University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Stein-Marten Pool

Forecasting loan default rates using macroeconomic variables

Actuarial and Financial Engineering

Master's Thesis (30 ECTS)

Supervisor: PhD Meelis Kääarik

Tartu 2025

FORECASTING LOAN DEFAULT RATES USING MACROECONOMIC VARIABLES

Master's thesis

Stein-Marten Pool

Abstract

The aim of this thesis is to model and forecast Estonian loan default rates using macroeconomic variables. Debt portfolio analysis and forecasting are important parts of credit risk management and are used in capital planning, risk assessment and scenario analysis. The selection of the final model and its suitability with the financial supervisors are discussed, highlighting the pros and cons of machine learning and simpler statistical models. The thesis also highlights the real-life value of default rate forecasting and describes its usage from the perspective of an Estonian bank - Coop Pank.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Keywords: loan defaults, forecasting, macroeconomic variables

MAKSEVIIVITUSE MÄÄRA PROGNOOS MAKROMAJANDUSLIKELT NÄITAJATELT

Stein-Marten Pool

Lühikokkuvõte

Antud teadustöö eesmärk on modelleerida ja prognoosida Eesti laenude makseviivituste määrasid makromajanduslikelt näitajatelt. Võlas oleva portfelli analüüs ja prognoosimine on oluline osa krediidiriski juhtimisest ja seda kasutatakse nii kapitali planeerimisel, riskide juhtimises, kui ka stsenaariumite analüüsil. Töös on käsitletud lõpliku mudeli valikut ja selle sobivust järelvaatajaga, tuues välja positiivsed ja negatiivsed küljed nii masinõppel, kui lihtsamatel statistilistel mudelitel. Töös tuuakse välja ka reaaleluline väärtus makseviivituste määrade prognoosimisel ning kirjeldatakse nende kasutust Coop Panga näitel.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: makseviivitused, prognoosimine, makromajanduslikud näitajad

Acknowledgements

I would like to express my deepest gratitude to my supervisor, PhD Meelis Käärik, for his invaluable guidance, constructive feedback, and consistent support throughout the course of this thesis. His expertise and advice have been fundamental to the development of this work.

I am also especially grateful to Professor Raul Kangro for his assistance and insightful suggestions in the forecasting section, particularly in the time series modeling. His input significantly improved the methodological part of the thesis.

My sincere thanks also goes to Theodora Nõmmik, who helped improve the readability and presentation of this work.

Finally, I would also like to thank Coop Pank, especially Kerstin Loss, for providing a practical context for the analysis.

Contents

Introduction **4**

1 Default rate modeling in the financial sector **5**

1.1 Default rate modeling 5

1.2 Regulatory requirements 6

1.2.1 The Estonian Financial Supervisory Authority 6

1.2.2 ICAAP 6

1.3 Usage of default rate models in Coop Pank 7

2 Datasets **9**

2.1 Modeling dataset description 9

2.2 Forecasting dataset description 10

3 Statistical models **11**

3.1 Regression models 11

3.2 Beta regression 12

3.3 ARIMA and SARIMA 13

3.4 ARIMAX and SARIMAX 14

4 Default rate forecasting **16**

4.1 Consumer loans 18

4.2 Home loans 22

4.3 Business loans 24

5 Discussion **26**

Conclusions **28**

References **29**

Appendix 1. Data sources **32**

Appendix 2. Time series models forecasts **33**

Introduction

The aim of this thesis is to find models for default rate forecasting. In this thesis, loan default rate is defined as a portfolio rate that is over 90 days in debt. Loan default rate is a widely used indicator for credit risk and its situation in the financial sector. The default rates of different loan types are often examined separately, for example consumer loans, business loans and home loans, due to their different behaviors and simplicity of accrediting.

Default rate modeling holds significant practical value in the banking and financial sector. Besides being necessary for regulatory documents, they are also used in internal documents and analysis. To highlight the real-life value of modeling and forecasting, this thesis also describes how default rate forecasts are used in Coop Pank, an Estonian bank.

Default rate modeling is based on historic macroeconomic and default rate data. Eesti Pank forecasts are used as a base scenario for the default rate forecasting. Scenario testing, such as stress and base scenarios, is also discussed in this thesis, which can be performed using the models fitted in this thesis.

Credit risk models, including default rate modeling, are supervised by the financial supervisory authority, which indicates that not all models could be used for forecasting. More complex models, such as machine learning models, could show better accuracy than simpler models like regressions or time series models but may not be chosen for the forecasting model due to lacking interpretability and explainability. The process of selecting the final model and forecasting is the main focus of the thesis.

1 Default rate modeling in the financial sector

1.1 Default rate modeling

Various studies indicate that macroeconomic variables such as GDP growth, unemployment rate and inflation are effective predictors for forecasting nonperforming loans. Models based on EU data demonstrate a correlation between debt portfolio ratio and macroeconomic variables. In addition to those variables, short-term historical data of overdue loans is also strongly correlated with the future occurrence of non-performing loans, making it a useful indicator for forecasting. (Staehr and Uusküla, 2017)

Default rate models can be used for different scenarios by employing scenario-based forecasts of macroeconomic variables for scenario testing. For example, a base scenario with forecasts of macroeconomic variables can be used for real life forecasting, however a stress scenario of macroeconomic forecasts can be used for stress-testing. It is useful for scenario testing and for example, risk assessment, where forecasts show the possible outcomes of the stress scenario. (Coop Pank, 2025c)

In scenario testing, it is important that all the variables in the model, on which the forecasts of default rates are made, come from the same scenario. This is important because models are trained on a single real-world scenario - historical data - where macroeconomic variables interacted and influenced default rates in the specific way they did at that time. Therefore, for the forecasts, also one single scenario should be used to get the default rate forecast which would be the best statistical prediction of the future, if that scenario happened. For example, if we use macroeconomic variables from stress scenarios provided by both the European Banking Authority and the Estonian Ministry of Finance, the resulting future dataset would contain forecasts based on different levels of stress. This could lead to inconsistencies in

the default rate forecast, potentially reducing its reliability. (Coop Pank, 2025a)

1.2 Regulatory requirements

Regulatory requirements are a key driver of default rate modeling. Banking supervision and regulatory requirements are meant to safeguard the stability of the banking system and protect clients from bank failure. Thus, financial supervision oversees credit risk modeling to ensure correctness in risk-taking and credit scoring. (Malikkidou and Strohbach, 2025)

