) 36 months (b) 60 months

Figure 9 – Earliest credit line distribution per time maturity.



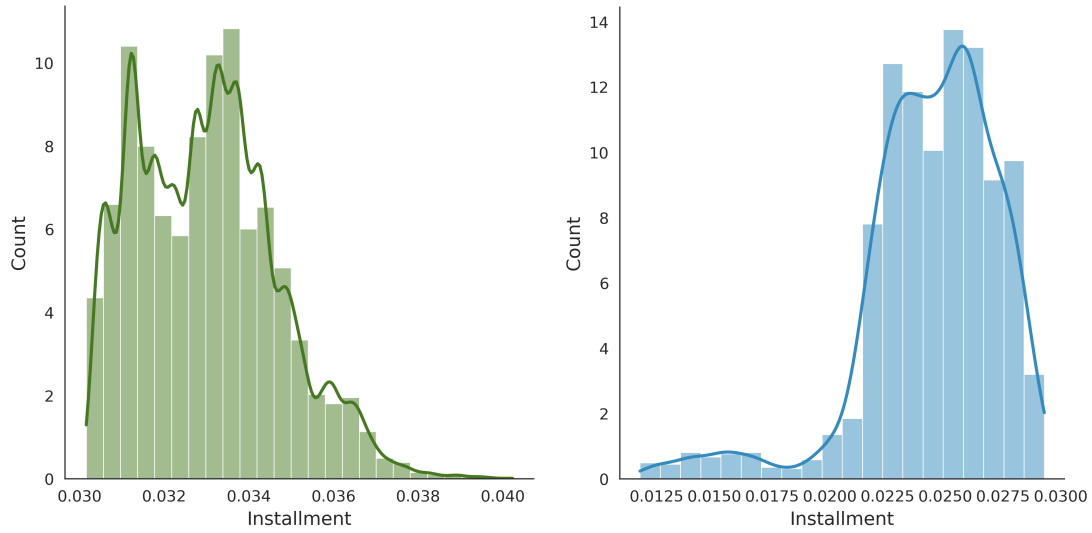
(a) 36 months (b) 60 months

Figure 10 – Total number of accounts distribution per time maturity.



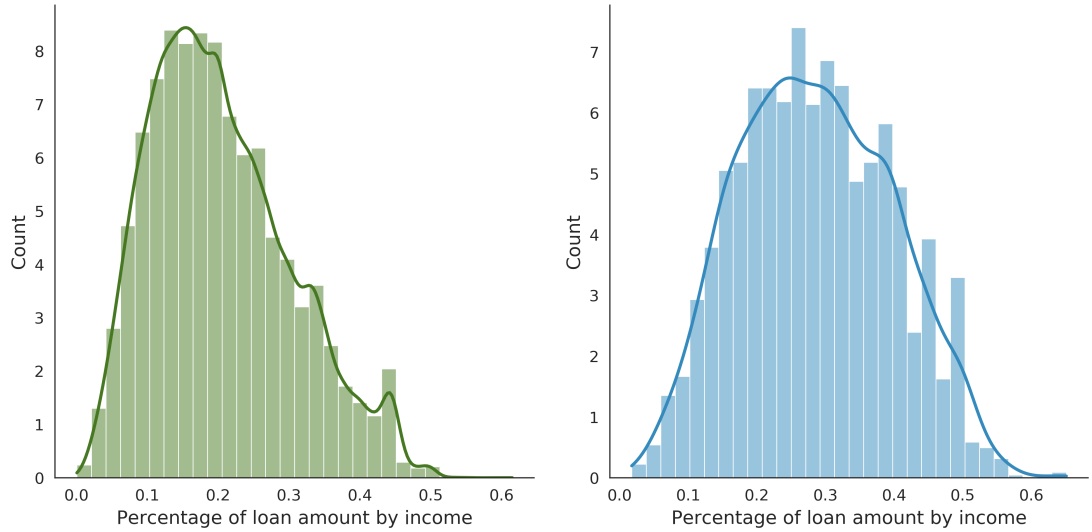
(a) 36 months (b) 60 months

Figure 11 – Loan amount distribution per time maturity.



(a) 36 months (b) 60 months

Figure 12 – Installment distribution per time maturity.



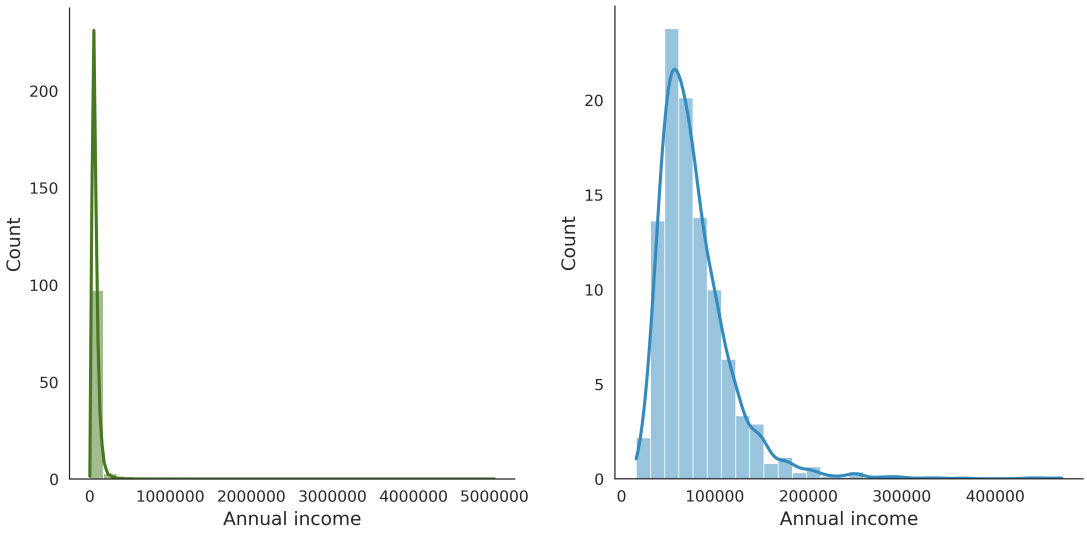
(a) 36 months (b) 60 months

Figure 13 – Percentage of loan by income distribution per time maturity.

value below 0.03 in 36-month time. In line with the loan amount (Figure 4), this behavior is caused by the time horizon defined, since the total amount can be diluted in more installments.

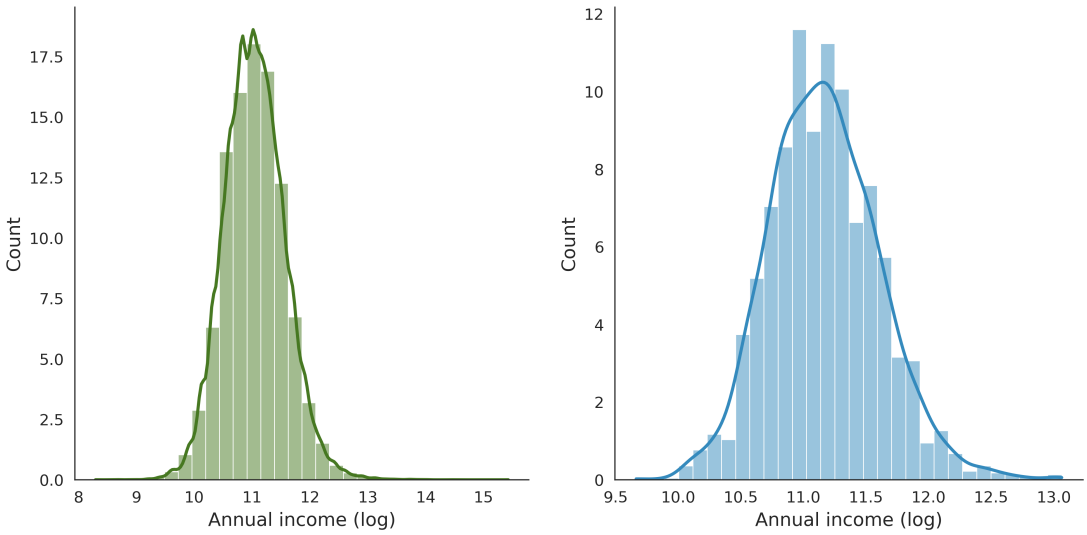
The annual income presents some extreme values, specially in three year operations, resulting in a highly right-skewed distribution. In order to adjust the scale, we choose to work with the log form of annual income in both terms, which yields a more normal-form distribution.

Therefore, these distributions presents a natural behavior considering their time horizon. It's expected that 36-months operations shows more extremely values, since it



(a) 36 months (b) 60 months

Figure 14 – Annual income distribution per time maturity.



(a) 36 months (b) 60 months

Figure 15 – Log annual income distribution per time maturity.

Table 5 – 36-month operations

Variable	Value	Total	Default	
			No	Yes
Verification Status	Not Verified	43%	90 %	10 %
	Source Verified	30%	89 %	11 %
	Verified	27%	89 %	11 %
Home Ownership	Mortgage	46%	91 %	9 %
	Own	9%	88 %	12 %
	Rent	45%	87 %	13 %

Table 6 – 60-month operations

Variable	Value	Total	Default	
			No	Yes
Verification Status	Not Verified	66%	82 %	18 %
	Source Verified	22%	76 %	24 %
	Verified	12%	77 %	23 %
Home Ownership	Mortgage	58%	78 %	22 %
	Own	5%	71 %	29 %
	Rent	37%	76 %	24 %

represents a larger share of contracts. The observed relative percentages of categorical variables, and their cross frequency with loan status are displayed on Table 1 (36 month) and Table 2 (60 month). Longer operations shows a default rate of 22,7 %, slightly higher than larger ones, with a rate of 10,6 %. The rate among possible classes of Verification Status and Home Ownership do not present large divergences. Since all borrowers did not bear the commitment in the original transaction, a higher share of mortgage and rent can be related to a greater difficulty in meeting all monthly commitments when part of the income is already compromised.

3.5 Results

In this section we analyze the evaluation metrics presented by all models. Different survival analysis techniques were used to study the behavior of default in refinancing credit card operations over the period agreed. The operations were separated according to two possible time duration, and then all algorithms were applied considering the features described above, ending up with 16 (8 x 2) model results. The performance comparisons were made using Concordance Index and Integrated Brier Score in both out-of-sample and out-of-time data-sets. We also added Kaplan-Meier results, which only consider the target variable for its calculation and results based on a completely random state, for base comparison purposes. In addition, we compare discrimination power over time from the best models evaluating their efficiency with a cumulative dynamic time-dependent AUC.

For 36 months time-horizon, the C-index of all Cox models showed the best results, with values aruond 0.652 (CoxRidge) and 0.648 (CoxPH, CoxLasso and CoxNet) , closely followed by CWGBSA, with a C-index of 0.647. The calibration power, represented by Integrated Brier Score, over the time of all models seems to be pretty close, as they all present values close to 0.069, apart from Survival Tree model which is ranked with the worst metric values (0.547 C-index and 0.086 IBS).

Regarding models with longer time-horizon operations, CoxPH, CoxLasso and CoxRidge achieves the best C-index scores with values of 0.637, 0.636 and 0.625 respectively. However, a larger difference was observed as the remaining highest values observed was of 0.615 from RSF, 0.613 from CWGBSA and 0.612 from CoxNet. Equivalent to 36-month operations, Survival Tree have also presented worse metric values in longer operations.

Table 7 – Out of sample results

Model	36 months		60 months	
	C-index	IBS	C-index	IBS
CoxPH	0.648	0.069	**0.637**	0.136
CoxRidge	**0.652**	0.069	0.625	0.136
CoxLasso	0.648	0.069	0.636	0.136
CoxNet	0.648	0.068	0.612	0.141
SurvTree	0.547	0.086	0.537	0.174
RSF	0.638	0.069	0.615	0.136
GBSA	0.637	0.069	0.597	0.138
CWGBSA	0.647	0.069	0.613	0.138
Random	0.500	0.253	0.500	0.245
Kaplan-Meier	-	0.071	-	0.141

We also look for model evaluation from a time-dependent perspective. The cumu-

lative dynamic AUC evaluates how well models can separate survival classes in sequential discrete periods of time. Figure 2.8 displays the evaluation over time during the 36 months agreed at the time contracting the operation. However, from the 18 month foward a higher performance is presented by CWGBSA, which increases at 25th month. In contrast, Figure 2.9 shows a dominant performance from CoxPH and CoxLasso.

The Cox Survival Regression models (with and without penalization) and CWGBSA demonstrated the best performances. Although a simple model outperforming more complexity ones may seems counter-intuitive, a possible explanation can arise from an explanatory power concentrated in one (or few) variable(s) as well as a weak non-linear relationship among covariates.

