Credit Default Forecasting with Machine Learning: Turning EBA Recommendations into Action

December 2023

Introduction

In today's data-driven financial and economic landscape, the evaluation of mortgage loan default risks has significantly evolved due to enhanced data insights and sophisticated model techniques. Predictive modelling techniques have risen to prominence due to their ability to enhance decision-making, streamline lending processes, and ultimately enhance transparency and risk management of financial institutions.

The EBA's follow-up report on Machine Learning for Internal Ratings-Based ("IRB") models ("EBA report") published in August 2023 [EBA 2023], (follow-up to the European Banking Authority ("EBA") report of November 2021 [EBA]) emphasises the prudent use of Machine Learning in IRB models, offers recommendations for this prudent use, and points out potential issues related to the General Data Protection Regulation and the Artificial Intelligence ("AI") Act. The aim of this follow-up report is to summarise the main conclusions from the consultation on the Discussion Paper ("DP") on Machine Learning ("ML") in the context of IRB models. According to the report, Machine Learning in the context of Credit Risk can serve various purposes and levels, including data preparation, risk differentiation, risk quantification, and internal validation. Recommendations suggest avoidance of excessive complexity, inclusion of non-predictive drivers, alongside emphasising proper interpretation, documentation, and addressing potential biases in Machine Learning models.

The aim of the report is to establish supervisory expectations regarding the coexistence and adherence of new, advanced Machine Learning models with the Capital Requirements Regulation ("CRR") in the context of IRB models used for calculating regulatory capital for credit risk. This follow-up report specifically focuses on more complex models than traditional techniques such as, regression analysis or simple Decision Trees, which are often less transparent and challenging to understand. In the credit risk context, these Machine Learning models have the potential to enhance predictive power and are already being used in internal models for credit approval processes.

The study conducted by Grant Thornton Cyprus Quantitative Risk ("GT QR") team in collaboration with SEKASA Technologies, investigates Logistic Regression, as a benchmark, against various Machine Learning models' performance in mortgage default prediction. Key points investigated are:

- Predictive accuracy: A comparison of accuracy between Logistic Regression and advanced ML models in predicting mortgage defaults in the light of EBA recommendations.
- Balancing complexity and explainability: Aim to find an optimal balance between model complexity and explainability as stated in the EBA follow up report.
- EBA Recommendations: A detailed examination of how Logistic Regression and advanced ML models can be used to predict mortgage defaults, considering the recommendations of the EBA follow-up report.

It is important to emphasise that this study aims to investigate the benefits and challenges of ML models in financial institutions in the context of the recommendation stated in the EBA follow-up report and not to identify the optimal fit for IRB modelling. For this reason, each model was run on the same data set without model specific feature optimisation.

Various methodologies are assessed in predicting the default on this dataset, including Gradient Boosting Trees ("XGBoost") and Neural Networks ("NN") in comparison with Logistic Regression which acts as a benchmark of a traditional statistical method. As part of our analysis, we also investigated other Machine Learning models (Decision Trees, Random Forests, Support Vector Machines and Stochastic Gradient Descent Classifiers). The latter are not presented in the report, as the results are similar to those of the Logistic Regression model and Gradient Boosting Trees algorithm.

This research provides insights that can be assessed by Financial Institutions ("FI") and specifically risk modelers, in expanding on the suite of methodologies already employed in risk modelling, in the light of the recommendations as presented in the EBA follow-up report. It also explores the trade-offs mentioned in the relevant report between model complexity, interpretability, and predictive accuracy in credit risk assessment. This study points out challenges in validation and peculiarities of the models that are relevant when applied to credit default forecasting.

The study utilises a comprehensive historical mortgage loan dataset. The dataset contains information on 50,000 U.S. residential mortgage borrowers across 60 timestamps. The target to be examined is a binary value (indicator) based on the probability of default predicted by the models ("default_time ").

The subsequent sections provide an in-depth exploration of the study's methodology and outcomes. Section 2. Data Preparation outlines the loading, feature construction, and preprocessing steps. Section 3. Model Building details the approach, incorporating Forward Stepwise Logistic Regression, Logistic Regression, XGBoost, and Artificial Neural Network Models. Section 4. Results and Discussion presents and analyzes the study's outcomes. The study concludes in Section 5. Conclusion and Future Work with a summary and suggestions for future research.

Data Preparation

The core step in the modelling process is data preparation, which involves organising and refining raw data in a methodical manner. To create a solid basis for additional research, data must undergo the necessary and meticulous alterations and modifications at this crucial stage. Considering this data set, this section seeks to demonstrate the necessary steps for improving the quality and suitability of the data. Additionally, it aims to aid the implementation of reliable and accurate forecasting models for estimating the likelihood that mortgage loans will default, and consequently, the target value, which is default_time.

A. Data Loading

The mortgage dataset includes data on 50,000 residential mortgage borrowers in the United States in a period of 60 years. The periods have been de-identified; a standard procedure in financial data analysis, to safeguard sensitive information. The dataset, containing 622,489 entries, considers loans that originated prior to the observation period, reflecting the complexity of mortgage origination in the actual world. The table below shows the 23 variables that make up the original data set.