1.2.1 The Estonian Financial Supervisory Authority

The Estonian Financial Supervisory Authority (Finantsinspektsioon) supervises the banks in Estonia. Finantsinspektsioon is autonomous in its decisions but is part of Eesti Pank's budget. Finantsinspektsioon is also responsible for supervising the financial sector's risk taking decisions, including credit risk. Therefore, default rate models included in the thesis are supervised by Finantsinspektsioon and must comply with the guidelines given by them. (Riigi teataja, 2025)

1.2.2 ICAAP

Internal capital adequacy assessment process (ICAAP) is one example of the practical need of default rate models. The ICAAP report is a document that gives detailed information about the bank's risk, stress testing results and capital planning. That way the bank's board and supervisors can evaluate the bank's capital adequacy in different scenarios and make future decisions based on ICAAP results. Default rate models can be used for stress scenario testing and risk evaluating and also for base scenario parts of the report like capital planning and general credit risk overview. In Estonia, banks send their ICAAP reports to Finantsinspektsioon. (European Central Bank, 2018)

1.3 Usage of default rate models in Coop Pank

The practical value of default rate modeling and forecasting is described by the credit risk department of Coop Pank AS. Coop Pank is a growing Estonian bank that is known for its dedication to improving life all around Estonia and its cooperation with Coop Estonia. (Coop Pank, n.d.)

Default rate models are widely used in Coop Pank and an important part in credit risk, financial and business planning. Different departments make forecasts for internal documents as well as external documents like regulatory required documents. Since the usage is wide and forecasts serve different purposes, the scenarios also vary between the forecasts. (Coop Pank, 2024a; Coop Pank, 2024b; Coop Pank, 2025c)

Regulatory-required documents like the ICAAP report and credit risk reports for Finantsinspektsioon may include both, base scenario forecasts and stress scenario forecasts. Stress scenario forecasts of default rates are used to predict debt portfolio activity in a stress scenario and therefore show the risk case of the portfolio. The base scenario is used for real life forecasts and show what is the best prediction of the debt portfolio when the macroeconomic indicators move as forecasted by real-life forecasts. Every regulatory document is supervised by Finantsinspektsioon which includes the evaluation of the default rate forecasts and modeling. Therefore, models used for forecasting must be statistically correct and the modeling process must be described in detail to make it understandable for both parties. (Coop Pank, 2025c)

Internal documents like financial plans and risk documents also include default rate forecasts. The financial plan takes into account loan provision and default estimates that are forecasted using the model. Default rate forecasts could also be used for edge case scenarios which indicate a macroeconomic situation where

the bank would default itself. Backward calculations like these could be useful for capital planning and risk assessment. (Coop Pank, 2024a)

Default rate forecasts are also part of the probability of default models in Coop Pank. Probability of default models are contract-based models that evaluate the risk of an individual contract defaulting. Additionally, to the contract-based and client-based variables, Coop Pank also incorporates default rate forecasts into its probability of default models, ensuring that model outcomes are sensitive to changes in the macroeconomic situation. (Coop Pank, 2025b)

2 Datasets

2.1 Modeling dataset description

The dataset used for modeling includes default rate data for three different loan types and macroeconomic data. Dataset was compiled from various sources, which are listed in Appendix 1. Data from Statistics Estonia (Statistikaamet) was retrieved via the official Statistikaamet database API (Statistikaamet, n.d.[b]). Additional data was collected from the listed webpages using the CSV or XLSX format. All data manipulation and the final dataset assembly were carried out using R software.

Several macroeconomic measures were selected based on their inclusion in Eesti Pank's base scenario forecasts. In the thesis, it was necessary to have forecasts of the macroeconomic variables available in order to use the model for forecasting. Macroeconomic variables in the dataset were:

- **GDP**: Chain index of Estonian seasonally and working day unadjusted gross domestic product (quarterly), with the reference year 2010.
- **CPI**: Consumer price index (monthly) in Estonia, with the reference year 1997.
- **Unemployment rate**: Unemployment rate (quarterly) in Estonia, where the unemployed population is divided by the total employment number.
- **Salary**: Average monthly gross wages (monthly) in Estonia, measured in euros.
- **Euribor**: 3-month Euribor rate (monthly).
- **EUR/USD rate**: Euro and United States dollar exchange rate (monthly).

- **Oil price**: Oil price (monthly) in United States dollars per barrel using Brent index.

Loan types for which default rates were forecasted were:

- **Home loan**: Over 90 days in debt home loans rate (monthly) of the Estonian home loans portfolio.
- **Consumer loan**: Over 90 days in debt consumer loans rate (monthly) of the Estonian consumer loans portfolio.
- **Business loan**: Over 90 days in debt business loans rate (monthly) of the Estonian business loans portfolio.

The dataset consists of data from December 2000 to December 2024. The starting point is chosen based on the availability of statistics and data points, meaning there is no data in the chosen sources for earlier time points. All variables have quarterly data, and some have monthly data. Due to this, the monthly data was calculated for all variables using the proportional difference between two quarters. Therefore the dataset has 289 months worth of data displayed in Appendix 1.

2.2 Forecasting dataset description

The dataset used for forecasting consists of the same macroeconomic variables used for modeling, but the values are for future time points. Macroeconomic forecasts are made by Eesti Pank and for the end of the years 2025, 2026 and 2027 (Eesti Pank, 2024). Since only annual forecasts are available for those years, monthly values are calculated using proportional differences between forecasts. The forecasting dataset does not contain default rates data because those are modeled and forecasted in this thesis. The forecasting dataset is also displayed in Appendix 1.

3 Statistical models

3.1 Regression models

A variety of statistical methods can be used for forecasting. One of the simplest and most commonly used statistical methods for modeling are regression models. For example, linear and logistic regression models are widely used in statistical modeling. These models use regressors to predict the response variable. Assumption testing is important before using regression models. These assumptions must be met to ensure the model's validity. One of the assumptions is that the observations have to be independent of each other. (Liang and Zeger, 1993)

Since the time series of default rates are modeled in this thesis, it can be shown that the values are dependent on the previous values violating the assumption of independence. Linear regression models with the default rate as response variables and default rate value with 1-step lag as regressors were fitted. R^2 values and p-values of the models are displayed in Table 1. It can be seen that default rates are dependent of the previous values and the model with only previous values has high predictive power, which means that the regression models are not a good fit for this thesis dataset.

