Master Study

Field of study Advanced Analytics – Big Data

Author's first name and surname

Bartosz Bogucki

Student's register No. 82801

Application of survival analysis in estimating probability of default: a benchmark study with logistic regression for mortgage loans

Master's thesis:
under the scientific supervision of
prof. dr hab. Bogumił Kamiński
Institute of Econometrics

Warsaw, 2023

Table of contents

1 Introduction

The assessment and management of credit risk are critical aspects of the banking industry. One of the primary concerns related to credit risk is the default event. Default occurs when a borrower is unable or unwilling to repay its contractual obligation. This can have significant consequences for banks, as it leads to loss of principal and interest, deteriorates the bank's asset quality and reduces its profitability. Therefore, accurate assessment and management of credit risk play a critical role in maintaining banks' financial stability and ensuring the safety of depositors' funds. Banks face the challenge of ensuring that they have sufficient level of credit risk provisions to cover credit losses and maintain solvency and ability to meet their obligations while optimizing income-generating opportunities.

After the global financial crisis, International Accounting Standards Board (IASB) initiated a project to replace International Accounting Standard (IAS) 39 (IASB, 2001) with International Financial Reporting Standard (IFRS) 9, which sets out the guidelines for recognizing and measuring financial instruments in a company's financial statements (IASB, 2014). IFRS 9 requires financial institutions to account for expected credit losses (ECL) on their loans (Cohen and Edwards, 2017). The implementation of IFRS 9 is anticipated to have a significant impact on the systems and processes of financial institutions (Beerbaum, 2015).

While Internal-Ratings-Based (IRB) approach for regulatory purposes employ Through-the-Cycle TTC approach, IFRS 9 emphasizes Point-in-Time (PiT) approach. TTC approach estimates credit losses based on the long-term average credit risk over the economic cycle, rather than at a specific point in time. On the other hand, PiT approach in IFRS 9 estimates credit losses based on the probability of default (PD) at a specific point in time, considering current information and forward-looking factors to assess the credit risk of financial assets.

The most crucial element that influences credit risk is the probability of default (Peláez et al, 2021). The use of statistical methods for prediction of default is now well-known (Bellotti and Crook, 2009). In particular, logistic regression has become a standard method for this task (Thomas et al, 2002). While there is extensive literature on cross-sectional approaches in credit risk modeling, research on survival analysis is relatively less advanced. Nevertheless, survival analysis provides valuable insights into credit risk by predicting the timing of default events (Louzada-Neto, 2006; Tong et al, 2012). As an alternative to logistic regression, Narain (1992) first introduced the idea of using survival analysis in the credit risk context. This approach enables a more dynamic assessment of credit risk, taking into account the time dimension and capturing the exact timing of defaults (Malik and Thomas, 2007). Survival models can provide predictions of dynamic probability of default (PD) over time, such as 12-month or lifetime PDs for loan portfolios, as required by IFRS 9 (Xia et al, 2021). The ability to predict dynamic PD allows financial institutions to accurately adjust credit risk provisions and collection strategies during the repayment of loans. Furthermore, survival models can easily incorporate time-dependent covariates, such as macroeconomic or behavioral factors (Bellotti and Crook, 2009).

The aim of the study is to verify the hypothesis that the application of survival analysis techniques for predicting PD yields a more discriminating model and more accurate PD estimates compared to logistic regression. Consequently, employing survival models leads to more precise estimates of expected credit loss in comparison to those obtained from logistic regression. To achieve this objective, the paper demonstrates the application of survival analysis in predicting the probability of default for estimating expected credit loss based on a portfolio of Italian mortgages using real-life commercial bank data, including information on defaults, collateral, customer application characteristics, transactions, and macroeconomic factors. The paper makes a valuable contribution by conducting a comparison between two approaches for predicting default probabilities: survival analysis and cross-sectional methods. Notably, there is a lack of such comprehensive comparisons in the existing literature. To achieve this objective, several survival models are selected,

including Weibull accelerated failure time, Log-Normal accelerated failure time, Cox proportional hazards and DeepHit. Additionally, logistic regression is chosen as the second approach, primarily because it is widely used in practice for estimating the probability of default (Kleinbaum and Klein, 2010). Furthermore, given a long-term nature of mortgage loans, the study explores survival approaches that can handle time periods beyond the training data to be able to predict lifetime PD. While prior literature has attempted to compare a range of available survival methods for predicting PD (Dirick et al, 2017), the study takes a step further by providing a comprehensive comparison of survival models for a different portfolio, introducing additional methods. It is important to highlight that existing literature mainly focuses on evaluating the effectiveness of survival models for short-term loans, where extrapolation is unnecessary due to the presence of observed times in the training sample. However, this study addresses the crucial need for extrapolating survival models for long-term loans, where data may be insufficient without the use of extrapolation techniques. Additionally, the research deviates from traditional out-of-sample validation approach used in prior studies. Instead, the paper compares model performances using both out-of-sample and out-of-time validation to assess model performance.

The rest of the paper is organized as follows. The methodological approach is presented in Section 2, a description of the data is given in Section 3, empirical results are discussed in Section 4, and concluding remarks are presented in Section 5. The source code for all the methods discussed in this paper can be accessed via the following GitHub repository: *https://github.com/bartekbogucki/Application-of-survival-analysis-to-estimate-PD*. OpenAI (2023) was used in the initial stages of this paper's preparation to generate a limited part of elementary code, which was subsequently extensively modified and adapted for specific research purposes and refined to obtain the final research results.

2 Methodology

This chapter introduces the methodology used in the study. It includes definitions of credit risk components and regulations, along with a description of the pre-processing methods, the models applied and the metrics used to evaluate the results.

2.1 Credit risk

Credit risk stems from the threat of default event by borrower, i.e., failing to meet contractual obligations, leading to financial loss incurred by lender. Credit risk is borne by the lender, while the ability to repay the loan is a borrower's characteristic. The goal of credit risk management is to maximize a bank's risk-adjusted rate of return by maintaining credit risk exposure within acceptable parameters (Basel Committee on Banking Supervision, 2000). For that purpose, entities must be able to measure risk and manage it trying to hedge or transfer that risk outside the entity. The effective credit risk management is a key aspect of a comprehensive approach to risk management and is critical to the long-term prosperity of any banking institution.

2.1.1 Definition of default

Default refers to the failure of a borrower to fulfill their contractual obligations, which typically involves the non-payment or inability to meet financial commitments such as interest or principal repayment on a loan. This failure to meet obligations leads to negative consequences, including financial losses for lenders. In this study, default is applied at the facility level, indicating that a default on one credit obligation by an obligor doesn't automatically classify all other credit obligations as defaulted. Instead, only the specific facility related to the default is treated as such.

The definition of default was introduced by the European Parliament and the Council Directive 2006/48/EC (2006), later replaced by Regulation (EU) No 575/2013 (2013). The definition of default of an obligor specified in these regulations includes, inter alia, the days

past due criterion for default identification and indications of unlikeliness to pay. However, in the absence of specific rules on these and other aspects of the application of the definition of default various approaches were adopted across institutions and jurisdictions. As a consequence a wide range of practices was observed (EBA, 2016).

After the financial crisis, in order to harmonize the approach EBA (2017) established tighter standards around the definition of default to increase the degree of comparability and consistency in credit risk measurement and capital frameworks across banks and financial institutions. The guidelines primarily address how banks identify defaults. Within the scope of this study, the new definition of default according to EBA (2016) is applied. As per this definition, a facility is considered to be in default when either one or both of the following events have taken place:

1. Facility is unlikely to pay.
2. Facility has a material overdue for more than 90 days past due.

The concept of 'unlikeliness to pay' refers to situations where a customer's financial distress prevents or is about to prevent them from meeting their credit obligations. In this study, it can be indicated by factors such as bankruptcy, distressed restructuring, fraud, or non-accrued status. Moreover, a facility is classified as defaulted if credit obligations above the threshold above the threshold remain overdue for more than 90 consecutive calendar days. The materiality threshold consists of both an absolute and a relative limit and is calculated as follows for retail exposures:

- An absolute limit of EUR 100, where overdue amount exceeds EUR 100.
- A relative limit of 1%, where overdue amount as a percentage of the total on-balance sheet exposure to the facility, surpasses 1%.

Once both these limits are breached for over 90 consecutive calendar days, a facility is considered to be in default. It's important to note that the count of days past due begins when both limits are breached.

2.1.2 IFRS 9 expected credit loss calculation

Given the impact of default events on lenders and the resulting credit losses, the calculation of loan provisions becomes imperative. International Financial Reporting Standard 9 (IFRS 9), established by the International Accounting Standards Board in 2014, outlines the requirements for measuring financial instruments and recognizing impairments (IASB, 2014). This framework was established in response to the financial crisis and replaced the previous International Accounting Standard (IAS) 39 (IASB, 2001). Criticisms of the incurred loss approach under IAS 39, which led to delayed recognition of loan losses, prompted standard setters to formulate accounting standards that allow for a more forward looking provisioning (BCBS, 2009; Financial Crisis Advisory Group, 2009; Financial Stability Forum, 2009).

In the standard that preceded IFRS 9, the incurred loss framework required banks to recognize credit losses only when evidence of a loss was apparent (BIS, 2017). The introduction of IFRS 9 brought a significant change by establishing an expected credit loss framework for impairment recognition. This new approach requires the timely identification of credit losses by considering past events, current conditions and forward-looking information (IASB, 2014). The amount of ECLs are recognized at each reporting date to reflect changes in an asset's credit risk. Unlike its predecessor, this methodology adopts a more forward-looking perspective, facilitating more prompt and accurate credit loss recognition (BIS, 2017).

IFRS 9 introduces 12 month ECL, which reflects the expected losses for assets that default over the next 12 months, and lifetime ECL, which reflects the expected losses for assets that default over their remaining maturity (BCBS, 2015). The calculation of provisions and the selection between 12-month and lifetime ECL for this purpose depends on the stage allocation, which indicates the level of credit risk associated with the assets (Aptivaa, 2016). In the context of this study, while the calculation of ECL follows IFRS 9

guidelines, no staging is introduced. The main focus of the research is the application of dynamic estimates of probability of default to estimate expected loss levels.

The measurement of expected credit loss as mandated by IFRS 9 is not limited to a single approach. Diverse techniques are applied to estimate ECL across various industries (Schutte et al, 2020). These techniques can divided into two main categories: direct methods and indirect methods. Direct methods involve directly modeling ECL using factors that drive ECL as explanatory variables (Schutte et al, 2020). On the other hand, indirect methods, referred to as loss component approaches, include techniques like simulation-based models and modularized models. Modularization is an approach that divides ECL into independent components that can be developed independently.

This study employs an indirect approach to calculate expected losses, incorporating components such as probability of default (PD), loss given default (LGD) and exposure at default (EAD). PD represents the likelihood that a debtor will default on its obligations, LGD refers to the percentage of funds that would be unrecoverable if the debtor defaults and EAD represents the unpaid amount owed by the debtor at the time of default. The expected credit loss is equal to the multiplication of all these components. To define ECL function over time, certain assumptions are made for simplification purposes. In this context, one assumption is that when a loan defaults, there is no recovery, meaning loss given default is equal to 1. Additionally, it is assumed that exposure at default at a specific time t and a horizon time Δ is equal to the outstanding amount at time t. Under these assumptions, the ECL function can be formulated as the probability of default multiplied by the outstanding amount (OS). It can be expressed by the following formula:

$$ECL(t+\Delta) = PD(t+\Delta) \cdot LGD(t+\Delta) \cdot EAD(t+\Delta) = PD(t+\Delta) \cdot OS(t). \quad (1)$$

2.2 Data pre-processing techniques

The study presents the results obtained from employing various default probability models on a portfolio of Italian real-life commercial bank mortgage data. Since certain models

require numerical predictive variables and the absence of missing values, pre-processing techniques to address missing values and convert categorical variables into numeric formats are applied prior to the modeling phase. Furthermore, data scaling is employed to prevent issues related to variable magnitudes, thereby ensuring that no individual feature disproportionately impacts distance calculations within the applied model.

CatBoost encoding (Prokhorenkova et al, 2019) is designed to overcome the obstacles of target encoding when dealing with categorical features. Target encoding is a prevalent technique that involves substituting a categorical attribute with the average target value corresponding to that category within the training dataset, coupled with the overall target probability across the entire dataset. Yet, this strategy can inadvertently result in target leakage, as it employs the target variable to explanatory variables. As a result, models employing this approach often struggle with overfitting. In contrast, CatBoost encoding introduces a more sophisticated strategy that incorporates an ordering principle to mitigate the challenges stemming from target leakage (Prokhorenkova et al, 2019). The computation of encoded value is based on observed historical data. Essentially, the target probability linked to the current feature is exclusively computed based on observations that come before it. The transformation formula for a specific observation i within category j in a feature f is given by:

$$EncodedValue_i(f,j) = \frac{Sum\left(y_i \middle| Category(f,j)\right) + Prior}{Count\left(y_i \middle| Category(f,j)\right)}, \quad (2)$$

where $Sum\left(y_i \middle| Category(f,j)\right)$ is the cumulative sum of target values for the specific category j in a feature f, extending up to the current one of observation i, Prior is a constant value calculated as the ratio of the sum of target values across the entire dataset to the total number of observations within the dataset, $Count\left(y_i \middle| Category(f,j)\right)$ is the accumulated count of category j in a feature f, extending up to the current one of observation i. This implementation is time-aware, therefore no random permutations are introduced. The sensitivity of this encoder to data ordering makes it well-suited for

addressing time series problems. By integrating these principles, CatBoost encoding effectively addresses the issues associated with target encoding, enhancing the model's ability to prevent target leakage and improve generalization when faced with previously unobserved data points.

To handle missing values, linear interpolation is applied, a technique that entails estimating and filling data gaps by computing values that lie between the nearest available data points. The methodology adopted in this study identifies missing values, locates the adjacent preceding and subsequent data points, and then calculates values that maintain a proportionate distribution between them. This estimation process is grounded in the assumption of a linear correlation between data points. The estimated values derived from the interpolation process are subsequently integrated into the dataset, resulting in a dataset that is complete.

Following the encoding and interpolation steps, the dataset undergoes scaling. Specifically, min-max feature scaling is employed, a data preprocessing method used to transform numerical features within a dataset into a predefined range. This process involves linearly adjusting the values of each feature to ensure they fall within a specified interval, specifically between 0 and 1. The primary objective is to establish a uniform scale for all features, which can be particularly advantageous for algorithms sensitive to the input feature's scale. The formula is represented as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{3}$$

where x is an original value, and x' is the scaled value. By applying these pre-processing techniques, the data is properly prepared for the modeling phase.

2.3 Probability of default models

Probability of default represents the likelihood that a counterparty will default within a certain period of time, i.e., it will not be able to fulfill its contractual obligations. Two

different approaches to estimating the probability of default are being studied: survival and cross-sectional methods. Multiple survival models are considered, including Weibull accelerated failure time, Log-Normal accelerated failure time, Cox proportional hazards and DeepHit. By considering the time dimension and the precise timing of default events, these models provide predictions of the probability of default over any time interval, such as 12-month or lifetime PDs, as required by IFRS 9 (Xia et al., 2021). Additionally, logistic regression is chosen as a cross-sectional method, mainly because it is widely used in practice to estimate the probability of default (Kleinbaum and Klein, 2010). Unlike survival models, using logistic regression it is possible to estimate PD only over a fixed time horizon, without considering the exact timing of default events, which can occur at any time during the remaining life of the contract.