Table 1: Original Dataset Variables

Variable Name	Description
Id	Borrower ID
Time	Time stamp of observation
orig_time	Time stamp for origination
first_time	Time stamp for first observation
mat_time	Time stamp for maturity
balance_time	Outstanding balance at observation time
LTV_time	Loan-to-value ratio at observation time (in %)
interest_rate_time	Interest rate at observation time (in %)
hpi_time	House price index at observation time (base year=100)
gdp_time	Gross domestic product (GDP) growth at observation time (in %)
uer_time	Unemployment rate at observation time (in %)
REtype_CO_orig_time	Real estate type condominium = 1, otherwise = 0
REtype_PU_orig_time	Real estate type planned urban development = 1, otherwise = 0
REtype_SF_orig_time	Real estate type single-family home = 1, otherwise = 0
investor_orig_time	Investor borrower = 1, otherwise = 0
balance_orig_time	Outstanding balance at origination time
FICO_orig_time	FICO score at origination time (in %)
LTV_orig_time	Loan-to-value ratio at origination time (in %)
Interest_Rate_orig_time	Interest rate at origination time (in %)
hpi_orig_time	House price index at origination time (base year=100)
default_time	Default observation at observation time
payoff_time	Payoff observation at observation time
status_time	Default (1), payoff (2), and nondefault/nonpayoff (0) observation at observation time

B. Feature Construction

According to the EBA Report, the need to combat overfitting, a common problem in Machine Learning, must be balanced by implementing standardised model development processes. This balance involves applying statistical techniques such as, feature selection and normalisation to ensure model stability, conducting rigorous out-of-time and out-of-sample tests, optimising hyperparameters, and evaluating the model's consistency with economic theory. In addition, expert judgement plays a critical role in feature selection and final model evaluation to ensure the completeness and soundness of the model from an economic perspective. To enhance the dataset and to ensure completeness and soundness in feature selection over the course of this study, new features were created, including time on book ("ToB"), remaining time to maturity ("TtM"), absolute change in interest rate ("IR_change"), percentage change in House Price Index ("HPI_pct_change"), and remaining balance as a percentage of the original balance ("RemBal_pct").These features are calculated based on the existing variables in the dataset.

As a part of dataset refinement, redundant or ineffective variables were systematically removed to enhance the dataset. The eliminated variables include: "id," "time," "first_time," "mat_time," "balance_orig_time," "payoff_time," "status_time," "orig_time," and "TtM ".

Data Preparation

C. Macro-Economic Features

Time-dependent macro-economic features were generated through the process of summarising and lagging variables such as Gross Domestic Product ("GDP") change and the Unemployment Rate ("UER"). These derived features aim to enhance the predictive capabilities of the model by offering supplementary information.

D. Data Preprocessing

The missing values in the data set were treated using simple imputation techniques. The initial missing value analysis revealed that the "LTV_time" variable consisted of missing values for "id" values ranging from id = 39722 to id = 39738, as well as for id = 49658. This encompassed a total of 270 entries in the dataset. Due to the limited number of entries affected, these were subsequently removed from the dataset. To address missing values in aggregated macro-economic features, the mean percentage change in GDP and UER was utilised for imputation. These approaches were employed to ensure the dataset's completeness for modelling. After the data preprocessing step, the dataset includes 622,219 entries and 31 variables with no missing values.

D. Univariate Analysis

This process identifies the relationship between the dependent variable and each one of the features in the variable set individually, to reveal associations or dependencies, often utilising statistical tests and visual aids. A Univariate Analysis was completed between each variable in the dataset and the dependent variable "default_time". Table 2 below, categorizes them as "Selected," or "Not Selected" based on criteria such as chi-squared or t-tests and their p-values. The results are as shown in the table below. According to the results, aside from Real estate type condominium ("REtype_CO_orig_time") and Real estate type planned urban development ("REtype_PU_orig_time"), all other features have been selected.