Table 1: Default rate models overview with 1 lag

Loan type	Model p-value	R^2
Consumer loans	<0.0001	98.92%
Home loans	<0.0001	99.45%
Business loans	<0.0001	98.87%

3.2 Beta regression

More complex regression models, such as the beta regression model, are used to achieve more accurate results. A beta regression model with a logit link is used for data where dependent variable is between 0 and 1, as it is in this thesis. The beta regression model can be expressed as

$$g(\mu_i) = X_i^T \boldsymbol{\beta},$$

where g is a link function, μ_i is the mean of the distribution, X_i^T is a transposed vector of covariates and $\boldsymbol{\beta}$ is a vector of regression parameters. (Ferrari and Cribari-Neto, 2004)

For beta regression, the assumption of independence remains, meaning the assumptions are violated. Another beta regression assumption is that the error term variances in beta regression have to be constant and therefore variance of the error rates must be tested (Ferrari and Cribari-Neto, 2004). To test these assumptions, a beta regression model with a logit link was fitted to the previously described dataset. Logit link function is commonly used for regression models and can be expressed as

$$\mathit{logit}(u) = \ln\left(\frac{u}{1-u}\right),$$

where *logit* is the logit link function and u is the function parameter (Cox, 1958). The final model was found for all the loan types. Final model selection process was applied where all the macroeconomic variables with $p \geq 0.05$ were dropped one by one and all p-values in the final model were smaller than 0.05.

The assumptions of the final models were tested and the Breusch-Pagan test was used to determine if the error term variances were constant (Koenker, 1981). The results are displayed in Table 2, where p-values indicate that the error term vari-

ance assumption is not met for any of the models and thus beta regression cannot be used.

Table 2: Beta regression model assumptions test

Loan type	Breusch-Pagan test p-value
Business loans	<0.0001
Consumer loans	<0.0001
Home loans	0.0004

3.3 ARIMA and SARIMA

Various studies on default rate and debt portfolio rate modeling show that historical data points of the dependent variables often improve the accuracy of forecasting models. Therefore, time series models are often used for modeling financial data. (Staehr and Uusküla, 2017)

Time series models like ARIMA are used to model time series and forecast future values based solely on historical data. The autoregressive integrated moving average (ARIMA) model introduced in 1970 by Box and Jenkins, is used to forecast future values based on historical values of the same variable as a time series. The ARIMA model is used to show patterns with no seasonality and no white noise. ARIMA model can be displayed as

$$\phi_p(B)(1-B)^d Z_t = c + \theta_q(B)\varepsilon_t,$$

where B is lag operator, ϕ_p is p-th order autoregressive operator, θ_q is q-th order moving average operator, $(1-B)^d$ is d-th order differencing operator, Z_t is time series point at time t, c is constant and ε_t is residual error. It can be seen that the ARIMA model has three parameters and is displayed as ARIMA(p, d, q). (Box

and Jenkins, 1970)

Since the ARIMA model cannot detect seasonality in time series, a seasonal autoregressive integrated moving average (SARIMA) model may prove to be more appropriate. The SARIMA model also includes a seasonal component in addition to the ARIMA formula. The SARIMA model can be expressed as

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D Z_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t,$$

where $\Phi_p(B)$ is P-th order seasonal autoregressive operator, $\Theta_Q(B)$ is Q-th order seasonal moving average operator, $(1-B)^D$ is D-th order seasonal differencing operator and S is seasonal length. SARIMA model has therefore seven parameters and is displayed as ARIMA$(p,d,q)(P,D,Q)_S$. (Box and Jenkins, 1970)

3.4 ARIMAX and SARIMAX

Although, ARIMA and SARIMA models are often used for time series modeling, they cannot be used for the problem stated in this thesis. Since the aim of the thesis is to use macroeconomic variables in the model, using a model with external variables is needed. Therefore, it is necessary to look for more complex time series models such as ARIMAX and SARIMAX.

Autoregressive integrated moving average model with external variables and Seasonal autoregressive integrated moving average model with external variables, respectively ARIMAX and SARIMAX models, are used for time series modeling where external variables are taken into account. For response variable Y on time point t, both models can be expressed as

$$Y_t = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \omega_t$$

where X_j, $j \in 1, 2, ..., k$ are observations of the external variables corresponding to time point t, β_0 is intercept, β_j, $j \in 1, 2, ..., k$ are regression coefficients of external variables and ω_t is a variable that differs by the time series usage. Since several lags of external (macroeconomic) variables can be used for forecasting, the model with the set of selected lags T_j for each $j \in \{1, 2, ..., k\}$ can be expressed as

$$Y_t = \beta_0 + \sum_{j=1}^{k} \sum_{t_j \in T_j} \beta_{j,t_j} X_{j,t_j} + \omega_t$$

where X_{j,t_j}, $j \in 1, 2, ..., k$ are observations of the external variables corresponding to time point t_j, β_0 is intercept, β_{j,t_j}, $j \in 1, 2, ..., k$ are regression coefficients of external variables and ω_t is corresponding to the residuals of the chosen model. For ARIMAX model, ω_t can be expressed as

$$\omega_t = \frac{c + \theta_q(B)\varepsilon_t}{\phi_p(B)(1-B)^d}$$

and for the SARIMAX model, as

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^S)\varepsilon_t}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}.$$

It is easy to see that the ω_t for the ARIMAX and SARIMAX is derivable from the ARIMA and SARIMA formulas respectively. (Hamilton, 1994)

4 Default rate forecasting

Default rate modeling and forecasting were conducted using time series models described in the previous section. Separate models were fitted for each loan type and used for forecasting. The models were trained on the full historical dataset introduced in the dataset section.

The training and testing datasets were created to validate the final model selection. The training dataset consisted of the first 230 rows from the original dataset. The next 36 rows were chosen for the testing dataset since the aim of this thesis is to forecast the next 36 months' default rates. Each loan type's final model was validated on the testing dataset using ARIMA and beta regression models fitted on the training dataset. Root mean square error (RMSE) and mean absolute error (MAE) were used to test accuracy between the forecasts and the real default rates in the testing dataset.

On the training dataset, the final model selection process described below was applied to find (S)ARIMAX model. The (S)ARIMA model with the same AR and MA components as the (S)ARIMAX was fitted without the macroeconomic variables. The beta regression model was fitted using the default rate as the response variable and the selected macroeconomic variables, with their respective lags, as regressors.

Since the forecasted value must be positive, logarithmic transformation was applied to the default rate data prior to time series modeling. To assess stationarity, the Augmented Dickey-Fuller test was applied to all macroeconomic variables and default rate series individually (Cheung and Lai, 1995). As the test p-values were all greater than 0.05, the entire dataset was differenced. After differencing, all Augmented Dickey-Fuller test p-values fell below 0.01, confirming stationarity and

making the differenced dataset suitable for modeling.

