Enhancing Probability of Default Prediction

Non-Linear Modeling in Turbulent Economic Times

Victor Stern

Master's thesis
2024:E44

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Enhancing Probability of Default Prediction: Non-Linear Modeling in Turbulent Economic Times

Victor Stern

University supervisor:
Prof. Erik Lindström

Company supervisors:
Christoph Neuner & Felix Gustavsson, Nordea

Examiner:
Prof. Magnus Wiktorsson

LUND
UNIVERSITY

Centre for Mathematical Sciences
Lund University

June 5, 2024

Abstract

This thesis investigates the application of spline regression models to predict the Probability of Default (PD) under varying macroeconomic conditions, exploring whether these models can enhance predictive accuracy over traditional linear models and compare favorably to XGBoost. The study analyses the non-linear dynamics between PD and key macroeconomic indicators within a Swedish small-sized corporate loan portfolio from 2008 to 2022. Spline models, particularly cubic splines, are compared against linear models and XGBoost in terms of predictive performance.

The results indicate that while spline models show potential in capturing complex non-linear relationships, their performance in out-of-time validation does not consistently surpass that of linear models. However, spline models provide a more nuanced understanding of the interactions between PD and macroeconomic variables, which could be crucial during turbulent economic periods. XGBoost demonstrated superior accuracy and generalization capabilities, particularly in handling diverse macroeconomic conditions without predicting excessively high PD values.

Based on these findings, spline models can serve as an effective intermediary, balancing the interpretability of linear models with the higher accuracy of XGBoost. This work contributes to the credit risk modeling discourse, particularly within the IFRS 9 framework, suggesting that incorporating non-linear modeling techniques such as splines could offer a more flexible and potentially more accurate approach to PD prediction.

Keywords: Probability of Default (PD), Linear regression, Splines, XGBoost, Macroeconomic indicators, Credit risk modeling

Acknowledgements

I am profoundly grateful to my supervisors at Nordea, Christoph Neuner and Felix Gustavsson, for their consistent support and for serving as sounding boards throughout the course of this thesis work. Your willingness to discuss, refine ideas and provide constructive feedback has been crucial to my research.

Equally, my deepest appreciation goes to my academic supervisor at Lund University, Professor Erik Lindström. Your expert advice and directional guidance have been very helpful. It is with sincere gratitude that I acknowledge your role in making this thesis possible.

Contents

Chapter 1

Introduction

1.1 Background & Motivation

One of the main concerns of banks and other financial institutions is to guarantee financial stability and prevent events such as the financial crisis of 2008. To prevent these types of crashes, it is imperative that banks manage their credit risk - the risk that the bank's obligors fail to repay their loans. Due to the risk of a customer defaulting, the bank has a regulatory requirement to hold a certain amount of capital to cover this potential future loss. The size of this capital allocation is decided by modeling the Expected Credit Loss (ECL) for the bank's loan portfolios. One of the key factors that affect ECL is the Probability of Default (PD), which is the likelihood that a customer fails to make its scheduled repayment on a debt. Accurate modeling of PD is important to ensure sufficient and effective estimation of regulatory capital.

As a response to the 2008 financial crisis, the International Financial Reporting Standard 9 (IFRS 9) was introduced. The crisis highlighted significant deficiencies in the previous standard, IAS 39, particularly in the way that financial instruments were reported and impairment losses on financial assets were recognized. One of the key motivations for the introduction of IFRS 9 was the need for a forward-looking "expected loss" impairment model as opposed to the "incurred loss" model under IAS 39, which was criticized for delaying the recognition of credit losses (Frykström and Li, 2018). The expected loss model under IFRS 9 requires banks and other institutions to account for expected credit losses from when financial instruments are first recognized, and to update the amount of expected credit losses recognized at each reporting date to reflect changes in the credit risk of the financial instruments. This approach aims to address the issue of "too little, too late" provisions for loan losses and to provide more timely information about expected credit losses.

The method most widely used by financial institutions to predict PD is logistic and linear regression, where both systematic risks such as macroeconomic conditions as well as idiosyncratic risks, i.e. client specific risks are assumed to drive the default probability. In the banking sector, the calculation of PD is approached in three distinct manners: through-the-cycle (TTC), point-in-time (PIT), and a hybrid of both. The TTC methodology relies on individual risk factors, thereby remaining unaffected by prevailing economic conditions. Conversely, under the IFRS 9 framework, the PIT model modifies this stable PD to reflect more immediate circumstances, incorporating adjustments for economic downturns or growth phases, as illustrated in Figure 1.1. The IFRS 9 models use relevant macro variables such as gross domestic product (GDP), unemployment rate, house prices, interest rates and commodity prices, which will be the focus of this thesis (Frykström

and Li, 2018).

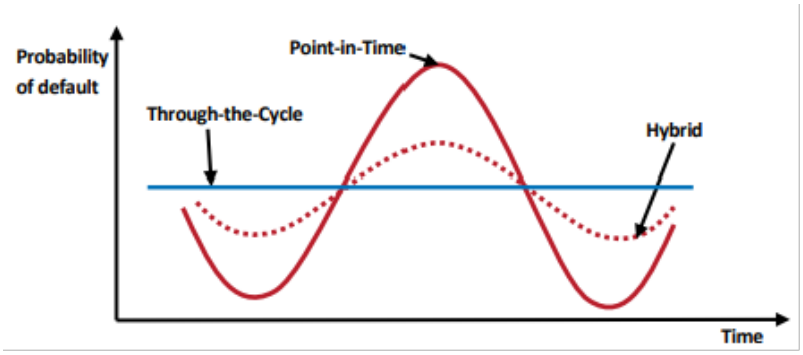


Figure 1.1: TTC vs PIT probability of default. From Frykström and Li (2018).

During extreme events, such as the Covid-19 pandemic, the statistical correlations between macroeconomic risk drivers and PD estimates, typically assumed to hold under normal conditions, may not remain valid. This results in erroneous predictions of the probability of default and in turn ECL estimates. To be able to handle such events in the future, for example in the case of another pandemic, capturing the changing dynamics between PD and its risk drivers under turbulent macroeconomic conditions are of key importance to ensure accurate loss predictions for banks. Since one of the assumptions of logistic regression is that there is a linear relation between log-odds of PD and the independent variables, other approaches need to be considered to be able to model different correlations under different states of the economy.

The problem with non-linearity has been highlighted by the European Banking Authority (2023). This authority, which monitors European banks' implementation of IFRS 9, states that the effect of non-linearity on the ECL estimates in 2021 is limited, which raises concerns that adverse macroeconomic scenarios are not taken into account. Thus, the ultimate ECL values might not comprehensively account for the uncertainties inherent in various macroeconomic projections and may fail to accurately represent the non-linear relationship between macroeconomic factors and the final ECL values. This indicates that there is a risk of inaccurately representing potential losses if the macroeconomic conditions significantly diverge from the initial baseline assumptions. Particularly in situations of high macroeconomic uncertainty, it is essential for financial institutions to properly incorporate the effects of non-linearity into their ECL calculations. Additionally, they express that there is a necessity for regulatory authorities to intensify their examination of these institutions' methodologies in reflecting such non-linear impacts (European Banking Authority, 2023).

Spline regression is a popular approach for modeling non-linear relations. Spline methods employ piecewise polynomial functions to model distinct segments of data. This approach segments the dataset into intervals, fitting each with its own polynomial function. These polynomials are then seamlessly joined at their boundaries, known as knots, creating a smooth overall function. This technique allows for a flexible and accurate representation of data, particularly useful in scenarios requiring interpolation or smoothing across diverse data trends (Brooks, 2019). A stylistic example of a simple first order spline regression can be seen in Figure 1.2.

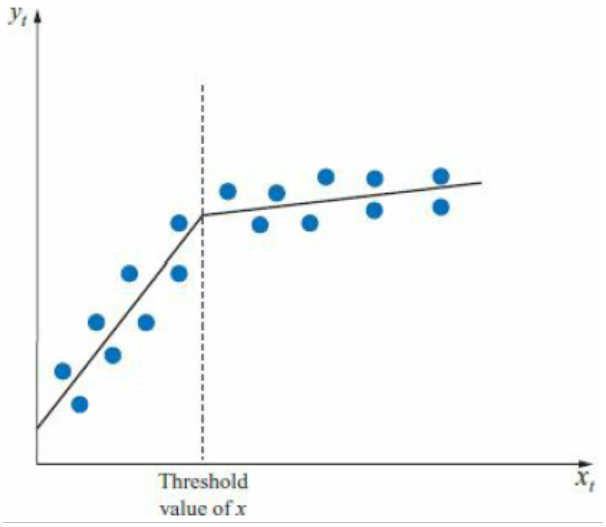


Figure 1.2: Example of a simple, linear spline model fitted to data showing a regime shift. From Brooks (2019, p. 587).

This thesis primarily investigates the application of spline techniques to model the non-linear relationships between PD and various macroeconomic factors, assessing whether these models offer superior predictive accuracy over traditional linear models. Additionally, machine learning methods such as XGBoost have been shown to outperform traditional linear credit risk models (Hild, 2021). Therefore, XGBoost will be considered for comparison to evaluate whether spline techniques can match the predictive accuracy of these machine learning models. This comparison will provide a broader perspective on non-linear modeling techniques and their effectiveness in capturing the dynamics of PD in response to macroeconomic changes.

1.2 Problem Statement

This thesis will address the following research questions:

- Can spline regression offer a more accurate model than an ordinary linear fit for the possibly non-linear dynamics between different macroeconomic variables and portfolio PD?
- Can spline models achieve predictive accuracy comparable to that of XGBoost in PD prediction?

1.3 Delimitations

The scope of this thesis is specifically focused on data obtained from a Swedish small-sized corporate loan portfolio. Consequently, the findings are primarily applicable to the Swedish market and entities that are characteristic of small corporate structures. It is also important to note that the research will concentrate exclusively on modeling of the default probability. PD is just one component of the ECL calculation, which also encompasses other critical parameters such as loss given default (LGD) and exposure at default (EAD). The investigation of LGD and EAD falls outside the ambit of this study. Further, the data used in the analysis is limited to the years 2008 - 2022.

1.4 Related Work

Several previous works have explored the area of PD modeling based on macroeconomic variables. Antonsson (2018) analysed the significance and effect of GDP, house price index, repo rate, and unemployment rate on default frequency in a Swedish retail credit portfolio between 2008 - 2015. The method used was multiple linear regression with ordinary least squares (OLS) to fit the predictor coefficients. Further, several time-lagged values of the mentioned variables were tested from a monthly lag of 1 to 13. The credit portfolio was segmented into three risk classes (low, medium, high) and a model was fitted to each segment. The work concluded that different lagged values of GDP and repo rate were the only statistically significant macroeconomic variables in explaining the variance of default frequency for the analysed Swedish retail credit portfolio.

In their investigation, Hild (2021) assessed the performance of various statistical methodologies for default classification (default or no default) within US mortgage loan portfolios, employing quarterly data between 2001 and 2015. Thus, the dependent variable was defined as the likelihood of a borrower defaulting within a given quarter (3-month PD), incorporating both macroeconomic and borrower-specific variables. The study compared traditional statistical approaches, including logistic regression (both with and without LASSO for variable selection) and linear discriminant analysis, against more contemporary machine learning techniques such as LightGBM and XGBoost. The comparative analysis was based on metrics such as the area under the ROC curve (AUC), Brier score, and the absolute error in the predicted PD. Their findings revealed that, generally, machine learning models surpassed traditional methods in performance, albeit at the expense of reduced interpretability of the model parameters. This motivates the scope of my thesis, as the focus of this study is on traditional, easily interpreted and explained, models - while incorporating flexible methodologies like splines for enhanced adaptability and also comparing its performance with similar machine learning methodologies.

Ali and Daly (2010) applied logistic regression to model PD with macroeconomic explanatory variables on both a US and an Australian portfolio with quarterly data from 1995 to 2009. They found that GDP, total debt-to-GDP ratio as well as short-term interest rates statistically significantly explained portfolio default rates in both countries. This provides useful input for the work of my thesis, by motivating the use of these macroeconomic variables for explaining default rates.

In their study, S. Li et al. (2022) present a novel approach to estimating corporate PD through a single-index hazard model utilizing penalized spline (P-spline) techniques, demonstrating superior predictive accuracy over traditional models. This methodology's effectiveness, particularly during the financial crisis of 2008, underscores the potential of spline regression to address the complex, non-linear dynamics between financial indicators and default probability. This research serves as a foundational reference for my thesis, where I aim to explore spline regression's applicability to portfolio-level PD modeling in relation to macroeconomic factors.

Bellotti and Crook (2012) analysed UK retail credit card data from 1999 to 2005, modeling LGD using both macroeconomic and account-level variables. While their period of analysis notably lacked major economic downturns, Bellotti and Crook (2012) established that macroeconomic variables such as bank interest rates and unemployment levels improved the model fit and were statistically significant in predicting LGD. Their study supports the inclusion of these variables in my thesis, as one can expect LGD and PD to be driven by similar macroeconomic factors.

This thesis builds upon previous research, such as the work by Antonsson (2018) focused on linear correlations between PD and macroeconomic variables, by applying non-linear techniques, such as those used by S. Li et al. (2022), to more accurately forecast default probabilities in unpredictable economic environments. This method not only builds on established findings but is also specifically tailored to the Swedish small corporates sector, aiming to enhance credit risk management approaches.

1.5 Outline

This theis is structured as follows: Chapter 2 describes the data used for modeling. Chapter 3 describes the methods used for data processing, as well as statistical and econometric theory related to the modeling. Chapter 4 presents the empirical results, model performance and evaluation. Chapter 5 discusses the results and its eventual drawbacks and limitations. Chapter 6 concludes key takeaways from the data analysis and modeling as well as gives proposals of possible further research.

Chapter 2

Data

The default data used in this thesis was provided by Nordea, and consists of the number of performing customers as well as the number of these customers that default within the next 12 month period, for a specific segment, at a given date. As mentioned earlier, the segment analysed is small corporates in Sweden. From this, the 12-month observed default rate (ODR) was calculated as the percentage-wise proportion of customers who defaulted out of all performing customers at each given month. This 12-month ODR is used as a proxy for 12-month PD and is thus the target variable in the modeling. The data is provided on a monthly frequency between the dates 2007-12-31 and 2022-06-30 which results in 175 data points. The period under study covers a variety of economic conditions, including the financial crisis of 2008, the Covid-19 pandemic, as well as relatively stable phases in the intervening years. Figure 2.1 illustrates the observed default rate over the entire time period. Note the pronounced shift around 2010, where there is a significant spike in the default rate over approximately 12 months, followed by a subsequent drop. This phase likely exerts a considerable influence on the models' estimations. Investigations did not yield obvious causes and no remedial actions was taken. Following the surge around 2010, the default rate exhibits a gradual decline until it reaches a pronounced peak starting in 2018. This is succeeded by a significant downturn, which coincides with the ending years of the pandemic. The Y-axis has been omitted and the data masked to maintain confidentiality. Consequently, the data shown here and in Chapter 4 has been altered from its original form.

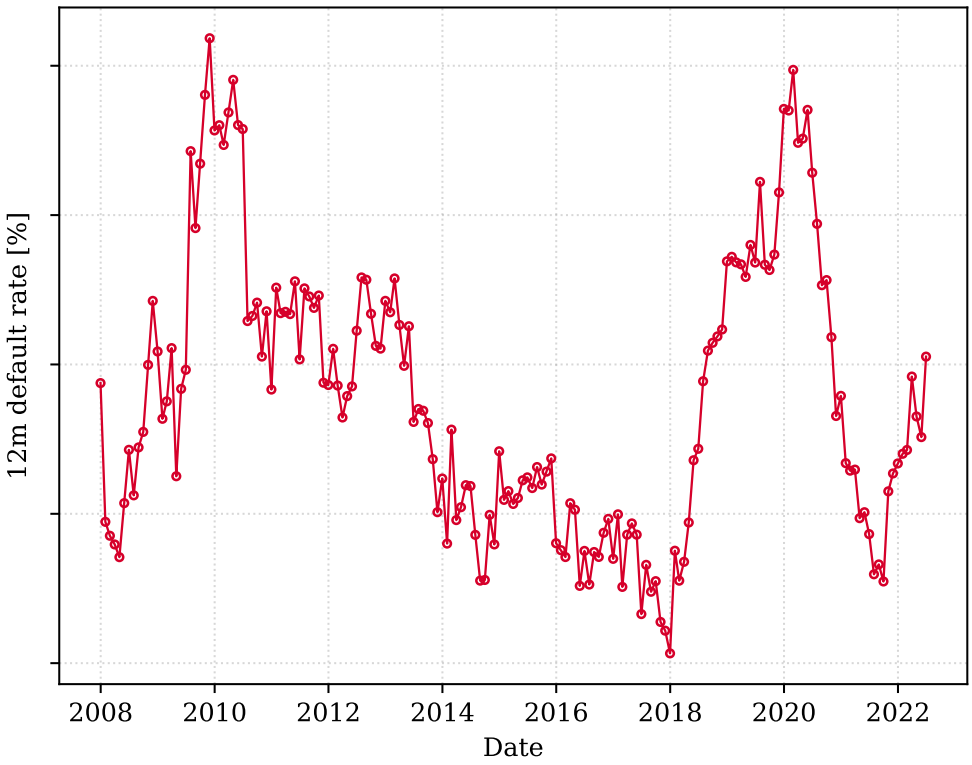


Figure 2.1: Trajectory of the observed default rate throughout the time period under study. Y-axis omitted and data is masked for confidentiality reasons.

The methodology used to calculate defaults over a 12-month horizon on a monthly basis inherently introduces autocrrelation in the default rate due to the significant overlap in sequential months. Despite this, the 12-month ODR is used as the dependent variable as this is in line with standard banking practices, where a 12-month PD is commonly used for credit risk assessment. To mitigate the impact of autocorrelation and ensure the robustness of the linear regression analysis, heteroskedasticity and autocorrelation consistent (HAC) estimators, such as those proposed by Newey and West (1987), are employed. This approach corrects for the presence of autocorrelation and potential heteroskedasticity in the error terms, providing more reliable standard error estimates.

The macroeconomic data was gathered from different public sources such as Sveriges Riksbank and Statistiska Centralbyrån (SCB). The selection of macroeconomic variables is based on the results of the statistical significance in explaining PD found in related works as well as on economic intuition. A total of 12 macroeconomic variables were considered in this study. A summary of the variables is seen in Table 2.1 below.

Table 2.1: Overview of macroeconomic variables

Variable (code)	**Description**	**Unit**	**Source**	**Frequency**
Repo Rate	Average monthly interest rate	%	Sveriges Riksbank	Monthly
UR	Unemployment rate, 15-74 years. Seasonally adjusted and smoothed	%	SCB	Monthly
FX_Rate	Exchange rate (USD/SEK). Monthly average	SEK per USD	Sveriges Riksbank	Monthly
GDP growth (YoY)	Year-over-year economic growth rate	%	SCB	Quarterly
CPIF	Consumer price index with fixed interest rate (1987=100)	Index	SCB	Monthly
OMXSPI	OMX Stockholm price index, end of period (29/12/95=100)	Index	Nasdaq Nordic	Quarterly
LT_Rate	10-year government bond yield, end of period	%	MarketWatch	Quarterly
ST_Rate	3-month treasury bill yield, average	%	Sveriges Riksbank	Quarterly
ELEC	Electricity producer price index (2019=100)	Index	Eurostat	Quarterly
GAS	Price of oil, gasoline	SEK per tonnes of oil equivalent	IEA	Quarterly
GC	Government consumption, real	MSEK	SCB / Trading Economics	Quarterly

Note: All variables pertain to the Swedish economy.