Table 2: Univariate Results

Variable	T_Test_P_Value	Wilcox_Test_P_Value	Variable_Status
balance_time	3.90344E-05	1.84980E-41	Selected
LTV_time	0.00000E+00	0.00000E+00	Selected
interest_rate_time	0.00000E+00	0.00000E+00	Selected
hpi_time	2.85848E-278	1.48749E-259	Selected
gdp_time	0.00000E+00	0.00000E+00	Selected
uer_time	8.44876E-96	3.76076E-97	Selected
REtype_CO_orig_time	9.54720E-01	9.54753E-01	Not Selected
REtype_PU_orig_time	3.10972E-01	3.15310E-01	Not Selected
REtype_SF_orig_time	4.41149E-03	4.62028E-03	Selected
investor_orig_time	1.09100E-03	1.47633E-03	Selected
FICO_orig_time	0.00000E+00	0.00000E+00	Selected
LTV_orig_time	4.86861E-192	5.14572E-161	Selected
Interest_Rate_orig_time	5.05508E-18	1.26019E-174	Selected
hpi_orig_time	0.00000E+00	0.00000E+00	Selected
ToB	4.79760E-258	7.22935E-49	Selected
IR_change	2.62284E-110	1.00869E-184	Selected
HPI_pct_change	0.00000E+00	0.00000E+00	Selected
RemBal_pct	0.00000E+00	0.00000E+00	Selected
GDP_change_t	0.00000E+00	0.00000E+00	Selected
UER_t	8.44876E-96	3.76076E-97	Selected
GDP_change_t_norm	0.00000E+00	0.00000E+00	Selected
UER_t_norm	8.44876E-96	3.76076E-97	Selected
GDP_change_t_norm_3	1.40261E-221	0.00000E+00	Selected
GDP_change_t_norm_6	1.33880E-71	1.49710E-148	Selected
GDP_change_t_norm_9	3.74331E-11	5.50900E-18	Selected
GDP_change_t_norm_12	4.70759E-43	8.53745E-83	Selected
UER_t_norm_3	2.48686E-19	9.43480E-51	Selected
UER_t_norm_6	1.13412E-129	0.00000E+00	Selected
UER_t_norm_9	2.60980E-207	0.00000E+00	Selected
UER_t_norm_12	2.69673E-213	0.00000E+00	Selected

Model Building

The EBA's follow-up report emphasises that the development of the models will be anchored in the four pillars of data management, technological infrastructure, organisation, governance, and analytical methodology. It also recognises the multiple applications of ML, ranging from data preparation to risk differentiation and quantification to internal validation.

Furthermore, the Report underlines the importance of prudent model construction by advising institutions to avoid overcomplication, encouraging statistical analysis of risk drivers, and, most importantly, advocating for the creation of a comprehensive summary document. Finally, the proposal emphasises the need to identify potential biases in the model, particularly overfitting to training data, while recognising that for credit risk, model parameters should generally remain stable due to the low frequency of changes compared to market risk.

For the context of this research, multiple models have been constructed to get a better understanding of the differences in performances, complexity and managing of overfitting. Before implementing each model, the data set was split into 70-30 train/test set, respectively. To ensure the reproducibility and consistency of the models, a seed was also defined to make the results reproducible and facilitate communication after each execution of the models.

The stratified train/set split is a method for splitting data that is broadly used in Statistics and Machine Learning. This technique splits a data set into a training and a test set (the split can vary according to the concept of study and the datasets in use), while maintaining the same class distribution as the original data. This ensures a representative sample for model evaluation, which is essential for accurate performance evaluation, especially for imbalanced data sets.

A. Logistic Regression

Logistic Regression is a statistical modelling technique used for binary classification tasks [M02], where the goal is to predict the probability of an event occurring, typically represented by a 'yes'/'no' outcome, or a similar indicator. It is widely employed in fields like Machine Learning, Statistics, and Epidemiology.

In Logistic Regression, a linear equation is used to model the relationship between one or more independent variables and the log-odds of the binary outcome. The logistic function, also known as the sigmoid function, is then applied to transform this linear combination into a probability score, bounded between 0 and 1. This probability score can be interpreted as the likelihood of the event occurring. Logistic Regression is valuable for making predictions, understanding the influence of different factors, and assessing the significance of predictors in a way that is both interpretable and computationally efficient, making it a fundamental tool in the predictive modelling toolbox.

As part of this study, a Logistic Regression model was created to predict the default probability of mortgages. The response variable, "default_time", and the predictor variables encompass a range of mortgage-related and macro-economic features. The model's performance was assessed by examining accuracy, ROC curves, AUC, precision, recall, and F1-score on both the training and test datasets.

Feature selection for this model was based on the results of the Forward Stepwise Logistic Regression model, univariate results, and expert judgement. The purpose of using Forward Stepwise Logistic Regression prior to modelling using Logistic Regression was to use the results as a method for feature selection. Forward stepwise Logistic Regression is a method for building a Logistic Regression model by adding one predictor variable at a time based on statistical criteria like p-values or the Akaike Information Criterion ("AIC"). It starts with an empty model and iteratively includes variables to improve model fit. The features that did not match the criteria and were not statistically significant, with a p-value <0.05, were removed from the model. For a more efficient model it is suggested to minimise the number of features in the model and optimise it accordingly.

Model Building

B. Gradient Boosting Trees

Gradient Boosting Trees is a powerful Machine Learning technique that combines the strengths of Decision Trees and Gradient Descent optimisation to create highly accurate predictive models. It trains a sequence of Decision Trees, each one designed to correct the errors of its predecessor. The model iteratively minimises the loss function, gradually improving its predictive capabilities. Gradient Boosting Trees are widely used in various fields, including regression and classification tasks, and are known for their ability to handle complex relationships in data and create robust, high-performance models.