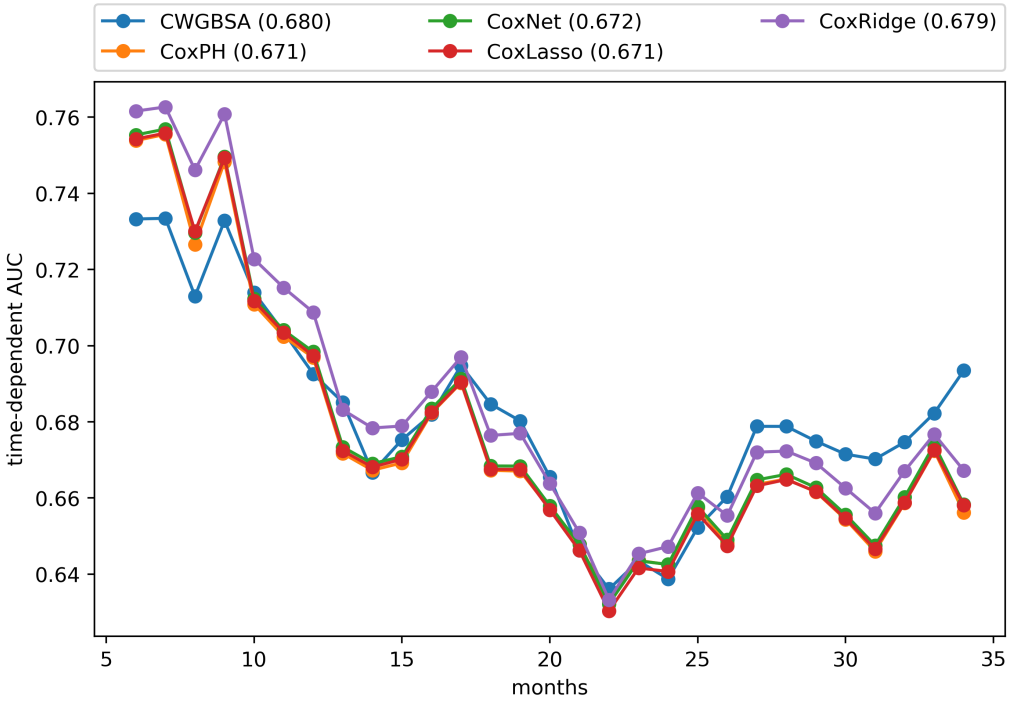


Figure 16 – Cumulative Dynamic AUC for 36-month operations

The best on the out-of-time performance was displayed by CWGBSA, (36 months) and CoxPH/CoxLasso/CoxRidge (60 months), indicating a good generalization power and an acceptable calibration efficiency. In this sense, the power to keep good results over period of times has little variation, which is reflected by IBS.

Out-of-time evaluation suggests that models maintain their performances on near future operations, with an improvement appointing to GWBSA which claims better performance on future operations. Although, structural changes given by market situations or intrinsic factors regarded features interactions can change expected results, and it require extra care when considering models at scale and production.

Dynamic AUC show that GWBSA has better discrimation power in every discrete period of time (Figure 2.10) for shorter operations. For 60 month operations, CoxPH and

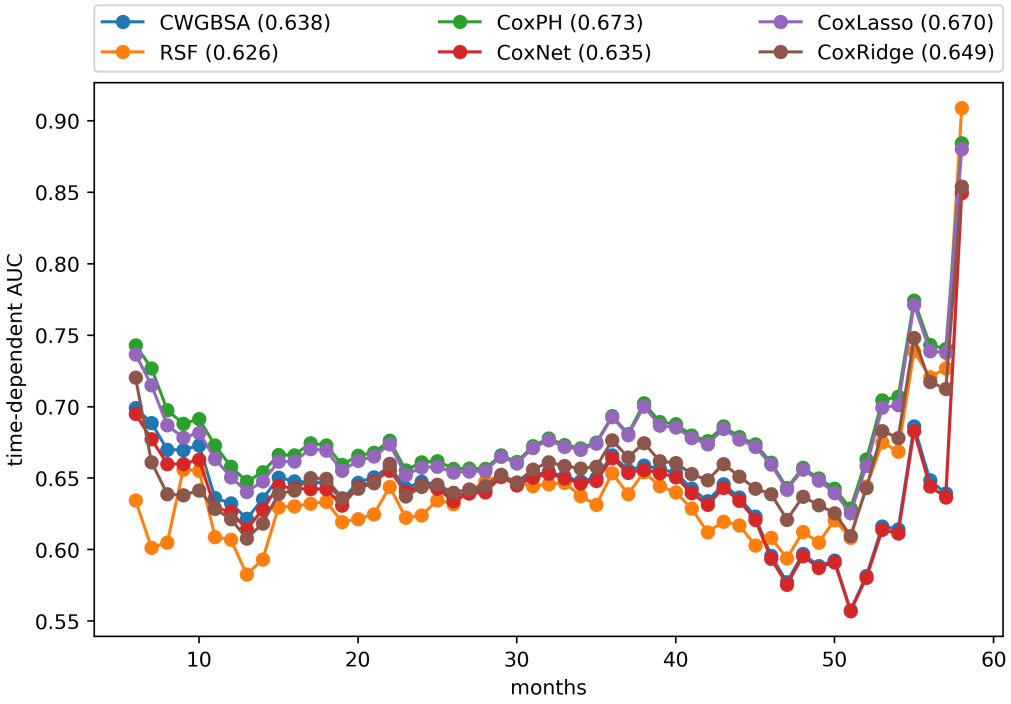


Figure 17 – Cumulative Dynamic AUC for 60-month operations

Table 8 – Out of time results

Model	36 months		60 months	
	C-index	IBS	C-index	IBS
CoxPH	0.678	0.083	**0.634**	0.138
CoxRidge	0.680	0.083	0.634	0.139
CoxLasso	0.680	0.083	0.634	0.138
CoxNet	0.678	0.083	0.621	0.144
SurvTree	0.577	0.092	0.533	0.184
RSF	0.677	0.083	0.613	0.140
GBSA	0.659	0.083	0.601	0.141
CWGBSA	**0.686**	0.083	0.621	0.140
Random	0.500	0.251	0.500	0.233
Kaplan-Meier	-	0.083	-	0.144

CoxLasso remains as the best models, closely followed by CoxRidge (Figure 2.11).

Considering the current dataset, results indicates that a boosting framework, with a component wise as a weak learner, provide better predictability considering operations shorter-terms operations. On the other hand, when a longer future is considered, Cox models (CoxPH, CoxLasso, CoxRidge) led to better predictability power. Once the operations are regarded to refinancing operations from customers who have already defaulted on previous loans, this difference on better models regarding different time operations can arise from intrinsic factors from this specific situation.

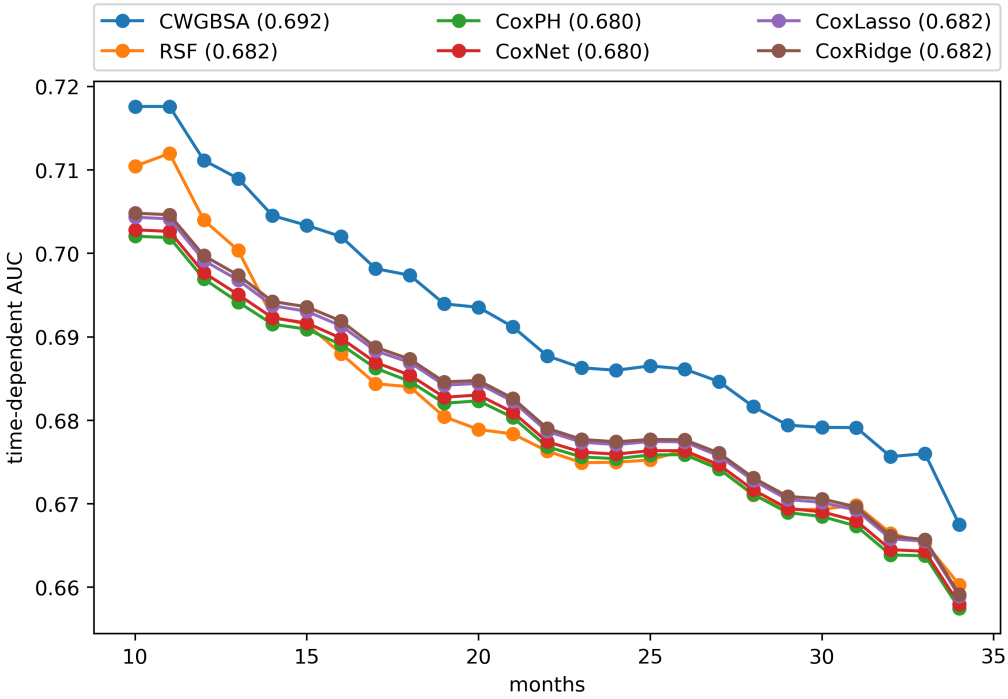


Figure 18 – Cumulative Dynamic AUC for 36-month operations

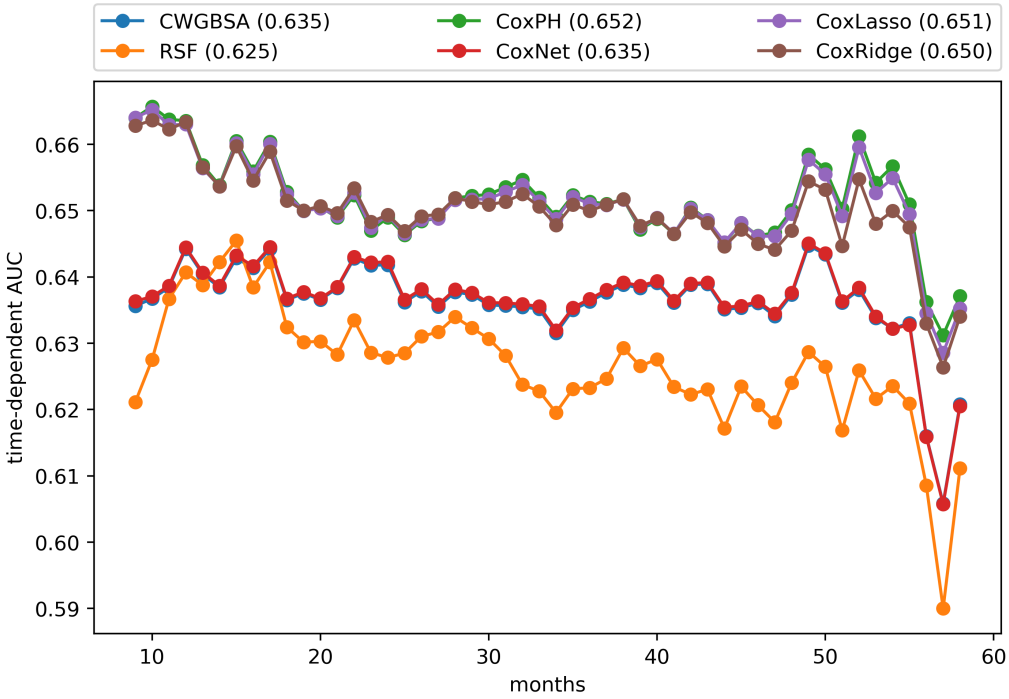


Figure 19 – Cumulative Dynamic AUC for 60-month operations

3.6 Conclusion

In this work we analyzed potential benefits provided by an approach combining Survival Analysis with statistical learning and machine learning methods, in order to adhere to requirements proposed by IFRS 9. This approach can generate point estimates over discrete period of times, providing dynamic probabilities regarded to probability of default during the period agreed. Therefore, such estimates can be of great contribution to address required regulation and to a better understanding of the phenom studied.

The analyzed data set consists of refinancing credit operations with several financial information regarding borrower history and characteristics of the current operation, such as interest rate, loan amount, number installments, etc. In according to survival methods, the time to default was considered as the target variable. Several models with different frameworks were fitted and compared considering suitable metrics.

Overall, four models were consistently on top ranked: Component Wise Gradient Boosting Survival Analysis (CWGBSA)., Cox Proportional-Harzards (CoxPH), Cox with Lasso penalty (CoxLasso), Cox with Ridge penalty (CoxRdige). However, their ranks varied according to horizon-time and different sets of test observations, with CWBSA displaying better results on shorter-time operations and CoxFamily models on longer-time ones.

4 Credit Risk Assessment with Machine Learning and Competing Risk Survival Analysis Models

4.1 Introduction

The study of the occurrence of a specific credit event in a lifetime context has become more important over the past years. The lifetime expected credit loss (Lifetime ECL), introduced by the International Financial Reporting Standard 9 (IFRS 9), implied the development of new credit models. More particularly, the models should measure the present value of potential losses that could arise from the default on an obligation throughout the life of the loan (BIS, 2017).

In this context, Survival Analysis (SA) techniques naturally arise as first-to-go method, where the objective is to study the occurrence of a certain event during a period of time. From a credit risk perspective, the event of interest is default during the lifetime of the loan. Narain (1992a) first introduced the use of survival analysis in credit scoring by estimating the probability of default in a 24-month loan dataset using an accelerated life exponential model. The author states that the use of estimated survival times supporting score ratings can improve credit-granting decision. The study of Narain (1992a) paved the way to many others, with more advanced survival methods applied to credit scoring (DIRICK; CLAESKENS; BAESENS, 2017).