2.3.1 Logistic regression

Logistic regression is a statistical model developed and popularized primarily by Joseph Berkson (1944). Particularly, binary logistic regression is employed within the study as the target variable has two possible outcomes: default or no default. Logistic regression models the likelihood of an event occurring by employing the log-odds, known as the logit, which is expressed as a linear combination of input variables. Odds here signify the probability of an event happening compared to the probability of it not happening. Mathematically,

$$\text{odds} = \frac{P}{1-P} \rightarrow logit(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_i^n \beta_i x_i, \tag{4}$$

where P is the probability of occurrence of the event, where β_0 is the intercept for the logistic regression, β_i is the logistic regression parameter for characteristic i, x_i is the value for characteristic i. Hence, the logistic sigmoid function is used to map the linear combination of variables to a probability value between 0 and 1:

$$\left(\frac{P}{1-P}\right) = e^{(\beta_0 + \sum_i^n \beta_i x_i)} \rightarrow P = \frac{e^{-(\beta_0 + \sum_i^n \beta_i x_i)}}{1 + e^{-(\beta_0 + \sum_i^n \beta_i x_i)}} = \frac{1}{1 + e^{-(\beta_0 + \sum_i^n \beta_i x_i)}}. \tag{5}$$

The parameters of a logistic regression are estimated by maximum-likelihood estimation (MLE). In general, MLE is an optimization technique used to determine the parameters of any assumed probability distribution from observed data. The underlying principle of MLE is to identify parameter values that maximize the likelihood function, which quantifies how probable it is to observe the given data given the model and its parameters. The likelihood function represents the probability of observing the given data for different parameter values. Mathematically, it can be expressed as the product of the probability density function (PDF) or probability mass function (PMF) for each observation, evaluated at the observed values with the parameters as inputs. In formulaic terms, if $p(x, \theta)$ represents the probability density function, where x is the data point and θ represents the parameters, the total likelihood for n data observations is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i, \theta). \tag{6}$$

Given that the output variable is a Bernoulli random variable, which can only assume two values (default or no default), the likelihood function for logistic regression takes the following form:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}. \tag{7}$$

The goal of MLE is to determine the parameter values that maximize the likelihood function, denoted as $\hat{\theta} = argmax(\mathcal{L}(\theta))$. To simplify the maximization process, it is more convenient to work with the natural logarithm of the likelihood function, known as the log-likelihood:

$$\ln\big(\mathcal{L}(\theta)\big) = \ln\left(\prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}\right) = \sum_{i=1}^{n} \ln\left(p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}\right). \tag{8}$$

The maximization process involves taking the partial derivative with respect to θ, setting it equal to zero, and solving for θ:

$$\frac{\partial(\ln(\mathcal{L}(\theta)))}{\partial\theta} = 0. \tag{9}$$

The solution to the above equation yields $\hat{\theta}$, which is referred to as the maximum likelihood estimate. Additionally, for a more extensive discussion on the logistic regression model and its estimation, see Kleinbaum and Klein (2010).

In the context of the study, for a fixed time, t, and a horizon time, Δ, PD can be defined as the probability that a loan will be defaulted no later than time $t + \Delta$. To compare logistic regression with survival models, PD estimates are compared to the true observed default events over the same one-year horizon for both approaches. In the case of logistic regression within a one-year time horizon, PD estimates are obtained using the target variable as whether the customer defaults in this 12-month following period. As discussed above, the probability of default in case of logistic regression is equal to the sigmoidal transformation of the linear function of the combination of explanatory variables:

$$PD(t+12) = \frac{1}{1 + e^{-(\beta_0 + \sum_i^n \beta_i x_i)}}. \tag{10}$$

2.3.2 Survival models

Survival analysis is a statistical method used to analyze and model the time until an event of interest occurs. In the context of the study, the event of interest is the occurrence of default, therefore results in the study are achieved by using the default indicator as the target variable. Survival analysis takes into account censored data, covering observations in which a default event has not been observed up to the time of data collection. In this study, the concept of censoring in the data applies to loans that did not encounter a default at the time of data collection. An observation is marked as censored if the default event did not occur during the study period, but occurred later, with the exact time not captured. Early repayment and mature cases are also considered censored observations.

The survival function, denoted as $S(t)$, gives the probability that default event has not occurred by time t. The survival function can be expressed as $S(t) = P(T > t)$, where T

represents the time until the event of default occurs. An important property of the survival function is that it is monotonically decreasing, as illustrated by the relationship $S(t_1)\rangle S(t_2)$, where $t_1\langle t_2$. This signifies that the probability of survival, i.e. the non-occurrence of a default event, decreases over time. The cumulative distribution function (CDF), denoted as $F(t)$, gives the probability that the event of interest has occurred by time t. Mathematically, $F(\mathrm{t}) = \mathrm{P}(\mathrm{T} \leq \mathrm{t})$. Given that the survival function can be represented as:

$$S(t) = P(T > t) = 1 - \mathrm{P}(\mathrm{T} \leq \mathrm{t}) = 1 - F(t). \tag{11}$$

Consequently, the relationship between the survival function and the cumulative distribution function is given by $F(t) = 1 - S(t)$. Furthermore, according to theoretical principles, the probability density function, denoted as $f(t)$, is the derivative of the distribution function. This implies that the density function represents the negative derivative of the survival function:

$$f(t) = \frac{d}{dt}\mathrm{F}(\mathrm{t}) = \frac{d}{dt}\big(1 - \mathrm{S}(t)\big) = -\frac{d}{dt}\mathrm{S}(t). \tag{12}$$

The hazard function represents the instantaneous rate at which an event of interest occurs at a specific point in time, given that the event has not occurred before that time. This function provides insights into how the risk of the default event changes over time. Mathematically, the hazard function is defined as:

$$\mathrm{h}(\mathrm{t}) = \lim_{\delta \to \infty} \frac{P(t < T < t + \delta | T < t)}{\delta}. \tag{13}$$

Additionally, applying the theorem of conditional probability and omitting suffix T, the hazard function can also be formulated using the survival function and the probability density function as follows:

$$\mathrm{h}(\mathrm{t}) = \frac{f(t)}{S(t)}. \tag{14}$$

Essentially, the survival function represents the probability of not experiencing the event up to time t, while the cumulative distribution function represents the probability of

experiencing the event up to time t. Thus, the probability of default is estimated by subtracting the probability of survival from 1:

$$PD(t + \Delta) = 1 - S(t + \Delta) = F(t + \Delta), \tag{15}$$

where $S(t)$ is the predicted probability of survival, which is estimated in accordance with a selected survival model. The survival models under consideration include Weibull accelerated failure time, Log-Normal accelerated failure time, Cox proportional hazards and DeepHit.

2.3.2.1 Accelerated failure time model

Accelerated failure time (AFT) model is a fully parametric variant of the survival model. In AFT models, explanatory variables impact the accelerated failure rate, denoted as λ, to either speed up or slow down the event occurrence compared to the baseline survival function. Mathematically, the survival function of the AFT model is expressed as follows:

$$S(\mathrm{t}|x_i) = S_0(t\lambda) = S_0\left(te^{\sum_i^n \beta_i x_i}\right), \tag{16}$$

where β_i is the coefficient for characteristic i, x_i is the value for characteristic i, $S(\mathrm{t}|x_i)$ is the survival function at time t for a given set of explanatory variables x_i, $S_0(t)$ is the baseline survival function without any influence from explanatory variables (where $\sum_i^n \beta_i x_i$, is 0), $e^{\sum_i^n \beta_i x_i}$ represents the acceleration factor, which can either decelerate ($0 < e^{\sum_i^n \beta_i x_i} < 1$) or accelerate ($e^{\sum_i^n \beta_{i*} x_i} > 1$) the event time. The hazard function in AFT models is given by:

$$h(\mathrm{t}|x_i) = \lambda h_0(t\lambda) = e^{\sum_i^n \beta_i x_i} h_0\left(te^{\sum_i^n \beta_i x_i}\right), \tag{17}$$

$h(\mathrm{t}|x_i)$ is the hazard function at time t for a given set of explanatory variables x_i, and $h_0(t)$ is the baseline hazard function. The corresponding AFT model can be expressed in a regression form as a log-linear model for the natural logarithm of the event time T (Collett, 2003):

$$\ln(T) = \sum_i^n \beta_i x_i + \sigma\varepsilon, \tag{18}$$

where ε is a random variable with density function $f_0(\varepsilon)$ and the corresponding baseline survival function $S_0(\varepsilon)$.

It is important to emphasize that parametric survival models, which include AFT models, have the ability to make predictions beyond the maximum observed survival time of training data due to the fact that survival times are assumed to follow a parametric distribution. Consequently, AFT model does not assume that all events occur within the predefined duration grid specified by the training data.

Similarly to logistic regression, the parameters of AFT models are determined using the maximum likelihood estimation technique. Following the same approach, a subject observed to fail at t contributes $f(t|\theta)$ to the total likelihood, where f is the density function, and the facility which survives after time t contributes $S(t|\theta)$ to the total likelihood. For n data observations the formula is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} (f(t|\theta))^{y_i} (S(t|\theta))^{1-y_i}. \tag{19}$$

Consequently,

$$\ln\big(\mathcal{L}(\theta)\big) = \sum y_i f(t|\theta) + \sum (1 - y_i) S(t|\theta). \tag{20}$$

Additionally, to avoid overfitting, L_2 regularization is introduced. L_2 regularization adds a penalty term to the negative natural logarithm of the total likelihood $\mathcal{L}(\theta)$. The penalty term is proportional to the sum of the squared coefficients of the model. The coefficient of this penalty is set to 10. L_2 will not yield sparse models and all coefficients are shrunk by the same factor. Subsequently, during the optimization process, the goal is to minimize this regularized loss. It can be expressed as

$$\tilde{\mathcal{L}}(\theta) = -\ln\big(\mathcal{L}(\theta)\big) + \delta \sum_{i=1}^{n} {\beta_i}^2, \delta = 10. \tag{21}$$

This encourages the optimization process to find parameter values that are less likely to overfit the data, namely $\hat{\theta} = argmin(\tilde{\mathcal{L}}(\theta))$.

Narain (1992) introduced AFT models as the first survival models within the context of credit risk, offering an alternative to logistic regression. For a more detailed discussion of AFT models, please refer to Kleinbaum and Klein (2011). These AFT models consider a wide range of survival distributions characterized by event times that exhibit logarithmic linearity. This paper specifically employs two AFT models: Weibull AFT model and Log-Normal AFT model.

Weibull AFT model assumes that the event times follow a Weibull distribution. Survival function is given by:

$$S(\mathrm{t}) = e^{-(\lambda t)^k}, \tag{22}$$

where k is a scale parameter and λ is a shape parameter. Using the relationship $\sigma = \frac{1}{k}$, it can be shown that a Weibull-distributed random event time $T = e^{\sum_i^n \beta_i x_i + \sigma\varepsilon}$ corresponds to a survival function:

$$S(\mathrm{t}|x_i) = e^{-\lambda t^{\frac{1}{\sigma}}}, \tag{23}$$

where $\lambda = e^{-\frac{\sum_i^n \beta_i x_i}{\sigma}}$ is the reparameterization used to incorporate the explanatory variables (Dirick et al, 2017).

Log-Normal AFT model assumes that random event time, T, follows a Log-Normal distribution. The Log-Normal distribution is characterized by its logarithm being normally distributed. In this context, it implies that the logarithm of the survival times follows a normal distribution. This distribution is frequently selected when the underlying survival times exhibit positive skewness, indicating that events tend to happen later in time. This is particularly relevant in situations such as mortgage loan defaults. The survival function of Log-Normal AFT model is defined as:

$$S(\mathrm{t}|x_i) = 1 - \phi\left(\frac{\ln(T) - \sum_i^n \beta_i x_i}{\sigma}\right), \phi \sim std.Normal\ CDF. \tag{24}$$

2.3.2.3 Cox proportional hazards

Cox proportional hazards model (Cox, 1972) is another frequently employed technique in survival analysis. It is considered semi-parametric because it combines both parametric and non-parametric elements in its structure. The Cox PH model has a parametric component in the sense that it makes an assumption that the hazard rate for any individual is proportional to the hazard rates of other individuals, and this proportionality remains constant over time. In other words, the effect of predictor variables on survival is assumed to be constant over time. The model also incorporates a non-parametric aspect by making no assumption on baseline hazard function and estimating the effect of covariates on the hazard function without specifying a particular distribution for the survival times. The hazard function is given by:

$$h(t|x_i) = \lambda h_0(t) = e^{\sum_i^n \beta_i x_i} h_0(t), \tag{25}$$

and the survival function is

$$S(t|x_i) = e^{-e^{\sum_i^n \beta_i x_i H_0(t)}}, \tag{26}$$

where $H_0(t)$ is the cumulative baseline hazard function. To estimate baseline hazard and parameters Breslow's method is used. The conditional probability of the event given that at least one subject from risk set, denoted as R, also has an event at t can be obtained from Bayes rule and is given by

$$\frac{h_i(t)}{\sum_{k\in R} h_k(t)} = \frac{\lambda_i h_0(t)}{\sum_{k\in R} \lambda_k h_0(t)} = \frac{\lambda_i}{\sum_{k\in R} \lambda_k}. \tag{27}$$

Therefore, the total likelihood for all subjects is given by

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \frac{\lambda_i}{\sum_{k\in R} \lambda_k} = \prod_{i=1}^{n} \frac{e^{\beta x_i}}{\sum_{k\in R} e^{\beta x_k}}. \tag{28}$$

Consequently, log-likelihood is represented as

$$\ln\big(\mathcal{L}(\beta)\big) = \sum\left(\beta x_i - \ln\left(\sum_{k\in R} e^{\beta x_k}\right)\right). \tag{29}$$

Furthermore, in the context of the study, L_2 penalty with a coefficient of 10 is also introduced for the Cox PH model. The best estimate $\hat{\beta}$ used to estimate baseline hazard is thus determined by minimizing the regularized loss function which incorporates negative natural logarithm of the total likelihood and L_2 regularization. Consequently, the expression for the baseline hazard in the Cox PH model can be formulated as

$$h_0(t) = \sum_t \frac{y_i}{\sum_n e^{\hat{\beta} x_i}}. \tag{30}$$

The Cox PH model was first used in the credit context by Banasik et al (1999). For a more extensive discussion on Cox PH model, see Kleinbaum and Klein (2011). Important to note is that without making any assumptions Cox Proportional Hazards model cannot make predictions beyond the longest survival time observed in the training data unless certain assumptions are made. For observations with a duration exceeding the maximum duration observed in the training data (78 months), an assumption is therefore introduced that the survival probability for such cases is equal to the survival probability observed for these facilities at a duration of 78 months.