To identify significant lags of the macroeconomic variables, the prewhitening process was employed. For each macroeconomic variable, an appropriate (S)ARIMA model was estimated using the auto.arima() function in R (Rdocumentation, n.d.) and residuals of the fitted model were computed. The same model was applied to the default rate time series and its residuals were also calculated. These residuals were used in the cross-correlation function (CCF) plot to detect statistically significant lags between the macroeconomic variable and default rate series (Brockwell and Davis, 1991).

Significant lags identified through CCF plots were incorporated as regressors in the final forecasting model. Two approaches were explored to specify the autoregressive (AR) and moving average (MA) components. First, the auto.arima() function was applied to the default rate series with the lagged macroeconomic variables as external regressors. Second, a multiple linear regression (MLR) model was fitted using the default rate as the response variable and the lagged macroeconomic variables as predictors. Residuals from the MLR model were used to plot the autocorrelation function (ACF) and partial autocorrelation function (PACF), which assisted in identifying suitable AR and MA components.

Default rate values of each loan type were forecasted using the final models. Due to the inclusion of lagged default rates in the (S)ARIMAX models, the next 36 months were forecasted iteratively, one step at a time. All the forecasted values were exponentiated (raised to the power of e) to reverse the initial logarithmic transformation and return the results to their original scale.

4.1 Consumer loans

The modeling procedure described above was applied to consumer loan default rates. As an illustration of the macroeconomic lag selection process, Figure 1 presents the CCF plot between GDP and consumer loan default rates. For this example, the final model used to calculate residuals for both the GDP and consumer loan default rate series was ARIMAX(2,1,1)(1,1,2)[12].

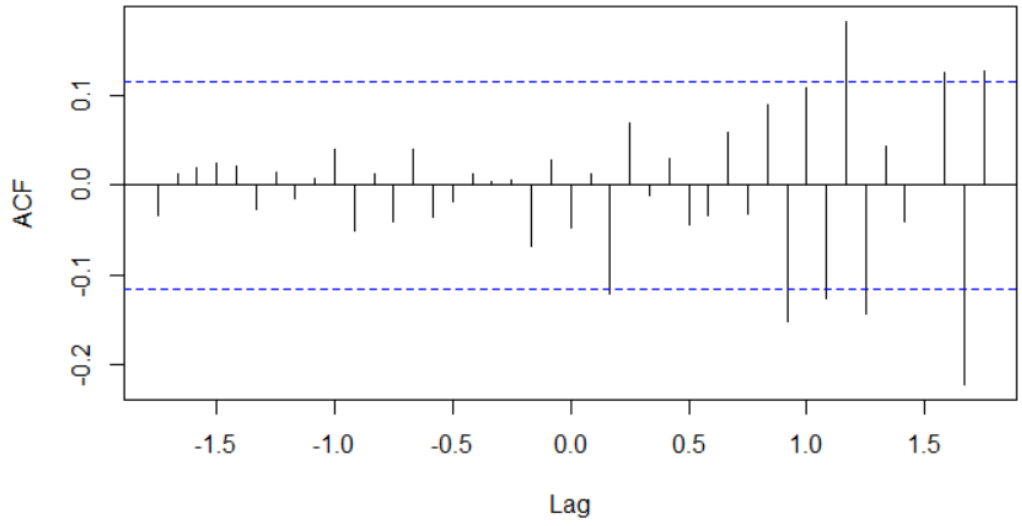

Figure 1: CCF Plot between GDP and consumer loan default rate

The x-axis in Figure 1 represents the lag number, where the lag 1 corresponds to one year, or 12 months. Every vertical spike in the graph shows the strength and direction of correlation at a specific lag. Spikes extending beyond the significance bounds indicate statistically significant correlations between the prewhitened GDP and the residuals of the consumer loan default rate at the corresponding lag. Each lag in the plot corresponds to a one-month interval. Negative lags represent GDP values that precede the default rate value, while positive lags represent GDP values following it. In this modeling process, only non-positive lags were considered,

since the future values of GDP cannot be used in real-time forecasting models. The significant lags for the final model are those where the spikes exceed the 95% confidence level (blue dashed line) in the CCF plot. Therefore, in the case of GDP, no significant non-positive lags were identified, meaning that GDP was ultimately not included as a regressor in the final consumer loan default rate model.

This process was completed with each macroeconomic variable. Significant macroeconomic variables with their respective lags in the consumer loan model were:

- CPI - 13
- Unemployment rate - 17
- Oil price - 0
- EUR/USD rate - 21

Initially, the auto.arima() function was fitted on the consumer loan default rates with macroeconomic variables and their respective lags. The final model from the function output was ARIMAX(1,1,0) with drift and the formula can be expressed as:

$$\begin{aligned}(\Delta Z)_t = &- 0.0113 - 0.1509 \cdot (\Delta Z)_{t-1} + 0.0118 \cdot (\Delta CPI)_{t-13} \\ &- 2.6316 \cdot (\Delta Unemployment\,rate)_{t-17} + 0.0016 \cdot (\Delta Oil\,price)_t \\ &- 0.3787 \cdot (\Delta EUR/USD\,rate)_{t-21},\end{aligned} \tag{1}$$

where Z is log-transformed consumer loan default rate, ΔZ denotes the first order difference of Z and t represents a certain time point.

For the second model selection method, the previously described MLR model was fitted. The ACF and PACF were applied on the MLR model residuals and are displayed respectively in Figure 2 and Figure 3.