Technological advancements and data availability have provided new tools and information to tackle the issue of estimation of the Lifetime ECL. Some of these developments relate to machine learning (ML) algorithms embedded in survival analysis models. These machine learning survival analysis mechanisms can be used in many applications, in several fields, for instance, to increase user retention (decreasing churn rate), to predict cross-selling opportunities (HARRISON; ANSELL, 2002), to leverage business strategy (KAUFFMAN; WANG, 2001), to estimate the prediction of purchase of online games (YANG et al., 2019). The SA framework can also help financial institutions comply with regulators' guidelines in credit risk management, such as the IFRS 9 (BIS, 2017).

In addition to the advantage of allowing the estimation of a curve representing the risk of default through time, survival analysis can be used to model more than one event. By modelling mutually exclusive events as competing risks, survival analysis methods can be enhanced. From a loan credit risk context, the survival analysis adjusted by competing risks can, for instance, assess both default and prepayment events. Loans therefore can

be analyzed by the event of default, which is the main concern in risk management, and also by the event of prepayment of the loan, situation in which credit risk ceases to exist. Both default and prepayment can be associated with losses.

Default is likely to be more severe due to potential losses in interest and principal value of the loan. However, prepayment also may bring losses, due to unearned interest on the remaining installments (LI et al., 2023). In addition, since prepayment may be unexpected, the cash surplus may be invested at lower interest rates. Hence, the modelling of competing risks can enhance the understanding of potential risk-adjusted performance of credit portfolios, enabling better expected profit strategies for financial institutions.

Competing risks models have already been investigated in studies concerning credit risk, for instance, with personal loan (BANASIK; CROOK; THOMAS, 1999; STEPANOVA; THOMAS, 2002) and mortgage applications (DENG; QUIGLEY; ORDER, 2000; AGARWAL; AMBROSE; LIU, 2006; THACKHAM; MA, 2022; STEINBUKS, 2015). In this study, we use a dataset of refinancing operations, which brings an interesting aspect due to borrower profile. Operations consist of borrowers who had already defaulted. Therefore, the study does not configure a traditional application of scoring model, but instead seek to contribute on understanding the context of debt renegotiation.

The approach adopted in past studies mainly focused on CoxPH (COX, 1972) and its adaptation to a competing risk framework (LUNN; MCNEIL, 1995). However, we follow another approach by embedding a machine learning technique based on boosting into a competing risk survival analysis framework. Therefore, our study has two contributions: i) we focus our analysis on a dataset of renegotiated transactions of defaulted loans, and ii) we introduce a novel ensemble model combining machine learning, more specifically, the boosting algorithm, within the context of competing risks in finance literature. The competing risks relate to credit risk and to prepayment risk.

The study is structured as follows. In the next section, we discuss credit risk and regulatory standards that imply the use of survival analysis techniques. Then, we analyze the adjustments to embed a machine learning technique within the context of competing risks in survival analysis. We apply the derived algorithm in a dataset of renegotiated loans to assess how the proposed model behaves comparing with other techniques. Finally, we conclude the study indicating implications and limitations and suggesting future research.

4.2 Related works

The financial industry is one field where survival analysis modelling approaches are specially useful, as they can provide additional information to support credit scoring decisions. The approach allows the analysis of time-to-event data, when there is an interest in the time to the occurrence of an event. In credit risk, the event of interest is the default,

and the time-to-event would represent when a particular default will likely to happen.

Standard credit scoring methodologies express the probability of default in terms of a binary classification problem (LI et al., 2023). The borrower is classified as "good" or "bad" depending on the estimated probability of default and a given threshold. In survival analysis framework, credit analysis can be translated not just to "if" a borrower will default, but "when" a borrower will default (BANASIK; CROOK; THOMAS, 1999). The possibility of building a predictive model that takes into account the "if" and "when" questions naturally complies with international regulations, such as the IFRS 9.

International Financial Reporting Standard 9 (IFRS 9) was released in 2014 and became effective since 2018, substituting the International Accounting Standard 39 (IAS 39). IFRS 9 incorporated a forward-looking approach for loss allowances calculation. It requires financial institutions to adhere to this forward-looking perspective for expected loss impairment models.

The IFRS 9 indicates mechanisms to calculate provisions, by assessing ECL considering the entire time horizon of the financial instrument, and making adjustments in the Profit and Loss (P&L) account (GORNJAK, 2020). The method involves checking whether there has been a significant increase in risk since initial recognition.

Considering the relevance of identify not only if but also when a default can occur, according to Dirick, Claeskens e Baesens (2017), early studies explored survival analysis techniques in credit risk investigating parametric accelerated failure time (AFT) survival methods or non-parametric baseline approach based on Cox Proportional Hazards (Cox PH) model (e.g. Narain (1992a), Banasik, Crook e Thomas (1999), Stepanova e Thomas (2001), Stepanova e Thomas (2002), Bellotti e Crook (2009), Cao Ricardo (2009), Zhang e Thomas (2012)). Other studies introduced mixture cure mechanisms (e.g. Tong, Mues e Thomas (2012), Dirick, Claeskens e Baesens (2015)) in survival analysis.

In particular, (DIRICK; CLAESKENS; BAESENS, 2017) using credit datasets from European banks, compare results of different configurations of traditional survival techniques based on AFT and CoxPH and mixture cure models. The study identify that models with single event mixture cure and spline adjustment in the hazard function lead to better credit scoring. More recently, machine learning algorithms begin to be incorporated in survival analysis. For instance, Ishwaran et al. (2008b) proposes Random Survival Forest, a random forest method for right-censored data, Binder et al. (2009) develop CoxBoost, an adaptation of boosting algorithm to Cox models, and Chen et al. (2013) build the GBMCI (gradient boosting machine for concordance index) (BAI; ZHENG; SHEN, 2021). Although many studies aimed at applications in medicine and health, credit risk emerges naturally as an area to explore machine learning with survival analysis, due the characteristics of the probability of default within a given period of time.

One example of study on credit is from Bai, Zheng e Shen (2021) that propose a nonparametric ensemble tree model (GBST) coupling survival tree models with a gradient boosting algorithm. Using two different large datasets, the results suggest that the GBST leads to better classification metrics when compared with other machine learning survival models such as Random Survival Forest (ISHWARAN et al., 2008b), CoxBoost (CHEN et al., 2013), Conditional Inference Survival Forest (CIF) model (WRIGHT; DANKOWSKI; ZIEGLER, 2017), and DeepHit based on deep neural networks (LEE et al., 2018).

Finally, in survival analysis, competing risks are relevant, since there may be events that preclude the event of interest from happening (GESKUS, 2015). For instance, in medicine the focus of the study could be related to the risk of an individual getting cancer with a specific treatment, then, death is a competing risk. In credit analysis, prepayment can be a competing risk for default (LI et al., 2023). As Schuster et al. (2020) suggest, biased results can emerge when survival data is analyzed without taking into account competing risks.

However, even though competing events are relevant, is a less well-known element of survival analysis (SCHUSTER et al., 2020). In this context, although some papers explore survival analysis and competing risks in credit (e.g., Banasik, Crook e Thomas (1999), Stepanova e Thomas (2002), Agarwal, Ambrose e Liu (2006), Li et al. (2023)), there are fewer studies that embed machine learning techniques (e.g., Lee et al. (2018), Frydman e Matuszyk (2022)).

Considering refinancing operations, many studies focus on mortgages or home equity line of credit (HELOC). (TRACY; WRIGHT, 2016) applied Cox competing risk models to investigate how mortgage payment reduction from the Home Affordable Refinance Program (HARP) affects the probability that the borrower defaults after having refinanced. The authors suggests that refinancing can have a positive impact in loss mitigation. (CHEN et al., 2018) analyze the risk of re-default on Federal Housing Administration (FHA) modified loans. Authors finds suggests that modified loans are more likely to default compared to identical loans with no modifications.

There has been few studies on the assessment of expected losses on refinancing operations of usual credit lines. In our study, we focus on credit card refinancing exploring a boosting approach embedded in survival analysis techniques with competing risks.

4.3 Machine Learning Survival Analysis for Competing Risks

In this study, we propose a machine learning survival analysis model that takes into account competing risks. We incorporate a boosting mechanism to assess competing risks during training, in a survival analysis setting. Although in this study we apply the method to credit risk and prepayment risk, the proposed model is suitable to any two

competing risks.

In our study, we aim at analysing credit risk loans that have two relevant elements: (i) the borrower default can occur in any moment until maturity and (ii) the borrower can prepay the loan in any moment until maturity. In the occurrence of any of the two events, potential credit risk ceases to exist, as the risk of not complying with the loan has been realized with the default or there is no credit risk anymore, as the loan was fully paid in advance.

4.3.1 Survival Analysis

Changes resulting from the new regulation implied adaptations in the estimation of PD, which is one of the most important risk component in credit risk analysis (VANĚK; HAMPEL, 2017). More specifically, the need to calculate Lifetime ECL requires a method to analyze credit risk not only on a given time, for instance, at maturity or after a year, but also throughout all the period of the loan.

In this context, Survival Analysis methodology can be considered a feasible and appealing approach, since it allows to tackle the default problem from a different perspective. SA models allows to assess whether as well as when a default will occur (BANASIK; CROOK; THOMAS, 1999).

The focus of SA methods is on the time T until an event occurs (e.g., default). Observations that did not experience the specified event are called censored observations. Usually SA data are represented by a pair of random variables (T, C). In the absence of competing risks, the censoring variable C takes the value 1 if the event of interest was observed or 0 if it is a censored observation. When $C = 1$, T refers to the time of occurrence of the event of interest and when $C = 0$, T refers to the time at which the observation was censored.

The function $S(t)$ represents the probability of not having experienced a given event until time t (i.e., the probability to survive until time t) and is given by:

$$S(t) = P(T > t) \tag{4.1}$$

Therefore, the cumulative distribution is defined as the probability of an observation do not survive until time T (COLOSIMO; GIOLO, 2006), that is, $F(t) = 1 - S(t)$, and the probability density function is $f(u) = -\frac{d}{du}S(u)$ (DIRICK; CLAESKENS; BAESENS, 2017).

Additionally, the hazard function, which represents the instantaneous risk, is expressed as:

$$h(t) = \lim_{\delta_t \to 0} \frac{P(t \leq T < t + \delta_t \mid T \geq t)}{\delta_t} = \frac{f(t)}{S(t)} \tag{4.2}$$

which can also be written in terms of the survival function (4.1) and the probability density function (4.2). From these equations, the cumulative hazard function can be defined as:

$$H(t) = \int_0^t h(t)dt = \int_0^t \frac{f(t)}{S(t)}dt = \int_0^t \frac{d\{1 - S(t)\}}{S(t)}dt = -\log\{S(t)\} \tag{4.3}$$

Since $S(t) = e^{-H(t)}$, there is an one-to-one correspondence between the hazard rate $h(t)$ and the cumulative risk distribution $F(t)$ (ANDERSEN; KEIDING, 2012).

4.3.2 Cox Proportional Hazard

The Cox Proportional-Hazards model (COX, 1972) allows to incorporate covariates information into a censored regression model and is one of the most traditional approaches based on time-to-event techniques. It consists of a semi-parametric model for the model is composed by two components: a non-parametrics base hazard λ_0 and a parametric component $g(X\beta)$. The parametric component is usually used as $g(X\beta) = \exp(X\beta)$ (COLOSIMO; GIOLO, 2006). Therefore, the model is given by:

$$\lambda(t) = \lambda_0(t) \exp(X\beta) \tag{4.4}$$

where x_i is a vector of observed data and β is a $p \times 1$ vector of parameters for each covariable. Proportional-hazards comes from the assumption that the ratio of failure rates among two individuals is constant over time. For instance, considering $\lambda_i(t)$ and $\lambda_j(t)$ representing the failure rate of two individuals at time t, it follows that:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp(x_i\beta)}{\lambda_0(t) \exp(x_j\beta)} = \exp(x_i\beta - x_j\beta), \tag{4.5}$$

where the ratio of failure rate is constant and independent of time. For parameter estimation, (COX, 1972; COX, 1975) proposed a partial likelihood without the semi-parametric component. The partial likelihood is a product of all terms associated to different failure times, i.e.:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(X\beta)}{\sum_{j \in R(t_i)} \exp(X\beta)} \right)^{\delta_i}, \tag{4.6}$$

where δ_i is the censoring indicator, taking value $\delta_i = 1$, if an event is observed and $\delta_i = 0$, in case of censoring. The risk set $R(t_i)$ is composed by individuals who have not yet failed until time t_i. Values of β that maximize the partial likelihood function are

obtained by solving the system defined by $U(\beta) = 0$, where $U(\beta)$ is the score vector of first-order derivatives of $l(\beta) = \log(L(\beta))$.