2.3.2.4 DeepHit

DeepHit (Lee et al, 2018) is a deep learning-based approach designed for survival analysis. The model employs a deep neural network to model the distribution of survival times directly. This neural network takes as input the covariates associated with each individual and learns a complex mapping between the input features and the distribution of survival times. DeepHit makes no assumptions about the underlying stochastic process and allows for the possibility that the relationship between covariates and risk changes over time. The method has been extended to handle competing risks, i.e. settings in which there is more

than one possible event of interest (Lee et al, 2018), but in the context of the study it is employed only for single event of default occurrence.

DeepHit is characterized by its discrete-time modeling approach, where in the context of the study time is divided into monthly intervals to align with the monthly granularity of the available data. DeepHit is obtained by parameterizing probability mass function (PMF) using a neural network. The probability of an event occurring at each time interval is determined by applying a softmax function to the output of the neural network:

$$f(t_j|x_i) = \frac{e^{\phi_j(x_i)}}{1 + \sum_i^n e^{\phi_i(x_i)}}, \tag{31}$$

where $\phi(x_i)$ represent a neural network that takes the covariates x_i as input and gives m outputs, each corresponding to a discrete time-point t_j, i.e., $\phi(x) = \{\phi_1(x), \ldots, \phi_m(x)\}$. The survival function of DeepHit is expressed as

$$S(t_j|x_i) = \frac{1}{1 + \sum_i^n e^{\phi_i(x_i)}}. \tag{32}$$

Consequently, the negative log-likelihood is given by

$$\mathcal{L}_1 = -\sum_{i=1}^{n} (y_i \, ln\left(\frac{e^{\phi_j(x_i)}}{1 + \sum_i^n e^{\phi_i(x_i)}}\right) - (1 - y_i) \, ln\left(\frac{1}{1 + \sum_i^n e^{\phi_i(x_i)}}\right)). \tag{33}$$

This is essentially the same negative log-likelihood as presented by Lee et al (2018), but with only one type of event (Kvamme and Borgan, 2019). Additionally, it is important to highlight that in contrast to the work by Lee et al (2018), the negative log-likelihood in (33) allows for survival beyond the maximum observed time t in the training data (Kvamme, B., Borgan Ø., 2019). Furthermore, DeepHit combines the log-likelihood with a ranking loss aimed at enhancing its discriminative capabilities (Kvamme et al, 2019):

$$\mathcal{L}_2 = \sum_{i,j} y_i I\{T_i < T_j\} e^{\frac{\hat{S}(T_i|x_i) - \hat{S}(T_i|x_j)}{\sigma}}. \tag{34}$$

To train DeepHit, we minimize a total loss function, denoted as $\mathcal{L}_{\text{total}}$, which is the sum of the two aforementioned losses. Mathematically,

$$\mathcal{L}_{total} = \alpha\mathcal{L}_1 + (1-\alpha)\mathcal{L}_2, \tag{35}$$

where α and σ from (34) are hyperparameters. For an in-depth exploration of DeepHit, please refer to Lee et al (2018).

In the context of the study, specific parameter values are employed, $\alpha = 0.4$ and $\sigma = 2.5$. The neural network used is a fully connected feedforward architecture, often referred to as a dense or multi-layer perceptron (MLP). In this type of network, every neuron in a given layer is connected to each neuron in the subsequent layer. The network comprises a single hidden layer containing 700 nodes. To prevent overfitting, a dropout rate of 0.25 is applied, randomly deactivating a quarter of the neurons during training. The initial learning rate is set at 0.0001, and the model undergoes training for 10 epochs, meaning the entire training dataset is processed forward and backward through the network for 10 complete iterations. The batch size used is 2900, which represents the number of data samples processed in each training iteration.

2.4 Evaluation metrics

The assessment of the models' performance considers its discriminatory power and goodness of fit. The analysis of discriminatory power is aimed at ensuring that the ranking of facilities that results from PD model estimates appropriately distinguishes between good and bad clients (i.e., those who do not default and those who default). The discriminatory power is assessed by calculating area under the curve (AUC). The analysis of goodness of fit is aimed at ensuring that a particular model adequately predicts the occurrence of defaults, i.e., that PD estimates are reliable predictors of default rates. To assess the accuracy of PD estimates Brier score (BS) and integrated calibration index (ICI) are calculated. Within the assessment, PD estimates are compared with observed default events over a one-year horizon. Moreover, the paper employs correlation analysis to

evaluate the predictive consistency of PD predictions over a one-year horizon obtained from different models, aiming to determine similarities and discrepancies between these forecasts. This evaluation is performed by calculating Spearman's rank correlation coefficient.

The AUC, an abbreviation for the area under the receiver operating characteristic (ROC) curve, is a common measure for assessing a model's capacity to distinguish between positive and negative instances across various probability thresholds (Bradley, 1997, Hanley and McNeil, 1982). The ROC curve visually depicts a model's performance across diverse classification thresholds, with the horizontal axis representing the false positive rate (FPR) and the vertical axis depicting the true positive rate (TPR), also known as sensitivity or recall. TPR quantifies the proportion of correctly predicted positive instances relative to the total actual positive instances, while FPR signifies the proportion of incorrectly predicted positive instances relative to the total actual negative instances. By plotting TPR against FPR for various threshold values, the ROC curve is generated, and the AUC represents the area beneath this curve. For a predictor f, the AUC can be expressed as:

$$AUC(f) = \frac{\sum_{a_i \in D^0} \sum_{a_j \in D^1} (I[f(a_i) < f(a_j)] + 0.5 \cdot I[f(a_i) = f(a_j)])}{|D^0| \cdot |D^1|}, \tag{36}$$

where D^0 is the subset of observations of facilities that have not defaulted, D^1 is the subset of observations of facilities that have not defaulted, $I[f(a_i) < f(a_j)]$ is an indicator function that returns 1 if the score of the not defaulted observation a_i is less than the score of the defaulted observation a_j, and 0 otherwise, $I[f(a_i) = f(a_j)]$ is an indicator function that returns 1 if the scores of a_i and a_j are equal, and 0 otherwise. The AUC value falls within the range of 0 to 1, where a higher AUC signifies superior model discriminatory power. An AUC of 0.5 signifies a random classifier, whereas an AUC of 1 signifies a perfect classifier.

The Brier score (Brier, 1950) is a metric used to evaluate the accuracy of probabilistic predictions made by a model. This is accomplished by calculating the mean squared

difference between predicted probabilities and the actual outcomes. The formulation of the Brier score, denoted as BS, is as follows:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2, \tag{37}$$

where n is the total number of observations, p_i is predicted probability assigned to a specific observation i by the model, o_i is the actual outcome for observation i (0 if no default occurred and 1 if default occurred). Lower Brier scores indicate better accuracy of predicted probabilities, while higher scores indicate poorer performance. A Brier score of 0 indicates perfect calibration, meaning that the predicted probabilities match the actual outcomes perfectly.

Similar to the Brier score, the integrated calibration index (ICI) is a metric used to assess the accuracy of a model's predictions. In the context of the study, to calculate the ICI, the predicted probabilities are sorted and divided into 10 equally sized intervals. For each interval, the average predicted probability and the observed average of the occurrence of default are calculated. The ICI is calculated as the sum of the absolute difference between the predicted and observed default probabilities in each interval (Austin and Steyerberg, 2019). Mathematically, it can be formulated as:

$$ICI = \frac{1}{n_{bins}}\sum_{1}^{n_{bins}}\left|P_{observed}(i) - P_{predicted}(i)\right|, n_{bins} = 10, \tag{38}$$

where n_{bins} is the number of bins used for calibration assessment, assumed to be equal to 10 in the context of the study, $P_{observed}(i)$ is the ratio of observed positive (defaulted) observations in bin i, $P_{predicted}(i)$ is the average predicted probability of the model in bin i. A lower ICI indicates better calibration, meaning that the model's predicted probabilities are well matched to actual outcomes.

Spearman's rank correlation coefficient (Spearman, 1904) is a nonparametric measure designed for assessing rank correlation between two variables. Specifically, it evaluates the strength and direction of the monotonic relationship between two sets of values, such as

PD predictions derived from two different models in the context of this study. This coefficient helps to understand how well the relationship between these forecasts can be described using a monotonic function, regardless of whether the relationship is linear or not. Unlike Pearson's correlation coefficient, which assesses linear relationships, Spearman's rank correlation focuses on the relationship between the ranks of the data points rather than the actual values of the variables. For a given sample of size n, each data point in each variable is assigned a rank, with the smallest value receiving the lowest rank and the largest value receiving the highest rank. In case of ties, the tied observations receive the same average rank. Spearman's rank correlation coefficient, denoted as ρ, is calculated as follows:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}, \tag{39}$$

where d_i is the difference between the ranks of corresponding data points in the two variables. The resulting ρ value falls within the range of -1 to 1. A positive ρ value near 1 indicates a strong monotonic positive relationship, suggesting that as one variable increases, the other tends to increase. A negative ρ value near -1 indicates a strong monotonic negative relationship, implying that as one variable increases, the other tends to decrease. When ρ value is close to 0, it implies a weak or no monotonic relationship between the two variables. Additionally, the p-value is used to assess the significance of the correlation. It represents the likelihood of obtaining Spearman correlation coefficients as extreme as the one calculated from the datasets if there was no actual correlation between the variables. The two-sided p-value is computed under the assumption of a t-distribution with two degrees of freedom, as outlined by Zar (1972). The t-statistic (t) used in this calculation is derived from the Spearman rank correlation coefficient (ρ) and the sample size (n) through the formula:

$$t = \rho \cdot \sqrt{\frac{n - 2}{(\rho + 1)(\rho - 1)}}. \tag{40}$$

This helps determine whether the observed correlation is statistically significant, given the null hypothesis that two variables, specifically the PD predictions of two different models, are uncorrelated.

In essence, Chapter 2 delves into the methodology of the study by introducing key definitions related to credit risk and regulatory aspects, along with a description of the data pre-processing techniques, the models applied and the metrics used to evaluate the results. The explained methodology ensures that the results of the study can be effectively and comprehensibly presented, thus establishing the foundation for subsequent chapters that delve into data analysis and empirical findings.

3 Data

The data represents a sample of Italian commercial bank's real-life mortgage portfolio granted to private individuals. This section provides an overview of the portfolio, including information on the size of the portfolio in terms of number of facilities, number of defaults and outstanding, as well as data preparation steps for modeling purposes, which involve splitting and pre-processing the data.

3.1 Portfolio overview

The Italian commercial bank's real-life mortgage portfolio sample spans from January 2013 to December 2021. It consists of 2 179 456 observations, including 50 892 distinct clients who started their credit history in January 2013 or later, with data available until December 2021. The dataset includes information such as the default indicator, collateral details, customer application characteristics, transaction information, and macroeconomic variables at monthly granularity. There are a total of 70 variables in the sample. A detailed description of all variables can be found in Appendix A.

The mortgage loans in this portfolio have long terms of up to 30 years. The effective maturity of these loans is influenced by factors like prepayments and amortization. While prepayments do impact the exposure's evolution over time, the focus of this paper is primarily on estimating the probability of default. Another crucial aspect to consider is loan seasoning, which refers to the known effect observed in retail mortgages where the riskiness of loans changes over time. The adequacy of capturing seasoning effects in the model is assessed through PD estimate testing.

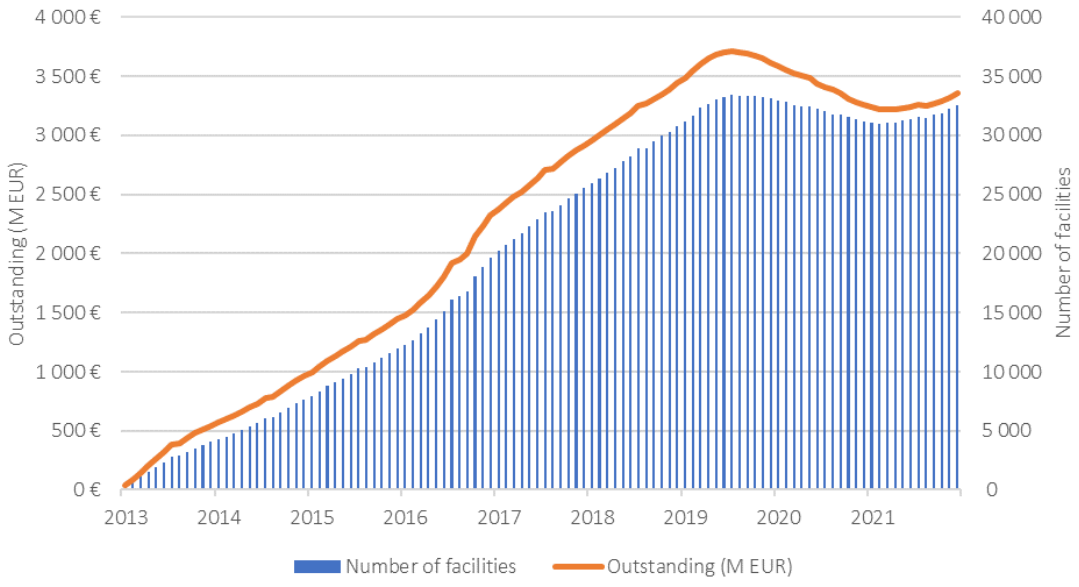
To determine the characteristics of the portfolio, the number of facilities, number of customers and outstanding amounts over time are calculated:

Table 1 Number of facilities, customers, and outstanding amount over time

Reporting date	Facilities		Customers		Outstanding	
	#	% change	#	% change	M EUR	% change
31-Dec-13	4 052	NA	4 052	NA	545.75 €	NA
31-Dec-14	7 681	89.56%	7 679	89.51%	969.46 €	77.64%
31-Dec-15	11 980	55.97%	11 977	55.97%	1 446.50 €	49.21%
31-Dec-16	19 684	64.31%	19 667	64.21%	2 321.84 €	60.51%
31-Dec-17	25 543	29.77%	25 493	29.62%	2 919.04 €	25.72%
31-Dec-18	30 812	20.63%	30 740	20.58%	3 446.18 €	18.06%
31-Dec-19	33 147	7.58%	33 070	7.58%	3 617.93 €	4.98%
31-Dec-20	31 222	-5.81%	31 137	-5.85%	3 260.84 €	-9.87%
31-Dec-21	32 557	4.28%	32 446	4.20%	3 354.84 €	2.88%

Source: own study

Figure 1 Evolution of facilities and outstanding amount over time



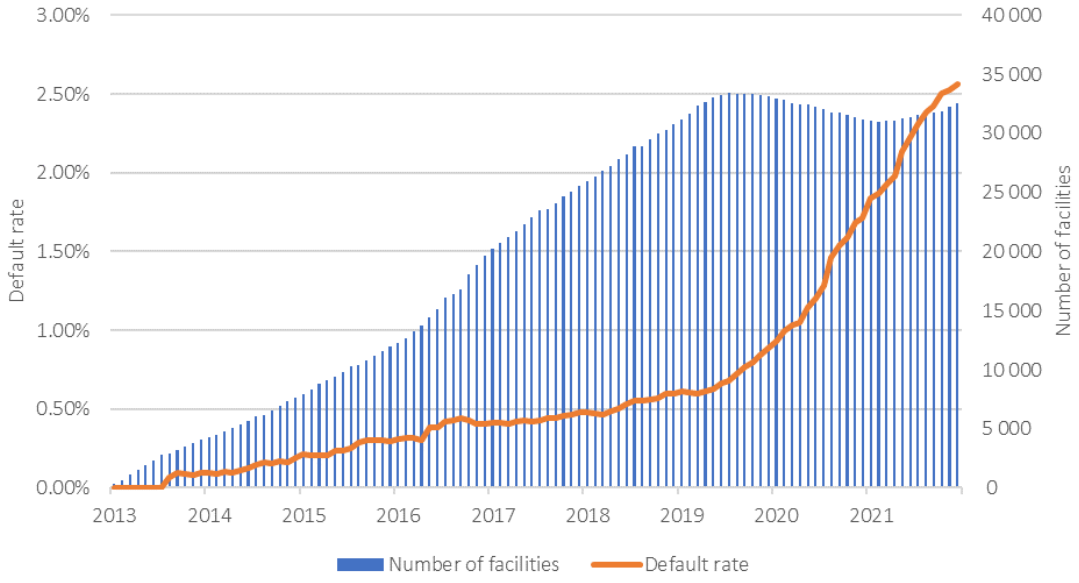
Source: own study

Based on the results presented in Table 1 and Figure 1, the portfolio experienced significant increase in the number of facilities and outstanding amount until 2019. This growth can be primarily attributed to the sample structure, which includes clients who initiated their credit history from January 2013 onwards. As a result, the introduction of new loans since 2013 led to a rise in the number of facilities, which consequently contributed to an increase in outstanding amount. Moreover, the growth of outstanding amount follows similar patterns as the growth of number of facilities.