Since many of the variables are acquired at a quarterly frequency, this data must be imputed to provide a consistent monthly time series, which is described in section 3.1.1.

Chapter 3

Theory & Methodology

This chapter explains the theory and methods used in the analysis, covering everything from how the data was prepared to the techniques for choosing variables, fitting the models, and evaluating their performance. It outlines the key steps taken to ensure that the analysis is solid and the results are trustworthy.

3.1 Data Handling

Before constructing the model for PD as a function of the macroeconomic variables, the raw data was analysed to avoid missing or irregular values, and to ensure it aligned with the assumptions of the regression techniques used.

3.1.1 Imputation

For the macroeconomic variables used in this thesis, the frequency at which the variables were observed differed between monthly and quarterly frequency. To avoid a major loss of information leading to less powerful estimates in the models, the choice was made to disaggregate the quarterly time series to a monthly frequency. The benefits of disaggregation was weighed against the potential drawbacks of modeling on interpolated data - which is not common practice at financial institutions like Nordea. However, as the purpose of this thesis is more a proof of concept rather than to produce a regulatory compliant model ready for use, the pros of more data were considered greater than the cons of lesser data quality. In this thesis, due to it's restrictive nature, the quarterly macroeconomic series were converted using linear interpolation.

However, for further research, more sophisticated methods are available for such conversions, as demonstrated by the works of Cuche and Hess (1999) and Chow and Lin (1971). Specifically, Chow and Lin (1971) introduced a statistical technique that employs regression analysis on related higher-frequency indicators to interpolate lower-frequency GDP data into more granular monthly data. This method not only enhances the accuracy of the interpolated series but also aligns closely with the underlying economic trends on a monthly basis. Building on the advancements of Chow and Lin (1971), Cuche and Hess (1999) further refined the methodology for estimating monthly GDP figures by employing a general Kalman filter framework. This approach, demonstrated through their work on Swiss data from 1980 to 1998, incorporates related series and addresses the challenges of non-stationarity within the data. By leveraging the flexibility and dynamic updating capabilities

of the Kalman Filter, they provide a robust model that significantly improves the interpolation of lower-frequency GDP data.

3.1.2 Lagged macroeconomic variables

As stated by Bellotti and Crook (2012), the dependence between a macroeconomic variable and credit risk may not be immediate. For example, a change in interest rate may affect an obligors payment capacity several months later. Therefore, lagged versions of the macroeconomic variables with a lag from 1 to 12 months were included in the modeling.

3.1.3 Exploratory data analysis

To analyse the characteristics of the raw data, an exploratory data analysis (EDA) was conducted. EDA was first introduced by Tukey (1977), and consists of visually and graphically presenting the data with for example scatter plots to spot missing data, outliers and overall trends and distributions of the dataset. Since the data was distributed over a number of years, the raw data was visualized over time using line plots with markers and analysed for outliers and trends. In the default dataset, used to calculate the ODR, some outliers in the number of performing customers were found. In one of the data points, there was a decrease of 2,000 performing customers, which unexpectedly returned to normal levels in the following month. The disappearance and quick return of such a large number of customers was considered unreasonable, indicating a potential fault in the data due to reasons not explored further. To correct this, the specific data point was replaced with the mean of the number of performing customers in the two surrounding months.

To investigate the relationship between the default rate and the selected macroeconomic variables, scatter plots with regression lines were utilized, using the `seaborn` library's `pairplot` function. This analysis aimed to uncover potential non-linear relationships among the variables, suggesting that splines and generalized additive models (GAMs) might offer a more accurate depiction of these dynamics compared to traditional linear models. This approach allowed for a quick overview of how macroeconomic factors interact with the probability of default, highlighting the complexity of these relationships and the importance of adopting flexible modeling techniques to capture them effectively.

3.1.4 Stationarity tests

A stationary process is characterized by a constant mean, variance and auto-correlation structure (Brooks, 2019). Therefore, a stationary process does not show any obvious trends over time and crosses the mean frequently. A key requirement for modeling time series data is that all variables - both dependent and independent - should exhibit stationarity. If the data is not stationary, it can result in misleading models indicating a strong relationship between variables, with a high R^2 and significant coefficients, even when there is no actual connection between the variables. Using stationary variables ensures that the relationships modeled are consistent over time. Consequently, the macroeconomic variables and the observed default rates were evaluated for stationary behaviour before they were used in the modeling.

One of the most widely used test for stationarity is the Augmented Dickey-Fuller (ADF) test. The objective of the test is to test the null hypothesis $H_0 : \psi = 0$ versus $H_1 : \psi < 0$ in

$$\Delta y_t = \psi y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + u_t \tag{3.1}$$

where $\Delta y_t = y_t - y_{t-1}$ and u_t is the white noise zero mean, constant variance error term. The inclusion of p lagged differences of the dependent variable, Δy_t, serves to absorb any temporal dependencies or patterns in the dependent variable (Brooks, 2019). This is done to make certain that the error term u_t does not exhibit autocorrelation. If the null hypothesis H_0 is rejected, the series does not contain a unit root, providing evidence that the series is stationary. The test statistic is defined as

$$DF_\tau = \frac{\hat{\psi}}{SE(\hat{\psi})} \tag{3.2}$$

which is the estimate from (3.1) divided by its standard deviation. This statistic is compared to critical values for the test distribution, and the null hypothesis is rejected if it is less than the critical value at a given significance level. The p-value in the ADF test measures the likelihood of seeing our test results if the null hypothesis were actually true. If this p-value is below a threshold, such as 5%, we have sufficient evidence to reject the null hypothesis and conclude that the data is stationary. For this thesis, the ADF test was conducted in Python with `adfuller` function from the `Statsmodels` package.

Another test commonly used to test stationarity of a time series is the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski et al., 1992). While the ADF test focuses on detecting a unit root as evidence of non-stationarity, the KPSS test approaches the problem from the opposite direction. Its null hypothesis, H_0, posits that the series is stationary around a deterministic trend or level, against the alternative hypothesis, H_1, that the series is non-stationary due to a unit root. The test assumes the series can be decomposed to a deterministic trend, a random walk and a stationary error as:

$$y_t = \xi t + r_t + \varepsilon_t \tag{3.3}$$

where r_t is the random walk defined as

$$r_t = r_{t-1} + u_t \tag{3.4}$$

with iid errors $u_t \sim \mathcal{N}(0, \sigma_u^2)$ and intercept r_0. The null hypothesis is simply $\sigma_u^2 = 0$. If ξ is set to zero, the test is done on level stationarity instead of around a trend. The KPSS test statistic is defined as follows:

$$\eta = \frac{\sum_{t=1}^{T} S_t^2}{T^2 s^2(l)} \tag{3.5}$$

where S_t is the partial sum of deviations from the sample mean, T is the number of observations, and $s^2(l)$ is an estimate of the long-run variance of y_t. The test statistic thus assesses the level of stationarity of the series by examining the severity of its deviation from a deterministic trend or level, rejecting the null hypothesis of stationarity for large values. For this thesis, the KPSS test was conducted using the *kpss* function from the `Statsmodels` package in Python.

In their study, Schlitzer (1995) concluded that the individual power of the ADF and KPSS test are low, especially for small samples sizes. However, they show that a combined ADF-KPSS approach can be used to reduce the number of erroneous conclusions. Therefore, the KPSS test is employed

in tandem with the ADF test to provide a robust framework for determining stationarity. While the ADF test identifies the absence of a unit root to suggest stationarity, the KPSS test verifies that the series does not exhibit structural changes or a stochastic trend. As underscored by Brooks (2019), the time series should be concluded stationary from both tests for the results to be robust. Therefore, in this study, the variables that did not pass both tests were considered non-stationary.

Variables that proved to be non-stationary from the joint ADF and KPSS tests, must be transformed to become stationary before being used in the modeling. There were mainly two methods used to transform the variables. For macroeconomic variables expressed as a percentage (such as interest rates), differencing from last quarter was calculated as

$$x_{\text{diff,t}} = x_t - x_{t-3} \tag{3.6}$$

and for the other variables (such as indices or prices) the growth rate from last quarter was calculated as

$$x_{\text{growth,t}} = \left(\frac{x_t - x_{t-3}}{x_{t-3}} \right) \times 100 \tag{3.7}$$

where x_t is the value in the current month and x_{t-3} is the value from three months prior.

Despite the tests suggesting non-stationarity in the 12-month observed default rate series, the decision was made to use it as the dependent variable in the modeling without employing differencing transformations to achieve stationarity. This strategy was chosen based on that models aimed at predicting changes in PD would demonstrate weak accuracy in the long-term predictions. Additionally, given the default rate's intrinsic constraint within the 0 to 1 range and grounded in economic logic, it is logical to infer that the PD will maintain long-term stationarity, as it cannot indefinitely increase. However, going forward, it is acknowledged that the development of a model is based on a response variable exhibiting potential non-stationary behaviour, which might not fully align with the assumptions of regression analysis.

3.1.5 Train-test split

The macroeconomic data and observed default rates exhibited atypical patterns starting from 2020, likely due to the Covid-19 pandemic. In this period, the macroeconomic conditions were worsening, yet contrary to expectations of a surge in the default rate, it unexpectedly declined. To accommodate this anomaly in the PD model training, the dataset was partitioned into three distinct periods. Data from 2008 to 2018 was allocated as a training set to capture the pre-pandemic economic conditions. For testing, two distinct subsets were created: the first, spanning from 2018 to 2020, encapsulates a period considered to reflect relatively stable conditions. The second test set, ranging from 2020 to mid-2022, encompasses the timeframe where the pandemic's impact was most pronounced, resulting in irregular data trends. This division allows for a more nuanced evaluation of the PD model's robustness across varying parts of the economic cycle.

As explained in section 3.3.2, cross-validation in the training set is also employed which is a more sophisticated version of train-test split used for model selection. However, this "in-sample" and "out-of-sample" data split is used in tandem with cross-validation to evaluate the performance of the chosen models on completely unseen data.

3.2 Modeling Theory

This sections explains the different theories and methods used for creating linear and non-linear models for the default probability.

3.2.1 Multiple linear regression on transformed PD

Multiple linear regression allows us to express the relationship between the explained variable y_t and the explanatory variables x_{kt} as:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \epsilon_t, \tag{3.8}$$

which can be written in matrix form as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1T} & x_{2T} & \ldots & x_{kT} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}$$

In this study, $\mathbf{y}$ contains the historical default rates, $\mathbf{X}$ the k different explanatory macroeconomic variables, $\boldsymbol{\beta}$ their corresponding coefficients as well as $\boldsymbol{\varepsilon}$ the residuals. The linear regression model makes the following key assumptions (Brooks, 2019):

- **Linearity:** The relationship between the dependent and independent variables is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedastic and normal errors:** The error terms are assumed to be independent and identically distributed, following a normal distribution with zero mean and constant variance across all levels of the independent variables.
- **No multicollinearity:** None of the independent variables should be highly correlated with any of the other independent variables.

To ensure that the output of the model is bounded between zero and one, since we want to predict a probability, we transform PD into log odds, as this operation enables us to maintain the unbounded characteristics necessary for linear regression. The transformation is defined as $y = \text{logit}(\text{PD}) = \ln\left(\frac{PD}{1-PD}\right) = \mathbf{X}\boldsymbol{\beta}$, which re-expresses the PD across the entire real number line. When the predictions are subsequently re-transformed, they revert to being confined within the interval $[0, 1]$. Note that log refers to the natural logarithm here. Therefore, the coefficients, $\beta_1, \beta_2, \ldots, \beta_k$, are interpreted as the change in the log odds of PD for a unit change in the corresponding macroeconomic variable, holding all other factors constant. The PD is thus calculated as:

$$P(\text{default}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}} \tag{3.9}$$

To find the coefficient estimates, $\hat{\boldsymbol{\beta}}$, ordinary least squares (OLS) estimates are used (Brooks, 2019). This is done by minimizing the residual sum of squares,

$$\mathbf{L} = \hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^T\mathbf{y} - 2\hat{\boldsymbol{\beta}}^T\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \tag{3.10}$$

which after differentiated w.r.t. $\hat{\boldsymbol{\beta}}$ and set to zero yields the coefficient estimates:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.11}$$

The significance of the parameter estimates are then tested using the traditional t-test or z-test, utilizing the standard errors of the coefficient estimates. The heteroskedasticity and autocorrelation consistent (HAC) standard errors, as introduced by White (1980) and further developed by Newey and West (1987), provide a way to adjust the variance-covariance matrix of the parameter estimates to account for possible heteroskedasticity and autocorrelation in the error terms. By employing these robust standard errors, hypothesis testing remains valid even when the classical assumptions of homoskedasticity and no autocorrelation are violated. Since the analysed default dataset is noted to have autocorrelation, HAC estimators are employed for the linear regression models in this thesis.

3.2.2 Univariate smoothing splines

To create a more flexible model, able to capture a non-linear dependence between the target and predictor variables, first let's consider the univariate model

$$y_i = f(x_i) + \epsilon_i, \tag{3.12}$$

where y_i represents the dependent variable, x_i the independent variable, f denotes an unknown smooth function supposed to capture the relationship between the two, and ϵ_i are iid $N(0, \sigma^2)$ errors (Wood, 2017). To be able to approximate the smooth function using the same estimation methods as in linear regression, f is expressed as

$$f(x) = \sum_{j=1}^{k} b_j(x)\gamma_j, \tag{3.13}$$

where $b_j(x)$ is a basis functions in a chosen space of functions, *basis*, of which f is supposed to be an element. Thus, $b_j(x)$ with coefficients γ_j are used to construct the function that estimates f. A simple example of a basis are all polynomials up to a certain degree k. The challenge with such a basis, however, lies in its asymptotic behaviour; polynomial functions can exhibit extreme values at the bounds of the domain, leading to poor extrapolation outside the local interval - commonly referred to as Runge's phenomenon.

To mitigate this problem, one can use a spline, which is a function built up of sections of polynomials, joined together at so called knot points (Wood, 2017). These polynomials are non-zero only in a close vicinity of its corresponding knot points, thus restricting its impact in the asymptotes. A spline function of any specified degree is the result of the weighted sum of these basis functions, which are polynomials of the same degree. A naive approach would be to use a piecewise linear basis, where $b_j(x)$ are "tent"-like functions, starting at 0, increasing to 1 at the corresponding knot point, to then decrease to 0 at the next. An issue with this spline basis is that the result is a non-smooth function with discontinuities in the first derivative, since it's built of first order polynomials (such as the spline seen in Figure 1.2). A better choice of basis, as described by De Boor (1978) and Wood (2017), is the piecewise cubic basis consisting of polynomials of order 3, which still today is the most popular choice for degree of a spline basis. The third order polynomials are connected so that they not only meet but also have matching first and second derivative at

the knot points, resulting in a function that is continuous up to the second derivative. Proof show that, out of all functions f which interpolate any set of data points, this cubic spline interpolation is the smoothest in terms of minimizing

$$\text{"wiggliness"} = \int_{x_1}^{x_n} f''(x)^2 \, dx \tag{3.14}$$

There are many ways to represent cubic splines, but a popular approach is to use the so called B-spline basis, which has the sought after local property that each basis function influences solely the interval across the $m+3$ neighboring knots, where $m+1$ is the polynomial order of the basis. The spline function with k evenly spaced knots can thus be expressed as

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\gamma_i, \tag{3.15}$$

where the basis functions are defined recursively as

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x) \quad \text{for } i = 1, \ldots, k \tag{3.16}$$

and the first order basis function as

$$B_i^{-1}(x) = \begin{cases} 1, & \text{if } x_i \leq x < x_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{3.17}$$

An illustration of a smooth cubic spline function constructed from this basis is seen in Figure 3.1 below.

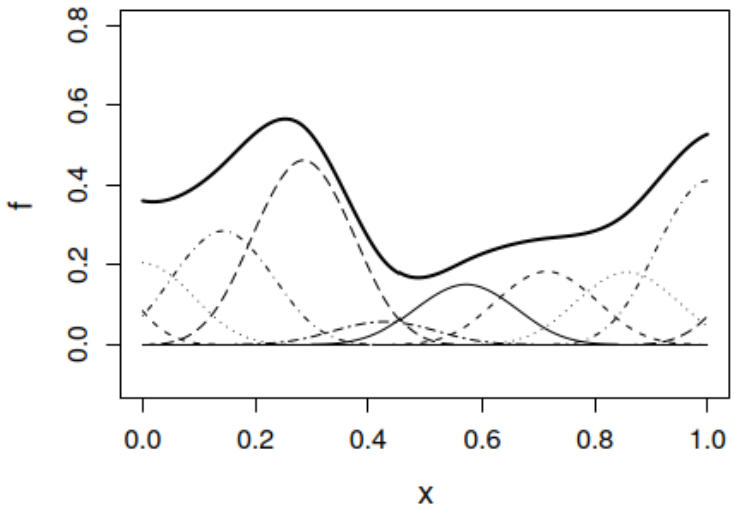


Figure 3.1: Rank 10 cubic B-spline. The thin curves show the weighted basis functions, $B_i^2(x)\gamma_i$, which summed together gives the smooth spline function, f, represented by the thick line. The knots are located where each basis function peaks. Adapted from Wood (2017, p. 205).

Substituting (3.15) into (3.12) yields a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ in terms of the basis coefficients γ_j with $\mathbf{X}$ now containing the basis functions as $\mathbf{X_t} = [1, B_1^2(x_t), B_2^2(x_t), ..., B_k^2(x_t)]$. To estimate

these coefficients and find $\hat{f}$, instead of solely minimizing the residual sum of squares as in linear regression, one minimizes the penalized least squares;

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(x)^2 \, dx \tag{3.18}$$

where λ is the non-negative smoothing parameter that balances the fit to the data against the smoothness of the spline function. A λ of zero leads to a spline that is free to pass as close as possible to each data point, potentially leading to a curve that overfits the data. As λ approaches infinity, f becomes a straight line. This hyperparameter needs to be finely tuned, often via cross-validation, to ensure a good fit to the trends of the data without capturing any noise that might be present. Assuming λ is known, using some numerical approximation of the second derivative, $\mathbf{D}$, and writing the integration as a quadratic form using the coefficients, the minimization objective of penalized least squares (3.18) can be written as;

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|^2 + \lambda \boldsymbol{\gamma}^T \mathbf{S} \boldsymbol{\gamma} \tag{3.19}$$

where $\mathbf{S} = \mathbf{D}^\mathbf{T}\mathbf{D}$. The closed form solution of the coefficient estimates can now be found:

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^T \mathbf{y} \tag{3.20}$$

Note however, that we have not chosen the dimension of the basis, k (which essentially is the number of basis functions and thus knots for each $f(x_i)$), the location of the knots nor the choice of smoothing parameter λ. Common practice is to choose k larger than thought to be necessary, and place the knots equally spaced over the independent variable, to then let λ and the penalized regression control the model flexibility and form of $\hat{f}(x)$ (Wood, 2017). Thus, in the `pyGAM` package used to implement GAMs in this thesis, the default value of k is set to 20, such that it is large enough for the spline to be able to capture the form of the true function while also maintaining computational efficiency. Although this value might seem large and prone to overfitting, considering the data used in this thesis, the penalty term will shrink the coefficients of many of the basis functions so that the final model has a lower dimension than this.