4.3.3 Competing Risks

The approach based on competing risks is adequate when there are two mutually exclusive events, i.e., the occurrence of one event implies the non-occurrence of the other. The (T, C) can be extended to $C = \{0, 1, 2, ..., k\}$ where $k \geq 2$ types of events are possible. When competing risks are present, the Cumulative Incidence Function (CIF) represents the probability of occurrence of a specific type of event before time t. Considering j competing events, the CIF for cause j is defined as (FRYDMAN; MATUSZYK, 2022):

$$F_j(t) = P(T \leq t, C = j) = \int_0^t P(T = t, C = j)dt = \int_0^t f_j(u)dt. \tag{4.7}$$

Thus, the probability that any event takes place before time t, is the sum of all j CIF:

$$F(t) = P(T \leq t) = \sum_{j=1}^{k} P(T \leq t, C = j) = \sum_{j=1}^{k} F_j(t) \tag{4.8}$$

When dealing with competing risks in Cox-PH regression models, there are two main methods for estimating the Cumulative Incidence Function (CIF) (AUSTIN; STEYERBERG; PUTTER, 2021): (i) modeling the cause-specific hazard by considering each event separately to estimate the CIF (KALBFLEISCH; PRENTICE, 2011), and (ii) modeling the Fine-Gray (FINE; GRAY, 1999) subdistribution hazard function, which enables a direct way for modelling the effect of covariates, considering both risks (e.g. $j = 2$) during CIF estimation. Each method have a defined hazard function for a specific event type: the cause-specific hazard function $h_j^{cs}(t)$ (equation 4.9) and the subdistribution hazard function $h_j^{sd}(t)$ (equation 4.9) (AUSTIN; FINE, 2017):

$$h_j^{cs}(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t, C = j | T \geq t)}{\Delta_t} \tag{4.9}$$

$$h_j^{sd}(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t, C = j | T \geq t \cup (T < t \cap C \neq j))}{\Delta_t} \tag{4.10}$$

The cause-specific hazard function is the instantaneous risk of event j in individuals who have not experienced any type of event until time t. The subdistribution hazard function, is the risk of event j considering individuals who have not experienced the specific j event until time t (AUSTIN; FINE, 2017). In this sense, the subdistribution hazard function proposed by Fine e Gray (1999) take into account individuals who have not experienced the primary event of interest, but, have experienced a competing event.

Austin, Lee e Fine (2016) suggest that, on one hand, subdistribution hazard models are better suited for clinical prediction models and risk-scoring systems, where there is a natural interest in estimating the absolute incidence of the primary event. On the other hand, cause-specific hazard models are more suitable when the objective is to assess epidemiological questions of etiology (AUSTIN; LEE; FINE, 2016). Furthermore, the former estimates cause-specific hazard functions for each competing event and derives the CIFs from there, while the latter allows directly estimate the CIF for the primary risk (FRYDMAN; MATUSZYK, 2022).

Fine e Gray (1999) proposes an adaptation for Cox partial likelihood, by changing the risk set R_j and adding weights w_j. The adapted likelihood is given by:

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(x_i\beta)}{\sum_{j \in R_i} w_{ij} \exp(x_j\beta)} \tag{4.11}$$

where R_i consists of observations that did not experience the primary event, even if they have experienced a competing risk event. The risk set is composed of observations who did not experience any event by time t and of those who experienced a competing risk event by time t, defined as (PINTILIE, 2006):

$$R_j(m) = [j; T_j \geq m \quad \text{or} \quad (T_j \leq m \text{ and the subject has experienced a competing risk event})]. \tag{4.12}$$

Additionally, the observations on the risk set are weighted by:

$$w_{ij} = \frac{\hat{G}(t_i)}{\hat{G}(\min(t_i, t_j))} \tag{4.13}$$

where $\hat{G}$ is the Kaplan-Meier estimate of the survivor function of the censoring distribution (PINTILIE, 2006). The weight goes to zero as the distance between the time point t_i and the time recorded for the competing risk event increases. Thus, observations that experienced a competing risk event do not participate fully in the likelihood (PINTILIE, 2006).

4.3.4 Boosting algorithm

Boosting is an ensemble method that sequentially fits models to the data, in which each subsequent model places more emphasis on the observations that were misclassified by the previous models (FREUND; SCHAPIRE, 1997). Friedman (2001) proposes Gradient Boosting Machine (GBM), a boosting framework that generalizes loss functions for regression problems. The GBM algorithm is depicted in 4.2 (RIDGEWAY, 1999):

Algorithm 4.2 *Gradient Boost algorithm (FRIEDMAN, 2001)*

Initialize $\hat{F}(x) = \min_\rho \sum_{i=1}^{n} \Psi(y_i, \rho)$
For m in $1, \cdots, M$ do

1. Compute the negative gradient as the working response

$$z_i = -\frac{\partial}{\partial F(x_i)} \Psi(y_i, \rho)\Big|_{F(x_i)=\hat{F}(x_i)} \tag{4.14}$$

2. Fit a regression model on z_i given covariates x_i
3. Choose a gradient descent step $\rho = \min \Psi(y_i, \hat{F}(x_i) + \rho f(x_i))$
4. Update $F(x)$ estimate as

$$\hat{F}(x) \leftarrow \hat{F}(x) + \rho f(x)$$

Friedman, Hastie e Tibshirani (2000) connects boosting with well-known statistical principles (e.g. additive modeling and maximum likelihood) and demonstrates how the method relates to algorithms used for fitting linear models, such as IRLS (Iteratively Reweighted Least Squares).

Ridgeway (1999) builds a generalization of boosting algorithms for the exponential family and proportional hazards regression models. The proposed generalization is based on Fisher scoring (FINE; GRAY, 1999), a variant of the Newton-Raphson optimizer. The author illustrates adaptations of the algorithm for generalized linear model under a framework proposed by Nelder e Wedderburn (1972).

For proportional hazards regression models, the illustration is made by allowing likelihood based loss functions in Friedman's gradient boosting machine. Thus, we can make use of Cox partial likelihood for fitting censored data. Considering the ideas described above, and that boosting fits nonlinear regression models (RIDGEWAY, 1999), $Enum2$ can be adapted for censored data by searching $F(x)$ to maximize Cox's log-partial likelihood (RIDGEWAY, 1999), by replacing $\Psi(y, F)$ with the $-\log PL(F|t, \delta, x)$, where:

$$\log PL(F|t, \delta, x) = \sum_{i=1}^{n} \delta_i \left[F(x_i) - \log \left(\sum_{j=1}^{n} I(t_j \geq t_i) e^{F(x_j)} \right) \right] \tag{4.15}$$

Therefore, the negative gradient is given by:

$$z_i = \delta_i - \sum_{j=1}^{N} \delta_j I(t_i \geq t_j) \frac{e^{\hat{F}(x_j)}}{\sum_{k=1}^{N} I(t_k \geq t_j) e^{\hat{F}(x_k)}} \tag{4.16}$$

Following the same steps proposed in 4.2, the algorithm for boosting CoxPH for censored data model is illustrated in Algorithm 4.3 (RIDGEWAY, 1999):

Algorithm 4.3 *Boosting algorithm for Cox's PH regression model (RIDGEWAY, 1999)*

Initialize $\hat{F}(x) = \min_\rho \sum_{i=1}^{n} \Psi(y_i, \rho)$
For m in $1, \cdots, M$ do

1. Compute the negative gradient as the working response

$$z_i = \delta_i - \sum_{j=1}^{N} \delta_j I(t_i \geq t_j) \frac{e^{\hat{F}(x_j)}}{\sum_{k=1}^{N} I(t_k \geq t_j) e^{\hat{F}(x_k)}} \tag{4.17}$$

2. Fit a regression model on z_i given covariates x_i
3. Choose a gradient descent step $\rho = \min \Psi(y_i, \hat{F}(x_i) + \rho f(x_i))$
4. Update $F(x)$ estimate as

$$\hat{F}(x) \leftarrow \hat{F}(x) + \rho f(x)$$

Gradient Boosting methods can also operate as a regularization framework (BÜHLMANN; HOTHORN, 2007). The core idea relies on a stepwise optimization of a function $F(.)$ in function space, by minimizing a loss function (BINDER; SCHUMACHER, 2008). This approach has been user for survival context, using Cox negative partial log-likelihood as loss function (BINDER; SCHUMACHER, 2008).

With componentwise least squares as base learner, in each m step, the negative gradient of the loss function is evaluated for the current estimate $F_m(x; \hat{\beta_m})$ (BINDER; SCHUMACHER, 2008). For each predictor variable, a simple linear regression is fitted to the gradient. Then, the coefficient of the predictor variable with the smallest sum of squares is updated. This can lead to many of the estimated coefficients being zero, resulting in sparse fits resembling Lasso-like approaches (BINDER; SCHUMACHER, 2008).

4.3.5 Boosting algorithm with subdistribution hazards competing risks

Considering the boosting approach to fit proportional hazards regression models propposed by Ridgeway (1999), a natural way to incorporate competing risks into a boosting framework is to replace $\Psi(y, F)$ with an adapted log-partial likelihood derived from (4.11).

Following this idea, Binder et al. (2009) proposed a competing risk boosting framework for high-dimensional data for fitting proportional sub-distribution hazards models. The study involves a context in which the number of covariates is greater than the number of observations, and a sparse vector of estimated parameters is desirable. With this, the authors implement componentwise boosting with penalized maximum partial likelihood, and incorporating previous boosting steps as an offset. Another adaptation occurs in a definition of sets of mandatory and optional covariates. Before each boosting

step, parameters referring to the mandatory covariates are updated simultaneously by one maximum partial likelihood Newton–Raphson step. In each boosting step, only one parameter corresponding to the optional covariates is updated.

Applications in credit risk scoring generally involve models with greater degrees of freedom on parameters, with the number of covariates being smaller than the number of observations. In this way, we proceed without the penalty term and restrictions on the covariates. Therefore, we can incorporate information of competing risks by considering the adapted likelihood proposed by Fine e Gray (1999). Hence, we wish to maximize the following log-partial likelihood:

$$\log FG(F|t,\delta,x) = \sum_{i=1}^{n} \delta_i \left[F(x_i) - \log \left(\sum_{j \in R_i} w_{ij} e^{F(x_j)} \right) \right] \tag{4.18}$$

where $j \in R_i$ if $(t_j \geq t_i)$ or $(t_j \leq t_i$ and individual j experienced a competing risk event). Taking the derivative with respect to $F(x_i)$, leads to:

$$\frac{\partial}{\partial F(x_i)} = \sum_{i=1}^{n} \delta_i \left[1 - \frac{w_{ij} e^{F(x_i)}}{\sum_{j \in R_i} w_{ji} e^{F(x_j)}} \right] \tag{4.19}$$

Similar to (4.16), the the negative gradient computed as the working is responses is given by:

$$z_i = \delta_i - \sum_{j=1}^{N} \delta_j I(i \in R_i) \frac{w_{ij} e^{\hat{F}(x_i)}}{\sum_{k \in R_j} w_{kj} e^{\hat{F}(x_k)}} \tag{4.20}$$

The final boosting algorithm that allows to incorporate information on secondary events is given in Algorithm 4.4 by:

Implementation of Algorithm 4.4 was made using the python language and the scikit-survival package (PÖLSTERL, 2020). We set up a development environment in order to adpat the loss functions. Main code changes was done in function `coxph_negative_gradien` from the file `_coxph_loss.pyx`.

4.4 Data and Method

This study analyzes data consisting of credit card refinancing operations of a US financial institution. Therefore, in contrast to traditional credit scoring applications, this research explores a different profile of borrowers. Instead of measuring the likelihood of a borrower defaulting on a new credit operation, we model the probability of default in refinancing borrowers who had already been delinquent in their credit card debt sometime

Algorithm 4.4 *Boosting algorithm for Fine-Gray adapted likelihood*

Initialize $\hat{F}(x) = \min_\rho \sum_{i=1}^{n} \Psi(y_i, \rho)$
For m in $1, \cdots, M$ do

1. Compute the negative gradient as the working response

$$z_i = \delta_i - \sum_{j=1}^{N} \delta_j I(i \in R_i) \frac{w_{ij} e^{\hat{F}(x_i)}}{\sum_{k \in R_j} w_{kj} e^{\hat{F}(x_k)}} \tag{4.21}$$

   where $w_{ij} = \frac{\hat{G}(t_i)}{\hat{G}(\min(t_i, t_j))}$

2. Fit a regression model on z_i given covariates x_i
3. Choose a gradient descent step $\rho = \min \Psi(y_i, \hat{F}(x_i) + \rho f(x_i))$
4. Update $F(x)$ estimate as

$$\hat{F}(x) \leftarrow \hat{F}(x) + \rho f(x)$$

in the past. As a result, our research advances knowledge of the credit risk phenomena in the context of a different borrower profile, i.e., a previous defaulter.

Due to confidentiality and strategic issues, we have access to refinancing operations that span from January 2012 to December 2015. Therefore, the outdated database precludes the disclosure of recent information, such as default rate, but allows the identification of outcomes based on real-world data using machine learning models embedded in survival analysis techniques.

The dataset consists of 135,718 operations with a time maturity of 36 months. We selected operations with issue dates starting up to 2014 (71,001) to train the model and separate those starting in 2015 (65,717) for an out-of-time evaluation. In addition, we selected a 10% sample of operations up to 2014 to reduce computational time, leaving the final train dataset with 7,146 observations.