In this study, a Gradient Boosting Model using XGBoost [CFA19] [C15] was set up and trained to predict mortgage defaults. What makes this framework special is its efficiency and performance optimisations, including techniques such as parallel processing, tree pruning and regularisation, which make it exceptionally fast and suitable for large amounts of data. It is well-regarded for its ability to handle imbalanced data, feature selection, and automatic handling of missing values. Additionally, XGBoost provides strong predictive accuracy and flexibility, making it a top choice in Machine Learning competitions and real-world applications, especially when it comes to tasks like Regression, Classification, and Ranking problems.

The model was trained using a subset of the features. The model's performance was examined using accuracy, ROC curves, AUC, precision, recall, and F1-score both on training and test data. For the XGBoost model, no further hyper parameter optimisation has been considered as the model was built based on the default values (in R package "xgboost") for learning rate, maximum depth of trees and boosting rounds. Notably, the features used in the XGBoost model are the same features used in the Logistic Regression model (Section 3.2) to have comparable results. The boosting was done based on decision tree booster for binary classification.

C. Artificial Neural Networks

Neural Networks are a class of Machine Learning models inspired by the structure and function of the human brain. They consist of interconnected artificial neurons organised in layers [LYG15]. Due to their layered structure, they are also referred to as deep learning algorithms. These networks have gained immense popularity in various fields due to their remarkable flexibility within their architecture. When dealing with artifacts in data, Neural Networks can learn to filter out noise and irrelevant information, adaptively improving data quality for more robust predictions. In the case of class imbalance, Neural Networks can be tailored to assign different levels of importance to minority and majority classes, mitigating the issue and promoting more balanced predictions. Moreover, their capacity to model complex relationships helps combat overfitting by learning relevant features and patterns in the data, reducing the risk of overly specific or erroneous generalisations. This adaptability makes Neural Networks invaluable in tackling a wide range of real-world problems and enhancing the quality and reliability of Machine Learning solutions.

It is common that Neural Networks are often referred to as black-box algorithms. However, this is only true to a limited extent. The reason that justifies this, is that such models capture non-linear relationships, which are harder to interpret compared to linear models e.g., Linear Regression. Therefore, to overcome these issues, relevant research around Neural Networks focused on the interpretability of such algorithms, producing numerous algorithms that make Neural Networks explainable and interpretable. Additionally, each prediction can be explained individually and no "rule set" is applied that must apply to all data sets.

As part of the architecture definition for this study, a systematic investigation of 800 Neural Networks architectures was performed using a complex parameter search method. The final architecture contains 194 parameters (178 trainable parameters and 16 non-trainable parameters) and applies a batch normalisation for regularisation.

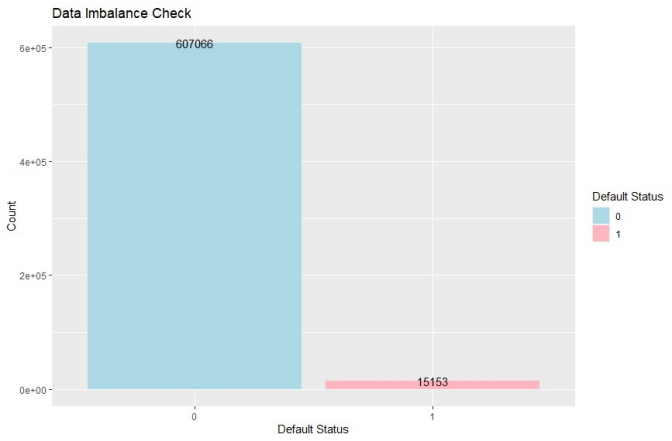
A total of 736 out of the 800 explored architectures, similar calibration issues as with the gradient boosting trees and the Logistic Regression (described in Section 4.1) were faced. The score on the threshold values was greatly optimised however, the actual credit default detection deteriorated significantly.

Results and Discussion

According to the EBA's follow-up Report, it is critical to achieve a balance between model performance and complexity. While increased complexity can improve performance, it often comes at the expense of model interpretability and understandability. It is critical to strike the right balance between performance improvements and added complexity and reduced interpretability.

In the previous sections, the data preparation and different modelling techniques were discussed. The results of the univariate analysis provided a wide range of variables that can be considered significant for our model and the target variable (default_time). As mentioned in the data preparation section, for the modelling a stratified sample of 70% train and 30% test has been considered and a seed (seed = 123) was set to make sure of the reproducibility and consistency in case of having an imbalanced data. Figure 1 illustrates the imbalance for the target variable, as 2,43% of the stratified sample are true defaults.

Figure 1: Data Imbalance Check



A. Interpretation and Assessment of the Results

In this section, the three final models are assessed and compared based on the validation metrics listed in Table 3. The table displays the accuracy, GINI and AUC metrics. Both scores are presented once each with the predicted probabilities ("prob") and the actual predicted classes ("binary"). Additionally, the section showcases the accuracy, true positives, recall, precision, and the F1 score, for a more detailed assessment of the results.

