UMEÅ UNIVERSITY

STRESS TESTING AN SME PORTFOLIO

Effects of an Adverse Macroeconomic Scenario on Credit Risk Transition Matrices

Siri Almqvist and Oskar Nordin

Master thesis, 30 credits

Master of Science in Industrial Engineering and Management, 300 credits

Spring term 2021

Stress Testing an SME Portfolio
Effects of an Adverse Macroeconomic Scenario on Credit Risk Transition Matrices
Department of Mathematics and Mathematical Statistics
Umeå University
SE - 907 87 Umeå, Sweden

Supervisor:
Marcus Olofsson, Umeå University

Examiner:
Antti Perälä, Umeå University

Abstract

The financial crisis of 2007-2008 was a severe global crisis causing a worldwide recession. One of the main contributing factors of the crisis was the excessive risk appetite of banks and financial institutions. Since then, regulatory authorities and financial institutions have directed focus towards risk management with the main objective to avert a similar crisis from occurring in the future. The aim of this thesis is to investigate how an adverse macroeconomic scenario would affect the migrations between risk classes of an SME portfolio, referred to as stress test.

This thesis utilises two frameworks, one by Belkin and Suchower and one by Carlehed and Petrov, for creating a single systematic indicator describing the credit class migrations of the portfolio. Four different regression model setups (Ordinary Least Squares, Additive Model, XGBoost and SVM) are then used to describe the relationship between macroeconomic indicators and this systematic indicator. The four models are evaluated in terms of interpretability and ability to predict in order to find the main drivers for the systematic indicator. Their corresponding prediction errors are compared to find the best model. The portfolio is stress tested by using the regression models to predict the corresponding systematic indicator given an adverse macroeconomic scenario. The probability of default, estimated from the indicator using each of the frameworks, are then compared and analysed with regards to the systematic indicator.

The results show that unemployment is the main driver of the risk class migrations for an SME portfolio, both from a statistical and economical perspective. The most appropriate regression model is the additive model because of its performance and interpretability and is therefore advised to use for this problem. From the PD estimations, it is concluded that the framework by Belkin and Suchower gives a more volatile estimate than that of Carlehed and Petrov.

Keywords: Stress test, SME, Transition Matrix, Credit Risk, Statistical Analysis, Machine Learning

Sammanfattning

Finanskrisen som ägde rum 2007-2008 var en global kris som inledde en omfattande lågkonjunktur världen över. En av de främst bidragande faktorerna till krisen var bankers och finansiella institutioners omättliga riskbenägenhet. Sedan dess har reglerande myndigheter och finansiella institutioner riktat större fokus mot riskhantering med det huvudsakliga målet att förhindra att en liknande kris återupprepas i framtiden. Målet med uppsatsen är att undersöka hur ett ogynnsamt makroekonomiskt scenario skulle påverka migrationer mellan riskklasser för en portfölj av Små och Medelstora Företag, vilket brukar kallas för ett stresstest.

Uppsatsen utnyttjar två ramverk, ett av Belkin och Suchower och ett av Carlehed och Petrov, för att ta fram en systematisk indikator som beskriver riskklassmigrationer för portföljen. Fyra olika regressionsmodeller (Ordinary Least Squares, Additive Model, XGBoost och SVM) används för att beskriva förhållandet mellan makroekonomiska variabler och den systematiska indikatorn för att hitta faktorerna med störst inverkan. För att stresstesta portföljen används regressionsmodellerna för att prediktera ett värde för den systematiska indikatorn vid ett givet ogynnsamt makroekonomiskt scenario. Sannolikheterna för fallissemang som ges av de två olika ramverken studeras och analyseras utifrån ett ekonomiskt perspektiv.

Resultaten visar att arbetslöshet har störst inverkan för riskklassmigrationer för en SME-portfölj, både från ett statistiskt samt ekonomiskt perspektiv. Den mest ändamålsenliga regressionsmodellen för skattning av den systematisk indikatorn är den additiva modellen på grund av dess lämplighet och tolkningsbarhet. Från estimeringen av sannolikheten för fallissemang kan det konkluderas att ramverket av Belkin och Suchower ger en mer volatil uppskattning relativt ramverket av Carlehed och Petrov.

Contents

List of Abbreviations

AIRB	Advanced Internal Ratings Based Approach
AM	Additive Model
BCBS	Basel Committee of Banking Supervision
BIS	Bank for International Settlements
EBA	European Banking Authority
FIRB	Foundation Internal Ratings Based Approach
MAE	Mean Absolute Error
MSE	Mean Square Error
OLS	Ordinary Least Square
PD	Probability of Default
PIT	Point-in-Time
RSS	Residual Sum of Squares
SME	Small and Medium Enterprises
SVM	Support Vector Machine
TTC	Through-the-Cycle

List of Figures

List of Tables

1 Introduction

The financial crisis of 2007-2008 was a severe global crisis that caused a worldwide recession. One of the main contributing factors of the crisis was the excessive risk appetite of banks and financial institutions. Since then, regulating authorities and financial institutions have directed focus towards risk management with the main objective to avert a similar crisis from occurring in the future. This thesis is written at a large bank in Sweden within the subject of risk management, and focuses on the effect from adverse macroeconomic scenarios on portfolios.

1.1 Background

A bank is a financial institution licensed to issue loans and provide financial services such as currency exchange. The largest undertaking of banks are loans, making credit risk, the risk of a counterparty defaulting on loans and derivative transactions, the greatest risk to manage [Hull, 2018, p.42]. In light thereof, there is a constant focus on risk management and assessment of risks and their continuous change. If the credit risk of a bank becomes too pronounced, the risk of default increases which can cause a ripple effect of defaulting financial institutions, causing disruption in the financial system. This is called systemic risk and is the reason domestic and international regulations are enforced to ensure financial stability in the market [Hull, 2018, p.348]. The demand for accurate estimations of credit risk for risk management and capital planning requires constant reevaluation and development of new risk models. This thesis seeks to develop a model for assessment of credit risk with a macroeconomic perspective.

1.1.1 Basel

Prior to the Basel accord, countries regulated national banks independently, usually by setting a minimum ratio of required capital to total assets [Hull, 2018, p.348] . This gave rise to asymmetric risk distribution where countries with less strict regulations were considered to have a competitive edge compared to those with stricter enforced capital regulations. The purpose of regulating banks is to ensure that they keep enough capital to withstand stressed market conditions. The lack of international regulations and the need to strengthen financial stability by improve worldwide banking supervision led to the first gathering of the Basel Committee of Banking Supervision (BCBS) in 1975 [BCBS, 2019]. The committee is part of an organisation for central

banks called the Bank for International Settlements (BIS) and initially consisted of 12 member countries. The first of three accords, Basel I, was implemented in 1988 as a first attempt to implement international risk based capital adequacy requirements [BCBS, 2019].

Standardized Approach The standardized approach, as presented in Basel I, is used for calculation of risk weighted assets, reflecting the accumulated credit risk of a bank. The approach provides risk classes with associated risk weights where, for example, loans to corporations has a risk weight of 100%, while loans to banks and government agencies has a risk weight of 20% due to being considered a sounder investment. The total risk weighted assets under the standardized approach by Basel I is given by

$$RWA = \sum_{i=1}^{N} w_i L_i,$$

where w_i denotes risk weights and L_i denotes the principal amount. According to Basel I, banks are required to hold 8% of the total risk weighted assets to ensure the ability to withstand a potential financial crisis. The required capital is therefore

$$\text{Required Capital} = 0.08 \times \text{RWA}.$$

Internal Rating Based Approach Basel II, an expansion of Basel I, allows for the use of internal models during the assessment of capital requirements. The capital requirements are the value at risk subtracted by the expected loss, and is regulated to be based on a one year horizon using a 99.9% confidence interval [Hull, 2018, p.363]. There are two options for assessment of capital requirements: the foundation internal rating-based approach (FIRB) and the advanced internal rating-based approach (AIRB). Under the AIRB approach, parameters as loss given default (LGD), probability of default (PD), exposure at default (EAD) and maturity may be estimated using internal models. On the other hand, under the FIRB approach, only the PD may be estimated using internal models [Murphy, 2008, p.289]. The regulated required capital is defined as

$$\text{Required Capital} = \sum_i \text{EAD}_i \times \text{LGD}_i \times (\text{WCDR}_i - \text{PD}_i),$$

where WCDR, which indicates the worst case default rate, is defined as

$$\text{WCDR}_i = \Phi[\frac{\Phi^{-1}(\text{PD}_i) + \sqrt{\rho}\Phi^{-1}(0.999)}{\sqrt{1-\rho}}],$$

where PD_i is the one-year default probability of the ith obligor, ρ is the copula correlation between the the pairs of obligors and Φ is the standard normal cumulative distribution function [Hull, 2018, p.364].

1.1.2 Rating Classification

Since credit risk accounts for the largest part of a bank's risk appetite, it is important to have an established process for evaluation of the degree of risk associated with various counterparties [Hull, 2018, p.42]. This is managed using a variety of methods, one being risk classifications. Risk classifications are widely used in the financial industry and can be applied to a company, portfolio or a bond. Credit ratings give an indication of the soundness of an investment by assessment of the creditworthiness of a counterparty. The rating indicates the probability of default by a specific counterparty within a certain time frame [Bluhm, 2003, p.13].

Risk classifications vary in form depending on the distributor, although a common disposition is ordered alphabetical ratings such as AAA, AA, A, BBB, BB, B, CCC, and D where D represents the state of default. This is the disposition of Standard & Poor, a company providing risk classifications for firms and companies. This credit rating disposition is used throughout the thesis to exemplify methods and results.

1.1.3 Transition Matrices

Transition matrices describe the probability of a counterparty of a specific rating, migrating to another rating during a certain time frame. Thus, transition matrices are used for analysis of historical changes in credit ratings. For example, in the case of the ratings of Standard & Poor, a transition matrix is structured as set out in Figure 1.

$$
\begin{array}{cccc}
\mathrm{P}_{AAA->AAA} & \mathrm{P}_{AAA->AA} & \cdots & \mathrm{P}_{AAA->D} \\
\mathrm{P}_{AA->AAA} & \mathrm{P}_{AA->AA} & \cdots & \mathrm{P}_{AA->D} \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{P}_{CCC->AAA} & \mathrm{P}_{CCC->AA} & \cdots & \mathrm{P}_{CCC->D}
\end{array}
$$

Figure 1: Disposition of a transition matrix.

Each row in Figure 1 symbolises the initial state of the credit rating for a potential counterparty, whereas each column represents ending up in a specific credit rating at a given point in time. Meanwhile, each cell in the matrix denotes the probability of a counterparty migrating from some credit rating at a given point in time to some credit rating at a later given point in time. The rightmost column represents migrations to the default state, meaning that the column as a whole is used to estimate overall probability of default, a useful statistic for assessment and management of credit risk.

1.1.4 Stress Testing

Following the financial crisis of 2008, flaws in the banking system were shed light on. As a result, stricter regulations were introduced by several institutions such as the European Banking Authority (EBA). EBA is an independent European authority, working to ensure effective and consistent regulations in the European banking sector. Their main objective is to sustain financial stability by defining harmonising rules for financial institutions in Europe [The European Banking Authority, 2016]. One of the rules enforced by EBA is stress testing. Stress tests are simulated hypothetical unfavorable scenarios the banks undertake to test their capability to withstand adverse financial shocks. The simulated shocks can be based on historical data such as the strong recession that followed the financial crisis of 2008 or strongly enhanced economic variables such as an increase in unemployment rate and a decrease in GDP [Hull, 2018, p.497-503].

EU-wide stress tests are conducted to evaluate the resilience of financial institutions during adverse market developments [The European Banking Authority, 2021]. The tests are performed to obtain an understanding of the amount of capital required for financial institutions to stay solvent during stressed scenarios and to provide valuable information in overall risk management and capital planning [Finansinspektionen, 2016]. Similar stress tests are also required and conducted on a domestic level for

financial institutions in Sweden [Riksbanken, 2019]. The domestic stress tests in Sweden are required mainly for the four largest banks and take five macroeconomic variables into account: GDP, property prices, inflation, unemployment, and equity prices [Riksbanken, 2020].

From an endogenous perspective, a bank can use the results from a stress test as a forward-looking management tool for identification, monitoring and assessment of risk [BCBS, 2018, p.8]. The results may be used to adjust risk appetite, financial and capital planning as well as liquidity and funding management. In addition, findings from the stress test may be used in internal capital adequacy assessments and as support for internal policies. One of the undertakings of a monetary authority is to ensure that the results from stress tests may be used as a supervisory tool. From an exogenous perspective, the results from the stress tests may therefore be used as a basis for macroprudential decisions. [BCBS, 2018, p.9].

1.2 Literature Framework

Two previous articles that this theses builds upon are "A One-Parameter Representation of Credit Risk and Transition Matrices" and "A Methodology for Point-in-Time-Through-the-Cycle Probability of Default Decomposition in Risk Classification Systems" [Belkin, Suchower, 1998], [Carlehed, Petrov, 2012]. Both papers address the problem of describing risk class migrations using only one parameter, but with different approaches. Belkin and Suchower use historic transition matrices as a base for the creation of a framework describing movements in migrations depending on macroeconomic changes. The framework includes a single parameter describing the migrations based on the matrices. Carlehed and Petrov utilise a method for describing the same parameter but with a base in historical default frequencies. Both frameworks are developed using the same data, but with a distinct difference in their methods. Carlehed and Petrov's method allows for use of a single index for estimation of the parameter while Belkin and Suchower's requires the whole transition matrix. Both of the papers are further explained in the Section 2.1.

There are advantages of both methods. When more observations are needed but historic transition matrices cannot be accessed, the method of Carlehed and Petrov is useful. The method of Belkin and Suchower is useful in that it allows to estimate the entire transition matrix, which is a clear advantage of this method over the method by

Carlehed and Petrov. In literature, as far as the authors of this thesis are concerned, there has been no previous research combining these two approaches. Therefore, the academic contribution of this thesis is to combine the methods and evaluate the resulting probability of default. Furthermore, as an extension of the two papers, the relationship between the systematic indicator parameter and macroeconomics variables is investigated in this thesis.