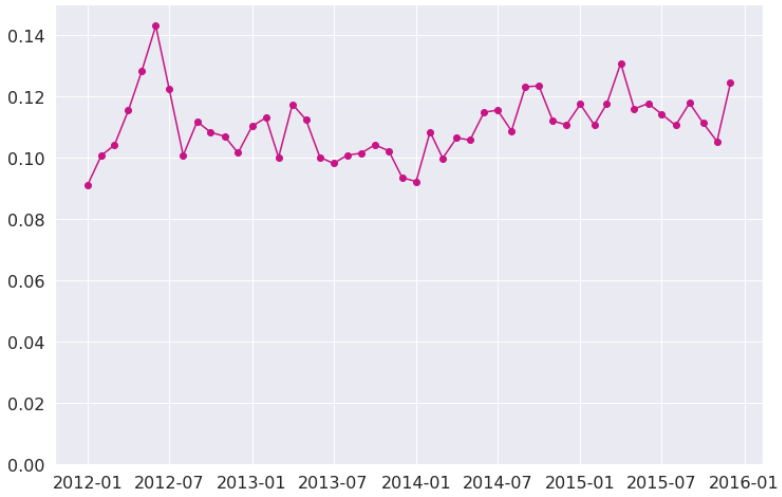
Default rate display values around 11%, with operations starting in 2013 presenting a lower rate of 10.36% (Table 9). Operations starting in 2012 present soaring early payment rates, of over 80% (Figure 20), while other periods keep a behavior with minor variations around 56%. This observed behavior can be a reflection of policies adopted by the financial institution.

For fitting predictive models, the following information, collected at the time of borrowing, was considered as covariates:

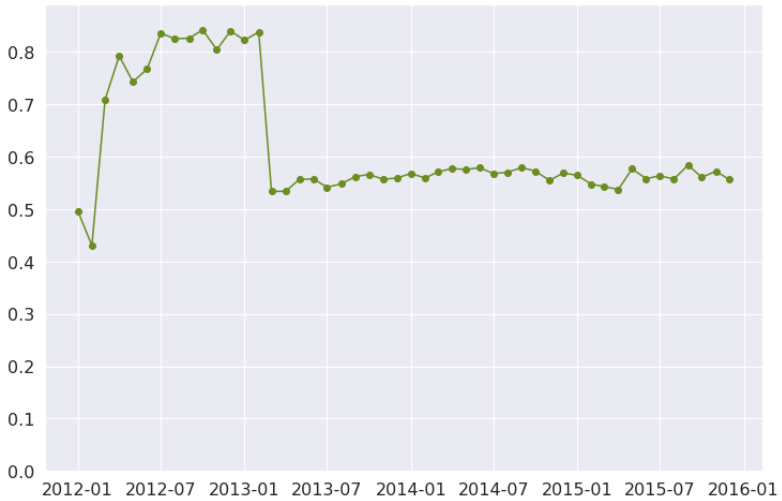
1. Loan amount: the listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be

Table 9 – Default rate and early payment rate by the year of issue

Year		Default		Early Payment	
	Total	Frequency	Percentage	Frequency	Percentage
2015	8708	963	11.60%	6724	56.07%
2014	24664	2555	11.08%	14477	57.00%
2013	37629	4172	10.36%	21449	58.70%
2012	65717	7626	11.06%	36849	77.22%



(a)



(b)

Figure 20 – Rate of (a) Default (b) Early Payment, by issue date (x axis)

reflected in this value.

2. Interest rate: interest rate of the loan.
3. Installment: the monthly payment owed by the borrower.
4. Employment length: employment length in years, ranging from zero to ten, where zero means less than one year and ten means ten or more years.
5. Home ownership: the home ownership status provided by the borrower during registration or obtained from the credit report (rent, own or mortgage).
6. Annual income: the self-reported annual income provided by the borrower during registration.
7. Verification status: indicate if income was verified or not.
8. Dti: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
9. Total acc: The total number of credit lines currently in the borrower's credit file.
10. Earliest credit line: time since borrower's earliest reported credit line was opened.
11. Loan percentage to income: a ratio computed as the loan amount on the annual income, reflecting the share of commitment of income with the loan.
12. Time to default or repayment (in months).
13. Default status: binary variable with 1 (default) or 0 (non default).
14. Repayment status: binary variable with 1 (default) or 0 (non default).

4.5 Results

In this section we compare results provided by fitted models. For the competing risk apporach we consider models based on on cause-specific (CS) and subdistribution hazard (SH). For CS models the secondary risk (early payment event) is assumed to be censored. For SH models, individuals with pre-payment event before time t remain in the risk set with an associated weight.

In credit risk context, ignoring the competing risk event of prepayment results in upwardly-biased estimate of the cumulative probability of default (FRYDMAN; MATUSZYK, 2022). In this way, we first evidence the importance of competing risk modeling

by comparing each survival model as if early repayment was not considered as secondary risk.

Figure 21 shows, for each model, the estimated curve of cumulative probability of default from a hypothetical renegotiation with 12% interest rate, assigned "rent" regarding ownership status and taking median values on all other covariates. For cause-specific models this is the Cumulative Incidence Function, as for the other models, it is represented by 1-predicted survival function. This shows that cause-specific models leads to a lower curve of cumulative probability of default with the same predictive power.

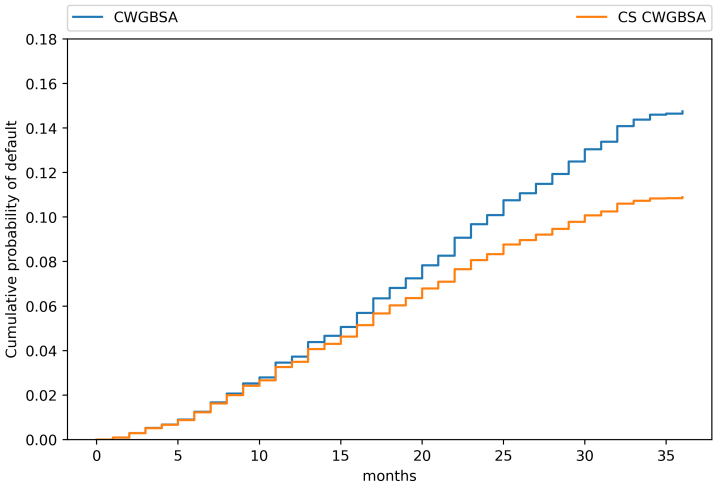
We evaluate predictive performance on both a test dataset and an out-of-date dataset. For performance comparison we compute three metrics commonly used to assess goodness of fit on survival models. The Concordance Index (HARRELL et al., 1982), which measure a rank correlation between estimated risks and observed times. The Integrated Brier Score (IBS), showing accuracy risk predictions over time. The Dynamic AUC providing a measure of calibration over time, by distinguishing observations who fail by time $t_i \leq t$ from those failing after time $t_i > t$.

Table 10 displays out-of-sample and out-of-time results. Models with a ComponentWise Gradient Boosting approach showed the best results, closely followed by Cox-PH. Gradient Boosting Survival Analysis was outperformed and every comparison.

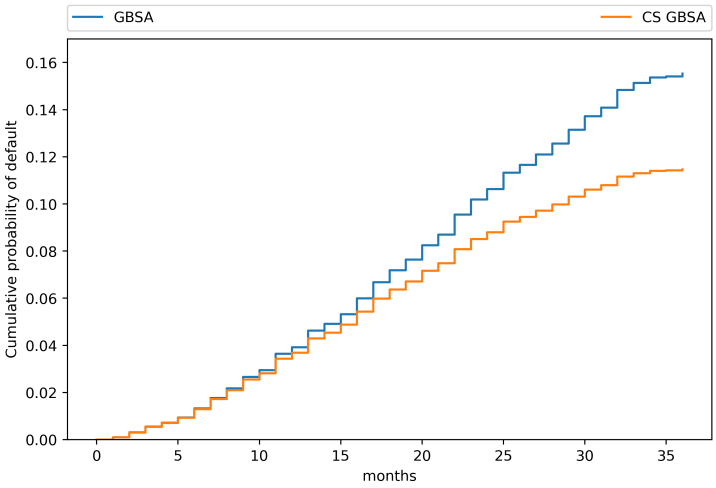
In an out-of-sample test, adaptation for subdistribution hazards showed slightly better results than cause-specific, with a C-index of 0.6462 (HS) over 0.6447 (CS), closely followed by CoxPH (0.6441) and GBSA (0.6313). IBS showed major differences in an out-of-sample test, with the best value of 0.0665 for SH CWGB and over twice this value for cause-specific models, such as, 0.1472 (CS CWGB), 0.1494 (CS GBSA) and 0.1691 (CS CoxPH). However, IBS in out-of-time comparison present similar values of over 0.9 with the best value achieved by CS CWGB (0.0898). This difference relate to the soared rate of early prepayment observed in operations issued in 2012, affecting the secondary event distribution within the time-period. Dynamic AUC for CWGB also show competitive numbers with similar values for SH (0.6607) and CS (0.6602), both approaches presenting better results than CoxPH (0.6574) and GBSA (0.6412).

In operations starting in 2015 a reversed behavior is seen among CWGB approaches, with CS with slilghty higher values than SH. However, both models outperformed Cox-PH and GBSA on all metrics considered. Out-of-sample dynamic AUC shoes greater disparitie with CWGB (0.709 CS and 0.6795 HS) over CoxPH (0.6789) and GBSA (0.6615).

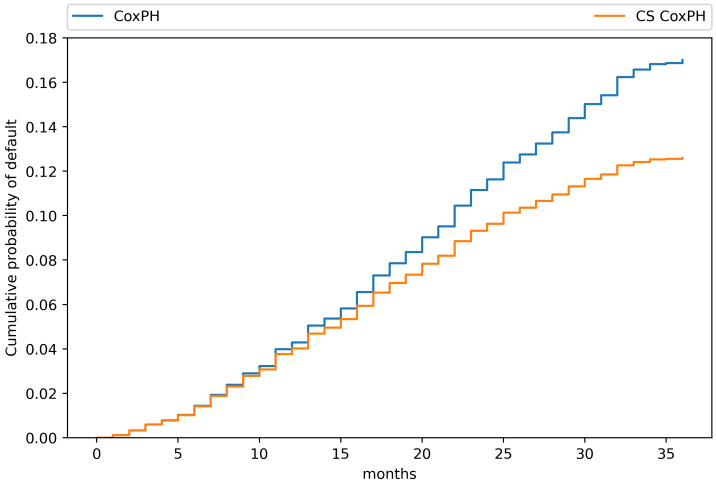
In general, we observe that ComponentWise Gradient Boosting models showed better performance on both scenarios, closely followed cause specific Cox Proportional Hazards, and Gradient Boosting Survival was outperformed in all comparisons. The loss



(a)



(b)



(c)

Figure 21 – Cumulative probability of default comparison when ignoring prepayment event for (a) CWGBSA (b) GBSA and (c) Cox-PH.

Table 10 – Out of sample and out of time results

Model	Out of sample			Out of time		
	C-Index	IBS	AUC	C-Index	IBS	AUC
SH CWGB	0.6462	0.0665	0.6607	0.6873	0.0961	0.6975
CS CWGB	0.6447	0.1472	0.6602	0.6887	0.0898	0.7009
CS CoxPH	0.6441	0.1691	0.6574	0.6711	0.0955	0.6789
CS GBSA	0.6313	0.1494	0.6412	0.6536	0.0937	0.6615

function adaptation to subdistribution hazards on CWGB showed comparative performance.

Aside from predictive power, it is also interesting to analyze default prediction. This can be achieved by comparing the cumulative incidence functions (CIFs), which provides an idea of the probability of failure (PINTILIE, 2006). We analyze predicted curves for loans with 7.5% (Figure 24), 10% (Figure 25) and 12% (Figure 26) interest rate, with home ownership status assigned as "rent" ad 'mortgage" and taking the median value on all other covariates. A higher curve is estimated by SH CWGBSA in all scenarios. While SH CWGBSA presents the same cumulative curve for both status of home ownership (keeping interest rate constant), it appears to be sensitive to interest rate level, with significant increase on the cumulative probability as higher rates are considered. This could be a reflection of the penalized approach lasso-like, leading to a prediction made by few covariates. Cause-specific models provides lower estimated curves and more mixed behavior of different interest rates and home ownership status. CS CWGBSA and CS GBSA presents a similar behaviuor, providing higher curves for "mortgage" when interest rate is 7% , and with a decreasing impact of home ownerhsip as higher rates are considered (with CS GBSA curve higher then CS CWGBSA in all scenarios). For CS CoxPH higher curves are observed for "rent" than for "mortgage" status, specially for higher interest rates. For instance, with 12% interest CS Cox-PH has the lowest curve for "mortgage" and the second highest for "rent".

4.6 Conclusion

The use of survival models for estimating default probabilities presents an attractive alternative for compliance with regulations such as IFRS9. Additionally, by incorporating the competitive risk of early prepayment, we can estimate lower default probability curves while maintaining the same predictive power. From the perspective of financial institutions, these smaller curves can lead to a lower required provision, thus positively impacting results such as shareholders' equity, bonuses, and dividends.