However, in 2020, the portfolio size declined due to a ban on granting new mortgages in the Italian banking sector, which lasted from March 2019 to October 2020. Since new loans were not granted and some customers repaid their loans, this resulted in a decrease in the number of loans and the total outstanding amount.

Additionally, it can be stated that there is approximately 1.002 facility per customer, indicating that usually the customer has only one mortgage. The probability of default is modelled at the facility level, as default is applied at the facility level.

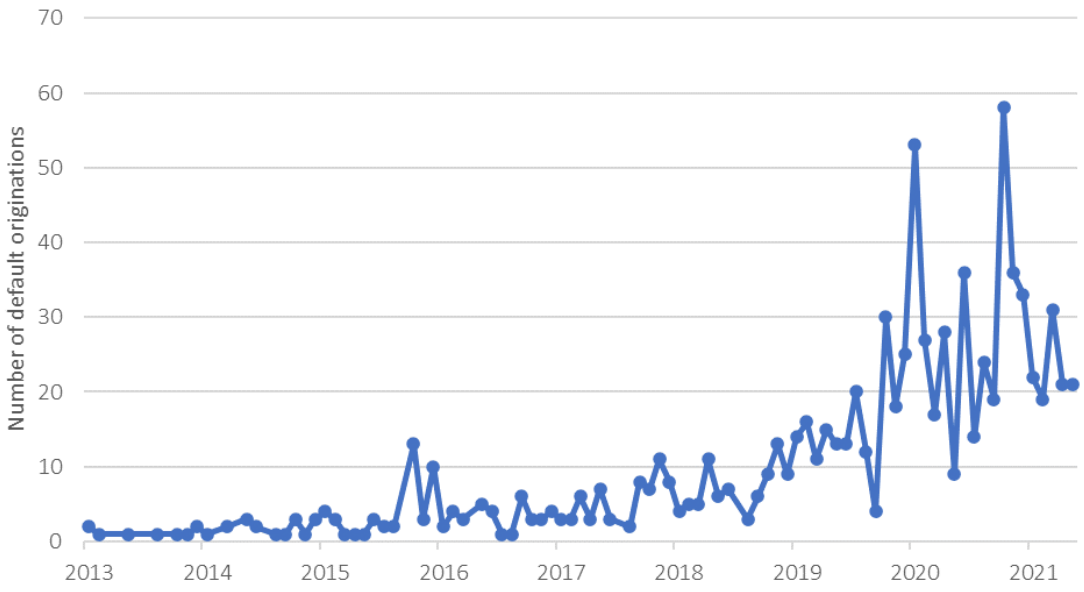
Figure 2 Default rate over time



Source: own study

Based on Figure 2, the default rate shows a gradual increase through 2019, followed by a significant increase in 2020 and 2021. This upward trend can be explained by the composition of the sample, which includes clients who initiated their credit history after January 2013 and are in the early stages of repayment during the first observed periods, potentially experiencing initial payment difficulties. As the portfolio matures, the default rate naturally increases. Additionally, it should be noted that the data assumes that the entire outstanding loan amount is considered a loss when a default event occurs, without taking into account any recoveries or partial losses. This assumption also implies that when a facility defaults, all subsequent observations on that facility are also considered to be in-default. The default rate gradually increased from 0.9% between 2013 and 2019 and then increased significantly in 2020 and 2021 reaching 2.56% by December 2021. To investigate the significant growth observed in 2020 and 2021, the number of defaults arising in those years is compared to the number in previous years.

Figure 3 Monthly granularity of number of default originations over time

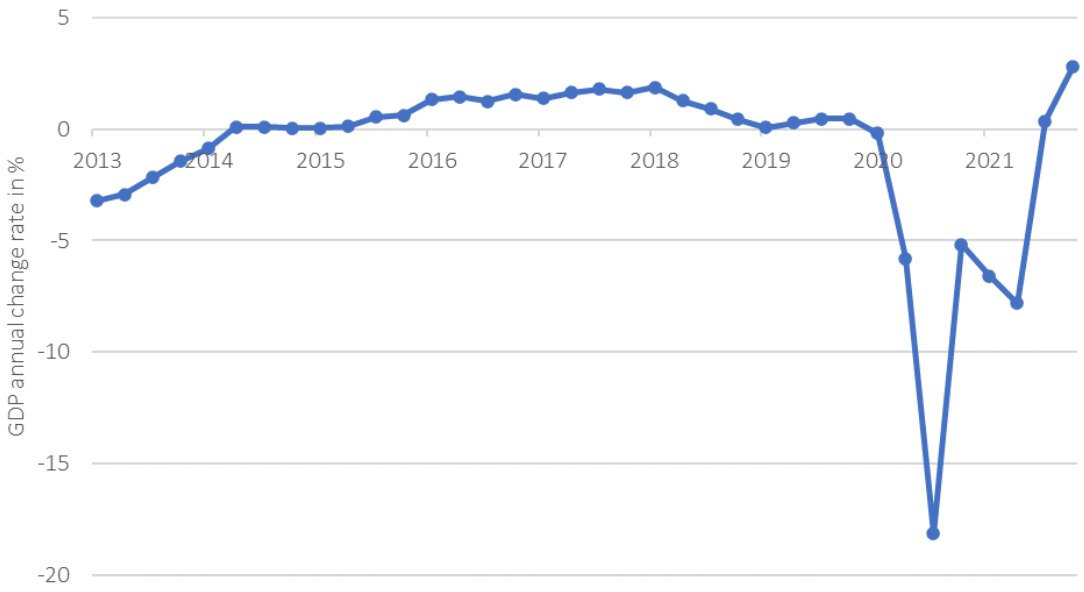


Source: own study

The upward trend of initiated defaults shown in Figure 3 is expected, given the larger customer base and the passing duration of loans. However, it is worth noting that the

number of default originations in 2020 and 2021 is significantly higher compared to 2019, despite a similar number of customers. It can therefore be stated that the sharp increase in the default rate in 2020 and 2021 is primarily due to the unusually high number of new defaults starting in these years. This increase in default originations is due to the occurrence of COVID-19 pandemic crisis, which is also reflected in unfavorable values of economic indicators in those years, including GDP in Italy.

Figure 4 Annual GDP change in Italy



Source: own study

As can be seen in Figure 4, Italy's GDP has a negative annual change from the first quarter of 2020 to the second quarter of 2021, a period when the number of default starts is the highest in the portfolio. It indicates a weak economic condition during that period, which contributes to the higher default rate during this time.

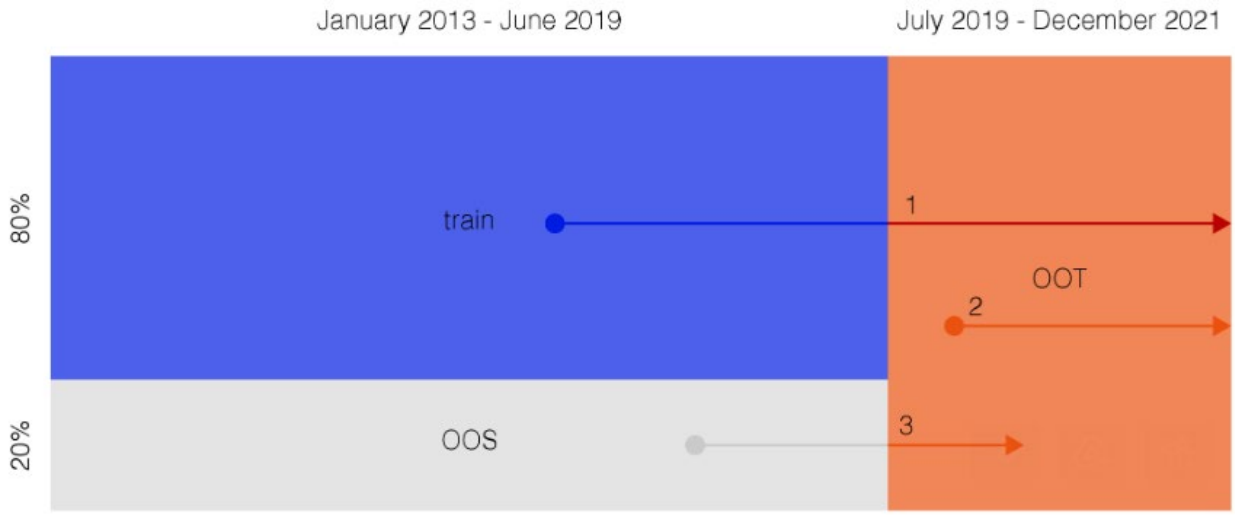
To further study the portfolio and model probability of default, it is necessary to split the sample into training and testing datasets and preprocess the data. These steps will be elaborated upon in the subsequent discussion.

3.2 Data preparation

3.2.1 Data splitting

For the period from January 2013 to June 2019, the data is divided into training and out-of-sample (OOS) samples at a ratio of 80:20 using an identifier-based grouping technique. This ensures that all observations with a specific facility identifier are exclusively assigned to either the training or test sample. Additionally, to evaluate the discriminatory power of a model as well as the accuracy of PD and ECL estimates an out-of-time (OOT) sample is created, consisting of observations from July 2019 to December 2021 for all clients who initiated their loans in July 2019 or later, as well as clients assigned to the OOS sample who still have a loan to repay later after June 2019. Observations of clients that started their history before June 2019 and is assigned to the train sample who still have a loan to repay later after June 2019 are excluded from the OOT sample.

Figure 5 Out-of-time sample definition



Source: own study

The splitting process, as shown in Figure 5, involves excluding observations after June 2019 for client marked as 1 from the OOT sample, including all observations for client 2, and including observations after June 2019 for client 3 in the OOT dataset. The training

sample has 971 758 observations, the OOS sample has 244 087 observations, and the OOT sample consists of 241 847 observations.

It is important to highlight that this study focuses on comparing two approaches for predicting the probability of default: survival models and logistic regression. Logistic regression is specifically designed to make predictions within a fixed time horizon, which, in this study, is set at a 12-month period. Consequently, a target variable in this case is whether the client defaults within this 12-month timeframe. By defining default over a one-year time horizon, it becomes difficult to use data on loans made in the last twelve months to predict risk parameters (Bonini and Caivano 2013). This challenge arises because the last year also includes information from the subsequent 12 months, which has not yet been observed. Therefore, logistic regression model is evaluated and compared to survival models based on data of the same time interval containing 133 405 observations from July 2019 to December 2020, rather than considering data up to December 2021.

The data splitting is performed prior to any imputation or encoding of character variables to prevent data leakage, ensuring that preprocessing steps are independently applied to each sample.

3.2.2 Data pre-processing

Multiple techniques already discussed in Section 2.2 are used to handle missing values, transform categorical variables and scale data. For certain variables, missing values are replaced with 0, as they indicate that there is no overdue amount. Other variables have a small percentage of missing values, with the highest being 2.2% for avg_duration_installment and 1% for fsi_20_class_imp, due to missing behavioral data for these specific variables. Rest of the variables have missing values of less than 1%. Missing values are interpolated using linear interpolation between the two nearest values. Categorical variables are transformed using Catboost encoder, a target-based categorical encoder that effectively captures the relationship between categorical variables and the target variable. To maintain consistency across features, the data is scaled using the min-

max method. This scaling technique brings all features to the same scale, eliminating potential biases arising from variations in ranges or units of the variables.

Based on the data overview provided, the data quality can be considered satisfactory. The expected trends are observed and the number of missing data is limited. Furthermore, the data is adequately prepared for modelling, ensuring that reliable results can be obtained.

4 Results

In this chapter, the results obtained from multiple models applied to a portfolio of Italian mortgages using real-life commercial bank data are presented. The models under consideration include logistic regression, Weibull AFT, Log-Normal AFT, Cox PH, and DeepHit described in Section 2.3. The default event information is used as a target variable, the duration of the loan is considered as a time variable, and the client identifier is used to link specific clients to their relevant data. The modelling involves multiple predictor variables, including collateral details, customer application characteristics, transactions, and macroeconomic factors. A total of 58 variables listed in Appendix A are used as predictor variables. The assessment of model performance considers three main aspects: the discriminatory power, the goodness of fit, and the impact on credit risk provisions. Furthermore, the paper employs correlation analysis to evaluate the predictive consistency of forecasts obtained from different models. The evaluation is conducted using the metrics described in Section 2.4.

4.1 Discriminatory power

The analysis of discriminatory power is aimed at ensuring that the ranking of facilities that results from PD model estimates appropriately distinguishes between good and bad clients (i.e., those who do not default and those who default). The discriminatory power is assessed by calculating area under the curve (AUC) at the portfolio level based on out-of-sample (OOS) and out-of-time (OOT) samples. PD estimates are compared to observed default events over a one-year horizon.

