ARCADA

An Analytics Process for Forecasting Expected Credit Losses for the Lifetime of Loans

Auto Loan Portfolios

Alexandros William Touvras

Degree Thesis

Big Data Analytics

2024

Degree Thesis

Alexandros William Touvras

An Analytics Process for Forecasting Expected Credit Losses for the Lifetime of Loans Auto Loan Portfolios

Arcada University of Applied Sciences: Big Data Analytics, 2024

Commissioned by:

Santander

Abstract:

The main goal of this study is to conduct a thorough assessment of expected credit risk losses from loan origination throughout the lifetime of loans by utilizing an integrated methodology that combines domain knowledge with advanced analytics. The data used, are extracted from Santander loan applications, consisting of auto loan data from Finnish consumers and companies and Danish borrowers. Following a correlation-based feature selection process, XGBoost is applied as a challenger model to logistic regression to forecast the early performance of loans. A second model that incorporates the early performance as input is trained to predict the expected losses at loan maturity. This nested method enables reforecasting and stress testing the predictions at an early stage of loan admission. The results demonstrate that the analytical process proposed may be adopted by portfolios of other countries and effectively incorporate loan-specific complexities. Furthermore, the XGBoost method shows improvement in the accuracy of predicting default and expected losses in auto loans. Additional insights into the adverse effects of unpredictable events on loan portfolios can be derived by including a systematic shock in credit risk predictions. Finally, model stability and transparency are maintained by including out-of-sample continuous monitoring and SHAP values.

Keywords:

Credit Risk; Expected Losses; Nordics; Auto Loans; Santander; XGBoost; SHAP values;

Contents

Table of Figures

Tables

Definitions and abbreviations

Abbreviation	Meaning
APR	Annual Percentage Rate
AUC	Area Under the Curve
BDS	Bad Debt Sale
CI/CD	Continuous Integration and Continuous Delivery
CFS	Correlation-based Feature Selection
CRM	Credit Risk Management
EAD	Exposure At Default
EBA	European Banking Authority
ECL	Expected credit Risk Losses
EU	European Union
FinTech	Financial Technology
FIs	Financial Institutions
FPR	False Positive Rate
GDPR	General Data Protection Regulation
GINI	Gini Coefficient
IAS 39	Incurred Loss Approach
IFRS 9	International Financial Reporting Standard 9
IRB	Internal Ratings-Based
KS	Kolmogorov-Smirnov
LGD	Loss Given Default
MAE	Mean Absolute Error
ML	Machine Learning
MoM	Month over Month
PD	Probability of Default
P2P	Person to Person
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
XGBoost	eXtreme Gradient Boosting

1 Introduction

1.1 Statement of the problem

The financial industry, specifically in consumer finance, faces the challenge of Credit Risk Management (CRM). The established techniques used to assess Expected Credit Losses (ECL) primarily employ traditional statistical methodologies that may not accurately account for the complexities inherent within our evolving financial system. As such, current economic conditions aspire to a more advanced and unified strategy incorporating both specialized financial knowledge with the power of data analytics.

Traditional methodologies frequently exhibit deficiencies in transparency, explainability and governance, thereby impeding their effectiveness in achieving regulatory compliance and establishing confidence among stakeholders. In an age of heightened regulatory scrutiny, it becomes imperative to acquire a more thorough comprehension of the complex interrelationships between variables within credit risk modeling.

Moreover, current approaches may exhibit insufficient flexibility to accommodate the dynamic macroeconomic landscape, thereby rendering Financial Institutions (FIs) susceptible to potential vulnerabilities. Additionally, inadequate mechanisms for both early warning performance assessment, and observing performance for the total lifecycle within loan portfolios could leave CRM open to unforeseen risks that are not identified during initial risk assessment or portfolio monitoring phases.

The objective of this work is to establish an advanced analytical approach that incorporates data analytics, early warning performance mechanisms and stress testing. This study seeks to present a proof-of-concept framework designed for CRM processes that can harmonize intricate financial complexities with up-to-date development in advanced analytic techniques related specifically to loan origination. The main goal and focus is to enhance resilience and adaptability within FIs as they navigate ever-changing risk environments. Furthermore, other retail banks can leverage similar tools aimed at systematic real time monitoring and efficient control over credit risk levels spanning from initial loan originations until its eventual maturity phase.

1.2 Background of Credit Risk Management

The development of lending practices reflects the dynamic and constantly evolving nature of finance. From ancient times to the present day, lending has progressed from a simple written commitment to repay to a complex and interconnected system that stimulates economic activity on a global scale. The future of lending holds the promise of further innovation and integration with emerging technologies, which will shape the way individuals and businesses access credit in the years ahead. Credit risk assessment is a fundamental aspect of lending and must therefore evolve in harmony with the ever-changing lending landscape. As lending practices become more sophisticated and technology-driven, CRM must adapt to effectively evaluate and mitigate risks, ensuring the ongoing stability and expansion of the financial system.

The banking sector is concerned with credit risk losses, which occur due to the possibility of financial losses arising from unforeseen events. Credit risk, also referred to as default risk, encompasses the inherent uncertainty associated with the complete repayment of loans by borrowers. Loan defaults may be triggered by numerous factors, including reduced income of borrowers, business failures, or fraudulent activities. The numeric estimation of Expected Credit Loss (ECL) involves the multiplication of three key components, namely the Probability of Default (PD), the bank's Exposure at Default (EAD), and the Loss Given Default (LGD), as described by Ghosh (2012).

Effective management of credit risk is vital for ensuring the stability and sustained performance of FIs. Historically, banks have relied on simpler models such as score models and logistic regressions to forecast and manage credit risk, particularly in the short term, typically within the first year of loan origination. However, these models, while straightforward to implement and validate, tend to underestimate risk as time progresses and uncertainties increase. To comprehensively evaluate the performance and profitability of a loan portfolio at origination, it is essential to estimate losses over the entire lifespan of a loan, rather than solely focusing on short-term performance.

The historical progression of CRM has undergone a significant transformation from a reliance on personal relationships and trust to a more quantitative and regulated discipline. According to Dionne (2013), the mid-20th century saw a pivotal moment in the development of CRM

with the introduction of credit scoring models and statistical tools, which enabled a more data-driven approach to credit risk assessment. The author further notes that the private industry's development of these tools was followed by the establishment of standardized risk-weighted capital requirements in banking regulations such as the Basel accord, emphasizing the importance of robust credit risk evaluation. However, the recurrence of financial crises, particularly the global financial crisis in 2008, has repeatedly exposed the limitations of existing managerial practices and industry regulation, leading to substantial regulatory reforms and revisions. The latest reform, the Basel III accord, places greater emphasis on risk-based capital requirements, stress testing, and liquidity management to enhance the resilience of financial institutions.

The evaluation of credit in loan origination is classified into two distinct approaches: qualitative and quantitative. Dastile et al. (2020) contend that historically, credit decisions were based on the 5C's approach, which involved questionnaires pertaining to the borrower's Character, Capital, Collateral, Capacity, and Condition. As described by Nationalbank O. (2004), qualitative models have employed rule-based techniques that rely on expert judgment and conditional (IF/ELSE) statements. These models incorporate traditional rating questionnaires where experts assess various risk factors and assign corresponding scores. Furthermore, qualitative systems may categorize borrowers into risk groups based on subjective evaluations.

In contrast, quantitative methodologies utilize historical data and statistical techniques to estimate credit risk. The Fair Isaac Corporation (FICO) was the first widely used credit scoring system that assessed an individual's creditworthiness based on their credit history and financial behaviors, providing lenders with a numerical representation of credit risk (Hurley & Adebayo, 2017). Dastile et al. (2020) have found that various statistical and machine learning methods are commonly employed in research related to credit risk assessment, including "linear discriminant analysis", "logistic regression", "k-nearest neighbor", "decision trees", "support vector machines", "artificial neural networks", "random forests", "bagging" and "boosting" models, "restricted Boltzmann machines", and "deep learning" models. Among these, logistic regression is the most frequently utilized method due to its simplicity and transparency.

Furthermore, there is an increasing interest in contemporary financial literature regarding the utilization of reinforcement learning methodologies in FIs. In conjunction with deep learning

algorithms, Singh et al. (2022) have noted that scholarly publications systematically classify a broad range of reinforcement learning and deep reinforcement learning methodologies.

The rise of big data and machine learning has brought in a new era of possibilities, enabling FIs to effectively harness and leverage vast volumes of information (Nguyen et al., 2022). Concurrently, the advancements in the field of Financial Technology (FinTech) have disrupted traditional practices by utilizing advanced algorithms and state-of-the-art technology, such as cloud computing, to enhance the accuracy and speed of credit risk assessment while also mitigating associated risks (Zhang et al., 2023). Figure 1 illustrates the evolution of qualitative and quantitative approaches as discussed in the preceding section.

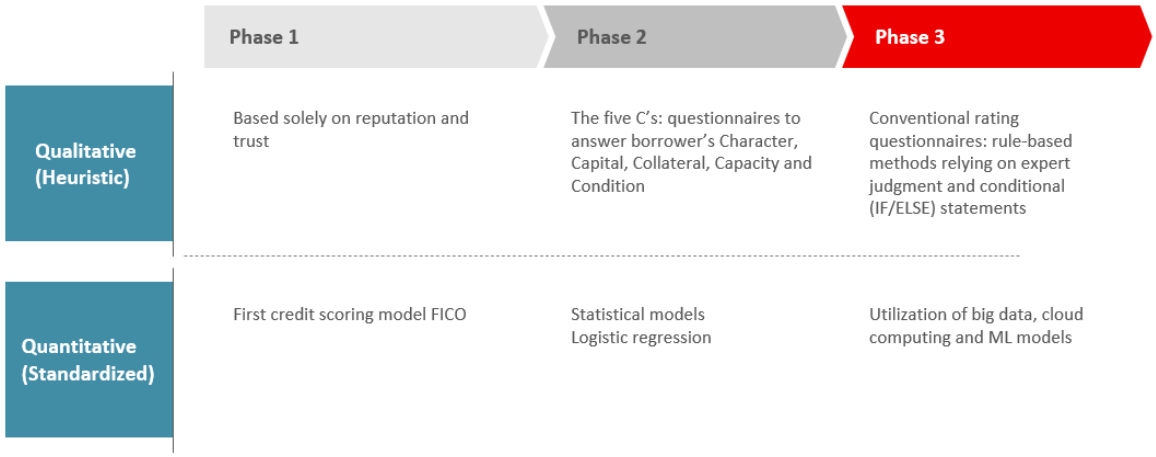


Figure 1. The progression of credit risk assessment[1]

The objective of this thesis is to establish a data-informed methodology for forecasting ECL over the entire loan duration at the time of loan origination. This methodology employs machine learning algorithms and statistical models to scrutinize a vast array of loan origination data, including borrower details, payment history, and external data sources. In addition, this project will integrate a re-forecasting technique based on initial observations of loan portfolio performance. In other words, re-forecasting will leverage the initial observations of the loan portfolio's early performance, utilizing observations of loan status after 3-months and 12-months of admission. This approach aims to update and refine the lifetime predictions by considering the evolving dynamics of the loans over time, ensuring that the forecasts remain

[1] *Information depicted in figure 1 is summarized from Dionne, 2013; Hurley & Adebayo, 2017; Dastile et al., 2020; Nguyen et al., 2022; Zhang et al., 2023.*

aligned with the most recent development and accurately reflect the changing conditions of the credit portfolio. Finally, stress testing is included to augment the insights obtained in the event of unfavorable economic circumstances that may affect borrower payment performance.

1.3 Purpose of the study

The objective of this research is to explore the complex terrain of credit risk in the consumer finance industry, utilizing an integrated methodology that combines financial expertise with data analytics. The emphasis is on conducting a thorough assessment of projected credit risk losses linked to loan origination throughout the entire duration of loans. Through the utilization of advanced concepts such as continuous integration and continuous delivery (CI/CD), automation, and optimization, this study endeavors to connect the domains of finance and data analytics, furnishing valuable perspectives for proficient risk evaluation and credit risk reduction.

Moreover, the study seeks to enrich its analytical framework by incorporating the integration of early warning performance to recalibrate results dynamically, enhancing the accuracy and responsiveness of risk assessments. Additionally, stress testing is examined, to ensure a robust evaluation of unforeseen risks. To align with the evolving landscape of CRM, this research explores the way to foster explainability in predictions of ECL.

The study aspires to contribute valuable insights with the investigation of the following research questions:

1. Is the developed analytics process, leveraging machine learning and statistical techniques to estimate ECL for the entire loan lifetime, transferable and effective when applied across loan portfolios of multiple Nordic countries?

2. Does the XGBoost method effectively challenge the accuracy of the logistic regression model while maintaining comparable explainability?

3. Can the incorporation of a systematic risk factor effectively serve as a proxy for unforeseeable adverse economic scenarios affecting the performance of an entire loan portfolio, and can stress testing based on this random shock provide valuable insights?

1.4 Ethical considerations and Limitations

This section delves into the ethical dimensions essential in conducting research on forecasting expected credit risk losses on loan origination. Recognizing the potential impact of the study on various stakeholders, ethical considerations are paramount to ensure responsible, transparent and sustainable practices.

Respecting data privacy is not just a regulatory obligation, but also a fundamental ethical commitment. The developed tool adheres to all relevant data protection regulations, securing the confidentiality of sensitive information. Personal identifiers are anonymized to prevent the identification of individuals, safeguarding their privacy. The developed modelling database does not contain address details except for postal codes. In addition, names, surnames, national identification numbers and any additional personal identifiers are removed. Therefore, identities of consumers are anonymized and no connection to persons is possible in the developed database, or the tool utilized for portfolio monitoring.

Given the integration of advanced technologies like machine learning, responsible use is a necessity. The study acknowledges the potential implications of technology on decision-making and commits to ethical AI practices. The models are designed with fairness and accountability in mind. Establishing fairness in the treatment of data and results is a core ethical principle, and the study avoids perpetuating bias and discrimination by striving for impartiality in the interpretation of results. Information related to gender or date of relocation to the country of residence, have been omitted from the database, and any inherent biases identified in the data are mitigated to the best extent possible.

Managing credit risk should be done with a mindset that considers the wider societal implications of its findings, such as promoting financial inclusivity by making sure that lending opportunities are available to all segments of the population. The objective is to establish an environment where no minority or sub-group is excluded from borrowing due to credit risk management and strategies. By strategically managing loan portfolios, FIs can customize their lending practices to be inclusive and responsive to the diverse needs of the community. This approach not only advances social equity but also contributes to overall economic well-being by empowering individuals to access the financial resources they require to prosper. An optimized loan portfolio management system serves as a catalyst for securing lending for all,

fostering an environment where economic opportunities are accessible to everyone.

This concise ethical framework highlights the dedication to conducting research with honesty, openness, and a profound sense of accountability to all parties concerned. Ethical considerations are dynamic and necessitate continual contemplation and modification. The researcher pledges to consistently examine ethical ramifications throughout the research journey. Any necessary alterations will be implemented to maintain the ethical principles stated in this current section.

The present study endeavors to introduce a novel tool in the realm of CRM practices. However, it is important to note certain constraints in the process. Primarily, the efficacy and applicability of the proposed tool may be contingent upon the particularity of the data accessible for analysis. The study is heavily reliant on the caliber and quantity of the financial datasets, the quality of the loan portfolio, and any limitations or predispositions within these datasets may impede the soundness of the conclusions.

Furthermore, the study's concentration on certain Nordic countries may limit the applicability of the results to a wider global setting. Disparities in regulatory landscapes, economic frameworks, and financial methodologies among various regions may require additional refinement and authentication of the suggested framework.

The ever-changing and unpredictable nature of financial markets and economic conditions present obstacles in developing a model that can effectively incorporate all conceivable risk factors. Although the proposed framework is extensive, it may not be impervious to unforeseeable disruptions or alterations in the financial environment that surpass the limitations of historical data, even with the inclusion of a stress testing methodology. It is important to acknowledge that CRM, due to its inherent nature, entails a certain level of uncertainty. Consequently, any model or framework utilized in credit risk is susceptible to the constraints of forecasting future events in an environment that is inherently uncertain. These recognized limitations present opportunities for future research and emphasize the necessity for continuous refinement and adaptation in the dynamic realm of finance.