In this paper, we have introduced a component-wise boosting framework that considers competing risks with subdistribution hazards in the credit risk literature. Unlike

Figure 22 – Out-of-sample cumulative Dynamic AUC

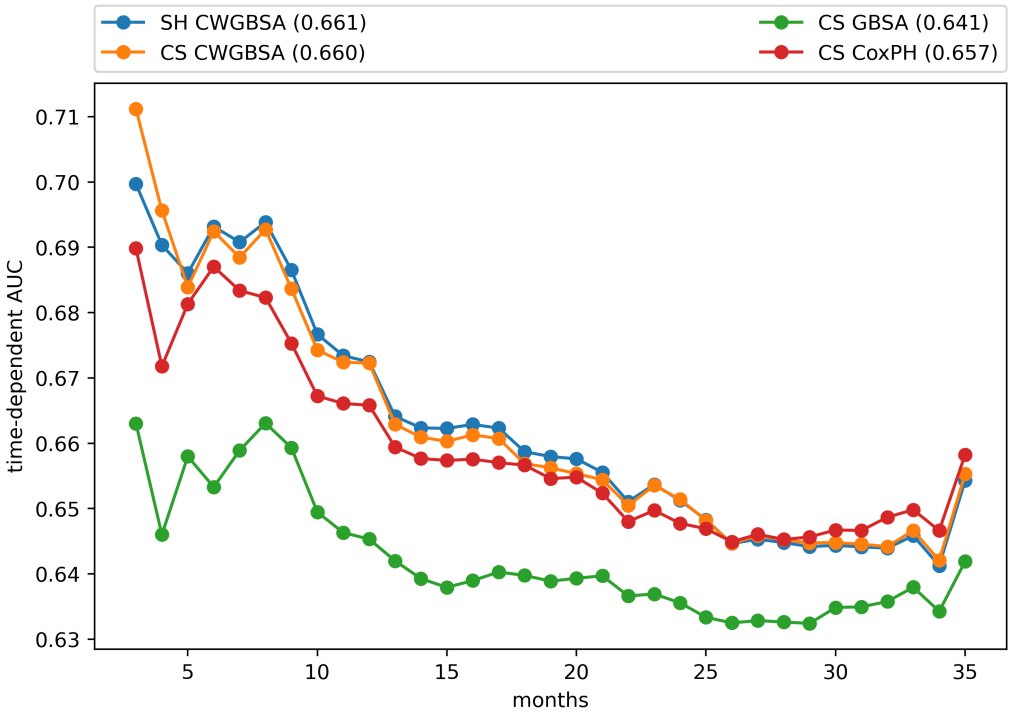


Figure 23 – Out-of-time cumulative Dynamic AUC

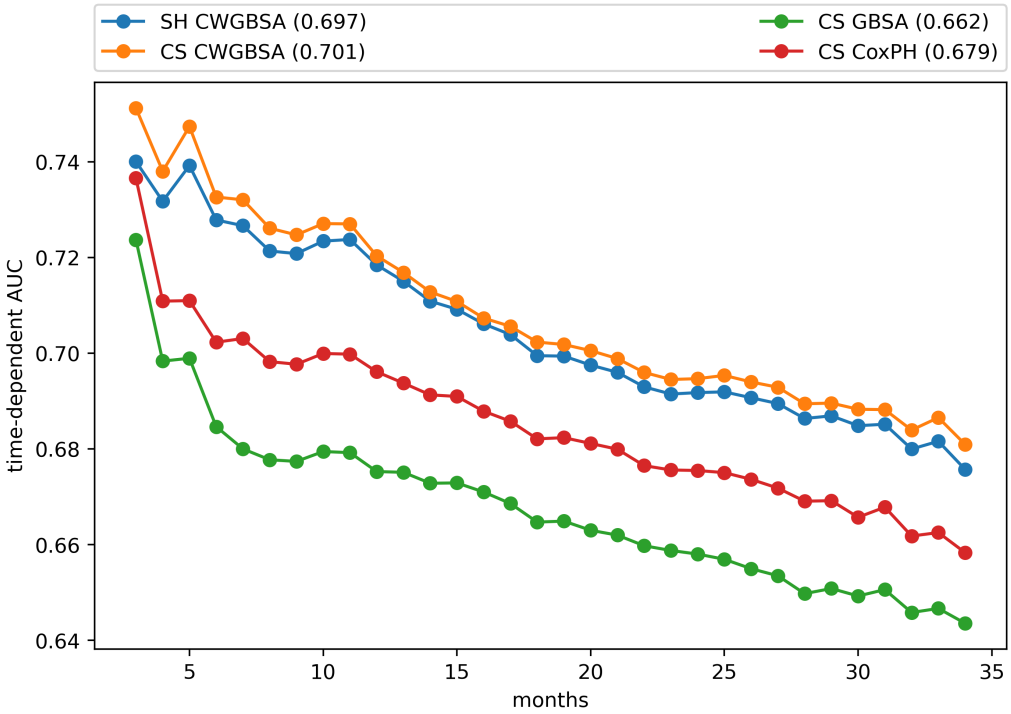
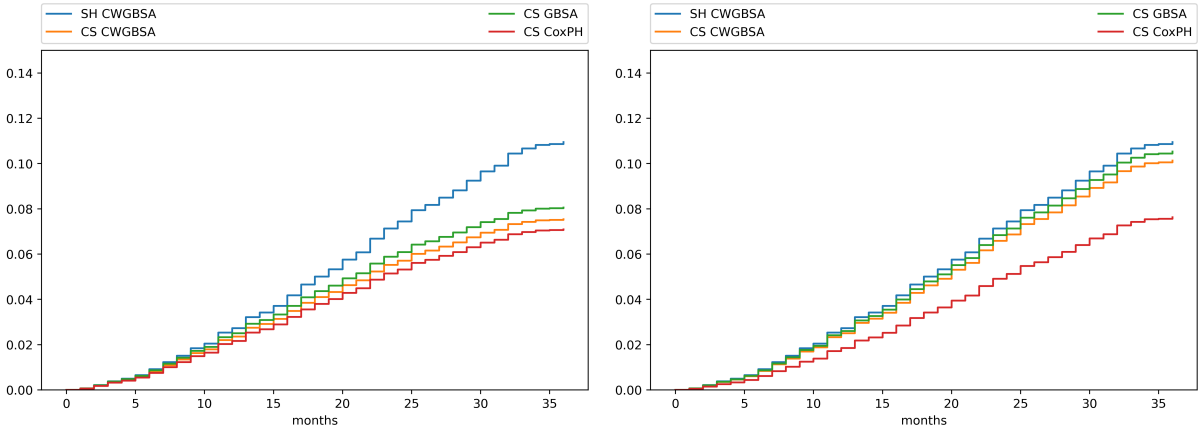
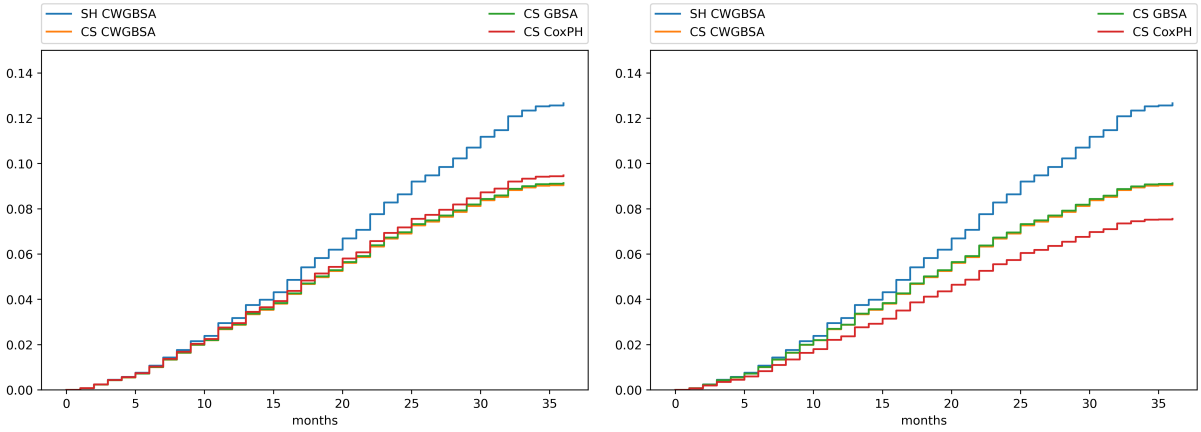


Figure 24 – Cumulative Probability of default for an operation with interest rate of 7% and home ownership assigned as (a) Rent and (b) Mortgage



(a) (b)

Figure 25 – Cumulative Probability of default for an operation with interest rate of 10% with home ownership assigned as (a) Rent and (b) Mortgage

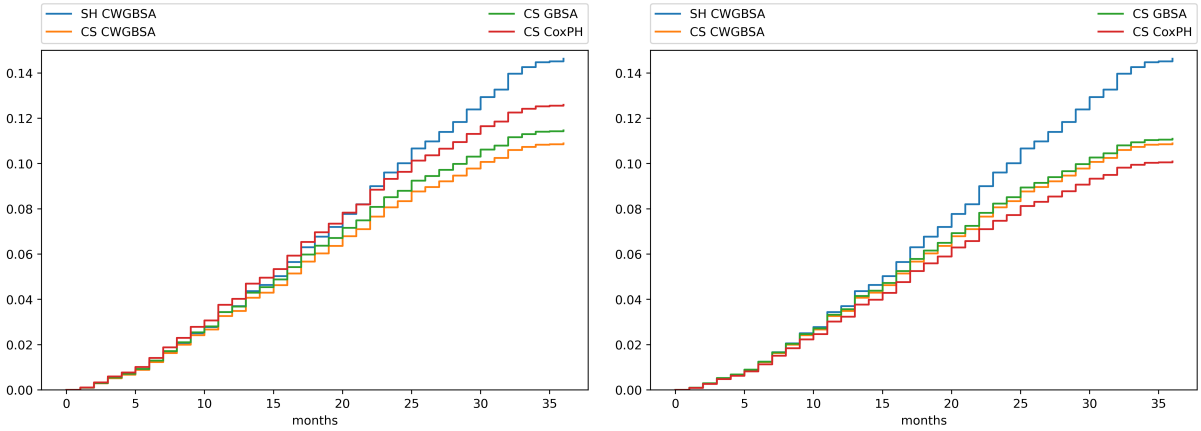


(a) (b)

studies applying a similar framework in other areas (BINDER et al., 2009), we do not consider an offset during the fit stage and use non-penalized loss functions. This choice is motivated by the uncommon occurrence of having more covariates than observations in the context of credit loans. We have demonstrated that adapting the loss function to include competing risks during estimation on the CIF yields comparative results compared to cause-specific models when analyzing a dataset of refinancing operations. For future studies, it would be interesting to test different base learners within the boosting framework.

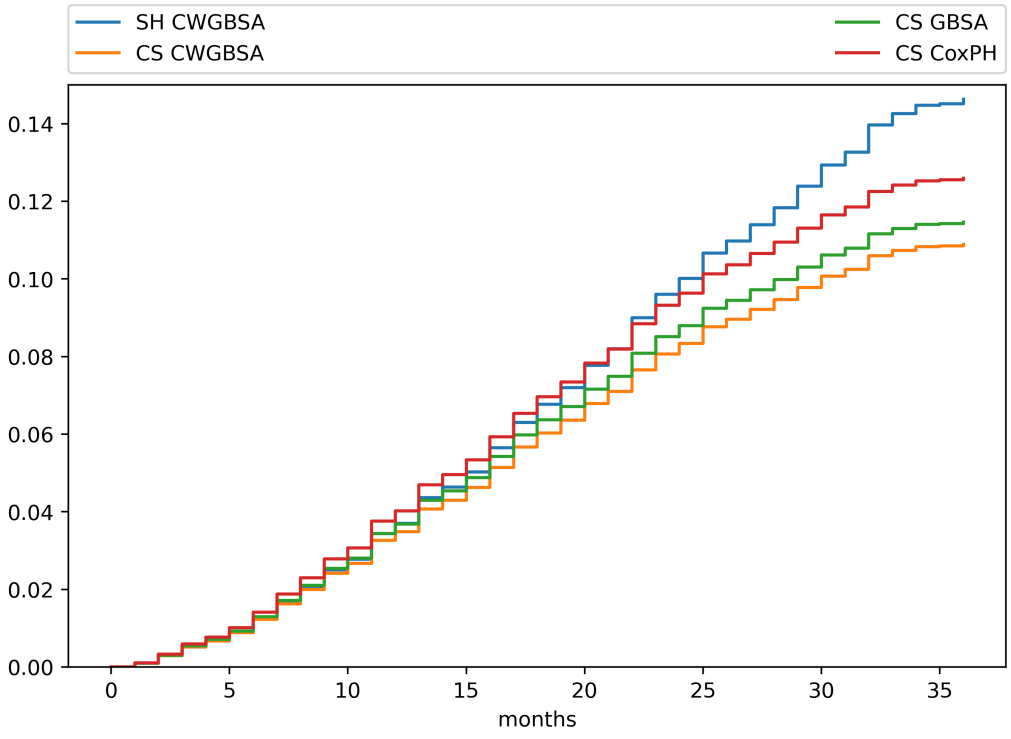
When considering different interest rates and home ownership statuses, SH CWGBSA demonstrated impacts on the curve solely by varying the interest rate. This observation may be attributed to the training structure, which generates predictions based on a limited number of covariates.

