Marte Aaberge
Jananni Johanraj

Improving Credit Management Practices: A Transdisciplinary Approach to Optimizing Risk and Profitability

Master's thesis in Cybernetics and Robotics
Supervisor: Adil Rasheed
Co-supervisor: Karl Johan Haarberg

June 2023

Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

NTNU
Norwegian University of
Science and Technology

Marte Aaberge
Jananni Johanraj

Improving Credit Management Practices: A Transdisciplinary Approach to Optimizing Risk and Profitability

Master's thesis in Cybernetics and Robotics
Supervisor: Adil Rasheed
Co-supervisor: Karl Johan Haarberg
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

NTNU
Norwegian University of
Science and Technology

Improving Credit Management Practices: A Transdisciplinary Approach to Optimizing Risk and Profitability

Jananni Johanraj and Marte Aaberge

Norwegian University of Science and Technology

Department of Engineering Cybernetics — NTNU 2023

Abstract

Effective credit management requires a delicate balance between maximizing profitability and minimizing risk. However, current credit management practices often fail to account for dynamic changes as well as interactions between internal and external factors, resulting in a lack of robust, actionable insights, sudden bankruptcies, and suboptimal decision-making. To address these challenges, this thesis proposes an enhanced credit management approach using systems theory by leveraging automation and dynamic forecasting of customer lifetime value (CLV). To achieve this, we conduct a comprehensive survey to map state-of-the-art credit practices in Norway, identify their strengths and challenges, and provide solutions to face these challenges.

Our research reveals that one of the most daunting aspects of credit management across Norwegian industries is accurately classifying customers and calculating their lifetime value. Additionally, our credit survey highlights a low degree of credit and analytical maturity among participants and industries, with the exception of bank and finance. This underscores the need for more interdisciplinary collaboration. By incorporating CLV into decision-making processes, financial institutions can optimize their credit risk management strategies and improve overall risk-adjusted profitability.

To substantiate our claims, we present empirical work on calculating CLV using statistical and machine-learning methods applied to time-series data from a Norwegian company. Introducing these methods and automation in the credit management process can contribute to improving the accuracy of credit assessment, thereby enhancing profitability and reducing risk. It is important to note that our approach represents one of many potential methods and serves as an initial step toward improving the credit management process.

As a future direction, we propose identifying and incentivizing the collection of easily accessible, relevant, enriched, and reliable data of high quality, in addition to incorporating more advanced methods, including multivariate panel data. Additionally, constructing models that enable companies to automate both credit limits and customer segmentation, as well as establishing early warning systems, would contribute to further advancements in the credit management process. By following these paths, we can continue to improve credit management, ultimately benefiting financial institutions and the overall economy.

Sammendrag

Effektiv kredittstyring krever en delikat balanse mellom å maksimere lønnsomhet og minimere risiko. Imidlertid mangler dagens praksis innen kredittstyring ofte evnen til å håndtere dynamiske endringer og samspillet mellom interne og eksterne faktorer, noe som resulterer i manglende robuste og handlingsrettede innsikter, plutselige konkurs og suboptimale beslutninger. For å håndtere disse utfordringene, foreslår denne oppgaven en forbedret tilnærming til kredittstyring ved å bruke systemteori og utnytte automatisering og dynamisk prognostisering av kundens livstidsverdi (CLV). For å oppnå dette gjennomfører vi en undersøkelse for å kartlegge kredittpraksis i Norge, identifisere styrker og utfordringer, og presentere løsninger for å håndtere disse utfordringene.

Undersøkelsen avdekker at en av de mest utfordrende aspektene ved kredittstyring i norske industrier er nøyaktig klassifisering av kunder og beregning av deres livstidsverdi. Vår kredittundersøkelse avslører også en lav grad av kreditt- og analytisk modenhet blant selskapene, med unntak av bank og finans. Dette understreker behovet for mer tverrfaglig samarbeid. Ved å inkorporere CLV i beslutningsprosesser, kan finansinstitusjoner optimalisere sine strategier for kredittrisikostyring og forbedre den overordnede risikojusterte lønnsomheten.

For å underbygge våre påstander, presenterer vi empirisk arbeid om beregning av CLV ved hjelp av statistiske og maskinlæringsmetoder basert på tidsseriedata fra et norsk selskap. Innføring av disse metodene og automatisering i kredittstyringsprosessen kan bidra til å forbedre nøyaktigheten i kredittvurderinger, dermed øke lønnsomheten og redusere risikoen. Det er viktig å merke seg at vår tilnærming representerer en av mange mulige metoder og fungerer som et første skritt mot forbedring av kredittstyringsprosessen.

Som videre arbeid foreslår vi en innsamling av lett tilgjengelige, relevante, berikede og pålitelige data av høy kvalitet, samt å inkorporere mer avanserte metoder, inkludert multivariate paneldata. I tillegg vil utvikling av modeller som muliggjør automatisering av både kredittgrenser og kundesegmentering, samt etablering av tidlige advarselssystemer, bidra til videre fremskritt innen kredittstyringsprosessen. Disse elementene kan bidra til å forbedre kredittstyringen, til syvende og sist til fordel for finansinstitusjoner og den overordnede økonomien.

Preface

This thesis is the result of our collaborative work with Business Intelligence Partner over the last six months. It concludes our five-year integrated Master of Science at the Department of Engineering Cybernetics at NTNU. Our objective has been to develop actionable credit management solutions that can drive meaningful improvements in the industry. This journey has been challenging, but has provided us with invaluable lessons and experiences.

We want to thank our supervisor, Professor Adil Rasheed, for his guidance and expertise. We also would like to thank Karl Johan Haarberg, PhD., for his valuable insights and constructive feedback, which has contributed significantly to the quality and depth of our research. Furthermore, we would like to thank our collaboration partner, Gunnar Haugen from Business Intelligence Partner, for excellent feedback and support, and for connecting us with industry partners. We would also extend appreciation to all participants in our survey, as well as valuable inputs from industry experts Glenn Gurrik and Kjell Strønen.

This thesis marks a milestone as the first Master's thesis in the field of Societal Cybernetics from the Department of Cybernetics and Robotics at NTNU. We are proud to be a part of this emerging field and hope that our work will contribute to the expanding field of knowledge and pave the way for future research in this area.

Finally, we thank our families and friends for their continuous support and encouragement.

Trondheim, June 2023

Jananni Johanraj and Marte Aaberge

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

Abbreviations

ADF	Augmented Dickey-Fuller
AFT	Accelerated Failure Time
AIC	Akaike Information Criterion
AML	Anti-Money Laundring
APT	Arbitrage Pricing Theory
AR	Autoregression
ARIMA	Autoregressive Integrated Moving Average
CAPM	Capital Asset Pricing Model
CCAPM	Consumption-based Capital Asset Pricing Model
CLV	Customer Lifetime Value
ECM	Error-Correction Model
GDP	Gross Domestic Product
GDPR	Global Data Protection Regulation
HP	Hodrick-Prescott
ICAPM	International Capital Asset Pricing Model
IQR	Interquartile Range
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
KYC	Know Your Customer
MA	Moving Average
MSE	Mean Squared Error
PP	Phillips-Perron
RADR	Risk-Adjusted Discount Rate
RMSE	Root Mean Squared Error
SML	Security Market Line
VaR	Value at Risk
VAR	Vector Autoregression
WACC	Weighted Average Cost Capital

1. Introduction

This thesis addresses the challenges in credit management, emphasizing the need for robust insights and decision-making that consider dynamic changes and interactions between the company, its customers, and external factors. The proposed approach leverages automation and dynamic forecasting of customer lifetime value (CLV) to optimize credit risk management strategies and improve profitability. Empirical work and survey findings highlight the importance of accurate customer classification and the lack of credit and analytical maturity, providing a foundation for future advancements in data collection, advanced methods, and automation in credit management.

1.1. Background and Motivation

Credit management plays a critical role in balancing profitability and risk within organizations. However, the occurrence of, e.g., payment delays, debt collection cases, and bankruptcies, despite assumed solid credit assessments, exposes inherent weaknesses in the existing credit management system and underscores the need for systematic improvements. In today's fast-paced business environment, making informed decisions based on actionable customer insights and trends is crucial. Unfortunately, many credit management practices are of static nature and lack adequate and high-quality data and thereby fail to account for the dynamic internal and external changes that constantly impact businesses. These dynamic elements directly influence a company's credit risk, necessitating the incorporation of feedback loops into the credit management process from a control system perspective.

Conversations with industry experts have revealed that decision-making in credit management often lacks a solid foundation based on reliable internal and external data and is influenced by a large degree of subjectivity. This thesis is motivated by these challenges and aims to enhance the predictability and profitability of credit management using automated decision support systems. Specifically, we propose dynamic forecasting of CLV using time series analysis as a method to establish the discounted, risk-adjusted profitability during a customer's lifetime with a company. By addressing the current needs and challenges as well as advancing best practice credit management, this thesis seeks to provide valuable and actionable insights for financial institutions. CLV, as a comprehensive metric encompassing forecast discounted sales and profitability, customer lifetime, and the associated discount factor (including, e.g., risk-free rate, liquidity premium, credit risk, customer churn, and shortfall risk), serves as a vital indicator that can inform strategic decisions, help to automize credit limit calculations, and improve resource allocation. Integrating CLV into credit risk management strategies will enable organizations to make more informed choices and optimize their credit management processes.

This thesis draws inspiration from the emerging field of societal cybernetics, which aims to describe, analyze, model, and simulate world dynamics using a holistic, transdisciplinary, and metadisciplinary approach. By fusing cybernetics with social sciences (including, e.g., neuroscience, psychology, sociology, statistics, economics, and econometrics), societal cybernetics seeks to construct digital twins of businesses and their underlying activities and processes, thus facilitating enhanced decision-making capabilities. This approach enables a deeper understanding of the interactive and dynamic interplay between internal and external factors that shape business operations. However, given the intricate relationships between customers, businesses, and external factors, addressing specific areas within this broad context, while maintaining a holistic perspective, poses a significant challenge.

In this thesis, we have delved into credit management, which in most industries, apart from the banking and finance industry, is a support business process that provides actionable insights with respect to overarching processes (i.e., processes tied to owners, the board of directors, and leaders), several parts of the core business (e.g., sales processes) and other support processes (e.g., accounting, and business administration). For the banking and finance industry, credit management is an important part of the core business processes. Our approach enables us to systematically investigate the mutual interactions between businesses, their customers, and external factors.

1.2. Contribution, Research Objectives and Research Questions

To the best of our knowledge, no previous study has comprehensively mapped the current state of credit management in Norway, identified associated challenges, and developed a dynamic credit risk system that integrates external and internal factors. Therefore, this study fills an important gap in the existing literature and provides valuable insights that have not been explored before.

1.2.1. Research Objectives

To narrow down the scope of the study, we have established a set of research objectives and formulated research questions that align with these objectives.

Primary objective: Address the current state of credit and analytical maturity across Norwegian industries and present methods, models, and techniques to improve this.

Secondary objectives:

- Conduct and present a survey to inform the reader of the current state of credit and analytical maturity across Norwegian industries.
- Present various statistical methods to improve credit and analytical maturity.
- Map the state of the art and shortcomings and present solutions for these.

1.2.2. Research Questions

- Are there significant differences between the industries, and is the bank and finance industry the benchmark for credit management?
- What methods can be utilized to maximize profitability and minimize risk in credit management?
- Which parts of the credit value chain should be optimized to improve credit management?

1.3. Structure of the Thesis

The thesis is structured as follows: Chapter 2 provides an introduction to credit management, mestrics for customer value measurement (including CLV), and statistical methodologies. Chapter 3 presents and discusses the results of a quantitative survey conducted across various companies in Norway. Based on these findings, Chapter 4 conducts an empirical analysis by employing machine learning algorithms, like ARIMA, linear regression, random forest, and XGBoost, and statistical models, like survival analysis, to forecast CLV using real customer credit data, addressing the identified challenges. The models are then evaluated, the major findings are critically examined, and their implications are discussed. The thesis is concluded in Chapter 5, by highlighting the progress, summarizing the contributions, and suggesting potential directions for future research in this evolving field.

2. Theory

2.1. Credit Risk

Credit risk refers to the potential loss creditors face when debtors fail to meet their obligations [1], including both the unexpectedness and uncertainty of extending credit to a customer [2]. Managing this risk is crucial for businesses that extend credit to customers, as it involves consistent decision-making related to customer targeting, resource allocation, and credit policies, all aimed at enhancing future profitability [3]. Effective credit risk management requires lenders to identify, measure, and mitigate these risks to make informed choices about which customers to maintain, nurture, or discontinue extending credit towards. Accurately estimating expected revenue and potential losses associated with extending credit is essential in this process. While various risk factors may come into play, the primary consideration is the probability that a customer will be unable to repay the credit extended to them.

Credit risk assessment approaches have evolved over time, with Altman's Z-score [4] being an early quantitative model. Structural models, pioneered by Black and Scholes [5] and expanded by Merton [6], focus on the relationship between asset value and debt, while reduced-form models estimate default risk without assuming the cause of credit risk premium. These models have contributed to the development of credit risk assessment, each offering distinct perspectives and methodologies [7].