Table 2 AUC values based on OOS and OOT samples at the portfolio level

Model	AUC OOS	AUC OOT
Logistic regression	94.32%	87.71%
AFT Weibull	95.44%	89.87%
AFT Log-Normal	95.14%	89.90%
Cox PH	95.11%	89.79%
DeepHit	**96.05%**	**90.30%**

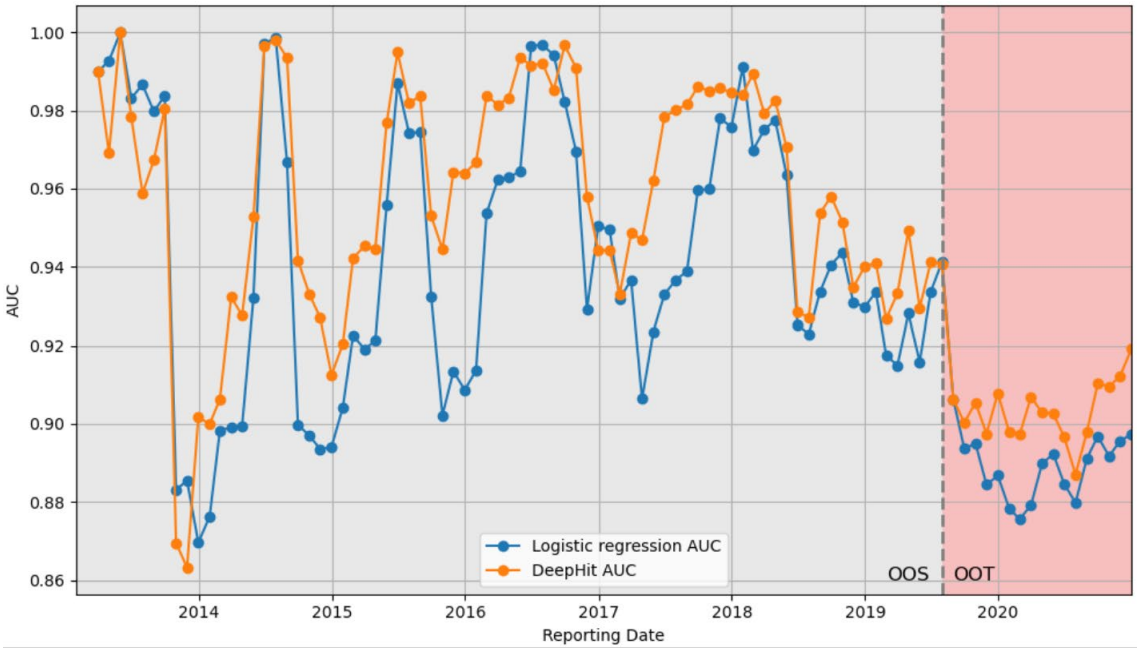
Source: own study

Both out-of-sample and out-of-time results indicate the outperformance of survival models over logistic regression in terms of discriminatory power at the portfolio level. This is reflected in higher AUC values obtained by each survival model for both samples, as given in Table 2. Notably, DeepHit model results achieve the highest AUC values for both OOS and OOT samples.

Moreover, considering the results of all models, there is a significant decrease in AUC values for the OOT sample compared to the OOS sample. This decrease can be attributed to the impact of COVID-19 pandemic crisis, which led to a higher number of defaults in the portfolio during this period, as indicated in Section 3., resulting in potentially unexpected defaults. Among the models, logistic regression results show the highest absolute drop of 6.61%, decreasing from 94.32% to 87.71%. On the other hand, the results of all survival models show an absolute decrease of less than 6%.

To further investigate this drop in discriminatory power, it is also assessed over time. Logistic regression results are compared with those of DeepHit model, which achieved the best discriminatory power at the portfolio level. The comparison is done by calculating AUC values for out-of-sample and out-of-time observations at monthly granularity.

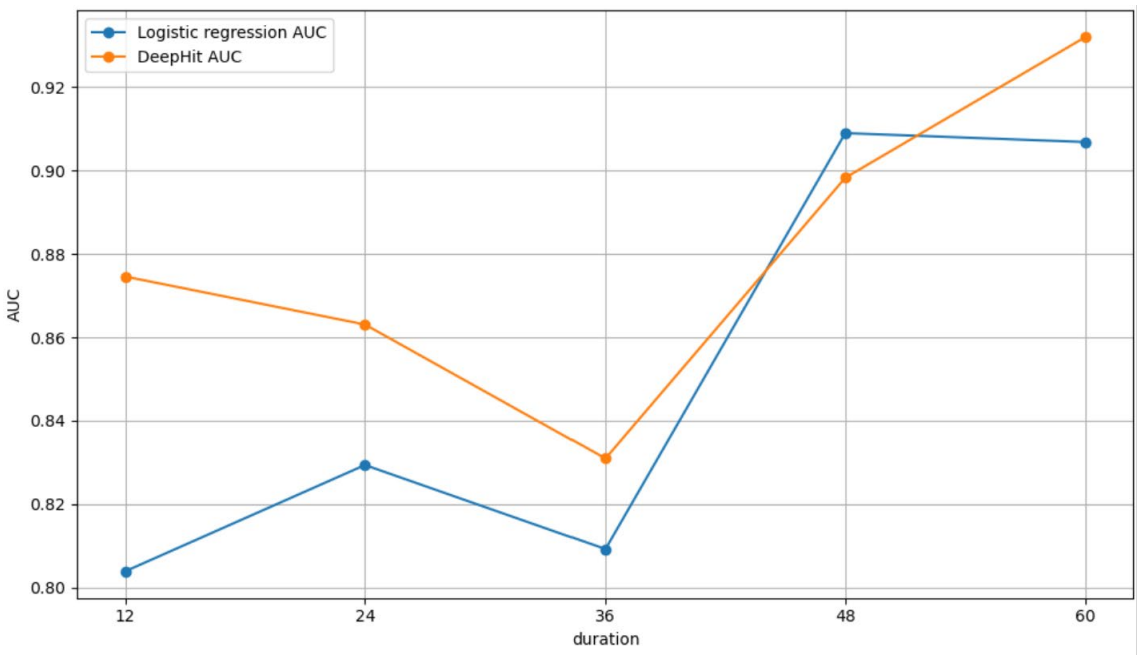
Figure 6 AUC values over time for logistic regression and DeepHit results based on OOS and OOT samples



Source: own study

Figure 6 clearly illustrates a substantial decline in the discriminatory power of logistic regression results for out-of-time observations, dropping from approximately 94% in the middle of 2019 to values below 90% afterward. Similarly, it reveals a noticeable decrease in the discriminatory power of DeepHit results, albeit with AUC values for out-of-time observations slightly higher than those of the logistic regression results. This similarity in the behavior of AUC values emphasizes that both models are experiencing reduced discriminatory power over time. To gain insight into the root causes of these declines, AUC values are calculated over the duration of clients in the portfolio based on the OOT sample.

Figure 7 AUC values over duration in the portfolio for logistic regression and DeepHit results based on the OOT sample



Source: own study

According to Figure 7, there is a significant discrepancy in the AUC values between new out-of-time observations of clients who have recently entered the portfolio and those of clients who have been in the portfolio for an extended period, for both logistic regression and DeepHit results. Specifically, the AUC values for new observations are notably lower compared to the observations of long-standing portfolio clients in the OOT sample. It can therefore be concluded that the drop in discriminatory power for both logistic regression and DeepHit results is primarily due to their relatively weaker ability to discriminate among new observations in the out-of-time sample. It further supports the possibility that this decrease may be linked to potentially unexpected default events stemming from the COVID-19 pandemic crisis that began to occur in the out-of-time sample.

For logistic regression results, the AUC for observations that have been in the portfolio for one year is slightly above 80%, while for those that have been in the portfolio for more

than 4 years, it exceeds 90%. As for DeepHit results, the AUC for observations that have been in the portfolio for one year is slightly below 88%, subsequently dropping to below 84% for those in the portfolio for 3 years, but then for those who have been in the portfolio for more than 4 years, it exceeds 90%. This means that the difference in AUC values between observations with different durations is much less significant for DeepHit compared to logistic regression results. It indicates that the discriminatory power of DeepHit estimates for new out-of-time observations performs better when compared to logistic regression. The DeepHit model exhibits greater consistency in discriminatory power across varying observation durations, making it more robust in distinguishing between good and bad customers in the out-of-time sample.

Furthermore, the discriminatory power is also assessed at the material sub-range level, specifically for two observation subgroups: those with arrears greater than 0 and those with no arrears. As some predictor variables include information about due payments, evaluating the models' performance in ranking clients both in arrears and without arrears is essential. The AUC values for out-of-time and out-of-sample samples are calculated for both logistic regression and DeepHit results.

Table 3 AUC values based on OOS and OOT samples at material sub-ranges level

Model	Bucket with arrears		Bucket without arrears	
	AUC OOS	AUC OOT	AUC OOS	AUC OOT
Logistic regression	**93.18%**	91.63%	78.44%	69.13%
DeepHit	91.12%	**92.05%**	**79.47%**	**72.98%**

Source: own study

The results presented in Table 3 reveals that on the sub-sample of observations without arrears, both models perform notably worse compared to the sub-sample of clients with arrears. However, it is worth highlighting that the PD estimates of the DeepHit model demonstrate a superior ability to distinguish between clients in the two sub-samples of out-of-time observations - those with and without due payments.

Based on the results, it can be concluded that survival models investigated outperform logistic regression in terms of discriminatory power, as all survival model results achieve higher AUC values for both OOS and OOT samples. This indicates that the results of survival models show a greater ability to distinguish between good and bad customers for out-of-sample and out-of-time observations. Furthermore, the highest AUC values are achieved by DeepHit estimates, which also show relatively good discriminatory power over time, and in the subgroup of clients without due payments compared to logistic regression results.

4.2 Goodness of fit

The analysis of goodness of fit is aimed at ensuring that a particular model adequately predicts the occurrence of defaults, i.e., that PD estimates are reliable predictors of default rates. To assess the accuracy of PD estimates Brier score (BS) and integrated calibration index (ICI) are calculated at the portfolio level based on OOS and OOT samples. PD estimates are compared to observed default events over a one-year horizon.

Table 4 BS and ICI values based on OOS and OOT samples at the portfolio level

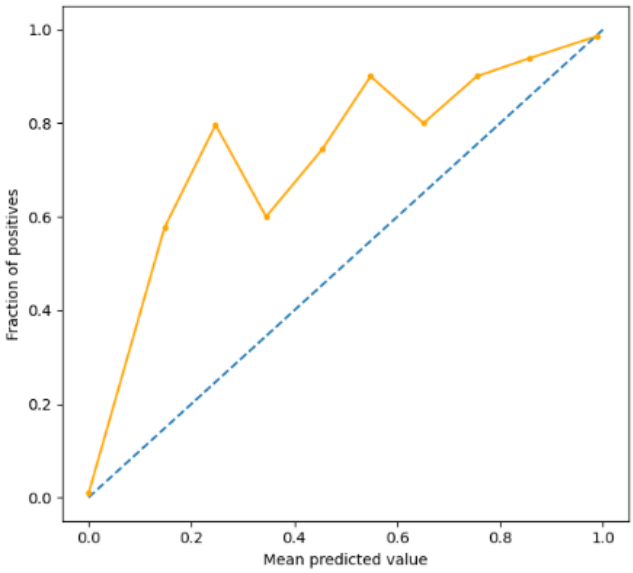
Model	OOS sample		OOT sample	
	BS	ICI	BS	ICI
Logistic regression	**0.0036**	**0.0012**	**0.0106**	0.0103
AFT Weibull	0.008	0.0051	0.0196	0.0105
AFT Log-Normal	0.007	0.0041	0.0178	0.0081
Cox PH	0.0083	0.0061	0.0219	0.0276
DeepHit	0.004	0.0014	0.109	**0.0074**

Source: own study

Logistic regression demonstrates good predictive accuracy on the OOS data, as evidenced by the lowest Brier score and integrated calibration index given in Table 4. However, its performance on the OOT sample exhibits a notable decrease in accuracy, as

indicated by the higher BS and ICI values. To further investigate the adequacy of logistic regression PD estimates on the OOT sample, a calibration plot is employed. This plot is created by dividing the predicted probabilities into 10 equal-width bins, ensuring that each bin covers the same range of predicted probabilities. For instance, if the first bin ranges from 0 to 0.1, subsequent intervals maintain the same width, so the next bin spans from 0.1 to 0.2, and so forth. Within each bin, the observed frequency of default occurrence and the average predicted probability are calculated. These two measures then serve as points to construct a calibration curve. The ideal calibration curve appears as a diagonal line, signifying well-calibrated predictions.

Figure 8 Calibration plot for logistic regression results based on the OOT sample



Source: own study

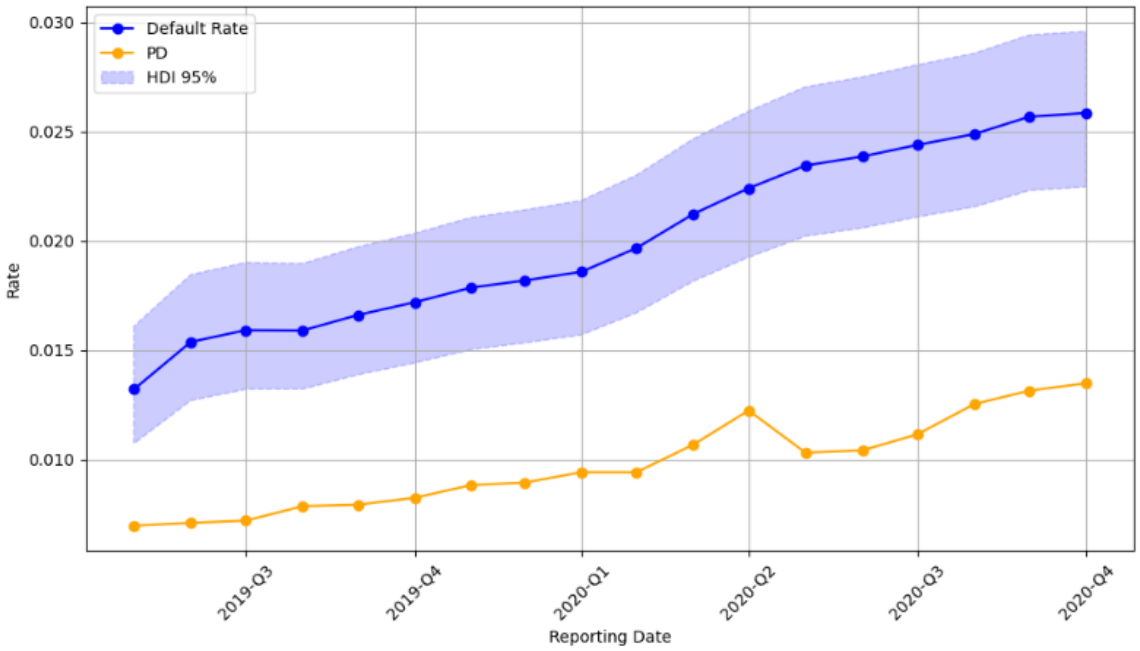
Based on Figure 8, the calibration curve lies above the diagonal line, indicating that PD estimates obtained by logistic regression are consistently underestimated for out-of-time observations at the portfolio level.

To evaluate the adequacy of these estimates over time, the average PD estimates per reporting date are compared with observed defaults using a binomial model with a Jeffreys prior (Brown et al 2007). Given the Jeffreys prior for the binomial proportion, the posterior

distribution is a beta distribution with shape parameters $a = D + \frac{1}{2}$ and $b = N - D + \frac{1}{2}$ (ECB, 2019). Here, N represents the number of customers for a specific reporting date, and D is the number of customers that have defaulted within that reporting date.