1.3 Project Description

In this project thesis, the relationship between macroeconomic variables and credit class migrations for a Small and Medium sized Enterprises (SME) portfolio is investigated. A framework used for credit risk representation is utilised in order to express the migrations using only one variable. In addition, different types of parametric and non-parametric regression models are used to examine the relationship between the macroeconomic variables and the credit risk representative variable. The models' inference and predictive abilities are scrutinised in terms of different evaluation metrics and from an economic perspective. An adverse macroeconomic scenario is applied to investigate each of the models' predicted changes in the risk of the investigated portfolio. Finally, the thesis seeks to compare the different models to determine the best performing one.

The research questions addressed in this thesis are:

- What are the main macroeconomic drivers for transitions of credit ratings?
- What statistical or machine learning technique is the most suitable for estimating the systematic indicator using the one-parameter framework?
- How does the different frameworks of Belkin and Suchower and Carlehed and Petrov affect the probability of default?

1.3.1 Purpose

The purpose of the thesis is to investigate the effects of macroeconomic changes on risk class migrations for an SME portfolio. The thesis therefore seeks to examine the possibility to use statistical and machine learning models for describing the effects of macroeconomic changes on risk class migrations. The aim is that the findings should provide the bank with a deeper understanding of what is affecting their risk, and to enable well argued risk management decisions.

1.4 Delimitations

The models created in this thesis are intended for use as a part of a larger process with the main goal of obtaining a value of the required economic capital. The input, further described in Section 3.1, is a set of transition matrices describing the migration probabilities of a portfolio on quarterly basis. The thesis is delimited to obtaining a model with corresponding stressed transition matrices and does not cover further investigation of the process such as the economics of the stress test. Further delimitation include the investigation of models for only one portfolio and country due to the limited time of the thesis. The thesis is very limited by the data in terms of number of observation. In addition, the model is intended to be used for the EBA stress test which includes a given set of macroeconomic variables. The thesis is therefore limited to investigate the impact of variables in this set.

2 Theory

In this section, the theory used in this thesis is presented. The first part introduces the framework of calculations for the response variable based on two papers, as previously described in Section 1.2. In the second part, the theory corresponding to the creation of the regression models for explaining the relationship between the response variable and the macroeconomic variables is given. Lastly, essential evaluation metrics and theory for hyperparameter tuning for the regression models are presented.

2.1 One-Parameter Representation of Risk

The thesis extends the framework for representation of credit risk and transition matrices introduced in two articles. The first article is "The One-Parameter Representation of Risk" by Belkin, Suchower [1998]. In the paper a normally distributed credit change indicator is assumed to be present in credit rating transition matrices. The credit change indicator, denoted X, is divided into two, an idiosyncratic component, Y, and a systematic component, Z. That is,

$$X = \sqrt{1-\rho}Y + \sqrt{\rho}Z, \tag{1}$$

where ρ is the assumed non-negative correlation between X and Z. Y and Z are assumed to be independent normal unit random variables and mutually independent. Conditional on an initial credit rating G at the beginning of a given time frame, the sample space of X can be partitioned into a set of disjoint bins $(x_g^G, x_{g+1}^G]$. The bins are defined such that the probability of X falling within a given interval is equal to the corresponding historical average transition rate to the corresponding rating. Hence, the bins are defined as

$$P(G,g) = \Phi\left(x_{g+1}^G\right) - \Phi\left(x_g^G\right), \tag{2}$$

where $P(G,g)$ denotes the historical average probability of transitioning from G-to-g and $\Phi(\cdot)$ represents the standard normal cumulative distribution function defined as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt.$$

The highest rated bin has an upper threshold of $+\infty$ and the lowest rated bin has a lower threshold of $-\infty$. A visual representation of the bins is given in Figure 2.

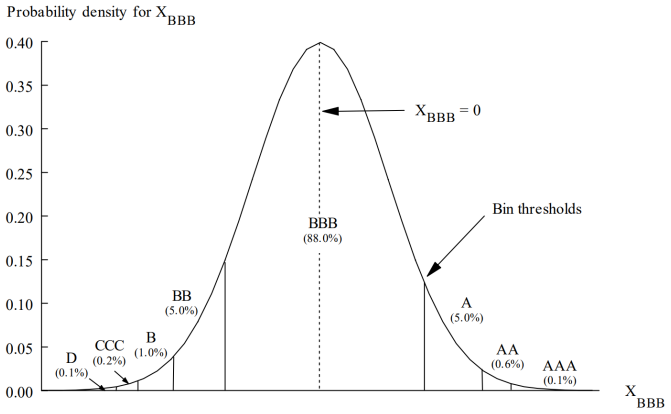


Figure 2: Diagram for exemplifying bin thresholds in a normal probability density function for credit ratings ranging from AAA to D.

Source: Belkin, Suchower [1998]

Given Equation 1, the historical average transition probabilities are acquired when $Z = 0$. For a given time frame, a value of Z is found using the corresponding average transition rates applied to Equation 2. Hence, Z describes the deviation of the transition probabilities from the average and is positive (negative) in times when the general probabilities of migration to superior credit ratings are higher (lower) than average. A higher (lower) Z corresponds to a lower (higher) probability of default. A positive Z can be though of as a flourishing economy.

The value of Z is denoted Z_t for a certain point in time t and is determined through minimization of the weighted mean squared discrepancies between the average transitions probabilities and the observed transition probabilities as

$$\min_{Z_t} \sum_{G} \sum_{g} \frac{n_{t,g} \left[P_t(G,g) - \Delta\left(x_{g+1}^G, x_g^G, Z_t\right) \right]^2}{\Delta\left(x_{g+1}^G, x_g^G, Z_t\right) \left[1 - \Delta\left(x_{g+1}^G, x_g^G, Z_t\right) \right]}, \tag{3}$$

where $n_{t,g}$ is the observed number of transitions for a time t from grade G, and

$$\Delta\left(x_{g+1}^G, x_g^G, Z_t\right) = \Phi\left(\frac{x_{g+1}^G - \sqrt{\rho} Z_t}{\sqrt{1-\rho}}\right) - \Phi\left(\frac{x_g^G - \sqrt{\rho} Z_t}{\sqrt{1-\rho}}\right). \tag{4}$$

The second article, by Carlehed, Petrov [2012], extends the framework by assessing the impact of current economic cycles and assumed time perspective in the initial credit ratings.

Point-in-Time Point-in-time (PIT) assumes that credit ratings are classified with a one year perspective, meaning that the rating is believed to remain in the same state for at least one year. The PIT probability of default (PIT PD) denoted $p_i(z)$, is the probability that a counterparty will default within 12 months given a systematic risk factor z.

Through-the-Cycle Through-the-cycle (TTC) assumes that the credit ratings are independent of market fluctuations. The TTC probability of default (TTC PD), denoted as q_i, is constant over time and does not fluctuate due to changes in the economic state (although it can oscillate due to fluctuations of individual obligors). TTC PD, denoted as q_i, is obtained through the average stationary PIT as

$$q_i = E_Z[p_i(Z)] = \int_{-\infty}^{\infty} p_i(z)\phi(z)dz, \tag{5}$$

where ϕ is the standard normal distributions density function.

By solving for Y in the standard one-factor Merton model in Equation 1, the hybrid probability of default is obtained by

$$p_i(Z) = P[X_i < B_i|Z] = \Phi(\frac{B_i - \sqrt{\rho}Z}{\sqrt{1-\rho}}), \tag{6}$$

where ρ is the correlation between X_i and Z, Φ is the standard normal cumulative distribution function and B_i is an obligor specific constant. The correlation is assumed to be constant over time and only depends on the sector, not on the individual obligor. Due to these assumed simplifications, the theory is applied and the following probability for a portfolio is obtained

$$p_P(Z) = \Phi(\frac{B - \sqrt{\rho}Z}{\sqrt{1-\rho}}), \tag{7}$$

where B denotes the average obligor. Inverting Equation 7 results in

$$Z_t = \frac{B - \Phi^{-1}(d_t)\sqrt{1-\rho}}{\sqrt{\rho}} \tag{8}$$

where d_t denotes the probability of default for time t.
Methods of moments is used for estimating the series Z. From Equation 7 it can be seen that

$$E[\Phi^{-1}(p)] = \frac{B}{\sqrt{1-\rho}} \tag{9}$$

and

$$V[\Phi^{-1}(p)] = \frac{\rho}{1-\rho}. \tag{10}$$

Only the default series d is of importance for obtaining the Z series under this framework. The default series is transformed by applying $\Phi(d)$ from which a mean (m) and standard deviation (σ) are derived. By application of this and solving for B, the following can be derived.

$$B \approx \frac{m}{\sqrt{1+\sigma^2}} \tag{11}$$

and

$$\rho \approx \frac{\sigma^2}{1+\sigma^2}. \tag{12}$$

When Equation 11 and 12 are inserted into Equation 8, the following is obtained

$$Z_t = \frac{m - \Phi^{-1}(d_t)}{\sigma}. \tag{13}$$

Equation 13 is used for obtaining a series $\{Z_t\}$ with an inherent indication of the impact from economic cycles and systematic risk.

In the case of an unknown time perspective of the credit ratings, the probabilities can be generalized and transformed to either TTC or PIT, respectively. This is done by deriving a parameter $0 \leq \alpha \leq 1$, denoting the degree of "PIT-ness" of the initial credit ratings where a value of 1 denotes a 100% PIT model. In the case of a hybrid model, α is in the interval $(0, 1)$ and the value of α is obtained by taken the difference between two points in time with respect to Z. Consider the two points in time

$$\phi^{-1}(p_{i,\alpha}(Z_1)) = \frac{B_i - \sqrt{\rho}\alpha Z_1}{\sqrt{1-\rho\alpha^2}} \tag{14}$$

and

$$\phi^{-1}(p_{i,\alpha}(Z_2)) = \frac{B_i - \sqrt{\rho}\alpha Z_2}{\sqrt{1-\rho\alpha^2}}. \tag{15}$$

The difference between the two points results in the right hand side of Equation 16 while the left hand side is the average of the inverse standard normal distribution density function of the default probability for all time points. α is obtained by solving Equation 16 for every time point.

$$\Delta(\frac{1}{P}\sum_{i\epsilon P}\Phi^{-1}(p_{i,\alpha}(Z))) = \frac{\sqrt{\rho}\alpha\Delta Z}{\sqrt{1-\rho\alpha^2}}. \tag{16}$$

For simplification purposes, the average of all αs is derived, making α constant over time. When applying this, a 100% TTC probability is obtained by deriving the expected value for the time period by mimicking previous steps as follows

$$q_i = \Phi[\sqrt{\rho}\alpha Z_t + \sqrt{1-\rho\alpha^2}\Phi^{-1}(p_{i,\alpha})]. \tag{17}$$

By repeating previous steps, the corresponding 100% TTC probability is transformed to a 100% PIT probability by applying Equation 18 as

$$p_i(z) = \Phi[\frac{\Phi^{-1}(q_i) - \sqrt{\rho}z}{\sqrt{1-\rho}}]. \tag{18}$$

The result from calculations using Equation 17 and 18 result in the PD as shown in Figure 3, where the PD for TTC is constant and the PD for PIT is more volatile.

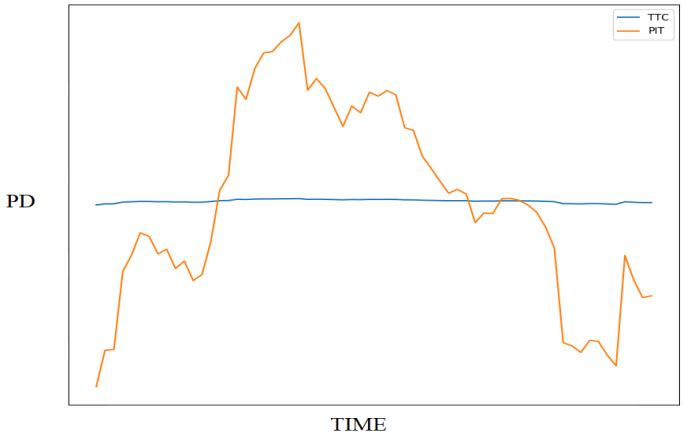


Figure 3: Line plot of typical behaviour for TTC PD and PIT PD with respect to time.

2.2 Regression Models

In this section, a set of regression models used to model the response variable Z to a set of predictors are presented with corresponding theory.

2.2.1 Linear Model

A multiple linear regression model is used to find a relationship between a response variable and several explanatory variables [Wood, 2017]. The response is a linear function of the unknown parameters and is expressed as

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{19}$$

$$\text{for } i = 1, 2, \ldots, n$$

where

- Y_i is the value of the response variable for the ith case.
- $\epsilon_i \sim^{iid} N\left(0, \sigma^2\right)$
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables.

- $X_{i,k}$ is the value of the kth explanatory variable for the ith case.

Interactions between explanatory variables are expressed as a product of $X's$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + \dots$$

$$\dots + \beta_{p-3} X_{i,p-2} + \beta_{p-2} X_{i,p-1} + \beta_{p-1} X_{i,p-2} X_{i,p-1} + \epsilon_i. \tag{20}$$

The model is linear due to having linear parameters β. When predicting a dependent variable Y, given known values of the explanatory variables X, Equation 19 is used with estimated coefficients.

T-Test A t-test is used when testing the hypothesis that a parameter β_k differ significantly from 0. It is often used as a decision tool for dropping one predictor from a model, such as:

$$\begin{cases} H_0 & : \beta_k = 0 \\ H_1 & : \beta_k \neq 0 \end{cases}$$

at significance level α. If the standard error of β_i is se, then, under the null hypothesis that $\beta_k = 0$, the test statistics is defined as

$$T = \frac{\hat{\beta}_k}{\text{se}\left(\hat{\beta}_k\right)} \sim t(n-p), \tag{21}$$

where $\hat{\beta}_k$ is the estimated parameter. The observed values of T-statistics are used for calculations and comparisons of critical values, or direct calculations of p-values [Wood, 2017]

2.2.2 Additive Model

An additive model (AM) is an extension of the framework for a standard linear model by allowing non-linear functions for the variables [Wood, 2017, p.131]. The model utilises a multiple linear regression defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i. \tag{22}$$

When extending the model, the linear components $\beta_j x_{ij}$, are replaced with a smooth non-linear function $f_j(x_{ij})$. Rewriting the model, it can be seen that AMs relate a response variable y to predictors, x_j as

$$y_i = \alpha + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots, \quad i = 1, \ldots, n, \tag{23}$$

where α is an intercept parameter, f_j are unknown smooth functions of the covariates x_k and $\epsilon_i \sim^{iid} N\left(0, \sigma^2\right)$. For further explanation of additive models, see Wood [2017]

Given known values of the explanatory variables, a fitted model is used for prediction by applying the estimated coefficients into Equation 23.