2.1.1. The Credit Value Chain

Companies that extend credit often have a credit value chain that is divided into sub-processes. In short, the credit-control system involves the initial assessment of potential debtors, acceptance or rejection decisions, establishing credit limits, monitoring and follow-up procedures, and recovering losses. Monitoring credit risk can occur at both client and portfolio levels. Expected default rates are used in credit risk measurement, incorporating the behavior of individual credit portfolio components. Confer Figure 2.1.

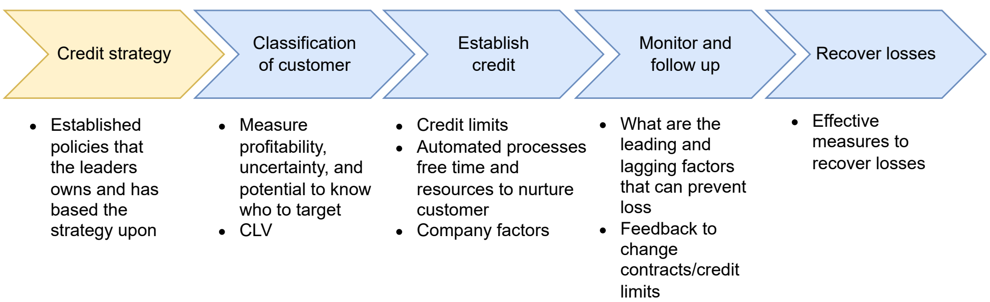


Figure 2.1.: The credit value chain with listed key components.

2.1.2. Various Measures of Credit Risk

Credit risk assessment and management involve employing diverse measures to evaluate the probability of borrower default and mitigate potential financial loss. These measures quantify credit risk by assigning values to factors such as borrower history, payment patterns, loan size, market conditions, and interest rate fluctuations, as well as the actual and anticipated business cycle. Key measures include the probability of default (PD), loss given default (LGD), exposure at default (EAD), and other risk measures (including structural models, reduced-form models, expected shortfall (ES), conditional expectile (CE), and value at risk (VaR) which estimates potential losses over a specific period [8]). In addition to risk mitigation, maximizing profitability is a crucial objective. CLV can serve as a valuable metric in credit risk management, where the estimated potential loss and profitability of the customer are accounted for.

2.2. Customer Lifetime Value (CLV)

Customer lifetime value (CLV) has primarily been utilized in marketing strategy and communication as a metric for customer segmentation, guiding target selection, resource allocation, and nurturing efforts [1]. By treating the customer as an asset and adopting an investment-oriented approach, CLV enables decision-making based on expected risk-adjusted returns.

By incorporating credit risk into CLV, it can serve as a fundamental metric in the financial service industry, quantifying the projected, net, risk-adjusted profit derived from a customer throughout their entire relationship with the credit extender. It provides a comprehensive perspective on the customer's potential revenue contribution, taking into account factors such as future customer activity, contribution margin, and customer acquisition costs. By evaluating the long-term value of customers, CLV facilitates resource allocation decisions and strategic planning within credit extension institutions with the aim of identifying and invest in profitable customers [9].

The mathematical representation of the CLV for all of a company's customers in the context of a credit extender can be observed in Equation 2.1,

$$CLV_{N_t,t} = \sum_{n=1}^{E_t(N_t)} \sum_{t=0}^{E_t(T_{i,t})} \frac{E_t\left[CF_{i,t}\right] - E_{t|default}\left[r \cdot D_{i,t}\right]}{(1+k_{i,t})^t} \tag{2.1}$$

where

- $E_t(N_t)$: The expected number of customers at time t.
- $E_t(T_{i,t})$: The expected future duration T of customer i with the company from time t.
- $E_t\left[CF_{i,t}\right]$: The expected net cash flow for customer i at time t, which may be decomposed into expected sales times expected profitability.
- $E_{t|d}\left[r \cdot D_{i,t}\right]$: The net exposure with respect to customer i at time t given customer default.
- $(1+k_{i,t})^t$: The discount factor k for customer i at time t which is used to calculate the present value of future cash flows.

- $k_{i,t}$: The risk-adjusted discount rate, which represents the expected required rate of return for customer i at time t.

This implies that the mathematical representation of the CLV for one customer in the context of a credit extender can be observed in Equation 2.2.

$$CLV_{i,t} = \sum_{t=0}^{E_t(T_{i,t})} \frac{E_t\left[CF_{i,t}\right] - E_{t|default}\left[r \cdot D_{i,t}\right]}{(1+k_{i,t})^t} \tag{2.2}$$

Drivers of CLV

Customer loyalty has been a focus area for companies to develop customer relationships to increase revenue. However, loyalty is not a direct measure of profitability [10]. It is of interest to investigate which factors that drive profitable customers in order to know which customers to focus on. The drivers are classified into exchange characteristics and customer heterogeneity, where the first are variables that define and describe relationship activities between the customer and the firm. These variables vary for different industries and include factors such as customer spending level, cross-buying behavior, focused buying, average inter-purchase time, and customer returns. Customer heterogeneity includes demographics, industry, and annual revenue. When considering the determinants of CLV in the context of credit, specific drivers can be identified as leading or lagging. Leading indicators will be customer satisfaction, customer sentiment, order frequency, and order value, whereas lagging indicators are credit score, invoice reminders, and additional notes or comments that provide insights into the customer's creditworthiness and risk profile. These factors play a significant role in assessing the potential risk-adjusted profitability of customers.

Use of CLV

The application of CLV can have a significant impact across various stages of the credit value chain, including customer selection, customer classification, credit establishment, monitoring and follow-up processes, as well as debt collection. It can also serve as a central part of the strategy which the policy is built upon. All operational decisions should be fundamental in the strategic plan of an organization. By e.g., incorporating CLV into customer classification strategies, companies can effectively identify and differentiate the most profitable customers. This enables them to set higher credit limits for its most lucrative customers while also allocating resources and directing efforts toward providing enhanced customer service and support.

Moreover, by gaining a comprehensive understanding of a customer's CLV, companies can optimize their credit limits and minimize potential risks. By identifying segments or customers as high-risk or unprofitable, companies can make informed decisions regarding credit extension.

Furthermore, CLV serves as a valuable guide in customer retention and development strategies. By identifying and nurturing customers with higher CLV, companies can prioritize building long-term relationships and thus enhance loyalty.

2.3. Methods and Models to Calculate CLV

Having established the components of CLV in a credit risk context (as described in Equation 2.1), the subsequent sections will delve into the application of statistical techniques in each of the components to forecast CLV.

2.3.1. Forecasting

Forecasting concerns predicting the future accurately, using available information such as historical data, knowledge of potential drivers, and potential influencing events. Forecasting is important in order to perform effective and efficient planning, and can be applied to many areas. The predictability of an event or a quantity depends on several factors, for instance, how well we understand the factors that contribute to it and how much high-quality data is available [11]. Whether qualitative forecasting or quantitative forecasting is appropriate to use depends on what data is available. In the context of credit, quantitative forecasting is preferred due to the availability of numerical data from the past and the reasonable assumption that past patterns will continue in the future. Various quantitative forecasting methods exist, including those based on cross-sectional data, panel data, and especially time series data, which will be the focus of this thesis. A time series represents a sequence of observations measured sequentially over time, either continuously or at discrete time intervals [12].

Time series

A time series model is typically on the form

$$X_t = f(X_{t-1}, X_{t-2}, X_{t-3}, .., error) \tag{2.3}$$

where t is the present time. The error term allows for random variation or noise.

Predictor variables are essential when it comes to time series forecasting. They are used as explanatory factors in models, and help explain what causes the forecasting variable [11], and are typically on the form:

$$X = f(variable_1, variable_2, variable_3, ..., error). \tag{2.4}$$

Dynamic regression models can integrate both time series and explanatory/mixed models. However, in terms of forecast accuracy, time series models often outperform other approaches [13]. A forecasting interval is often provided, representing the range of values the forecast variable can assume with a relatively high probability [11]. Once a forecast is generated, it is customary to estimate the central value within the range of possible values in the forecasting interval.

A common method used to analyze a time series, is time series decomposition. A time series can be represented as the sum of a deterministic trend component, a cyclical component, a periodic seasonality component and a stochastic irregularity component as introduced in [14]. Decomposition is performed by splitting the given time series Y_t, such that

$$Y_t = f(T_t, C_t, S_t, E_t), \tag{2.5}$$

where for a given t, T_t denotes the trend, C_t denotes the cycle, S_t denotes the seasonal components, and E_t corresponds to the error that results from the decomposition [15]. There are two common methods for time series decomposition: additive decomposition and multiplicative decomposition. Additive decomposition is useful when the trend component changes linearly over time, while the seasonality and error components are consistent over time [16]. The decomposition function for additive decomposition is:

$$f(T_t, S_t, E_t) := T_t + C_t + S_t + E_t. \tag{2.6}$$

Multiplicative decomposition is more useful when the trend changes non-linearly over time, and the seasonality and error components vary in proportion to the trend level [16]. The function for multiplicative decomposition is:

$$f(T_t, S_t, E_t) := T_t \times C_t \times S_t \times E_t. \tag{2.7}$$

Stationarity

Stationarity is a fundamental assumption in many time series models and statistical methods due to its importance in achieving reliable results [17]. Although several techniques exist to model non-stationary time series data, working with stationary series simplifies the modeling process and enhances result accuracy.

A time series is considered strictly stationary when the joint distribution of $X_{t_1}, ..., X_{t_n}$ is equal to the joint distribution of $X_{t_1+\tau}, ..., X_{t_n+\tau}$ for all $t_1, ..., t_n$ and τ, where X is the set of data points and τ is a time lag. This implies that the distribution of the stationary process remains unchanged when shifted by τ. In other words, a stationary time series has a time-invariant autocovariance, which is approximately equivalent to saying that a time series is stationary if it has a constant mean and variance over time.

Stationarity can can also be observed in an autocorrelation plot. The presence of autocorrelation in the time series can be identified with the Autocorrelation function (ACF). Autocorrelation can be extended to a time series Y_t, by it comparing a previous lagged value Y_{t-1}, i.e., a value that occurred at a specific time interval before the current observation. Thereby, the autocorrelation at lag k can be defined as [18]:

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}. \tag{2.8}$$

The complete set of autocorrelations is called the Autocorrelation Function.

Unit root

A unit root test investigates whether a time series variable is non-stationary and possesses a unit root. Differencing can help to achieve stationarity [19]. For instance, if the prices of a certain stock exhibit a unit root, using the first differences of these prices (i.e., the changes from one period to the next) instead of the original prices themselves may result in a stationary series. A variable is considered integrated of order d, denoted as $I(d)$, if it needs to be differenced d times to achieve stationarity, while a stationary variable is integrated of order 0, i.e., $I(0)$. If a non-stationary variable has to be differenced once to stationary, we say that the variable in question is integrated of order 1, i.e., $I(1)$ [20].

Consider the following model:

$$y_t = y_{t-1} + \varepsilon_t, \tag{2.9}$$

which represents a random walk model provided the residual is white noise (i.e., normally, independent, and identically distributed). Here, the value at time t equals the value at the previous time period plus a random error term, denoted as ε_t. In this case, ε_t is stationary ($I(0)$), while y is non-stationary ($I(1)$) as the first difference, $\Delta y_t = y_t - y_{t-1} = \varepsilon_t$.

This relationship can be generalized as:

$$y_t = \alpha y_{t-1} + \varepsilon_t, \tag{2.10}$$

If $|\alpha| < 1$, the variable y is stationary (i.e., $I(0)$). Conversely, if $\alpha = 1$, y is non-stationary (i.e., $I(1)$). Hence, formal tests for stationarity are conducted under the assumption of $\alpha = 1$ and are referred to as tests for a unit root [20].

A unit root implies that cointegration or (integer or fractional) differencing can transform a time series into a stationary series. A fractional unit root represents a form of non-stationarity characterized by a slowly decaying long-memory component [21]. Techniques such as fractional differencing may be necessary to address non-stationarity in such series but are outside the scope of this thesis. [22].

Several statistical tests are available to determine the presence of a unit root in a time series, including the Augmented Dickey-Fuller (ADF) test, the Phillips-Perron (PP) test, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, further explained in Appendices B.2.1, B.2.2 and B.2.3, respectively. These tests examine the autocorrelation or variance behavior of the time series to assess the presence of a unit root or stationarity.

Cointegration

If the data is found to be non-stationary by a unit root test, differencing and estimating using only differentiated variables, as proposed by Box and Jenkins [19], may be an incomplete approach as it may lead to the loss of valuable information from economic theory related to long-run equilibrium [20]. An alternative approach is to use an error-correction model (ECM), which combines data in different levels and differentiates them in the same equation. However, this may lead to spurious results. In some cases, the linear combination of $I(1)$ variables is indeed $I(0)$, resulting in cointegrated variables. Cointegrated variables can provide a framework for testing and estimating long-run relationships among economic variables, such as interest rates, prices and wages, imports, and exports, etc. The error correction term represents cointegration, and therefore, all cointegrated variables have an ECM representation.