Figure 26 – Cumulative Probability of default for an operation with interest rate of 12% and home ownership assigned as (a) Rent and (b) Mortgage



(a) (b)

Figure 27 – Predicted cumulative probability of default



Boosting has been shown to improve the prediction accuracy of survival analysis models, particularly for high-dimensional data (MAYR et al., 2014). However, the interpretability of the GBM can be challenging, as it combines many weak learners to make the final prediction. Therefore, it is important to carefully consider the trade-off between model accuracy and interpretability when using boosting in survival analysis.

References

AGARWAL, S.; AMBROSE, B. W.; LIU, C. Credit lines and credit utilization. *Journal of Money, Credit and Banking*, JSTOR, p. 1–22, 2006. Cited 2 times ins pages 56 e 58.

AKTER, S. et al. Algorithmic bias in machine learning-based marketing models. *Journal of Business Research*, Elsevier BV, v. 144, p. 201–216, maio 2022. Disponível em: <https://doi.org/10.1016/j.jbusres.2022.01.083>. Cited in page 16.

ANDERSEN, P. K.; KEIDING, N. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, Wiley Online Library, v. 31, n. 11-12, p. 1074–1088, 2012. Cited in page 60.

ANDREEVA, G. European generic scoring models using survival analysis. *Journal of the Operational Research Society*, Informa UK Limited, v. 57, n. 10, p. 1180–1187, out. 2006. Disponível em: <https://doi.org/10.1057/palgrave.jors.2602091>. Cited in page 36.

ANDREEVA, G.; ANSELL, J.; CROOK, J. Modelling profitability using survival combination scores. *European Journal of Operational Research*, Elsevier BV, v. 183, n. 3, p. 1537–1549, dez. 2007. Disponível em: <https://doi.org/10.1016/j.ejor.2006.10.064>. Cited in page 36.

APOSTOLIK, R. et al. *Foundations of banking risk: an overview of banking, banking risks, and risk-based banking regulation*. [S.l.]: John Wiley, 2009. Cited in page 14.

ASHOK, M. et al. Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, Elsevier BV, v. 62, p. 102433, fev. 2022. Disponível em: <https://doi.org/10.1016/j.ijinfomgt.2021.102433>. Cited in page 16.

AUSTIN, P. C.; FINE, J. P. Practical recommendations for reporting fine-gray model analyses for competing risk data. *Statistics in medicine*, Wiley Online Library, v. 36, n. 27, p. 4391–4400, 2017. Cited in page 61.

AUSTIN, P. C.; LEE, D. S.; FINE, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, Am Heart Assoc, v. 133, n. 6, p. 601–609, 2016. Cited in page 62.

AUSTIN, P. C.; STEYERBERG, E. W.; PUTTER, H. Fine-gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: cumulative total failure probability may exceed 1. *Statistics in Medicine*, Wiley Online Library, v. 40, n. 19, p. 4200–4212, 2021. Cited in page 61.

BAI, M.; ZHENG, Y.; SHEN, Y. Gradient boosting survival tree with applications in credit scoring. *Journal of the Operational Research Society*, Informa UK Limited, v. 73, n. 1, p. 39–55, jun. 2021. Disponível em: <https://doi.org/10.1080/01605682.2021.1919035>. Cited 3 times ins pages 37, 57 e 58.

BALAZY, K. et al. Prognostic model using a simple survival tree algorithm for patients undergoing palliative radiation. *International Journal of Radiation Oncology*∗*Biology*∗*Physics*, *Elsevier BV*, *v.* 105, *n.* 1, *p.* *E*581, *set.*2019. *Citedinpage*37.

BANASIK, J.; CROOK, J. N.; THOMAS, L. C. Not if but when will borrowers default. Journal of the Operational Research Society*, Taylor & Francis, v. 50, n. 12, p. 1185–1190, 1999. Cited 4 times ins pages 56, 57, 58 e 59.*

BAROCAS, S.; HARDT, M.; NARAYANAN, A. Fairness in machine learning. Nips tutorial*, v. 1, p. 2017, 2017. Cited 2 times ins pages 12 e 17.*

BAROCAS, S.; SELBST, A. D. Big data's disparate impact. California law review*, JSTOR, p. 671–732, 2016. Cited in page 17.*

BCBS. Studies on the Validation of Internal Rating Systems. 2005. Cited in page 34.

BCBS. Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version*. [S.l.: s.n.], 2006. 285 p. ISBN 9291316695. Cited in page 34.*

BELLE, V. V. et al. Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial Intelligence in Medicine*, Elsevier BV, v. 53, n. 2, p. 107–118, out. 2011. Disponível em: <https://doi.org/10.1016/j.artmed.2011.06.006>. Cited in page 37.*

BELLINI, T. Chapter 3 - lifetime pd. In: BELLINI, T. (Ed.). IFRS 9 and CECL Credit Risk Modelling and Validation*. Academic Press, 2019. p. 91 – 153. ISBN 978-0-12-814940-9. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780128149409000116>. Cited in page 40.*

BELLOTTI, T.; CROOK, J. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society*, v. 60, n. 12, p. 1699–1707, 2008. Cited 2 times ins pages 37 e 38.*

BELLOTTI, T.; CROOK, J. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society*, Informa UK Limited, v. 60, n. 12, p. 1699–1707, dez. 2009. Disponível em: <https://doi.org/10.1057/jors.2008.130>. Cited in page 57.*

BINDER, H. et al. Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics*, Oxford University Press, v. 25, n. 7, p. 890–896, 2009. Cited 3 times ins pages 57, 64 e 73.*

BINDER, H.; SCHUMACHER, M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC bioinformatics*, Springer, v. 9, p. 1–10, 2008. Cited in page 64.*

BIS. IFRS 9 and expected loss provisioning - Executive Summary*. [S.l.], 2017. Cited in page 55.*

BOU-HAMAD, I. et al. Discrete-time survival trees. Canadian Journal of Statistics*, v. 37, n. 1, p. 17–32, 2009. ISSN 03195724. Cited 2 times ins pages 39 e 40.*

BREIMAN, L. Algorithm cart. Classification and Regression Trees. California Wadsworth International Group, Belmont, California*, 1984. Cited in page 37.*

BREIMAN, L. Bagging predictors. Machine learning*, Springer, v. 24, p. 123–140, 1996. Cited in page 21.*

BREIMAN, L. Random forests. Machine learning, *Springer, v. 45, p. 5–32, 2001. Cited in page 21.*

BROWN, K.; MOLES, P. Credit risk management. K. Brown & P. Moles, Credit Risk Management, *v. 16, 2014. Cited in page 12.*

BUEHLMANN, P. Boosting for high-dimensional linear models. The Annals of Statistics, *Institute of Mathematical Statistics, v. 34, n. 2, p. 559–583, 2006. Cited in page 41.*

BÜHLMANN, P.; HOTHORN, T. Boosting algorithms: Regularization, prediction and model fitting. 2007. Cited in page 64.

CAO RICARDO, V. J. M. D. A. Modelling consumer credit risk via survival analysis. SORT, *v. 33, n. 1, p. 3–30, 2009. Disponível em: <http://eudml.org/doc/43039>. Cited in page 57.*

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, *v. 16, p. 321–357, 2002. Cited in page 22.*

CHEN, J. et al. Re-default risk of modified mortgages. International Real Estate Review, *Global Social Science Institute, v. 21, n. 1, p. 1–40, 2018. Cited in page 58.*

CHEN, Y. et al. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. Computational and Mathematical Methods in Medicine, *Hindawi Limited, v. 2013, p. 1–8, 2013. Disponível em: <https://doi.org/10.1155/2013/873595>. Cited 2 times ins pages 57 e 58.*

CHOPRA, A.; BHILARE, P. Application of ensemble models in credit scoring models. Business Perspectives and Research, *SAGE Publications, v. 6, n. 2, p. 129–141, abr. 2018. Disponível em: <https://doi.org/10.1177/2278533718765531>. Cited in page 36.*

CHUI M., H. B. S. A.; SUKHAREVSKY, A. The state of ai in 2021. retrieved from: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021 (accessed february 14th, 2022). 2021. Cited in page 11.

COHN, S. L. et al. The International Neuroblastoma Risk Group (INRG) classification system: An INRG task force report. Journal of Clinical Oncology, *v. 27, n. 2, p. 289–297, 2009. ISSN 0732183X. Cited in page 39.*

COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. *[S.l.]: Editora Blucher, 2006. Cited 2 times ins pages 59 e 60.*

COX, D. R. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), *Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Cited 3 times ins pages 38, 56 e 60.*

COX, D. R. Partial likelihood. Biometrika, *Oxford University Press, v. 62, n. 2, p. 269–276, 1975. Cited in page 60.*

CROOK, J. N.; EDELMAN, D. B.; THOMAS, L. C. Recent developments in consumer credit risk assessment. European Journal of Operational Research, *Elsevier, v. 183, n. 3, p. 1447–1465, 2007. Cited in page 14.*

DE-ARTEAGA, M.; FEUERRIEGEL, S.; SAAR-TSECHANSKY, M. Algorithmic fairness in business analytics: Directions for research and practice. Production and Operations Management*, Wiley Online Library, v. 31, n. 10, p. 3749–3770, 2022. Cited in page 15.*

DENG, Y.; QUIGLEY, J. M.; ORDER, R. V. Mortgage terminations, heterogeneity and the exercise of mortgage options. Econometrica*, Wiley Online Library, v. 68, n. 2, p. 275–307, 2000. Cited in page 56.*

DIRICK, L.; CLAESKENS, G.; BAESENS, B. An akaike information criterion for multiple event mixture cure models. European Journal of Operational Research*, Elsevier BV, v. 241, n. 2, p. 449–457, mar. 2015. Disponível em: <https://doi.org/10.1016/j.ejor.2014.08.038>. Cited in page 57.*

DIRICK, L.; CLAESKENS, G.; BAESENS, B. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society*, Informa UK Limited, v. 68, n. 6, p. 652–665, jun. 2017. Disponível em: <https://doi.org/10.1057/s41274-016-0128-9>. Cited 4 times ins pages 36, 55, 57 e 59.*

DJEUNDJE, V. B.; CROOK, J. Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. European Journal of Operational Research*, Elsevier, v. 271, n. 2, p. 697–709, 2018. Cited in page 36.*

DUAN, L. et al. Cluster-based outlier detection. Annals of Operations Research*, Springer, v. 168, n. 1, p. 151–168, 2009. Cited in page 11.*

DUROVIĆ, A. Macroeconomic Approach to Point in Time Probability of Default Modeling - IFRS 9 Challenges. Journal of Central Banking Theory and Practice*, v. 8, n. 1, p. 209–223, 2019. ISSN 23369205. Cited in page 38.*

FANTAZZINI, D.; FIGINI, S. Random survival forests models for SME credit risk measurement. Methodology and Computing in Applied Probability*, Springer Science and Business Media LLC, v. 11, n. 1, p. 29–45, maio 2008. Disponível em: <https://doi.org/10.1007/s11009-008-9078-2>. Cited in page 36.*

FINE, J. P.; GRAY, R. J. A proportional hazards model for the subdistribution of a competing risk. Journal of the American statistical association*, Taylor & Francis, v. 94, n. 446, p. 496–509, 1999. Cited 4 times ins pages 61, 62, 63 e 65.*

FINREGLAB. The Use of Cash-Flow Data in Underwriting Credit*. 2019. Https://finreglab.org/wp-content/uploads/2019/07/FRL$_R$esearch − Report$_F$inal.pdf. Citedinpage*17*.*

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997. Cited 2 times ins pages 21 e 62.*

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000. Cited in page 63.*

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. Annals of statistics*, JSTOR, p. 1189–1232, 2001. Cited 3 times ins pages 40, 62 e 63.*

FRYDMAN, H.; MATUSZYK, A. Random survival forest for competing credit risks. Journal of the Operational Research Society*, Taylor & Francis, v. 73, n. 1, p. 15–25, 2022. Cited 4 times ins pages 58, 61, 62 e 68.*

GARCÍA, V.; MOLLINEDA, R. A.; SÁNCHEZ, J. S. Index of balanced accuracy: A performance measure for skewed class distributions. In: SPRINGER. Pattern Recognition and Image Analysis: 4th Iberian Conference, IbPRIA 2009 Póvoa de Varzim, Portugal, June 10-12, 2009 Proceedings 4*. [S.l.], 2009. p. 441–448. Cited in page 23.*

GESKUS, R. B. Data analysis with competing risks and intermediate states*. [S.l.]: CRC Press, 2015. v. 82. Cited in page 58.*

GIFFEN, B. van; HERHAUSEN, D.; FAHSE, T. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research*, Elsevier BV, v. 144, p. 93–106, maio 2022. Disponível em: <https://doi.org/10.1016/j.jbusres.2022.01.076>. Cited in page 16.*