To determine the optimal λ, a common approach is to use cross-validation techniques, measuring the model's predictive performance for different λ and choosing the parameter that yields the lowest score. Ordinary cross-validation (OCV) involves sequentially leaving out each data point, fitting the model to the remaining data, and assessing the model's prediction for the omitted point. This is quantified by the OCV score:

$$V_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2$$

where $\hat{f}_i^{[-i]}$ represents the model's prediction fitted without the i-th data point, and y_i is the actual value. The objective is to select λ that minimizes this score, indicating superior predictive capabilities. However, fitting the model n times is computationally expensive. An efficient alternative to OCV is generalized cross-validation (GCV), which utilizes the influence matrix $\mathbf{A}$ and only needs to make one fit to each model (Golub et al., 1979). The GCV score is calculated as:

$$V_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[n - \mathrm{tr}(\mathbf{A})]^2}$$

where $\hat{f}_i$ is the predicted value from the model using all data points, $\text{tr}(\mathbf{A})$ is the trace of the influence (hat) matrix $A = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T$ which is used as an approximation of the degrees of freedom of the model, and n is the number of observations. Ultimately, this is an efficient approximation to the "leave-one-out" OCV. Thus, the optimal smoothing can be found through searching over a range of lambda values and choosing the one with the lowest V_g. Additionally, a more accurate approximation of the effective degrees of freedom is given by $2\text{tr}(\mathbf{A}) - \text{tr}(\mathbf{AA})$, which can lead to improved estimates of the GCV score and the smoothing parameter (Wood, 2017).

3.2.3 Additive models

To create a model where the target variable depends on several independent variables, an additive model can now be constructed, where smooth functions described in the previous section is fit to each independent variable as:

$$y_t = \alpha + f_1(x_{1t}) + f_2(x_{2t}) + ... + f_k(x_{kt}) + \epsilon_t \tag{3.21}$$

where α is a constant and $f_i(x_i)$ are cubic splines from (3.15) fit to each macroeconomic variable x_i for each data point $1 \dots t$. This can also be expressed in a linear form $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ where $\mathbf{X} = (\mathbf{1}, \mathbf{X_1}, ..., \mathbf{X_k})$ containing the basis functions for each independent variable and $\boldsymbol{\gamma}^{\boldsymbol{T}} = (\alpha, \boldsymbol{\gamma}_{\mathbf{1}}^{\boldsymbol{T}}, ..., \boldsymbol{\gamma}_{\boldsymbol{k}}^{\boldsymbol{T}})$ the corresponding basis coefficients. Consequently, the additive model can be estimated in a manner analogous to the univariate model by minimizing the penalized least squares, with one smoothing parameter per independent variable (assuming λ_i are known);

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|^2 + \sum_{i=1}^{k} \lambda_i \boldsymbol{\gamma}^T \mathbf{S}_i \boldsymbol{\gamma} \tag{3.22}$$

which results in the coefficient estimates and hat matrix:

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \left(\mathbf{X}^T\mathbf{X} + \sum_{i=1}^{k} \lambda_i \mathbf{S}_i \right)^{-1} \mathbf{X}^T\mathbf{y}, \\ \mathbf{A} &= \mathbf{X} \left(\mathbf{X}^T\mathbf{X} + \sum_{i=1}^{k} \lambda_i \mathbf{S}_i \right)^{-1} \mathbf{X}^T \end{aligned} \tag{3.23}$$

Therefore, due to the additive nature and similarity with the linear regression framework, much of the theory from multiple linear regression can be applied to additive models as well - for example confidence and prediction intervals as well as feature selection with p-values.

Furthermore, the inclusion of multiple functions within the model presents a challenge in determining unique estimates for each function, a so called "identifiability problem". Specifically, consider a model containing two functions: both f_1 and f_2 can only be estimated with an uncertainty of an additive constant. This is because the addition of a constant to f_1 and its simultaneous subtraction from f_2 would yield the same predictions from the model. Consequently, constraints must be applied to ensure the model's identifiability prior to the estimation process. A good choice is the sum-to-zero constraint (Wood, 2017):

$$\sum_{i=1}^{n} f_1(x_i) = 0 \tag{3.24}$$

This constraint solely adjusts the vertical positioning of f_1 to achieve a mean of zero, without altering its original shape or penalty value.

To determine the optimal smoothing parameters, a k-dimensional random grid search (where k is the number of macroeconomic variables) is implemented across 2000 points within the range $[10^{-3}, 10^{3}]$. This approach involves randomly selecting combinations of parameters from a uniform distribution and evaluating each using the GCV score (Servén and Brummitt, 2020). The combination yielding the lowest GCV score is selected as the optimal parameter set for the model.

I also experimented in finding the optimal smoothing parameters by implementing a time series cross-validation and testing all combinations of lambdas between $[10^{-3}, 10^{3}]$ with 30 equally spaced points and choosing the lambda combination that yielded the lowest average RMSE scores. This gave approximately the same optimal lambdas as the grid search using GCV.

3.2.4 Effective degrees of freedom

A measure of the complexity of a model is the effective degrees of freedom (EDoF), which corresponds to the effective polynomial degree of the smooth fit that is estimated by the spline function (Ventrucci and Rue, 2016). For a spline function, such as the one in equation (3.15), the EDoF is calculated as the trace of the hat matrix $\mathbf{A}$ in equation (3.23), and thus ranges from 0 (when $\lambda \to \infty$) to the number of basis functions and dimension of the spline term k (when $\lambda = 0$) (Wood, 2017).

3.2.5 XGBoost

As this thesis focuses on spline models but also includes a comparison with XGBoost, this section provides a brief introduction to the theory behind it. Short for eXtreme Gradient Boosting, XGBoost is a powerful and efficient implementation of gradient tree boosting tailored for both regression and classification problems, developed by Chen and Guestrin (2016).

XGBoost is an ensemble method that combines the predictions of multiple regression trees to make a final prediction. Each tree in the ensemble contributes an additive function $f_k(x)$ to the final prediction $\hat{y}_i$. The prediction for a given input x_i is:

$$\hat{y}_i = \eta \sum_{k=1}^{K} f_k(x_i) \tag{3.25}$$

where η is the learning rate, reducing the impact of each individual tree, and K is the number of trees (M. Zou et al., 2022). To ensure the model generalizes well and avoids overfitting, XGBoost incorporates a regularized objective function, $\mathcal{L}(\phi)$, defined as:

$$\mathcal{L}(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{k} \Omega(f_k) \tag{3.26}$$

This objective function combines a loss term, $l(\hat{y}_i, y_i)$, which measures the error between the predicted and actual values - typically the Mean Squared Error (MSE) for regression tasks - with a regularization term that penalizes the complexity of the model. The regularization term, $\Omega(f_k)$, is applied to each tree f_k to control its complexity, and is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3.27}$$

Here, γ penalises the total number of leaves in the tree, T, and λ controls the L2 regularization on the leaf weights (w). This regularization helps to keep the model simple, preventing overfitting by penalizing complexity and the influence of individual data points.

XGBoost trains the model additively. Starting with an initial prediction, it iteratively adds new trees to improve the prediction. At each iteration t, a new tree f_t is added to minimize the objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{3.28}$$

where $\hat{y}_i^{(t-1)}$ is the prediction from the previous iteration, and f_t is the new tree being added. To optimize this objective, XGBoost uses a second order Taylor approximation, simplifying the optimization process and allowing for the calculation of optimal weights for the tree leaves. For a fixed tree structure, the optimal weight of leaf j is given by:

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3.29}$$

where I_j represents the set of data points that fall into leaf j, g_i is the gradient and h_i the hessian of the loss function with respect to the prediction from the previous iteration $\hat{y}_i^{(t-1)}$. The respective optimal value of the regularized objective is:

$$\mathcal{L}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{3.30}$$

This value serves as an evaluation metric for the decision tree. It combines the fit of the tree to the data, reflected by the sum of gradients and hessians of the loss function, with the complexity of the tree, moderated by the regularization parameters λ and γ. A lower score indicates a better trade-off between accuracy and simplicity, signifying a more optimal tree structure. A greedy algorithm is typically used to build the tree, starting from a single leaf and iteratively adding branches to minimize the loss function while considering the regularization term. Specifically, each potential split is evaluated based on the reduction in the loss function, or gain, defined as:

$$\text{Gain} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{3.31}$$

where I_L and I_R are the subsets of data points for the left and right leaves after the split (children), and $I = I_L \cup I_R$ represents the parent leaf. Thus, if the gain of the split is lower than γ, the branch is not created. For further details of the XGBoost algorithm, see Chen and Guestrin (2016).

To use XGBoost, certain hyperparameters need to be set by the user:

- **Learning rate (η)**: This controls the step size at each iteration while moving towards a minimum of the loss function. A smaller value makes the model more robust but requires more boosting rounds.

- **Max depth**: The maximum depth of a tree. Increasing this value makes the model more complex and more likely to overfit.
- **Number of estimators (K)**: The number of boosting rounds. More boosting rounds improve the model's performance but also increase computation time and may lead to overfitting.
- **Gamma (γ)**: The minimum loss reduction required to make a further partition on a leaf node of the tree. The larger gamma is, the more conservative the algorithm will be.
- **Min child weight**: The minimum sum of instance weight (hessian) needed in a child. In simpler terms, it ensures that any node being split contains a sufficient number of data points. A larger value prevents overfitting.
- **Lambda (λ)**: L2 regularization term on weights, helping to reduce overfitting.
- **Alpha (α)**: L1 regularization term on weights.

These hyperparameters enable the user to tune the model according to the complexity and size of their data, to find a balance between bias and variance. One popular method for finding the optimal hyperparameters is GridSearchCV (M. Zou et al., 2022). By defining a range for each parameter, training a model for each configuration, and then evaluating each combination using cross-validation, the optimal set of hyperparameters can be determined.

3.2.6 Implementation

The linear models were implemented using the `statsmodels` package in Python. To implement GAMs for this thesis, the `pyGAM` package is utilized (Servén and Brummitt, 2020). This package is commonly employed for similar purposes, as evidenced by its use in studies such as Yang et al. (2021) and Siems et al. (2024). Although the `statsmodels` package also supports GAMs, it was not selected due to its inability to make predictions outside the range of the explanatory variables used in training. Additionally, the `mgcv` package by Simon Wood, which is referenced extensively as a primary source for GAM theory in this thesis, could be another good option. However, because this is implemented in `R`, it was not chosen for use in this work. The XGBoost models are implemented using the `xgboost` package in Python.

3.3 Model Criteria

This section describes the methods used to select the most suitable and well-performing models for predicting PD, given the analysed dataset. Throughout the model selection process, only models where all coefficients show significance using the p-value threshold of 0.05 are considered.

3.3.1 How many explanatory variables?

To limit the model's complexity and facilitate its interpretability, a cap was set on the number of independent variables to three. This means that only models with one, two, or three macroeconomic variables were considered. The rationale behind this approach is to ensure the models remain comprehensible and interpretable while still capturing the essential economic dynamics.

By adhering to this convention, the study aims to produce models that align with established modeling practices within the financial industry.

3.3.2 Time series cross-validation

As described by Hyndman and Athanasopoulos (2018), traditional K-fold cross-validation predicated on the assumption that data points are independent and identically distributed is not suitable for time series data due to the inherent auto-correlation that is often present. The correlation between training and testing sets together with shuffling could adversely affect the accuracy of generalization error estimates. Thus, a better option is to use cross-validation that preserves the temporal dimension of the data. One such option is so called chained or expanding window cross-validation. This technique modifies the traditional K-fold methodology by extending the training dataset with each iteration to include all data up to the Kth fold, while the $(K+1)$th fold serves as the test set. This method facilitates a realistic forecast scenario by leveraging historical data up to the point of each prediction period, thereby avoiding the inaccuracies associated with small or future-informed training sets. The forecasting model's effectiveness is assessed by calculating a performance metric, such as the root mean square error (RMSE), across all test folds, with the goal of selecting the model that demonstrates the lowest average RMSE and thus best generalization. Therefore, expanding window cross-validation from the `Scikit-learn` package is used in this study as illustrated in Figure 3.2 below.

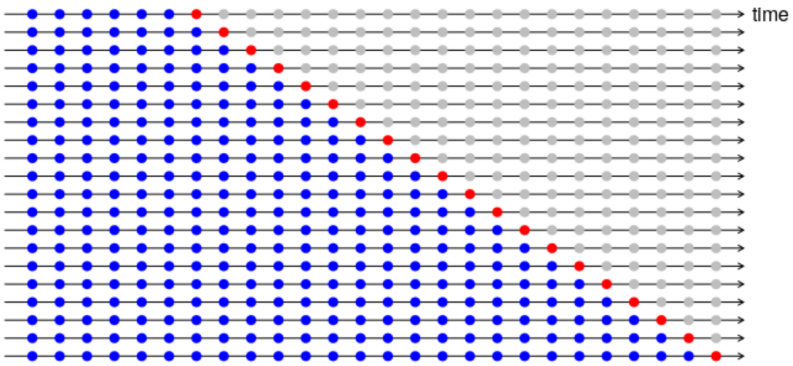


Figure 3.2: Illustration of expanding window cross-validation used for time series data. However, in this study K blocks of test sets are used rather than each individual data point. From Hyndman and Athanasopoulos (2018, p. 69).

3.3.3 Multicollinearity

One underlying assumption in regression analysis is the absence of perfect correlation between the independent variables. Should they be uncorrelated, adding or removing one should not affect the estimates of the others. However, there will always be some degree of correlation present among the explanatory variables. When two or more predictor variables in a multiple regression model are highly correlated - such that one can be linearly predicted from the others - the condition is referred to as multicollinearity (Brooks, 2019). This results in difficulties in distinguishing the individual contribution of each explanatory variable, leading to high coefficient standard errors and potentially causing insignificance, despite a high R^2 indicating a good model fit. Multicollinearity can also result in highly sensitive model coefficients, with the addition or removal of an indepen-

dent variable greatly changing the value or significance of the other variables.

A way of measuring multicollinearity is by calculating the Variance Inflation Factor for each independent variable, expressed as:

$$VIF_i = \frac{1}{(1 - R_i^2)} \tag{3.32}$$

where R_i^2 represents the coefficient of determination from regressing the explanatory variable x_i against all other independent variables in the model. A higher VIF indicates more severe multicollinearity between the variable in question and the rest of the explanatory variables in the model. The VIF has a minimum value of one when the tested variable is completely independent from other variables. Generally, a VIF under 5 suggests multicollinearity is not a concern, but a value of 5 or above signals a need for corrective measures. However, as a rule of thumb, some researchers consider a VIF of 10 as the threshold for significant multicollinearity that requires attention (Brooks, 2019).

Firstly, one could ignore the problem with multicollinearity if the model is satisfactory, since it may not always significantly affect the t-ratios of variables that would have been significant otherwise. Secondly, another approach is to remove one of the collinear variables to eliminate the problem. However, this could introduce omitted variable bias. Lastly, transforming the correlated variables into a ratio and using this ratio in the regression is an alternative strategy. In this study, the second approach is used. Hence, all models where any of the explanatory variables indicates a VIF greater than 5 were excluded from the model selection.

3.3.4 Intuitive signs

As previously mentioned, the macroeconomic factors included in the model selection process are selected through economic intuition and based on previous studies, hypothesized to be connected to the probability of default among corporates in Sweden. Thus, there is also a common conception of whether each of these variables are positively or negatively correlated to PD. To align this knowledge with the model selection, a requirement is set so that the sign of the independent variable is the same as the "expected" sign according to economic theory. Thus, no model included in the final selection has non-intuitive correlations between the macroeconomic variables and PD.

3.4 Model Evaluation

This section presents the different theories and methods applied to evaluate and compare the performance of the different models.

3.4.1 Accuracy metrics

Accuracy measures are used to compare different models' capacity to predict actual data, either from the training set (in-sample) or from future unseen data (out-of-sample). The measure best fit for evaluating in-sample might not be the one best fit for out-of-sample evaluation. Thus, one should choose measure based on how well it offers insight into out-of-sample accuracy (Makridakis, 1993). Further, there is no best overall measure that fits to all datasets. However, if such a measure were to exist it would take the form of a relative measure, that is, expressed as a percentage, as one

would otherwise compare numbers with different magnitude. As there is no "one size fits all", two different loss functions are used in this study to provide robust results. The two metrics used to compare the accuracy and predictive performance of the different models are: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). This is similar to the approach used by Brooks et al. (2001), where they employed different time series models to predict the FTSE 100 index movements based on futures prices and used Mean Absolute Error (MAE) and RMSE as measures of prediction errors.

RMSE is chosen as it is one of the most commonly used loss functions. It has the beneficial property that it ensures that the error's magnitude corresponds to that of the predicted quantity (Han et al., 2022). The measure also puts larger emphasis on outliers and is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.33}$$

Further, as described by Makridakis (1993), MAPE is a relative metric that integrates the most beneficial aspects of various accuracy measures. It is defined as

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.34}$$

As this is a relative measure, it is highly interpretable and has a lower bound of zero (Brooks, 2019). There are however a few flaws with this measure that needs to be addressed. Firstly, it is not symmetric in the sense that equal errors above and below the actual value does not yield the same MAPE. Secondly, it can inflate the percentage errors if the actual value, y_i, is small or close to zero as can clearly be seen by the denominator in (3.34). It is acknowledged that there are different adjusted version of MAPE mitigating these flaws, but to maintain the intuitiveness of the measure, the original MAPE is used in this study. Moreover, given that the target variable (PD) ranges from 0 to 1 across all data points, the amplification of errors for small values is expected to be consistently applied across all observations, suggesting that MAPE remains a suitable loss function for the dataset analysed in this thesis.

For a model to demonstrate strong generalization capabilities, it is essential that it performs well on both in-sample and out-of-sample data. Consequently, models that exhibit the lowest RMSE and mean absolute percentage error (MAPE) during in-sample cross-validation are subject to further evaluation on the out-of-sample data. The model that achieves the lowest loss scores across both datasets is then ultimately selected.

3.4.2 Regression assumptions

Further, the models under consideration are assessed for their adherence to the assumptions of regression analysis, as fulfilling these criteria is imperative to ensure the model's reliability for both predictive and inferential analyses. This includes verifying the independence and normality of the residuals for the validity of statistical tests, and examining for homoscedasticity to guarantee that the variance of the error terms is constant across different values of the independent variables. Following the principles outlined by Hyndman and Athanasopoulos (2018), there are two crucial attributes that warrant extra attention in residual analysis. *Firstly*, it is essential that residuals

exhibit no correlation; the presence of correlation suggests remaining predictive information in the residuals that could improve forecast accuracy. *Secondly*, residuals should have a mean of zero to prevent forecast bias; a non-zero mean indicates systematic overestimation or underestimation in the model predictions. The properties of homoscedasticity and normal distribution of the errors are useful but not essential.