Traditionally, p-value obtained from the cumulative distribution function of the beta distribution serves as a measure of the adequacy of the estimated PD. This p-value indicates whether the estimated PD falls within the equal-tail interval of the posterior with a confidence level of $1 - p$. However, the equal-tail interval might not always be the most meaningful posterior interval to consider. The Highest Density Interval (HDI) offers a more accurate and reliable measure of the estimated PD's adequacy by ensuring that it includes the region with the highest density of the posterior distribution (Kruschke, 2015). In contrast, the equal-tail interval might exclude such regions. Given the significance of the HDI as a more meaningful quantity, the test is based on this interval. PD estimate can be considered accurate if it falls within 95% of the HDI of the posterior distribution (Kruschke, 2010).

Figure 9 Logistic regression PD estimates and observed defaults on the OOT sample

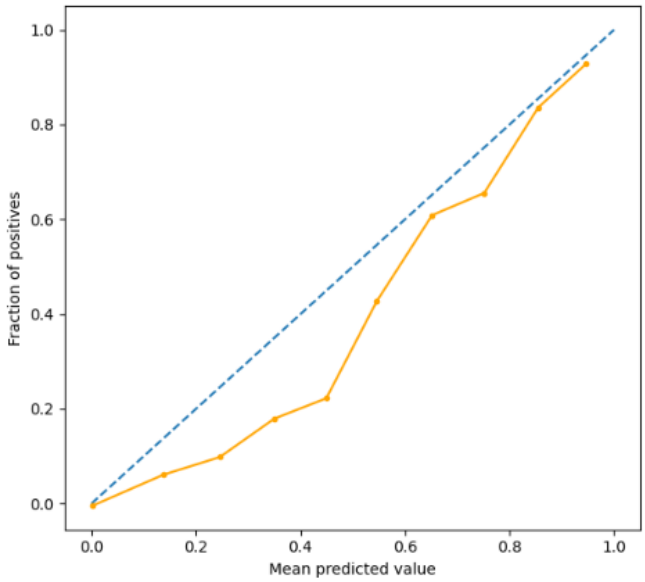


Source: own study

Based on the results presented in Figure 9, the average PD estimates obtained by logistic regression consistently fall below 95% HDI interval for all months in the OOT sample. Therefore, it can be concluded that PD estimates obtained from logistic regression are consistently and significantly underestimated over time for out-of-time period.

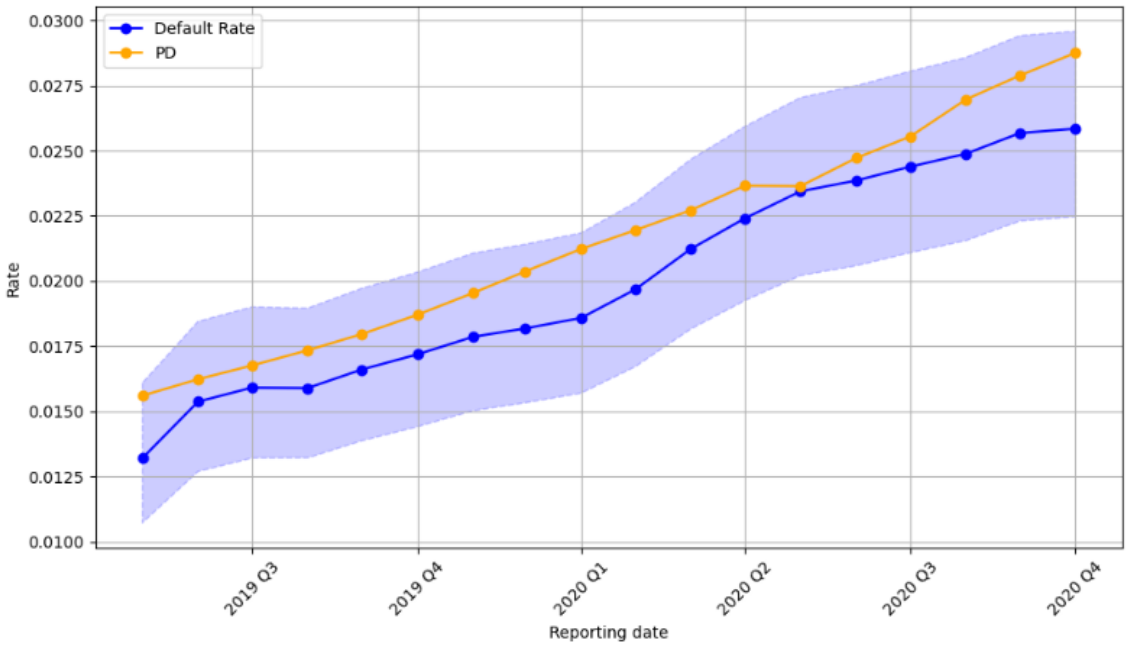
Among the models evaluated, DeepHit model results achieve the lowest integrated calibration index values at the portfolio level on OOT sample compared to other models, as shown in Table 4. To further evaluate the DeepHit results, a calibration plot and a comparison of PD estimates with observed defaults over time are presented in Figure 10 and Figure 11, respectively.

Figure 10 Calibration plot for DeepHit results on the OOT sample



Source: own study

Figure 11 DeepHit PD estimates and observed defaults on the OOT sample



Source: own study

Based on the calibration plot presented in Figure 10, the calibration curve for DeepHit results is significantly closer to a diagonal line than for logistic regression results, signifying that PD estimates of DeepHit model are more accurate at the portfolio level on out-of-time observations. Additionally, it can be observed that DeepHit PD estimates exhibit a slight overestimation at the portfolio level based on the OOT sample. Furthermore, Figure 11 demonstrates that the average PD estimates obtained by DeepHit consistently fall within the 95% HDI interval for each month in OOT sample. Therefore, it can be concluded that PD estimates of DeepHit model are accurate over time for the out-of-time period.

The results indicate that PD estimates obtained from logistic regression are consistently underestimated for the out-of-time sample, while PD estimates derived from DeepHit model are accurate on out-of-time observations. The accuracy of PD estimates directly influences the misestimation of expected credit loss when compared to observed loss.

4.3 Provisions impact assessment

The impact on credit risk provisions is evaluated by calculating expected credit loss and observed loss (OL). The expected loss determined over a specified time horizon, following the approach outlined in Section 2.1.3, is compared to the observed loss of the corresponding reporting date, which is equal to outstanding amount at default (OAD). It assumes that full outstanding amount is considered as the loss when a loan defaults, without any recovery or partial loss.

To assess the accuracy of the survival models and logistic regression estimates of expected credit loss, they are compared with observed losses over a one-year horizon in December 2020, corresponding to the value of observed loss in December 2021. The observed loss at this reporting date is 21.16 million euros.

Table 5 ECL estimates and OL difference over one-year horizon for December 2020

Model	ECL (M EUR)	Difference (M EUR)
Logistic regression	11.82 €	-9.34 €
AFT Weibull	8.32 €	-12.84 €
AFT Log-Normal	17.44 €	-3.73 €
Cox PH	36.53 €	15.36 €
DeepHit	**21.21 €**	**0.05 €**

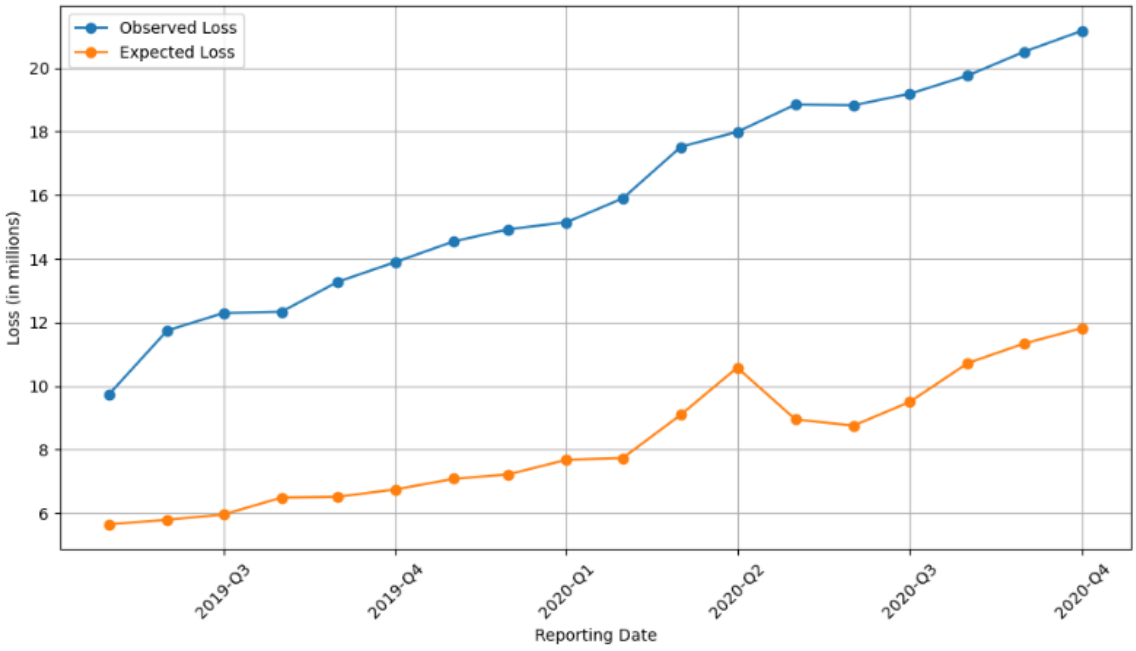
Source: own study

The results in Table 5 reveal that DeepHit model provides the most accurate one-year ECL estimate for December 2020, with the absolute difference between expected credit losses and observed losses being only 50 thousand euros. This negligible level of overestimation suggests that DeepHit estimates are highly reliable. In contrast, logistic regression ECL estimate is 11.82 million euros, leading to a substantial underestimation of over 9 million euros. This significant level of underestimation indicates that the logistic regression model does not adequately account for potential losses and can lead to

underprepared risk management strategies. Models that provide accurate or slightly conservative ECLs, such as DeepHit in this case, are more appropriate for risk management purposes and provisions calculation because they account for a wider range of potential losses, providing a more prudent and reliable risk assessment.

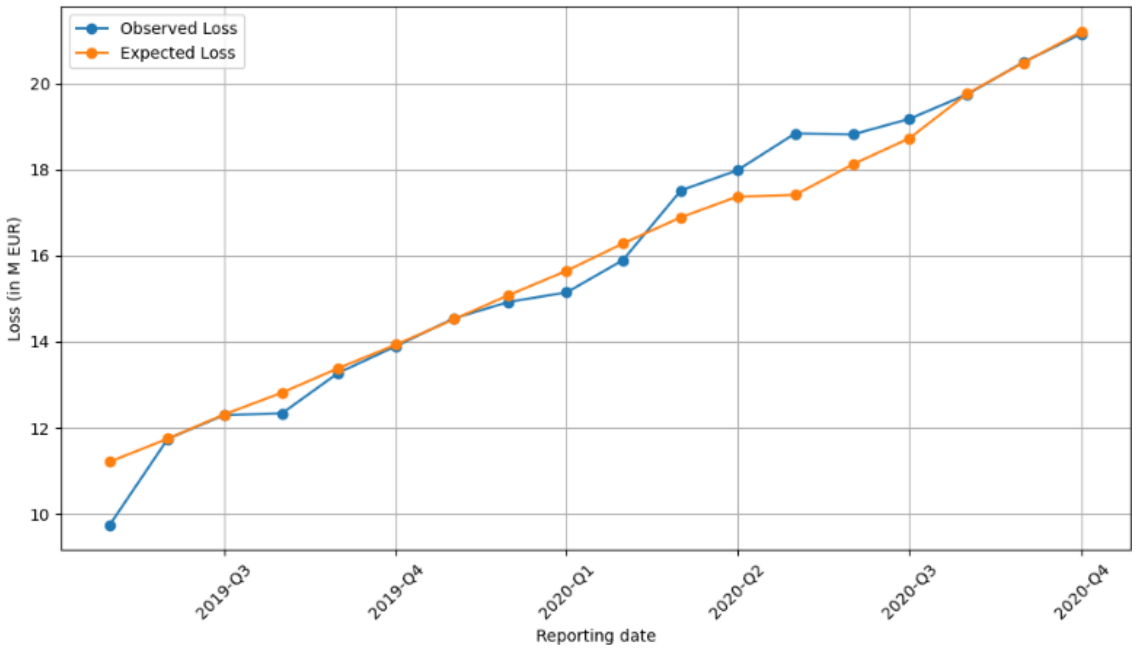
Additionally, the accuracy of expected loss estimates over one-year horizon obtained by logistic regression and DeepHit results is assessed over time using out-of-time observations.

Figure 12 Logistic regression ECL estimates and OL over one-year horizon on the OOT sample



Source: own study

Figure 13 DeepHit ECL estimates and OL over one-year horizon on the OOT sample



Source: own study

Based on the results presented in Figure 12 and Figure 13, ECL estimates derived from logistic regression consistently fall significantly below the observed loss over one-year horizon on out-of-time observations, while DeepHit ECL estimates are very close to the observed loss over time. Therefore, it can be concluded that expected credit loss estimates obtained from logistic regression are consistently underestimated over time for the entire out-of-time period, while DeepHit results provide accurate estimates of expected losses over one-year horizon on the OOT sample over time, which is consistent with the results of PD estimates accuracy assessment presented in Section 4.2. This suggests that DeepHit model is more accurate in its estimates and better predicts potential losses than logistic regression, making it a more reliable choice for assessing credit risk.

Furthermore, survival models offer a more dynamic approach to assessing credit risk. By considering the time dimension and the precise timing of default events, these models provide predictions of the probability of default over time, such as 12-month or lifetime PDs, as required by IFRS 9 (Xia et al, 2021). Financial institutions use these dynamic PD

estimates to adjust credit risk provisions and collection strategies throughout the loan repayment process, or to determine the loan term. Therefore, multi-year probability of default values are used to assess the accuracy of multi-year expected credit loss estimates. These estimates are then compared with observed losses over a time horizon of 0 to 8 years. This assessment is performed for each survival model using December observations from OOS and OOT samples.