GORNJAK, M. Literature review of ifrs 9 and its key parameters. Management*, v. 20, p. 22, 2020. Cited in page 57.*

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 160, n. 3, p. 523–541, 1997. Cited in page 36.*

HARRELL, F. E. et al. Evaluating the yield of medical tests. Jama*, American Medical Association, v. 247, n. 18, p. 2543–2546, 1982. Cited 2 times ins pages 41 e 69.*

HARRISON, T.; ANSELL, J. Customer retention in the insurance industry: Using survival analysis to predict cross-selling opportunities. Journal of Financial Services Marketing*, Springer, v. 6, n. 3, p. 229–239, 2002. Cited in page 55.*

HASTIE, T. et al. The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009. v. 2. Cited 3 times ins pages 20, 21 e 22.*

HIDO, S.; KASHIMA, H.; TAKAHASHI, Y. Roughly balanced bagging for imbalanced data. Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 2, n. 5-6, p. 412–426, 2009. Cited in page 22.*

International Accounting Standards Board. International Financial Reporting Standard 9 Financial instruments*. [S.l.], 2014. Cited 2 times ins pages 12 e 35.*

ISHWARAN, H. et al. Random survival forests. The annals of applied statistics*, Institute of Mathematical Statistics, v. 2, n. 3, p. 841–860, 2008. Cited in page 40.*

ISHWARAN, H. et al. Random survival forests. Annals of Applied Statistics*, v. 2, n. 3, p. 841–860, 2008. ISSN 19326157. Cited 2 times ins pages 57 e 58.*

JAGTIANI, J.; LEMIEUX, C. The roles of alternative data and machine learning in fintech lending: evidence from the lendingclub consumer platform. Financial Management*, Wiley Online Library, v. 48, n. 4, p. 1009–1029, 2019. Cited in page 16.*

JAMES, G. et al. An introduction to statistical learning. *[S.l.]: Springer, 2013. v. 112. Cited in page 20.*

KALBFLEISCH, J. D.; PRENTICE, R. L. Marginal likelihoods based on cox's regression and life model. Biometrika, *Oxford University Press, v. 60, n. 2, p. 267–278, 1973. Cited in page 38.*

KALBFLEISCH, J. D.; PRENTICE, R. L. The statistical analysis of failure time data. *[S.l.]: John Wiley & Sons, 2011. Cited in page 61.*

KALLUS, N.; MAO, X.; ZHOU, A. Assessing algorithmic fairness with unobserved protected class using data combination. Management Science, *INFORMS, v. 68, n. 3, p. 1959–1981, 2022. Cited in page 15.*

KAUFFMAN, R. J.; WANG, B. The success and failure of dotcoms: A multi-method survival analysis. In: CITESEER. Proceedings of the 6th INFORMS Conference on Information Systems and Technology (CIST). *[S.l.], 2001. Cited 2 times ins pages 15 e 55.*

KLINE, P.; ROSE, E. K.; WALTERS, C. R. Systemic discrimination among large u.s. employers. The Quarterly Journal of Economics, *Oxford University Press (OUP), v. 137, n. 4, p. 1963–2036, jun. 2022. Disponível em: <https://doi.org/10.1093/qje/qjac024>. Cited in page 15.*

KOZODOI, N.; JACOB, J.; LESSMANN, S. Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, *v. 297, n. 3, p. 1083–1094, 2022. ISSN 0377-2217. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377221721005385>. Cited 2 times ins pages 14 e 17.*

KUBAT, M.; MATWIN, S. et al. Addressing the curse of imbalanced training sets: one-sided selection. In: CITESEER. Icml. *[S.l.], 1997. v. 97, n. 1, p. 179. Cited in page 23.*

KUMAR, I. E.; HINES, K. E.; DICKERSON, J. P. Equalizing credit opportunity in algorithms: Aligning algorithmic fairness research with us fair lending regulation. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. *[S.l.: s.n.], 2022. p. 357–368. Cited in page 16.*

LEE, C. et al. DeepHit: A deep learning approach to survival analysis with competing risks. Proceedings of the AAAI Conference on Artificial Intelligence, *Association for the Advancement of Artificial Intelligence (AAAI), v. 32, n. 1, abr. 2018. Disponível em: <https://doi.org/10.1609/aaai.v32i1.11842>. Cited in page 58.*

LEFEBVRE-ULRIKSON, W. et al. Data Mining. *[S.l.: s.n.], 2016. ISBN 9780128047453. Cited in page 36.*

LESSMANN, S. et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, *Elsevier, v. 247, n. 1, p. 124–136, 2015. Cited in page 36.*

LI, Z. et al. The profitability of online loans: A competing risks analysis on default and prepayment. European Journal of Operational Research, *Elsevier BV, v. 306, n. 2, p. 968–985, abr. 2023. Disponível em: <https://doi.org/10.1016/j.ejor.2022.08.013>. Cited 3 times ins pages 56, 57 e 58.*

LUNN, M.; MCNEIL, D. Applying cox regression to competing risks. Biometrics, *JSTOR, p. 524–532, 1995. Cited in page 56.*

MACLIN, R.; OPITZ, D. An empirical evaluation of bagging and boosting. AAAI/IAAI, *Citeseer, v. 1997, p. 546–551, 1997. Cited in page 22.*

MAKHLOUF, K.; ZHIOUA, S.; PALAMIDESSI, C. Machine learning fairness notions: Bridging the gap with real-world applications. Information Processing &amp Management, *Elsevier BV, v. 58, n. 5, p. 102642, set. 2021. Disponível em: <https://doi.org/10.1016/j.ipm.2021.102642>. Cited 3 times ins pages 15, 16 e 17.*

MAYR, A. et al. The evolution of boosting algorithms. Methods of information in medicine, *Schattauer GmbH, v. 53, n. 06, p. 419–427, 2014. Cited in page 75.*

NARAIN, B. Survival analysis and the credit granting decision. Credit scoring and credit control, *Clarendon Press Oxford, v. 109, p. 121, 1992. Cited 2 times ins pages 55 e 57.*

NARAIN, B. Survival analysis and the credit granting decision. LC Thomas, JN Crook, DB Edelman, eds. Credit Scoring and Credit Control. *[S.l.]: OUP, Oxford, UK, 1992. Cited in page 36.*

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. Journal of the Royal Statistical Society: Series A (General), *Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. Cited 2 times ins pages 20 e 63.*

NILOY, N.; NAVID, M. Naïve bayesian classifier and classification trees for the predictive accuracy of probability of default credit card clients. American Journal of Data Mining and Knowledge Discovery, *Science Publishing Group, v. 3, n. 1, p. 1, 2018. Cited in page 36.*

NILSSON, N. J. Principles of artificial intelligence. Burlington, MA. *[S.l.]: Morgan Kaufmann, 2014. Cited in page 11.*

PARIZADEH, D. et al. Exploring risk patterns for incident ischemic stroke during more than a decade of follow-up: A survival tree analysis. Computer Methods and Programs in Biomedicine, *Elsevier BV, v. 147, p. 29–36, ago. 2017. Disponível em: <https://doi.org/10.1016/j.cmpb.2017.06.006>. Cited in page 37.*

PINTILIE, M. Competing risks: a practical perspective. *[S.l.]: John Wiley & Sons, 2006. Cited 2 times ins pages 62 e 71.*

PÖLSTERL, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research, *v. 21, n. 212, p. 1–6, 2020. Disponível em: <http://jmlr.org/papers/v21/20-729.html>. Cited in page 65.*

RIDGEWAY, G. The state of boosting. Computing science and statistics, *Citeseer, p. 172–181, 1999. Cited 3 times ins pages 62, 63 e 64.*

SANTOS, P. F.; SAAVEDRA, C. A. P. B.; KIMURA, H. Default prediction in special overdraft checking accounts using machine learning algorithm. 2023. Cited in page 18.

SCHUSTER, N. A. et al. Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. Journal of Clinical Epidemiology*, Elsevier BV, v. 122, p. 42–48, jun. 2020. Disponível em: <https://doi.org/10.1016/j.jclinepi.2020.03.004>. Cited in page 58.*

SETTLEMENTS, B. Bank for I. Humans keeping AI in check – emerging regulatory expectations in the financial sector*. FSI Insights on policy implementation, 2021. FSI Insights on policy implementation. Cited 2 times ins pages 14 e 31.*

SINGAPORE, M. Monetary Authority of. Implementation of Fairness Principles in Financial Instituion's use of Artificial Intelligence / Machine Learning*. Thematic Review, 2022. Thematic review. Cited 2 times ins pages 14 e 31.*

STEINBUKS, J. Effects of prepayment regulations on termination of subprime mortgages. Journal of Banking & Finance*, Elsevier, v. 59, p. 445–456, 2015. Cited in page 56.*

STEPANOVA, M.; THOMAS, L. Survival analysis methods for personal loan data. Operations Research*, INFORMS, v. 50, n. 2, p. 277–289, 2002. Cited 3 times ins pages 56, 57 e 58.*

STEPANOVA, M.; THOMAS, L. C. PHAB scores: proportional hazards analysis behavioural scores. Journal of the Operational Research Society*, Informa UK Limited, v. 52, n. 9, p. 1007–1016, set. 2001. Disponível em: <https://doi.org/10.1057/palgrave.jors.2601189>. Cited in page 57.*

SUPERVISION, B. C. on B.; SETTLEMENTS, B. for I. Principles for the management of credit risk*. [S.l.]: Bank for International Settlements, 2000. Cited in page 11.*

THACKHAM, M.; MA, J. On maximum likelihood estimation of competing risks using the cause-specific semi-parametric cox model with time-varying covariates–an application to credit risk. Journal of the Operational Research Society*, Taylor & Francis, v. 73, n. 1, p. 5–14, 2022. Cited in page 56.*

THOMAS, L.; CROOK, J.; EDELMAN, D. Credit scoring and its applications*. [S.l.]: SIAM, 2017. Cited in page 33.*

TONG, E. N.; MUES, C.; THOMAS, L. C. Mixture cure models in credit scoring: If and when borrowers default. European Journal of Operational Research*, Elsevier BV, v. 218, n. 1, p. 132–139, abr. 2012. Disponível em: <https://doi.org/10.1016/j.ejor.2011.10.007>. Cited in page 57.*

TRACY, J.; WRIGHT, J. Payment changes and default risk: The impact of refinancing on expected credit losses. Journal of urban Economics*, Elsevier, v. 93, p. 60–70, 2016. Cited in page 58.*

UEJIO, H. D.; BUREAU, C. F. P. Request for information and comment on financial institutions' use of artificial intelligence, including machine learning. 2021. (Docket No. CFPB-2021-0004). Cited in page 16.

VANĚK, T.; HAMPEL, D. The probability of default under ifrs 9: Multi-period estimation and macroeconomic forecast. Acta universitatis agriculturae et silviculturae mendelianae brunensis, *Mendelova univerzita v Brně, 2017. Cited in page 59.*

VERWEIJ, P. J.; HOUWELINGEN, H. C. V. Penalized likelihood in cox regression. Statistics in medicine, *Wiley Online Library, v. 13, n. 23-24, p. 2427–2436, 1994. Cited in page 38.*

WRIGHT, M. N.; DANKOWSKI, T.; ZIEGLER, A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Statistics in Medicine, *Wiley, v. 36, n. 8, p. 1272–1284, jan. 2017. Disponível em: <https://doi.org/10.1002/sim.7212>. Cited in page 58.*

XIA, Y. et al. A dynamic credit scoring model based on survival gradient boosting decision tree approach. Technological and Economic Development of Economy, *v. 27, n. 1, p. 96–119, 2021. Cited in page 37.*

XU, L. D.; LU, Y.; LI, L. Embedding blockchain technology into iot for security: A survey. IEEE Internet of Things Journal, *IEEE, v. 8, n. 13, p. 10452–10473, 2021. Cited in page 11.*

YANG, W. et al. Purchase prediction in free online games via survival analysis. In: IEEE. 2019 IEEE International Conference on Big Data (Big Data). *[S.l.], 2019. p. 4444–4449. Cited in page 55.*

YEH, I. C.; LIEN, C. hui. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, *Elsevier Ltd, v. 36, n. 2 PART 1, p. 2473–2480, 2009. ISSN 09574174. Cited in page 36.*

ZETTEN, W. van; RAMACKERS, G.; HOOS, H. Increasing trust and fairness in machine learning applications within the mortgage industry. Machine Learning with Applications, *Elsevier BV, v. 10, p. 100406, dez. 2022. Disponível em: <https://doi.org/10.1016/j.mlwa.2022.100406>. Cited in page 15.*

ZHANG, C.; LU, Y. Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, *Elsevier, v. 23, p. 100224, 2021. Cited in page 11.*

ZHANG, J.; THOMAS, L. C. Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. International Journal of Forecasting, *Elsevier BV, v. 28, n. 1, p. 204–215, jan. 2012. Disponível em: <https://doi.org/10.1016/j.ijforecast.2010.06.002>. Cited in page 57.*

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), *Wiley Online Library, v. 67, n. 2, p. 301–320, 2005. Cited in page 38.*