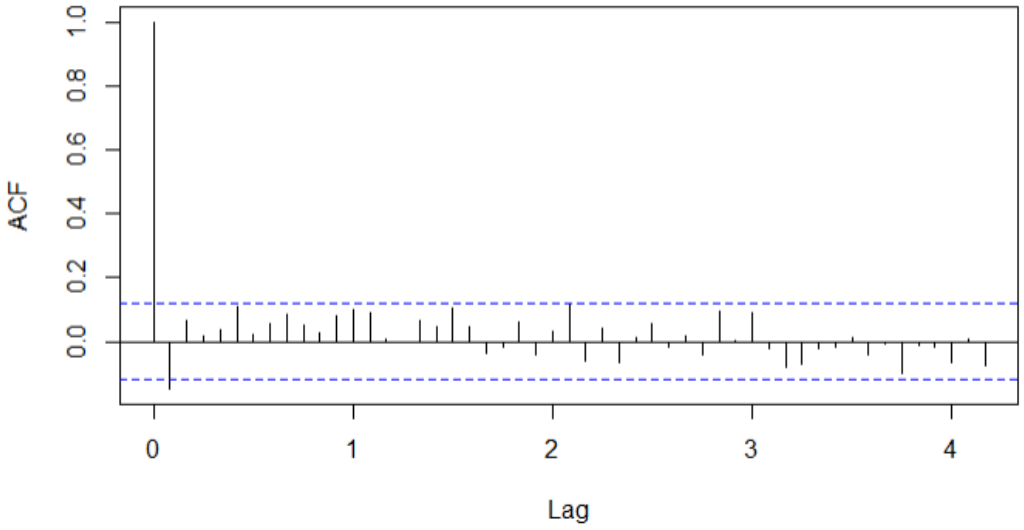

Figure 2: Consumer loan default rate MLR ACF plot

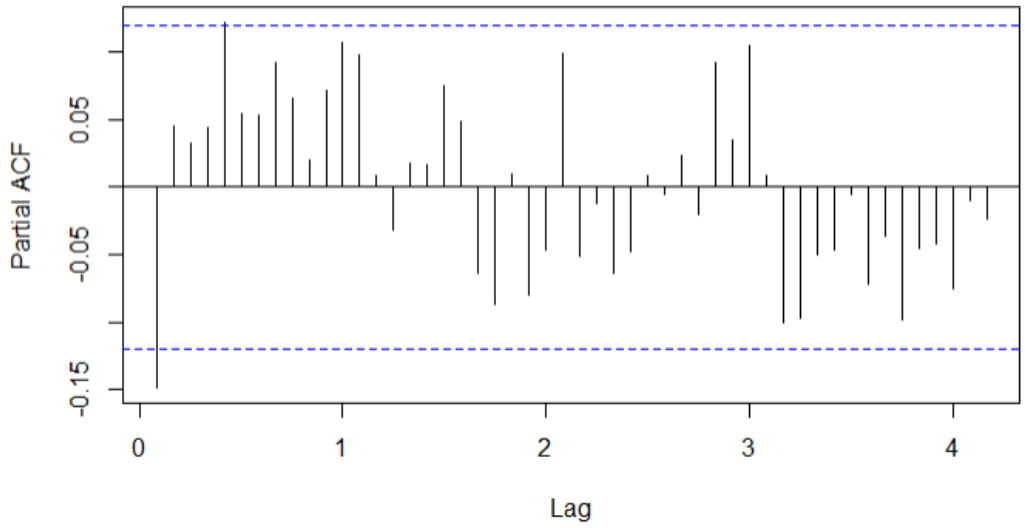

Figure 3: Consumer loan default rate MLR PACF plot

The model selected with the auto.arima function was the same that was selected

on the ACF and PACF plots. Therefore, the model described before was used for consumer loan default rate forecasting. Forecasts were calculated using the forecasting dataset and Equation 1. The forecast values are presented in Appendix 2 and the plot with the historical and forecasted values is shown in Figure 4.

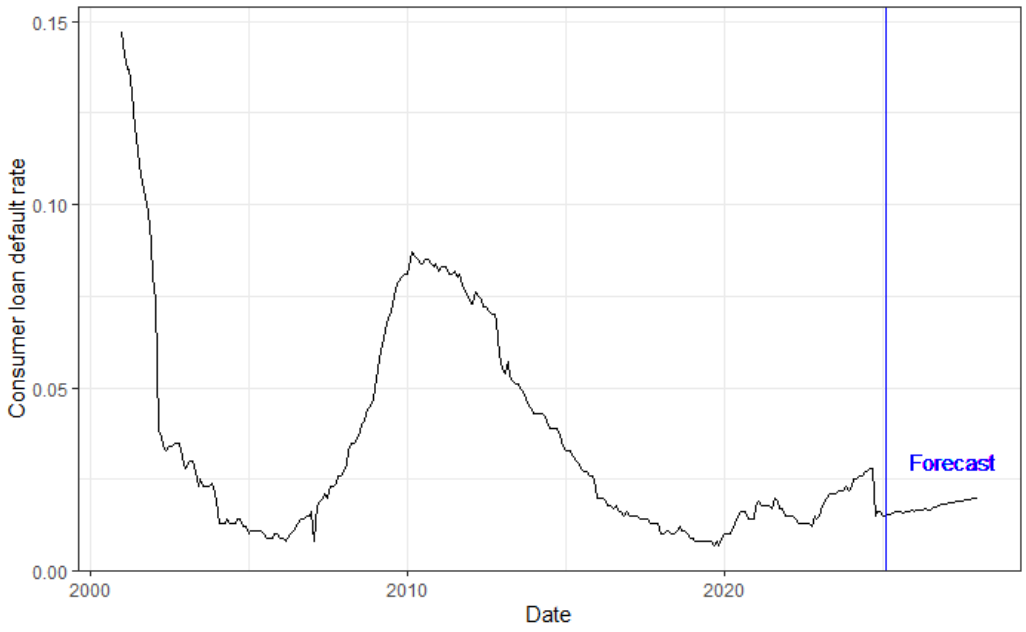

Figure 4: The consumer loan default rate data with forecasts

Previously described model validation process was applied to the consumer loan default rate models. The final model on training dataset was ARIMA(1,1,0) with macroeconomic variables with their respective lags:

- CPI - 17
- Oil price - 0
- EUR/USD rate - 11

All the fitted models and their RMSE values are displayed in Table 3. As expected, the ARIMAX model yielded the lowest RMSE and MAE values, followed

by ARIMA and beta regression models.

Table 3: Consumer loan model validation RMSE values

Model type	RMSE value	MAE value
ARIMAX	$2.04 \cdot 10^{-5}$	$3.89 \cdot 10^{-3}$
ARIMA	$2.94 \cdot 10^{-5}$	$4.97 \cdot 10^{-3}$
Beta regression	$4.33 \cdot 10^{-5}$	$5.76 \cdot 10^{-3}$

4.2 Home loans

The model selection process described previously was also applied to home loan default rates. Macroeconomic variables with their respective lags in the home loan default rate model are:

- GDP - $4, 5, 6$
- CPI - 7

The final model for the log-transformed home loan default rates was ARIMAX(1,1,0), which was again chosen with two overlapping methods. The model equation can be expressed as:

$$\begin{aligned}(\Delta Z)_t = &0.2513 \cdot (\Delta Z)_{t-1} - 0.0188 \cdot (\Delta GDP)_{t-4} \\ &+ 0.0138 \cdot (\Delta GDP)_{t-5} - 0.0198 \cdot (\Delta GDP)_{t-6} \\ &+ 1.8462 \cdot 10^{-3} \cdot (\Delta CPI)_{t-7},\end{aligned} \tag{2}$$

where Z is log-transformed home loan default rate, ΔZ denotes the first order difference of Z and t represents a certain time point.