2 Literature Review

The literature review has been organized in a manner that facilitates a thorough comprehension of the fundamental concepts and perspectives that are relevant to the research objectives.

The present review comprehensively examines critical aspects of credit risk. Initially, it delves into pertinent regulatory frameworks in the European Union (EU) (Section 2.1), recognizing the pivotal role of regulation in shaping risk management practices. Subsequently, the literature lists significant factors that determine loan default prediction (Section 2.2), providing valuable insights into the multifaceted nature of credit risk. Furthermore, the review analyzes the intricacies of feature selection and modeling (Section 2.3), highlighting the challenges and considerations involved in developing robust credit risk models. Finally, the scope of the review is broadened to encompass topics such as achieving explainability and stability in predictive models (Section 2.4), underscoring the significance of model interpretability in the financial sector and loan origination.

2.1 Regulatory Frameworks in the European Union

The pivotal function of regulation lies in guaranteeing that technological progressions, such as FinTech, Big Data, and ML, are utilized in a manner that fosters societal acceptance and forestalls the emergence of unanticipated risks within the financial sector (Nguyen et al., 2022; Zhang et al., 2023). This segment presents a synopsis of significant regulatory frameworks, encompassing European perspectives, and their impact on models for assessing credit risk.

The Internal Ratings-Based (IRB) approach is an integral component of credit risk assessment, particularly within the broader Basel framework on banking regulation. By enabling banks to customize their credit risk models to their specific portfolios, the IRB approach enhances accuracy in evaluating credit risk losses. The IRB framework encompasses the following parameters, Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EAD), Maturity (M), and Size (S) (Penikas, 2015), and is essential for understanding how FIs utilize their internal models to estimate credit risk and effectively manage capital requirements.

The International Financial Reporting Standard 9 (IFRS 9) was implemented on January 1, 2018, replacing the Incurred Loss Approach 39 (IAS 39) (PricewaterhouseCoopers, L. L. P., 2017). IFRS 9 represents a notable change in accounting methodology, shifting from a

traditional incurred loss approach to a forward-looking model that prioritizes the management of expected loss and timely credit risk assessment (Porretta et al., 2020). This shift underscores the importance of anticipating and managing credit risk, aligning accounting standards with the evolving risk management practices of the broader financial industry.

The European Banking Authority (EBA) has played an essential role in driving reform efforts within the EU. Drawing on the principles outlined in the Basel accord, the EBA's guidelines on loan origination cover a broad range of areas, including credit risk management, loan origination and risk monitoring frameworks, and stress testing practices (EBA, 2020). These guidelines are designed to promote the standardization of risk management practices across the EU. Additionally, the EBA underscores the importance of effectively measuring, analyzing, and continuously monitoring credit risk drivers. It is imperative that the credit risk management function regularly report their analytical findings to the management body, thereby promoting transparency and informed decision-making within FIs.

The significance of consumer protection regulations has been on the rise in the CRM domain, particularly in the realm of retail lending. The European Union's "General Data Protection Regulation" (GDPR, 2022) and similar regulations have brought to the forefront concerns pertaining to data privacy and equitable lending practices.

The European Union's AI Act of 2023 marks a noteworthy achievement in the regulation of artificial intelligence (AI), encompassing its application within credit risk models. This legislation is designed to promote ethical and responsible deployment of machine learning (ML) models, while also ensuring their robustness and reinforcing the regulatory framework within the financial sector.

2.2 Key factors determining loan default prediction.

In the realm of credit risk assessment, a myriad of features is examined in loan default predictions. These variables span from contractual loan characteristics, applicant characteristics, institutional characteristics, and other relevant statistics.

In a recent study by Rao et al. (2022), a risk assessment mechanism for consumer auto loans is introduced. The study utilizes an extensive set of forty features containing information about

the loan and the borrower. These variables encompass details such as age, identity information, loan characteristics, bureau data like bureau score, number of active accounts, status of other loans, credit history and more. The collective information utilized in this study aligns with the conventional data typically employed in credit risk assessment purposes.

In addition to loan and applicant characteristics, Croux et al. (2020) delve into the significance of employment area, and average taxable income per county or zip code area, shedding light on the multidimensional nature of factors influencing credit risk. In their paper, various macroeconomic indicators were incorporated, including daily Treasury bill rates, the volatility index (VIX), current and real GDP, economic uncertainty levels sourced from the Policy Uncertainty website, and returns of the Russell 2000 Index to approximate market performance. Their findings reveal a significant association between macroeconomic variables and person to person (P2P) lending default likelihood. Specifically, they observed that while risk premium and returns on the Russell 2000 Index heightened default risk, greater volatility in equity options during the month of loan origination appeared linked to a reduced likelihood of default.

Moreover, an increase in GDP growth was associated with a lower probability of default. Kelly and O'Malley (2016) explored the impact of macroeconomic factors on mortgage defaults. They considered regional unemployment rates, house prices and interest rate levels, and similarly found that the tested variables can significantly influence default rates, underscoring the relevance of macroeconomic elements in default prediction models. However, it is essential to note that there are gaps within the existing literature, as highlighted by Dastile et al. (2020). A notable limitation pertains to the absence of studies examining the role of macroeconomic variables in credit risk prediction. Variables such as inflation and unemployment, which are integral elements of the macroeconomic landscape, may exert a substantial influence on a debtor's payment performance.

The intricacies of credit risk assessment extend beyond traditional financial metrics. The research conducted by Cortés et al. (2016) underscores that even weather conditions, as well as the characteristics of firms, loan officers, and borrowers, play a substantial role in determining loan approval rates and, subsequently, the performance of approved loans. For instance, variables like cloud cover, overcast and loan officer compensation may significantly

impact the mood and, by extension, the decision-making process of loan officers, influencing the credit risk landscape.

Therefore, the need for an integrated approach advocating the inclusion of not only loan and applicant characteristics, but additional features that explain the general economic state such as macroeconomic data, information that proxy for the mood of loan officers and typical characteristics of FIs is justified in the literature for the development of credit risk models. EBA (2020) underscore the significance of including macroeconomic variables in the credit risk assessment as well. Nevertheless, the methodology for incorporating aggregated values, like the ones proposed in this review, may not be straightforward, and the outcomes might pose challenges for integration into the modeling processes.

2.3 Feature selection and modelling

Considering the escalating dimensionality of data, the selection of key features to enter the modeling phase is paramount. This process entails identifying and retaining the most influential variables while discarding irrelevant or redundant ones. By doing so, feature selection mitigates overfitting, enhances the model's generalization to unseen data, and facilitates computational efficiency, which is particularly important when dealing with large datasets (Dhal & Azad, 2021). In their investigation, Zhou et al. (2021) employed a feature selection method prior to a classifier and achieved superior results compared to other methods, demonstrating a notable improvement in accuracy. Overall, feature selection constitutes a fundamental strategy for optimizing model outcomes across diverse machine learning applications. Nevertheless, it is imperative to acknowledge the potential risk of selection bias associated with these applied methodologies (Atif & Salmi, 2022).

The correlation-based feature selection (CFS) algorithm, as introduced by Hall, M. A. (2000), is a straightforward and effective means of reducing the number of features utilized as input for a model. This method operates as a filtering technique, whereby feature selection is performed prior to the modeling phase. Furthermore, it is a process that evaluates and ranks subsets of features, rather than individual features.

The metric utilized to determine the optimal subset of features, known as merit, is a direct

computation predicated on the supposition that the subset exhibiting the greatest correlation with the target variable, while simultaneously ensuring that the individual features remain uncorrelated with one another, will yield the highest level of prediction accuracy. The derivation of the merit calculation can be obtained from equation (1) as presented below:

$$merit_s = \frac{k * \overline{r_{cf}}}{k + k * (k-1) * \overline{r_{ff}}} (1).$$

The term "merit" refers to the value of a subset of features, denoted as "s", that comprises a specific number of features, "k". $\overline{r_{cf}}$ is the average correlation of the features with the target and $\overline{r_{ff}}$ is the average correlation amongst the features themselves.

Within the ever-evolving realm of credit risk assessment, a multitude of models have undergone testing to effectively predict and manage credit risk. Logistic regression, a well-established statistical technique renowned for its straightforwardness and comprehensibility, has traditionally been employed in the prediction of loan defaults (Shi et al., 2022).

According to Kelleher et al. (2015), logistic regression is an error-based learning algorithm that is specifically designed to address classification problems. This model is dependent on the coefficients assigned to independent variables and an error function, which is utilized to estimate the degree to which the model's output aligns with the actual target values. During the initial phase, arbitrary values are assigned to each coefficient, and through an iterative process, these values are methodically adjusted to optimize the accuracy of predictions.

Upon obtaining the optimal set of coefficients and feature values (z) through the iterative process, the sigmoid function (σ) -also referred to as the logistic function- is employed to convert this linear combination into probabilities. The sigmoid function guarantees that the model's output falls within the confines of [0,1], which is consistent with the anticipated outcome of a binomial classification problem. The mathematical representation of the sigmoid function is demonstrated in equation (2),

$$\sigma(z) = \frac{1}{1 + e^{-z}} \ (2)$$

and visually represented in figure 2 below.

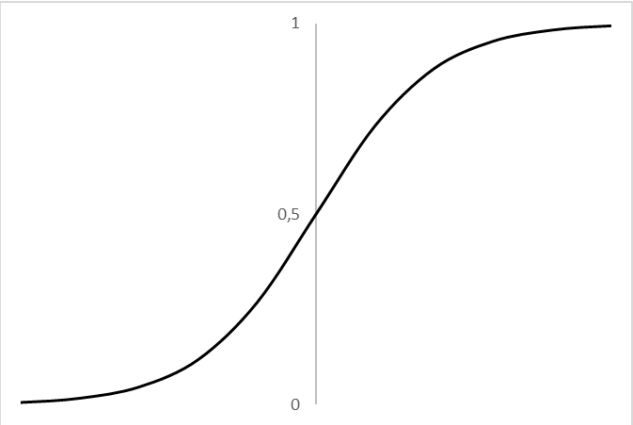


Figure 2. The sigmoid function (σ)

Recent development in the ML field, has introduced more sophisticated methodologies in modelling, such as ensemble tree methods. In a systematic literature review, Dastile et al. (2020) conclude that ensemble methods outperform single classifiers in credit risk forecasting. A noteworthy ensemble tree methodology is the work of Chen and Guestrin (2016), who introduced "eXtreme Gradient Boosting" (XGBoost), a highly efficient and effective tree boosting system. In simplified terms, their model utilizes a tree ensemble method, summing up the weighted predictions of each tree structure produced. Figure 3 below, provides an example of a tree ensemble.

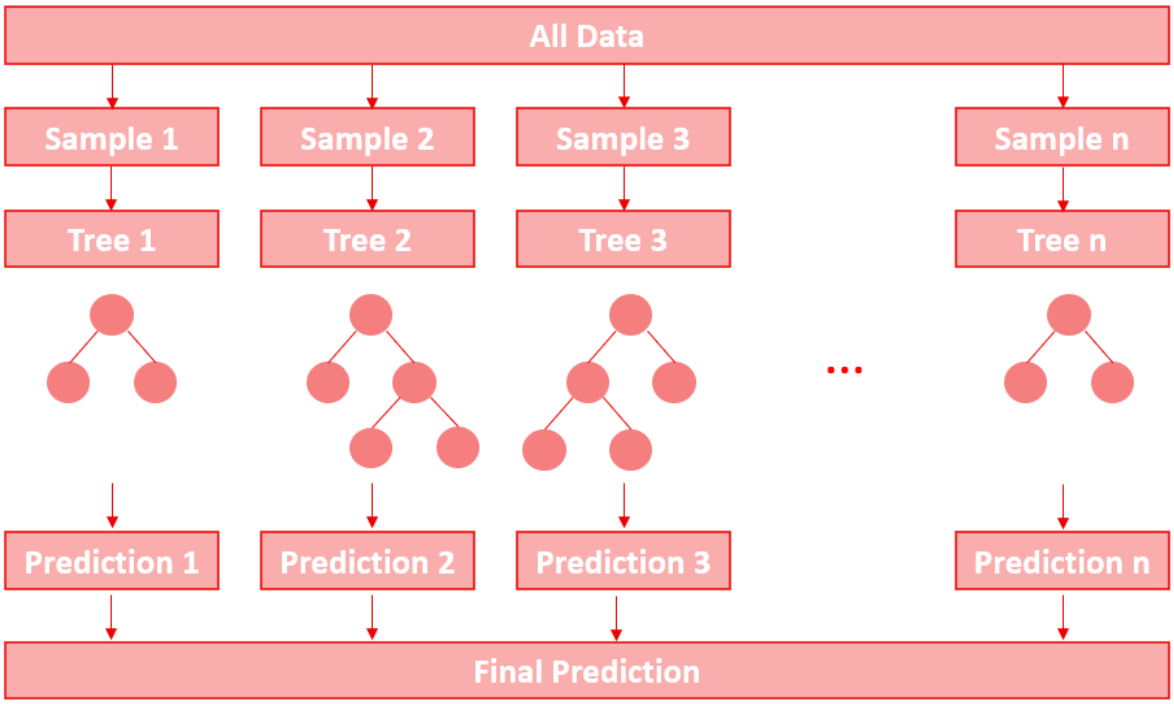


Figure 3. Tree ensemble method

A differentiable objective function, L (2), aggregates the output from the weighted predictions (loss function), measuring the difference between the prediction $\hat{y}$ and the target y. Model complexity is penalized by an additional term Ω in the objective function. The goal is to minimize the loss function, optimizing both leaf weights and tree structure quality to secure high accuracy, while maintaining model complexity to the minimum to reduce model variance.

$$L = \sum_{i}^{n} (\hat{y}i, yi) + \sum_{k}^{m} (\Omega) \; (3)$$

To address overfitting, XGB Boost employs techniques such as shrinkage and feature subsampling. Shrinkage reduces the weight of each tree structure, allowing subsequent tree additions to have an impact on predictions. Feature subsampling further mitigates overfitting and contributes to a reduction in computational requirements.

The identification of tree splits is facilitated by the Basic Exact Greedy Algorithm. This algorithm filters features based on their values and estimates the base score to determine the optimal split for model optimization. For large datasets, an Approximate Algorithm is employed, determining splits based on quantiles and aggregations of continuous variable

buckets. Additionally, the model methodology accounts for missing values or frequent zeros in input features through a technique known as Sparsity-aware Split Finding, which learns the best approach to oversee these occurrences.

The system architecture of the model is designed to ensure efficiency by reducing computational requirements, while enhancing parallelization and scalability. It is worth mentioning that Rao et al. (2022) provide a practical use-case of how XGBoost can positively impact loan default prediction specifically in auto loans.

2.4 Explainability and Stability in Predictive Models

Achieving transparency in predictive models is crucial, as trust in a model's predictions is directly correlated with its utilization. Users are more likely to embrace and apply a model if they have confidence in its outputs (Ribeiro et al., 2016). Furthermore, enhancing explainability not only fosters trust but also contributes to the improvement of model performance and the expansion of knowledge derived from the model (Lundberg & Lee, 2017). As outlined by Gosiewska et al. (2021), trustworthy and accessible predictive models must meet several essential requirements, including high performance, auditability, explainability/transparency, and automaticity. These criteria collectively contribute to the development of models that not only make accurate predictions but also ensure transparency, accountability, and ease of use in real-world applications.

To ensure acceptable accuracy in model predictions, it is important to utilize appropriate metrics that specifically apply to the problem in question. As argued by Kochański (2022), it is a common occurrence in the literature and an industry best practice to employ the Receiver Operating Characteristic (ROC) curve for assessing the discriminatory power of a loan application scoring model. The ROC curve is a graph illustrating the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for a binary classifier. It proves particularly valuable when evaluating the performance of a classifier, especially in datasets with imbalanced targets.