Table 6 Relative difference between ECL and OL for Weibull AFT results

Time horizon/ reporting date	0	1	2	3	4	5	8	7	8	Average by date
13-Dec	83%	-21%	-58%	-62%	-61%	-65%	-76%	-87%	-90%	-49%
14-Dec	-6%	-39%	-42%	-37%	-43%	-59%	-77%	-83%		-48%
15-Dec	-37%	-30%	-20%	-24%	-45%	-68%	-76%			-43%
16-Dec	-29%	-5%	-3%	-26%	-56%	-66%				-31%
17-Dec	-9%	2%	-19%	-51%	-61%					-28%
18-Dec	-5%	-19%	-49%	-59%						-33%
19-Dec	-23%	-49%	-57%							-43%
20-Dec	-54%	-61%								-58%
21-Dec	-63%									-63%
Average by horizon	-16%	-28%	-35%	-43%	-53%	-65%	-76%	-85%	-90%	Overall -42%

Source: own study

Table 7 Relative difference between ECL and OL for Log-Normal AFT results

Time horizon/ reporting date	0	1	2	3	4	5	8	7	8	Average by date
13-Dec	-95%	-82%	-74%	-58%	-37%	-26%	-37%	-59%	-66%	-59%
14-Dec	-81%	-70%	-46%	-11%	10%	-2%	-35%	-44%		-35%
15-Dec	-72%	-47%	-5%	26%	19%	-18%	-28%			-18%
16-Dec	-52%	-10%	29%	32%	-3%	-10%				-2%
17-Dec	-18%	20%	27%	-2%	-6%					4%
18-Dec	13%	18%	-7%	-8%						4%
19-Dec	12%	-11%	-10%							-3%
20-Dec	-19%	-18%								-18%
21-Dec	-25%									-25%
Average by horizon	-37%	-25%	-12%	-4%	-3%	-14%	-33%	-52%	-66%	Overall -22%

Source: own study

Table 8 Relative difference between ECL and OL for Cox PH results

Time horizon/ reporting date	0	1	2	3	4	5	8	7	8	Average by date
13-Dec	-92%	-84%	-77%	-56%	-15%	45%	147%	33%	-13%	-12%
14-Dec	-85%	-75%	-48%	8%	86%	201%	123%	60%		34%
15-Dec	-78%	-52%	5%	89%	200%	142%	128%			62%
16-Dec	-58%	-4%	79%	187%	143%	149%				82%
17-Dec	-16%	59%	159%	124%	137%					93%
18-Dec	41%	130%	101%	116%						97%
19-Dec	105%	82%	97%							95%
20-Dec	59%	73%								66%
21-Dec	52%									52%
Average by horizon	-8%	16%	45%	78%	110%	134%	133%	47%	-13%	Overall 53%

Source: own study

Table 9 Relative difference between ECL and OL for DeepHit results

Time horizon/ reporting date	0	1	2	3	4	5	8	7	8	Average by date
13-Dec	-60%	-75%	-80%	-69%	-46%	-16%	5%	-20%	-28%	-43%
14-Dec	-73%	-75%	-60%	-28%	14%	42%	19%	14%		-18%
15-Dec	-76%	-60%	-27%	18%	48%	30%	37%			-5%
16-Dec	-64%	-31%	13%	45%	32%	47%				7%
17-Dec	-38%	3%	33%	23%	39%					12%
18-Dec	-8%	19%	10%	27%						12%
19-Dec	7%	0%	16%							8%
20-Dec	-14%	0%								-7%
21-Dec	-13%									-13%
Average by horizon	-38%	-27%	-14%	2%	17%	26%	20%	-3%	-28%	Overall -9%

Source: own study

Tables 6-9 present the relative differences, which are defined as the amount of difference between the expected loss and the observed loss divided by the observed loss, over different time horizons. The results indicate that the DeepHit model yields the lowest relative differences, with an overall average relative difference of -9%. This suggests that the DeepHit model provides the most accurate estimates overall. Additionally, when

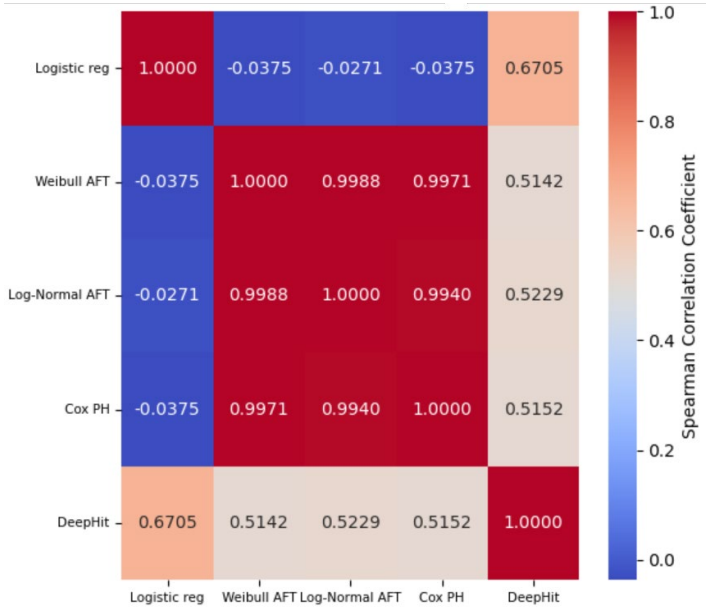
excluding the relative differences between ECL estimates obtained from December 2013 and OL over all horizons due to the limited number of 784 distinct facilities at this reporting date for the tested sample, the overall average relative difference for DeepHit results decreases to only -1%.

The results show that the differences between expected and observed loss for DeepHit estimates are negligible, indicating satisfactory performance. This makes DeepHit a reliable choice for assessing credit risk and calculating provisions. Following the evaluation of the model's performance, the relationships between various model results are investigated by conducting a correlation analysis for the PD predictions.

4.4 Correlation analysis of predictions

The analysis of correlation between predictions is conducted to assess the predictive consistency of PD estimates obtained from various models, aiming to determine similarities and discrepancies between the results. The correlation analysis involves the calculation of Spearman's rank coefficient between PD estimates over a one-year horizon derived from different model, considering out-of-time observations at the portfolio level.

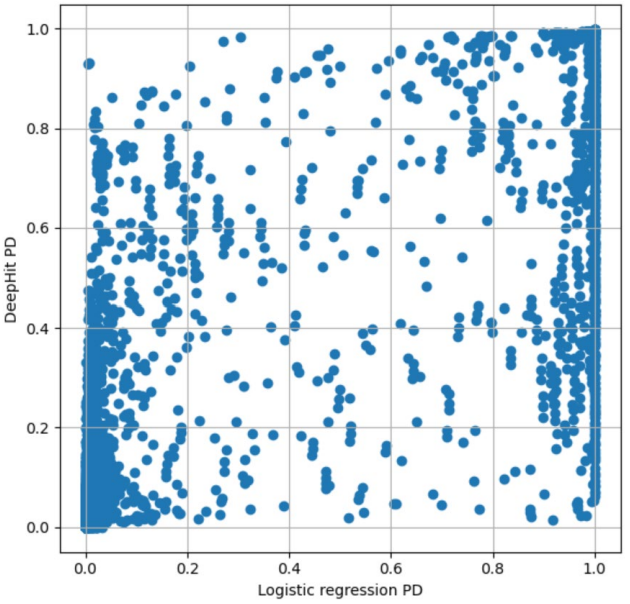
Figure 14 Correlation matrix between PD estimates on the OOT sample



Source: own study

The results, as depicted in Figure 14, demonstrate a significant and highly positive correlation between the predictions generated by the Log-Normal AFT, Weibull AFT, and Cox PH models. This strong positive correlation suggests a consistent and monotonic relationship, highlighting the alignment in their predictions. In contrast, the correlation values between these models and the PD predictions from the logistic regression model approach close to 0, signifying a weak correlation between them. Notably, the predictions of the DeepHit model show a moderate positive correlation with the predictions of all the aforementioned models, slightly above 50% when compared to the predictions of Log-Normal AFT, Weibull AFT, and Cox PH predictions, and slightly above 65% when compared to the predictions of logistic regression. This suggests that the predictions of the DeepHit model have the highest level of correlation with the logistic regression PD estimates.

Figure 15 Scatterplot between DeepHit and logistic regression PDs on the OOT sample



Source: own study

As shown in Figure 15, there are a limited number of observations where the DeepHit PD values are the highest while the logistic regression PD values are the lowest, as well as the reverse scenario where the DeepHit PD values are the lowest while the logistic

regression PD values are the highest. This suggests that there is a correlation between these two sets of PD estimates. It is also important to note that the calculated p-values associated with all the correlation coefficients are significantly below the 0.05 threshold. These low p-values indicate a rejection of the null hypothesis in favour of the alternative hypothesis, indicating that the predictions of all models are statistically significantly correlated with each other. In addition, it should be noted that the correlation results at the portfolio level are primarily influenced by the different strengths of the correlation between the PD predictions and the duration of the facility in the portfolio variable.

Table 10 Correlation between the PD predictions and duration variable on the OOT sample

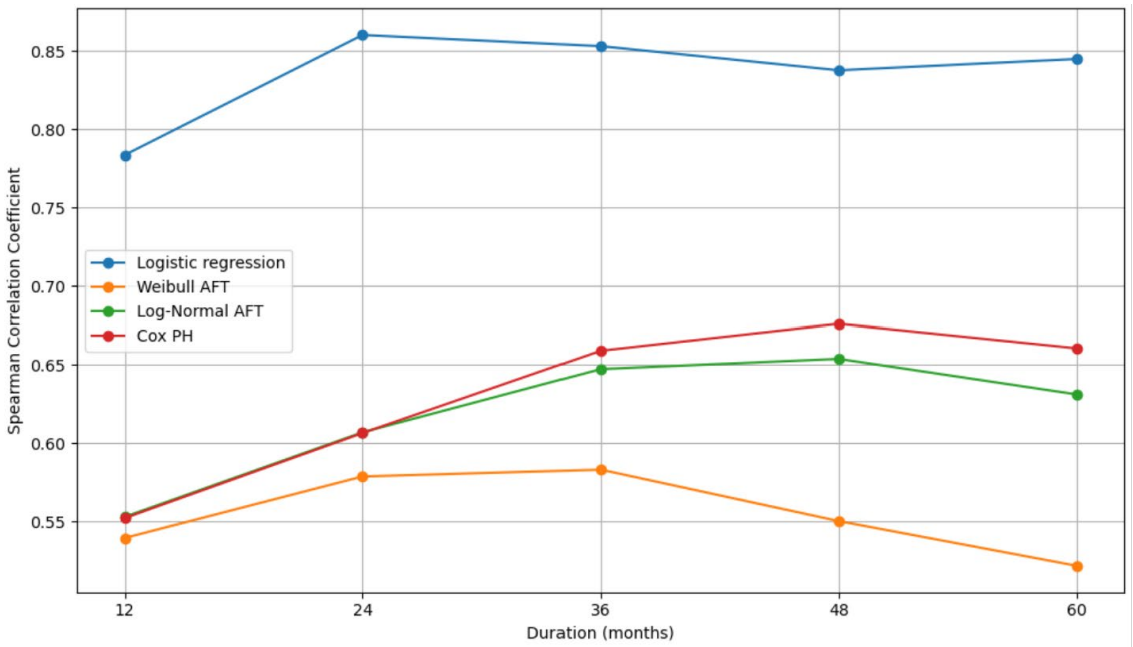
Model	Spearman's rank correlation coefficient
Logistic regression	-0.0595
AFT Weibull	0.9974
AFT Log-Normal	0.9934
Cox PH	0.9980
DeepHit	0.4979

Source: own study

According to the results presented in Table 10, the PD predictions obtained from the Log-Normal AFT, Weibull AFT and Cox PH models show a very strong positive correlation with the duration of the loan. This substantial correlation is in line with expectations, given that survival models incorporate the time component, which in this context is defined as the duration of the facility in the portfolio. In contrast, predictions from the DeepHit model show a moderate level of correlation, while predictions from logistic regression show a weak correlation with the facility duration variable. This discrepancy in correlation strength highlights the different modelling approaches. Logistic regression, designed as a model primarily suited to cross-sectional data, may struggle to capture the time-dependent patterns inherent in the data. DeepHit, on the other hand, incorporates the time component, defined as the duration of the loan, by using a neural network architecture to

generate predictions that have a more moderate but still relevant correlation with the facility duration variable. It is therefore important to also investigate the correlation coefficients at a more granular level of facility duration in the portfolio. As the predictions of the DeepHit model give the most comparable correlation results among the other models, the PD predictions of the DeepHit model are compared with the predictions of the other models for the OOT period.

Figure 16 Correlation between DeepHit PD estimates and PD estimates from other models over duration in the portfolio on the OOT sample



Source: own study

As can be seen in Figure 16, the results at the portfolio duration level tend to be higher on average than those obtained for the entire portfolio. This suggests that once the influence of results correlated with duration is mitigated, higher correlation results are obtained for the DeepHit model results. The strongest correlations between DeepHit PD estimates and estimates from other models are observed when comparing them to the results of the logistic regression model. Specifically, the Spearman's rank correlation coefficient between DeepHit and logistic regression is around 85% for facilities with a duration of 24 months or longer, and slightly below 80% for facilities with a duration of 12

months. When DeepHit predictions are compared with those obtained using other models, it can be seen that the Log-Normal AFT and Cox PH models produce similar correlation coefficients. These coefficients show a consistent trend similar to that observed in logistic regression, with higher correlation coefficients occurring at longer durations, reaching around 65% at durations greater than 36 months. The lowest correlation with DeepHit PD estimates comes from the predictions of the Weibull AFT model, hovering consistently around 55% across all durations.

The strong positive correlations observed between certain models indicate their consistency in predicting default probabilities. Conversely, models with weak correlations may benefit from further scrutiny, potentially employing ensemble methods to improve model performance. Additionally, the importance of considering specific variables in the assessment of prediction correlations is highlighted by recognizing the influence of the duration of the loan variable. The assessment of correlations at different duration levels helps to highlight areas of stronger or weaker consistency, further suggesting the potential of using ensemble methods to improve model performance. Less correlated forecasts, resulting from different types of model error, could potentially perform better when combined. It is also noteworthy that the DeepHit model predictions consistently shows a moderate correlation with the results of other models.

5 Conclusions

In this paper, the application of multiple survival analysis techniques in predicting the probability of default to estimate expected credit loss is assessed against the results of logistic regression based on a portfolio of Italian mortgages using real-life commercial bank data. The assessment of these models takes into account their discriminatory power and goodness of fit, by comparing their probability of default estimates with observed default events over a one-year horizon. The findings of the study reveal that survival models consistently outperform logistic regression in terms of discriminatory power, as evidenced by higher area under the curve values for both out-of-sample and out-of-time observations. These results indicate that the survival models exhibit a greater ability to distinguish between good and bad clients (i.e., those who do not default and those who default). Notably, the DeepHit model stands out as particularly effective, achieving the highest AUC values and demonstrating robust discriminatory power, especially over time and within the subset of clients with no overdue payments, as compared to logistic regression results. Furthermore, the findings reveal that logistic regression consistently provides underestimated PD estimates for out-of-time data, while the DeepHit model offers accurate PD estimates. This results in a consistent underestimation of ECL estimates using logistic regression results, leading to significant deviations from observed losses over a one-year horizon for out-of-time observations. In contrast, ECL estimates derived by the DeepHit model accurately predict losses over a one-year horizon for out-of-time data, accurately reflecting observed losses. These results suggest that the DeepHit model is a reliable choice for assessing credit risk and predicting potential losses, not only improving the discriminatory power and accuracy of ECL estimates compared to logistic regression, but also enabling more dynamic estimates due to its survival approach features, which are consistent with the criteria of IFRS 9 requirements. Other survival models, compared to logistic regression, do not perform particularly better in terms of calibration accuracy than logistic regression, and their PD and ECL estimates also deviate from observed default and loss values.