Basis Functions A set of basis functions defines a space of functions such that each element in the space is a linear combination of the basis functions. Choosing a basis corresponds to choosing a number of basis functions. The jth basis function is denoted by $b_j(x)$ [Wood, 2017, p.120].

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

where f is the smooth function of the covariate x, b_j is the jth basis function and β_j is the jth unknown parameter.

Smoothing Spline A spline curve is constructed by concatenation of basis functions and can accordingly be described as a curve of two or more joined curves. The location where two curves are joined is called a "knot" and can be at or beyond the limits of the data, called a boundary knot in the latter case. When the boundary knot is unconstrained, a common spline, called the B-spline, is obtained and agreement of the second derivatives imply smooth joints. The smoothing spline assesses a knot to each point in the data making the total number of basis functions one less than the total number of points in the data implicating a quite computationally expensive algorithm. The smoothing spline f is obtained by minimizing the residual sum of squares and adding a penalising term λ to the integral of the squared second derivative

$$\sum_{i=1}^{n} \left(y_i - g\left(x_i\right)\right)^2 + \lambda \int g''(x)^2 dx, \tag{24}$$

where the non-negative penalty parameter λ controls the trade-off between the smoothness and the model of fit [James, 2017, p.277]. The integral is evaluated over the range of x_i. If $\lambda \to \infty$, then the f is a straight line estimate whereas $\lambda = 0$ gives an unpenalised regression spline estimate [Wood, 2017, p.126].

Cubic Smoothing Spline A cubic spline basis spreads its knots evenly through the covariate values, optimizing the computationally expensive smoothing spline. It is especially useful if y_i is measured with noise. It can then be beneficial for x_i, y_i data to be smoothed instead of interpolated. It can be fitting to treat $g(x_i)$ as n free parameters of the cubic spline instead of setting $g(x_i) = y_i$. Equation 24 is then minimized and the resulting $g(x)$ is a smoothing spline minimizing

$$\sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 + \lambda \int f''(x)^2 dx. \tag{25}$$

Thin Plate Spline When there is noisy data that needs an estimation for a smooth function with multiple predictors, a thin plate splines is a general solution to the problem. It can handle any number of predictors and allows for selecting the order of derivatives when measuring the "wiggliness" of the function. The thin plate spline finds the function that best satisfies the conflicting goals of making $\hat{f}$ smooth while still matching the data. A thin plate spline differs from other basis functions because it does not offer the choice of knot positions or selecting basis functions since it naturally emerges from the smoothing problem's mathematical statement. However, thin plate splines are computationally heavy. This is because the cost is equal to the cube of number of parameters and the number of parameters is as large as the number of unique predictor combinations. It is defined as

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(|| \mathbf{x} - \mathbf{x}_i ||) + \sum_{j=1}^{M}, \alpha_j \phi_j(\mathbf{x}), \tag{26}$$

where δ and α are the coefficient vectors that need to be estimated. For further definitions, see "Generalized Additive Models: an Introduction with R. 2ND ed." [Wood, 2017, p.152]. Subject to $\mathbf{T}\boldsymbol{\delta} = \mathbf{0}$ is δ which is a linear constraint, where $T_{ij} = \phi_j(x_i)$. The functions ϕ_i, having $M = \binom{m+d-1}{d}$ functions, span the space of the polynomials in $\Re^d$ of degree, less than m and are linearly independent. The ϕ span the function space where J_{md}, the wiggliness penalty, is zero, meaning that the functions are considered completely smooth in the null space of J_{md} [Wood, 2017, p.152-153].

General Cross Validation Generalized cross validation (GCV) is used to estimate the smoothing parameter λ and for model comparison. In essence, it is an approximation of the ordinary cross validation but with the advantage of being computationally

efficient and less time consuming but still producing good results [Wood, 2017, p.129]. It can therefore reflect an error of how well a model is fitted. It is defined as

$$\mathcal{V}_g = \frac{n \sum_{i=1}^{n} \left(y_i - \hat{f}_i\right)^2}{[n - \mathrm{tr}(\mathbf{A})]^2}, \tag{27}$$

where $\mathbf{A} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}$ is the influence matrix, EDF $= \tau = \mathrm{tr}(\mathbf{A})$ are the effective degrees of freedom and $\mathbf{W}$ are the weights [Wood, 2017, p.129].

F-Test An F-test is used for verifying multiple types of hypotheses and simultaneous significance tests. When using additive models this test can be used effectively when testing hypothesis of smoothing functions being significantly different from constant, as

$$\begin{cases} H_0 & : f = constant \\ H_1 & : f \neq constant \end{cases}$$

In general, the F-test is defined by letting p be number of parameters in the large model Ω and $q < p$ be number of parameters in the small model ω. Then F is defined as

$$F = \frac{\left(SS_{res,\omega} - SS_{res,\Omega}\right) / \left(df_\omega - df_\Omega\right)}{SS_{res,\Omega}/df_\Omega}, \tag{28}$$

$$F = \frac{\left(RSS_\omega - RSS_\Omega\right) / \left(df_\omega - df_\Omega\right)}{RSS_\Omega/df_\Omega}, \tag{29}$$

where SS_{res} is the Residual Sum of Squares defined as $SS_{\mathrm{res}} = \sum_i \left(y_i - f_i\right)^2$ and df denotes the degrees of freedom. Formally, H_0 is then rejected at significance level α if $F > F_\alpha\left(df_\omega - df_\Omega, df_\Omega\right)$

2.2.3 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a popular machine learning gradient boosted tree implementation. The supervised learning algorithm is often used for regression or classification and uses ensemble learning methods based on weak prediction models [Chen, 2016]. For further explanations of XGBoost, see Chen [2016].

Decision Tree Decision trees are commonly used machine learning models constructed by creating rules for decision making according to some criteria. Trees are built by nodes which can be split based on which value and variable would be most

favourable according to the criteria. Such criteria can be, but is not limited to, minimizing a loss or maximizing a gain. Figure 4 visualizes an example of the output from a two-dimensional tree with one root node, three decision nodes and five terminal nodes [James, 2017, p.305].

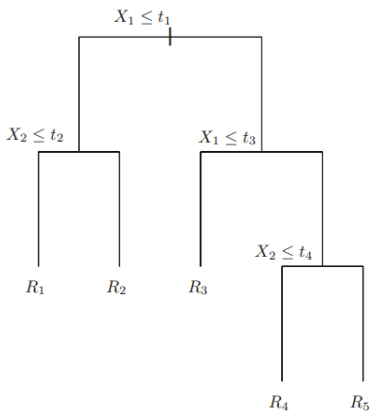


Figure 4: Typical structure of a two-dimensional decision tree.
Source: James [2017]

Decision trees are used to solve problems for either classification or regression [James, 2017, p.304]. The difference between the two lies in the form of the response variable with regression trees having a qualitative response while classification trees have a quantitative response.

A regression tree is created by dividing the predictor space for $X_1, X_2, ..., X_p$ into J distinct and non-overlapping regions, $R_1, R_2, ..., R_J$. For every observation that falls within the region of R_J, the same prediction is made which is the mean of the response values for the training observations in R_J. Let $R_1, ..., R_J$ be the regions that minimize the residual sum of squares (RSS) and

$$\sum_{j=1}^{J} \sum_{i \epsilon R_J} (y_i - \hat{y}_{R_J})^2, \tag{30}$$

where $\hat{y}_{R_J}$ is the mean response for the training set in the jth region. The recursive binary splitting, also known as the top-down greedy approach is used due to it being computationally infeasible to consider all possible partitions for the feature space. The approach initially starts at the top and sequentially splits the predictor space into two.

In every split, the best split at that moment is made, with no regards for optimal future outcome, making the approach greedy. For the recursive binary splitting, the predictor X_j is initially selected along with the thresholds s. The predictor space is split into two regions $X|X_j < s$ and $X|X_j > s$, leading to the greatest feasible reduction in RSS. That is, for any j and s, the pair of half-planes is defined as

$$R_1(j,s) = \{X|X_j < s\} \tag{31}$$

and

$$R_2(j,s) = \{X|X_j \leq s\}. \tag{32}$$

Equation 33 is minimized for the optimal value of j and s

$$\sum_{i:x_i \epsilon R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \epsilon R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{33}$$

where $\hat{y}_{R_1}$ is the mean response of the training observations in $R_1(j,s)$ and $\hat{y}_{R_2}$ is the mean response of the training observations in $R_2(j,s)$. The process is repeated for each resulting region in order to find the optimal threshold and minimize the RSS.

Regularized Learning Objective For a given data set, $\mathcal{D} = (\mathbf{x}_i, y_i)$ $(\mid \mathcal{D} \mid = n, \mathbf{x}_i \epsilon \mathbb{R}^m, y_i \epsilon \mathbb{R})$ with n observations and m features, let a tree ensemble model predict the output using K additive functions as

$$\hat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon \mathcal{F} \tag{34}$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}(q : \mathbb{R}^m \rightarrow T, w \epsilon \mathbb{R}^T)$ denotes the space of regression trees. q represents the individual tree's structure and T is the number of leaves. w_i is defined as the ith leaf's score. For the final prediction, the summation of the score for the corresponding leaves is derived. For the set of functions F, minimization of the regularized objective in Equation 35 is preformed.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k), \tag{35}$$

where

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

and l is a differentiable convex loss function with measures the difference between the target y_i and the prediction $\hat{y}_i$ and Ω is penalization term of the complexity for the model.

Gradient Tree Boosting Due to the parameters in Equation 35 being functions, traditional optimization methods cannot be applied. The model is instead trained in a sequential order by letting $\hat{y}_i^{(t)}$ be the prediction at the tth iteration of the ith instance and choosing the optimal function for the specified loss function. This is achieved by adding f_t and minimizing

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)) + \Omega(f_t). \tag{36}$$

The general setting can swiftly be optimized using a second-order approximation

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y_i}^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t), \tag{37}$$

where the first and second order gradient statistics on the loss function are $g_i = \delta_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}$ and $h_i = \delta^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$. The following simplified objective at step t is obtained by removing the constant terms

$$\bar{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t). \tag{38}$$

Denote $I_j = \{i | q(x_i) = j\}$ as the set of instances of leaf j. Equation 38 follows by expanding Ω as

$$\begin{aligned} \tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{i=1}^{T} [(\sum_{i \epsilon I_j} g_i) w_j + \frac{1}{2} (\sum_{i \epsilon I_j} h_i + \lambda) w_j^2] + \gamma T. \end{aligned} \tag{39}$$

The optimal weight w_j^* of leaf j is be computed for a fixed structure $q(x)$ as

$$w_j^* = -\frac{\sum_{i \epsilon I_j} g_i}{\sum_{i \epsilon I_j} h_i + \lambda}, \tag{40}$$

where the optimal corresponding value is calculated by

$$\tilde{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \epsilon I_j} g_i)^2}{\sum_{i \epsilon I_j} h_i + \lambda} + \gamma T. \tag{41}$$

Equation 41 is used as a measure of the quality, q, of a tree structure as a scoring function. It has similarities with the impurity score used for evaluation of decision trees with the exception of being derived for more general objective functions. Under normal conditions it is usually impossible to enumerate all viable tree structures q. XGBoost therefore uses a greedy algorithm starting from a single leaf while adding branches to the tree iteratively. Let I_L and I_R be assumed to be instance sets of right and left nodes after the split. If $I = I_L \cup I_R$, then the loss reduction after the split is be defined as

$$\mathcal{L}_{split} = \frac{1}{2}[\frac{(\sum_{i\epsilon I_L} g_i)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{(\sum_{i\epsilon I_R} g_i)^2}{\sum_{i\epsilon I_L} h_i + \lambda} - \frac{(\sum_{i\epsilon I} g_i)^2}{\sum_{i\epsilon I} h_i + \lambda}] - \gamma. \tag{42}$$

2.2.4 Support Vector Machine

A support vector machine, SVM, is a machine learning algorithm developed in the 1990s computer community and has been proved to be valuable in many settings. It attempts to perceive a line in a multidimensional space such that the line fits the observations as close as possible, with regards to a margin. The penalty of being outside of the margin is controlled by adjusting a hyperparameter for penalty. SVM handles non-linearity by using certain kernel functions [Awad, 2015, p.68].

Maximal Margin Classifier SVM is originally a generalization of the maximal margin classifier which uses a linear p-dimensional hyperplane for classification of a data set. A hyperplane divides the data set into two subsets. The hyperplane has a margin to the the observations and the most optimal hyperplane has the largest margin. The maximal margin classifier is obtained by solving the optimization problem in Equation 44. It is defined as

$$\begin{aligned} &\underset{b,\mathbf{w}}{min} \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i = 1, ..., N, \end{aligned} \tag{43}$$

where the matrix $\mathbf{w}$ is the parameters and b is the boundary for the points x_i.

Soft Margin Classifier For the case when separation by a hyperplane for a set of observation is not feasible, alternatively when misclassification leads to greater classification of remaining observations, support vector classifiers, also known as soft marginal classifiers, are favourable. A slackness parameter is introduced for each

observation, $\xi_i \geq 0$. With respect to the slack parameter, an extra penalty term is added resulting in the optimization problem

$$\underset{b,\mathbf{w},\xi}{min}\frac{1}{2}\mathbf{w}\mathbf{w}^T + C\sum_{i=1}^{N}\xi_i \tag{44}$$

$$\text{s.t } y_i(\mathbf{w}^T\mathrm{x}_i + b) \geq 1 - \xi \text{ and } \xi_i \geq 0 \ \forall i,$$

where C denotes a non-negative hyperparameter that determines the tolerance of severity and number of violations of the hyperplane. As C increases, the more tolerant the model becomes towards misclassifications and therefore the cost decreases.

Support Vector Machine Classification of non-linear relationships for the predictors and outcome requires an enlarged feature space. This condition is satisfied by using functions for the predictors such as quadratic, cubic or higher-order polynomials. Support vector machines enlarge the feature space by using kernels, an efficient computational approach that quantifies the similarities between observations [James, 2017, p.337]. A function $\varkappa(\cdot,\cdot) : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is a kernel function if, for arbitrary $x_1, ..., x_N \epsilon \mathbb{X}$ and $a_1, ..., .a_N \epsilon \mathbb{R}$,

$$\sum_{i,j=1}^{N} a_i a_j \varkappa(x_i, x_j) \geq 0. \tag{45}$$

A type of kernel is the radial basis function, RBF, defined as

$$\varkappa(x, y) = exp(-\frac{||x-y||^2}{2\sigma^2}), \tag{46}$$

where $||\cdot||$ is the ℓ_2 norm.