To eliminate a unit root, one must first test for cointegration, and if it is found, ECM estimation can be performed. By regressing a set of $I(1)$ variables onto the others, the residual produced is $I(0)$. However, using a single equation to represent the system assumes that the variables are exogenous, which may not always be true. Additionally, if there are three or more cointegrating relationships in the equation being estimated, using the method described may not be appropriate because there is no obvious way to identify the cointegrating relationships. In such cases, a more appropriate method is to adopt more general formulations, where each variable is modeled in terms of the lagged values of all the other variables, resulting in a vector autoregressive model (VAR).

There are several methods to test for cointegration relationships. The ADF test is limited to testing two time series at a time, while the Johansen test (B.2.4) can be used to test up to 12 time series. In the Johansen test, we check whether lambda has a zero eigenvalue. When all the eigenvalues are zero, that would mean that the series are not cointegrated, whereas when some of the eigenvalues contain negative values, it will imply that a linear combination of the time series can be created, which would result in stationarity.

ARIMA (p,d,q) models

When constructing forecasting models, it is often more favorable to capture the nuances of a complex reality rather than assuming the fitted model represents the absolute truth. There are various approaches available to make these approximations, including autoregressive (AR) models, moving average (MA) models, and autoregressive integrated moving average (ARIMA) models.

Autoregressive models of order p, AR(p) : In AR models, the variable of interest is forecast by employing a linear combination of its past values [11]. These models are characterized by the number of time lags, denoted as p, which are included in the autoregression. A general p-th order autoregressive process, denoted as AR(p), can be expressed as follows:

$$Y_t = c + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + ... + \Phi_p Y_{t-p} + \varepsilon_t, \tag{2.11}$$

where ε_t is normally distributed with mean 0 and variance 1, i.e., white noise [23]. c is a constant and Φ_j are parameters to be determined. However, there are constraints on the allowable values of Φ_j [18]:

- For p=1, -1 $< \Phi_1 < 1$
- For p=2, -1 $< \Phi_2 < 1$, $\Phi_2 + \Phi_1 < 1$ and $\Phi_2 - \Phi_1 < 1$
- For $p \geq 3$, more complicated conditions hold.

These constraints ensure that the autoregressive model remains stable and avoids issues such as explosive behavior or divergence in the forecasts.

Moving average models of order q, MA(q) : MA models utilize past forecast errors within a regression-like framework, as opposed to incorporating past values of the forecast variable in a regression [11]. A general finite-order moving average process of order q, denoted as MA(q), can be expressed as follows:

$$Y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}, \tag{2.12}$$

where ε_t is normally distributed with mean 0 and variance 1, i.e., white noise [23]. c is a constant and θ_j are parameters to be determined. As with AR models, there are restrictions on the allowable values of θ_j [18]:

- For q=1, -1 $< \theta_1 < 1$
- For q=2, -1 $< \theta_2 < 1$, $\theta_2 + \theta_1 < 1$ and $\theta_2 - \theta_1 < 1$
- For $q \geq 3$, more complex conditions hold.

These restrictions ensure the stability of the MA model and prevent issues such as explosive or divergent forecasts. By appropriately selecting the parameter values within the specified ranges, MA models can effectively capture the dependencies in the forecast errors, leading to accurate forecasts and valuable insights in forecasting tasks.

Autoregressive integrated moving average (ARIMA) models : Combining differencing with autoregression and moving average models yields the ARIMA models [11]. These models include an integration (I) component, which reverses the differencing process, in addition to the AR and MA components. ARIMA models are particularly effective for short-term forecasting, which aligns with our objective. The complete representation of an ARIMA(p, d, q) model is as follows:

$$y't = c + \Phi 1 y't - 1 + ... + \Phi p y't - p + \theta 1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{2.13}$$

Here, p denotes the order of the autoregressive part, d represents the degree of differencing applied, and q indicates the order of the moving average part. The series $y't$ corresponds to the differenced series, and the predictors encompass lagged values of yt as well as lagged errors. The same constraints applicable to AR and MA models are also applicable to ARIMA models to ensure model stability and valid forecasts.

2.3.2. Customer Duration

Customer duration, which refers to forecasting how long a customer will stay with a company, is a crucial aspect of credit risk management and can supply the classification of customers, as well as CLV calculation. Various methods, including survival analysis, can be employed to address this forecasting problem. Survival analysis is a statistical method used to model the probability of events occurring over time, such as customer default or customer attrition [24], [25].

Survival models offer several advantages over traditional classification and regression methods. They can handle censored data where the event of interest (e.g., customer default) has not yet occurred, and they allow for the inclusion of time-varying covariates [26]. This flexibility enables a more comprehensive analysis of customer duration compared to other approaches. In credit risk management, the probability of default is influenced by macroeconomic variables that vary over time, posing challenges for logistic regression models [27]. Survival models, on the other hand, can effectively incorporate such time-dependent covariates.

To forecast default probability, the duration from the start of the customer relationship until the end or default date is calculated, taking into account both censored and non-censored customers. Additionally, customer characteristics such as company size, credit score, historical payments, and market cycles may also influence customer duration. The goal is to estimate the expected future duration T of a customer i with the company from time t, denoted as $E_t(T_{i,t})$, and generate either a survival function or a hazard function.

The survival function expresses the probability of the event not occurring, i.e., in our context, equivalent to the customer staying with the company beyond time t. The survival function is defined as follows:

$$S(t) = Pr(T > t), \tag{2.14}$$

where T is a random lifetime and is a non-increasing function. This function can be estimated using non-parametric, semi-parametric, or parametric methods. The Kaplan-Meier estimator, a non-parametric method, is used to estimate the survival function from lifetime data, which is particularly useful in handling right-censored data. However, it doesn't accommodate covariates. In parametric methods, the Accelerated Failure Time (AFT) model posits that certain factors accelerate or decelerate the time to an event, assuming a specific statistical distribution for survival times. An example of this is the Weibull distribution model, which assumes survival times follow a Weibull distribution, allowing for constant, increasing, or decreasing failure rates. Despite the insights these models offer, the semi-parametric Cox proportional hazards (Cox-PH) model provides additional flexibility. This model can handle time-varying covariates and quantifies the effects of predictor variables as multiplicative factors applied to a baseline hazard function, making it particularly useful in credit risk management [28].

The Cox-PH model's hazard function represents the event probability at a given time and quantifies predictor variable effects. It doesn't assume a specific distribution, offering flexibility. This model's hazard function, denoted $h(t, X)$, is defined as:

$$h(t, X) = h_0(t) \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right), \tag{2.15}$$

where $X_i = (x_{i1}, \dots, x_{ip})$ represents the predictor variables and β_j the j-th predictor variable's coefficient. The baseline hazard rate function, $h_0(t)$, captures the hazard rate with all predictor variables at zero.

The Cox-PH model assumes proportional hazards, meaning the hazard ratio is time-independent. However, some macro variables, like interest rates and GDP, are time-dependent and may enhance credit risk assessment accuracy [29]. Considering such variables, along with explanatory variables, in modeling improves prediction accuracy.

Survival analysis assumes the independence and non-informativeness of different event survival times, ensuring unbiased estimation. The hazard function informs the current risk after a customer survives a high-risk period—a crucial aspect for effective credit risk management.

Previous literature has explored different methods for modeling customer duration, including estimation of survival functions [30], parametric models like AFT and the Weibull distribution, and the semi-parametric Cox-PH function [28]. These approaches have proven to be valuable tools in analyzing customer behavior and effectively managing credit risk.

Regularization techniques

Regularization techniques such as lasso (ℓ_1) and ridge (ℓ_2) prevent overfitting in statistical modeling and machine learning by adding a penalty term to the loss function. They address issues like multicollinearity, common in data sets with many features, thereby enhancing model stability.

Ridge regularization encourages smaller, evenly distributed coefficients, reducing the influence of individual features and mitigating the impact of multicollinearity [31]. The objective function for ridge regression is:

$$\min_{\beta} \left[\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right], \tag{2.16}$$

where y_i is the response variable, x_i is the predictor vector for the i-th observation, β is the coefficient vector, λ is the regularization parameter, and p is the number of predictors.

Lasso regularization induces coefficient sparsity, serving as a feature selector by setting some coefficients to zero. Its objective function is:

$$\min_{\beta} \left[\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \tag{2.17}$$

y_i, x_i, β, λ, and p carry the same meaning as in ridge regression.

Both methods control the coefficients' magnitude through the regularization parameter λ, allowing tailored trade-offs between fit and regularization.

2.3.3. Exposure

Exposure is a critical metric for evaluating counterparty credit risk and estimating potential losses in the event of default [1]. It quantifies the level of vulnerability that a company faces from counterparties failing to fulfill their obligations. There are two types of exposure measures: gross exposure and net exposure. Gross exposure, excluding collateral, is commonly used to assess the maximum potential loss.

Net exposure represents the total amount at risk of loss if a customer defaults on their credit obligations. It is typically calculated as the expected portfolio loss given default. While credit losses are expected to occur occasionally, the goal is to minimize this expected loss. The most commonly used equation to calculate expected net exposure is:

$$\text{Expected loss} = \text{PD} \times \text{LGD} \times \text{EAD}, \tag{2.18}$$

where PD denotes the probability of default, EAD represents exposure at default, and LGD denotes the loss given default. This equation provides a numerical estimation of the predicted net exposure. The probability of default (PD) measures the likelihood that a borrower will be unable or unwilling to repay their debt fully. It is an estimate of the probability of default within a specified time horizon and is closely related to the borrower's credit score.

Loss given default (LGD) refers to the proportion of the total exposure that cannot be recovered by the lender following a borrower's default. It represents the share of an asset that is lost when default occurs. Measures such as mortgage security, credit insurance, and bail can help reduce this proportion. Exposure at default (EAD) is the total value to which a lender is exposed when a borrower defaults. For fixed exposures, it is equal to the current outstanding monetary amount. Therefore, it represents the maximum potential loss a creditor may face when a borrower defaults on a loan.

2.3.4. Risk

Risk factors can include the likelihood of a customer defaulting on their payments or declaring bankruptcy, and by incorporating an estimation of risk, financial institutions can better understand the overall value of a customer while accounting for potential losses. The risk-adjusted discount rate (RADR) is the denominator of the CLV equation, as expressed in Equation 2.1:

$$(1 + k_{i,t})^t,$$

representing the expected required rate of return for customer i at time t.

The required rate of return is used to transform a cash flow into value by discounting. More specifically, it is used to calculate the present value of future cash flow in order to evaluate the worth of the company. According to Gjesdal and Johnsen [32], the required return is defined as the capital market's expected return on alternative investments with the same risk as the company. This definition emphasizes four conditions: expected return, alternative investments, capital market, and same risk.

The expected return is based on expected values, i.e., the probability-weighted averages of optimistic and pessimistic estimates. As the required rate of return concerns a future, unknown return, the cash flow values must be expected values, when discounting future cash flows for a project, a business, or a customer. When measuring historical earnings, results for several periods should be used to smooth out random, annual variations.

Alternative investments refer to the fact that the required rate of return is an opportunity cost. Hence, the return must compensate the investors for what they could otherwise have earned from corresponding, risky investments. The required rate of return concerns all capital that requires a financial return, i.e., the employed capital. This concept does not include interest-free debt, whether it is free or has its return covered in operating results.

A capital market that is open, integrated, and efficient is required, as the required return requirement is based on risk equivalent market investments, which presupposes that the market really reflects the investor's alternative return (i.e., the opportunity cost). This will be indirectly reflected in an increase in the required return's risk, to the extent that illiquidity or irreversibility affects the variability of the cash flow.

The required return depends on the risk of the company, customer, or project, i.e., the possibility of deviation from the expected return to take greater risks. Investors are assumed to be risk averse and will thus require a higher expected return to take greater risks. Risk aversion also implies that the investors are well-diversified, i.e., that they have distributed their capital over several risky investments. This reduces the required return for individual placements, as the investors will only demand compensation for the risk that also characterizes other placements, i.e., market risk or cyclical risk.

Measuring Risk-Adjusted Discount Rate (RADR)

There are several ways to measure the risk-adjusted discount rate (RADR), including weighted average cost of capital (WACC), the Treynor-Sharpe-Mossin capital asset pricing model (CAPM), Lucas-Breeden consumption-based capital asset pricing model (CCAPM), Merton's international capital asset pricing model (ICAPM), and Ross arbitrage pricing theory (APT), e.g., confer [33].

Capital Asset Pricing Model (CAPM) The CAPM approach is one of the most common approaches to determining the risk-adjusted discount rate. In the capital asset pricing model, confer Treynor [34], Sharpe [35], Lintner [36], and Mossin [37], the linear relationship between the investors' minimum required return on investment, in stock market securities and business operations, and its systematic risk is represented by

$$E\left[r_i\right] = r_f + \beta_i \cdot (E\left[r_m\right] - r_f), \tag{2.19}$$

where $E\left[r_i\right]$ is the required return on financial asset i, r_f is the risk-free rate of return, β_i is the beta value for financial asset i, and $E\left[r_m\right]$ is the average return in the capital market.