Few metrics like The Area Under the Curve (AUC) and the GINI coefficient can be calculated to summarize information from the ROC curve. AUC quantifies the entire area beneath the ROC curve, providing a comprehensive measure of classifier performance. Additionally, the

GINI coefficient offers a simple transformation that scales the metric between 0 and 1. The GINI calculation is derived from function (4) as described below:

$$GINI = 2 * AUC - 1 \ (4)$$

In addition, the PD that maximizes the Max Kolmogorov-Smirnov (KS) statistics can be estimated, to identify the optimal PD cut-off point. It is a measure of the separation between the binary classes of the target in the ROC curve and can provide an initial recommendation to identify the optimal separation point by ensuring that the model output effectively distinguishes between default and non-default cases (Fang & Chen, 2019).

Lastly, the estimation of how well the PD models follow the monthly aggregated average default rate levels in each portfolio, involves the use of two key metrics, the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). MAE is a measure of the average difference between the predicted values and the true values. RMSE is a measure of the average squared difference between the predicted values and the true values. For both metrics, a lower value indicates a more accurate model. Interestingly, Willmott & Matsuura (2005) argue that when compared to RMSE, MAE should be the metric of choice when measuring the degree of average error in models, by listing its favorable statistical properties.

In addition to RMSE and MAE metrics, a visual validation for the modeled PDs over the entire loan term can be achieved by aggregating the historical development of the default rate into various curves based on the time periods when the loans were financed. These curves illustrate the performance of the portfolio over time, starting from the month when the loans were initially financed. An illustrative example of such a graph is provided in Figure 17 in the Results section, where each curve represents a pool of loans that were financed during the same quarter of the year.

An approach for achieving explainability of complex "black box" models is the Local Interpretable Model-Agnostic Explanations (LIME) technique (Ribeiro et al. 2016), where a simpler linear model is trained to explain the original model output. In a theoretical analysis conducted by Garreau and Von Luxburg (2020), it was observed that LIME effectively identifies key features; however, it tends to omit significant parameters in the process.

Lundberg and Lee (2017) propose a unified measure of feature importance called SHapley Additive exPlanations (SHAP). SHAP values meet three essential requirements for model explainability: local accuracy, missingness, and consistency. Local accuracy ensures that predictions from a simplified model used to interpret the original model match the original predictions of the complex model's output. Missingness indicates that features not considered in the original model cannot have an impact in the simplified model. SHAP values are consistent, meaning that they are not dependent on the architecture of a simplified model but solely on model input.

In addition to transparent and interpretable models, monitoring after implementation is another important aspect of maintaining stable conditions in which advanced analytics methods operate. Monitoring during production is a proactive approach to ensure that machine learning models continue to provide stable results, remain accurate, in a dynamic and ever-changing environment. Population Stability Index (PSI) is a metric used to assess whether the distribution of a specific variable has changed over time (Yurdakul and Naranjo, 2020). PSI is calculated by dividing observations of a feature or the predictions of a model into N groups. The frequencies of observations in each feature are summed up and compared between two samples: the sample used in the modeling phase versus the out-of-sample observations.

2.5 Stress testing

A "black swan" refers to an unpredictable and rare event that has a significant impact. Such events are often characterized by their extreme rarity, high impact, and the difficulty of predicting them beforehand (Taleb, N. N., 2020). The concept underscores the limitations of traditional credit risk models and emphasizes the importance of being prepared for unforeseen and highly impactful events.

In the field of finance and credit loss estimation, the distribution of credit losses is skewed and have fat tails (Galaasen et al., 2020b). In other words, defaults happen rarely but with huge impacts, therefore systems need to be built in a way that are more prepared to withstand the unexpected. Following the approach of Vasicek (2002), portfolio losses contain two determinants, the systematic risk factor and the idiosyncratic risk of each asset. By applying a shock to the systematic risk factor and assessing the impact on credit losses could therefore be a way to stress-test PD predicted by the models initially. Stress testing the credit risk of loan

portfolios is also mentioned as compliance and reporting obligation of FIs in the EBA (2020) guidelines, under the CRM and internal control frameworks section.

In conclusion, understanding and preparing for unexpected events are crucial for building robust risk management in loan portfolios. The rarity and significant impact of events like the 2008 financial crisis, and the COVID-19 pandemic reveal the limitations of traditional credit risk models, emphasizing the importance of readiness and adaptability of FIs in the face of unpredictable events.

Overall, the literature portrays a dynamic and evolving landscape in credit risk assessment, navigating the intersections of technology, regulation, and methodological advancements. More specifically, it highlights the importance of a multifaceted approach that encompasses regulatory frameworks, comprehensive feature analysis, advanced modelling techniques, explainable models, stable operating environments, and stress testing.

3 Research Methodology

This section outlines the structure of the research methodology, offering a roadmap for understanding the approach, data collection, and overall analytics methods employed. The goal is to guide the reader through the key components of the research process.

The research framework is detailed and outlines the chosen methodology. The emphasis on system architecture highlights its importance in facilitating a flexible and scalable framework. The data collection and preprocessing phase includes the gathering of data, handling of missing data, and encoding of categorical variables. The modeling and validation section covers feature selection methods and introduces logistic regression and XGBoost models. The discussion of model output and insights emphasizes early and lifetime performance, portfolio monitoring, and stress scenarios. Validation measures and ongoing monitoring mechanisms, such as the PSI and model drift analysis, are provided to ensure model stability and accuracy. The methodology also addresses explainability, ensuring transparency and understanding in the model's outputs. This structured approach aims to provide a comprehensive and rigorous investigation across all aspects of the research.

3.1 Design of Research Framework

The primary objective of this research is to establish a comprehensive and transparent data analytics framework that can accurately predict ECL on loan origination throughout the entire lifespan of loans. This will entail utilizing machine learning algorithms and statistical methodologies to scrutinize extensive loan origination data, which will include borrower demographics, payment behavior, and other several factors. Additionally, the methodology incorporates stress testing and early warning performance analysis to augment the precision of expected loss estimates. Moreover, the project endeavors to demonstrate the feasibility of a CI/CD approach in FIs through practical implementation.

The research framework's design is the foundation of any research project. It serves as a blueprint for the entire study, outlining the research questions, objectives, and methodology. The design should be carefully crafted to establish a systematic analysis, with clear and concise objectives that are aligned with the research questions. Additionally, the design should incorporate appropriate data collection and statistical methods, as well as ethical considerations. A well-designed research framework is essential for producing reliable and

valid research findings that can contribute to the advancement of knowledge in the field. The research methodology comprises the following fundamental elements.

The analytical framework consists of employing a methodology that incorporates quantitative analysis in conjunction with heuristic data-informed insights. This framework facilitates the application of machine learning algorithms and statistical models, while simultaneously integrating financial expertise and regulatory considerations.

The process of collecting and preprocessing data involves the gathering of a variety of datasets that encompass borrower demographics, financial indicators, and macroeconomic variables. This is achieved through the implementation of a standardized data collection process. To ensure the quality of the data, various preprocessing steps are undertaken, such as addressing missing data and transforming categorical variables.

The implementation of sophisticated machine learning models, such as logistic regression and XGBoost, are implemented. During the development phase, feature selection techniques are used to ensure the inclusion of relevant variables. To establish high accuracy for out of sample populations, comprehensive validation procedures and metrics are utilized.

The integration of methods to achieve transparency in predictive models, with a specific emphasis on interpretability, is a critical aspect of transparency and explainability. The utilization of SHapley Additive exPlanations (SHAP) is being explored as a means of enhancing the comprehension of model outputs.

Finally, the implementation of a stress testing the forecasts is included in the analytical tool to assess the loan portfolios' resilience in the face of adverse economic conditions.

3.2 System Architecture

The system architecture has been crafted to coordinate the complete data analytics lifecycle, encompassing streamlined data collection, model development, validation, and monitoring of both model and loan portfolio performance. This comprehensive approach guarantees a scalable, maintainable, and adaptable infrastructure for an advanced analytics process.

The process of data collection and preprocessing has been strategically developed to effectively acquire and refine data from a variety of sources. This involves the implementation of data pipelines and preprocessing modules that guarantee the accuracy, uniformity, and preparedness of the data for analytical purposes.

The modeling and validation entail the ML pipeline. This necessitates a solution that can seamlessly incorporate diverse modeling methodologies. To promote the CI/CD model delivery approach, the system architecture must embody flexibility. This involves a framework that can effortlessly update algorithms and models without disrupting established functionalities.

The utilization of versioned model endpoints and code repositories enhances the scalability and reinforces the maintainability of the system. This form of version control enables reproducibility, auditing, and the capability to revert to prior states if required.

The loan application data and model output are consolidated in formats that are specifically tailored to facilitate the subsequent visualization and monitoring process. Through the systematic aggregation of data, the system can provide a comprehensive, unified, and flexible perspective of the analytics process. In terms of data visualization, the system architecture incorporates tools and frameworks that enable the creation of informative and interactive visual representations of data and model output. It also supports the integration of visualizations, thereby creating a user-friendly interface for stakeholders to comprehend intricate data patterns, model output, and derive insights.

Finally, the risk management and monitoring components of the system architecture involves the implementation of monitoring tools, warranting continuous surveillance of model performance, data quality, and effective CRM. The architecture supports month to month monitoring to promptly address drift in KPIs, such as model accuracy metrics (AUC, GINI & RMSE), feature stability (PSI), portfolio default rates and general trends. Drift in model accuracy and/or feature instability is an indication for recalibration or re-modeling. As a general good practice, the models should be re-trained at least once a year, to keep the predictions updated with added information and to avoid any implications in the model predictions. The system architecture of the analytics process is illustrated in the flowchart presented in Figure 4 below.

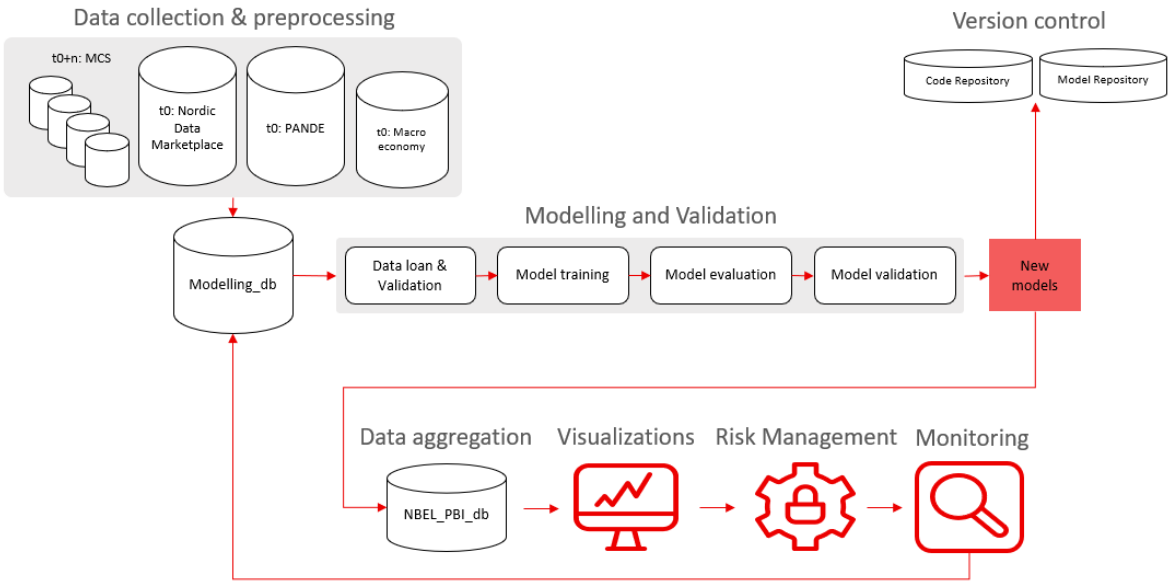


Figure 4. System architecture flowchart

3.3 Data collection & preprocessing

This section presents an overview of the data collection process, including the types of data, sources, and methodology, providing clarity on data origin and nature.

Gathering the data, involves collecting information from diverse sources to build a foundation for analysis and exploration. Most features were fetched from a newly developed relational database, the Nordic Data Marketplace, developed by a data management team and contains information of four countries (Finland, Denmark, Norway, and Sweden). The definitions of features in this database are aligned for all four countries and therefore, it provides a standardized source of loan application information and accelerates the data collection process.

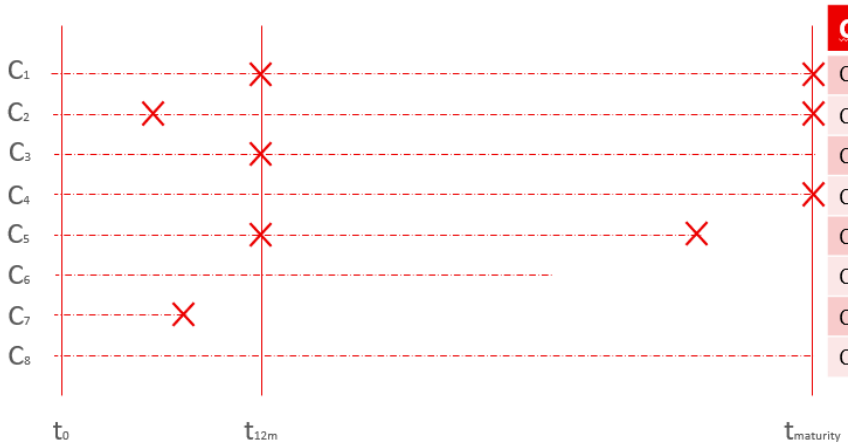
Due to the early inception phase of the Nordic Data Marketplace, not all available information was accessible during the time of data collection. Therefore, variables are gathered from another Nordic database, called PAnNordic Decision Engine. This database contains additional variables stated by the customer in the application phase, as well as data from external sources, such as private and public credit bureaus that were not yet available in the Nordic Data Marketplace. Both databases operate under General Data Protection Regulation (GDPR) requirements and are therefore compliant with regulatory standards regarding consumer protection.

The early performance and the lifetime target variables denoted as *bad_90* and *bad_90_lt* respectively, were collected from local databases from tables called Monthly Contract Snapshot tables for each country, respectively. These tables gather monthly information regarding the loan balance, payment performance, delayed payments, months on book and therefore, provide sufficient information to extract 90+ days past due flags for each contract.

Both, *bad_90* and *bad_90_lt* binary target value is determined at a defined point-in-time. Prior to that specific point in time, the contract might have entered in and out of default status multiple times. However, the binary outcome is assigned a value of 1 only if the contract exhibits a 90+ days past due status, at the designated point in time, which is 12 months after loan initiation for the early performance target and at loan maturity for the lifetime target. This approach verifies that the output of the trained models considers recoveries throughout the loan's lifetime, making it an approximation of PD*LGD. Recoveries outside of the lifetime of the loan and bad debt sales (BDS) are not considered in this study.

Finally, in the event where the loan is not defaulted during the specified point in time, the binary targets are equal to zero. The loans need to have reached the 12 months on book and maturity respectively, otherwise the contracts are omitted from model training. However, in the case of lifetime performance, loans that have been repaid in full before the initial loan maturity are also identified as not defaulted. If the binary classification is missing, then the contract is not considered in the training and testing dataset.

Figure 5 illustrates eight scenarios of contracts (C_n) from the loan origination moment (t0) to 12 months on book (t_{12m}) and up to loan maturity ($t_{maturity}$). As observed, a contract may be classified as default, no default or out of scope, based on the contract status at 12 months and maturity for *bad_90* and *bad_90_lt*, respectively. Default status in the time horizon of each contract is depicted as the red x-mark.



C_n	Bad_90	Bad_90_lt
C_1	Default	Default
C_2	No Default	Default
C_3	Default	No Default
C_4	No Default	Default
C_5	Default	Out of scope
C_6	No Default	Out of scope
C_7	Out of scope	Out of scope
C_8	No Default	No Default

Figure 5. Contract status examples at 12 months on book and at maturity.

Figure 6 presents the targets and groups of features in their original state, before encoding. The groups of features are borrower characteristics, payment history, vehicle information, company information, and other relevant information. As requested by the thesis commissioner, the feature names are confidential. Pseudonyms are given to each feature, based on the group they belong to.