To check the normality of the residuals, a graphical approach is to use a quantile-quantile plot (QQ-plot) which plots the quantiles of the empirical distribution against the quantiles of the theoretical distribution that the data is assumed to follow (Chambers et al., 2018). Should the assumption hold true and the residuals follow a normal distribution, the points on the plot are expected to align closely with the line $y = x$.

A statistical test that can be used in tandem with the QQ-plot is the Jarque-Bera test with the null hypothesis H_0 : Residuals are normally distributed against H_1 : Residuals are not normally distributed (Brooks, 2019). A high p-value thus indicates normal errors and the residuals are thus considered normal if the JB p-value is above 0.05.

Further, autocorrelation often appears in the residuals when one applies a regression model to data with a time series structure - violating the assumption of independent errors in the estimated model (Hyndman and Athanasopoulos, 2018). The independence of the residuals can be tested using the Durbin-Watson (DW) test, with the null hypothesis $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$, where ρ is from the regression on two subsequent error terms:

$$\epsilon_t = \rho\epsilon_{t-1} + \nu_t \tag{3.35}$$

where $\nu_t \sim N(0, \sigma^2)$. However, for computational reasons, the test statistic is calculated directly from the error terms as

$$DW = \frac{\sum_{t=2}^{n}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{n}\epsilon_t^2} \approx 2(1 - \rho) \tag{3.36}$$

For further details on the last approximation, see Brooks (2019), chapter 5. Given that $-1 \leq \rho \leq 1$, it follows that $0 \leq DW \leq 4$, where $DW = 0$ signifies perfect positive autocorrelation, $DW = 4$ signifies perfect negative autocorrelation and $DW = 2$ indicates no autocorrelation in the residuals. Since the DW test statistic does not follow a standard statistical distribution, a p-value cannot be directly calculated. However, there are upper and lower critical values presented by Durbin and Watson (1951) which can be used for hypothesis testing.

Another way to analyse the autocorrelation in the residuals is through the autocorrelation function (ACF) plot which visualizes the autocorrelation function as a function of lag length (Box et al., 2015), together with a 95% confidence interval. A high value outside the confidence interval at a certain time step indicates strong correlation between the residuals at that interval, suggesting significant influence of past values on the current value.

3.5 Linear Model Development Process

This section details the practical steps undertaken in the development and selection of the predictive models for PD. The process began with preliminary testing of the individual significance of each macroeconomic variable. This was achieved through single-factor models to ensure that

only variables with a consistent impact on PD were considered for further analysis. Subsequently, all possible combinations of one, two, and three significant variables were created, utilizing time series cross-validation to assess model performance in-sample as well out-of-sample. This iterative approach aimed to identify the model configuration that provided the best balance between complexity and predictive accuracy, as measured by RMSE and MAPE.

3.5.1 Single-factor significance testing

The initial phase of model development focused on the individual assessment of each macroeconomic variable's predictive power regarding PD. This involved constructing single-factor linear models for each variable and evaluating their performance through 5-fold cross-validation on the in-sample dataset. Variables were considered significant and retained for further analysis if their associated models demonstrated significance in at least 50% of the cross-validation folds, using a p-value threshold of 0.05.

3.5.2 Multi-factor models

Following the identification of individually significant variables, the next step were the exploration of variable combinations. Models comprising each combination of one, two, or three macroeconomic variables were constructed and evaluated. The models were evaluated using time series cross-validation in-sample as well as trained on the whole in-sample test to then be evaluated on the two out-of-sample periods.

3.5.3 Final model selection

Lastly, the models that did not fulfill the criteria described in Section 3.3 were removed. Out of the remaining adequate models, which were sorted on average CV RMSE, the model that exhibited the lowest RMSE and MAPE on the out-of-sample data was chosen as the best model for predicting PD. This model was then further analysed according to Section 3.4.2, to assess its compliance with the assumptions of linear regression and to confirm its overall suitability.

3.6 Cubic Spline Model Development Process

A major aspect of the cubic spline model development involved determining the appropriate macroeconomic variables for inclusion in the model. The decision was whether to solely incorporate variables that exhibited significance during the linear model development phase or to also include variables that were initially discarded. Therefore, firstly the linearly significant variable combinations were tested, to see if a non-linear dependence could improve the linear models' predictive accuracy.

Secondly, the variables that were initially discarded due to insignificance were also fit with a cubic spline model, with the hypothesis of their potential to capture non-linear relationship that a linear model may not have adequately been able to capture.

3.6.1 Variables from linear model selection

For each combination of variables deemed adequate from the linear model development phase, cubic spline functions were fit for each variable. For each GAM model, a grid search to find the optimal smoothing parameter yielding the lowest GCV score was conducted, selecting an optimal smoothing parameter for each macroeconomic variable. The evaluation criterion remained the same as for the linear models, using time series CV to gauge model performance in-sample as well as testing on the two out-of-sample periods using RMSE and MAPE.

The spline models were then sorted on average CV RMSE and the model with the lowest out-of-sample RMSE and MAPE was selected as the optimal spline model. This selected model was then reviewed to assure its adherence to the principles of regression analysis, confirming its overall appropriateness for the study's objectives, as detailed in Section 3.4.2. Furthermore, partial dependence plots showcasing the possibly non-linear dependence between each macroeconomic variable and PD with 95% confidence intervals were created. These partial dependence plots could then be used to analyse whether the smooth function fit through penalized least squares is too "wiggly" (λ too low) or smooth enough (λ okay) through a visual inspection.

3.6.2 Variables that indicate a non-linear relationship

As previously mentioned, a second hypothesis was that a reasonable approach would be to not only fit GAMs to the variables deemed significant in the linear regression, but to also test the variables rejected in the linear single-factor significance test. The Regression Specification Error Test (RESET), developed by Ramsey (1969), is a method for assessing the adequacy of linear regression models. It evaluates whether incorporating non-linear transformations of the independent variables could enhance the model's ability to predict the dependent variable. This approach is based on the premise that if these non-linear transformations significantly contribute to explaining the dependent variable, it suggests that a simple linear model may not accurately capture the underlying relationship, indicating a potential for improvement by considering polynomial or other non-linear forms. The test is done on a regression using higher orders of the fitted linear model $\hat{y} = \beta_0 + \beta_1 x$ together with the original regression:

$$y = \alpha_1 + \alpha_2 \hat{y}^2 + \alpha_3 \hat{y}^3 + \cdots + \alpha_k \hat{y}^k + \beta_1 x + \epsilon \tag{3.37}$$

and then testing the joint significance of the α's using an F-test. If the null hypothesis that all $\alpha_k = 0$, $k \geq 2$ is rejected, this suggests that the linear model is misspecified and non-linear transformations could be a better approach, such as a GAM.

Therefore, the variables that were not used in the linear model development phase was tested using the RESET test. The variables for which the null hypothesis was rejected was then included in the selection for a GAM model with up to 3 variables in a similar manner as previously described and sorted based on average RMSE.

3.6.3 Variables from best linear model

Further, a model with the exact same macroeconomic variables as the linear model with the lowest RMSE was chosen to investigate if a spline model with the same variables could improve the

predictive accuracy.

Lastly, a model with these same variables but with a mix of linear and spline terms to the different variables were tested as a "mixed" model.

3.7 XGBoost Model Development Process

Similar to R. Li et al. (2020), backward feature elimination was used to select which macroeconomic variables to include in the first XGBoost model. This process starts with a model that includes all macroeconomic variables and then step-by-step eliminates the variable with the lowest importance (F-score) until only two variables remain. The final model therefore consists of the two variables most frequently used in splits for predicting PD. Additionally, an XGBoost model using the same variables as the best linear model was created for comparison purposes with the other modeling techniques.

To ensure a fair comparison between the predictive performance of spline models and XGBoost, the decision was made to limit the number of macroeconomic variables in the XGBoost models to only two. This aligns with the linear and spline models, which also contain only two or three variables, to maintain interpretability. Although XGBoost is typically used with a larger number of predictor variables, this constraint was necessary to create a balanced comparison and evaluate the models under similar conditions.

As described in Section 3.2.5, a grid search using the `GridSearchCV` method was used to optimise the XGBoost model's hyperparameters. The parameter grid included three values for each of the following hyperparameters: max depth, gamma, lambda, alpha, min child weight, number of estimators, and learning rate. For each point on the grid, a 3-fold cross-validation was applied, with the negative mean squared error as the evaluation metric. Therefore, the combination of hyperparameters that yielded the XGBoost model with the lowest average MSE was selected.

Given that the dataset comprises only around 150 data points and the model includes only two variables, the hyperparameters for the grid search were set conservatively to avoid overfitting. The max depth values were set to 2, 3 and 4, while the n_estimators values were set to 3, 7 and 10. Initially, larger hyperparameters were tested, but this led to significant overfitting. This selection ensures the model remains appropriately simple for the limited data.

Chapter 4

Results

This chapter presents the empirical findings of the data analysis, model selection procedures for the linear and non-linear models as well as the evaluation results of the best fitted models.

4.1 Data Transformations and Significance

From the stationarity tests, many of the macroeconomic variables indicated non-stationarity. They were thus transformed according to Section 3.1.4. The macroeconomic variables that were included in the final models, after transformation to yield stationarity, can be seen in Appendix A.

Among the total 143 macroeconomic variables considered (including their lagged versions), 42 were deemed significant following the single-factor significance test outlined in Section 3.5.1. Consequently, these variables were incorporated into the model development process. These 42 variables are different lagged versions of GDP growth, unemployment rate (UR), electricity producer price index growth (ELEC), government consumption (GC) growth, repo rate (differenced), short-term and long-term interest rate (ST_Rate, LT_Rate) (differenced) and share price index growth (OMXSPI).

4.2 Linear Regression Models

An overview of the top 5 linear models based on average cross-validation RMSE is presented in Table 4.1 below.

Table 4.1: Top 5 linear models based on average CV RMSE.

Features	**CV RMSE**	**CV MAPE**	**OOT 1 RMSE**	**OOT 1 MAPE**	**OOT 2 RMSE**	**OOT 2 MAPE**
UR, GC growth lag 3	0.001426	20.76	0.002863	28.48	0.002711	32.18
GDP growth, UR lag 1, GC growth lag 3	0.001499	21.14	0.002869	28.52	0.002752	29.32
UR lag 2, GC growth lag 3, OMXSPI growth lag 12	0.001570	22.83	0.002766	27.92	0.002782	31.36
GDP growth, UR lag 1	0.001576	22.49	0.003031	30.26	0.002647	29.91
UR lag 3, OMXSPI growth lag 12	0.001670	24.49	0.002907	29.28	0.003033	34.04

Note: Several models share variables in different combinations and lags. However, models with identical variable combinations, differing only by a single lag in one variable, are excluded from this table as they are considered nearly identical.

It is important to note that the model with the lowest average RMSE from cross-validation might not have the lowest RMSE and MAPE on out-of-time set 1. However, average RMSE from cross-validation is considered a more reliable metric because it evaluates model performance across multiple test sets, whereas OOT 1 is just a small dataset spanning 2 years and can be more affected by outliers and, hence, has more uncertainty attached to it. Consequently, based on Table 4.1, the model exhibiting the lowest average RMSE from in-time cross-validation is chosen for detailed examination, as it is deemed to best generalize to new data, aligning with the thesis's objective. A summary of this model is shown in Table 4.2.

Table 4.2: Summary of the best linear model.

Feature	Coefficient
Constant	-x (0.482)
UR	y*** (0.061)
GC growth lag 3	-z** (0.041)
Durbin-Watson	0.221
Jarque-Bera p-value	0.293

Note: Coefficients omitted for confidentiality. Standard errors are reported in parentheses.
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

The predicted versus actual probability of default across the training and testing periods for the optimal linear model is depicted in Figure 4.1. This figure shows that the linear model manages

fairly well to capture the trends of the data used for training, up until 2018, but its performance during the first OOT period (beginning in 2018) indicates a lack of response. The second OOT period presents an atypical scenario: despite macroeconomic indicators suggesting an expected rise in PD (due to Covid-19), observed PD values actually decreased, likely due to extraordinary government interventions. This anomaly makes the period less critical for model evaluation, as the inverse relationship observed is not typical of the correlation between macros and PD.

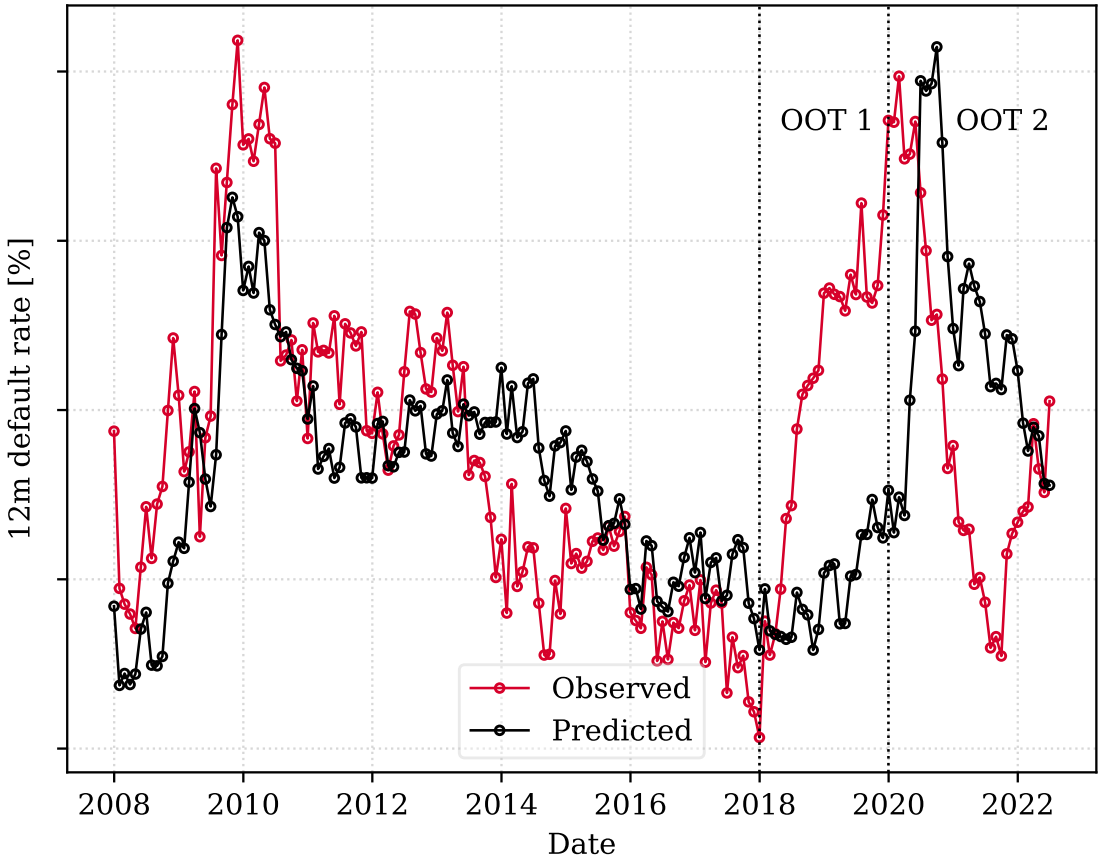
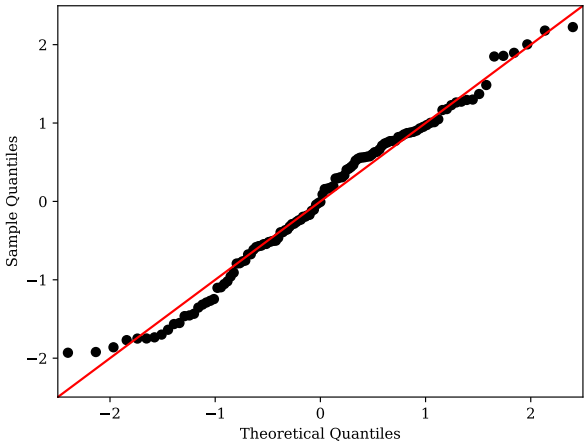


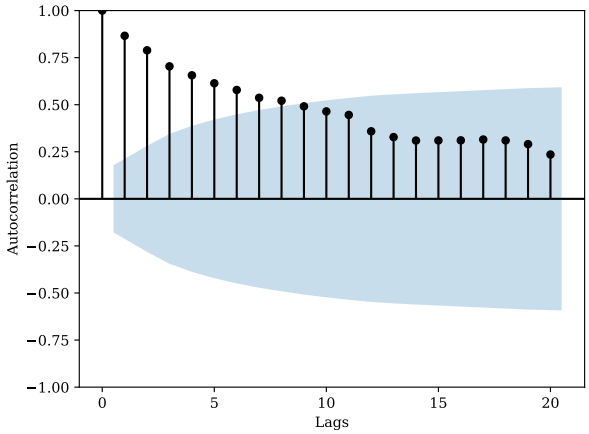
Figure 4.1: Actual versus observed PD for the linear model with the lowest average RMSE from cross-validation. Trained on all data up until 2018 (before OOT 1). Y-axis omitted and data masked for confidentiality reasons.

The QQ-plot and ACF of the residuals for the linear model are illustrated in Figure 4.2 below. From 4.2a it is seen that the residuals indicate normality since the observations follow the diagonal line well, except for some observations in the tails. Furthermore, the Jarque-Bera p-value from Table 4.2 indicates that the null hypothesis of normally distributed errors should not be rejected.

Notably, Figure 4.2b reveals strong autocorrelation among the residuals, with several lags showing values significantly different from zero outside the blue confidence interval. This persistent autocorrelation in the residuals suggests that the model may not be capturing all the predictive structure in the data, which could be due to time-dependent processes or omitted variables. Probably this stems from the inherent autocorrelation in the 12m PD used as the target variable. These findings indicate an opportunity for model enhancement to account for the temporal dependencies not yet addressed in the current model formulation.



(a) QQ-plot for the residuals of the linear model with the lowest average RMSE from cross-validation.



(b) ACF plot for the residuals of the linear model showing strong autocorrelation, with 95% confidence intervals in blue shading. Observations outside these intervals carry significant autocorrelation.

Figure 4.2: Residual diagnostics for the linear model: (a) QQ plot, and (b) ACF plot.

4.3 Cubic Spline Models

Table 4.3 summarizes the different spline models considered, yielding the lowest average RMSE for each selection of macroeconomic variables. The following sections give a more detailed description and evaluation of each model.

Table 4.3: Summary of the spline models yielding the lowest average RMSE from each variable selection.

Features	**EDoF**	**CV RMSE**	**CV MAPE**	**OOT 1 RMSE**	**OOT 1 MAPE**	**OOT 2 RMSE**	**OOT 2 MAPE**
GDP growth lag 3, UR lag 3	10.6	0.001855	24.50	0.003319	30.26	0.005492	66.71
ST_RATE diff lag 12, GC growth lag 12	8.3	0.001725	25.78	0.002366	24.54	0.002420	23.83
UR, GC growth lag 3	9.5	0.001848	27.98	0.003085	30.23	0.003772	45.80
UR, GC growth lag 3	5.7	0.001398	20.87	0.003003	29.48	0.003810	46.17

Note: The last two models are (1) with spline terms for both variables and (2) with a spline term on the first variable and linear term on the second variable. EDoF shows the total effective degrees of freedom.