The home loan default rate forecasts were calculated using the forecasting dataset

and Equation 2. The forecasted values are presented in Appendix 2. The historical and forecasted values are visualized in Figure 5.

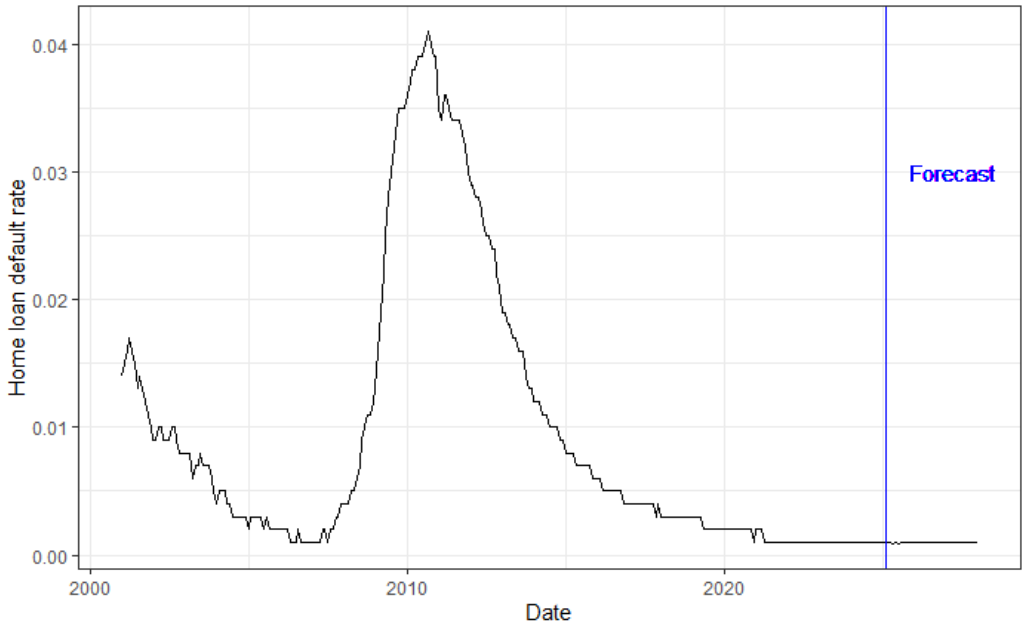

Figure 5: The home loan default rate data with forecasts

Model validation process was applied to the home loan default rate models. The final model on the training dataset was ARIMA(1,1,0) with macroeconomic variables with their respective lags:

- CPI - 7
- OCPI - 8

RMSE values of all the fitted models are displayed in Table 4. As expected, ARIMAX showed the lowest RMSE value, following ARIMA and beta regression models.

Table 4: Home loan model validation RMSE values

Model type	RMSE value	MAE value
ARIMAX	$1.28 \cdot 10^{-7}$	$2.37 \cdot 10^{-4}$
ARIMA	$1.37 \cdot 10^{-7}$	$2.62 \cdot 10^{-4}$
Beta regression	$1.33 \cdot 10^{-5}$	$1.12 \cdot 10^{-2}$

4.3 Business loans

Finally, for the business loan default rate modeling, the same model selection procedure was applied. The final model was again the same with both model selection procedures and the selected model for forecasting the log-transformed business loan default rates was ARIMAX(0,1,0)(1,0,0)[12]. The macroeconomic variables and their corresponding lags used in this model are:

- GDP - 18
- Oil price - 0

The final model can be expressed as:

$$\begin{aligned}(\Delta Z)_t = &0.2449 \cdot (\Delta Z)_{t-12} - 7.3581 \cdot 10^{-4} \cdot (\Delta GDP)_{t-18} \\ &+ 2.6634 \cdot 10^{-3} \cdot (\Delta Oil\,price)_t,\end{aligned} \tag{3}$$

where Z is log-transformed business loan default rate, ΔZ denotes the first order difference of Z and t represents a certain time point.

Forecasts of the business loan default rates were calculated using the forecasting dataset and Equation 3. The forecasted values are presented in Appendix 2 and plotted with the historical values in Figure 6.

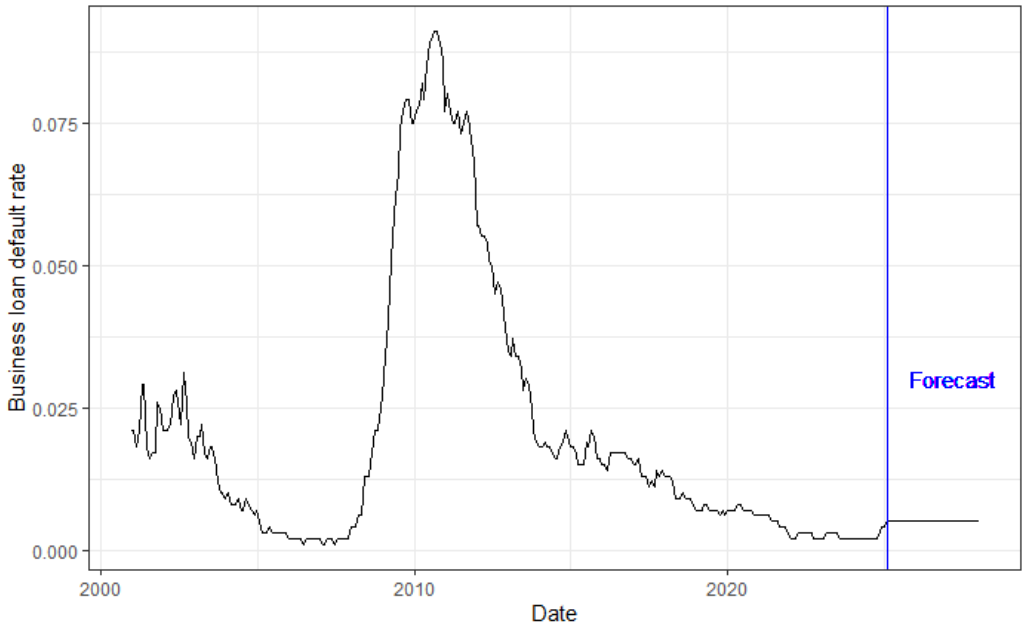

Figure 6: The business loan default rate data with forecasts

Business loan default rate model was also validated with the same validation process. The final model on training dataset was ARIMA(0,1,0)(2,0,0)[12] with macroeconomic variables with their respective lags:

- CPI - 4

RMSE values of all the fitted models are displayed in Table 5. Model with the lowest RMSE value was SARIMAX, following ARIMA and beta regression models.