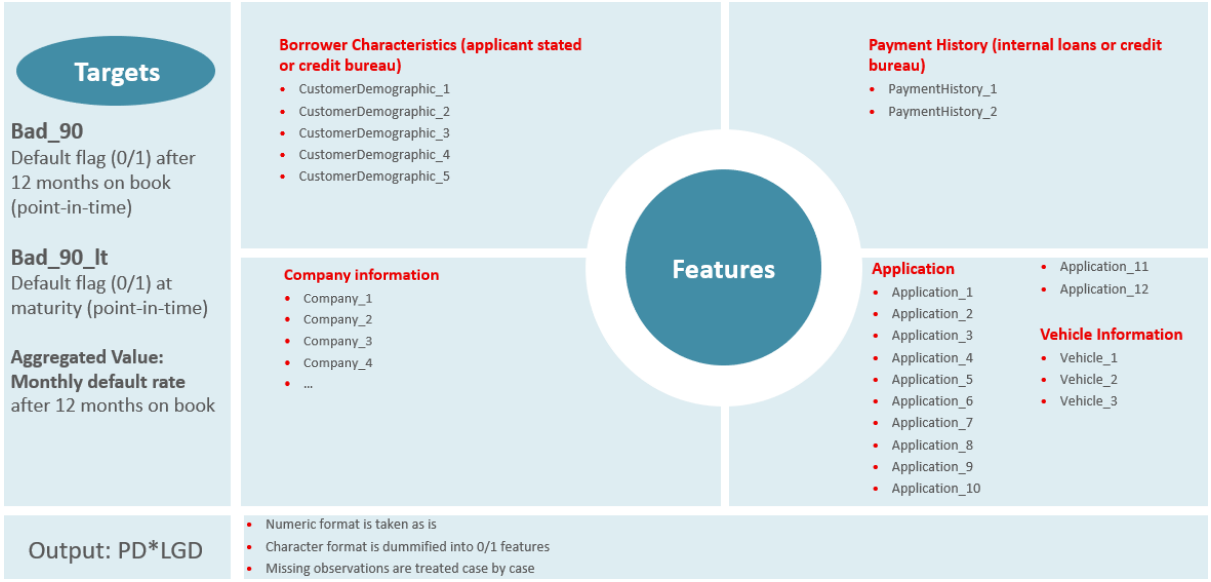


Figure 6. Targets and groups of features

A comprehensive list of features with their descriptive statistics after encoding, have been attached in the Appendix for each country separately.

Most models are not designed to manage input with missing observations, necessitating the treatment of such values. During the preprocessing phase, features with a missing proportion exceeding 50% were removed. The remaining observations were manually filled, by assessing the reason for missing values in each feature. For instance, missing observations for the feature "Application_1" indicate that there has never been an application in the past, indicating a new customer and replacing them with zero is a logical and meaningful choice.

Observations with extreme values in certain features were excluded because model development may be impacted negatively by outliers. Filtering was based on the following criteria: Consumer age should be less than 100, loan duration should be 24 or higher, the residual value percent should be between 0 and 100, and down payment percent should also fall within the range of 0 to 100.

Encoding categorical variables is a crucial preprocessing step in both data analysis and machine learning. This procedure transforms categorical variables into a numerical format, facilitating algorithmic interpretation. Given that most machine learning algorithms struggle with character values, encoding ensures the standardization of dataset features, promoting compatibility with various methods without encountering compatibility issues.

Clustering is a machine learning technique that operates in an unsupervised manner to identify patterns in data and group together similarities. This method involves categorizing data points into clusters, which are subsets of observations sharing common traits. By doing so, clustering methods enhance the understanding of the underlying organization of data. Leveraging clustering algorithms, such as k-means, hierarchical clustering, and DBSCAN, can create new features that capture the underlying complexity of the dataset. These derived features often serve as valuable inputs for machine learning models, aiding in the identification of patterns that might be challenging to discern through traditional means. Incorporating clustering in feature engineering can enhance a model's ability to generalize and make predictions by introducing a higher level of abstraction and representation of the inherent structures present in the data.

Initial results from identifying clusters with the K-means methodology failed in grouping the population of borrowers into segments that would enhance the classification of good and bad

contracts. As such, further experimentation is required to find the correct methodology in incorporating significant clusters that would significantly improve model accuracy.

Before initiating the modelling phase, the dataset is randomly split into a 70:30 distribution of train and test sample. Normalization of the features is not required since either logistic regression or XGBoost methods are sensitive to features' scale.

3.4 Modelling, output, and insights

The CFS method is adopted as a feature selection process. This is to create a simple interpretable initial model for the early performance model with as few inputs as possible to remove any irrelevant variables and noise from the dataset. In addition, since features entering the early performance model are excluded in the next model that estimates lifetime performance, many of the features are preserved to be used in the next modelling phase.

The logistic regression is used as a benchmark (champion) model in the predictions of the target classes, for the early performance output and the lifetime performance, respectively. The XGBoost model is used as a challenger, in both the early performance and the lifetime PD forecasts.

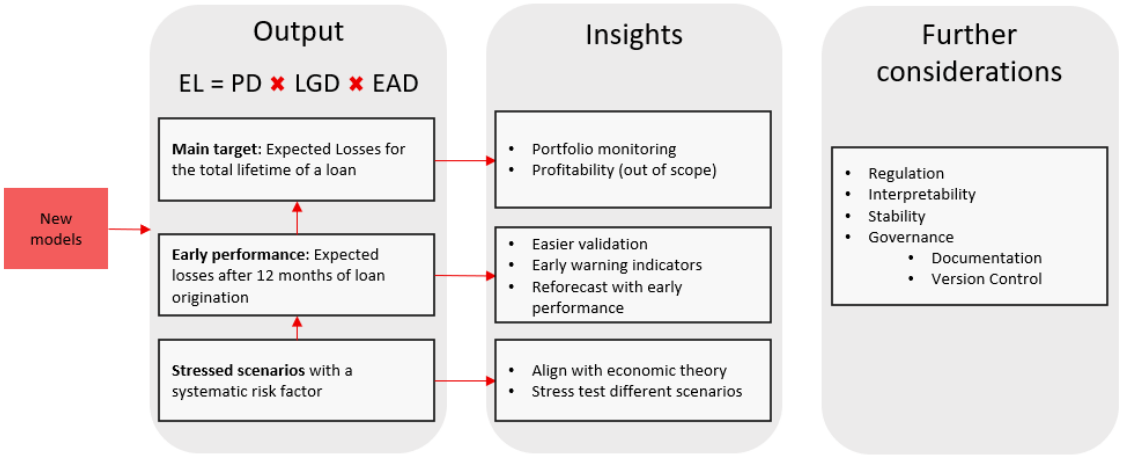
The output from the early performance model is calculated first so that the predictions align with the loan portfolio performance one-year after loans are paid out. Metrics such as thirty days late after three months on book and ninety days late after twelve months on book are investigated in parallel to detect subtle shifts or anomalies in the initial stages, enabling timely responses to emerging risks. By emphasizing early performance evaluation, the model becomes a dedicated tool, mitigating potential risks before they escalate. The predictions derived from the early performance (PD) are also easier to validate since they can be confirmed within just one year. In contrast, the output from the lifetime model requires the loan to be either fully repaid or its performance observed at the maturity of each loan. The average maturity for the loan portfolios under examination is 60 months for Finland and 84 months for Denmark, making it challenging to validate predictions for loans that are disbursed today, for example.

Predicting ECL for the total loan lifetime, is utilized in influencing the strategic decision-making of loan portfolio management. This involves continuous monitoring of the portfolio's

composition and risk profile. Insights into profitability can be derived, validating that pricing is aligned with risk and market conditions. In addition, the model output contributes to optimizing pricing strategies, guiding decisions on new business cases with car dealers. Additionally, a re-forecasting methodology based on initial loan performance is explored on the ECL levels, allowing for dynamic adjustments based on evolving trends in initial stages by observing loan performance after 3 months on book and 12 months on book, respectively. Effectively leveraging the model's output in these areas enhances overall portfolio performance and aligns risk management with general goals in the organization.

A simple stress testing method is applied to assess the portfolio's resilience and robustness under a hypothetical random shock in the economy that affects all loans in the portfolio. By introducing an economic shock to the predictions, the impact of an adverse scenario on the quality of the portfolio can be evaluated. This approach is adopted to address questions such as what would happen to the ECL predictions if an unlikely but highly negative scenario, not observed in the past, were to occur in the economy where the loan portfolio operates. By simulating extreme scenarios, potential vulnerabilities are identified, allowing for preparations and risk mitigation strategies. Understanding the model's behavior in stressed environments is paramount for validating its reliability and providing stakeholders with insights into potential worst-case scenarios.

Model output and insights are summarized in figure 7, with further measurable and qualitative considerations like model governance, regulation, explainability and model stability included in the output and insights derived.



Abbreviation	Parameter	Estimation technique
PD*LGD	Probability of Default*Loss Given Default	logistic regression / XGBoost
EAD	Exposure At Default	based on historical balance rate at default

Figure 7. Analytical service model output, insights and further considerations

3.5 Validation

ROC curves are used to visually evaluate the performance of the binary classifiers. The AUC values summarize the overall performance of the classifiers, where a higher AUC value indicates better accuracy. The train/test ROC curves and accuracy metrics for the Finnish and Danish portfolios for consumers and companies are illustrated in chapter 4.

In addition, the analysis incorporates the KS statistic, while also estimating the PD value in which the KS statistic is maximized. This metric assesses the alignment between the cumulative distribution functions of the predicted PDs for default and non-default cases. The KS statistic is used as a measure of the discriminative power of a classifier and the PD-point at which KS is maximized is where the cut-off between default and no-default is optimal.

The RMSE and MAE are commonly employed to compare the mean deviation between PD and consolidated default rates monthly. The comprehensive evaluation of these indicators facilitates a multifaceted validation procedure, providing a nuanced comprehension of the credit risk models' efficacy across diverse dimensions.

3.6 Monitoring

The consistent surveillance of machine learning models in operation is a crucial aspect of preserving model precision and stability. In dynamic real-world settings, where data distributions, user behaviors, and external factors are subject to change, models may encounter obstacles such as model drift. Monitoring enables the timely identification of issues, enabling prompt interventions to uphold precise predictions, while also promoting the enduring triumph of machine learning models throughout their lifespan.

The solution developed for assessing ongoing performance incorporates three fundamental concepts. First, the PSI, monitoring potential model drift by visual comparison of predictions against the target, and continuous update of accuracy metrics with observations after modelling. PSI is employed to evaluate shifts in the distribution of features and predictions for the out-of-time sample, providing valuable insights into potential changes in the underlying population. Second, model drift, which refers to worsening performance over-time, highlighting the significance of maintaining continuous model validation. Third, accuracy metrics serve as quantitative measures to assess the effectiveness and precision of the models even in post-production. Regularly assessing PSI, monitoring for model drift, and utilizing accuracy metrics collectively contribute to a comprehensive framework for maintaining the reliability and adaptability of credit risk models over time.

It is important to mention that a limited number of observations in the out-of-sample period may result in a volatile PSI value. Therefore, a representative sample needs to be gathered, before any decisions are made based on the PSI level of features and predictions. Evaluating the stability of features and model predictions one year after implementation should be adequate. However, the data scientist responsible for modeling must closely monitor the performance of the models trained in this analytics process and suggest actions based on potential changes in the population of the features, during the model inception phase. It is a standard industry practice to consider re-modeling or re-calibration of a model when the PSI exceeds 0.25 for a specific feature in the model, a level that indicates a significant shift in the population of the variable examined (Yurdakul and Naranjo, 2020).

Visual monitoring of model performance is accomplished through the construction of a graph that compares key metrics, including the observed default rate, predicted probability of default,

and Upper and Lower Confidence Levels (UCL and LCL) at a 95% significance level. These graphs provide a dynamic and real-time assessment of the model's performance over time. The observed default rate reflects the target the model needs to achieve, while the predicted probability of default provides valuable insights into the model's forecasting capabilities. The UCL and LCL serve as critical benchmarks, enabling the identification of significant deviations and potential drift in the model's predictive power. By continuously monitoring and comparing these metrics, any shifts in the model's performance can be detected in a timely manner and tackling such issues as early as possible. This approach enhances the model's adaptability to changing patterns and safeguards its reliability.

It is imperative to consistently monitor accuracy metrics throughout the entire lifespan of the models. In the event of a notable decline in accuracy metrics, particularly GINI, it is necessary to consider remodeling. A GINI value below 0.40 indicates a significant deterioration in accuracy and thus, re-modeling is recommended.

An assessment for recalibration/remodeling at least once per year, will ensure that the latest information from the loan portfolios is used in predictions, preventing any deterioration in the accuracy and model stability.

3.7 Credit Risk Management

The visualizations built in the monitoring tool contain a deep dive page with a graph that depicts the historical performance for each portfolio in each country, respectively. Specifically, the graph illustrates the Month on Month (MoM) development in four metric groups, the early performance, the lifetime performance, the new business expected losses and finally the risk-adjusted Annual Percentage Rate (APR).

The early performance group depicts MoM development of the 12-month actual default rate, the modeled 12-month PD, LCL and UCL at 95% confidence level and the stressed PD at 99% confidence after a hypothetical systematic shock occurs in the economy.

The lifetime performance group illustrates the MoM development of the lifetime PD, the LCL at a 95% confidence level, the UCL at a 95% confidence level, the stressed PD considering shocks that occurred in the early performance, and two reforecasts of the lifetime PD after

observing the actual performance of the loan portfolios at three months and twelve months on books. The twelve-month performance is aligned with the lifetime model parameter derived from the early performance model and can be an input into the lifetime model as is in the reforecast. However, the 3-month performance first needs to be transformed into a twelve-month default rate and then inputted into the reforecast model. The transformation involves refactoring the 3-month performance into a 12-month performance using a calculated factor based on the historical difference between the two metrics.

The new business expected losses group contains MoM development of the lifetime expected losses, which is derived from the Lifetime PD multiplied by the estimated EAD, based on historically observed balance levels of defaulted loans. The conversion rates are specified in table 1. For loans with duration equal to forty-eight months or higher the EAD rate is kept at 7.5% as a conservative measure. A three-month reforecast and a twelve-month reforecast of the lifetime expected losses is included, which both are derived from the actual performance of the loans after three and twelve months on books and used as input in the lifetime PD. Lastly, the adverse scenario after a systematic adverse shock is introduced in the economy is depicted in the new business expected losses. The early performance derived from the stressed scenario estimated from a shock into the economy that impairs the early performance, which consequently increases the lifetime PD and finally negatively impacts the lifetime expected losses.

Loan Duration (Months)	EAD FI	EAD DK
12	62.68%	50.63%
24	34.96%	22.71%
36	15.06%	11.86%
≥48	7.50%	7.50%

Table 1. EAD estimations

Lastly, APR visualizes the MoM development of the effective interest rate, and the effective interest rate after risk adjustment, which is the subtraction of the annualized lifetime expected losses from the effective interest rate. The annualized lifetime expected losses can be derived from the calculation (5) below:

$$\frac{Lifetime\ PD}{maturity} * 365 \quad (5),$$

where maturity is the loan maturity expressed in number of days. The reason that the lifetime PD is annualized is so that it is comparable with the effective interest rate which is a 12-month interest rate that is the return on the loan. APR visualizations are not presented in this study.

3.8 Explainability

Linear models like logistic regression are typically easy to interpret, with feature coefficients of the independent variables that have entered the model, providing an intuitive understanding of a model's output. However, recent advancements in ML have introduced methodologies such as ensemble trees and neural networks, often referred to as "black box" models. These models create complex, non-linear relationships between features and the target, making it challenging for humans to intuitively interpret them. Hence, a methodology to maintain model interpretability and transparency becomes crucial.

SHAP values provide a method for attributing model predictions to individual features, offering a clear and interpretable breakdown of a model's output and can provide insight to whether predictions make logical sense, and models meet ethical and regulatory standards.