However, the CAPM is often deemed unrealistic due to invalid model assumptions. According to Watson & Head [33], four assumptions are underlying the CAPM. First, all investors are assumed to hold well-diversified portfolios. This means that investors will only require a return for the systematic risk of their portfolios since unsystematic risk has been diversified and can be ignored. Not all investors are well-diversified.

Second, a single-period transaction period is assumed. This means that the CAPM assumes a standardized holding period, usually one year, to make the returns on foreign securities comparable. The use of derivatives, different investor horizons, different underlying cash flows, etc., implies that a single period limits the investors' investment universe to the extent that it becomes restrictive and impractical. However, even though many investors hold securities for longer than a year, returns on securities are usually quoted annually.

Third, it is assumed that investors can borrow and lend at a risk-free rate of return. This assumption stems from the mean-variance portfolio theory of Markowitz [38], from which the CAPM was developed. The risk-free rate of return corresponds to the intersection of the security market line (SML) and the ordinate axis. Hence, the SML is a graphical representation of the CAPM formula. Investors' different credit scores imply that everyone cannot borrow and lend at the same risk-free rate of return.

Fourth, it is assumed that the capital market is perfect. This means that all securities are valued correctly and that their returns will plot on the SML, which requires that there are no transaction costs, that perfect information is freely available to all investors, and that all investors are risk-averse, rational, and desire to maximize their own utility. In addition, many buyers and sellers in the market are required.

2.3.5. Accuracy Measures

In the following analyses and discussions, various statistical metrics and tests are utilized to assess and validate the accuracy of the forecasting models and to check the assumptions underlying these models. Below is a brief introduction to the key tests and measures, with more detailed explanations in Appendix B.

For hypothesis testing, we utilize a range of tests, including the Shapiro-Wilk test for normality, the Chi-Square test for categorical data, the Kolmogorov-Smirnov test and Jarque-Bera test for checking a distribution against a normal distribution, and the Mann-Whitney U test and Kruskal-Wallis test for comparing independent samples.

For time series analysis tests, we use the Johansen test to determine the cointegration rank in a time series system, which helps us understand the long-run relationships between the variables. The main measures used for evaluating the accuracy of sales and customer duration forecasts include Mean Squared Error (MSE), Root Mean Squared Error (RMSE) (B.2.5), and the R^2 statistic (B.2.6). MSE and RMSE measure the average magnitude of the forecast errors, with RMSE being especially useful in contexts where significant errors are undesirable. R-squared, on the other hand, indicates the goodness of fit of a model, i.e., the proportion of variance in the dependent variable that the independent variables can explain.

We also use the Akaike Information Criterion (AIC) (B.2.7) to compare different potential models for our data. The AIC considers both the goodness of fit and the complexity of the model, helping us avoid overfitting. Finally, we use the Durbin-Watson statistic (B.2.8) to validate our regression models' assumptions. In addition, this test helps us detect the presence of autocorrelation in the residuals from a regression analysis, which can indicate issues with the model.

These tools will be invaluable in our subsequent analyses, allowing us to assess the performance of our forecasting models and validate the assumptions underlying our statistical approach.

3. Survey

3.1. Introduction

The first part of this thesis aims to explore credit management strategies employed by companies to maximize profitability and to identify the challenges they encounter in the process. To achieve this objective, we conducted a comprehensive survey targeting companies from five distinct industries in Norway. The survey, with a total of 15 participating companies, was designed to yield valuable insights into current credit management practices and to assess the extent of analytical maturity, a key prerequisite for effective credit management.

The survey was conceived and designed as the culmination of three major preparatory steps. Initially, an extensive literature review was undertaken to study credit management in depth. This was followed by conducting interviews with several leading practitioners in the field of credit management, which offered a nuanced understanding of the credit management process and its various components in the credit value chain. Drawing on the insights gained from the literature review, the expert interviews, and our analytical interpretation of the gathered knowledge, we designed a survey tailored to investigate and evaluate credit management practices within the targeted industries in Norway.

The survey was divided into five distinct sections, each designed to evaluate credit maturity, analytical maturity, and to provide answers to the research questions outlined in Section 1.2.2. These sections are as follows:

- **Credit strategy**: The credit strategy of the business is normally laid down in the credit policy and forms the basis for how credit is handled in the company. The credit policy outlines goals, rules, and guidelines for handling credit, including e.g., roles, mandates, and incentives.

- **Classification and analyses of customers**: For the company to treat customers optimally, it is important for the business to have an overview of and be able to analyze, among other things, sales, profitability, risk, customer behavior, and sales potential for various customers, as well as the development of customer statuses (i.e., potential, new, existing active, existing passive, lost, and former customers), customer segments, customer categories (e.g., good customers, potentially good customers, bad customers, potentially bad customers, and other customers), and customer portfolios over time.

- **Establishing credit**: Establishing credit involves several different and normally manual and time-consuming processes. Hence, the more of its manual credit processes the business can automate, the more time and resources it will free up to address and improve customer service, and other important issues.

- **Monitoring and following up credit**: Monitoring and following up on approved credits is important. In this context, it may be useful to establish leading, coincident, and lagging credit indicators so that the company can establish early warning systems regarding credit risk, etc., e.g., to establish actionable insights, save resources, and prevent losses.

- **Recovering losses**: When the company has customers who default, it is important that the company has procedures, tools, and competencies that allow it to recover most of its losses as effectively as possible.

A detailed list of survey questions is provided in Appendix A.1.

3.2. Choice of Industries

The survey was completed by 15 participants who hold leadership positions within credit management across various industries. The inclusion of companies from different industries served multiple purposes. Firstly, it aimed to provide a comprehensive understanding of credit management practices in Norway. Secondly, it aimed to identify possible trends and patterns that may exist across industries. Lastly, it aimed to gain insights into how businesses of different sizes and in various industries handle credit management and the challenges they encounter.

The banking and finance industry was specifically included in the survey due to its core competency in credit management. By comparing this industry with others, we could identify potential differences in credit management processes. As a result, companies from other sectors that also work with credit were added to the list of potential participants. The survey participants identified themselves within the following industries:

- Manufacturing
- Electricity, gas, steam, and air conditioning supply
- Construction
- Wholesale and retail trade; repair of motor vehicles and motorcycles
- Transportation and storage
- Information and communication
- Financial and insurance activities
- Professional, scientific, and technical activities
- Administrative and support service activities

The participants labeled their companies within these industries, but in order to facilitate a more organized analysis, these industries were grouped into five generalized industries illustrated in Figure 3.1.

It's important to note that there is a difference in the number of companies within each industry. Given the relatively small number of overall respondents and the potential variation in the representativeness of respondents' answers across industries, it is crucial to acknowledge that the small and asymmetric samples may impact the interpretation and analysis of the survey results.

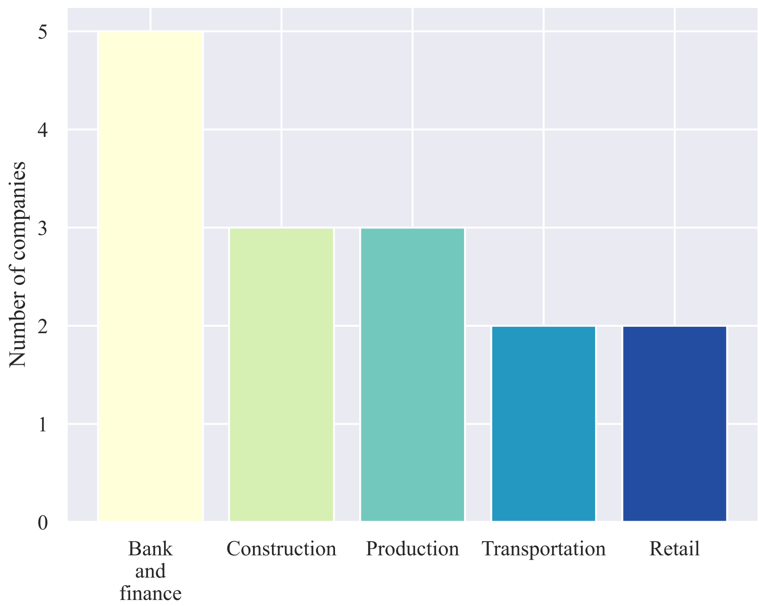


Figure 3.1.: The number of industries that participated in the survey and the distribution of companies within them. The distribution is not equal among all industries.

3.3. Methodology

In this section, we present the methodology employed for conducting the survey.

3.3.1. Preparation

To fulfill the research objectives outlined in Section 1.2.1, we conducted extensive research on credit management and developed a survey as described in Section 3.1.

We employed purposeful and convenience sampling methods to identify the ideal survey participants. Purposeful sampling involves selecting specific cases expected to provide rich and relevant information, guiding the selection of industries and companies within each industry [39]. Our goal was to include a diverse range of industries and companies of various sizes to minimize non-response bias. Individuals within these companies were identified through the network of Business Intelligence Partner or through online research, targeting knowledgeable individuals with sufficient expertise in credit management. Additionally, convenience sampling was used, with the final participants being those who accepted the invitation to contribute to the research. Convenience sampling selects participants based on their availability or accessibility to the researcher rather than through a random or systematic process [40].

To collect data, we utilized Effect IO, a survey tool designed to standardize data collection and exports [41]. This tool facilitated the efficient distribution of survey questionnaires and related information to the participants. The Effect IO platform, accessible via a hyperlink, provided an introduction to the project and survey, contact information, and the survey questionnaire divided into sections (see Appendix A.1). Participants were able to select options ranging from 0 to 10 for each question. The platform allowed for pretesting, monitoring of respondent activity, data retrieval, and follow-ups with non-respondents to improve response rates. The final data set was converted into a CSV file for further analysis.

3.3.2. Data Processing

Upon completion of data collection, the collected survey data were subjected to a series of processing steps to facilitate analysis and interpretation. First, we assessed the completeness of the data set and identified a single missing value. To address this missing value, we employed the k-nearest neighbors (k-NN) algorithm. The k-NN algorithm, a non-parametric supervised learning method [42], was used in a regression context, where the missing value was estimated as the average of its nearest neighboring values.

Following the completion of data imputation, the processed data set was prepared for analysis in accordance with the research objectives outlined in Section 1.2.1. To enable statistical data visualization, we utilized Seaborn, a Python data visualization library [43]. Descriptive statistics were applied to provide insights into the data features. It is important to acknowledge that many statistical methods rely on certain assumptions about the data. In cases where the underlying statistical distribution deviates from mesokurtic and exhibits skewness, such as platykurtic or leptokurtic, the use of the median as a measure of central tendency may be preferred over the arithmetic mean. As our data showed skewness, the median was used during our analysis. Moreover, normality tests were conducted to validate the appropriateness of statistical analyses and to prevent potential misinterpretations of the results. Additionally, statistical tests were employed to examine potential variations across industries and identify significant patterns or trends.

Normality Tests

Assessing the normality of data is essential to ensure the validity and reliability of statistical analyses, particularly for parametric procedures that assume a specific data distribution, such as the mesokurtic normal distribution. Deviating from the normality assumption can lead to unreliable or invalid results.

There are several methods available to assess normality, including graphical methods (e.g., histograms, box plots, and confidence limit plots), numerical methods, e.g., estimates of skewness and kurtosis (B.1.1), and formal normality tests. While numerical methods can provide some indication of normality, the Jarque-Bera (B.1.2) skewness test and formal normality tests, such as Shapiro-Wilk (B.1.3), Kolmogorov-Smirnov (B.1.4), and Anderson-Darling tests (B.1.5), offer more objective and reliable assessments.

The choice of a specific normality test depends on the data characteristics, such as sample size, skewness, and kurtosis. For small data sets like the survey results in this thesis, Razali and Wah [44] recommend the Shapiro-Wilk or Anderson-Darling tests due to their higher power in detecting departures from normality. The Shapiro-Wilk test examines whether a data set is drawn from a normal distribution, while the Anderson-Darling test measures the deviation of the sample data from a normal distribution. The null hypothesis for both tests is that the data is normally distributed, and the alternative hypothesis is that it is not. If the p-value, a measure of evidence against the null hypothesis, is below the chosen significance level, the null hypothesis is rejected, indicating non-normality of the data.

To obtain a comprehensive assessment of normality, it is advisable to use multiple normality tests in conjunction with graphical and numerical methods. Additionally, the choice of the significance level (α) is critical for interpreting the results of normality tests. In this thesis, a significance level of 0.05, following Pearson [45], was employed.

Significance tests

When analyzing data that does not follow a normal distribution, measures of central tendency such as medians are more appropriate than means [46]. Therefore, median tests are preferred for skewed data as they are more robust in such cases.

To investigate significant differences between groups, one approach is to compare the confidence interval around the median of one variable with the confidence interval of another variable or industry. Confidence intervals can be calculated using bootstrapping, a statistical technique that involves resampling the data to generate simulated samples [47]. Bootstrapping is commonly used for estimating standard errors and constructing confidence intervals. Visual comparison of the confidence intervals can provide insights into whether the groups differ significantly.

For non-parametric data analysis, when assumptions of parametric tests (such as the t-test or ANOVA) are not met, the Mann-Whitney U test (B.1.8) and the Kruskal-Wallis H test (B.1.9) are suitable alternatives. These tests do not require the data to follow a normal distribution or have equal variances.