From top to bottom in the table, the models are referenced to as spline model 1, 2, 3 and 4. For ref-

erence, ST_Rate is short-term interest rate (3-month T-bill yield), GC is government consumption and UR is unemployment rate.

4.3.1 Variables from linear selection

Let us take a closer look at spline model 1. Table 4.4 presents a summary of the cubic spline model that achieved the lowest average RMSE on cross-validation, utilizing the same variable pool selected during the linear regression model development phase. The combination of macrovariables that yielded the lowest average RMSE was GDP growth lag 3 and unemployment rate lag 3.

Table 4.4: Summary of spline model 1.

Feature	Smoothing parameter (λ)	EDoF
GDP growth lag 3^{***}	38.4	5.8
UR lag 3^{***}	3.1	4.8
Durbin-Watson	0.65	
Jarque-Bera p-value	0.85	
GCV	0.0134	

$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

Smoothing parameters were selected from grid search with a 2000 x 2 grid drawn from a uniform distribution on $[10^{-3}, 10^{3}]$ using GCV as selection criteria.

Figure 4.3 illustrates the partial dependence plot for each macroeconomic variable included in spline model 1, accompanied by 95% confidence intervals. The left plot shows a nearly quadratic relationship between PD and GDP growth lag 3. One explanation for this shape is as follows: *Firstly*, negative GDP growth often correlates with higher default rates, a common occurrence during economic crises. *Secondly*, unusually high GDP growth may result from increased government spending, also typically seen in crisis situations. This spending can temporarily inflate GDP figures. However, there may be a delayed effect on the default rates, which remain high until the impact of this spending is realized in the economy.

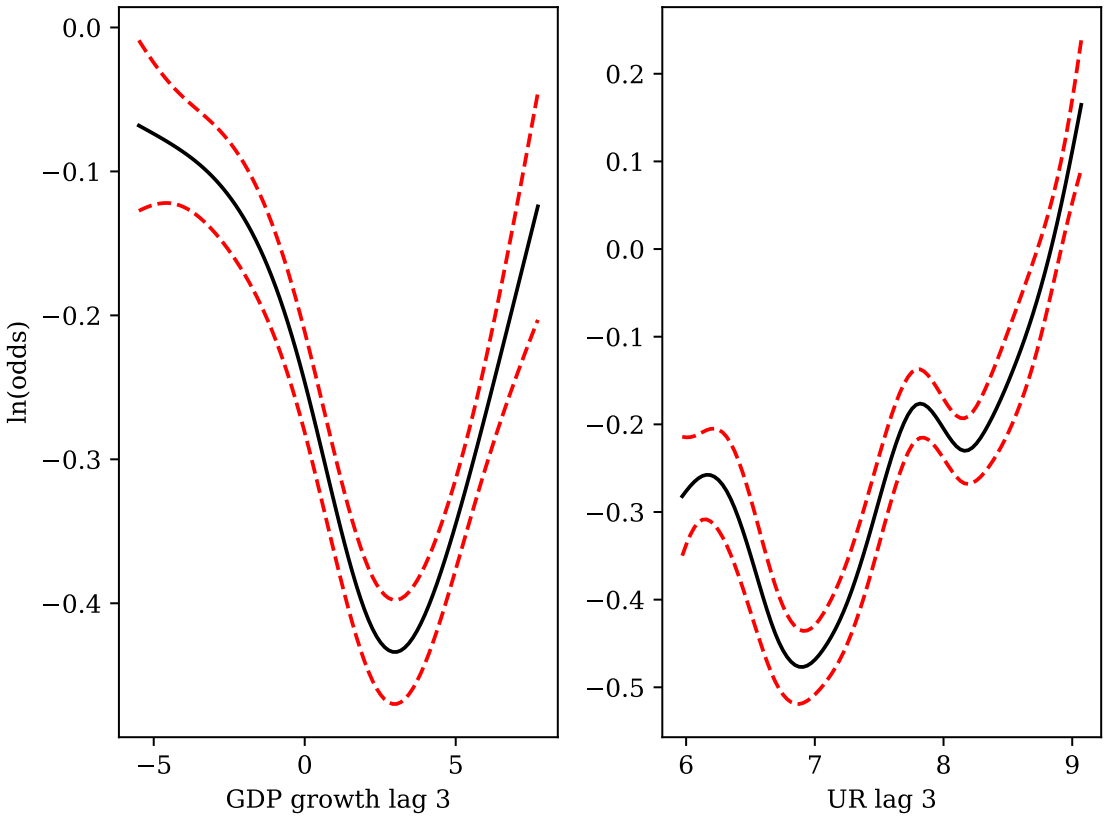


Figure 4.3: Partial dependence plots for spline model 1. 95% confidence intervals in red. $\lambda_1 = 38.4$, $\lambda_2 = 3.1$. Note the bump in right plot (UR $\approx$ 8), which may be due to undersmoothing.

Figure 4.4 displays the predicted versus actual probability of default across the training and testing periods for spline model 1. Note the spikes in the predictions in out-of-time period 2, probably due to unseen values of the macroeconomic factors, i.e., values lying outside the intervals of the independent variables the model was trained on. Note further that the limits on the y-axis changed from Figure 4.1 due to these outliers.

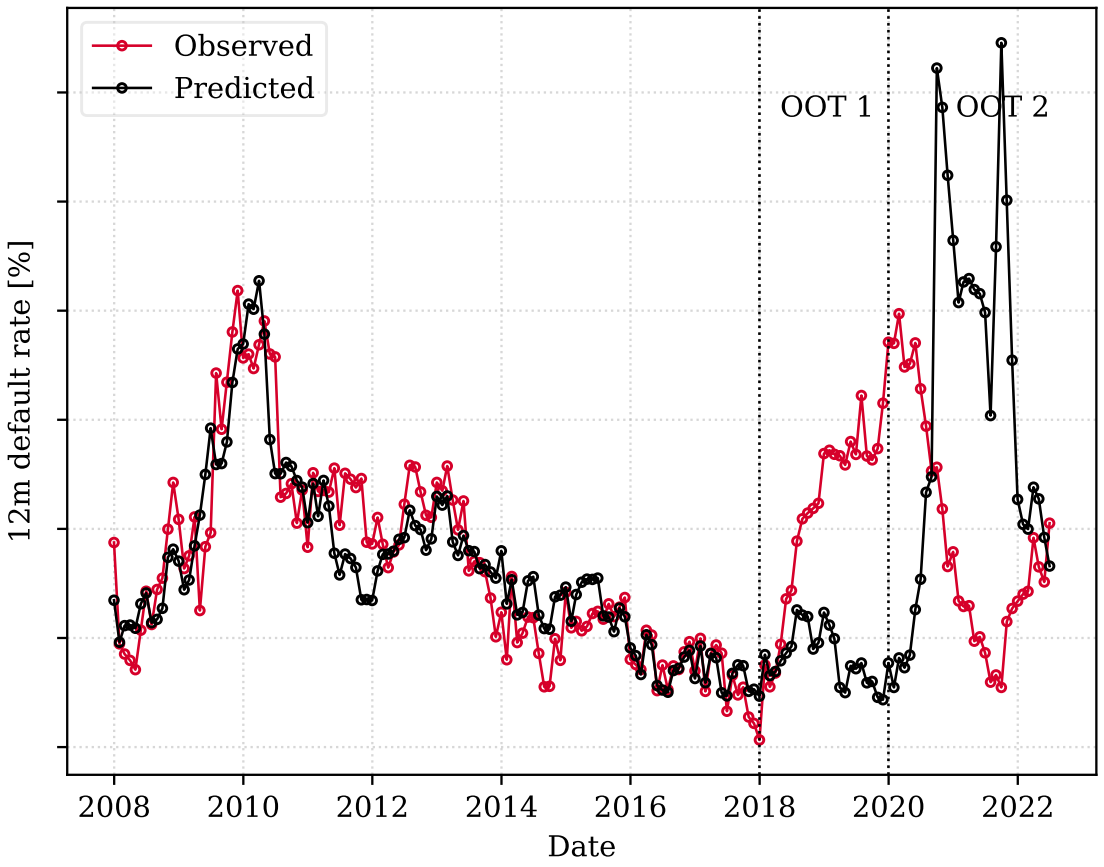
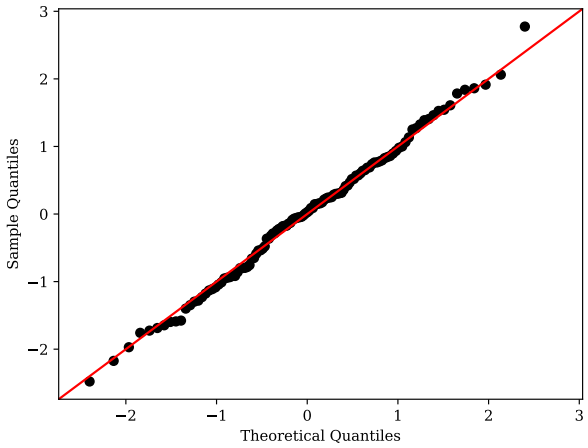
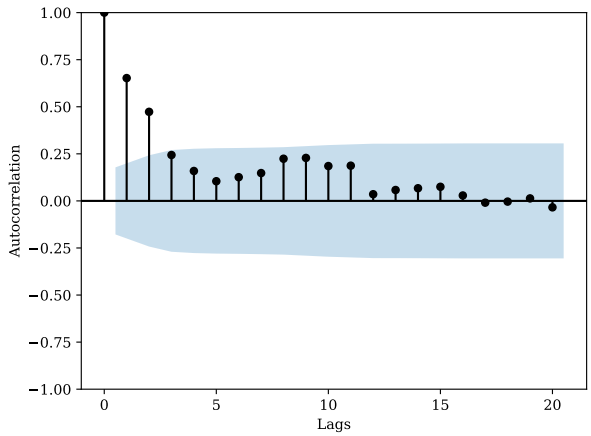


Figure 4.4: Actual versus observed PD for spline model 1 with the lowest average RMSE from cross-validation. Trained on all data up until 2018 (before OOT 1). Y-axis omitted and data masked for confidentiality reasons.

The QQ-plot and ACF plot for the residuals of spline model 1 are seen in Figure 4.5. The residuals seem to follow a normal distribution since the quantiles in the sample follow the theoretical quantiles very well. Furthermore, the residuals from spline Model 1 exhibit autocorrelation at shorter lags compared to the linear model, possibly suggesting that it captures the data's temporal structure more effectively in the immediate term.



(a) QQ-plot for the residuals of spline model 1 with the lowest average RMSE from cross-validation.



(b) ACF plot for the residuals of spline model 1 with the lowest average RMSE from cross-validation, with 95% confidence intervals in blue shading. Observations outside these intervals carry significant autocorrelation.

Figure 4.5: Residual evaluation plots for spline model 1: (a) QQ-plot and (b) ACF plot.

4.3.2 Variables rejected from linear RESET test

Table 4.5 presents a summary of spline model 2, which is the model that achieved the lowest RMSE and MAPE on out-of-sample set 1, incorporating variables that were not used in the linear regression selection and was rejected in the linear RESET test, described in section 3.6.2. The variables yielding the lowest average RMSE from CV are short-term interest rate lag 12 (differenced) together with government consumption growth lag 12.

Table 4.5: Summary of spline model 2.

Feature	Smoothing parameter (λ)	EDoF
ST Rate diff lag 12***	3.5	7.1
GC growth lag 12**	102.5	1.2
Durbin-Watson	0.28	
Jarque-Bera p-value	0.57	
GCV	0.0385	

$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$

The partial dependence plots for each macroeconomic variable included in spline model 2 is illustrated in Figure 4.6, accompanied by 95% confidence intervals.

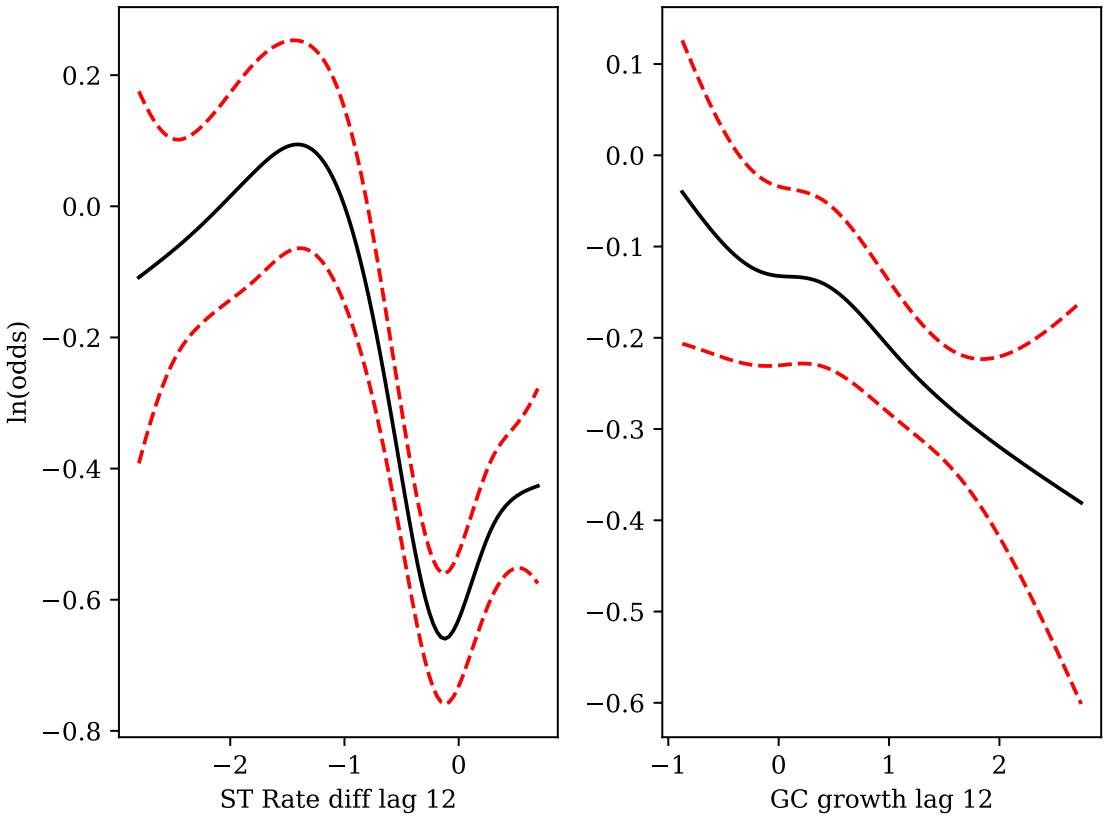


Figure 4.6: Partial dependence plots for spline model 2. 95% confidence intervals in red. The y-axis represents the marginal effects. $\lambda_1 = 3.5$, $\lambda_2 = 102.5$.

Figure 4.7 displays the predicted versus actual probability of default across the training and testing periods for spline model 2. The "wiggly" pattern observed in the predicted PD is due to the nature of the macroeconomic variables that are used in the model. As variables defined as relative and absolute change, respectively, they inherently exhibit volatility, leading to difficulties in capturing the overall trends of the observed PD.

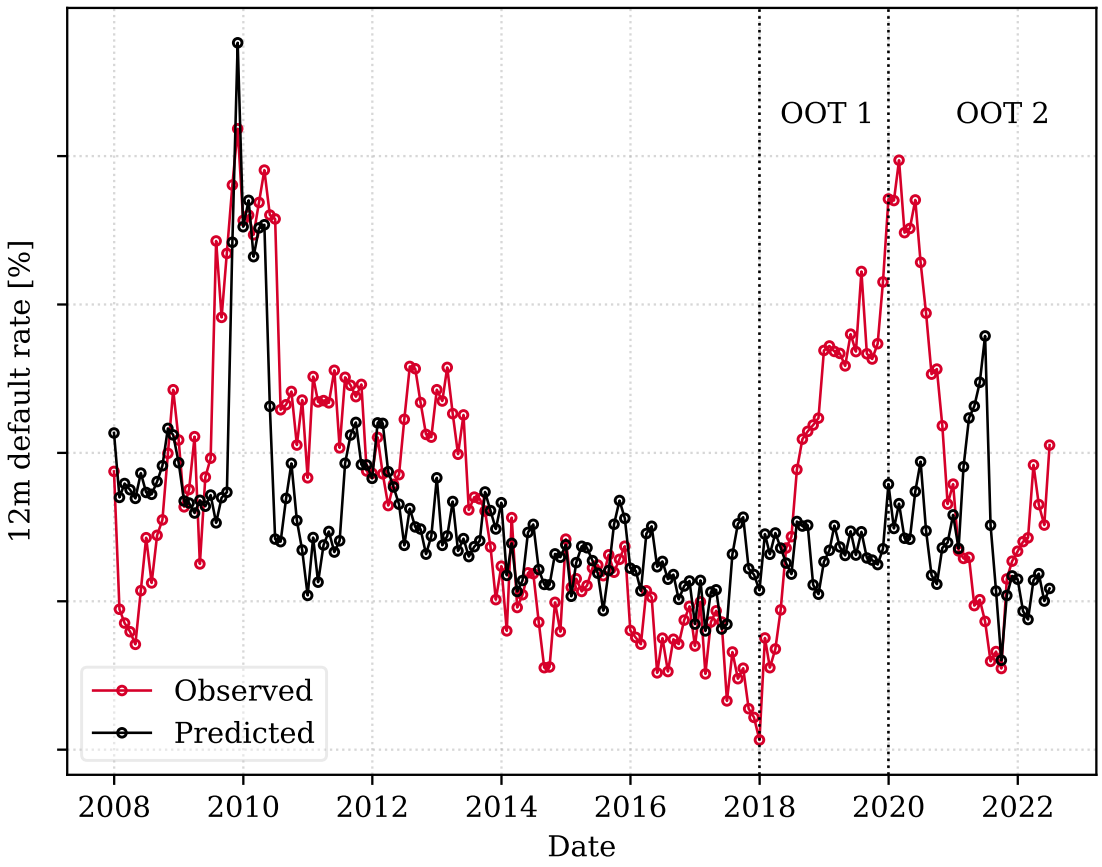
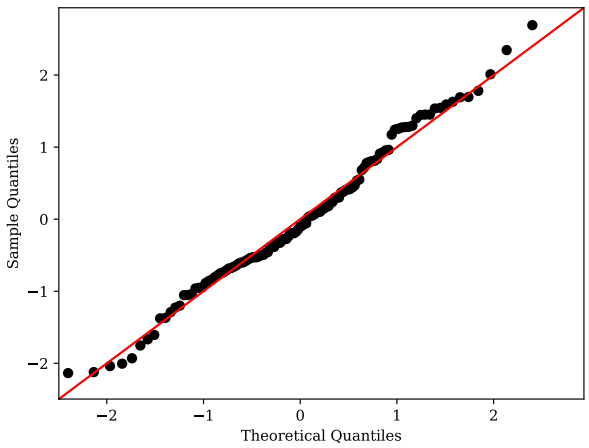
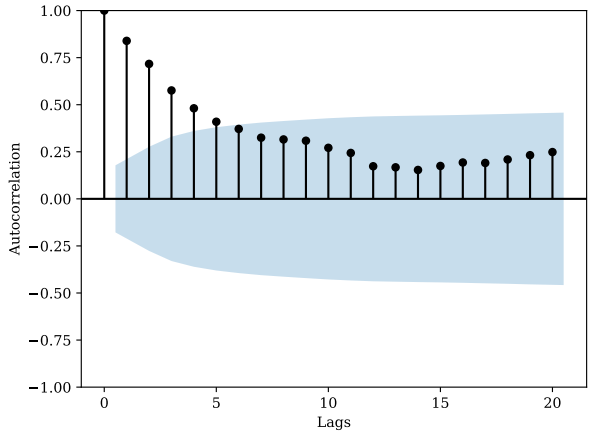


Figure 4.7: Actual versus observed PD for spline model 2 with the lowest average RMSE from cross-validation. Trained on all data up until 2018 (before OOT 1). Y-axis omitted and data masked for confidentiality reasons.