3.9 Stress testing

To assess the impact of an adverse economic scenario on expected credit risk losses, a horizontal random shock is simulated by introducing a systematic risk factor. The random parameter ε follows a normal distribution N with a mean of 0.0125 and a variance of 0.0075. One thousand scenarios are simulated based on the parameter ε, affecting all loans in a portfolio. The simulated scenarios are sorted, and the scenario sitting in the 990th worst position is chosen as the systematic risk factor for each loan. The simulated stressed PD is then computed as the initial 12-month PD plus 20% of the simulated systematic risk factor. The mean and variance of parameter ε and the 20% weight are manually chosen to introduce a reasonable level of shock to the early performance PD. In future iterations, these parameters should be selected to simulate actual economic crises, creating an environment as close to reality as possible.

As detailed in the CRM in section 3.7, the stressed early performance PD from the stressed scenario is incorporated into the lifetime PD model, resulting in the stressed lifetime PD. This, in turn, adversely impacts the lifetime expected losses and creates the stress tested ECL.

4 Results

Tables 2-5 show the performance of the two models, logistic regression (Log reg) and XGBoost, on train and test samples for Finnish and Danish consumer portfolios. XGBoost is a slightly better model on all samples and portfolios in both countries. However, GINI and the KS statistic observe a slight deterioration in the test samples, an indication an increase of variance in XGBoost models. Overall, both logistic regression and XGBoost show acceptable results for predicting default in Finnish and Danish consumers, especially in 12-month predictions. XGBoost has a slightly better performance for models that forecast the lifetime performance.

Model	Sample	GINI	Max KS Statistic	Max KS at PD
Log reg	Train	0.632904	49.2191	0.00574231
Log reg	Test	0.632106	47.8988	0.00571142
XGBoost	Train	0.695010	53.6071	0.00497773
XGBoost	Test	0.662824	49.7898	0.00694051

Table 2. Finnish consumers, 12-month models

Model	Sample	GINI	Max KS Statistic	Max KS at PD
Log reg	Train	0.572244	42.9249	0.0191238
Log reg	Test	0.552710	40.9516	0.0193992
XGBoost	Train	0.648408	48.9070	0.0238367
XGBoost	Test	0.615108	46.5832	0.0190045

Table 3. Finnish consumers, lifetime models

The lifetime performance compared to the early performance, suffers some decrease in the accuracy levels in the consumer portfolios of both countries. While the test GINI of the 12-month model for the Finnish consumers in the XGBoost model is equal to 0.6628, the GINI of the lifetime model is equal to 0.6151. The drop in accuracy for the Dannish model is higher, from 0.5974 to 0.5484. This behavior is somewhat expected, since uncertainty increases with time and the information retreived during the loan application, might be less relevant after years.

Model	Sample	GINI	Max KS Statistic	Max KS at PD
Log reg	Train	0.575202	45.4916	0.0123681
Log reg	Test	0.597445	47.2837	0.0100311
XGBoost	Train	0.635428	49.0635	0.0115133
XGBoost	Test	0.630072	50.2078	0.0101800

Table 4. Danish consumers, 12-month models

Model	Sample	GINI	Max KS Statistic	Max KS at PD
Log reg	Train	0.448693	33.4059	0.0529299
Log reg	Test	0.439644	33.2611	0.0530594
XGBoost	Train	0.543425	40.8818	0.0543673
XGBoost	Test	0.511383	38.6207	0.0535503

Table 5. Danish consumers, lifetime models

In a similar manner, the performance of both ML models for Finnish company portfolios are presented in tables 6 and 7. XGBoost predictions seems to outperform the forecasts of the logistic regression, especially in the lifetime models.

Model	Sample	GINI	Max KS Statistic	Max KS at PD
Log reg	Train	0.513238	39.0664	0.00787102
Log reg	Test	0.475579	35.7344	0.00854154
XGBoost	Train	0.601972	44.9673	0.00720392
XGBoost	Test	0.548426	40.9798	0.00859984

Table 6. Finnish companies, 12-month models

Model	Sample	GINI	Max KS Statistic	Max KS at PD
Log reg	Train	0.556112	42.6390	0.0321504
Log reg	Test	0.540695	43.3969	0.0300244
XGBoost	Train	0.680095	53.2117	0.0291171
XGBoost	Test	0.623693	48.1225	0.0245443

Table 7. Finnish companies, lifetime models

All in all, the models have managed to capture consumer and company characteristics to a high degree, since the accuracy levels are acceptable, and the variance among the train and test samples is low.

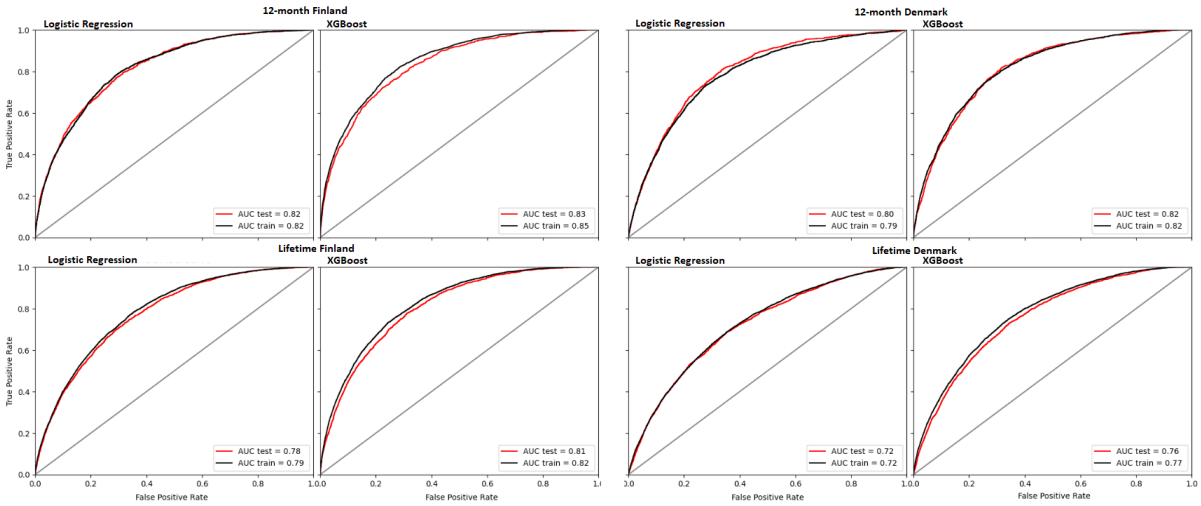
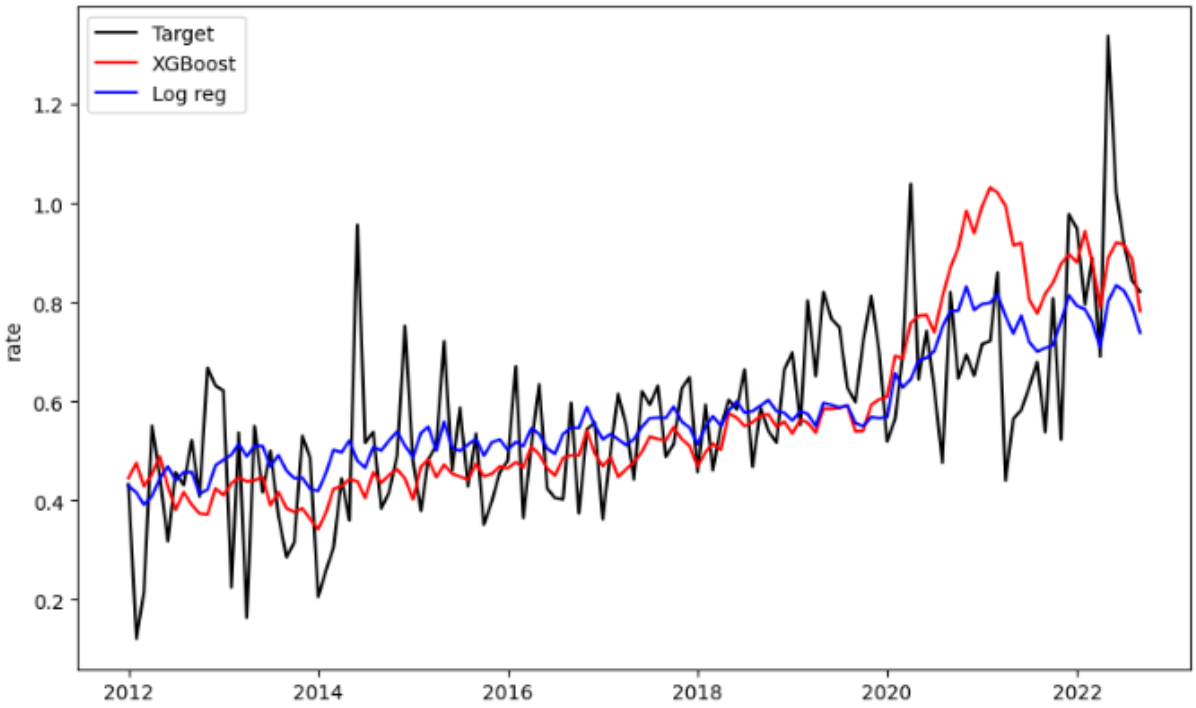


Figure 8. Logistic regression and XGBoost ROC curves Finland and Denmark, consumers

Figures 9 demonstrates the comparison between monthly consolidated 12-month default rate and the average model output of the logistic regression and XGBoost models for Finnish consumers. Similarly, figure 10 illustrates the comparison between monthly consolidated lifetime performance and the average predictions of the logistic regression and XGBoost models. Figures 11-12 show the comparison for Danish consumers and figures 13-14 for Finnish companies. Even though the target lines contain a lot of noise and are extremely volatile from one month to the other, the models manage to follow the trend of each portfolio's performance sufficiently.

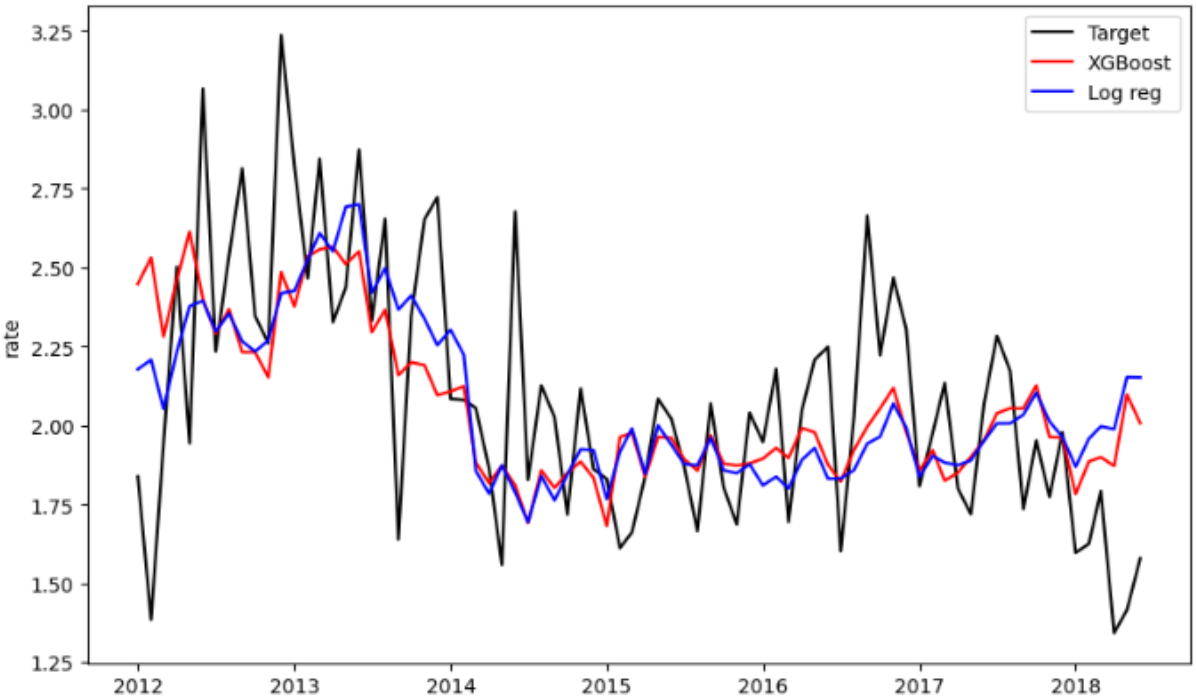
MAE and RMSE metrics are calculated to assess the goodness of fit for both methods in all segments. On the one hand, by observing the figuress, it is clear that both models follow the trend of the average default rate in a very similar manner. On the other hand, the XGBoost method returned lower MAE and RMSE values in all segments, but the difference in the metrics are extremely low.

The "corona effect" is visible in some of the portfolios. During the pandemic, consumers managed to increase their savings and most likely were able to fulfill their commitments to a greater extent, which is also visible in the 12-month performance of the Danish consumers segment during 2020 and in a later period, during 2021, for the Finnish consumers. However, the worsening performance during 2022 due to high inflation, interest rate increases, the general macroeconomic landscape and the war in Ukraine has had a negative impact on the Finnish portfolios of both consumers and companies during 2022 and onwards. As expected, since no macroeconomic variables have been used in these models, the predictions do not manage to catch the deterioration of default rates observed in 2022. The models trained predict the portfolio performance based on the information received during the time of the application, and external factors that may have an impact on the payment performance of consumers and companies are not incorporated.



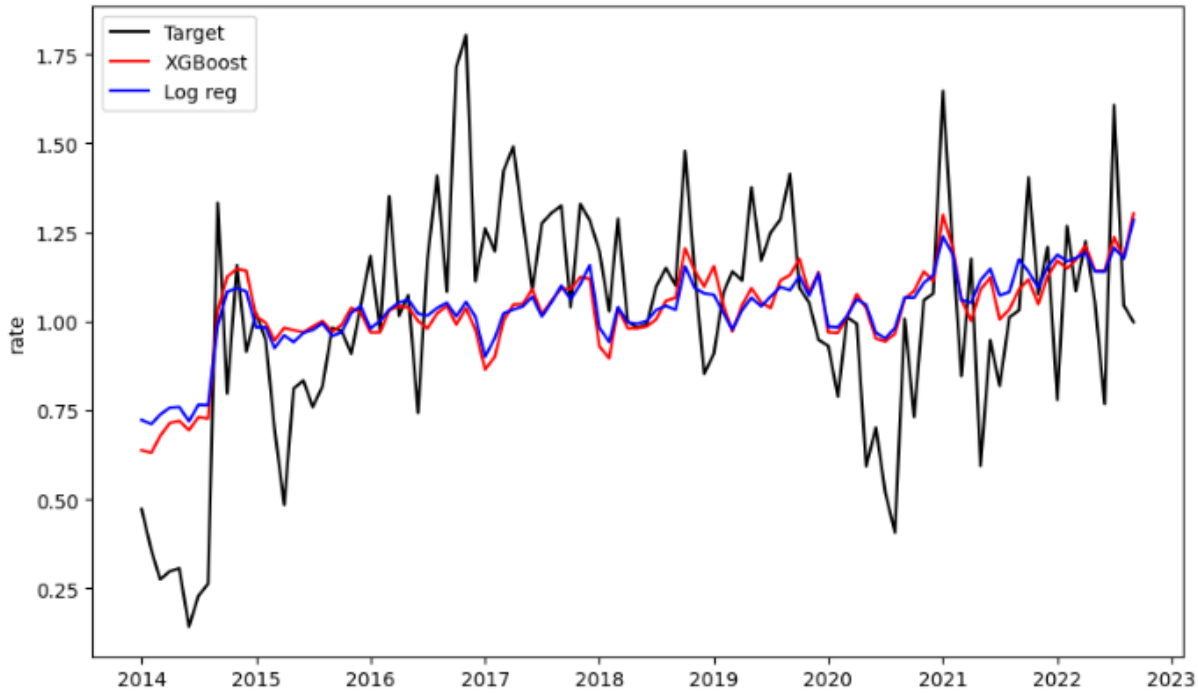
Metric	XGBoost	Log reg
MAE	0.0114	0.0115
RMSE	0.0753	0.0760