Support Vector Regression For a regression problem, a generalization of the classification problem is established where a continuous-valued output is returned by the model. For multidimensional data, x is augmented by one and b is included in the w vector

$$y = f(x) =< w, x > +b = \sum_{j=1}^{M} w_j x_j + b, y, b \epsilon \mathbb{R}^M \tag{47}$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + bx, w \epsilon \mathbb{R}^{M+1}. \tag{48}$$

This function approximation is formulated as an optimization problem in support vector regression (SVR) that is intended to find the narrowest space surrounding the surface while the prediction error is minimized. The objective function is produced through the former condition as

$$min_w \frac{1}{2}||w||^2, \tag{49}$$

where $||w||$ is the approximated magnitude of the normal vector to the surface [Awad, Khanna, 2015, p.68].

2.3 Evaluation Metrics

Evaluation metrics for regression models provide necessary information of their performance and fit. They are useful tools when comparing models and in determining if a model has tendencies of over- or underfit. The three evaluation metrics used in the thesis are given in this section.

2.3.1 Mean Squared Error

The mean squared error (MSE) is a common metric used for model evaluation of performance in terms of how similar its predictions are to the observed values. It is calculated by squaring the residuals, implicating that large errors have a greater impact on the MSE compared to smaller errors. It is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{50}$$

where Y are observed values, $\hat{Y}$ are predicted values and n the number of observations [James, 2017, p.29].

2.3.2 Mean Absolute Error

The mean absolute error (MAE) is also a common metric for evaluation of the performance of a model used to determine the extent to which predictions are similar to observed values. For MSE, calculations entail absolute value of the residual, giving an indication of the actual observed difference. It is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|, \tag{51}$$

where Y are the observed values, $\hat{Y}$ are the predicted values and n the number of observations [James, 2017, p.29].

2.3.3 R-Squared

R-Squared can be interpreted as the proportion of variance in the dependent variable that is predictable from the independent variables. It is useful for determining how well a model is fitted to the observed data where a higher value would indicate a better model since a larger portion of the variance in the observed data then is explained by the model. It is defined as

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{52}$$

where $SS_{\text{tot}} = \sum_i (y_i - \overline{y})^2$ and $SS_{\text{res}} = \sum_i (y_i - f_i)^2$ [Wood, 2017].

2.4 Hyperparameter Tuning

Hyperparameters of a model are the external characteristics that cannot be estimated from the data. They are set prior to the learning process, in contrast to parameters estimated from the data. The values of hyperparameters are of importance for the structure and accuracy of a model. To achieve optimal values of hyperparameters, several methods presented in this section can be applied.

2.4.1 K-Fold Cross Validation

K-fold cross validation is a commonly used method for estimation of model performance on in-sample predictions for machine learning models. The method randomly partitions the data set into k sets of as close to equal size as possible and fit the model on $k-1$ of the sets. The kth set is used for testing. The procedure is performed k times, omitting one set from training each repetition [James, 2017, p.181]. The mean squared error is calculated for each k, resulting in an average error estimation of

$$MSE_{CV} = \frac{1}{k} \sum_{i=1}^{k} MSE. \tag{53}$$

Other metrics like the MAE can also be used in the same way.

2.4.2 Grid Search

A grid search has a large reach and tests many combinations of hyperparameters to find optimal values. When a grid search is conducted, a vector of estimated values for each parameter is created. The grid search tests every feasible combination of values for the model and returns optimal values for the hyperparameters.

2.4.3 Bayesian Search

Bayesian search is more thorough compared to a grid search. In contrast to the extensive trial in a grid search, Bayesian search bases the selection on previous evaluation results and proceeds with the most promising estimate. Based on previous results, a probabilistic model is formed that actively select and map the hyperparameter to the probability of an objective function's score

$$P(score \mid hyperparameters).$$

3 Method

In this section, the method implemented for the thesis is presented. First, the data is presented and elucidated, covering the investigation of the internal data of the migration matrices as well as the external data for the macroeconomic variables. Thereafter, the method of the estimation of Z is presented, giving an indication of the market status at each time point on a quarterly basis. Due to the data consisting of a reasonably short time interval, Section 3.2.1 provides the method of prolonging the series, Z_t, using loss rate and inflation to create a stabilised training basis for the models. Then, a set of regression models used to model the macroeconomic variables to Z are assessed in Section 3.3, initially elucidating on the concept of a "good model" and the main corresponding factors used for determination in this thesis. Next, the degree of point-in-time of the credit ratings is obtained using theory from Carlehed and Petrov in Section 3.4. Following is the calculation method for the predictions of Z when assuming a stressed market scenario provided by EBA. From the predictions, the probabilities of default and overall migration matrices are calculated for 100% TTC as well as 100% PIT based on the predicted $\hat{Z}$ for each of the models. Figure 5 visualises the process using a flow chart.

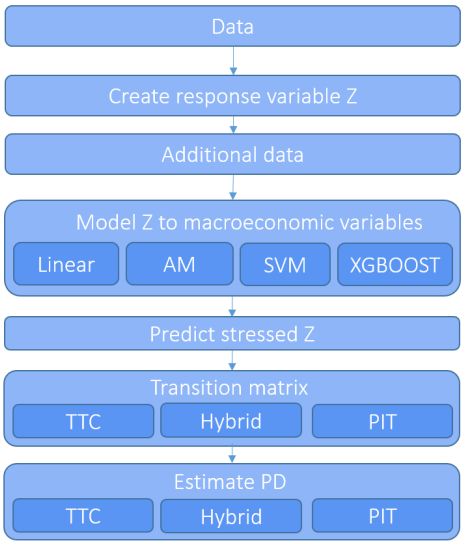


Figure 5: Flowchart of the methodology disposition for the thesis.

3.1 Data

The data in this thesis contains frequencies of migrations for each risk class and is structured as shown in Figure 6.

$$\begin{matrix} \mathrm{N}_{AAA->AAA} & \mathrm{N}_{AAA->AA} & \cdots & \mathrm{N}_{AAA->D} \\ \mathrm{N}_{AA->AAA} & \mathrm{N}_{AA->AA} & \cdots & \mathrm{N}_{AA->D} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{N}_{CCC->AAA} & \mathrm{N}_{CCC->AA} & \cdots & \mathrm{N}_{CCC->D} \end{matrix}$$

Figure 6: Structure of a standard frequency transition matrix.

The available data includes migration matrices from 2005 to 2020 on a quarterly basis. The thesis is focuses on a portfolio containing migrations for an SME portfolio. SME represents 99% of all business in the EU using the definition as having less or equal turnover to €50m or have less or equal to 250 employees [European Commission, 2020]. The specific portfolio is chosen due to its large size along with the relatively even risk class distribution.

The data set includes macroeconomic variables for the time period between 1985 and 2020. The variables are chosen in accordance with the EBA 2021 EU-wide stress test in order to allow testing of the models using the EBA scenarios [European Banking Authority, 2021]. The variables are potential drivers of the macroeconomic state and are therefore suitable for stress testing purposes. The data contains no missing values and includes the following macroeconomic variables:

Table 1: Macroeconomic variables for stress tests provided by EBA.

Macroeconomic variable
GDP
Unemployment rate
Harmonised Index of Consumer Prices
Stock index OMXSPI
Long-term rate
Residential real estate index
Commercial real estate index
5-year SEK Swap rate

The GDP denotes the gross domestic product which is a common measure for representation of added value through production of services and goods in a country, in this case denoted by the change on a quarterly basis. The unemployment rate is the percentage rate of people between the ages of 15-74 years not currently employed or occupied otherwise, but available for work. The data is seasonally adjusted and smoothed. The Harmonised Index of Consumer Pricing (HICP) used in this thesis is KPI which is considered one of the most common measures for compensation and inflation calculations in Sweden. It describes the average price development for the private consumption [SCB, 2021]. The stock market is represented using change of the Swedish index (OMXSPI) an all-share index including all listed shares on OMX Nordic Exchange Stockholm [Nasdaq, 2021]. The long term-rate expresses the long term interest rate of a 10-year government bond. Real estate is depicted using two variables, residential real estate (price trends for one and two-dwelling houses intended for permanent living [SCB, 2020]) and commercial real estate (CRE) (property exclusive provision of workspace and other business related purposes such as office space, hotels, malls, restaurants). The final and eighth variable is the rate of a swap between a fixed interest rate and a floating interest rate such as STIBOR.

The dimensions of the data largely depends on the time span during which it was collected. Even though the use of risk classes is traced back to the beginning of the 20th century, the regulations regarding it have changed immensely, affecting the meaning of given risk classes over time. As previously stated, introduction of IRB

models in Basel II led to further modification of ratings, limiting the time span of coherent accessible data. Due to this, additional data is required to be collected from a similar environment as the original data. This causes a trade-off between the benefits of enlarging the data set using more extensive data based on different circumstances, and using data from comparable circumstances although of smaller size.

3.1.1 Correlation

The macroeconomic variables are examined using Pearson's correlation. A correlation close to 1 or -1 indicates similarities in the provided information of the variables. A threshold of +-0.85 is used for determining removal of a variable, the removed variable in the pair is the one possessing the highest correlation with any other variable. This results in a smaller model of fewer variables with close to equal information. However, it is conducted with the purpose of addressing potential problems with multicollinearity rather than as a means for dimension reduction. Multicollinearity, defined as one predictor being able to linearly predict another, can result in difficulty in the reliability of the estimates of the model parameters [Alin, 2010].

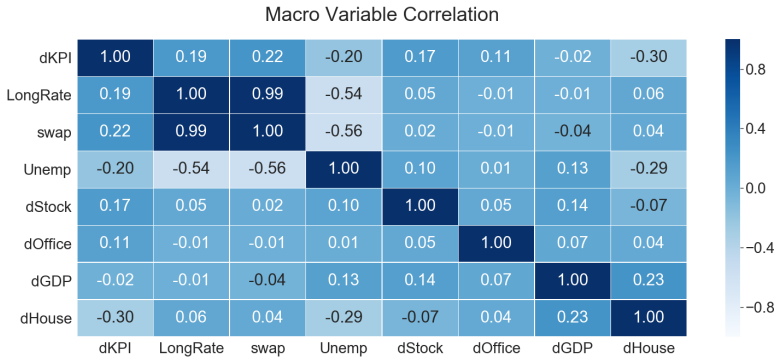


Figure 7: Correlation matrix of macroeconomic variables.

As can be seen in Figure 7, the correlation between the 5Y swap rate and the long term-rate is close to 1. Swap rate possess the second highest correlation with any other variable and is therefore removed from the data set. Removal of swap rates from the data set hopefully resolves any problems regarding multicollinearity.

3.2 Estimation of One-Parameter Z

To find the relationship between a response variable and the macroeconomic variables, a time series describing the economic state Z is needed. As initially described in the theory section, two main articles are used in this regard. The method from the paper "The one parameter representation of risk" is based on theory of Merton and the one factor model described in Equation 1 [Belkin, Suchower, 1998].

The initial data consists of transition matrices on quarterly basis from 2005 to 2020. There is a strong time dependency in the frequency matrices because the number of counterparties increase with time, causing complications when modelling if not handled. The frequencies of data are therefore transformed to probabilities by dividing each row with the sum of the same row, removing the time dependency. The transformed matrices describe the probability of migrating to a given risk class. Its structure is visualised in Figure 1.

The series Z is obtained by applying the theory from the article by Carlehed and Petrov given in Section 2.1. The sole focus of the method lies on the rightmost column denoting migrations to the default state. For each time step t, the summation of the rightmost column is retrieved, obtaining the overall probability of default for the period, not considering from nor to what risk class the migration describes. The obtained vector, denoted d, is the probability of default for each time step. A transformation of d is conducted by applying the standard normal cumulative distribution function, $\Phi^{-1}(d)$. From the transformed vector, a mean and standard deviation are derived which are inserted into Equation 13, resulting in a series Z.

Figure 8 shows a fictional time series for Z values based on quarterly bankruptcy statistics in Sweden for limited companies [Carlehed, Petrov, 2012]. The black line represents values for Z corresponding to the (left y-axis) and the grey line represents the percentage rate for the annualised bankruptcy frequency corresponding to the (right y-axis). Z can be interpreted as the state of the market and thereby, simplified, as the inverse of the probabilities of defaults for the same period of time.

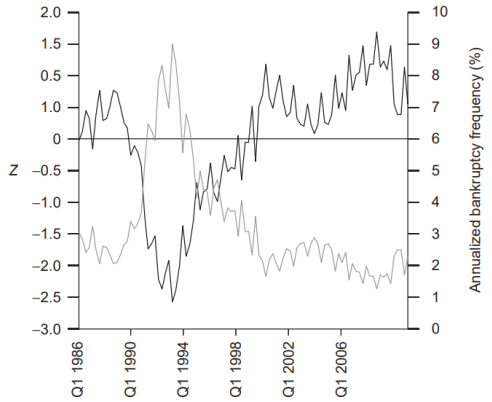


Figure 8: Diagram depicting typical values for Z (black curve) and annualised bankruptcy frequency (grey curve) based on Swedish bankruptcy statistics for limited companies on a quarterly basis from 1986 to 2010.

Source: Carlehed, Petrov [2012]

3.2.1 Additional Data

The amount of data a model is trained on directly correlates to the magnitude of the corresponding error of the predictions, up to some point. The data from the original source is limited and therefore alternative approaches are studied for enlarging the data set such as by simulation or by use of other macroeconomic indicators to emulate Z. There are predominantly two main complementary techniques used for comparing time series. The first one utilises correlation where a high correlation between the investigated macroeconomic index and the original data set indicates similarities and support the use of the emulating index. The second option is visual investigation using a plot of the movements.