The Mann-Whitney U test is used when comparing two independent groups with a small sample size. It assesses whether the medians of these groups differ significantly and is further explained in Appendix B.1.8.

The Kruskal-Wallis H test is employed when comparing the median values of three or more independent groups. It examines whether there is a significant difference among these groups. This test serves as an alternative to the one-way ANOVA when the data is not normally distributed or when the assumption of homogeneity of variances is violated. A brief explanation of the Kruskal-Wallis H test is provided in Appendix B.1.9.

Both tests yield p-values, which indicate the probability of observing the results by chance alone if the null hypothesis were true. If the p-value is lower than the chosen significance level (e.g., 0.05), it suggests that the observed differences are statistically significant. In this thesis, a significance level of 0.05 was selected as the threshold for determining statistical significance.

3.3.3. Ethical Considerations

Throughout the process of conducting, analyzing, and reporting on the survey, we prioritized adherence to research ethics guidelines, Norwegian personal data legislation, and the EU's General Data Protection Regulation (GDPR). We took measures to ensure the ethical treatment of the participants and the protection of their personal data.

Informed consent was obtained from all participants, as we provided a clear explanation of the study's purpose and procedures prior to the survey. Participants had the opportunity to ask questions and seek clarification before deciding to participate. Additionally, we assured the participants that all responses, analyses, and related processes would strictly comply with research ethics guidelines, Norwegian personal data legislation, and GDPR.

The confidentiality and anonymity of the respondents were crucial considerations. To safeguard privacy, we anonymized all survey responses and aggregated them into industry groups rather than reporting data for individual companies. Furthermore, raw data was excluded from the report to prevent any identification of individual participants. These measures maintained the privacy and confidentiality of the survey participants while enabling meaningful analysis of the collected data.

3.4. Survey Results

The survey data can be analyzed from different perspectives and may provide insights into the companies' credit management, as well as their credit value chain, credit processes, credit management maturity, and analytical maturity. To answer the research objectives described in Section 1.2.1, we investigated a set of hypotheses on the basis of the survey results. These hypotheses will be tested by processing the survey data, performing various statistical tests, and visualizing the results, followed by an analysis, interpretation, and discussion of the results. But firstly, the distribution of the data is addressed.

Normality tests

The normality of the survey data was assessed using several statistical tests, including the Shapiro-Wilk test, which is known for its robustness in detecting normality in small sample sizes. Table 3.1 shows that the p-value for the Shapiro-Wilk test was less than 0.05, indicating that the null hypothesis of normality was rejected. This result implies that the survey data is not normally distributed. Consistent with the Shapiro-Wilk test, other statistical tests, including the Anderson-Darling, with test statistics of [0.57, 0.65, 0.779, 0.909, 1.081] reject the null hypothesis on a significance level of 5%, D'Agostino's K-squared (B.1.6), Jarque-Bera, Kolmogorov-Smirnov, and Chi-square (B.1.7) tests, also rejected the null hypothesis with small p-values. Therefore, we conclude that the survey data is not normally distributed, and we must use non-parametric tests to analyze the data further, as well as the median as the measure of central tendency for the remaining survey result analysis.

Table 3.1.: Results of statistical tests confirm that the survey results are not normally distributed, which supports our assumption of overconfidence bias in the data.

Test	**Statistic**	**p-value**
Shapiro-Wilk	0.855	6.80e-19
Anderson-Darling	19.572	-
D'Agostino's K-squared	118.764	1.62e-26
Jarque-Bera	44.013	2.77e-10
Kolmogorov-Smirnov	0.812	4.56e-291
Chi-square	811.757	1.29e-29

Due to the limited sample size within each industry and the unequal distribution, normality tests within individual industries are not included, as their reliability would be compromised.

Apart from formal tests, data distribution can be evaluated through visual representations, such as box plots and violin plots. The skewness observed in Figure 3.4 for all five DELTA factors indicates that the data is not normally distributed. The asymmetric shape of the violins further corroborates the non-normality. In Figure 3.3, skewness is also evident, with whiskers extending to only one side of the box in several instances.

3.4.1. Hypothesis 1: Credit management maturity

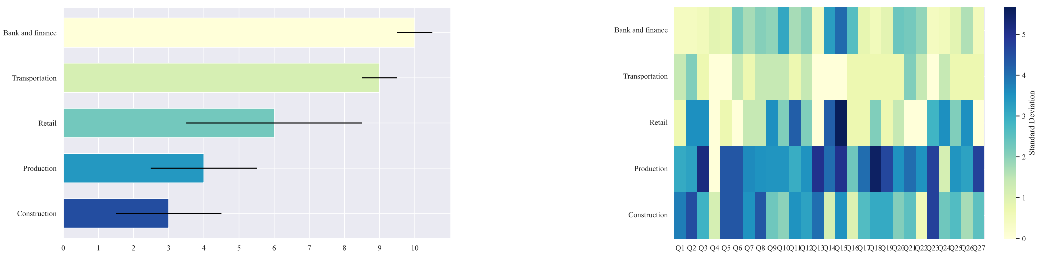
The purpose is to determine if the credit management maturity score in the banking and finance sector is significantly higher than the same measure in other industries.

We define each industry's credit management maturity score as the uniformly weighted median score for all questions in the credit survey, including credit strategy, classification and analyses of customers, establishing credit, monitoring and following up credit, and recovering losses. Our pre-survey hypothesis is defined as follows:

H0. The banking and finance industry's credit management maturity score is significantly higher than other industries.

Visual inspection

Figure 3.2a displays the overall survey scores distributed across each industry, with bar plots representing the median scores. Companies in the banking and finance industry have the highest score, closely followed by transportation. Conversely, construction has the lowest scores, with retail and production situated in the middle of the scale.



(a) Barplot displaying survey scores across industries, reflecting credit management maturity. The banking and finance industry leads, followed by transportation, with narrower confidence limits compared to other sectors.

(b) The heat map shows the variance for each question for each industry. As we see, the banking and finance industry and the transportation industry have less variance in their scores than the other industries.

Figure 3.2.: The median for each industry's score across all questions is visualized with an error band computed using bootstrapping. The heat map shows the variance in each question for the different industries.

We have also included bootstrapped confidence intervals for each industry to display the precision and accuracy of the survey responses within each industry and for each question. For the banking and finance industry, as well as the transportation industry, the confidence intervals are relatively narrow. The fact that the respondents' answers cluster around relatively similar scores in these two industries could indicate that the credit survey scores are a relatively accurate measure of the population measure in those industries. In contrast, the remaining industries have significantly wider confidence intervals, signifying less agreement on the state of current credit management practices.

Figure 3.2b presents a heat map that more clearly visualizes the standard deviation, enabling an examination of the variability among industries and individual questions. Darker areas represent greater dispersion in responses for each question (x-axis) within each industry (y-axis), which may suggest diverse perspectives and experiences on the relevant topic, leading to less consistent answers. Lighter areas, on the other hand, indicate less dispersion and more consistent responses within each industry. The heat map confirms that production has the highest variability, while transportation has the least, as indicated in Figure 3.2a. Furthermore, the heat map reveals that question 4 (i.e., leaders base all their decisions on facts - data, analyses, and models) has the lowest standard deviation among all industries, while question 15 (i.e., automatic credit limits) has the highest standard deviation. A more detailed heat map displaying the standard deviations for each cell can be found in Appendix A.2, in Figure A.1.

To get an indication of whether there are any significant differences between the different industries when it comes to credit management maturity, we visually inspect whether the confidence intervals for the overall median scores in the different industries overlap each other. Since the confidence interval of the banking and finance industry seemingly does not overlap with the median scores of the transportation industry or the other industries, it is tempting to argue that the banking and finance industry has a higher credit management maturity than the other industries. The transportation industry also seems to be better than retail, production, and construction. However, to examine our hypothesis, we need to employ formal statistical tests.

Statistical tests

The results in Table 3.2 indicate that significant differences exist between all industries, as suggested by the Kruskal-Wallis H p-value. However, further investigation with the Mann-Whitney U test reveals that the median scores in the banking and finance industry and the transportation industry are not significantly different (p-value = 0.1441). We also see that the banking and finance industry and the transportation industry both have significantly higher credit management maturity than the retail, production, and construction industries.

Table 3.2.: Formal tests for significant differences between the industry's credit maturity. Results indicate that the industries marked in green are significantly different, and the industries marked in red are not.

Test	Industries compared	p-value
Mann-Whitney U	Bank vs Transportation	0.1441
Mann-Whitney U	Bank vs Retail	0.0294
Mann-Whitney U	Bank vs Production	0.0284
Mann-Whitney U	Bank vs Construction	0.0294
Kruskal-Wallis H	All Industries	0.0042
Mann-Whitney U (Bonferroni-corrected)	Banking vs Transportation	0.1441
Mann-Whitney U (Bonferroni-corrected)	Banking vs Retail	0.0294
Mann-Whitney U (Bonferroni-corrected)	Banking vs Production	0.0284
Mann-Whitney U (Bonferroni-corrected)	Banking vs Construction	0.0294
Mann-Whitney U (Bonferroni-corrected)	Transportation vs Retail	0.0286
Mann-Whitney U (Bonferroni-corrected)	Transportation vs Production	0.0294
Mann-Whitney U (Bonferroni-corrected)	Transportation vs Construction	0.0286
Mann-Whitney U (Bonferroni-corrected)	Retail vs Production	0.3807
Mann-Whitney U (Bonferroni-corrected)	Retail vs Construction	0.1465
Mann-Whitney U (Bonferroni-corrected)	Production vs Construction	0.5590

Findings

Our analysis of credit management maturity across different industries suggests that the banking and finance industry, apart from the transportation industry, exhibits greater proficiency in credit management compared to other industries, which was indicated by our visual inspection in Figure 3.2, and further confirmed by the Kruskal-Wallis H test.

However, the Mann-Whitney U tests do not support our initial hypothesis (H0) that the banking and finance industry is significantly better at credit management than the transportation industry. Hence, our null hypothesis is formally rejected. In short, the banking and finance industry does have higher credit management maturity than the retail, production, and construction industries, but not compared to the transportation industry.

Discussion

It is not surprising that the different industries exhibit varying degrees of credit management maturity, given their different utilization of credit in their day-to-day businesses. According to leading credit experts, the construction industry's immaturity in credit management is well-known and costly and will have to be addressed over time. The retail and production industries, on the other hand, sell on credit with lower credit risk on their customers, while the transportation industry and the banking and finance industry utilize credit management on a daily basis.

However, it was still slightly surprising that the transportation industry scored itself almost as high as the banking and finance industry. One plausible explanation for this is the fact that the transportation industry has a different reference frame when it comes to credit management than the banking and finance industry. In short, what is considered excellent in the transportation industry may not even be considered mediocre in the banking and finance industry when it comes to credit management, while also remembering that credit management is part of the core competency and the core business process of the banking and finance industry. Another reason is the ever-present overconfidence bias (i.e., the tendency for persons to believe that they are better than what they really are, confer [48]), which is accentuated in the transportation industry compared to the banking and finance industry as the former has a larger set of unknowns and unknowables than the banking and finance industry when it comes to credit management. We also believe caution is warranted as the survey results are based on the subjective nature of people's opinions. Insights from our interviews with credit and industry-leading experts corroborated these assertions.

All in all, we believe credit experts are correct in stating that the Norwegian credit management maturity is significantly lower than the international state-of-the-art benchmark. We also believe that the survey scores are inflated and would be adjusted down for all industries if we were to employ more objective measures, but more so for transportation, retail, production, and construction than banking and finance. Finally, despite our null hypothesis being rejected, we maintain that it is still valid to conclude that the banking and finance industry represents the benchmark when it comes to credit management in Norway.

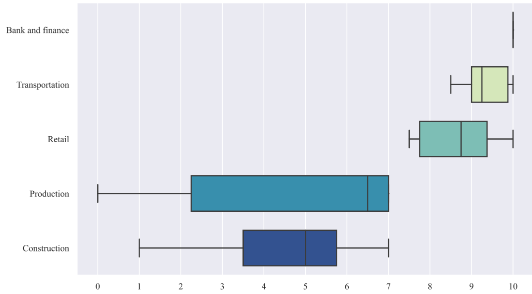
3.4.2. Hypothesis 2: Credit strategy and the credit value chain

The credit value chain consists of various parts that contribute differently to the credit process. Therefore, it is crucial to investigate whether there are statistically significant differences between companies and industries in terms of the entire credit value chain and its different parts, such as establishing credit, monitoring credit, preventing defaults, and debt collection. This analysis can help determine the credit management maturity of companies and compare them effectively.

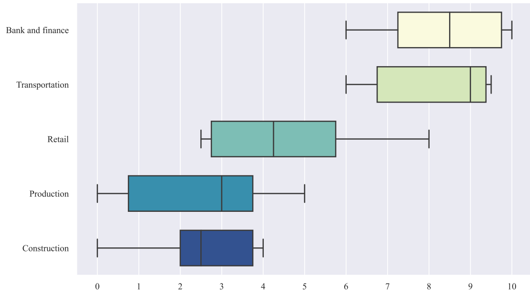
H0. The banking and finance industry's credit management maturity score is significantly higher than other industries in each part of the credit value chain.