The AUC is calculated in both ways, as the calculation of the AUC on binary values reduces rich probabilistic information to simplistic class labels (0 or 1), resulting in a loss of valuable insights. This approach is also highly dependent on the choice of the threshold, making it challenging to assess the model's performance across different operating points. However, the calculation of AUC on probability values has the following disadvantages:

- **Insensitive to threshold and optimal threshold:** AUC is insensitive to the choice of classification threshold, which can be a disadvantage when the optimal threshold is crucial. It does not offer guidance on selecting the right threshold for specific use cases where different levels of sensitivity or specificity are required.
- **Class imbalance and misleading high scores:** AUC may not adequately address class imbalance. In highly imbalanced datasets, it can yield high scores, even if the model performs poorly on the minority class. The ROC curve focuses on true positives and false positives but doesn't consider class distribution, potentially leading to misleading evaluations.
- **Interpretation Challenges:** Interpreting AUC values can be challenging. While higher AUC scores generally indicate better discrimination, there's no universally accepted scale for defining what constitutes a good or excellent AUC score, making it less straightforward to gauge model performance.

The results presented in Table 3, have been calculated on the holdout sample, which serves as an independent test set. The holdout sample, in this context, refers to the independent test set, used to evaluate the performance of the model presented in the table above. An examination of the results in Table 3 reveals a problem in binary classifications: Gradient Boosting Models and the Logistic Regression reveal their superiority over the Neural Network concerning GINI and AUC calculated on predicted probabilities. Conversely, the Neural Network outperforms the other models when considering the AUC and GINI calculated on the binary values. The absence of default detection in the Logistic Regression and Gradient Boosting Tree Model, despite their relatively good AUC calculated with probabilities, shows a calibration issue.

Results and Discussion

Table 3: Performance Results (for more information on technical terms see 6. Glossary)

	Logistic Regression	XGBoost	NN
Accuracy	0.975	0.975	0.863
AUC (Prob)	0.731	0.784	0.696
GINI (Prob)	0.463	0.568	0.392
True Positives	0	0	1093
AUC (Binary)	0.499	0.499	0.572
GINI (Binary)	0.00	0.00	0.145
Recall	0.00	0.00	0.243
Precision	0.00	0.0	0.093
F1	NaN	NaN	0.145

Calibration issues in Machine Learning models pertain to the situation where the predicted probabilities generated by a model fail to precisely represent the actual likelihood of an event happening. In simpler terms, these predicted probabilities may be inaccurately calibrated, which can have a significant impact on the model's dependability, especially in the context of binary classification tasks. Accurate calibration of probabilities is of paramount importance for informed decision-making and evaluating the associated risk in model predictions. This issue can be traced back to the high-class imbalance and can be attributed to an incorrectly set threshold.

Since the model must recognise possible mortgage defaults to create real added value for financial institutions, the AUC and the GINI calculated on probabilities can be disregarded for a final evaluation. As a result, the Neural Network outperforms the other two models.

Considering these factors, it is crucial to use a combination of evaluation metrics and consider the specific context and requirements of the business problem at hand. While AUC provides valuable insights into the overall performance of the model, it should not be the sole basis for assessing the model's effectiveness, especially when the practical implications and specific use cases are of utmost importance.

B. Explainability of Machine Learning Models

Explainability/Interpretability are crucial for risk management, regulatory compliance, and ensuring that model outputs align with established standards and guidelines. FIs are expected to implement methodologies and tools that allow stakeholders, including regulators and customers, to comprehend the factors influencing model predictions. In this study we made use of SHAP (SHapley Additive exPlanations) [LL17] to ensure explainability for the predictions of each model.

SHAP is highly regarded for its ability to offer clear and intuitive explanations of feature importance. It is widely used in data science and Machine Learning to make complex models more understandable and trustworthy.

In contrast to other explainability methods, SHAP is a model-agnostic technique used to explain the predictions of various Machine Learning models, such as Decision Trees, Random Forests, and Neural Networks. It quantifies the importance of each feature in a prediction by leveraging SHAP values, a concept from cooperative game theory. SHAP values provide a systematic way to distribute the "credit" for a prediction among the input features, considering all possible feature combinations and assessing how each feature contributes to the model's output for a given instance. This makes SHAP a versatile and interpretable tool for understanding model behaviour and feature importance across different Machine Learning models.

Results and Discussion

The figures below illustrate the SHAP Feature importance for the Machine Learning models and the feature importance based on the coefficients of the Logistic Regression that have been examined for this study. The displayed SHAP values represent average values, as these values are created individually for each default prediction. Thus, the bar plots presented below show which input variable has the highest average influence on the prediction (calculated across all cases). The coefficients of the Logistic Regression are determined globally across the data set and do not have to be averaged.

The importance of the input variables alone does not automatically indicate a meaningful narrative. In the case of Machine Learning algorithms, they should be validated individually for each case with regard to the narrative by experts. One example of this is that some mortgage default predictions are justified with a high FICO score, but a high FICO score should be assigned to customers who have a low credit default risk. However, if we review other features in the combination, it often becomes clear that these customers often have exceptionally high loans. In this combination, explainability represents a meaningful narrative.