The investigated indicator is internal loss rate, defined as premiums not earned due to counterparties not meeting their obligations divided by total earned. It gives an indication of the general probability of default further back by providing more data. In general, a higher loss rate indicates a higher PD and vice versa. The data needs to be scaled since the loan to value ratio differs to great extent from 1980 to the beginning of the 1990s. To account for the varying loan to value ratio, the total amount of internal lending with regards to inflation is used to scale the loss rate. The inflation and lending adjusted loss rate behaves similarly to the PD from the original

data. In order to facilitate the comparison with the Z series, the adjusted loss rate is standardised using Equation 13. By doing this, the correlation between the original time series Z and the Z based on the adjusted loss rate for the years 2005-2020 is equal to 0.68. By investigation of their movement, noticeable similarities in pattern is observed, although the fluctuations of the loss rate are somewhat amplified. A new response is created with a combination of inflation and lending adjusted loss rate and the original Z data. This new time series will henceforth in the thesis be referred to as Z.

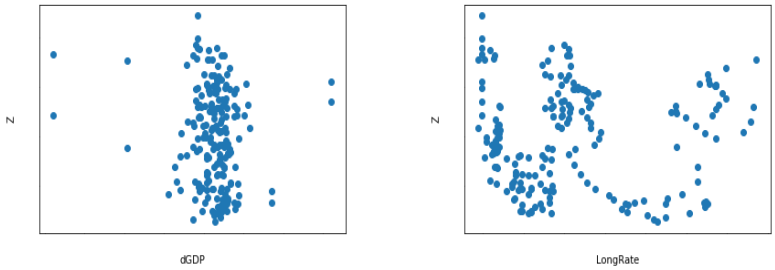
The purpose of including the new data is to create a more stable model and thus decrease fluctuations of the size of the residuals depending on which subset is excluded in a cross validation. The forgoing data set used when creating the models is therefore a combination of the original data Z originating from the transition matrices from 2005 to 2020, with the addition of the inflation and lending adjusted loss rate from 1985 to 2005.

3.3 Regression Models

In this section, the method for constructing regression models using a variety of approaches is presented. Recall that the aim is to find the relation between the systematic variable Z and the corresponding macroeconomic variables. The predictions of the model should be independent of time, whereby what happened before and after the stressed economic scenario should not matter to the model or the prediction.

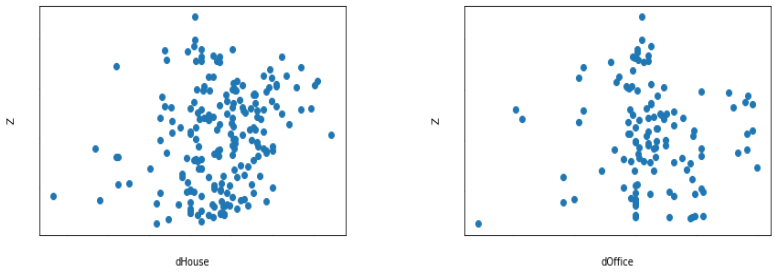
3.3.1 Visualisation of Data and Macroeconomic Variables

The scatter plots of the response Z to each of the macroeconomic variables presented in Figure 9, 10, 11 and 12 serve as support when evaluating the models in terms of macroeconomic effect on the response.



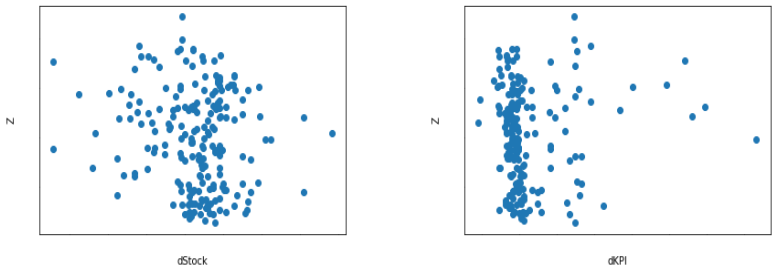
(a) (b)

Figure 9: Scatter plot of the response Z against the macro variables. (a) Change in GDP against Z. b) Long term-rate against Z.



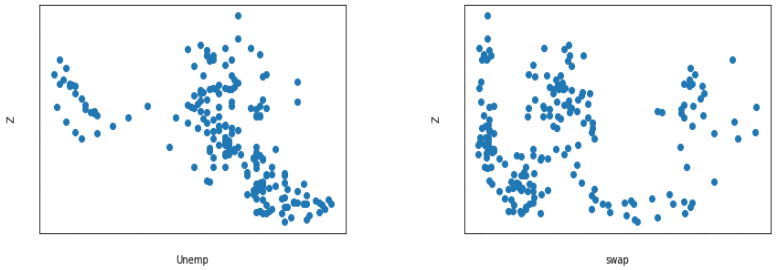
(a) (b)

Figure 10: Scatter plot of the response Z against macroeconomic variables. (a) Change in house prices against Z. (b) Change in commercial real estate against Z .



(a) (b)

Figure 11: Scatter plot of the response Z against macroeconomic variables. (a) Change in the stock index OMXSPI against Z. (b) Change in HICP, KPI against Z.



(a) (b)

Figure 12: Scatter plot of the response Z against macroeconomic variables. (a) Unemployment rate against Z (b) Swap rate against Z.

3.3.2 Evaluation of Models

There are several indicators to use when comparing models in order to declare the optimal one. MSE and MAE, as described in the theory section, attempt to give an error measurement based on the residuals of the fitted model, with squared or the absolute residuals, respectively. Concerning these measurements, a smaller value is preferred when deciding on the best model since it implicates closer average predictions to the observed values. R-squared, defined in Equation 52, is also a measurement of fit. It can be interpreted as the proportion of the variance of a dependent variable explained by the independent variables. It takes the sum of squares of residuals with respect to the sum of squares of the difference between each observation and the mean of the response into account. Hence, a greater value indicates a fitted model that more closely explains the same variance as the original data and therefore is more likely to be an appropriate model.

When interpreting the coefficients it is important to know how the independent variables, in this case the macroeconomic variables, affect the dependent variables, in this case the Z. When evaluating these, the effect on the dependent variable is studied and then compared to the general perception of how it is described in economic literature, if such perception exist. Considering this, a regression model is sought after since the dependent variables effect can easily be interpreted while for tree based models, it is not as clear. Therefore, when evaluating the model, the interpretability of the effect of dependent variables for each model has to be taken into account. Figure 13 shows each of the included models for this thesis and their interpretability in relation to each

other. Also included in Figure 13 is the flexibility of each model which corresponds to which type of functions the model can emulate. Least Squares in Figure 13, which in this thesis is referred to as the linear model, can only take linear relations into account while each of the other models can account for much more flexible ones [James, 2017, p.25].

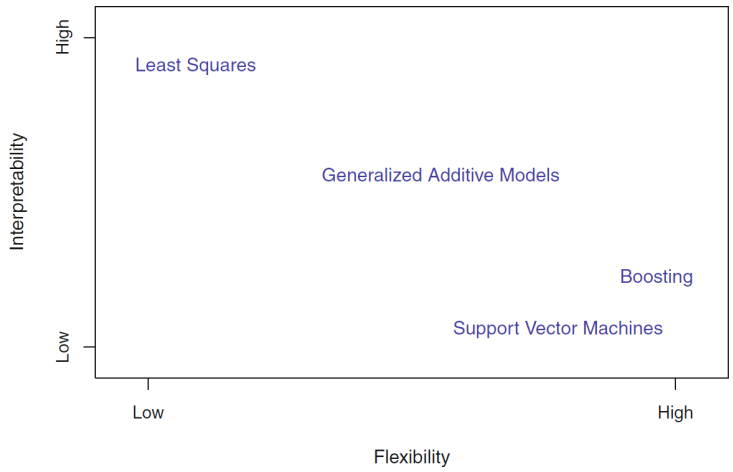


Figure 13: Plot of interpretability against flexibility for regression models. The figure has been adjusted to fit the chosen methods.

Source: James [2017]

Cross validation is commonly used for tuning the hyper-parameters of a model. However, since the data set for this thesis is small, cross-validation is used for determination for the stability of the predictions depending on the partitioning of training and test sets. If the variance is high then the the model is less stable which has to be taken into account when using the model. Cross-validation is therefore used for evaluating the models where the average test and training errors give an indication of the stability of the model.

The bias-variance trade-off has to be considered for model evaluation [James, 2017, p.34]. Variance corresponds to the amount by which the predictions would change given that the model would be trained using two different data sets. When minimising, the wanted results are small changes for different data sets. On the contrary, bias corresponds to the error which occurs when trying to estimate a complex model

using a too simple model [James, 2017, p.35].

3.3.3 Linear Baseline Model

The first model is a linear regression model using ordinary least square to estimate the unknown parameters. It is used as a baseline model due to its simplicity and strong interpretability in regards to the effect from the predictors on the response. In the event that the errors and values for metrics described in the previous section for this model are smaller or equal to the errors of different models the linear baseline model would be preferred.

Initial scatter plots of each of the macroeconomic variables to the response Z visualised in Section 3.3.1, show no indication of a linear relationship. However, since a linear regression model is easy to interpret in terms of coefficients, the model is still investigated as a baseline. To further investigate interactions among the macroeconomic variables, scatter plots for each combination of the variables are created. Some interactions are observed to have patterns indicating a non-random relationship.

The model is fitted with Z as response and all of the macro variables as predictors, represented by Y and $\boldsymbol{X}$ in Equation 19. The model is refitted using backwards stepwise selection of the full model with both the main effects and the interactions. For every model, a t-test according to Equation 21 is made for each of the predictors and the term with the highest (most non-significant) p-value with a confidence of 95% is removed. This is applied to the full model where the interaction effects are removed first until all remaining interactions are significant. The same procedure is applied for the main effects. However, if a linear predictors is the least significant according to t-test of predictors Equation 21, but its corresponding macro variable is a part of a significant interactions, the predictor is not removed and the second most non-significant predictor is removed instead, if any exist. This procedure results in the model presented in Section 4.1.1.

To be able to sufficiently rely on the created model the following assumptions are investigated:

- Independent residuals
- Constant variance among the residuals

- Normally distributed residuals.

To be able to determine if the approximated errors (residuals) are independent, it is important to see if there are any structural problems within the model. Figure 14 shows the residuals from the fitted model plotted against an index which in this case represents time. The pattern tends to not be fully random since the variance with respect to time is not constant. It indicates that a non-linear relationship has not been accounted for with the current model.

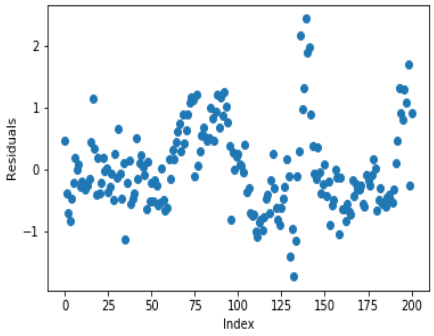


Figure 14: Scatter plot of index against residuals for the full linear model.

When evaluating the variance of a scatter plot with the fitted values and residuals, a constant variance among the observations indicates a solid model in terms of standard deviation among forecasted errors, called heteroscedasticity. Considering Figure 15, a trend of smaller variance in the lower and higher fitted values than around 0 is observed. Hence, the residuals seem to be somewhat more extreme in the middle of the fitted values. Since there is such a small set of data points to evaluate, a small set of outliers can affect the constant variance assumption with a large magnitude. However, this pattern is, most likely, too prominent to be a result of outliers and hence there is problem with heteroscedasticity.

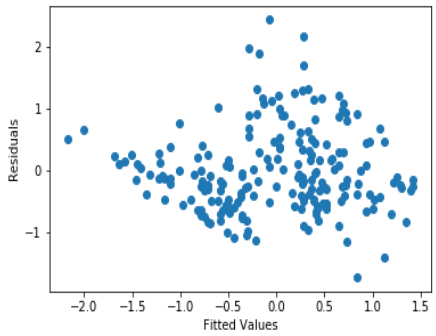


Figure 15: Scatter plot of fitted values and residuals for the full linear model.

To ensure that the model predicts as expected, the normality assumption of the residuals is important. For this thesis, the assumption is checked in two ways. The first approach is reviewing the normal quantile-quantile plot in Figure 16 by making a scatter plot of the theoretical quantiles against the sample quantiles. If the observed values follow the 45 degree red line closely, they are normally distributed. Figure 16 shows a pattern indicating that the sample quantiles of the residuals follow the theoretical quantiles closely, with the exception of a few outliers.

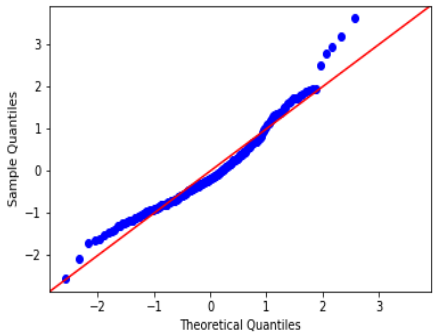
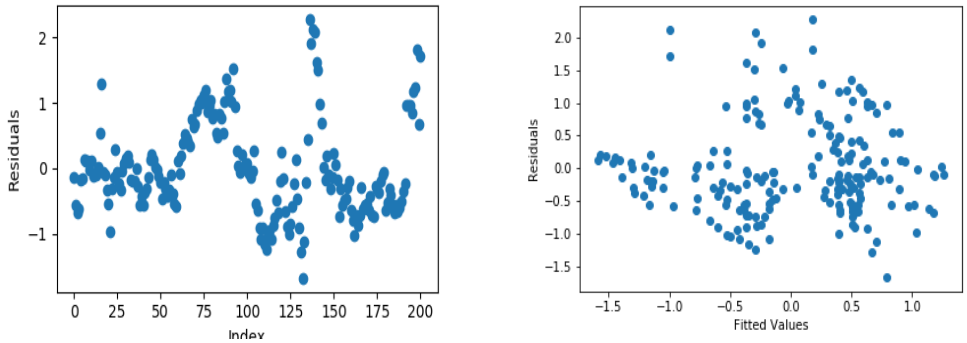


Figure 16: Quantile-Quantile plot of the residuals for the full linear model.

Since it cannot not be concluded that the residuals are independent, nor that the residuals have constant variance, it is unclear how reliable the predictions from the model are.

When assessing the stability of the model, cross validation is used. Cross validation is normally used for parameter tuning for machine learning algorithms. However, with regards to the small data set used for this thesis, cross validation is used for evaluation of the stability of the model depending on the partition for the training set. Because of the small size, leaving out a few observations or not in the training process cause significant change in the parameter estimation, the goal is to see how much. The cross validation uses is a K-fold with 10 folds described in the Section 2.4.1 which results in 10% of the observations observations being left out each time for validation while the parameter estimation is performed on the rest. This procedure results in the mean squared error, mean absolute error, and r-squared presented in Table 6.