Visual inspection

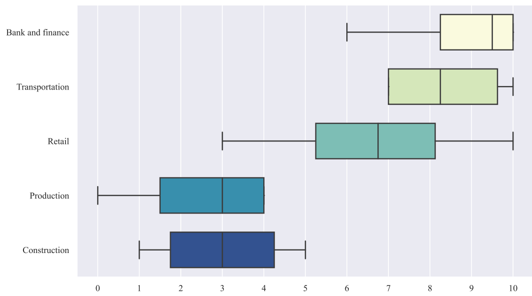
Figure 3.3 presents the median scores for each industry, similar to Figure 3.2a. However, Figure 3.3 displays the results for each component of the credit value chain, specifically the five sections outlined in Section 3.1: credit strategy, customer classification and analysis, credit establishment, credit monitoring and follow-up, and loss recovery. The median scores are visualized using box plots, enabling easier examination of data distribution and variability.



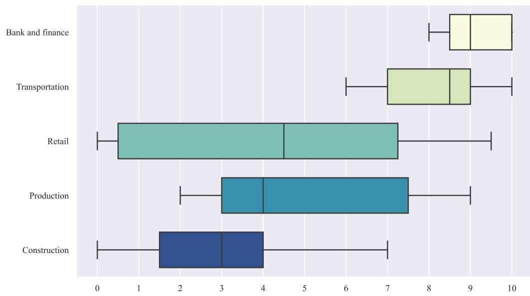
(a) Credit strategy.



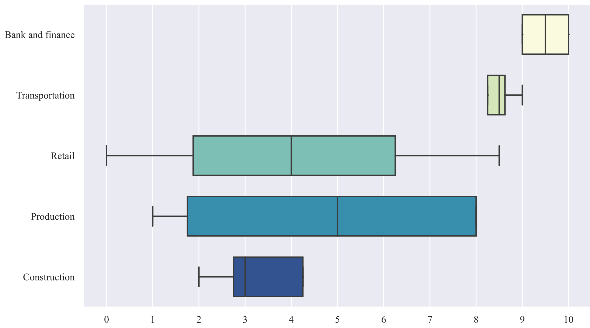
(b) Classification and analysis of customers.



(c) Establishing credit.



(d) Monitoring and following up credit.



(e) Recovering losses.

Figure 3.3.: Box plots displaying the consistent order of median scores across the credit value chain for each industry, with banking ranking the highest, followed by transportation, retail, production, and construction. The wider whiskers for retail, production, and construction suggest greater response variability. The visual differences between the industries could indicate that the differences are statistically significant.

Each subplot contains five boxes representing the industries. The boxes signify the interquartile range (IQR), which measures the variability of the data as the range between the first and third quartiles. The whiskers' minimum and maximum values indicate the data set's range, while the median is represented by the line inside the box, reflecting the data's central tendency.

Banking and finance consistently achieve the highest scores across all categories, closely followed by transportation. These two industries, along with construction, exhibit less variability in their scores, as their IQRs are consistently narrower than the others. The primary trend reveals that banking and finance, as well as transportation, attain high scores. In contrast, construction scores are lowest, and retail and production fall in the middle of the scale, corroborating Figure 3.2. A notable observation is that all industries have their medians in the upper part of the scale in Figure 3.3a, displaying scores related to a credit strategy. Banking and finance, in particular, achieve the highest possible score for all questions in this category. This aligns with the expectation that banking excels in credit strategy, being their area of expertise. Generally, all industries demonstrate lower and, in some cases, more dispersed scores in the other parts of the credit value chain.

Boxplots can be compared across industries to explore potential statistically significant differences, and overlapping whiskers can be identified. Generally, banking and finance, as well as transportation, have medians relatively close to one another in each category, suggesting similar central tendencies. The remaining industries — i.e., retail, production, and construction —also exhibit similar central tendencies in customer classification and analysis (Figure 3.3b) and credit monitoring and follow-up (Figure 3.3d). However, a significant difference exists in credit strategy between banking and finance, transportation, retail as a group, and production and construction. Another observation is that the whiskers of banking and finance boxes and construction boxes never overlap, indicating a significant difference between these industries. Finally, regarding loss recovery, the banking and finance industry whiskers do not coincide with any other industry, meaning that banking serves as the benchmark in this value chain component. However, it is crucial to remember that box plots alone cannot confirm statistical significance. Therefore, statistical tests should be conducted to verify the observed differences before drawing conclusions.

Statistical tests

Although the visual inspection implied that there could be significant differences between banking and finance and construction, the Kruskal-Wallis H test results, as presented in Table 3.3, indicate that there are no statistical differences within the various sections of the survey. The p-values are all above 0.05. Hence the null hypothesis cannot be rejected. These findings suggest that the responses across the different groups in each section were statistically indistinguishable.

Table 3.3.: Results of Kruskal-Wallis H test for each survey section. No values are below the chosen significance level of 5%, meaning there is not enough evidence to say that there is a statistically significant difference between the medians in the different industries.

Survey section	p-value
Credit strategy	0.9055
Classification	0.2123
Establishing credit	0.2141
Monitoring credit	0.1378
Recovering losses	0.3416

Findings

Our analysis of credit maturity across the different sections of the credit value chain concludes that although bank and finance is not statistically different in a better manner than the other industries, which rejects the null hypothesis H0. However, the visual inspection demonstrates that the banking and finance industry has the highest scores across all sections of the survey and the credit value chain.

Discussion

While the statistical tests reveal no significant difference between the medians, this may be due to the significant variation in answers across the bottom three industries. This variation, in turn, is a consequence of the small sample size. Nevertheless, the maturity of the industries is consistent with the overall order in Figure 3.2a. The primary trend reveals that banking and finance, as well as transportation, attain high scores, while construction scores are lowest, and retail and production fall in the middle of the scale. The bank and transportation industries exhibit consistently shorter whiskers than the bottom three industries, which suggests that the latter ones are less confident and consistent with their credit processes. The highest medians were observed in credit strategy, which may indicate that the industries find it easier to formulate policies than to put credit management into practice. These findings support the hypothesis (H0) that the banking and finance industry serves as the benchmark for credit management.

3.4.3. Hypothesis 3: Credit analytical maturity

We propose that a high degree of analytical maturity is essential for exceptional credit management, making it crucial to assess and pinpoint this maturity across various industries. By measuring analytical maturity, companies and industries can be compared to one another. We utilize the DELTA Model, as cited in [49], to evaluate analytical maturity levels. To gain deeper insights into the analytical maturity of different companies, we categorize select survey questions into the five factors of the DELTA model, which are outlined in Section 3.4.3. These factors are allocated among the DELTA attributes, as illustrated in Table 3.4.

H0. The banking and finance industry's DELTA score is significantly higher than other industries.

Analytical maturity refers to an organization's ability to effectively gather, interpret, and use data to inform decision-making and improve business outcomes [50]. One method for assessing the analytical maturity is the STO DELTA PAI Model developed by Haarberg [49], as presented in [51]:

The DELTA Model is an analytical maturity model and describes capabilities and assets needed in order to succeed with analytics incentives [49]. The five factors are grouped under the acronym DELTA:

D for accessible, high quality data

E for an enterprise-wide orientation

L for analytical leadership

T for strategic targets

A for analysts

For an organization to be analytically mature, all of these factors should work together and be equally positioned. It may be challenging if there is a lack of any of the DELTA factors, e.g., if the analysts in the organization are highly competent and analytically mature, but they do not have access to high-quality data. For organizations to orient themselves on the scale of analytical maturity, there is a five-stage model of progress [52]:

- *Stage 1: Analytically Impaired.* The organization lacks one or several of the prerequisites for serious analytical work, such as data, analytical skills, or senior management.
- *Stage 2: Localised Analytics.* There are pockets of analytical activity within the organization, but they are not coordinated or focused on strategic targets.
- *Stage 3: Analytical Aspirations.* The organization envisions a more analytical future, has established analytical capabilities, and has a few significant initiatives underway, but progress is slow – often because some critical DELTA factor has been too difficult to implement.
- *Stage 4: Analytical Companies.* The organization has the needed human and technological resources, applies analytics regularly, and realizes benefits across the business. But its strategic focus is not grounded in analytics, and it has not turned analytics into a competitive advantage.
- *Stage 5: Analytical Competitors.* The organization routinely uses analytics as a distinctive business capability. It takes an enterprise-wide approach, has committed and involved leadership, and has achieved large-scale results.

The DELTA Model will be our base for assessing analytical maturity in this thesis. Becoming an analytical competitor is not necessary or appropriate for all organizations at any given time. In certain situations, such as when there is limited time, the history is misleading, the decision maker has considerable experience, or when the variables cannot be measured, using this method may not be practical. However, most organizations will, in general, benefit from becoming more analytical.

Table 3.4.: Questions distributed within attributes in the DELTA Model.

Attribute	Questions
Data	Access to very easily accessible, reliable, relevant, and enriched data of very high quality
Enterprise	A credit policy that everyone follows fully
Leadership	Leaders who base all their decisions on facts (i.e., data, analyses, and models)
Targets	A credit policy with very clear goals, guidelines, incentives, mandates, and consequences for deviations from the credit policy
Analysts	Automatic and very good and easily accessible overview of customer statuses, customer segments, customer categories, and customer portfolios
	Automatic and very good methods and models for selecting and prioritizing customers
	Automatic and very good methods and models for calculating the customer's lifetime value (CLV)
	Automatic and very good methods and models for obtaining actionable insights wrt. customer behaviour, customers' preferred content, customers' preferred channels, and phase in the customer value chain (KYC)
	Automatic and very good methods and models for uncovering money laundering (AML)
	Automatic and very good methods and models for relevant external factors (e.g., the business cycle, market development, and competition) wrt. the business's decisions, processes, and analyses

Visual inspection

Figure 3.4 presents the scores from all companies with regard to the DELTA Model in a violin plot which shows the distribution of scores and the probability density for each DELTA factor. The violin plot combines box plots and kernel density plots as it consists of a boxplot, and the width of the violin is proportional to the density of data points for each DELTA factor. As the boxplots in Figure 3.3, the thin black line represents the data range from minimum to maximum value, the thicker box indicates the IQR and the white dot in the IQR shows the median value. The violin plot extends the data range to show the distribution more clearly.

The skewness indicates that all companies are analytically mature, as the probability density is higher at the upper side of the scale. The Target attribute stands out from the others, as the IQR (black thick box) is shorter, indicating that there is a smaller range among the scores related to the question about clearness in the credit policy, and these answers are also more concentrated in the higher end of the scale. Another observation is that the median for the questions related to analysts and automation is slightly lower than the others. This indicates that there is a lower analytical maturity connected to automation, and this will be investigated further in Section 3.4.5. To get more insight into the analytical maturity of the companies and to answer the research questions more precisely, the DELTA scores from the banking industry are compared to the DELTA

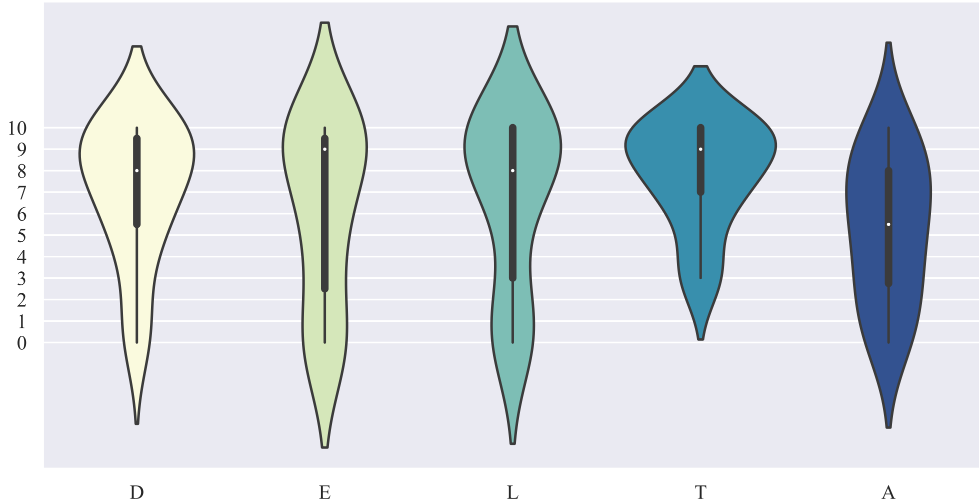
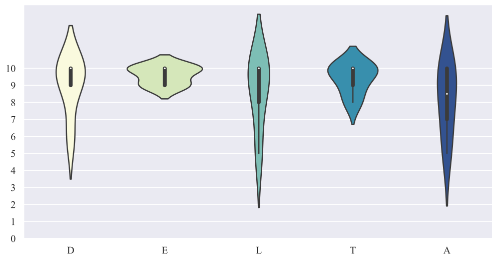
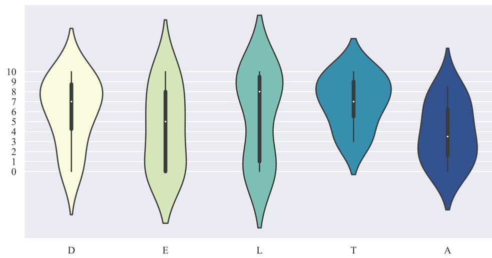


Figure 3.4.: Violin plot showing the distribution of DELTA attributes across the questions listed in Table 3.4. The results indicate a high degree of analytical maturity, as evidenced by the mean positioned high on the thicker black line. The distribution of all DELTA attributes is skewed.

scores from the rest of the industries in Figure 3.5.