For Logistic Regression, only one of these narratives is mapped and can therefore be interpreted as a fixed rule set that is applied to all customers in the same way. Although this increases interpretability, it can lead to confusion in special cases that do not correspond to the broad mass of data sets. On this basis, it poses a particular challenge in data sets with a high-class imbalance.

Figure 2: Feature Importance Plot for XGBoost Model

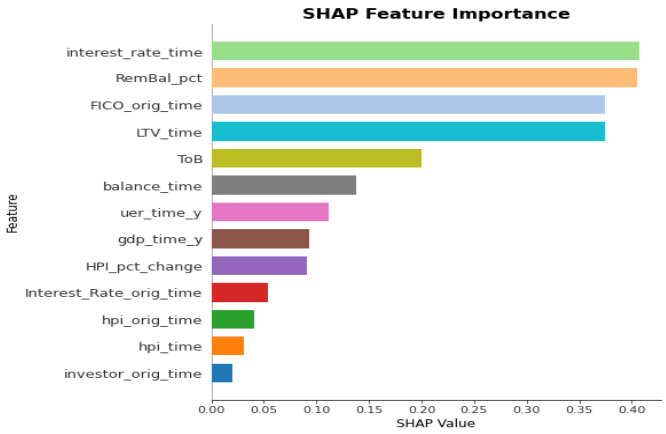


Figure 3: Coefficient Plot for Logistic Regression Model

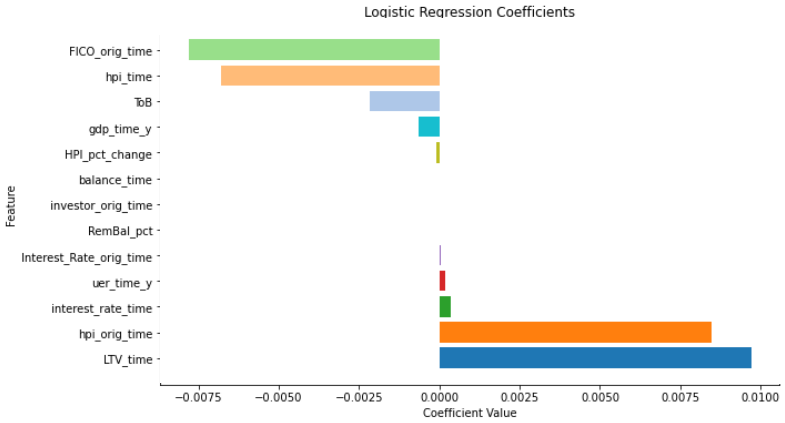
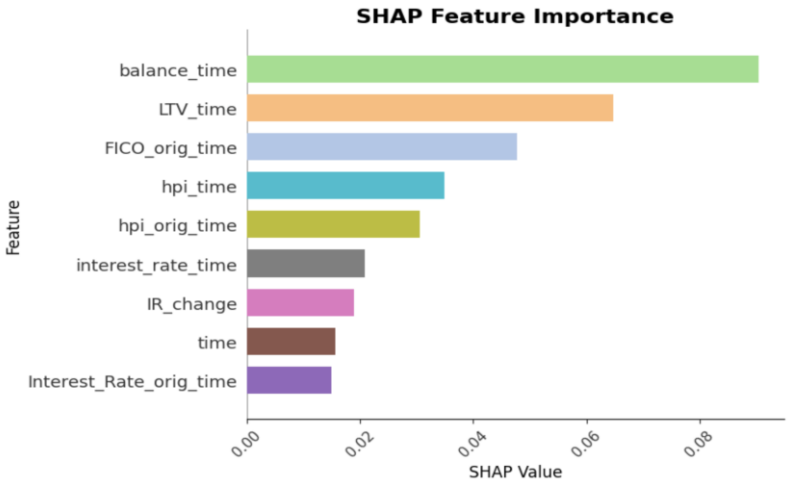


Figure 4: Feature Importance Plot for the NN (Shap-Values are normalised)



Conclusion and Future Work

This study adhered to the recommendations outlined in the EBA's follow-up report on Machine Learning for IRB models. The focus was on avoiding excessive complexity and excluding non-predictive factors. The study also emphasised the importance of appropriate interpretation, thorough documentation, and the elimination of potential biases in Machine Learning models.

In this study, the exploration of trade-offs mentioned in the EBA's follow-up report between model complexity, interpretability, and predictive accuracy in credit risk assessment revealed noteworthy insights. The study conducted shows that the Neural Network algorithm, designed for credit default detection for the purposes of this study, achieves results that correspond qualitatively to state-of-the-art research [OL21], while a variety of Neural Networks with other architectures as well as, Logistic Regression and XGBoost models did not achieve fully comprehensive results by analysing only single scores. Despite yielding good values for predicted probabilities, these models often failed to recognise actual loan defaults.