When investigating scatter plots of the response Z to each of the macroeconomic variables, some of the combinations show patterns generally considered more random than others. The macroeconomic variables that show clear patterns of some non-random relationships are unemployment, long term-rate and house price index. Therefore, given that the previously presented model had problems with its assumptions, the noise that some of the other macroeconomic variables may have added could be avoided. A new model using only unemployment, long term-rate and house price index is therefore fitted and presented in Section 4.1.1. Each of the previously presented assumptions are also checked for this model. Considering the scatter plots in Figure 17, they show similar tendencies as the previous model. The pattern is the same but the residuals do not have the same spread as in Figure 15. Since the residuals do not seem to be independent, the corresponding assumption of independence does not hold. Regarding the constant variance assumption, there is a clear heteroscedastic pattern in the fitted values to the residuals plot, as in Figure 15. Therefore, by the same reasoning as for the previous model, the constant variance assumption does not hold.



(a) (b)

Figure 17: Scatter plot of index and residuals & of fitted values and residuals for the small linear model.

The quantile-quantile plot in Figure 18 is almost identical to the one for the previous model shown in Figure 16 and hence the same reasoning is made for this model. Thus, the residuals are assumed to be normally distributed.

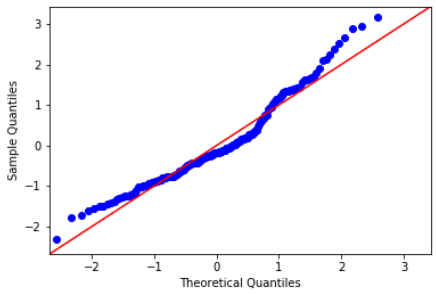


Figure 18: Quantile-Quantile plot of the residuals for the small linear model.

A 10-fold cross validation is also used for this model for investigation of the stability. For the cross-validation, the MSE, MAE and R-squared are calculated and presented in Table 6.

3.3.4 Additive Model

By using the arguments of non-linearity in the previous section from the linear model, an additive model is a natural next step. The main objective for this model is to incorporate the assumed non-linear effect that some of the macroeconomic variables

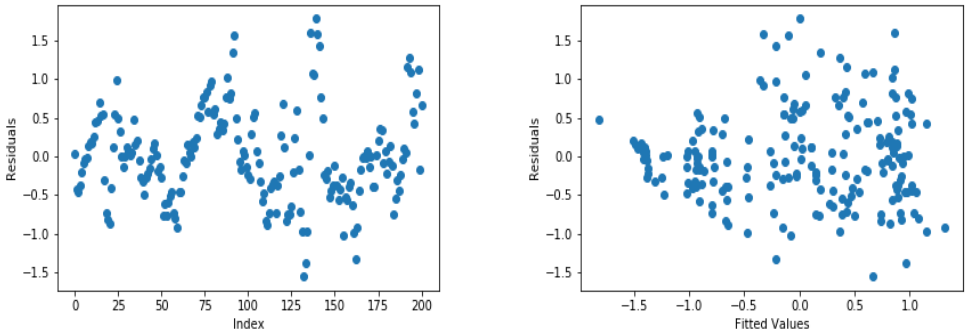
have on the response Z that the linear model could not account for. From the scatter plots of Z to the macroeconomic variables, tendencies of non-linear effect for both unemployment rate and long term-rate are pronounced since neither of them seem to increase or decrease in a linear way. The interaction between the long term-rate and the unemployment plotted against the response Z shows a pattern containing non-linear indications. Smoothing splines for additive model are used in order to account for non-linear complex relationships between the response and the predictor. An AM is desirable, given that there are non-linear relationship in the response to the macro, since an AM can model complex relationship while maintaining the ability to be fairly interpretable in terms of effect of predictors.

The aim is to fit an additive model using the methods presented for the linear model where the f is represented in such a way that the corresponding response of the additive model depends linearly on some unknown smooth functions of some predictors. To do this, a set of basis functions must be chosen. In order to find the best model, several models are created by fitting one model each, including only the main effects of unemployment and long term-rate as well as a model containing both the main effect and one with the interaction. The house prices index is also tested for all the models but has non-significant effect and is therefore not further included. The models are then compared in terms of GCV presented in Section 2.2.2 and the one with the smallest values, corresponding to the best fit, is chosen. This results in the model shown in Section 4.1.2. The additive model is then adjusted by simultaneously tuning each of the smoothing splines dimensions of their bases and order of penalty corresponding to unemployment and long term-rate. It is done in order to fit the observations as close and smooth as possible. The basis for the chosen smoothing splines is a thin plate regression spline for unemployment, as presented in Equation 26 and a cubic regression spline for long term-rate, as presented in Equation 25. Both with the penalised sum of squares presented in Equation 24. Hence, the smooth functions are defined according to $f_1(x) = \sum_{i=1}^{k_1} b_{1i}(x)\beta_{1i}$ and $f_2(z) = \sum_{i=1}^{k_2} b_{2i}(z)\beta_{2i}$ where β_{1i} and β_{2i} are unknown coefficients, $b_{1i}(x)$ and $b_{2i}(z)$ are sets of known basis functions as previously defined.

Each model is evaluated using the GCV-score presented in Section 2.2.2 where a smaller value indicates a better model since it corresponds to a smaller estimation of the cross validation error. Each of the smooth terms are jointly tested with statistical significance according to the F-test in Equation 29 and the null-hypothesis of equality

of zero with a confidence of 95%.

For the model to be reliable, the residuals need to be, as for the linear model, independent and have constant variance. From Figure 19, the same pattern as for the linear model is observed. There are some trends of non-independent residuals in Figure 19 plot (a). Considering plot (b) in the same Figure, low variance is found in the lower values of the fitted values and higher variance in the higher parts. Hence, the residuals are neither independent, nor do they have constant variance.



(a) (b)

Figure 19: (a) Scatter plot of index and residuals. & (b) Scatter plot of fitted values and residuals for the additive model.

For this additive model, the assumptions for normality also applies, just as for the linear model, since the mean of the response as well as the residuals were assumed to be normally distributed.

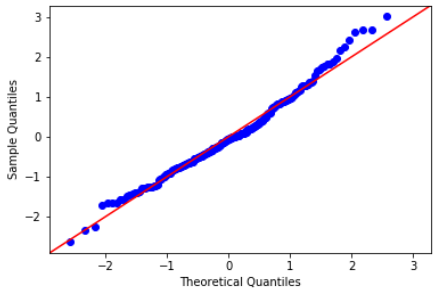


Figure 20: Quantile-Quantile plot of the residuals for the additive model.

The final assumption is the identifiability constraint. Since f_1 and f_2 are used in an additive model, they can be estimated within an additive constant. Imposed on the smooth terms presented is the zero-sum identifiability constraint where $f_1 = \sum_i f_1(x_i) = 0$ and $f_2 = \sum_i f_2(x_i) = 0$. This constraint is checked by reviewing the smoothed functions presented in Figure 21 on page 49 where the average of the smoothing functions is set around zero. From the figure, it appears to be the case and thus, this constraint seems to hold. The stability of the model is then evaluated using a 10-fold cross validation and its resulting MAE, MSE and r-squared are presented in Table 6. The final model is presented in Section 4.1.2.

3.3.5 Extreme Gradient Boosting Model

Extreme gradient boosting, XGBoost is a widely used tree-based machine learning algorithm utilising the gradient boosting algorithm and regularisation. In contrast to the linear model, this approach can incorporate non-linear relationships into the model

XGBoost has a large number of hyperparameters however eight are chosen for this model which can be seen in Table 2.

Table 2: Hyperparameters for XGBoost.

Hyperparameter
Number of estimators
Max depth
Learning rate
Min child weight
Max delta step
Colsample by node
Alpha

The hyperparameters have to be set before the training of the model begins and the set of hyperparameters focuses on the build of the trees as well as regularisation and sampling. First the number of estimators is included, denoting the number of trees and is equivalent to the number of boosting rounds. The second hyperparameter is the *max depth* that describes the maximum number of nodes that are allowed from the root to a leaf of a tree for a base learner. Deeper trees allow for more complex

relationships to be modelled although the splits become less relevant with the increasing depth and too deep trees easily lead overfitting. *The learning rate*, or the shrinkage factor as it is also called, is a technique to slow down the learning by applying a weighing factor for the corrections of residual errors by newly added trees of the model. A high *learning rate* increases the time to fit but will most likely cause the model to overfit the training set, especially for a small data set, while a smaller value of the *learning rate* will make the computations slower but often results in a better solution. *Min child weight* is the minimum required weight when a new node is created in a tree. A smaller *min child weight* allows for more complex trees due to creating nodes corresponding to fewer samples however it is also more likely to overfit. A larger min child weight therefore entails a more conservative algorithm. *Max delta step* describes the largest step the estimation of a tree's weight is allowed to be. At default it is set to zero, meaning that no constraint is active. A positive value makes the creation of the tree more conservative. *Colsample by node* denotes the fraction of columns that are randomly sampled from for each node. A smaller subsample ratio for the *colsample by node* limits the amount of variables that are used at each split which can decrease overfitting. Alpha denotes the L1 regularisation term on the weights. When *alpha* is increased the model will be more conservative.

Taking wisdom from the linear models and the additive model, only a subset of the whole data set is used for this model. The variables included are those showing signs of a non-random relationship and hence provides information to the model. Therefore the model is based on unemployment rate, long term-rate and residential house price index.

To find the best fitting model and its corresponding hyperparameters, a two-step procedure is used for implementation. The first one is a grid search as described in theory Section 2.4.2. For this procedure, a set of values for each of the hyperparameters is set with regards to the data, covering a large range of values. All combinations of the hyperparameters are tested and the combination resulting in the smallest feasible error is selected. Thereafter, a Bayesian search, as described in Section 2.4.3, is conducted around the particular area for the value of each hyperparameter obtained in the grid search. This procedure increases the accuracy of the hyperparameters. The metric used for evaluation of the grid search and Bayesian search is mean squared error using a 10-fold cross validation as described in theory Section 2.4.1.

3.3.6 Support Vector Machine Model

Support vector machine (SVM) is a machine learning algorithm for regression and classification. The creation of this model is initiated by selecting values for the hyperparameters. As previously stated, hyperparameters are estimated prior to training the model and are meant to establish the best possible conditions for the model, in the case of SVM, only two hyperparameters can be estimated. The first one is the cost parameter, C which is used to adjust the penalty for an observation inside the margin. Smaller values indicate a smaller penalty and higher values indicate a higher penalty. Therefore, the hyperparameter C has a large effect on the resulting model and thus need to be properly tuned. The second hyperparameter, γ, states how much curvature the decision boundary is allowed to have. A high value for γ entails more curvature and vice versa. These two variables need to be estimated to minimise the error while at the same time minimising both variance and bias.

As in the XGBoost model, the SVM model is based on the subset containing only the unemployment rate, long term-rate and house price index. The hyperparameters are estimated using the same method as previous models, initiated with an extensive grid search of comprehensive intervals to find a reasonable range for each of the hyperparameters followed by a Bayesian search based on surrounding intervals to obtain more exact values. The final hyperparameters are presented in Table 5.

3.4 Prediction of Transition Matrices

This section present the process of predicting transition matrices and corresponding PD's.

3.4.1 Estimation of PIT-ness

As presented in the theory section according to Carlehed and Petrov, the α in Equation 16 is defined as the degree of PIT. To convert the transition matrices from a hybrid perspective to either PIT or TTC, the degree of PIT of the hybrid ratings is required. To obtain α for the credit ratings of the data from this thesis, Equation 16 is solved by using the originally estimated Z and correlation ρ according to Equation 12. The resulting α is 0.984.

3.4.2 Prediction of Stressed Scenario

The EBA adverse scenario for 2021 is used for evaluation of the effects from an adverse macroeconomic shock on the portfolio. The adverse macroeconomic scenario is intended to resemble a severe (but realistic) scenario. It includes the macroeconomic variables shown in Table 3.

Table 3: Macroeconomic variables for EBA adverse 2021 scenario(%).

dGDP	**Unemp**	**dKPI**	**dHouse**	**dOffice**	**longrate**	**dStock**	**Swaps**
-2.0	13.3	0.7	-5.3	-22.5	-0.77	-50	-0.17

No monetary or fiscal policy reactions are assumed in the scenario beyond what is already in place. When quantifying and analysing the effects of the scenario the models presented in Section 3.3 are used for estimation of the response $\hat{Z}$. The predictions are made with each model using the respective section of prediction in the theory section. The predicted values of the response Z are presented in Table 7.

Prediction of Hybrid Transition Matrix and PD When predicting transition matrices for each perspective (PIT, TTC, hybrid) given a $\hat{Z}$, the average transition matrix for the interval is required. For each predicted $\hat{Z}$ a quarterly specified matrix reflecting the market fluctuations is obtained by shifting the basis transition matrix with respect to $\hat{Z}_t$. This is done by taking the inverse cumulative distribution function of the cumulative sum for every row, obtaining the average binned transition matrix. Every bin represent a cut off point in the distribution function of the transition for each risk class. The shift is applied by first subtracting the given value for $\hat{Z}$ from the value of the binned matrix in the rightmost column followed by taking the cumulative distribution function over the expression and transforming it back to a probability. For the remaining values in the matrix, the market adjusted probability is obtained by taking the cumulative sum over the difference between the given binned value subtracted by $\hat{Z}$ and the cumulative sum of the columns for the risk classes considered lower than the current. The result is an adjusted matrix reflecting the fluctuations in the market at time t, through a shift $\hat{Z}_t$ for a hybrid model. The probability of default is assumed to be the mean of the rightmost column, representing the overall transition from the different risk classes to the state of default. The corresponding probability of default is presented in Table 8.

Using the method of Carlehed and Petrov, the PD is calculated using Equation 13 and is presented in Table 9.

Prediction of 100% TTC Transition Matrix When determining the transition matrix and probability of default for a 100% TTC model, Equation 17 is used. The Equation takes the transition probabilities from the hybrid model, p_i into account and for each position in the matrix, representing every migration from all combinations between the risk classes, calculates a new probability. The new matrix of values represents the long term, through-the-cycle transition probability to migrate to a given risk class. The probability of default for the 100% TTC is then determined as the overall sum of the last column of the matrix representing the transition from every risk class to the state of default. Using the method of Carlehed and Petrov, the PD is calculated using Equation 17 and is presented in Table 9.