The QQ-plot and ACF plot for the residuals of spline model 2 are seen in Figure 4.8. The residuals seem to follow a normal distribution since the quantiles in the sample follow the theoretical quantiles very well. Further, similar to spline model 1, the residuals from spline model 2 exhibit autocorrelation at shorter lags compared to the linear model, possibly suggesting that it captures the data's temporal structure more effectively in the immediate term. However, observe the non-zero autocorrelation at lag 20, potentially indicative of non-stationary behaviour of the macroeconomic variables underlying the model.



(a) QQ-plot for the residuals of spline model 2 with the lowest average RMSE from cross-validation. Trained on all data up until 2018 (start of OOT 1).

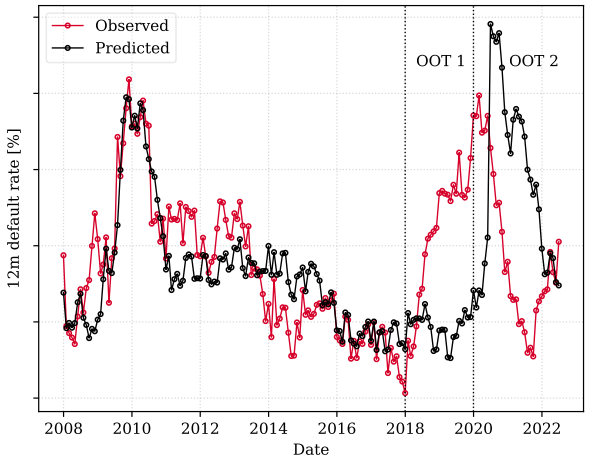


(b) ACF plot for the residuals of spline model 2 with the lowest average RMSE from cross-validation, with 95% confidence intervals in blue shading. Observations outside these intervals carry significant autocorrelation.

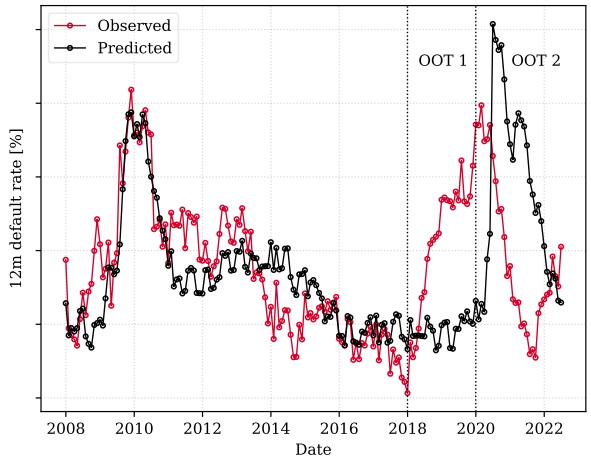
Figure 4.8: Residual evaluation plots for spline model 2: (a) QQ-plot and (b) ACF plot.

4.3.3 Variables from best linear model

Below is a summary of spline model 3 and 4, which are the models using the same variables as the linear model with the lowest RMSE. Model 3 has spline functions on both variables while model 4 contains a mix of linear and spline terms. Figure 4.9 depicts the actual versus predicted values from spline model 3 and 4.



(a) Spline model 3: Splines on both UR and GC growth lag 3.

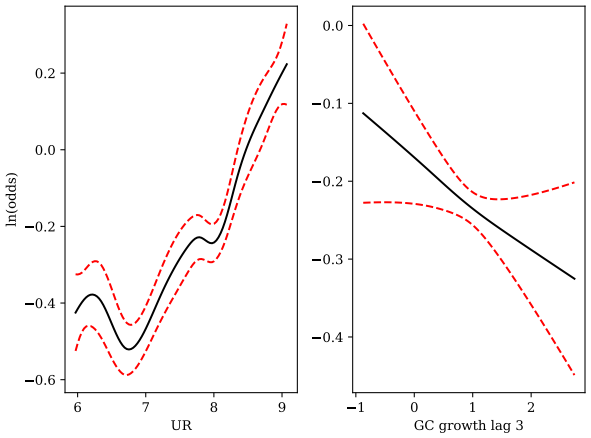


(b) Spline model 4: Spline on UR and linear term on GC growth lag 3.

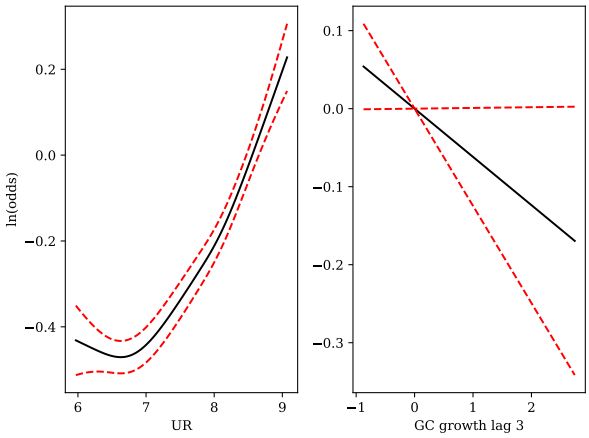
Figure 4.9: Comparative predictions from spline models 3 and 4. Trained on all data up until 2018 (before OOT 1). Y-axis omitted and data masked for confidentiality reasons.

The partial dependence plots is shown in Figure 4.10 below. Note that the spline term of GC

growth lag 3 has such a large penalizing factor that the fit reduces to almost a linear effect, which makes model 3 very similar to model 4 where a linear term was explicitly fit. The zero width of the confidence interval stems from the sum to zero constraint presented in equation (3.24), exactly determining where the straight line passes through zero.



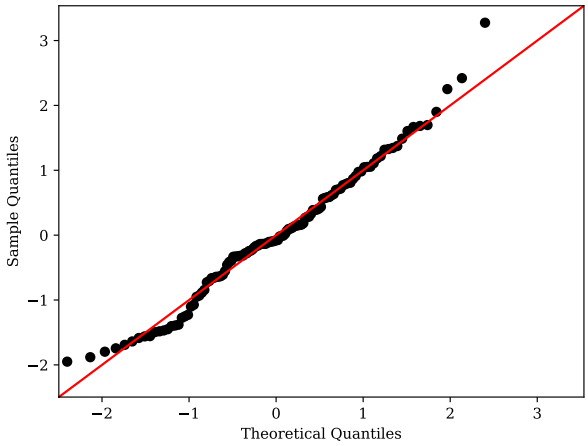
(a) Spline model 3: Splines on both UR ($\lambda_1 = 1.7$) and GC growth lag 3 ($\lambda_2 = 759$).



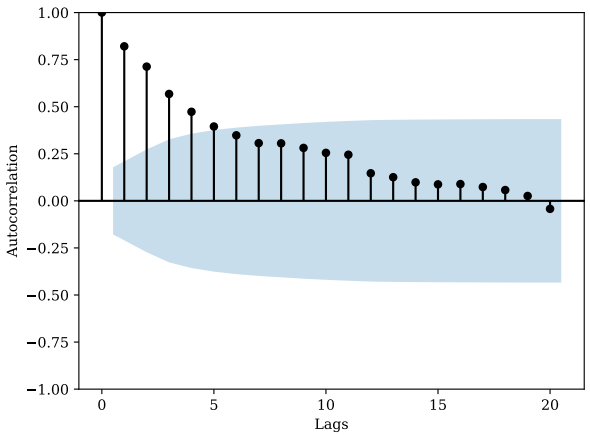
(b) Spline model 4: Spline on UR ($\lambda_1 = 80$) and linear term on GC growth lag 3.

Figure 4.10: Partial dependence plots of spline models 3 and 4.

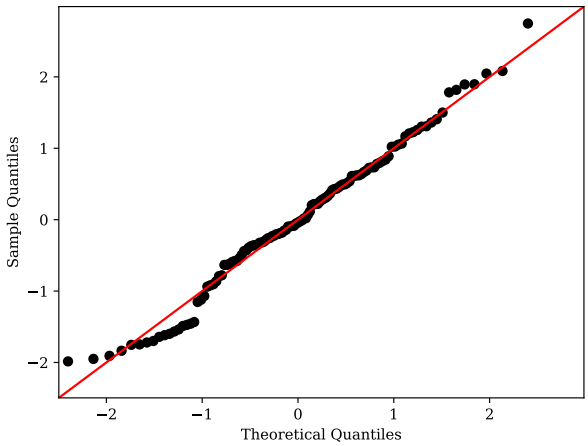
The residual analysis plots are seen in Figure 4.11. The ACF plots indicate that the mixed model has slower decaying autocorrelation in the residuals, and the QQ plots indicate residuals that follow a normal distribution. The autocorrelations tend towards zero with increasing lag lengths, which is good.



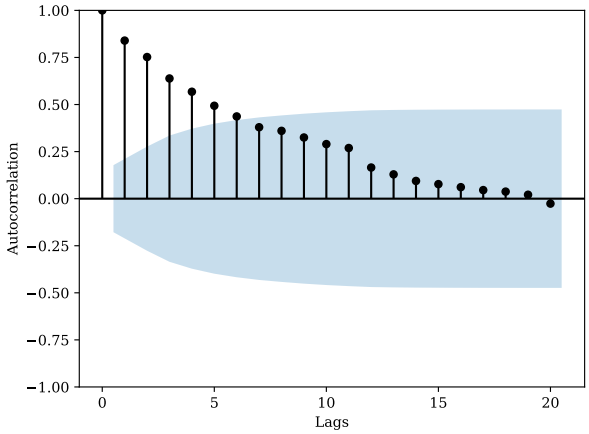
(a) QQ plot for spline model 3.



(b) ACF plot for spline model 3.



(c) QQ plot for spline model 4.



(d) ACF plot for spline model 4.

Figure 4.11: QQ and ACF plots for the residuals of spline models 3 and 4.

4.4 XGBoost Models

Table 4.6 below summarises the accuracy metrics of the two XGBoost models. Note that the CV RMSE value of the first model is lower than those of any of the linear and spline models.

Table 4.6: Summary of the XGBoost models.

Features	**CV RMSE**	**CV MAPE**	**OOT 1 RMSE**	**OOT 1 MAPE**	**OOT 2 RMSE**	**OOT 2 MAPE**
UR, ST Rate diff lag 9	0.001364	18.33	0.002736	27.74	0.003103	37.19
UR, GC growth lag 3	0.001540	22.02	0.003013	29.84	0.003292	39.01

Note: The first model uses the backward-selected variables, while the second uses the variables from the best linear model.

4.4.1 Variables from backwards selection

Table 4.7 below summarises the first XGBoost model. When comparing the Durbin-Watson statistic to those of the other models, this XGBoost model exhibits the lowest level of autocorrelation in the residuals. Additionally, the Jarque-Bera p-value indicates that the residuals follow a normal distribution.

Table 4.7: Backwards XGBoost model hyperparameters and residual statistics.

Hyperparameter	**Value**
Learning Rate	0.25
Max Depth	3
Number of Estimators	10
Gamma	0
Min Child Weight	3
Lambda (L2 Regularization)	1.5
Alpha (L1 Regularization)	0
Durbin-Watson	0.99
Jarque-Bera p-value	0.4

Figure 4.12 displays the predicted versus actual PD across the training and testing periods for the XGBoost model with variables from backwards selection. The model captures the upward trend of PD in OOT 1 more accurately than the linear and spline models. It also avoids predicting excessively high values in OOT 2, as opposed to the spline models.

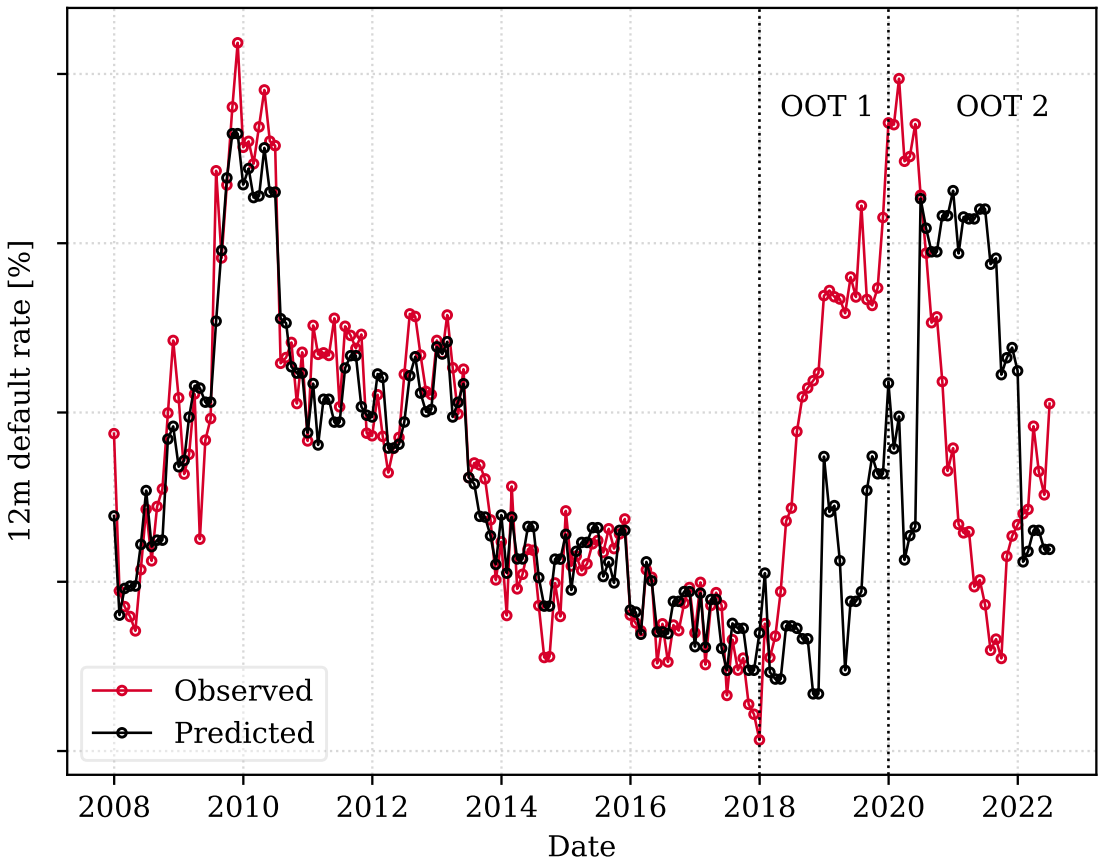
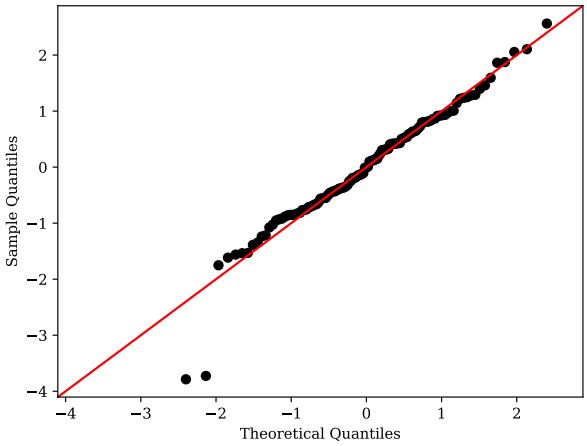
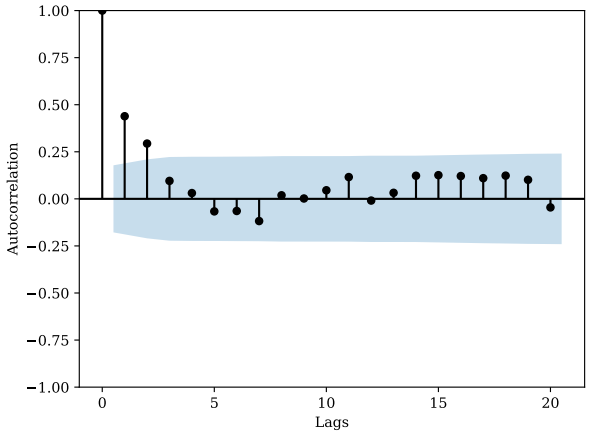


Figure 4.12: Actual versus observed PD for XGBoost backwards selected model. Trained on all data up until 2018 (before OOT 1). Y-axis omitted and data masked for confidentiality reasons.

The residual analysis plots for the backward-selected XGBoost model are shown below. Figure 4.13a shows that the residuals follow a normal distribution well, except for two outliers, and Figure 4.13b indicates that the autocorrelation decays quickly, at approximately the same rate as in spline model 1. These two outliers in the residuals are omitted when calculating the residual test statistics in Table 4.7.



(a) QQ-plot for the residuals.



(b) ACF plot for the residuals, with 95% confidence intervals in blue shading.

Figure 4.13: Residual evaluation plots for XGBoost model with backwards selection of variables: (a) QQ-plot and (b) ACF plot.

4.4.2 Variables from best linear model

Table 4.8 below summarises the second XGBoost model, using the same macroeconomic variables as the best linear model.

Table 4.8: Hyperparameters and residual statistics for the XGBoost model with the same variables as the best linear model.

Hyperparameter	**Value**
Learning Rate	0.3
Max Depth	3
Number of Estimators	7
Gamma	0
Min Child Weight	1
Lambda (L2 Regularization)	1.5
Alpha (L1 Regularization)	0.01
Durbin-Watson	0.38
Jarque-Bera p-value	0.19

Figure 4.14 displays the predicted versus actual PD across the training and testing periods for the XGBoost model with the same variables as the best linear model.