The results underscore the clear need for comprehensive validation using different scores and techniques. Furthermore, the study emphasised that low scores should not automatically induce uncertainty. When classifying the results, stakeholders should be guided by appropriate baselines, taking into consideration the results in the context of the dataset properties, such as information level and distributions.

There are challenges in the validation process of the models specifically for credit related predictions, that need to be considered by the business owner. Additionally, the models should be examined based on the independent validation units to check if the models are in line with the regulations, fit to business, and that they are reliable models. According to the EBA follow-up report, explainability, complexity, and overfitting are the challenges that need to be resolved while using Machine Learning methods for credit related predictions and they must also be addressed in the model validation process. In quantitative testing for FI, there are individual validation handbooks that specify the thresholds, regulations, and other necessary fields that are required to be revisited and updated in light of the possible use of new machine learning methodologies.

Regarding the interpretability of Machine Learning algorithms, a decisive advantage, which is the individual treatment of data sets, is listed. This is a decisive advantage over linear models, as a fixed rule set becomes increasingly inaccurate as the amount of data increases. At the same time, the non-linear relationships in the models and consequently their explainability pose a challenge. Understanding non-linear relationships can be challenging for humans because our intuition is often built on linear thinking, which assumes that changes in one variable result in proportional changes in another. Non-linear relationships, however, do not follow this straightforward pattern. They involve complex, often unexpected interactions between variables, making it difficult to predict outcomes based on intuition alone.

Summary Documents

In the report, it is recommended to create at the end of the model building and application a concise summary document that simplifies the model's explanation, identifies the key drivers of the model, and elucidates the primary relationships between risk drivers and model predictions.

This document is intended for all relevant stakeholders, including internal staff who utilise the model for their specific purposes. The summary documents for the models created and trained as part of this study show various feature meanings and explain the models used in a brief and comprehensible form:

Summary Document Logistic Regression

Model Name: Logistic Regression

Description

Logistic regression is an interpretable method for predicting binary outcomes. It's commonly used in various fields like machine learning and epidemiology. For mortgage default prediction, a logistic regression model was built, incorporating mortgage-related and macro-economic features.

Key model driver

Outstanding balance at observation time, Interest rate both at origination and observation time, House price index at the time of observation and origination and HPI percentage change, Gross domestic product (GDP) growth at observation time (in %) , Unemployment rate at observation time (in %) , Investor borrower = 1, otherwise = 0 , FICO score at origination time (in %) , Loan-to-value ratio at observation time (in %) , Time on Book, Percentage of Remaining Balance

Main relationship between risk drivers and model prediction

Based on the estimates, an increase in total balance and Loan to value at observation time, house price index all at the time of origination, interest rate at time of observation and house price index percentage change are associated with an increase in the likelihood of default probability. Conversely, an increase in house prices index, GDP, unemployment rate all at time of observation and FICO score, interest rate at the origination, time on book, and remaining balance percentage change corresponds to a decrease in the likelihood of default.

Summary Document XGBoost

Model Name: XGBoost

Description

XGBoost, an algorithm for structured data, utilizes an ensemble of decision trees to correct errors iteratively. It efficiently handles imbalanced data and missing values. Default hyperparameters were used, and the model mirrored logistic regression features for comparative analysis. The selected booster for this model was decision tree and the number of iteration was set to 350.

Key model driver

Outstanding balance at observation time, Interest rate both at origination and observation time, House price index at the time of observation and origination and HPI percentage change, Gross domestic product (GDP) growth at observation time (in %) , Unemployment rate at observation time (in %) , Investor borrower = 1, otherwise = 0 , FICO score at origination time (in %) , Loan-to-value ratio at observation time (in %) , Time on Book, Percentage of Remaining Balance

Main relationship between risk drivers and model prediction

Based on the estimation results, an increase in total balance and Loan to value at observation time, house price index all at the time of origination, interest rate at time of observation and house price index percentage change are associated with an increase in the likelihood of default probability. Conversely, an increase in house prices index, GDP, unemployment rate all at time of observation and FICO score, interest rate at the origination, time on book, and remaining balance percentage change corresponds to a decrease in the likelihood of default.

Summary Document Neural Network

Model Name:

Artificial Neural Network

Description

Neural networks weight each data element individually. This step is performed several times at the same time and thus generates differently weighted data combinations. These data combinations are also weighted and converted by mathematical transformations to a score, which can be interpreted as a kind of probability indicating how likely the target value (the credit default) will be reached. The analysis of the weights and how the weights were determined, allows a detailed understanding of the relevant factors for each individual credit default prediction.

Key model driver:

Outstanding balance at observation time, Loan-to-value ratio at observation time (in %), FICO score at origination time (in %) , Interest rate at origination time (in %), House price index at observation time (base year=100)

Main relationship between risk drivers and model prediction:

Variables Associated with Likelihood Increase of Default: Outstanding balance at observation time, Loan-to-value ratio at observation time (in %)Interest Rate at origination time (in %), House price index at observation time (base year=100)

Decrease in the value FICO score at origination time (in %) lead to an increase in default probability.