Prediction of 100% PIT Transition Matrix From the transition matrix with a TTC-perspective, the 100% PIT transition matrix is obtained by applying Equation 18. The TTC-probabilities for every transition are inserted into the Equation and a matrix with a PIT-perspective is obtained. From this, the PIT PD is derived as the sum of the last column representing the total migration from every risk class to the state of default. Using the method of Carlehed and Petrov, the PD is directly calculated using Equation 18 and is presented in Table 9.

4 Results

This section presents the regression models and their corresponding predicted $\hat{Z}$ and probability of default, given the stressed EBA 2021 scenario.

4.1 Model Result

In this section, the resulting models and their corresponding values of the evaluation metrics are presented.

4.1.1 Linear Models

Full Model Recall from the method section that the first linear model included all macroeconomic variables and combinations of interactions. The full linear model, achieved by step-wise backward elimination and therefore only includes significant coefficients, is expressed as

$$\hat{Z} = 1.59 - 0.59X_{dGDP} - 1.394X_{Unemp} - 1.327X_{longterm-rate} + 0.37X_{dHouse} + 0.712X_{dOffice} + 1.284X_{dGDP}X_{Unemp} + 0.573X_{dGDP}X_{longterm-rate} - 0.0933X_{longterm-rate}X_{dOffice} - 0.182X_{Unemp}X_{dOffice} + 0.932X_{Unemp}X_{longterm-rate} + 0.533X_{dGDP}X_{dHouse} \quad (54)$$

The β values are randomized for confidentiality purposes.

Small Model The small linear model using step-wise backward elimination with only unemployment, long term-rate and house price index is presented in Equation 55. All included terms are significantly different from zero.

$$\hat{Z} = 0.043 - 1.25X_{Unemp} - 1.3372X_{longterm-rate} - 0.93X_{dHouse} + 0.96X_{Unemp}X_{longterm-rate} \quad (55)$$

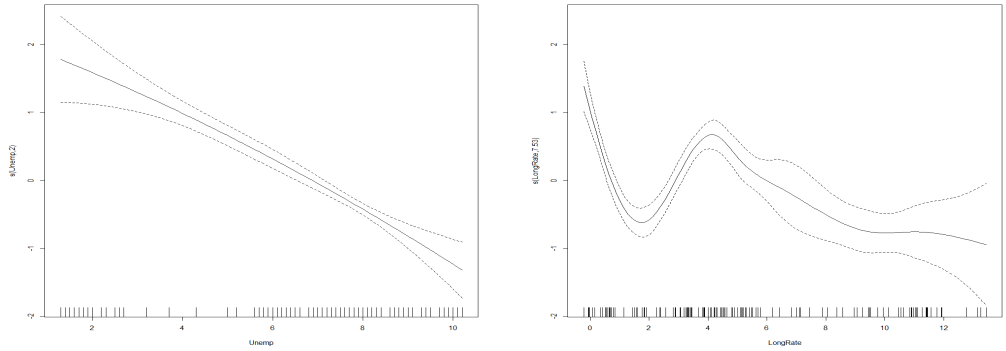
The β values are randomized for confidentiality purposes.

4.1.2 Additive Model

The additive model using only unemployment rate and long term-rate results is expressed in Equation 56. Recall from the method section that the included terms are chosen according to the lowest GCV-score.

$$\hat{Z} = -0.01 + f_1(Unemp_i) + f_2(longrate_i) \quad (56)$$

where f_1 is a smooth function with thin plate regression spline and f_2 is a cubic regression spline.



(a) Unemployment (b) long term-rate

Figure 21: (a) Plot of the fitted main effect of unemployment to the Z for the additive model & (b) Plot of the fitted main effect of long term-rate to the Z for the additive model.

4.1.3 Extreme Gradient Boosting Model

The tuned hyperparameters using grid- and Bayesian search for the extreme gradient boosting model are presented in Table 4. A representation of the final tree is presented in Figure 22.

Table 4: Values of the hyperparameters for XGBoost.

Hyperparameter	**Value**
Number of estimators	11
Max depth	3
Learning rate	0.86
Min child weight	1.35
Max delta step	0.45
Colsample by node	0.64
Alpha α	0

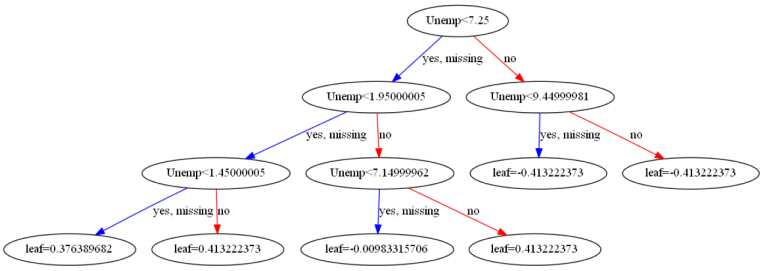


Figure 22: Resulting decision tree for XGBoost.

4.1.4 Support Vector Machine Model

The tuned hyperparameters using grid- and Bayesian search for the support vector machine algorithm are presented in Table 5, using the non-linear radial basis function as kernel.

Table 5: Hyperparameters for SVM.

Hyperparameter	**Value**
C	0.648
γ	0.000563
Kernel function	Radial basis function

4.1.5 Measure of Fit Metrics

Table 6 show the errors from the 10-fold cross validation measured in mean squared error and mean absolute error as well as the R-squared coefficient of determination. The metrics in Table 6 are calculated for each model for estimation of the response Z using macroeconomic variables.

Table 6: Measure of fit metrics for all models.

Metric	**OLS Full**	**OLS Small**	**AM**	**XGBoost**	**SVM**
CV MSE Train	0.453	0.518	0.320	0.190	0.416
CV MSE Test	0.442	0.484	0.324	0.657	0.743
CV MAE Train	0.525	0.558	0.438	0.333	0.55
CV MAE Test	0.510	0.536	0.460	0.660	0.712
R^2	0.596	0.471	0.662	0.804	0.580

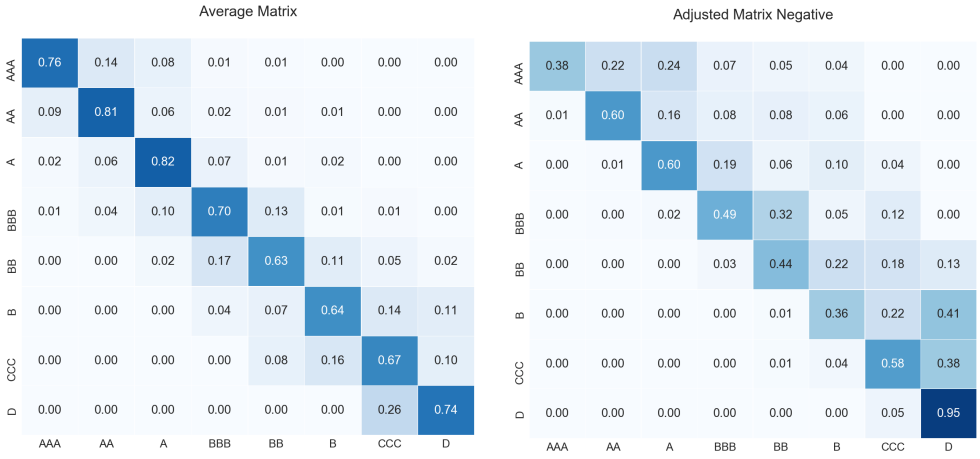
4.2 Stress Test Results

The predicted $\hat{Z}$ for the stressed adverse EBA 2021 scenario is presented for each model in Table 7.

Table 7: Predicted values for adverse scenario $\hat{Z}$ for different models. The rightmost column shows the $\hat{Z}$ using an ensemble of the AM, SVM and XGBoost models.

Metric	**OLS Full**	**OLS Small**	**AM**	**XGBoost**	**SVM**	**Mean**
$\hat{Z}$	-3.62	-4.43	-0.422	-0.884	-0.521	-0.608

Figure 23 visualizes two fictitious heat maps, the left one describing the average transition matrix for the portfolio and the right describing the adjusted transition matrix given a unfavourable year in terms of macroeconomic variables, with a Z = -1. The figure exemplifies adjusting of transition matrices. The PD calculated from the transition matrices are presented in 8



(a) (b)

Figure 23: Heat maps depicting the probabilities of migration for each risk class. (a) Average transition matrix for $Z = 0$. (b) Adjusted transition matrix given $Z = -1$. Values are fictional for confidentiality purposes.

Table 8 shows the probability of default from a TTC, hybrid and PIT perspective using the predicted $\hat{Z}$ in Table 7 for the stressed EBA 2021 scenario. The PDs are calculated using the rightmost column, the state of default, of the predicted transition matrices shown in Figure 23.

Table 8: The resulting probability of default from each of the corresponding $\hat{Z}$ using the predicted transition matrices. The rightmost column shows the $\hat{Z}$ using an ensemble of the AM, SVM and XGBoost models.

Metric	**OLS Full**	**OLS Small**	**AM**	**XGBoost**	**SVM**	**Mean**
PD TTC (%)	38.4	53.0	1.24	2.71	1.48	1.72
PD Hybrid (%)	48.3	62.7	1.34	3.24	1.64	1.93
PD PIT (%)	57.3	69.8	1.46	3.87	1.83	2.22

Table 9 presents the probability of default for a TTC, hybrid and PIT perspective using predicted $\hat{Z}$ in Table 7. Recall that values are calculated using Equation 13 by Carlehed and Petrov.

Table 9: Resulting probability of default for corresponding $\hat{Z}$ using the method of Carlehed and Petrov.

Metric	**OLS Full**	**OLS Small**	**AM**	**XGBoost**	**SVM**	**Mean**
PD TTC (%)	1.90	1.90	1.90	1.90	1.90	1.90
PD Hybrid (%)	5.20	6.40	2.03	2.35	2.10	2.16
PD PIT (%)	5.25	6.51	2.04	2.36	2.10	2.16

5 Discussion

In this section, the results and findings of the thesis are discussed and analysed. The section is divided based on objective of focus, first on the regression models and then on the economic consequences implied by the stressed economic scenario.

5.1 Discussion of Modelling

The framework for creating a macroeconomic indicator variable Z is based on the assumption that the credit change indicator X is normally distributed, as seen Figure 2. This assumption is fairly strong and uncertainty lies in whether it holds or not. It has a large impact on the applied shifts in the transition matrices and also the probability of default, which has to be taken into consideration. Considering the one-parameter framework, the assumption of one variable being able to reflect the whole movement of the portfolio is quite naive. The method implies that the applied shift occur according to a normal distribution which can raise questions regarding the method and corresponding assumptions. However, when creating a model with the purpose of inference and the independent variables effect on the dependent, there are clear advantages of having one single response variable assumed to be normally distributed. This simplifies answering questions regarding which variable affects the changes in the portfolio. However, for predictive purposes it has to be considered that a small error in the predicted value for $\hat{Z}$ can cause large changes in the transitions of the migration matrices and the probability of default. With this in mind, the regression models can now be investigated.

Four different methods for creating regression models for describing the Z series are created in this thesis. The first method, used for the creation of two linear baseline models, is linear regression. From the model in Section 4.1.1, it can be seen that some of the macroeconomic variables do not have an effect and they are therefore not included in the presented model. However, some variables do have relations that violates economic assumptions and theory. For example, an increase in GDP has a negative effect on the response which implies a higher than average probability of defaults and a larger degree of transitions from higher risk classes to lower. This violates previous research that shows that GDP has a positive effect on the development of SMEs [Woźniak, Duda, Gasior, Bernat, 2019]. This is the case for several of the predictors for the linear models. When the residuals for both of the linear models

are investigated, as described in Section 3, neither of them satisfies all of the assumptions. The violated assumptions are connected to independence of residuals and their variance. The heteroscedasticity does not necessarily make the OLS estimator biased but it leads to biased standard error of the coefficients, which therefore affects the t-test [Kaufman, 2013, p.13]. Thus, the estimates of the coefficients cannot with certainty be reliable and neither can the models. The impact of the violation of the constant variance assumption varies depending on the degree of heteroscedasticity. Consequently, how much the violation impacts the t-tests is not known. When investigating the linear models' predicted values for $\hat{Z}$ presented in Table 7, it can be seen that neither of the linear models have reasonable predictions, given that the worst observed values for the series of Z, as seen in the response to macroeconomic scatter plots, is around -1.7. This prediction is most likely the result of using linear models for non-linear relationships in addition to predictions using values of independent variables that are not in the sample and have not been seen before. Consequently, with regards to the violation of the assumptions and the presumed non-linear relationship, the linear models do not seem to be the best models for this problem.

The additive model includes unemployment and the long term-rate to estimate the response. The main effect of the independent variables, visualised in Figure 21, shows an escalation in downwards steepness as unemployment increases. The long term-rate pattern with respect to the response violates accepted economic theory since values of around 0 and 4 would indicate higher Z but lower around all other values. This is discussed further in the economic implication discussion. The assumptions of the residuals of the additive model are still not satisfied. Since the non-linear effects most likely have been taken into account, then either the explanatory variables included cannot explain the response, or there is some underlying problem with the data. There is a possibility that the additional data is the problem since the larger fluctuations in the residuals corresponds to the transitions between different data sets. When the error measurements for the cross validation in Table 6 are investigated it can be seen that the model is neither over- or underfitted due to the difference between the training and test errors being very small. The result of the predicted value $\hat{Z}$ for the additive model is in a reasonable range with regards to the observed data. However, it may seem quite small in absolute value given that the stressed scenario is adverse and the corresponding $\hat{Z}$ may not reflect that severity.

The tree from the XGBoost model is presented in Figure 22 and, even though un-

employment, long term-rate and house prices are included, the best XGBoost model uses unemployment exclusively. Hence, according to this model, unemployment rate should solely be used to describe the migrations between risk classes for an SME portfolio. The errors from the cross validation of the XGBoost model presented in Table 6 shows a quite large difference between the errors for the training and test. The model is therefore overfitted and not stable and there is a greater uncertainty in what range the error would be when using the model on out-of-sample data. The predicted $\hat{Z}$ for the stressed scenario, as shown in Table 7, is considered to be reasonable and a Z with that value would reflect an economic situations which, by comparison to previous years of unfavourable macroeconomic situations, is quite severe. What has to be taken into account however is that a tree model cannot successfully predict values outside of the sample and therefore, it is most likely to predict a $\hat{Z}$ with an unemployment as close to the out-of-sample unemployment as possible.