Table 5: Business loan model validation RMSE values

Model type	RMSE value	MAE value
ARIMAX	$0.98 \cdot 10^{-5}$	$2.41 \cdot 10^{-3}$
ARIMA	$1.16 \cdot 10^{-5}$	$2.77 \cdot 10^{-3}$
Beta regression	$2.28 \cdot 10^{-4}$	$1.5 \cdot 10^{-2}$

5 Discussion

In a recent research paper by the European Banking Authority, random forest models outperformed simpler, widely used models - specifically logistic regression - in forecasting banking data. In the aforementioned paper, the authors note that improvements in predictive accuracy may come at the expense of interpretability and explainability. Although machine learning models present challenges in terms of explainability, the most accurate model should still be used if possible. Model selection is influenced by regulatory requirements, which may prefer simpler models. However, in the future, if machine learning models and their interpretation improve, more complex models could be used. (Malikkidou and Strohbach, 2025)

Since the data and forecasts in this thesis are based on Estonian data, all modeling must adhere to the expectations of the Estonian financial supervisory authority, Finantsinspektsioon. This institution oversees the models and forecasts used in regulatory documents, assessing their validity and interpretability. According to Coop Pank's experience, Finantsinspektsioon tends to favor simpler models over complex machine learning models that are difficult to interpret. Moreover, default rate forecasts are used in numerous annual calculations and reports subject to regulatory oversight. Based on both Finantsinspektsioon's recommendations and Coop Pank's experience, time series models are used in this thesis for default rate forecasting.

For internal analyses and calculations that are not subject to regulatory review, more accurate but less interpretable models may still be beneficial. These models can enhance forecasting accuracy, but may hinder effective communication and shared understanding within the organization. In this thesis, a single forecasting model is recommended for both external regulatory documentation and internal corporate use. However, the needs and context of each use case should still be

carefully considered.

The final models used for forecasting default rates are all (S)ARIMAX models, fitted on differenced datasets. The business loan final model was a SARIMAX model with yearly seasonality and a seasonal autoregressive component. In contrast, the consumer and home loan default rate models do not include any seasonal components. Each loan type's model also incorporates different lags of different macroeconomic variables, indicating that the relationships between macroeconomic indicators and default rates vary across loan types. It was validated that the final models show better forecasting accuracy than fitted ARIMA and beta regression models.

The default rate forecasts presented in the forecasting section of the thesis indicate that under the base macroeconomic scenario, the predicted default rate forecasts remain largely consistent with last year's values. While the business and home loan default rates remain relatively stable, with only a slight decrease, the consumer loan default rates noticeably increase in the next three years compared to the latest observed value. These results are expected, given that the base scenario reflects macroeconomic conditions similar to those of the previous year.

As discussed in this thesis, the default rate models can also be used for stress scenario forecasting. This involves applying macroeconomic forecasts from a stress scenario to the same models developed during the forecasting stage. Since the scenario affects only the input data used in forecasting, not the model itself, any number of alternative scenarios can be tested using the established models.

Conclusions

The aim of this thesis was to model and forecast default rates in Estonia, focusing on three loan categories: home loans, consumer loans and business loans. A variety of modeling approaches were considered, including regression models, time series models and machine learning models. The models incorporated macroeconomic variables, including those with available forecasts, to enhance predictive accuracy. The primary outcome of the thesis was the selection of appropriate models and the generation of base scenario forecasts for each loan type.

These forecasts are valuable tools in the financial sector, supporting internal analysis, regulatory reporting, capital planning, risk assessment and scenario analysis. Although this thesis focused on base scenarios, the chosen models allow for future testing under alternative scenarios. The real-life applicability of the forecasting models was illustrated using Coop Pank as an example.

After evaluating different methods, time series models were ultimately selected for their simplicity, interpretability and effectiveness. A separate model was fitted for every loan type to provide tailored forecasts. The final models included an ARIMAX model for consumer loans and SARIMAX models for home and business loans. The resulting base scenario forecasts indicated that default rates are expected to remain relatively stable in near future.

While machine learning models theoretically offer improved accuracy, their complexity and lower transparency can hinder their adoption in regulated environments. Interpretability and explainability remain crucial for practical implementation. As machine learning models become more interpretable, their use in forecasting may increase. However, for the purposes of this thesis, time series models were the most appropriate due to their balance of performance and clarity.

References

Box, George and Gwilym Jenkins (1970). *Time Series Analysis: Forecasting and Control*.

Brockwell, Peter J. and Richard A. Davis (1991). *Time series: Theory and Methods*.

Cheung, Yin-Wong and Kon S Lai (1995). "Lag order and critical values of the augmented Dickey–Fuller test". In: *Journal of Business & Economic Statistics*.

Coop Pank (n.d.). *About Coop Pank*. URL: `https://www.cooppank.ee/en/coop-pank/about-coop-pank` (visited on 04/13/2025).

– (Dec. 2024a). "Allahindlused". Principles of credit claim assessment.

– (Nov. 2024b). "Pandikirjaportfelli stressitestide läbiviimine". Conducting Stress Tests on the Covered Bond Portfolio.

– (May 2025a). "Makseviivituse määra modeleerimine". Description of the default rate modeling.

– (Apr. 2025b). "PD mudelite dokumentatsioon". Probability of default models documentation.

– (Mar. 2025c). "Stressitestide läbiviimine". Stress-testing process.

Cox, David R (1958). "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2, pp. 215–232.

Eesti Pank (n.d.[a]). *Tähtajaks tasumata laenude jääk (miljon eurot)*. URL: `https://statistika.eestipank.ee/#/et/p/650/r/2670/2459` (visited on 04/13/2025).

– (n.d.[b]). *Valuutakursid*. URL: `https://www.eestipank.ee/valuutakursid` (visited on 04/13/2025).

Eesti Pank (Dec. 2024). *Rahapoliitika ja Majandus 4/2024*. URL: https://www.eestipank.ee/publikatsioonid/rahapoliitika-ja-majandus/2024/rahapoliitika-ja-majandus-42024 (visited on 04/13/2025).