Starting from these findings, it would be interesting to further explore the application of deep learning-based survival analysis models, such as DeepHit, and evaluate their performance on a wider range of data sets and banking portfolios. This could include assessing their robustness in different economic climates or regions to determine their generalizability. Additionally, investigating the interpretability of these deep survival models and understanding the factors influencing their predictions could provide valuable information for risk management practitioners and regulators. Furthermore, the analysis highlights the potential benefits of using ensemble methods, as the models produce PD estimates with different levels of correlation strength. Models with strong positive correlations demonstrate consistency in predicting defaults, while those with weaker correlations may benefit from further investigation, potentially incorporating ensemble methods to improve model performance. Predictions with lower correlations, arising from different types of model error, have the potential to achieve improved performance when combined.

References

Aptivaa, 2016. Building blocks of impairment modeling.

Austin, P. C., Steyerberg, E. W., 2019. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in medicine, 38(21), 4051-4065.

Banasik J., Crook J., Thomas L., 1999. Not if but when will borrowers default. The Journal of the Operational Research Society 50(12): 1185–1190.

Bank for International Settlements, 2017. IFRS 9 and expected loss provisioning - Executive Summary.

Basel Committee on Banking Supervision, 2000. Principles for the management of credit risk. 1–26.

Basel Committee of Banking Supervision, 2009. Guiding Principles for the Replacement of IAS 39.

Basel Committee on Banking Supervision, 2015. Guidance on accounting for expected credit losses.

Beerbaum, D., 2015. Significant increase in credit risk according to IFRS 9: Implications for financial institutions. International Journal of Economics and Management Sciences, 4(9), 1-3.

Bellotti, T. and Crook, J., 2009. Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society, 60(12), pp.1699-1707.

Berkson, J., 1944. Application of the logistic function to bio-assay. Journal of the American statistical association, 39(227), 357-365.

Bonini, S., Caivano, G., 2013. The survival analysis approach in Basel ii credit risk management: modeling danger rates in the loss given default parameter. Journal of Credit Risk Volume 9 (1), 101–118.

Bradley, A. P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

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

Brown, L. D., Cai, T. T., & DasGupta, A, 2001. Interval estimation for a binomial proportion. Statistical science, 16(2), 101-133.

Collett D., 2003 Modelling Survival Data in Medical Research. 2. CRC Press; Boca Raton.

Cox, D.R., 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B, 34, 187-202.

Directive 2006/48/EC, 2006. Directive relating to the taking up and pursuit of the business of credit institutions.

Dirick, L., Claeskens, G. and Baesens, B., 2017. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 68, pp.652-665.

EBA, 2016. Guidelines on the Application of the Definition of Default under Article 178 of Regulation (EU) No 575/2013. EBA/GL/2016/07.

EBA, 2017. Guidelines on PD Estimation, LGD Estimation and the Treatment of Defaulted Exposures. EBA/GL/2017/16.

ECB, 2019. Instructions for reporting the validation results of internal models IRB Pillar I models for credit risk.

Financial Crisis Advisory Group, 2009. Report of the Financial Crisis Advisory Group.

Financial Stability Forum, 2009. Report of the Financial Stability Forum on Addressing Procyclicality in the Financial System.

Hanley, J. A., McNeil, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

IASB, 2001. IAS 39 Financial Instruments: Recognition and Measurement,

IASB, 2014. IRFS9 Financial Instruments: Project summary.

Kleinbaum D. G., Klein M., 2010. Logistic Regression, A Self-Learning Text, Springer, 3rd Edition.

Kleinbaum D. G., Klein M., 2011. Survival Analysis: A Self-Learning Text, Third Edition. Statistics for Biology and Health. Springer.

Kruschke, J., 2010. Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Cambridge, MA: Academic Press.

Kruschke, J., 2015. Doing Bayesian Data Analysis: A Tutorial Introduction with R, JAGS and Stan, 2nd edn. Academic Press, London.

Kvamme, B., Borgan Ø., 2019. Continuous and Discrete-Time Survival Prediction with Neural Networks.

Kvamme, H., Borgan, Ø., Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression.

Louzada-Neto, F., 2006. Lifetime modeling for credit scoring: A new alternative to traditional modeling via survival analysis. Tecnologia de Cr´edito (Serasa) 56, 8–22.

Malik, M., Thomas, L., 2007. Modeling Credit Risk of Portfolio of Consumer Loans. School of Management at the University of Southampton.

Narain, B., 1992. Survival Analysis and the Credit Granting Decision. In: Thomas, L.C., Crook, J.N. and Edelman, D.B., Eds., Credit Scoring and Credit Control, OUP, Oxford, 109-121.

OpenAI, 2023. ChatGPT-3.5.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., V., Gulin, A., 2019. CatBoost: unbiased boosting with categorical features.

Regulation (EU) No 575/2013, 2013. Regulation on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012.

Schutte, W. D., Verster, T., Doody, D., Raubenheimer, H., Coetzee, P. J., 2020. A proposed benchmark model using a modularised approach to calculate IFRS 9 expected credit loss.

Spearman, C., 1904. The proof and measurement of association between two things.

Tong, E. N., Mues, C., Thomas, L. C., 2012. Mixture cure models in credit scoring: If and when borrowers default. European Journal of Operational Research 218 (1), 132–139.

Xia, Y., He, L., Li, Y., Fu, Y. and Xu, Y., 2021. A dynamic credit scoring model based on survival gradient boosting decision tree approach. Technological and Economic Development of Economy, 27(1), pp.96-119.

Zar, J. H., 1972. Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association, 67(339), 578-580.

List of tables

List of figures

Appendix A: List of variables

Name of the variable	Meaning of the variable	Description of the variable
cid	Instrument identifier	An identifier applied by the institution to uniquely identify each instrument under a contract. Each instrument must have one instrument identifier. This value will not change over time and cannot be used as the instrument identifier for any other instrument under the same contract.
reporting_date	reporting date	The reporting reference date of the observation, i.e., the last day of the reporting month.
default_adj	Default indicator	Indicator of default event.
intodefault_12	In-default indicator (12 months following)	Indicator for exposures that have the defaulted status within the next 12 months following the reporting reference date.
default_start_date	Default related information	Date of the default of the instrument, as used for the purpose of estimation of PD.
default_exit_date	Default related information	Date of the end of default of the instrument.
first_def_date	Default related information	First date of the default.
duration	Duration	Duration of client in the portfolio.
min_rep_date	Duration related information	Minimum reporting date of a facility.
mx_rep_Date	Duration related information	Maximum reporting date of a facility.
Outstanding	Outstanding value	Outstanding principal amount at reporting date.
Arrear_amount_adj	Sub-range determination	Past due amount at reporting date.
absolute_breach_past_due_amnt_o	Predictor variable	The arrear amount on the absolute breach start date.
Arrear_Nr_Instalments_adj	Predictor variable	Number of instalments unpaid at reporting date.
average_overdue_12m_imp	Predictor variable	Average overdue amount in the last 12 months, calculated by

		only considering the months in which an overdue payment was recorded.
avg_amt_used_last_12m_imp	Predictor variable	Average of the sum of the amounts used in relation to the customer in the last 12 months.
avg_arrear_amount_12	Predictor variable	Average value of past due amount at reporting date over a time horizon of 12 months.
avg_CA_tot_eop_balance_3	Predictor variable	Average value of sum of all ends of period balances of all current accounts at reporting date over a time horizon of 3 months.
avg_duration_installment	Predictor variable	Average duration of existing installment credit lines. Calculated starting from the original durations of the individual credit lines.
avg_SA_tot_eop_balance_ca_3	Predictor variable	Average value of sum of all ends of period balances of all saving accounts at reporting date over a time horizon of 3 months.
CA_num_cca	Predictor variable	Number of current accounts open at reporting date.
CA_tot_eop_balance	Predictor variable	Sum of all ends of period balances of all current accounts at reporting date.
channel_2_adj	Predictor variable	Detailed channel of distribution.
Cover_value_adj	Predictor variable	The value of collateral at reporting date (after adjustments).
diff_abs_breach_end_rep_date	Predictor variable	Difference in months between the absolute breach end date (which is the most recent date when the sum of the total amounts past due on any credit obligation of the facility or obligor fell below or was equal to the absolute threshold) and reporting date.
diff_abs_breach_str_rep_date	Predictor variable	Difference in months between the absolute breach start date (which is the most recent date when the sum of the total amounts past due on any credit obligation of the facility or obligor

		fell below or was equal to the absolute threshold) and reporting date.
diff_rel_breach_end_rep_date	Predictor variable	Difference in months between the relative breach end date (which is the most recent date when the sum of the total amounts past due on any credit obligation of the facility or obligor fell below or was equal to the relative threshold. This field can only be provided if the relative breach start date is provided) and reporting date.
dsr_bik_0	Predictor variable	Sum of the obligations of the subject on the processing date divided by monthly sum of salary for all borrowers and guarantors.
dsr_bik_12	Predictor variable	Sum of the obligations of the subject 14 months before the processing date divided by monthly sum of salary for all borrowers and guarantors.
fsi_20_class_imp	Predictor variable	Has a value from 1 (serious financial stress) to 14 (minor financial stress) (after adjustments).
IMPORTOMUTUO_adj	Predictor variable	Amount disbursed (after adjustments).
Instalment_amount_adj	Predictor variable	Next month instalment amount (after adjustments)
ltv_adj	Predictor variable	Original Loan to Value calculated as outstanding divided by collateral value at the moment of the application.
max_arrear_amount_12	Predictor variable	Max value of past due amount at reporting date over a time horizon of 12 months.
max_arrear_nr_instalments_12	Predictor variable	Max value of # instalments unpaid at reporting date over a time horizon of 12 months.
max_average_overdue_12m_imp	Predictor variable	Maximum for the first and second borrower of average overdue amount in the last 12 months, calculated by only

		considering the months in which an overdue payment was recorded.
max_avg_amt_used_last_12m_imp	Predictor variable	Maximum for the first and second borrower of average of the sum of the amounts used in relation to the customer in the last 14 months.
max_max_n_insolvenc_cr_lines_imp	Predictor variable	Maximum for the first and second borrower of maximum number of missing payments in the last 26 months (Installment, Cards).
max_n_insolvencies_cr_lines_imp	Predictor variable	Maximum number of missing payments in the last 24 months (Installment, Cards).
max_SA_num_ca_3	Predictor variable	Max value of number of saving accounts open at reporting date over a time horizon of 3 months
max_Total_Obligations_T0_imp	Predictor variable	Maximum for the first and second borrower of sum of the obligations of the subject on the processing date.
min_arrear_amount_12	Predictor variable	Min value of past due amount at reporting date over a time horizon of 12 months
min_arrear_nr_instalments_12	Predictor variable	Min value of # instalments unpaid at reporting date over a time horizon of 12 months
min_CA_tot_eop_balance_3	Predictor variable	Min value of sum of all ends of period balances of all current accounts at reporting date over a time horizon of 3 months
n_active_installment_imp	Predictor variable	Total number of existing installment credit lines.
n_active_non_rev_cards_imp	Predictor variable	Total number of existing Charge Card credit lines.
n_active_personal_loans_imp	Predictor variable	Total number of existing personal loans.
n_reporting_fis_active_cr_li_imp	Predictor variable	Number of reporting institutions for credit lines with phase AC and role R or C (Borrower or Coborrower).
n_reporting_fix_CH_imp	Predictor variable	Total number of reporting institutions for credit lines with

		role R or C (Borrower or Coborrower).
NUM_CUST	Predictor variable	Number of customers (borrowers and guarantors) assign to credit at reporting date.
OCCUP_PROF_FIRST_FINAL_2	Predictor variable	Occupation of main borrower.
overdue_amt_active_installme_imp	Predictor variable	Total overdue amount relating to existing installment credit lines.
period_fix_interest	Predictor variable	Fixed interest period.
region_adj	Predictor variable	Region of the collateral.
rel_breach_pst_due_amt_o	Predictor variable	The Relative Breach Past Due Amount is the sum of the total amounts past due on any credit obligation of the limit or customer (including principal, interest, and fees) on the given Relative Breach Start Date. This field can only be provided if the relative breach start date is provided.
Remain_loan_period	Predictor variable	Difference in months between reporting date and maturity date.
SA_max_eop_balance_ca	Predictor variable	maximum balance at end of period (reporting date) between all the saving accounts.
SA_out	Predictor variable	Sum of all ends of period balances of all saving accounts at reporting date divided by outstanding principal amount at reporting date.
SA_tot_eop_balance_ca	Predictor variable	sum of all ends of period balances of all saving accounts at reporting date.
sum_instal_reddito	Predictor variable	Instalment value divided by monthly sum of income of first and second borrower and first and second guarantor.
sum_outst_reddito	Predictor variable	Outstanding value divided by annual sum of income of first and second borrower and first and second guarantor.
TIPO_CONTRATTO_first_final	Predictor variable	Type of contract of first borrower.
TIPOMUTUO_adj	Predictor variable	type of mortgage.

Tot_borrower_coborrower_exp_imp	Predictor variable	Overall exposure from credit lines for which the subject has the role of borrower or co-borrower (direct risk).
Tot_borrower_coborrower_over_imp	Predictor variable	Overall overdue for credit lines for which the subject has the role of borrower or co-borrower (direct risk).
total_due_credit_imp	Predictor variable	Total monthly amount deriving from the due installments of all active installment credit lines.
Total_Obligations_T0_imp	Predictor variable	Sum of the obligations of the subject on the processing date
GDP	Predictor variable	Gross domestic product annual change in Italy
RHP	Predictor variable	Residual house price index in Italy
UR	Predictor variable	Unemployment rate in Italy

Abstract

The assessment and effective management of credit risk is of crucial importance within the banking industry, given that default events, i.e. failures of clients to meet their contractual obligations, pose a significant threat to financial stability. The study evaluates the application of multiple survival analysis techniques in the context of credit risk, including classical survival analysis techniques, in particular accelerated failure time and Cox proportional hazards models, as well as DeepHit, a deep learning model adapted to survival analysis. The evaluation involves predicting the probability of default (PD) to estimate expected credit losses (ECL), comparing the performance of these survival models with the results of logistic regression based on real-life data of a commercial bank's Italian mortgage portfolio. The assessment of the models takes into account two critical aspects: their discriminatory power and goodness of fit, by comparing their probability of default estimates with observed default events over a one-year horizon. It is found that survival models outperform logistic regression in terms of discriminatory power. Notably, the DeepHit model stands out as highly effective, achieving the highest AUC values and demonstrating robust discriminatory power. Furthermore, the study reveals that logistic regression results consistently underestimate PD estimates for out-of-time data, while the DeepHit model provides accurate PD estimates. As a result, this leads to more accurate ECL estimates when employing the DeepHit model. These findings indicate that the DeepHit model is a reliable choice for assessing credit risk and predicting potential losses, not only improving the discriminatory power and the accuracy of ECL estimates compared to logistic regression, but also allowing for more dynamic estimates due to its survival approach features, which are in line with the criteria of IFRS 9 requirements for measuring financial instruments and recognizing impairments.