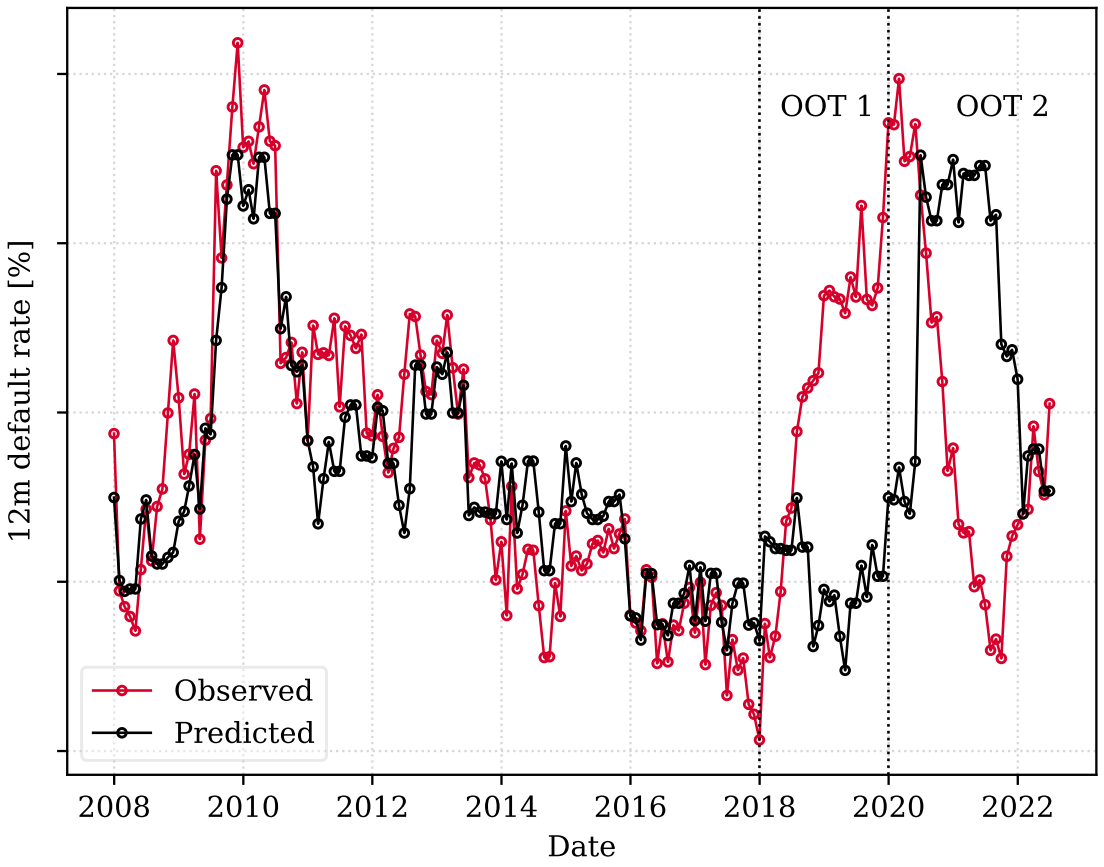
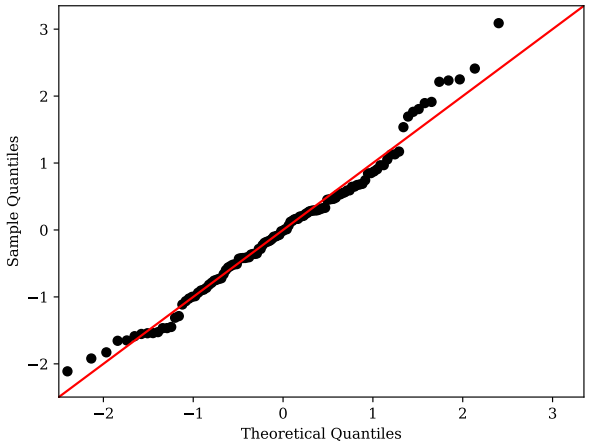
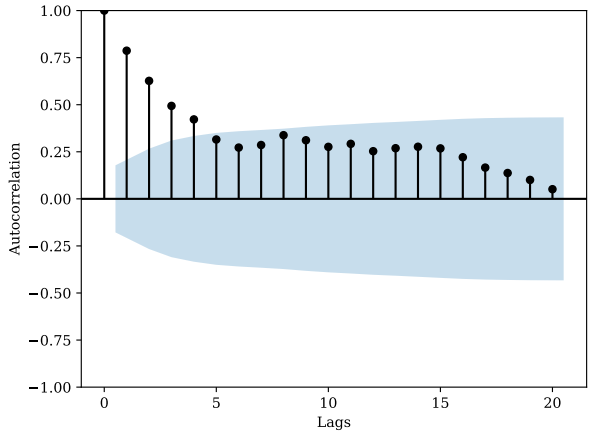


Figure 4.14: Actual versus observed PD for XGBoost with same variables as the best linear model. Trained on all data up until 2018 (before OOT 1). Y-axis omitted and data masked for confidentiality reasons.

The residual analysis plots for the model are shown below. Figure 4.15a shows that the residuals follow a normal distribution well, except for values above the first quantile, and Figure 4.15b indicates autocorrelation decaying at a slower rate than for the backwards selected XGBoost model.



(a) QQ-plot for the residuals.



(b) ACF plot for the residuals, with 95% confidence intervals in blue shading.

Figure 4.15: Residual evaluation plots for XGBoost model with the same variables as the best linear model: (a) QQ-plot and (b) ACF plot.

As the XGBoost model with backwards selection of variables performs best, it is selected for further comparison with the other models.

4.5 Model Comparison

As previously mentioned, the purpose of this thesis is to assess whether spline models have superior predictive accuracy over linear models, as well as how they perform relative to XGBoost. To assess the generalization capabilities of the linear model versus the various spline models on unseen data, this section includes a closer examination of the two out-of-time periods. It specifically evaluates the accuracy of each model's predictions against the observed default rates within these time frames. Figure 4.16 shows the predicted versus observed default rates during OOT 1 for the different models. During this period, there is a rise in the observed default rate that none of the models seem to be able to capture fully. However, it can be seen that the XGBoost model demonstrates the smallest errors during this period, indicating its superior performance.

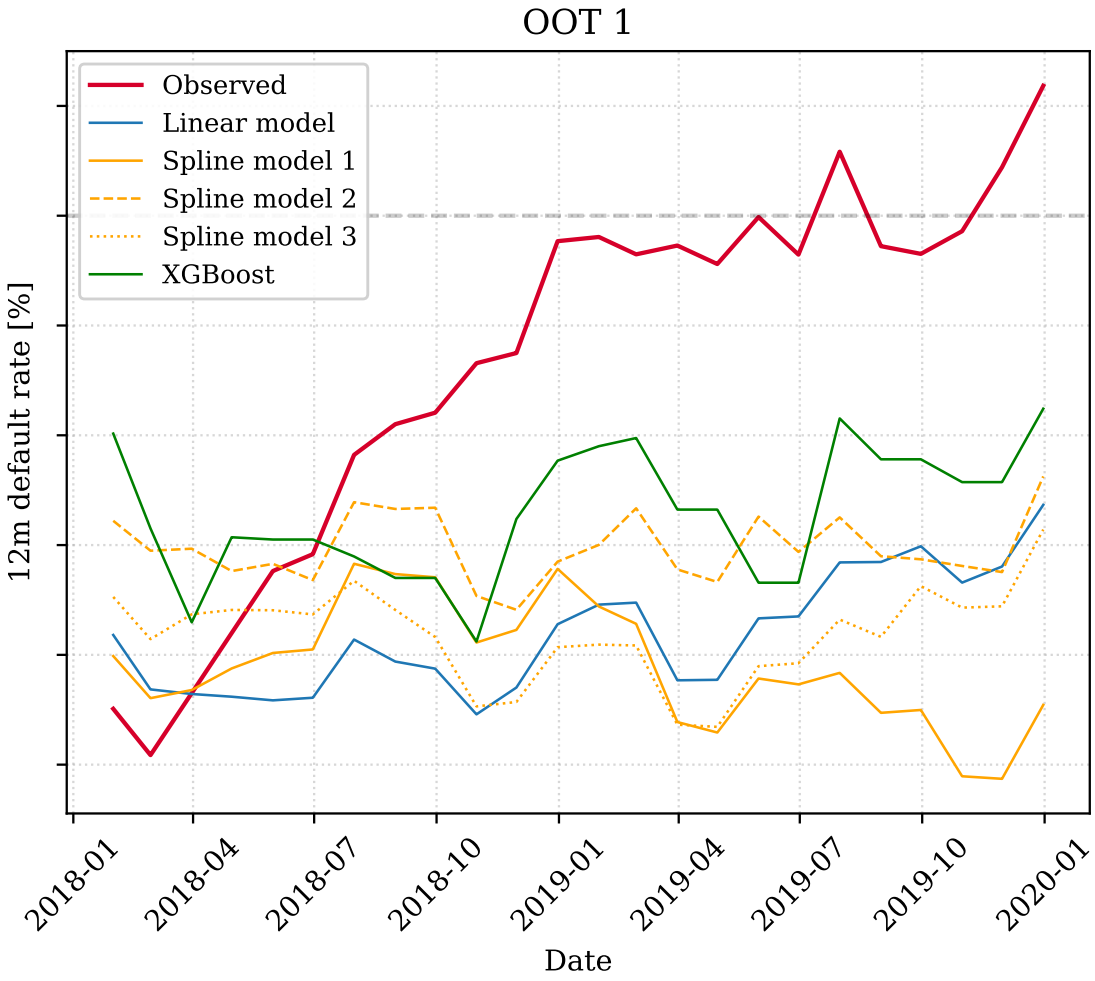


Figure 4.16: Predictive performance of all models on out-of-time period 1. Note the horizontal line to use as a reference for comparison with plot of OOT 2. Y-axis omitted and data masked for confidentiality reasons.

Figure 4.17 shows the predicted versus observed default rates during OOT 2 for the different models. A notable aspect is the spikes in the predictions from spline model 1, where unusually large predictions likely result from the model encountering previously unseen values of the macroeconomic variables. This underscores spline models' sensitivity to unseen data. In contrast, XGBoost demonstrates a more robust handling of these unseen data points, maintaining stable and more accurate predictions. Moreover, note the different limits of the y-axes in the two figures.

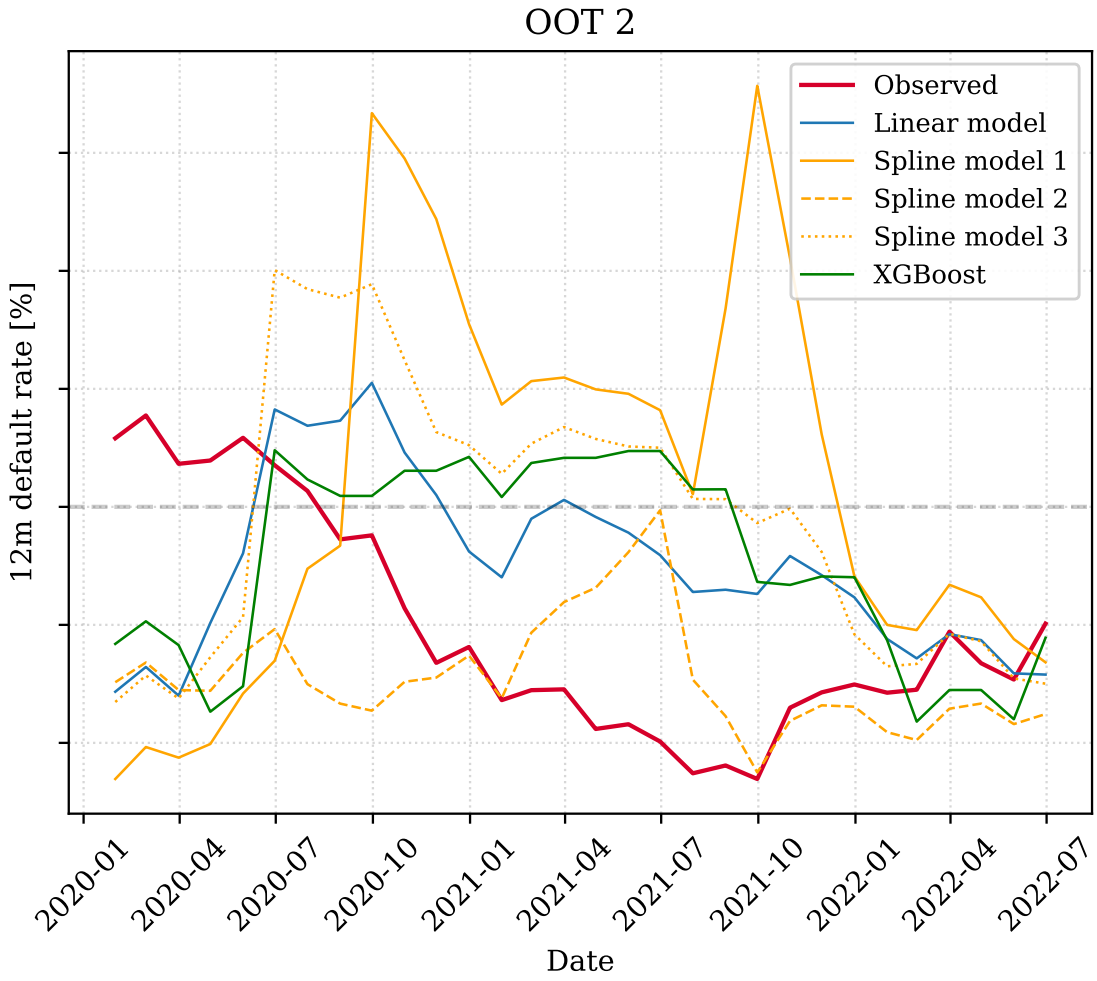


Figure 4.17: Predictive performance of all models on out-of-time period 2. Note the horizontal line to use as a reference for comparison with plot of OOT 1. Y-axis omitted and data masked for confidentiality reasons.

Chapter 5

Analysis & Discussion

5.1 Interpretation of Results

For the convenience of the reader, here is a recap of the spline models discussed in the following section.

Table 5.1: Descriptions of the spline models.

	Description
Spline Model 1	Variables selected from the pool of significant variables from linear single-factor significance test.
Spline Model 2	Variables rejected from the linear RESET test.
Spline Model 3	Variables from best linear model. Splines on both variables.
Spline Model 4	Variables from best linear model. Mix of linear and spline terms.

5.1.1 Predictive accuracy

All spline models, with the exception of spline model 2, show a higher pseudo-R^2 compared to the linear model - indicating a better fit to the training data; however, they do not achieve a better cross-validation RMSE, apart from spline model 4. Additionally, these models do not demonstrate any large improvements in predictive accuracy for OOT periods 1 and 2. This suggests that the spline models may have overfitted to the training data.

Although spline model 2 exhibits the lowest RMSE in out-of-time periods 1 and 2, an inspection of the predicted values in Figure 4.16 reveals that its apparent superiority in RMSE in OOT 1 is attributed to the model's tendency to fluctuate around a higher mean value, rather than effectively capturing underlying data trends. Additionally, the other models fail to predict the increasing PD values observed in the period. The same reasoning applies to OOT 2 (seen in Figure 4.17), where the superior predictive accuracy of spline model 2 appears to be more a result of chance rather than it being a better model.

Analysing the different models' performance during periods of the Covid-19 pandemic (OOT 2), only spline model 2 shows an improvement from the linear model in terms of RMSE and MAPE. Spline models 1 and 3 yielded higher predictions compared to the linear model during the period, aligning with expectations of rising default rates during the pandemic (although the observed default rate decreased). This suggests that while the spline models did not enhance predictive accuracy for OOT 2 in this study, they could potentially excel at forecasting extreme values in future crises, particularly if trained on a larger dataset of similar scenarios, especially considering the pandemic's unique economic distortions due to stimulus measures. These results thus support the statement by Hyndman and Athanasopoulos (2018, Chapter 5.8) that "cubic splines usually give a better fit to the data. However, forecasts of y become unreliable when x is outside the range of the historical data".

In summary, spline model 4 has the lowest RMSE from cross-validation on the in-time set, showing a slight increased accuracy from the best linear model, and spline model 2 exhibits the lowest RMSE and MAPE on both out-of-time periods. The performance of the spline models analysed in this study thus seem to be similar, if not worse, than that of the linear models.

5.1.2 Non-linear dynamics and smoothing

The partial dependence plots, specifically the function estimated for GDP growth lag 3 in Figure 4.3, indicate that there might be a non-linear relationship between this macroeconomic variable and PD. The function derived for the unemployment rate lag 3 appears to suggest a predominantly linear relationship, where the observed fluctuations may be due to insufficient amounts of data across all values of the independent variable rather than due to inherent non-linearity in the underlying relationship. Should the relationship between a macroeconomic factor and PD be linear, Figure 4.10a and 4.6 illustrate that the application of penalized least squares combined with the smoothing parameter, determined through GCV, reduces the dimension of the spline to yield a linear function.

Furthermore, Figure 4.10 demonstrates that the spline terms are capable of capturing a nearly linear yet dynamic response across different regions of the independent variable, such as the unemployment rate in this case. To test this adaptability was a central objective of this study, given the hypothesis that introducing such "thresholds" where the correlations between a macroeconomic variable and PD change could enhance predictive accuracy under turbulent conditions.

As Wahba and Wang (1995) describe, when the sample sizes are small, there is a probability that the GCV estimator of the smoothing parameter λ will yield an extremely small value, resulting in severe under-smoothing of the function $\hat{f}(x_i)$. This issue was encountered in this study, as seen in the low λ values and resulting functions depicted in Figures 4.3, 4.6, and 4.10a. It can be questioned whether these are truly the optimal smoothing parameters, since the function fits appear to be relatively "wiggly", potentially modeling noise rather than capturing an underlying trend between the macroeconomic variable and PD. This issue could be mitigated by having a larger sample of data, or by introducing a more robust estimator for the optimal smoothing parameter, such as the restricted maximum likelihood (REML) estimator, as suggested by Wood (2017). The REML estimator is less prone to over-fitting as its optimum tends to be more pronounced than for GCV, and it is less likely to create phantom minima in the absence of genuine signals in the data. Different smoothing estimators were not tested in this study due to lack of implementation in the `pyGAM` package. Another mitigation to this issue would be to use a better estimator of the

effective degrees of freedom expressed in Section 3.2.2 for a more accurate GCV score estimate.

5.1.3 Residual analysis

The diagnostics of the residuals reveal that while the models generate errors with a normal distribution (indicated by Jarque-Bera p-values larger than 5% and by the QQ-plots), there is substantial autocorrelation across various lag lengths for all models. The reason for this is most likely the inherent autocorrelation in the target variable - being a 12-month PD at a monthly frequency. However, based on the Durbin-Watson test statistics and the ACF plots, the spline models demonstrate reduced autocorrelation in the residuals compared to the linear models. This suggests that the spline models may be better at accounting for the underlying patterns in the data.

To address the time series dependency in the target variable and mitigate autocorrelation issues apparent in all models, introducing an autoregressive term of order 1 (AR(1)) was considered. This adjustment could potentially enhance the accuracy of one-step-ahead predictions. However, it might also impair the performance of long-term forecasts, particularly if the model is intended for use over extended periods, such as projecting 10 to 15 years ahead. Additionally, incorporating lagged values of the dependent variable in a regression model violates the non-stochastic assumption of the independent variables, potentially resulting in biased coefficient estimates for smaller datasets (Brooks, 2019). Therefore, it was ultimately decided not to include this term in the models.

5.1.4 Final model selection

Considering the RMSE values, spline model 4 emerges as the preferable choice due to its superior generalization abilities as suggested by the cross-validation in the in-time sample. This model seems to strike an effective balance by integrating a spline term that captures the non-linear dynamics of the unemployment rate, alongside a linear term for government consumption. One drawback is that the model predicts excessively high values in out-of-time period 2, likely due to encountering macroeconomic values in the test data that were not present in the training set combined with the uncertain asymptotic behaviour of spline models. This issue could potentially be mitigated in the future by expanding the training dataset to include a broader range of macroeconomic conditions and extreme data points, thus enhancing the model's robustness and accuracy in predicting under diverse scenarios.