Figure 9. Monthly consolidated 12-month default rates – Finnish consumers



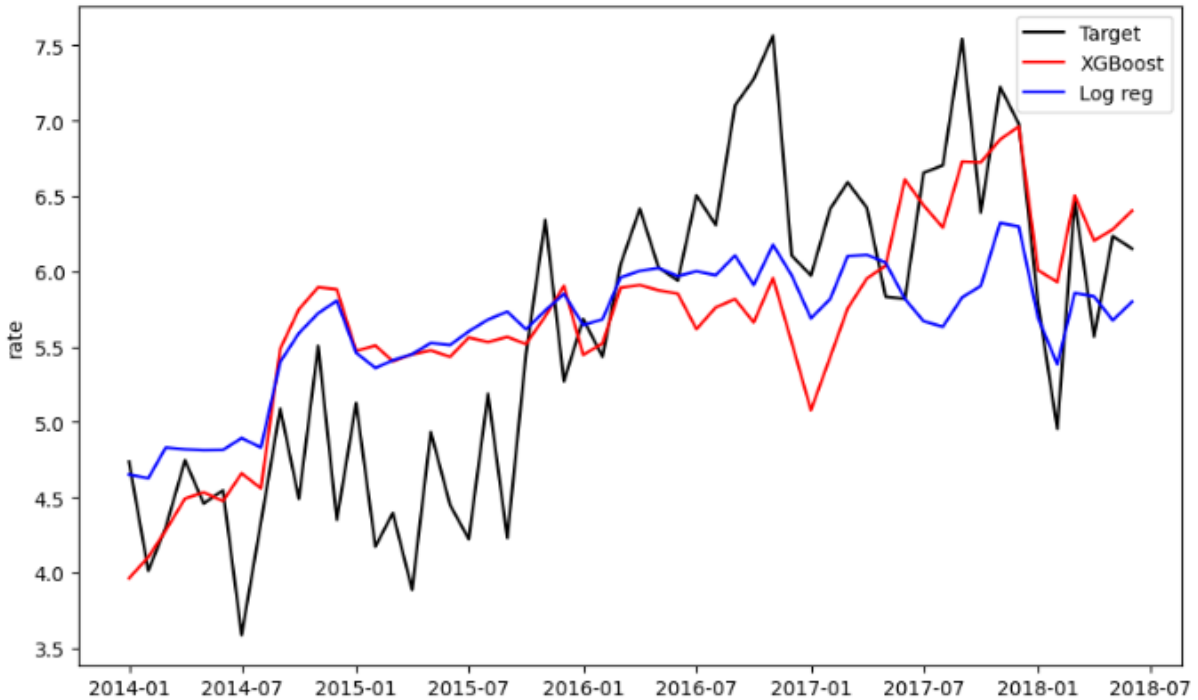
Metric	XGBoost	Log reg
MAE	0.0374	0.0380
RMSE	0.1365	0.1378

Figure 10. Monthly consolidated lifetime default rates – Finnish consumers



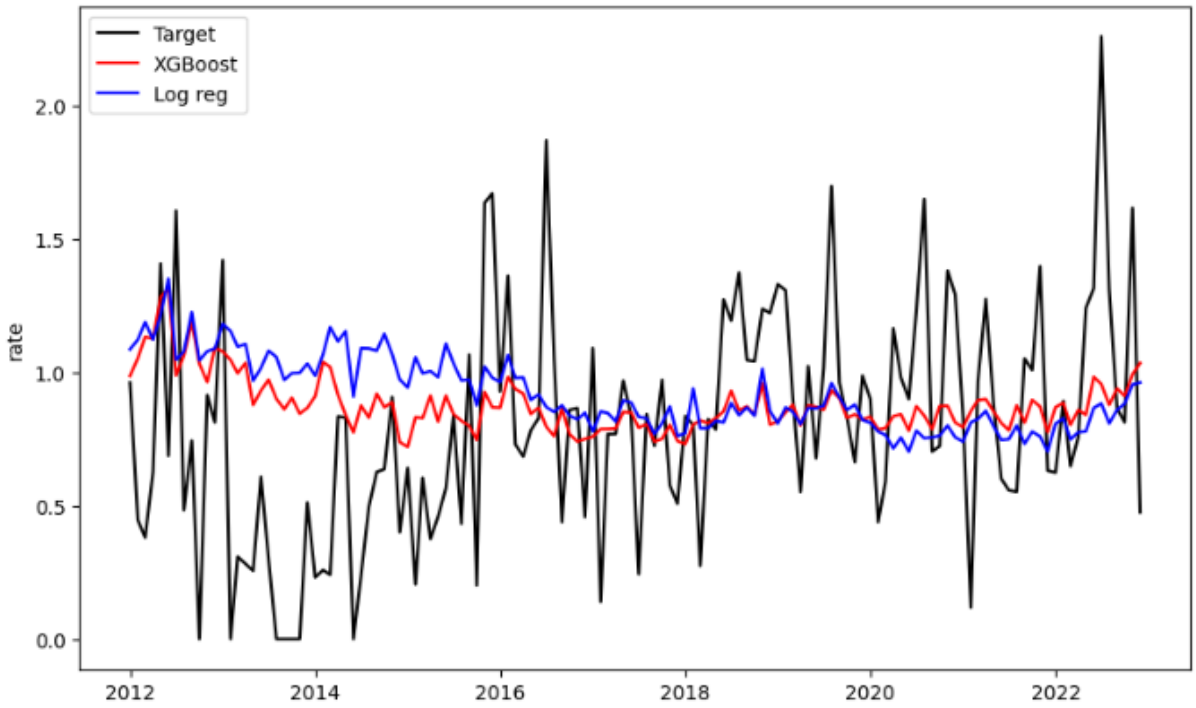
Metric	XGBoost	Log reg
MAE	0.0198	0.0200
RMSE	0.0994	0.0999

Figure 11. Monthly consolidated 12-month default rates – Danish consumers



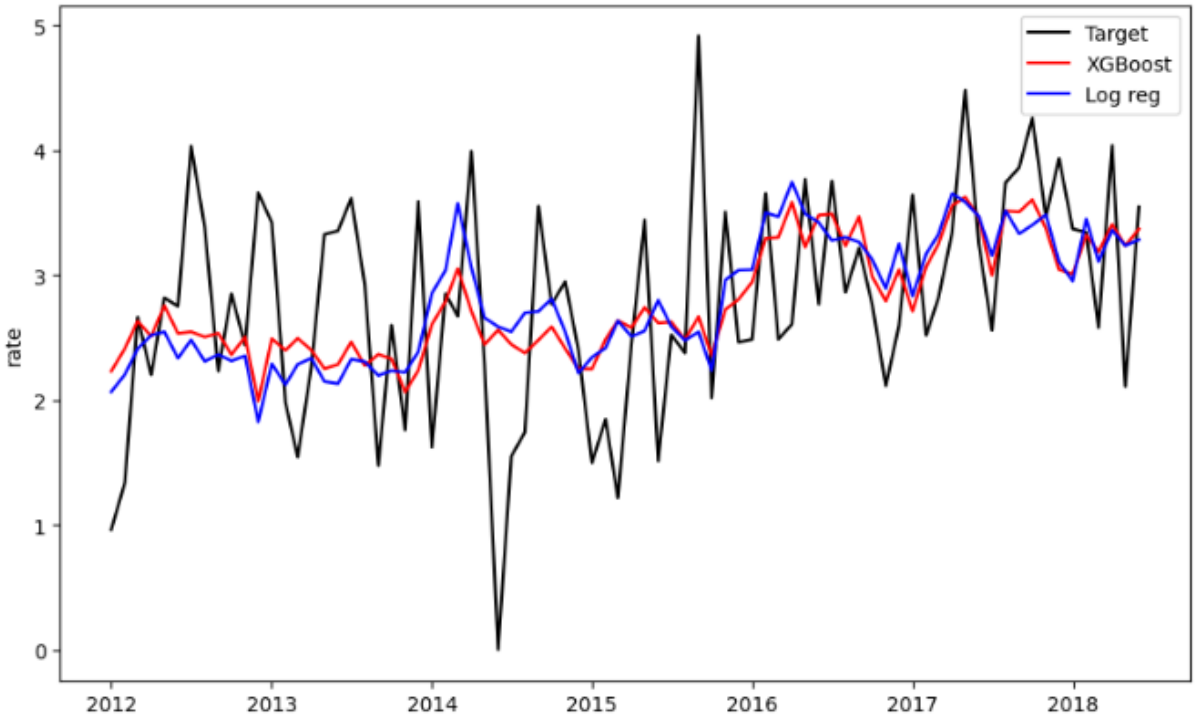
Metric	XGBoost	Log reg
MAE	0.0996	0.1021
RMSE	0.2219	0.2253

Figure 12. Monthly consolidated lifetime default rates – Danish consumers



Metric	XGBoost	Log reg
MAE	0.0168	0.0170
RMSE	0.0915	0.0918

Figure 13. Monthly consolidated 12-month default rates – Finnish companies



Metric	XGBoost	Log reg
MAE	0.0538	0.0553
RMSE	0.1629	0.1659

Figure 14. Monthly consolidated lifetime default rates – Finnish companies

As an attempt to stress test the predictions, a systematic shock is introduced in the model output to simulate an adverse scenario. Such an unlikely event is not observed in the past, is not related to the individual specific risk, but if materialized it can impact the portfolio's performance significantly. By introducing this shock an estimation of the impact on credit losses can be calculated. The results from the simulated shock are presented with dotted and dash lines in figures 15 and 16. The stressed scenarios are only estimated for the last twelve months, so the months where the actual early performance is not observed yet.

In addition, the thiner lines in the figures portrait the reforecasted predictions based on 3-month and 12-month days past due payment of loan contracts. This update of the initial model predictions is a proactive approach in re-assessing newly available information in the predictions and keeping the predictions relevant. The reforecasted lines are very highly correlated with the initial predictions, and fluctuate between the LCL and UCL limits. Furthermore, the majority of the reforecasted values lie below the initial predictions, indicating some conservatism in the models, which is seen as a positive trait from a credit risk monitoring point of view. In a posible future scenario, where reforecast values are seen to deteriorate to levels higher than what was initially predicted, this would be considered as an early warning indicating that the performance of the loan portfolio is underperforming, triggering a request for further deep dive analysis,in order to understand the cause of this occurance.

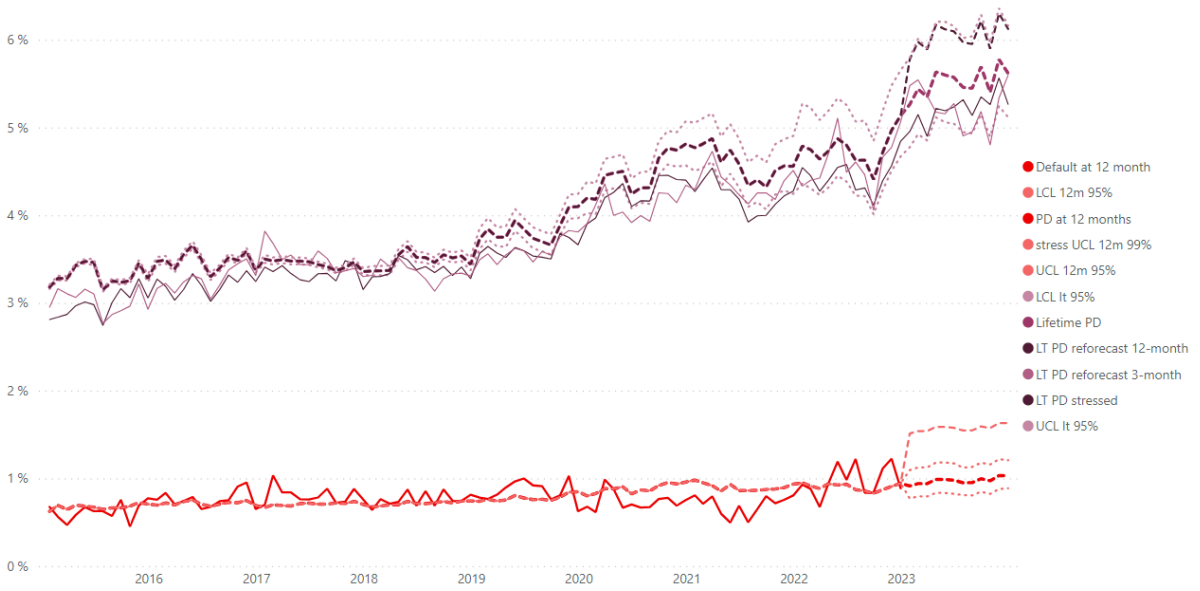


Figure 15. Deep dive - Early and lifetime performance, all portfolios

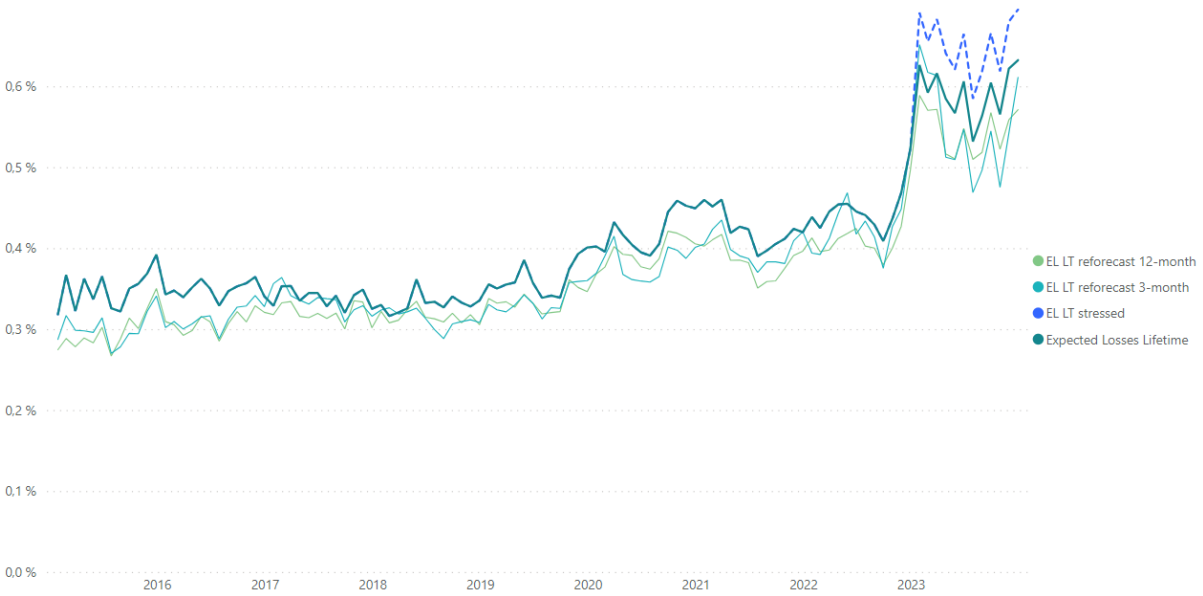


Figure 16. New Business Expected Losses, all portfolios

One of the sections in the visualization tool, is the depiction of model fitting for the 12-month and lifetime performance (left side of figure 17) and the development of default levels based on how long the contracts have been in loan books (rights side of figure 17), connecting the predictions of the early and lifetime performance in one graph. This way, model stability and a one-to-one comparison with the historical performance of the loan portfolios is established.