The SVM model is hard to interpret in terms of each predictors effect on the response. Nevertheless, the variables used for the model are unemployment, long term-rate and house prices which may give an indication of how the response is predicted since the combination of those resulted in the best model. The errors presented in Table 6 for the SVM model show signs of overfitting since the cross validation training error is smaller than that of the test sets. Thus, the model cannot be reliable in terms of how large an error is expected to be for out-of-sample predictions. The prediction $\hat{Z}$, presented in Table 7, is considered a bit too high for an adverse scenario and the SVM can potentially be seen as somewhat restrictive in its predictions.

When the models are compared, the linear ones are excluded since they clearly cannot explain the non-linear effects that the independent variables have on the response. Hence, by investigation of the AM, XGBoost and SVM model, it can be seen that the AM has the lowest error for both training and test as well as not being overfitted. Therefore, the AM model is the most stable as well and it also has the lowest average error. When the R^2 for the models is compared, it can be seen that the XGBoost is the best model since it has the highest value and therefore would be able to best explain the variance in the original data. However, it has to be considered that XGBoost is the model with the largest difference in training and test error and it could be one of the reasons why it has the highest R^2. To clarify, the R^2 is based on the variance training set and since the model performs better on the training set and is overfitted, the R^2 will naturally be high. It becomes eminent, given that the

AM model seems to have the lowest error and best fit in terms of the bias-variance trade-off, that the AM model has the highest predicted value of Z, which may seem like the least reasonable value of the three models. However, since it seems to be the best model in terms of fit and error, then the predicted value, which may seem small, might be the most accurate for the given economic scenario.

The additional data originating in the loss rate has the effect of enabling more stability in the models. With the benefit of more stable models, the choice of using additional data can be argued to be better even if the result is a slightly higher average error. The major concern is how to interpret the residuals if they are a result of a mix in the data used as the response. A way to avoid that problem is to use a model not subject to assumptions, but with the trade-off of having to use less interpretable models, like XGBoost or SVM. Consequently, there is a trade-off between using a model with violation of assumptions and a lower average error or a model without any necessary assumptions on data and a higher error and overfit.

5.2 Discussion of Economic Implications

The focus of this thesis lies in estimation of transition matrices for a portfolio in stressed market conditions through a univariate random variable Z that describes the portfolio movements corresponding to the market. First it must be decided to what extent the movements in the portfolio stem from movements in the market. The SME can initially be assumed to have a reasonably high correlation with the market which would entail a sound base for the construction of a model with good predictive abilities. This constitutes one of the arguments for the choice of a portfolio to model. By visual examination of the macroeconomic indicator variable Z, the movements fluctuate over time and, by comparison to other indexes, the series seems to match the overall movement quite well, with a slight lag.

The lag can stem from a few different sources. Either the initial credit risk rating models, used to create the data, do not react fast enough to changes in the macroeconomic variables, or the portfolio does not fluctuate from the changes at all. Another reason could be that the portfolio has a delayed reaction due to a chain of movements in the financial system. If the portfolio does not react to changes in the economic variables, it would indicate a lower degree of PIT for the portfolio than the observed $\alpha = 0.984$ in Section 3.4.1. Such a value of α indicates that the initial

model has a very high degree of PIT and should therefore react reasonably fast to financial fluctuations. As previously stated, the lag could however stem from the initial data, meaning that it might be too stable. The portfolio could then mostly consists of companies with stable finances that are not as affected by fluctuations in the market due to their resources to withstand an economic downturn, which could cause a delay in the response. By definition, the TTC estimate should not change in regards to external factors such as the macroeconomic variables. However, this is not the case for the PD of the different models as can be seen in Table 8, based on the method by Belkin and Suchower. It can be seen that the TTC PD is always slightly lower than the hybrid PD, which is lower than the PIT PD. This relation is expected for values of Z below 0, but not that the TTC should change to the extent that it does depending on the predicted response. For the predicted PD using the method by Carlehed and Petrov, as presented in Table 9, the TTC is lower and has exactly the same value for all predicted $\hat{Z}$. This is desired since it should be constant and shows that the method works. Since the same does not apply for the PD calculated from the transition matrices, the TTC and PIT calculations cannot be reliable since the TTC PD is not constant. However, the same relation, $TTC < Hybrid < PIT$, for the corresponding PD can be seen in Table 8. Consequently, the method of using the transition matrices for calculating the PD could be useful if larger fluctuations are desired. This can be observed since the XGBoost has a low Z and also a lower PD using the transition matrix than when using the method of Carlehed and Petrov. An extreme case of this can be observed in for the linear models. The contrary is true for the SVM and the AM models.

The predictor variables are plotted against Z in Section 3.3.1 and all but three seem to have random effect on the response. One with a vague observable pattern is the residential real estate variable which has a mostly random pattern and very weak tendencies. It is therefore not included in three of the models and the additive model because there is no visible pattern to replicate. In the results it is therefore concluded that the variables that seem to have an effect on the actual fluctuations are unemployment rate and long term-rate. The long term-rate however seems to be somewhat unlikely to possess a predictive ability in financial theory judging by its movement with time as well as with Z. The models are trained on data from 1985 up to 2020 during which time the long term-rate had a almost constant continuous decrease from 10% down to 1% [FRED, 2021]. With this in mind, it seems unlikely that this rate has an indicative correlation with the movements in the portfolio or the market. It is

therefore suggested to not be incorporated in future models. The models in this thesis are and base in statistical evidence to which the long term-rate could be argued for to include. The unemployment rate is, on the contrary to the long term-rate, included in various studies that have shown it to be the main indicator for movements of an SME portfolio [Bekeris, 2012]. This strengthens the outcome of the models that solely base the prediction of a stressed Z on the unemployment rate.

The stressed $\hat{Z}$ result in a PD estimate for every model, presented in Table 8. By reviewing the results it can be concluded that the resulting PD predictions during stressed market conditions for the XGBoost model, SVM model and the additive model are in a reasonable range based on historical data. However, given that the adverse scenario is out-of-sample, the resulting PD should intuitively be larger than that of any observed data. The mean of the PD presented in Table 8 is in the upper limits of the historical observed values which can be interpreted as the adverse scenario having an adverse result on the PD. The linear models' corresponding PD are however not plausible in any case due to the high values for their estimates and can therefore not be recommended for future models. This is because the linear models are used to model non-linear relationships.

It would be beneficial to have knowledge of the exact markets the companies in the portfolio act in and from there define the variables that would be reasonable to assume to have a predictive ability. Such information is not available for this thesis. The stress tests are conducted based on the EBA stressed scenario which uses a standardised numbers of macroeconomic variables that are the same for every bank in every country. This conditions the models to use the same macroeconomic variables even though other variables or indexes might have higher correlation and effect on the portfolio. The use of such variables could potentially increase the predictive ability and result in more useful information to base risk appetite on. In the case of a delayed effect of some macroeconomic variable on the response, the stress test does not allow such variables to be included since the test should be independent of previous states. For the current situation it has to be possible to test the models for any given time point and to receive a stressed prediction at the same time point without any information of the previous time point.

The EBA adverse scenario is inspired by the current situation of a global pandemic and therefore the scenario use extreme values for variables that are assumed to be the

most affected, such as unemployment rate, GDP, real estate and equity prices. It is a more extreme version of the current financial state. During the recent year, drastic negative changes have been observed in these variables, and the market has partially been stabilised by regulatory bodies and external institutions in various efforts. The data used in the models includes intrinsic monetary or fiscal policy reactions that can have an effect on the historical transitions in the data set. Hence, any historical transitions excluding those reactions are not known. In addition, it can be assumed that the data used for creating the models possess an intrinsic effect from monetary or fiscal policy reactions and in extension, the predictions may therefore assume such reactions, at least to some unknown degree. The extreme values that are adopted in the EBA stress tests are meant to be adverse. In practice, this means that the models have not been trained on any data similar to it before, complicating the prediction, especially for the tree-based models. It can cause the models to fail to predict such movements and only base the prediction on the previous worst outcome, which is not extreme enough for the stressed variables.

The stress tests are conducted on behalf of regulatory bodies such as EBA and the Swedish financial supervisory authority. The various actors have different intentions and desired outcomes when stress testing. The bank prioritises a model that has the ability to replicate a fair reaction to maximise the use of resources and be fully invested. For this purpose, the method by Belkin, Suchower [1998] would be more favourable. The regulatory bodies want a strong reaction which would entail a larger amount of held capital to ensure financial stability. For this purpose the method by Carlehed, Petrov [2012] is recommended since the reactions are amplified and the result is more extreme compared to the other method.

6 Conclusion

From the results in this thesis, it can be concluded that the Swedish SME-portfolio is highly affected by changes in unemployment rate and the long term-rate. However, the change in long term-rate in Sweden the last 30 years has been steadily declining, resulting in relations which are not reasonable for inference purposes. Therefore, unemployment rate can, and should, be the only covariate for such a model out of the macroeconomic variables included in this thesis. If one model is to be preferred, it is the additive model. It is fairly easy to see the effect of each of the covariates, it has a good balance in the bias-variance trade-off and the lowest general error when tested using cross validation. The one-parameter framework can be good for inference since it allows for univariate models which are easy to interpret. However, it relies heavily on a normal distribution assumption of the risk class migrations, which is convenient from a mathematical point of view, but most likely does not hold in reality. Therefore, with respect to the large changes an error of the predicted $\hat{Z}$ has on the resulting PD, it may not be the best framework to use for prediction in out-of-sample scenarios. The use of the parameter estimation of Carlehed and Petrov enables additional data from single indexes to be used which makes the models more stable. The probability of default from the predicted transition matrices are, when using the method by Belkin and Suchower based on transition matrices, higher for lower values of Z and vice versa, than when using the one by Carlehed and Petrov based on default frequencies. It can be concluded that usage of transition matrices as a base for probability of default gives a more volatile estimation while the use of the method by Carlehed and Petrov gives a more stable estimation.

6.1 Suggestions for Further Research

For future research it would be favourable to train the models on larger data sets that cover more extreme fluctuations and therefore gives a better indication of what is a more common state of the economy. If the data is biased then the model will be biased. It would also be interesting to try different methods when sampling data for the macroeconomic indicator series, Z, in the case where a richer data set is unobtainable. As previously discussed, the models are very dependent on the initial modelling of the macroeconomic indicator variable and errors will be included in the next model as well. A closer estimation of Z will improve the prerequisites for the other models. As for the XGBoost-model specifically, it would be recommended to

include new theory enabling a regression to model data outside of the sample data. This would predict a closer estimation of the movements of the data rather than choosing the most extreme point available in the data set.

References

Hull. *Risk Management and Financial Institutions*. Wiley, 2018.

BCBS. History of the Basel committee, 2019. URL `https://www.bis.org/bcbs/history.htm`. Accessed: 2021-02-10.

Murphy. *Understanding Risk; The Theory and Practice of Financial Risk Management*. Chapman Hall/CRC, 2008.

Wagner Bluhm, Overbeck. *An introduction to credit risk modelling*. Chapman Hall, 2003.

The European Banking Authority. Eba at a glance, 2016. URL `https://www.eba.europa.eu/about-us/eba-at-a-glance`. Accessed: 2021-04-29.

The European Banking Authority. Eu-wide stress testing, 2021. URL `https://www.eba.europa.eu/risk-analysis-and-data/eu-wide-stress-testing`. Accessed: 2021-04-29.

Finansinspektionen. Stress test methodology for determining a capital planning buffer, 2016.

Riksbanken. The riksbank's method for stress testing banks' capital, 2019.

Riksbanken. The riksbank's stress test of banks' capital – an update, 2020.

BCBS. Stress testing principles, 2018. URL `https://www.bis.org/bcbs/publ/d428.pdf`. Accessed: 2021-02-18.

Belkin, Suchower. A one-parameter representation of credit risk and transition matrices, 1998.

Carlehed, Petrov. A methodology for point-in-time-through-the-cycle probability of default decomposition in risk classification systems, 2012.

Wood. *Generalized Additive Models: an Introduction with R. 2ND ed.* Chapman Hall (2017), 2017.

Hastie Tibshirani James, Witten. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017.

Guestrin Chen. Xgboost: A scalable tree boosting system, 08 2016.

Awad. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress, 2015.

Awad, Khanna. *Efficient Learning Machines. Theories, concepts and applications for engineers and system designer*. Apress open, 2015.

European Commission. Sme definition, 2020. URL https://op.europa.eu/en/publication-detail/-/publication/5849c2fe-dcd9-410e-af37-1d375088e886. Accessed: 2021-04-20.

European Banking Authority. Eba launches 2021 eu-wide stress test exercise, 2021. URL https://www.eba.europa.eu/eba-launches-2021-eu-wide-stress-test-exercise. Accessed: 2021-04-01.

SCB. Konsumentprisindex (kpi), 2021. URL https://www.scb.se/hitta-statistik/statistik-efter-amne/priser-och-konsumtion/konsumentprisindex/konsumentprisindex-kpi/. Accessed: 2021-04-19.

Nasdaq. Omx stockholm pi (omxspi), 2021. URL https://indexes.nasdaqomx.com/Index/Overview/OMXSPI. Accessed: 2021-04-19.

SCB. Real estate price index, 2020. URL https://www.scb.se/en/finding-statistics/statistics-by-subject-area/housing-construction-and-building/real-estate-prices-and-registrations-of-title/real-estate-prices-and-registrations-of-title/pong/tables-and-graphs/real-estate-price-index-annually-1981100/. Accessed: 2021-04-19.

Alin. Multicollinearity, 2010. URL https://doi.org/10.1002/wics.84. Accessed: 2021-04-27.

Woźniak, Duda, Gasior, Bernat. Relations of gdp growth and development of smes in poland. *Elsevier*, 159:2470–2480, 2019.

Kaufman. *Heteroskedasticity in Regression: Detection and Correction*. SAGE Publications, 2013.

FRED. Long-term government bond yields: 10-year: Main (including benchmark) for the euro area, 2021. URL https://fred.stlouisfed.org/series/IRLTLT01EZM156N. Accessed: 2021-05-12.

Rokas Bekeris. The impact of macroeconomic indicators upon sme's profitability. *Ekonomika*, 91:117–128, 01 2012. doi: 10.15388/Ekon.2012.0.883.