In discussing spline models versus linear models, an important consideration worth emphasizing is the trade-off between predictive accuracy and model interpretability. Spline models may offer enhanced accuracy; however, the clarity of linear models cannot be overlooked, especially in environments such as banking, where interpretability is crucial due to stringent regulatory requirements. Evaluating whether the improved performance is substantial enough to justify transitioning from a well-understood linear model to a more complex spline model is essential. In this study, the improved cross-validation RMSE of around 0.003 percentage units of spline model 4 may not be large enough to give up the best linear model.

5.1.5 Comparison with XGBoost

The XGBoost model with variables selected through backward elimination, with its CV MAPE of 18.33% compared to 20.87% and 20.76% of the best spline and linear models respectively, shows that it is more accurate and generalizes better than the other models for predicting PD. It also yielded lower MAPE than all other models (except for one spline model) in out-of-time period 1, indicating a better fit to unseen data. Furthermore, as seen in Figure 4.17, XGBoost avoids predicting excessively high values when encountering previously unseen data, demonstrating superior robustness and generalization compared to the spline models. From the Durbin-Watson statistic, it is closer to 2, indicating better handling of autocorrelation in the residuals. This demonstrates that, although it is originally designed to handle extensive datasets and many variables, XGBoost can work effectively also for smaller datasets and only two dependent variables if the hyperparameters are tuned correctly.

This result shows that XGBoost can find complex non-linear dependencies between the macroeconomic variables and PD, offering an even more flexible modeling approach than spline models. However, the greatly reduced interpretability of this machine learning model is a significant downside. While spline models provide a balance between capturing non-linear relationships and maintaining some level of interpretability, XGBoost's complex nature makes it challenging to explain the underlying decision process to stakeholders, especially in highly regulated industries like banking where model transparency is crucial. Therefore, despite XGBoost's superior performance in terms of predictive accuracy, the choice of model should consider the trade-off between accuracy and interpretability based on the specific application requirements.

Additionally, the implementation and use of XGBoost were simpler than fitting a good spline model, particularly when dealing with numerous variables. XGBoost does not require careful checking of partial dependence plots for all macroeconomic variables to ensure the smoothing parameters are correctly set. This makes XGBoost a more practical choice for situations with a high number of independent variables.

In conclusion, the linear models were the least accurate, the spline models showed intermediate performance, and XGBoost was the most accurate in predicting PD. This confirms the hypothesis that spline models can serve as a good intermediary between the highly interpretable but biased linear models and the accurate yet less interpretable machine learning "black box" models.

5.2 Limitations of the Study

One aspect that can be explored further in this study is the quality of the data. As mentioned in section 2, many of the originally quarterly observed macroeconomic variables were interpolated to acquire a monthly frequency on all time series. Consequently, much of the data for the independent variables consists of interpolated rather than directly observed values. This interpolation might impact both the reliability and quality of the data, potentially affecting the robustness of the models trained and tested on it. As described in section 3.1.1, there are more accurate imputation methods than simple linear interpolation that could be used to mitigate these issues. Another approach would be to simply build the models on quarterly frequency to avoid having to impute.

A further limitation arises from the application of the GCV method used to find the optimal smoothing parameter, which is primarily designed for non-time-series data and assumes independent and identically distributed observations of the target variable. In this study, the observed default rate clearly does not fulfill these assumptions as the observations have autocorrelation. This discrepancy may undermine the suitability of GCV, potentially affecting its effectiveness and the validity of estimating the optimal smoothing parameter.

Another challenge in this study is associated with the target variable, namely the default rate. As it is computed based on performing customers that vary each month, the variations in default rate could be attributed to changes within Nordea's portfolio or the departure of customers for reasons unrelated to their credit performance, thereby altering the number of performing customers. This variability introduces uncertainties in the representativeness of this calculated default rate as a "proxy" for actual PD. Ideally, analysing a consistent portfolio or the same set of customers over a period would provide a clearer indication that the observed changes in default rate reflect actual probabilities of default, rather than fluctuations in customer numbers. Additionally, since the PD is calculated on a 12-month basis but reported monthly, there is inherent autocorrelation within the target variable, complicating the modeling process without the use of specific time-series models. The default rate series also indicated non-stationarity from the joint ADF & KPSS test. Despite these complexities, the decision was made to maintain this frequency of reporting for the PD without differencing, to preserve the long-term predictive capabilities and relevance of the models. Another drawback of using monthly data is its high volatility, with the number of defaults varying significantly from month to month, which could potentially be mitigated by aggregating the data quarterly.

Finally, another limitation of this study stems from the restricted time period covered by the data, which includes only a few distinct crises. Unfortunately, the dataset does not encompass enough diverse crisis scenarios to thoroughly train the models on such data. The implications of this are seen in the spline models' predictions in the out-of-time periods. However, with an extension of the dataset to include additional years or more varied crisis data in the future (including wider value ranges of the independent variables), one could get a better understanding of how the spline models' performance might yield an improvement over linear models.

It is also important to acknowledge the inherent difficulty in making predictions during a crisis. Such periods involve unexpected developments and deviations from normal patterns, making accurate modeling and forecasting challenging. Crises are defined by their anomalous nature and the breakdown of established relationships. Therefore, while models can be trained to account for past emergencies, their ability to predict future ones remains limited due to the possibly unique and unforeseen circumstances that define each new crisis.

Chapter 6

Conclusion

6.1 Answering the Research Questions

This thesis aimed to evaluate the efficacy of spline regression compared to traditional linear models and "black box" ML models such as XGBoost in predicting the probability of default under varying macroeconomic conditions. The central research questions were:

1. Can spline regression models provide a more accurate prediction of PD than linear models, reflecting non-linear dynamics between macroeconomic variables and PD?

2. How do spline models compare to XGBoost in terms of predictive accuracy?

The analysis lend support to the expectation that spline models, particularly spline model 4 (which consists of a mix of linear and spline terms), displayed a slightly improved accuracy in cross-validation on the in-time dataset compared to linear models. However, this did not translate into superior performance during out-of-time periods, suggesting potential overfitting to the training data. The spline models' inability to consistently outperform linear models suggests limitations in their current configuration, particularly in handling extreme values outside the range of the training data.

Overall, while spline models hold promise for modeling non-linear relationships, their application in PD prediction requires careful consideration of their tendency to overfit and their sensitivity to the range of input data.

Moreover, the utility of spline models extends beyond predictive accuracy. These models can provide critical insights into the relationships between macroeconomic variables and PD. Spline models can serve as a tool for preliminary analyses to reveal partial dependencies and unusual patterns in the data. By visualizing these relationships, practitioners can determine whether the interaction between a macroeconomic variable and PD exhibits non-linear characteristics before committing to more restrictive model forms like linear regression.

In comparison to XGBoost, the results demonstrated that XGBoost outperformed both linear and spline models in predictive accuracy. With an average MAPE from cross-validation of 18.33% (compared to 20.76% and 20.87% for the optimal linear and spline model respectively), XGBoost showed superior accuracy and generalization capabilities. It avoided excessively high predictions on unseen data, highlighting its robust handling of diverse macroeconomic conditions. However,

the trade-off is the reduced interpretability of the model, which can be a drawback in industries requiring high transparency.

In conclusion, while spline models were slightly more accurate than linear models, they did not achieve the same level of predictive performance as XGBoost. This demonstrates that spline models can serve as a valuable intermediary, balancing interpretability and accuracy between traditional linear models and more complex machine learning approaches like XGBoost.

6.2 Future Research

For future research, it would be interesting to analyse the stability of spline models in response to new data. Similarly to how one tests the stability of coefficient estimates of a linear model, it would be interesting to analyse how the smooth fits of the spline models evolve as additional data points are introduced. Such an analysis was not included in this thesis due to time constraints, but it is critical for understanding the robustness of these models over time.

Future research could also investigate the impact of using different spline basis functions on model fit and performance. This study utilized penalized B-splines; however, exploring other bases such as the "thin plate regression splines" as used by B. Zou et al. (2016), which do not require selecting knot locations, could enhance model objectivity and robustness. Thin plate regression splines is based on radial basis functions (RBF), as explained by Naumann et al. (2020), which inherently accommodate multivariate inputs, making them particularly suitable for modeling complex, non-linear interactions between multiple risk factors in credit risk analysis. As stated earlier, RBFs do not require the selection of knot locations, thereby increasing model robustness and objectivity by eliminating a source of potential bias.

Lastly, investigating if different estimators for the smoothing parameter λ, such as the REML method, could yield improvements over the GCV method for the endeavors of this study would be of interest.

Bibliography

Ali, A., & Daly, K. (2010). Macroeconomic determinants of credit risk: Recent evidence from a cross country study. *International Review of Financial Analysis*, *19*(3), 165–171.

Antonsson, H. (2018). *Macroeconomic factors in probability of default* [Master's thesis, KTH Royal Institute of Technology].

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *28*(1), 171–182.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.

Brooks, C. (2019). *Introductory econometrics for finance* (4th ed.). Cambridge University Press.

Brooks, C., Rew, A. G., & Ritson, S. (2001). A trading strategy based on the lead–lag relationship between the spot index and futures contract for the ftse 100. *International Journal of Forecasting*, *17*(1), 31–44.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (2018). *Graphical methods for data analysis*. Chapman; Hall/CRC.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chow, G. C., & Lin, A.-l. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics*, 372–375.

Cuche, N. A., & Hess, M. K. (1999). *Estimating monthly gdp in a general kalman filter framework: Evidence from switzerland* (tech. rep.). Working Paper.

De Boor, C. (1978). *A practical guide to splines* (Vol. 27). Springer-Verlag New York.

Durbin, J., & Watson, G. S. (1951). Testing for serial correlation in least squares regression. ii. *Biometrika*, *38*(1/2), 159–177.

European Banking Authority. (2023). Final report on IFRS 9 implementation by EU institutions [Accessed: 2024-02-20]. https://extranet.eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2023/1063709/Final%20Report%20on%20IFRS9%20implementation%20by%20EU%20institutions.pdf

Frykström, N., & Li, J. (2018). *IFRS 9 – the new accounting standard for credit loss recognition* (Economic commentaries) (Accessed: 2024-02-20). Sveriges Riksbank. https://www.riksbank.se/globalassets/media/rapporter/ekonomiska-kommentarer/engelska/2018/ifrs-9--the-new-accounting-standard-for-credit-loss-recognition.pdf

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Hild, A. (2021). *Estimating and evaluating the probability of default - a machine learning approach* [Master's thesis, Uppsala University].

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1), 159–178.

Li, R., Cui, L., Fu, H., Meng, Y., Li, J., & Guo, J. (2020). Estimating high-resolution PM1 concentration from Himawari-8 combining extreme gradient boosting-geographically and temporally weighted regression (XGBoost-GTWR). *Atmospheric Environment*, *229*, 117434.

Li, S., Tian, S., Yu, Y., Zhu, X., & Lian, H. (2022). Corporate probability of default: A single-index hazard model approach. *Journal of Business Economic Statistics*, *41*, 1–32.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529.

Naumann, C., Glänzel, J., & Putz, M. (2020). Comparison of basis functions for thermal error compensation based on regression analysis–a simulation based case study. *Journal of Machine Engineering*, *20*.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, *31*(2), 350–371.

Schlitzer, G. (1995). Testing the stationarity of economic time series: Further monte carlo evidence. *Ricerche Economiche*, *49*(2), 125–144.

Servén, D., & Brummitt, C. (2020). pyGAM Documentation.

Siems, J., Ditschuneit, K., Ripken, W., Lindborg, A., Schambach, M., Otterbach, J., & Genzel, M. (2024). Curve your enthusiasm: Concurvity regularization in differentiable generalized additive models. *Advances in Neural Information Processing Systems*, *36*.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Ventrucci, M., & Rue, H. (2016). Penalized complexity priors for degrees of freedom in bayesian p-splines. *Statistical Modelling*, *16*(6), 429–453.

Wahba, G., & Wang, Y. (1995). Behavior near zero of the distribution of gcv smoothing parameter estimates. *Statistics & Probability letters*, *25*(2), 105–111.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Wood, S. N. (2017). *Generalized additive models: an introduction with R* (2nd ed.). Chapman and Hall/CRC.

Yang, Z., Zhang, A., & Sudjianto, A. (2021). Gami-net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognition*, *120*, 108192.

Zou, B., Chen, J., Zhai, L., Fang, X., & Zheng, Z. (2016). Satellite based mapping of ground PM2.5 concentration using generalized additive modeling. *Remote Sensing*, *9*(1), 1.

Zou, M., Jiang, W.-G., Qin, Q.-H., Liu, Y.-C., & Li, M.-L. (2022). Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting. *Materials*, *15*(15), 5298.

Appendix

Evolution of macroeconomic variables in data period

Below the evolution of all macroeconomic variables that were included in the final models selected is presented, over the entire training and test sets used.

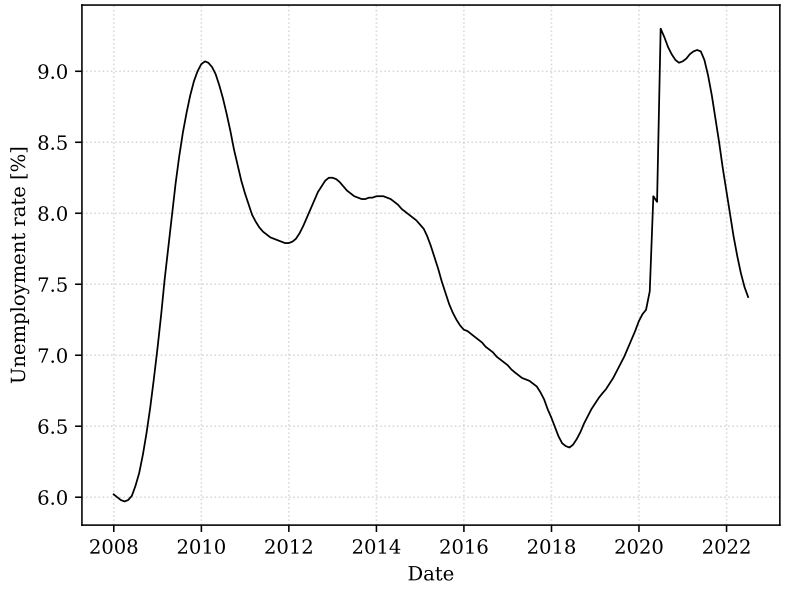


Figure 1: Trajectory of the unemployment rate throughout the time period under study.

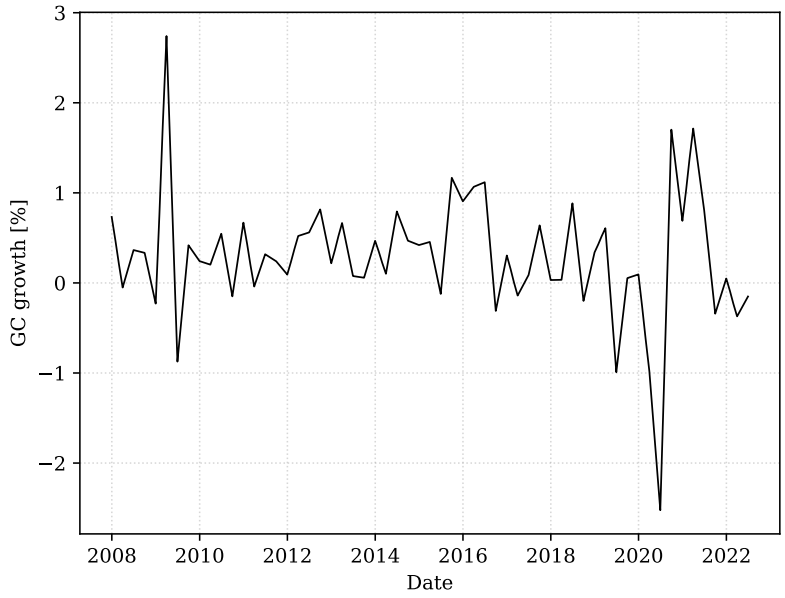


Figure 2: Trajectory of the government consumption growth throughout the time period under study.

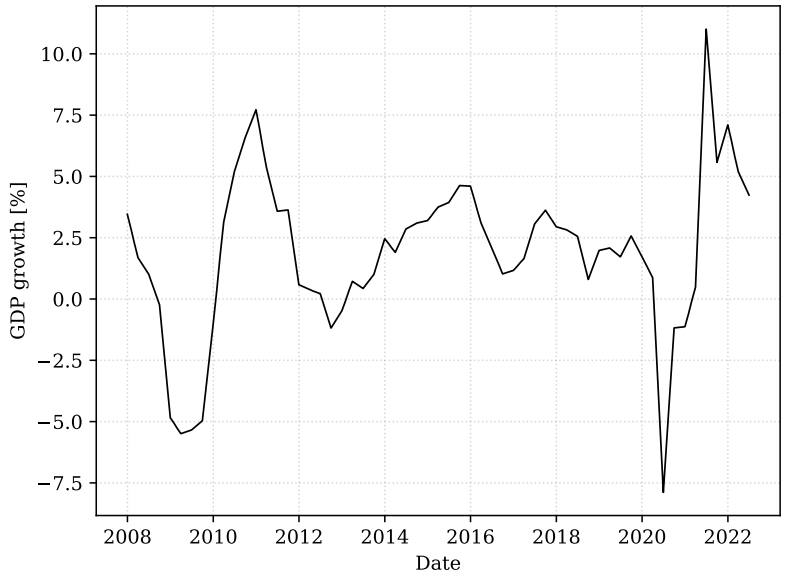


Figure 3: Trajectory of the GDP growth throughout the time period under study.

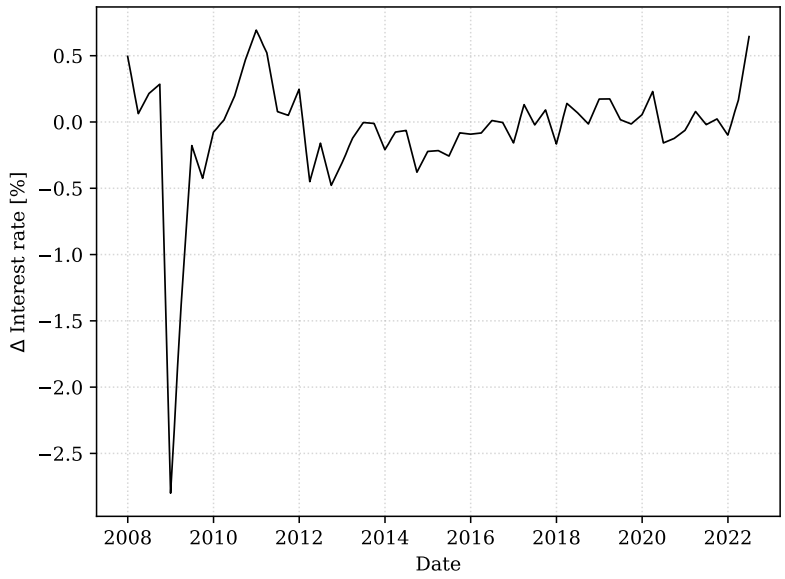


Figure 4: Trajectory of the differenced short-term interest rate throughout the time period under study.

Master's Theses in Mathematical Sciences 2024:E44
ISSN 1404-6342

LUTFMS-3499-2024

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/