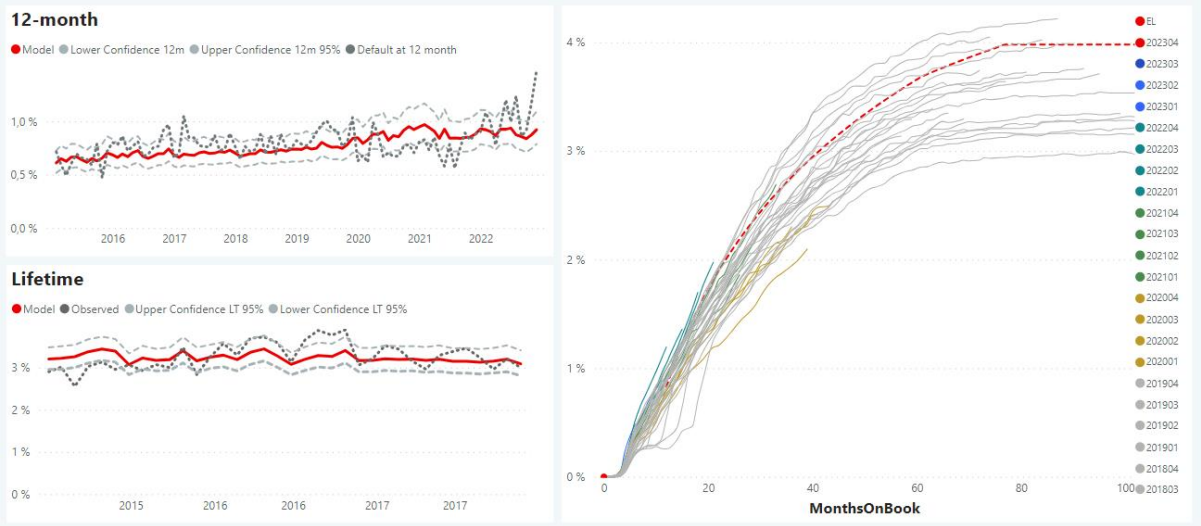


Figure 17. Monitoring - Model fitting

In addition to model fitting, drift in the distribution of the features that have entered the models is examined. In an example showcased in figure 18, the population of the first feature shows a shift in the distribution for applications financed during 2023, as observed in the deciles five

to nine. This shift has led to a PSI level above 0,25 and therefore this particular model needs to be recalibrated, with some adjustments in the modelling sample to grasp the change in the population of this feature. Otherwise, the model needs to be re-trained without this feature. This action will ensure model stability in out-of-sample observations and provide trust in future predictions.

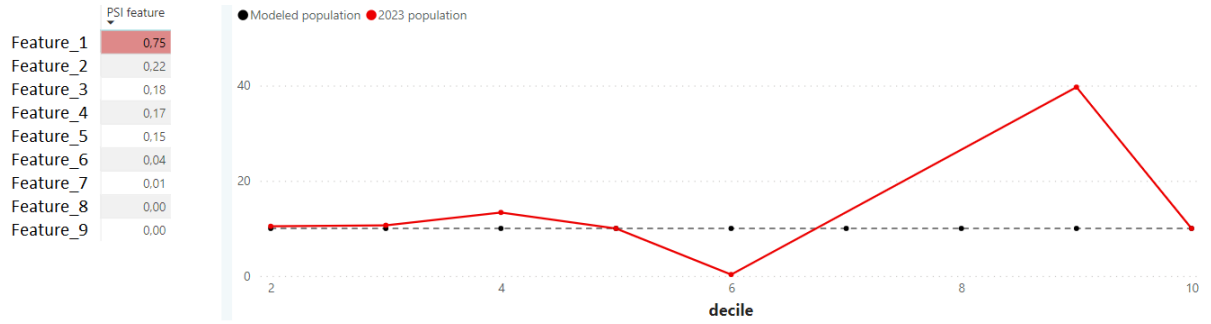


Figure 18. Monitoring – Population Stability

To explain the XGBoost model output, the SHAP values are measured. In figure 19, four graphs are presented, which showcase an example for one of the models trained in the analytical process, the 12-month XGBoost for Finnish consumers. The first graph on the top left, shows the summarized absolute shap values providing a feature importance measure for each independent variable in the model. On the top right, SHAP values are shown in a beeswarm plot, providing further insight into how each feature has impacted the predictions, based on feature value and SHAP value. On the bottom left, the average impact of each feature to the model output can be found. Finally on the bottom right, a waterfall graph for an individual loan application is investigated and the impact of each feature value to the prediction is compared to the prediction of the total population. SHAP values can be summarized and assessed in various ways and provide further insights to the ML developer about any particular bias in the predictions. Similarly, figure 20 depicts the SHAP values for the lifetime model of the same segment.

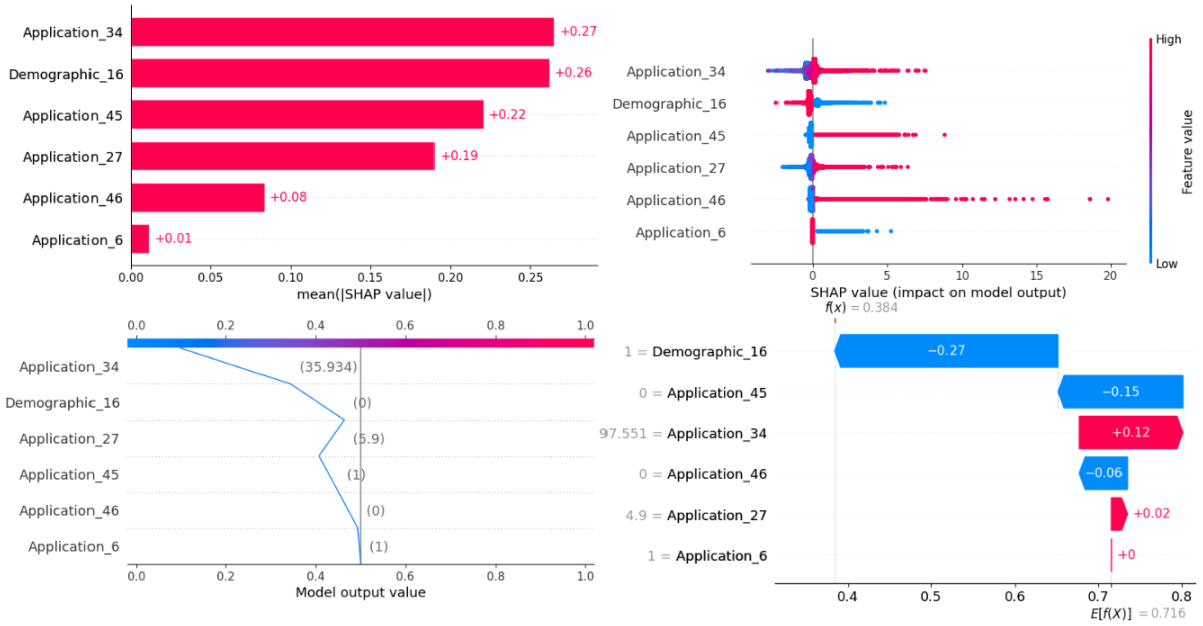


Figure 19. SHAP Value graphs, Finnish consumers, 12-month XGBoost model

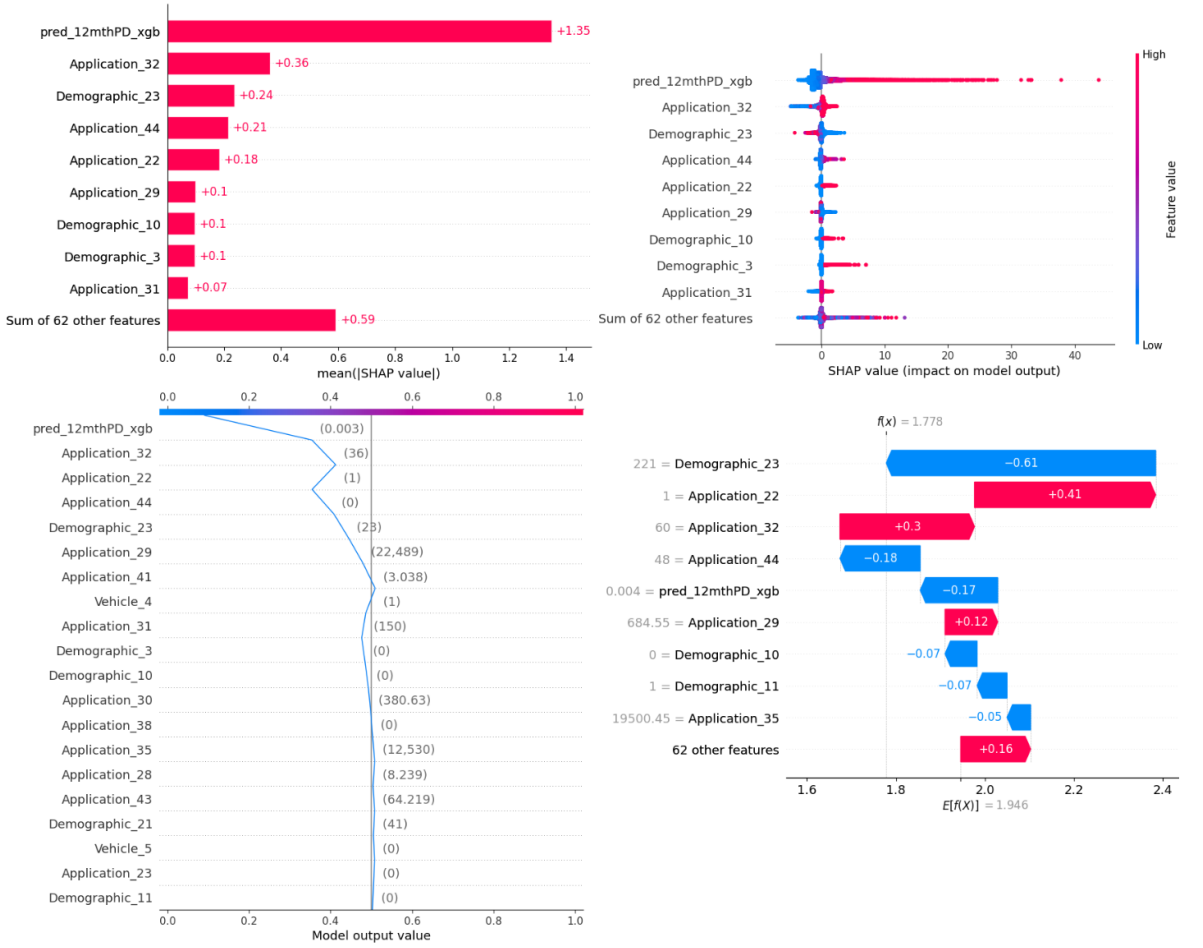


Figure 20. SHAP Value graphs, Finnish consumers, lifetime XGBoost model

5 Conclusion and Discussion

This study yields key findings in response to the research questions. Firstly, the developed analytics process, demonstrates effectiveness and transferability when applied across auto loan portfolios of two Nordic countries. Secondly, the XGBoost algorithm effectively challenges the accuracy of the logistic regression model as evidenced by metrics such as GINI, KS, and AUC, while maintaining comparable explainability by using SHAP values. Thirdly, the incorporation of a systematic risk factor serves as a valuable proxy for unforeseeable adverse economic scenarios that impact the performance of an entire loan portfolio. Stress testing the model predictions, based on this random shock, provides valuable insights to stakeholders. However, the results fall short of establishing a clear link between macroeconomic information and the impact it has on credit risk losses of auto loan portfolios.

The purpose of this work was to propose a proof-of-concept analytical process designed for managing credit risk in auto loan portfolios of the total loan lifecycle. This process aimed to harmonize intricacies in credit risk methodologies with the addition of the latest developments in advanced analytics.

The system was built with maintaining a robust methodology, and a development approach that includes automation, a CI/CD mindset and scalability. The incorporation of processes for model and code repositories, out-of-sample monitoring, and a method for model explainability, places the development of the system's pipeline into the spotlight and enforces an end-to-end transparent methodology. The adoption of this proposed tool provides support for enhanced risk management strategies and insights are derived for data-informed decision-making in terms of determining portfolio management and overall risk mitigation in the realm of consumer finance.

Future development and improvement of this methodology will enhance credit risk optimization in the area of loan origination. Additional data mining, feature engineering and the continuous incorporation of new statistical and ML methodologies in feature selection and modelling are just a few of the areas that can potentially yield improvements in loan origination and portfolio monitoring. The impact of the macroeconomic landscape in loan portfolios should be incorporated in a more standardized and data driven approach and harmonized with a theoretical framework that incorporates the systematic risk factor.

References

Atif, D., & Salmi, M. (2022). The Most Effective Strategy for Incorporating Feature Selection into Credit Risk Assessment. *SN Computer Science, 4(2)*. https://doi.org/10.1007/s42979-022-01500-7

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv. https://doi.org/10.48550/ARXIV.1603.02754

Cortés, K. R., Duchin, R., & Sosyura, D. (2016). Clouded judgment: The role of sentiment in credit origination. *Journal of Financial Economics.* https://doi.org/10.1016/j.jfineco.2016.05.001

Croux, C., Jagtiani, J., Korivi, T., & Vulanović, M. (2020). Important factors determining Fintech loan default: Evidence from a lendingclub consumer platform. *Journal of Economic Behavior and Organization, 173*, 270–296. https://doi.org/10.1016/j.jebo.2020.03.016

Dastile, X., Çelik, T., & Potsane, M. M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing, 91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Dhal, P., & Azad, C. (2021). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence, 52(4)*, 4543–4581. https://doi.org/10.1007/s10489-021-02550-9

Dionne, G. (2013). Risk Management: History, definition and critique. *Social Science Research Network.* https://doi.org/10.2139/ssrn.2231635

EU AI Act: first regulation on artificial intelligence | News | European Parliament. (2023, August 6). https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

EBA (2020). Guidelines on loan origination and monitoring. European Banking Authority. https://www.eba.europa.eu/regulation-and-policy/credit-risk/guidelines-on-loan-origination-and-monitoring

Fang, F., & Chen, Y. (2019). A new approach for credit scoring by directly maximizing the Kolmogorov–Smirnov statistic. Computational Statistics & Data Analysis, 133, 180–194. https://doi.org/10.1016/j.csda.2018.10.004

Galaasen, S., Jamilov, R., Juelsrud, R., & Rey, H. (2020b). Granular credit risk. https://doi.org/10.3386/w27994

Garreau, D., & Von Luxburg, U. (2020). Explaining the explainer: A first theoretical analysis of LIME. HAL (Le Centre Pour La Communication Scientifique Directe). https://hal.science/hal-03233013

General Data Protection Regulation (GDPR) – Official Legal text. (2022, September 27). General Data Protection Regulation (GDPR). https://gdpr-info.eu/

Gosiewska, A., Kozak, A., & Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems, 150*, 113556. https://doi.org/10.1016/j.dss.2021.113556

Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. *International Conference on Machine Learning, 359–366.* https://researchcommons.waikato.ac.nz/bitstream/10289/1024/1/uow-cs-wp-2000-08.pdf

Hurley, M., & Adebayo, J. (2017). CREDIT SCORING IN THE ERA OF BIG DATA. Yale Journal of Law and Technology, 18(1), 5. https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1122&context=yjolt

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine learning for Predictive data analytics: algorithms, worked examples, and case studies.* https://dl.acm.org/citation.cfm?id=2815672

Kelly, R. J., & O'Malley, T. (2016). The good, the bad and the impaired: A credit risk model of the Irish mortgage market. *Journal of Financial Stability, 22*, 1–9. https://doi.org/10.1016/j.jfs.2015.09.005

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Neural Information Processing Systems, 30*, 4768–4777. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Nationalbank, O. (2004). Guidelines on Credit Risk Management: Rating Models and Validation.

Nguyen, D. K., Sermpinis, G., & Stasinakis, C. (2022). Big data, artificial intelligence and machine learning: A transformative symbiosis in favour of financial technology. *European Financial Management, 29(2)*, 517–548. https://doi.org/10.1111/eufm.12365

Penikas, H.I. (2015). History of Banking Regulation as Developed by the Basel Committee on Banking Supervision in 1974-2014 (Brief Overview).

Porretta, P., Letizia, A., & Santoboni, F. (2020). Credit risk management in bank: Impacts of IFRS 9 and Basel 3. *Risk Governance and Control: Financial Markets & Institutions.* https://doi.org/10.22495/rgcv10i2p3

PricewaterhouseCoopers, L. L. P. (2017). IFRS9, Financial Instruments: Understanding the Basics.