(a) DELTA boxplot for bank industry.



(b) Boxplot for the rest of the industries.

Figure 3.5.: Violin plots comparing DELTA model scores: banking and finance industry vs. other industries (transportation, construction, production, and retail). Figure 3.5a shows skewed data with high analytical maturity in the banking and finance industry, while Figure 3.5b displays more evenly distributed data with lower analytical maturity and greater variability across the other industries.

Comparing Figure 3.5 to Figure 3.4, it is clear that Figure 3.5b is similar to Figure 3.4, while Figure 3.5a deviates significantly from the total plot. This indicates that the credit analytical maturity of the banking industry is higher than for the other industries. Firstly, the box plots inside the violins are significantly shorter than those placed higher on the scale for banking compared to the other industries. The IQRs are also shorter, meaning that there is less variability in the scores. The medians are 10 for all DELTA attributes, except for the analysts attribute, indicating that this is an area for more improvement among the banking companies, as well as for the other industries.

Secondly, the violins themselves are slimmer for data, leadership, and analysts, meaning that there is less variability in these answers. However, enterprise and targets have wider shapes, indicating more dispersion among the banking companies at these attributes.

Statistical tests

The statistical tests, presented in 3.5, indicate that there is a significant difference in the analytical maturity between the banking and finance sector and the rest of the industries, as confirmed by both the violin plots and the Mann-Whitney U test. The choice of the Mann-Whitney U test was appropriate because the DELTA scores for both groups did not follow a normal distribution. The test resulted in a low p-value, providing strong evidence to reject the null hypothesis and confirm the significant difference in analytical maturity between the two groups.

Table 3.5.: Results of the Mann-Whitney U test imply that there are significant differences between the analytical maturity of banks and finance and the rest of the industries.

Test	**U statistic**	**p-value**
Mann-Whitney U test	1012.00	0.0000

Findings

Our evaluation of analytical maturity using the DELTA model reveals, through both visual inspection and formal statistical tests, that the banking and finance industry performs significantly better than the other industries in terms of analytical maturity. Consequently, we do not reject the hypothesis. The analysis offers valuable insights into the analytical maturity of the survey participants. Results in Figure 3.4 indicate that all companies exhibit analytical maturity, as the probability density is higher towards the upper end of the scale. A comparison between the DELTA scores of the banking industry and other industries shows that the banking sector has greater analytical maturity. The Mann-Whitney U test results further support this finding, revealing a statistically significant difference between the analytical maturity of the banking industry and the remaining industries.

Discussion

The overall high median scores in the violin plots in Figure 3.4 are influenced by the high scores of the banking and finance industry. The comparison in Figure 3.5 suggests that the banking and finance sector excels in analytical maturity compared to other industries.

Median scores for questions related to analysts and automation were slightly lower than for other DELTA factors, both for banking and finance and the remaining industries, indicating a lower level of analytical maturity in these areas. This section of the DELTA model focuses on analysts and incorporates automated models to derive insights from customer data. The low scores in this section can be attributed to the limited degree of automation in credit management within Norwegian financial institutions, as well as the minimal or nonexistent calculation of customer lifetime value. This aspect of the credit management process is the most complex and valuable, warranting improvement to optimize credit maturity.

The box plots inside the violins for the banking industry were notably shorter than those for other industries, signifying less variability in scores. Furthermore, the medians for all DELTA attributes were 10, with the exception of the analysts attribute, suggesting that this area needs further improvement for both banking and other industries. This could be because industries that consider credit management a secondary business process may not prioritize analytical maturity, unlike the banking and finance sector, which regards it as a core process.

3.4.4. Hypothesis 4: Checking for consistency

To scrutinize the company's credit management, credit strategy and credit policy, checks and balances can be done for possible exaggerations and to check for consistency. To do so, the questions will be compared for each industry. For instance, if the two first questions within the credit strategy (everyone follows the credit policy, and the credit policy has very clear goals) differs significantly, there is an inconsistency. If a respondent provides a high score on one of them and a low on the other, the respondent is not consistent and the credit policy is thus not clear.

H0. All survey participants are consistent in their answers.

Visual inspection

Figure 3.6 is a line plot with the median industry score on the y-axis and the survey questions on the x-axis. The plot connects all the data points with lines, visualizing the answer trend for each industry. In this way, the consistency among the answers can be investigated. The plot can be analyzed in general and for each survey section. The distribution of the questions is as follows:

- Q1-Q6: Credit Strategy
- Q7-Q12: Classification and analysis of customers
- Q13-Q16: Establishing credit
- Q17-Q23: Monitoring and following up credit
- Q24-Q27: Recovering losses

In general, there are no straight lines in the plot, meaning that there is inconsistency among the answers to some extent, which is as expected. However, there is a difference in score variability among the industries. As observed several times, there is less variability in the answers from banking and transportation, as their scores ranges between 6 and 10. Retail ranges from 0 to 10, production from 0 to 9, and construction from 0 to 8, meaning that the ranges for the three latter are at least doubled from the two first.

Diving into the survey sections, banking and transportation are most consistent in their answers regarding credit strategy and recovering losses, showing that there is more inconsistency in the three first parts of the credit value chain , i.e., classification and analyses of customers, establishing credit and monitoring and following up credit. For all industries in total, the spikes are most significant in the section about monitoring and following up losses.

Figure 3.6.: Line plot illustrating the variation in survey responses across industries. Banking and finance, along with transportation, demonstrate consistent answers, whereas the remaining industries exhibit greater variability.

Findings

The survey results reveal varying scores across industries, which is expected given the diverse challenges they face. However, the banking and transportation sectors display scores within significantly smaller ranges compared to the other industries. Monitoring and addressing losses is the part of the value chain with the most inconsistent responses from all industries.

Discussion

The smaller variation in the responses from the banking and finance sector may imply consistent confidence throughout the credit management processes. However, this could also be a result of overconfidence bias, considering the high scores. A high degree of variability exists among the responses within the retail, production, and construction industries. This variability may arise from different interpretations of the questions, varied levels of knowledge or experience related to the topics, subjectivity, or actual differences in performance. All industries exhibit minimal variation in the first two questions about credit policy, which could suggest consistency in their interpretation of the questions. Nonetheless, retail and production exhibit the most variation throughout the questions, which could be attributed to substantial differences in the quality of their credit management or possible inconsistency and dishonesty in their responses. The considerable variation in retail's and production's responses to questions 4 and 5 implies that credit policies may not be well-established at the executive level in these businesses.

3.4.5. Hypothesis 5: The degree of automation

Automation plays a crucial role in maximizing revenue, optimizing resource allocation, and improving operational efficiency within companies. To assess the degree of automation across industries, we analyzed whether there were statistically significant differences in the median scores of respondents' answers to automation-related questions. These questions served as a proxy for measuring the level of automation in different companies and industries. The survey included 11 questions specifically related to automation:

- Automatic and very good methods and models to optimize credit risk given the customers' risk-adjusted profitability and credit limits as well as external factors.
- Automatic and very good and easily accessible overview of customer statuses, customer segments, customer categories, and customer portfolios.
- Automatic and very good methods and models for selecting and prioritizing customers.
- Automatic and very good methods and models for calculating the customer's lifetime value (CLV).
- Automatic and very good methods and models for obtaining actionable insights wrt. customer behavior, customers' preferred content, customers' preferred channels, and phase in the customer value chain (KYC).
- Automatic and very good methods and models for uncovering money laundering (AML).
- Automatic and very good methods and models for relevant external factors (e.g., the business cycle, market development, and competition) wrt. the business's decisions, processes, and analyses.
- Automatic credit limits.
- Automatic and very good methods and models for calculating the effects of loss prevention activities.
- Automatic and very good methods and models for calculating risk-adjusted profitability (including the costs associated with trouble and delay).
- Automatic and very good methods and models for calculating customer sentiment.

Based on these questions, we formulated and tested the following hypothesis:

H0. Banking and finance utilize automation to a significantly higher degree compared to the other industries.

Visual inspection

To assess the degree of automation across different industries, we multiplied the scores for automation-related questions by 10 and converted them into percentages for each company within each industry. The results are presented in Figure 3.7.

There are considerable differences between the industries. Banking has the highest degree of automation. Transportation has the second-highest degree of automation, and together with banking, their levels of automation are more than double those of the other industries.

Statistical tests

The results of the statistical tests conducted on the automation data are summarized in Table 3.6. Both the non-parametric Kruskal-Wallis H test and the Friedman test (B.1.10) reject the null hypothesis of equal medians, indicating significant differences among the industries in terms of their automation scores. The pairwise comparisons

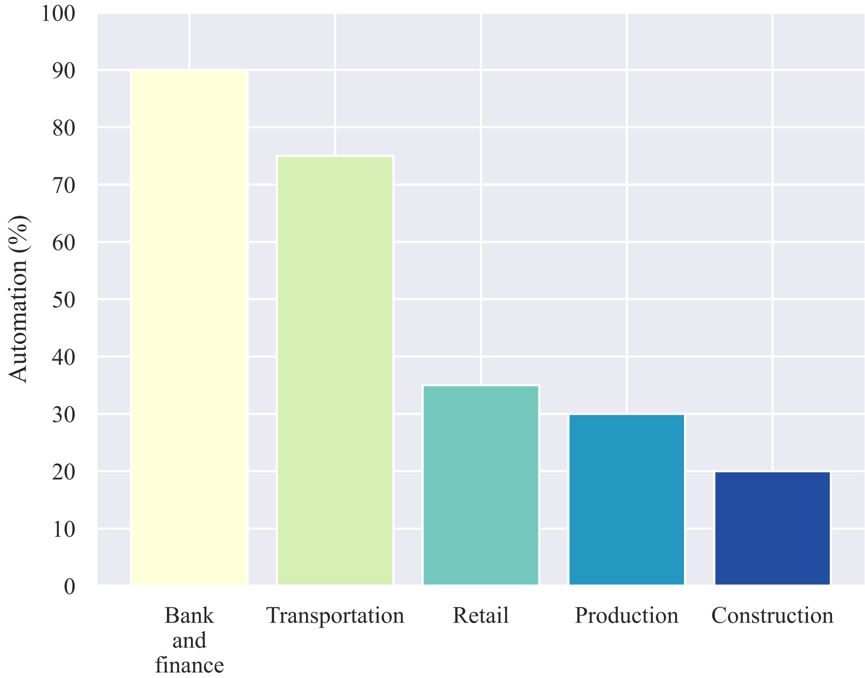


Figure 3.7.: The banking industry exhibits the highest level of automation, followed by transportation.

using the Mann-Whitney U test show that all industry comparisons, except for the bottom three, exhibit significant differences in their automation scores. Specifically, the banking and transportation industries show higher automation levels compared to the retail, production, and construction industries. The bottom three industries demonstrate lower levels of automation, and their scores are not statistically different from each other.

Findings

The formal tests reject the hypothesis that the banking and finance sector is statistically superior to the transportation sector, but the tests do support this assertion for the remaining industries. This pattern can be observed consistently throughout the data and hypothesis testing.

Discussion

Figure 3.7 demonstrates a significant variation in the level of automation across different industries, as confirmed by the statistical tests. The banking and finance sector exhibits the highest level of automation, followed by the transportation industry, while the other three industries show notably lower levels of automation.

The overall low level of automation among Norwegian companies in various industries suggests that there is significant room for improvement to optimize credit management. The banking industry's automation level is more than four times higher than that of the construction industry, indicating a substantial gap between banking and other sectors, and highlighting the potential for improvement and optimization of existing systems. The questions related to automation assess the competence of extracting insights from customer data. The results imply that the industries may lack the necessary competence

Table 3.6.: Formal tests to check for statistical differences in the level of automation across industries.

(a) The non-parametric tests conducted on the data reject the null hypothesis of equal medians, indicating that there are one or more industries that exhibit significant differences compared to the other industries.

Test	**Statistic**	**p-value**
Kruskal-Wallis H	31.080	0.000003
Friedman	32.019	0.000002

(b) Significant differences were observed in all industry comparisons except for bank and transportation, as well as the bottom three industries. Industries marked in green indicate statistical differences, while those marked in red do not reject the null hypothesis.

Comparison	**U**	**p-value**
Banking vs transportation	82.0	0.161
Banking vs retail	106.5	0.002
Banking vs production	118.5	0.000
Banking vs construction	118.5	0.000
Transportation vs retail	101.0	0.008
Transportation vs production	116.5	0.000
Transportation vs construction	116.5	0.000
Retail vs production	68.0	0.644
Retail vs construction	77.0	0.291
Production vs construction	73.5	0.404

to derive these insights from their customer data, a conclusion that is also supported by interviews with industry-leading experts.