European Central Bank (n.d.). *Euribor 3-month - Historical close, average of observations through period, Euro area (changing composition), Monthly*. URL: https://data.ecb.europa.eu/data/datasets/FM/FM.M.U2.EUR.RT.MM.EURIBOR3MD_.HSTA (visited on 04/13/2025).

– (Nov. 2018). *ECB Guide to the internal capital adequacy assessment process (ICAAP)*. URL: https://www.bankingsupervision.europa.eu/press/supervisory-newsletters/newsletter/2019/html/ssm.nl190213_3.en.html (visited on 04/13/2025).

Ferrari, Silvia and Francisco Cribari-Neto (2004). *Beta regression for modelling rates and proportions*. URL: https://www.ime.usp.br/~sferrari/beta.pdf (visited on 04/13/2025).

Hamilton, James D. (1994). *Time series analysis*.

Koenker, Roger (1981). *A note on studentizing a test for heteroscedasticity*.

Liang, K Y and S L Zeger (May 1993). *Regression Analysis for Correlated Data*. URL: https://www.annualreviews.org/content/journals/10.1146/annurev.pu.14.050193.000355 (visited on 04/13/2025).

Malikkidou, Despo and Wolfgang Strohbach (Jan. 21, 2025). *Predicting bank distress in Europe using machine learning and novel difinition of distress*. URL: https://www.eba.europa.eu/sites/default/files/2025-01/09a91f9a-ee49-4b83-b995-36ff736926d8/Staff%20Papers%20-%20Predicting%20bank%20distress%20in%20Europe.pdf (visited on 04/13/2025).

Rdocumentation (n.d.). *auto.arima: Fit best ARIMA model to univariate time series*. URL: https://www.rdocumentation.org/packages/forecast/versions/8.23.0/topics/auto.arima (visited on 04/13/2025).

Riigi teataja (Jan. 17, 2025). *Finantsinspektsiooni seadus*. URL: https://www.riigiteataja.ee/akt/12898417?leiaKehtiv= (visited on 04/13/2025).

Staehr, Karsten and Lenno Uusküla (Sept. 2017). *Forecasting models for non-performing loans in the EU countries*. URL: https://www.eestipank.ee/en/publications/working-papers/2017/102017-karsten-staehr-lenno-uuskula-forecasting-models-non-performing-loans-eu-countries (visited on 04/13/2025).

Statistikaamet (n.d.[a]). *Statistika andmebaas*. (Visited on 04/13/2025).

– (n.d.[b]). *Statistika andmebaasi API*. URL: https://stat.ee/et/statistikaamet/tarkvaraarendajale (visited on 04/13/2025).

World Bank Group (n.d.). *Commodity Markets*. URL: https://www.worldbank.org/en/research/commodity-markets (visited on 04/13/2025).

Appendix 1. Data sources

Table 6: Macroeconomic historic data sources

Variable	Citation	Table
GDP	(Statistikaamet, n.d.[a])	RAA0012
CPI	(Statistikaamet, n.d.[a])	IA02
Unemployment rate	(Statistikaamet, n.d.[a])	RAL0012
Salary	(Statistikaamet, n.d.[a])	PA92
Euribor	(European Central Bank, n.d.)	-
EUR\USD rate	(Eesti Pank, n.d.[b])	-
Oil price	(World Bank Group, n.d.)	-

Table 7: Default rate data sources

Default rates	Citation	Row name for dept over 90 days
Home loan	(Eesti Pank, n.d.[a])	Housing loans
Consumer loan	(Eesti Pank, n.d.[a])	Consumer credits
Business loan	(Eesti Pank, n.d.[a])	Non-financial corporations

Historical and future datasets are seen in the files below. Datasets are given as files due to the high number of rows and columns:

Historical dataset (file)

Forecasting dataset (file)

Appendix 2. Time series models forecasts

Table 8: Forecasted default rate values

Year	Month	Home loans	Consumer loans	Business loans
2025	1	0.0987 %	1.4918 %	0.5012 %
2025	2	0.0965 %	1.5558 %	0.5011 %
2025	3	0.096 %	1.5579 %	0.5007 %
2025	4	0.093 %	1.577 %	0.5005 %
2025	5	0.1003 %	1.6365 %	0.5003 %
2025	6	0.0911 %	1.6356 %	0.4999 %
2025	7	0.0987 %	1.6279 %	0.5001 %
2025	8	0.0989 %	1.598 %	0.5004 %
2025	9	0.0988 %	1.608 %	0.5004 %
2025	10	0.0988 %	1.6136 %	0.5001 %
2025	11	0.099 %	1.6427 %	0.4997 %
2025	12	0.0988 %	1.6269 %	0.4992 %
2026	1	0.0989 %	1.64 %	0.499 %
2026	2	0.0991 %	1.6439 %	0.4989 %
2026	3	0.0989 %	1.6655 %	0.4986 %
2026	4	0.099 %	1.6818 %	0.4984 %
2026	5	0.0991 %	1.6777 %	0.4981 %
2026	6	0.0989 %	1.678 %	0.4977 %
2026	7	0.0986 %	1.7219 %	0.498 %
2026	8	0.0987 %	1.7556 %	0.4979 %
2026	9	0.0982 %	1.7791 %	0.4977 %

Year	Month	Home loans	Consumer loans	Business loans
2026	10	0.0979 %	1.8028 %	0.4975 %
2026	11	0.098 %	1.8144 %	0.4974 %
2026	12	0.0975 %	1.8245 %	0.4972 %
2027	1	0.0973 %	1.8489 %	0.4971 %
2027	2	0.0973 %	1.8593 %	0.497 %
2027	3	0.0968 %	1.8683 %	0.4969 %
2027	4	0.966 %	1.89 %	0.4968 %
2027	5	0.0966 %	1.9007 %	0.4972 %
2027	6	0.0961 %	1.9108 %	0.4966 %
2027	7	0.0959 %	1.934 %	0.4965 %
2027	8	0.0959 %	1.9469 %	0.4964 %
2027	9	0.0953 %	1.9574 %	0.4963 %
2027	10	0.095 %	1.9823 %	0.4962 %
2027	11	0.095 %	1.9967 %	0.4961 %
2027	12	0.0945 %	2.0097 %	0.4959 %

Non-exclusive licence to reproduce thesis and make thesis public

I, Stein-Marten Pool,

1. grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis Forecasting loan default rates using macroeconomic variables, supervised by Meelis Kääärik;

2. grant the University of Tartu a permit to make the work specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in point 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Stein-Marten Pool
19/05/2025