Glossary

Accuracy: Accuracy is the proportion of correct predictions to the total number of predictions. Formula:

$$(Number\ of\ Correct\ Predictions)\ /\ (Total\ Number\ of\ Predictions)$$

AUC: AUC (Area Under the Curve) is a metric that measures the performance of a Machine Learning model on a binary classification task by evaluating the area under the Receiver Operating Characteristic (ROC) curve. In AUC, the "classes" refer to the two categories in a binary classification: the positive class represents what the model should predict, while the negative class represents the opposite or lack thereof. The AUC value indicates how well the model can distinguish between these classes, i.e. how well it is able to correctly recognise positives and negatives. A higher AUC value (closer to 1) means better discrimination between the classes.

Akaike Information Criterion: The AIC is a metric used for model selection, especially in the context of Linear Regression and other statistical models. It balances the trade-off between the goodness of fit of the model and the complexity. The AIC value quantifies the quality of a model by considering how well it fits the data while penalising the number of parameters in the model. Lower AIC values indicate a better trade-off between model fit and simplicity, so a model with the lowest AIC value is preferable in relative terms.

F1: The F1 score combines the model's precision and recall into a single metric, providing a balance between both measures.

GINI: GINI measures the inequality between the values of a frequency distribution. It is used, for example, in conjunction with Decision Trees to assess the impurity or homogeneity of a data set.

P-Value: The p-value is a statistical measure used to determine the strength of evidence against a null hypothesis.

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Formula:

$$(True\ Positives)\ /\ (True\ Positives\ +\ False\ Positives)$$

Recall: Recall measures the proportion of actual positives that were correctly identified by a model. Formula:

$$(True\ Positives)\ /\ (True\ Positives\ +\ False\ Negatives)$$

ROC: (Receiver Operating Characteristic): The ROC curve is a graph showing the performance of a classification model at different classification thresholds. It plots the rate of true positives (sensitivity) against the rate of false positives (1 - specificity).

True Positives: True positives are the number of correctly predicted positive instances by a model.

True Negatives: True negatives are the number of correctly predicted negative instances by a model.

False Positive: A false positive occurs when a model incorrectly predicts a positive outcome when the actual result is negative.

False Negative: A false negative occurs when a model incorrectly predicts a negative outcome when the actual result is positive.

References

[EBA2023]www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf

[EBA]www.eba.europa.eu/sites/default/documents/files/document_library/About%20Us/Annual%20Reports/2021/1035237/EBA%202021%20Annual%20Report.pdf

[CFA19] Carmona, Pedro, Francisco Climent, and Alexandre Momparler. "Predicting failure in the US banking sector: An extreme gradient boosting approach." International Review of Economics & Finance 61 (2019).

[C15] Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." R package version 0.4-2 1.4 (2015).

[LL17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[LYG15] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[M02] Menard, Scott. Applied Logistic Regression analysis. No. 106. Sage, 2002.

[OL21] Ojha, V., Lee, J. Default analysis in mortgage risk with conventional and deep Machine Learning focusing on 2008–2009. Digit Finance 3, 249–271 (2021).

Disclaimer on Data Usage

The use and processing of the data in this study was carried out independently of a commercial project. The data processing serves the sole purpose of highlighting challenges and potentials in the application of Machine Learning, considering the current guidelines of the European Banking Authority (EBA) on the use of Machine Learning.

Contact

Grant Thornton Quantitative Risk team and SEKASA Technologies have teamed up for an investigation to unravel the latest advances in default risk assessment as presented in this report. Feel free to contact us using the below information.

Andreas Spyrides
Director, Head of Quantitative Risk Cyprus
T +357 22 600 270
E andreas.spyrides@cy.gt.com

Sebastian Niehaus
Co-Founder and CTO – SEKASA Technologies
T +357 95 501 007
E sn@sekasa-technologies.com

Katharina Brunkhorst
Founder and CEO – SEKASA Technologies
T +357 95 712 140
E kb@sekasa-technologies.com

Lukas Majer
Director, Head of Quantitative Risk Spain
T +353 (0)1 646 9006
E lukas.majer@ie.gt.com

Maria Yiasouma
Assistant Manager, Quantitative Risk Cyprus
T +357 22 600 161
E maria.yiasouma@cy.gt.com

Kimia Mirsalehi
Assistant Consultant, Quantitative Risk Cyprus
T +357 22 600 172
E kimia.mirsalehi@cy.gt.com

Grant Thornton Offices in Cyprus, United Kingdom, Ireland, and Spain

linkedin.com/company/**grantthorntoncy**
twitter.com/**grantthorntoncy**
linkedin.com/company/sekasa-technologies-ltd
facebook.com/**grantthorntoncyprus**
instagram.com/**grantthorntoncy**
sekasa-technologies.com

grantthornton.com.cy