Rao, C., Liu, Y., & Goh, M. (2022). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model. *Complex & Intelligent Systems, 9(2)*, 1391–1414. https://doi.org/10.1007/s40747-022-00854-y

Ribeiro, M., Singh, S., & Guestrin, C. (2016) Association for Computational Linguistics " Why Should I Trust You? " Explaining the Predictions of Any Classifier. https://api.semanticscholar.org/CorpusID:13029170

Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications, 34(17),* 14327–14339. https://doi.org/10.1007/s00521-022-07472-2

Singh, V., Chen, S., Singhania, M., Nanavati, B., Kar, A. K., & Gupta, A. (2022). How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda. *International Journal of Information Management Data Insights, 2(2),* 100094. https://doi.org/10.1016/j.jjimei.2022.100094

Taleb, N. N. (2020). Statistical Consequences of fat tails: real world Preasymptotics, Epistemology, and applications. RePEc: *Research Papers in Economics.* https://ideas.repec.org/p/arx/papers/2001.10488.html

Vasicek, O. (2002). The distribution of loan portfolio value. risk, 15(12), 160-162.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30, 79–82.* https://doi.org/10.3354/cr030079

Yurdakul, B., & Naranjo, J. D. (2020). Statistical properties of the population stability index. *The Journal of Risk Model Validation.* https://doi.org/10.21314/jrmv.2020.227

Zhang, Y., Ye, S., Liu, J., & Du, L. (2023). Impact of the development of FinTech by commercial banks on bank credit risk. *Finance Research Letters, 55*, 103857. https://doi.org/10.1016/j.frl.2023.103857

Zhou, Y., Uddin, M. S., Habib, T., Chi, G., & Yuan, K. (2021b). Feature selection in credit risk modeling: an international evidence. Ekonomska Istrazivanja-economic Research, 34(1), 3064–3091. https://doi.org/10.1080/1331677x.2020.1867213

Appendices

Appendix I: Features and descriptive statistics

Country	Portfolio	Feature	Observations	MAX	MEAN	MIN	CORRbad_90_lt	CORRbad_90
FI	Consumer	Demographic_1	606236	1	0,84	0	-0,01	-0,01
FI	Consumer	Demographic_2	606236	1	0	0	0	0
FI	Consumer	Demographic_3	606236	1	0,07	0	0,01	0,01
FI	Consumer	Demographic_4	606236	1	0	0	0,01	0,01
FI	Consumer	Demographic_5	606236	1	0	0	0	0
FI	Consumer	Demographic_6	606236	1	0,03	0	0	0
FI	Consumer	Demographic_7	606236	1	0	0	0	0
FI	Consumer	Application_1	606236	1	0,97	0	-0,06	-0,04
FI	Consumer	Application_2	606236	1	0,03	0	0,06	0,04
FI	Consumer	Application_3	606236	1	0,97	0	-0,06	-0,04
FI	Consumer	Application_4	606236	1	0,01	0	0,04	0,02
FI	Consumer	Application_5	606236	1	0,02	0	0,04	0,03
FI	Consumer	Application_6	606236	1	0,99	0	-0,06	-0,04
FI	Consumer	Application_7	606236	1	0	0	0,01	0
FI	Consumer	Application_8	606236	1	0,01	0	0,06	0,04
FI	Consumer	Demographic_8	606236	1	0,96	0	-0,02	-0,01
FI	Consumer	Demographic_9	606236	1	0,04	0	0,02	0,01
FI	Consumer	Demographic_10	606236	1	0,13	0	0,02	0,01
FI	Consumer	Demographic_11	606236	1	0,48	0	-0,04	-0,03
FI	Consumer	Demographic_12	606236	1	0,32	0	0,01	0,02
FI	Consumer	Demographic_13	606236	1	0,02	0	0,03	0,03
FI	Consumer	Demographic_14	606236	1	0,01	0	0	0
FI	Consumer	Vehicle_1	606236	1	0,05	0	0	0
FI	Consumer	Vehicle_2	606236	1	0,02	0	0	0
FI	Consumer	Vehicle_3	606236	1	0	0	0	0
FI	Consumer	Vehicle_4	606236	1	0,9	0	0	0
FI	Consumer	Vehicle_5	606236	1	0,03	0	0,01	0,01
FI	Consumer	Application_9	606236	1	0,79	0	-0,01	-0,01
FI	Consumer	Application_10	606236	1	0	0	0	0
FI	Consumer	Application_11	606236	1	0	0	0	0
FI	Consumer	Application_12	606236	1	0	0	0	-0,01
FI	Consumer	Application_13	606236	1	0,01	0	-0,01	-0,01
FI	Consumer	Application_14	606236	1	0	0	-0,01	0
FI	Consumer	Application_15	606236	1	0,01	0	-0,01	0
FI	Consumer	Application_16	606236	1	0	0	0	0
FI	Consumer	Application_17	606236	1	0	0	0	0
FI	Consumer	Application_18	606236	1	0	0	0	0
FI	Consumer	Application_19	606236	1	0	0	0	0
FI	Consumer	Application_20	606236	1	0,03	0	0,03	0,04
FI	Consumer	Application_21	606236	1	0	0	0	0
FI	Consumer	Application_22	606236	1	0,16	0	0,01	0

FI	Consumer	Application_49	606236	1	0	0		0
FI	Consumer	Application_50	606236	1	0,93	0	-0,04	-0,03
FI	Consumer	Application_51	606236	1	0,04	0	0,02	0
FI	Consumer	Application_52	606236	1	0,03	0	0,04	0,04
FI	Consumer	Application_24	606236	1	1	0	0	-0,01
FI	Consumer	Application_25	606236	1	0	0	0	0,01
FI	Consumer	Application_26	606236	1	0	0	0	0
FI	Consumer	Demographic_15	606236	1	0	0	0,01	0
FI	Consumer	Demographic_16	606236	1	0,68	0	-0,08	-0,06
FI	Consumer	Demographic_17	606236	1	0,01	0	0	0
FI	Consumer	Demographic_18	606236	1	0,25	0	0,07	0,06
FI	Consumer	Demographic_19	606236	1	0	0	0	0
FI	Consumer	Demographic_20	606236	1	0,02	0	0,01	0,01
FI	Consumer	Application_44	606236	99999	123,34	0	0,05	0,05
FI	Consumer	Demographic_21	606236	119	40,85	0	-0,04	-0,03
FI	Consumer	Application_43	606236	96,84	19,73	0	-0,08	-0,05
FI	Consumer	Application_35	606236	995000	17637,66	430,78	0	-0,01
FI	Consumer	Application_30	606236	85300	293,46	9,93	-0,01	-0,01
FI	Consumer	Application_41	606236	102,97	2,13	0,1	0	0,02
FI	Consumer	Application_34	606236	142,78	81,62	3,16	0,09	0,06
FI	Consumer	Demographic_23	606236	698	86,36	0	-0,06	-0,05
FI	Consumer	Application_46	606236	98	0,32	0	0,06	0,06
FI	Consumer	PaymentHistory_1	606236	182	0,03	0	0,03	0,02
FI	Consumer	Application_45	606236	91	0,17	0	0,11	0,09
FI	Consumer	PaymentHistory_2	606236	2114283,83	73,36	0	0,02	0,02
FI	Consumer	Application_38	606236	45012406	122858,44	0	0,04	0,02
FI	Company	Demographic_1	98867	1	0,75	0	0	0
FI	Company	Demographic_3	98867	1	0	0	0	0,02
FI	Company	Demographic_5	98867	1	0	0	0	0
FI	Company	Demographic_6	98867	1	0	0	-0,01	0
FI	Company	Demographic_7	98867	1	0	0	0	0
FI	Company	Application_2	98867	1	0,1	0	0,04	0,04
FI	Company	Application_3	98867	1	0,9	0	-0,04	-0,04
FI	Company	Application_4	98867	1	0,03	0	0,03	0,02
FI	Company	Application_5	98867	1	0,07	0	0,03	0,03
FI	Company	Application_6	98867	1	0,95	0	-0,03	-0,02
FI	Company	Application_7	98867	1	0,01	0	0,01	0,01
FI	Company	Application_8	98867	1	0,03	0	0,03	0,02
FI	Company	Company_1	98867	1	0	0		0
FI	Company	Company_2	98867	1	0	0		0
FI	Company	Company_3	98867	1	0	0		0
FI	Company	Company_4	98867	1	0,01	0		0
FI	Company	Company_5	98867	1	0	0		0
FI	Company	Company_6	98867	1	0	0		0
FI	Company	Company_7	98867	1	0,7	0	0,03	0,02
FI	Company	Company_20	98867	1	0,28	0	-0,03	-0,02

FI	Company	Vehicle_1	98867	1	0,07	0	-0,01	0
FI	Company	Vehicle_2	98867	1	0,01	0	0,02	0,03
FI	Company	Vehicle_3	98867	1	0,01	0	0	0
FI	Company	Vehicle_4	98867	1	0,58	0	-0,04	-0,01
FI	Company	Vehicle_5	98867	1	0,33	0	0,04	0,01
FI	Company	Application_24	98867	1	0,68	0	0,05	0,02
FI	Company	Application_25	98867	1	0,27	0	-0,04	-0,02
FI	Company	Application_26	98867	1	0,05	0	-0,02	-0,01
FI	Company	Demographic_16	98867	1	0,73	0	0,01	0
FI	Company	Demographic_17	98867	1	0	0	0	0
FI	Company	Demographic_18	98867	1	0,02	0	-0,02	0
FI	Company	Demographic_19	98867	1	0	0	0	0
FI	Company	Demographic_20	98867	1	0	0	0	0
FI	Company	Company_21	98867	1	0,26	0	0,02	0
FI	Company	Company_23	98867	1	0,28	0	-0,04	-0,03
FI	Company	Company_24	98867	1	0,09	0	-0,03	-0,02
FI	Company	Company_25	98867	1	0,07	0	0,03	0,02
FI	Company	Company_22	98867	1	0,01	0	0,01	0
FI	Company	Company_26	98867	1	0,1	0	0,03	0,03
FI	Company	Company_27	98867	1	0,04	0	0,02	0,02
FI	Company	Company_28	98867	1	0,04	0	0,01	0,01
FI	Company	Company_29	98867	1	0,1	0	-0,01	0
FI	Company	Application_44	98867	99999	454,76	0	0,06	0,05
FI	Company	Company_30	98867	125	11,37	0	-0,06	-0,04
FI	Company	Company_35	98867	1768900	1306,24	121	-0,01	-0,02
FI	Company	Company_47	98867	535,44	0,01	-1	-0,01	-0,01
FI	Company	Company_34	98867	2969000	1278,29	-54	-0,03	-0,03
FI	Company	Company_46	98867	202	0,02	-1	-0,01	-0,01
FI	Company	Application_43	98867	96,89	12,84	0	-0,01	-0,02
FI	Company	Company_40	98867	347800	144,88		-0,01	-0,02
FI	Company	Company_44	98867	732,5	0,19	-1	0	0
FI	Company	Company_37	98867	6454305	1299,83		-0,05	-0,03
FI	Company	Company_48	98867	1207	0,12	-1	0	0,01
FI	Company	Application_35	98867	499337,16	31074,02	517,07	-0,02	-0,02
FI	Company	Company_41	98867	131967	-3,46		0,01	0,01
FI	Company	Company_52	98867	9	0	-0,93	0	0
FI	Company	Application_41	98867	104,95	2,29	0,08	-0,05	-0,01
FI	Company	Company_38	98867	312585	221,3	0	0	-0,01
FI	Company	Company_49	98867	231	0,03	-1	0	0
FI	Company	Company_15	98867	0	0	0		
FI	Company	Application_34	98867	121	87,69	3,12	0,02	0,02
FI	Company	Company_33	98867	223844	75,19		-0,02	-0,02
FI	Company	Company_45	98867	978	0,33	-1	0	0
FI	Company	Company_31	98867	99999	36,41	0	-0,04	-0,03
FI	Company	Company_54	98867	100	0	-1	0	0
FI	Company	Application_46	98867	43	0,75	0	0,04	0,03

FI	Company	PaymentHistory_1	98867	82	0,08	0	0,02	0,02
FI	Company	Application_45	98867	34	0,12	0	0,05	0,05
FI	Company	Company_42	98867	3845134	3169,29	-11	-0,02	-0,02
FI	Company	Company_53	98867	485,43	0,03	-1	0	0
FI	Company	Company_36	98867	10667650	2633,7	-14	-0,01	-0,02
FI	Company	Company_50	98867	108	-0,02	-1	-0,01	0
FI	Company	Company_39	98867	4013224	1149	106	-0,01	-0,02
FI	Company	Company_51	98867	202	0,04	-1	-0,01	-0,01
FI	Company	Application_38	98867	53824547	913679,26	0	0,01	0,02
DK	Consumer	Demographic_1	317574	1	0,01	0		
DK	Consumer	Demographic_3	317574	1	0	0		
DK	Consumer	Demographic_4	317574	1	0	0		
DK	Consumer	Demographic_5	317574	1	0	0		
DK	Consumer	Demographic_6	317574	1	0	0		
DK	Consumer	Demographic_7	317574	1	0	0		
DK	Consumer	Application_1	317574	1	0,98	0	-0,03	-0,02
DK	Consumer	Application_2	317574	1	0,02	0	0,03	0,02
DK	Consumer	Application_47	317574	1	0,13	0	-0,03	-0,02
DK	Consumer	Application_3	317574	1	0,98	0	-0,02	-0,02
DK	Consumer	Application_4	317574	1	0	0	0	0,01
DK	Consumer	Application_5	317574	1	0,02	0	0,02	0,01
DK	Consumer	Application_6	317574	1	1	0	0,01	0
DK	Consumer	Application_7	317574	1	0	0	-0,01	0
DK	Consumer	Application_8	317574	1	0	0		0
DK	Consumer	Demographic_8	317574	1	0,97	0	-0,07	-0,03
DK	Consumer	Demographic_9	317574	1	0,03	0	0,07	0,03
DK	Consumer	Demographic_24	317574	1	0,24	0	0,01	0
DK	Consumer	Demographic_11	317574	1	0,45	0	-0,1	-0,06
DK	Consumer	Demographic_12	317574	1	0,27	0	0,08	0,05
DK	Consumer	Demographic_13	317574	1	0	0		0
DK	Consumer	Application_49	317574	1	0,02	0	0	0
DK	Consumer	Application_50	317574	1	0,93	0	-0,04	-0,02
DK	Consumer	Application_23	317574	1	0,04	0	0,05	0,02
DK	Consumer	Application_24	317574	1	0,94	0	-0,05	-0,02
DK	Consumer	Application_25	317574	1	0,04	0	0,07	0,03
DK	Consumer	Application_26	317574	1	0,02	0	-0,01	-0,01
DK	Consumer	Demographic_25	317574	1	0,03	0	-0,02	-0,01
DK	Consumer	Demographic_15	317574	1	0,04	0	0,02	0,02
DK	Consumer	Demographic_16	317574	1	0,5	0	-0,14	-0,07
DK	Consumer	Demographic_18	317574	1	0,4	0	0,12	0,06
DK	Consumer	Demographic_19	317574	1	0	0		0
DK	Consumer	Demographic_21	317574	96	40,52	0	-0,11	-0,07
DK	Consumer	Application_43	317574	87,64	16,22	0	-0,06	-0,03
DK	Consumer	Application_32	317574	180	76,87	3	0,02	-0,01
DK	Consumer	Demographic_2	317574	0	0	0		
DK	Consumer	Application_40	317574	100	3,34	0	0,03	0,02

DK	Consumer	Application_41	317574	25,09	1,46	0	0,02	0,04
DK	Consumer	Application_34	317574	295,26	92,91	12,73	0,07	0,04
DK	Consumer	Application_46	317574	22	0,05	0		0,03
DK	Consumer	Application_45	317574	20	0,03	0		0,03
DK	Consumer	Application_33	317574	1175173	4640,01	0	0,05	0,02
DK	Consumer	Application_37	317574	3553364,06	6715,29	-882706	0	0
DK	Consumer	Application_38	317574	50218820	183788,09	0	0,05	0,02
DK	Consumer	Application_39	317574	581	4,28	0	0,02	0