3.4.6. Hypothesis 6: The most significant credit challenges

Identifying the most significant credit challenges for companies and across industries can be achieved by examining the survey questions and parts of the value chain that receive the lowest scores overall. These low scores also indicate areas where the respondents are least satisfied. Investigating these challenges is essential for determining which parts of the credit value chain require improvement. Once these areas have been identified, potential solutions for enhancing these aspects can be developed and implemented.

H0. Automation represents the most substantial challenge for financial institutions concerning credit management.

Visual inspection

Figure 3.8 displays the combined median scores for all industries arranged in ascending order, to provide an indication of the most significant challenges. The median was chosen as the measure of central tendency due to the non-normal distribution of the data. As the plot shows, banking and finance significantly contribute to the overall median score. This is to investigate how the alleged benchmark is compared to the other industries, as well as the possible overconfidence in the banking and finance industry. Therefore, excluding this industry can provide a more representative view of the most significant challenges in credit management among Norwegian industries. By doing so, it becomes evident that the remaining industries suffer from these challenges more significantly, thus providing a more accurate representation of the credit management state in Norway.

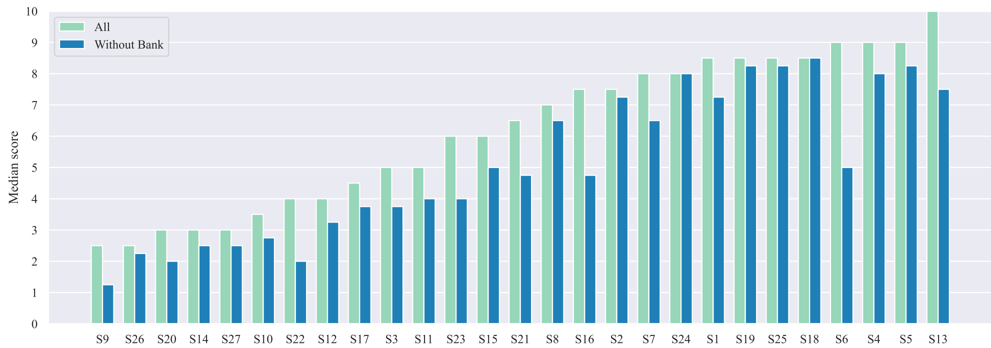


Figure 3.8.: Barplot of questions with the lowest average scores, highlighting significant challenges in questions 9, 26, 20, 14, and 27.

The five bottom questions, shown in Table 3.7, are from four different sections of the survey, indicating that challenges occur in all parts of the credit value chain. There is one from each of the three first parts of the credit value chain, i.e., classification and analyses of customers, establishing credit and monitoring and following up credit, and two s from recovering losses.

Table 3.7.: The five questions with the lowest median scores across all industries. When banking is included, it clearly lowers the median scores.

Question	**All industries**	**Without bank**
9 - automatic and very good methods and models for calculating the customer's lifetime value	2.5	1.25
26 - very good collaterals for all customers	2.5	2.25
20 - automatic and very good methods and models for calculating the effects of loss prevention activities	3.0	2.00
14 - no manual credit procedures	3.0	2.50
27 - very good methods and models for calculating the probability of loss recovery, default probability, expected loss given default, and net exposure at default	3.0	2.50

Figure 3.9 provides an overview of the challenges encountered across different industries within the credit value chain. The plot shows that there is a limited range of medians, which can be attributed to the considerable variation within each section. Figures 3.6 and 3.2b further support this observation by highlighting the varying responses provided by companies across different sections. Consequently, the medians for the various sections in Figure 3.9 exhibit lower variation. It is worthwhile to examine which section of the credit value chain presents the most significant challenges. Classification has the lowest median, as seen in both the bars of Figure 3.9, whether or not bank and finance are included.

Figure 3.9 shows that the classification section of the credit value chain presents the most significant challenges. This aligns with the finding from Table 3.7 and Figure 3.8

that computing customer lifetime value is a challenge. Notably, the classification section has the highest number of s related to automation, reinforcing the need for increased automation in this area.

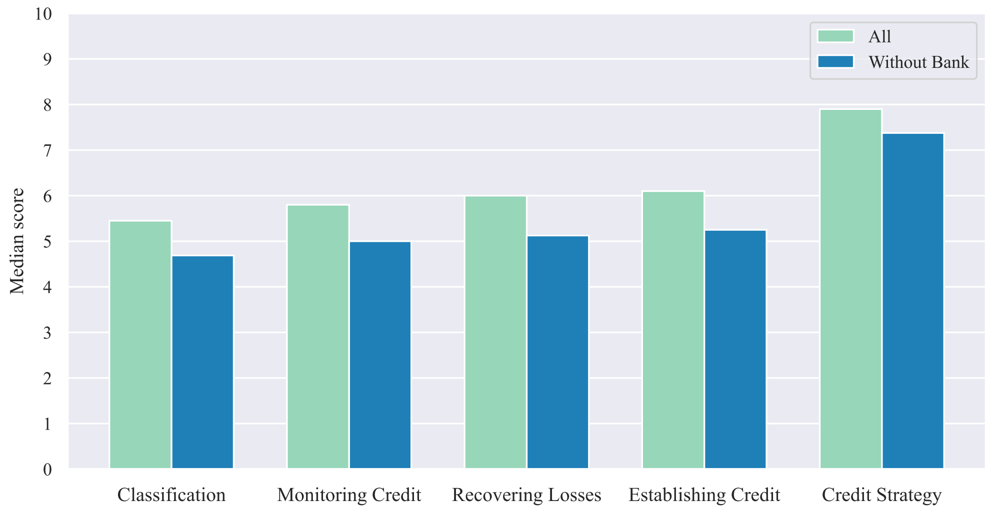


Figure 3.9.: The industries appear to have the most challenges in the classification section of the credit value chain and less with credit strategies.

Findings

The analysis of credit challenges in Norwegian industries reveals that challenges are present throughout the credit value chain, and there is a lack of automation in credit management processes. Therefore, we can conclude that the hypothesis is not rejected due to automation being four out of the five bottom questions.

Discussion

Most questions with low scores relate to automation, highlighting a lack of automation in credit management processes. This observation is consistent with the results from the previous hypothesis. Figure 3.9 corresponds with Table 3.7 and Figure 3.8, emphasizing the challenge in computing customer lifetime value, which falls under the classification of customers. Notably, this section also contains the highest number of s concerning the level of automation. These findings reinforce the hypothesis that most industries have low levels of automation. Interestingly, the highest score is for credit strategy, suggesting that industries find it easier to develop policies than to implement credit management practices effectively. These findings are supported by our interviews with industry-leading experts, who pointed out a low level of automated procedures in credit management and a prevalence of manual processes. This situation can result in a significant amount of time spent on unprofitable customers, which may not be the most efficient use of resources.

3.5. Summary of Findings

The analysis and interpretation of the survey results provide valuable insights into credit management practices in different Norwegian industries and contribute to answering the research questions in Section 1.2.2. The main findings are:

- The banking and finance industry is the benchmark for credit management as they consistently have higher scores compared to the other industries. However, transportation closely follows banking and finance and thus also emerges as a leading industry based on the survey results.
- The companies seem to perform better when it comes to credit strategy compared to the other parts of the survey, i.e., the credit value chain. This apparent decoupling between the credit strategy and the credit value chain is at odds with what leading practitioners say, as most acknowledge that in other industries than banking and finance, and transportation, there is a lack of clear, robust, and transparent strategy, roles, mandates, incentives, rules, guidelines, checks and balances, and repercussions in case of potential breaches. In fact, in other industries, most of the credit decisions have to be made by the credit department, as they receive inadequate guidance from the top management group. Another possible explanation is that the credit strategy is in place but rarely used in practice, and thus the daily credit tasks become relatively complex.
- There is overall room for improvement concerning analytical maturity.
- During the analysis of the survey's lowest median scores, automation emerged as a significant challenge in the credit management process. Specifically, the computation of customer lifetime value was highlighted as the most complex aspect of credit management, irrespective of whether or not the banking and finance industry was considered.

It is crucial to acknowledge that the findings of this study are based on survey results, which the subjectivity and uneven distribution of survey participants across industries may influence. Even though the banking and finance industry emerged as the benchmark in credit management, it is important to recognize that more experienced and leading foreign credit management firms would not rate any Norwegian firm conducting credit management anywhere near a perfect score of 10. An overall score of 10 implies flawless credit management and complete mastery of all aspects of credit management. Conversations with leading experts in the credit management field indicate that the Norwegian banking and finance industry has ample room for improvement in all aspects of the credit value chain.

Analytical maturity, as assessed using the DELTA model, is another area where the banking and finance industry outperforms the other industries. However, there is still room for improvement in the analysts and automation attributes for both the banking and finance industry, as well as for the other industries. Enhancing analytical maturity could lead to better decision-making and more efficient credit management processes across all industries.

By adopting automation and optimizing credit management processes, industries can reduce the time spent on unprofitable customers and allocate resources more efficiently. Given these findings, especially the challenges associated with automation and calculating CLV, the next chapter of the thesis will focus on addressing these critical aspects, with the aim to provide a framework that can work as possible solutions. Furthermore, this model will serve as a stepping stone towards enhancing automation in the credit management process, thus optimizing resource allocation and improving overall credit management performance across industries. With the insights gained from the survey results, we are now prepared to delve into the practical aspect of credit management and explore potential directions for improvement.

4. Applied Credit Risk Modeling

This chapter builds upon the findings from our credit survey presented in Chapter 3, where we identified several areas where there is room for improvement regarding both credit management and managing the different parts of the credit value chain.

Limited time does not permit us to dive into all the identified shortcomings regarding credit management. However, among the various aspects of credit management, the computation of customer lifetime value (CLV) emerged as the most demanding, interesting, complex, and critical area. Therefore, we have chosen to attempt to develop methods for calculating CLV.

As described in Chapter 2, we divide CLV into five components; forecasting sales, forecasting profitability, expected customer duration, net exposure, and the discount factor. The following sections will briefly explain the methods used to calculate each component, the empirical results, and the findings we derive from them.

4.1. Experimental Set-up

4.1.1. Data Sources

The data used in the modeling tasks are gathered from two primary sources. Our first data source was macroeconomic data obtained from Statistics Norway's statistics bank [53]. This data consist of historical values and forecasts for population, wage growth, 3-month Norwegian Interbank Offered Rate (Nibor), mainland GDP, etc. We chose not to use total GDP as it is customary to exclude petroleum and natural gas in macroeconomic analyses in Norway, primarily because oil and natural gas production typically are highly volatile [54]. Our second data source was the company data set received from a participating company in our credit survey. This includes invoice data, customer relationship duration, country information, and other relevant metrics. Please refer to Appendix C.1 for an overview of the provided data.

4.1.2. Data Pre-processing

The company data set consists of monthly observations for 15,848 unique customers in a time-series format. However, the data quality varies significantly, which impacts the calculations. The data set has 179,168 rows with considerable missing data, confer Table 4.1. To minimize data loss and prevent biased parameter estimates, appropriate imputation methods are employed following the guidelines of [55]. Missing end dates are treated as censored data, particularly in survival analysis. The customer name column is not used due to the EU's GDPR, Norwegian law, research ethical considerations, and varying data quality, including missing, temporary, and notes entries. Rows with missing postcodes and countries are removed from the analysis. Additionally, there is missing data for invoice amounts, costs, invoice reminders, and reminder degrees, as some companies have no activity during the observation period. The missing data can affect the representativeness of the data and, thus, our empirical results.

Table 4.1.: A comprehensive review of the variables included in the data set along with the corresponding counts of missing values. The considerable number of missing entries across various variables underscores the extensive gaps in the data.

Variable	**Missing values**
Period	0
Customer ID	0
Customer Name	31
Postcode	2211
Country	1226
Customer Group	5
Industry	44382
Risk Segment	123963
End Date	94913
Start Date	0
Calculated Credit Limit	67959
Invoice Frequency	0
Invoiced Amount	126749
Cost	126749
Other Costs	126749
Invoice Reminders	168944
Invoice Reminder Degree	168944

Testing for stationarity

Before proceeding with the modeling, it is essential to address the stationarity of the time series as most forecasting models presuppose stationary data, see Chapter 2.3.1. There is an abundance of stationarity tests, and each involves several steps. Apart from the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root test, other unit root tests' null hypothesis is that there is a unit root (i.e., that the times series is non-stationary), which is contrary to other statistical tests such as Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP). The tests are explained in further detail in Appendix B.2. In short, we utilized the following unit root test procedure:

First, we plot the series to be able to determine the correct test parameters in the ADF unit root test.

Second, a unit root test is conducted to determine the order of integration of the raw times series. This test helps determine if the times series requires differencing to achieve stationarity. The test can yield both integer unit roots and fractional unit roots, however, we will not assume fractional unit roots in this thesis. The potential existence of a cointegrating equation that could render the series stationary, brings us to the third step.

If a unit root is present, we perform a Johansen cointegration regression test to check if the variables are cointegrated. This step identifies any potential long-term relationships among the variables. When identifying the cointegrating equation, we consider the trend, cycle, and seasonality components of the time series, according to Wold's 1938 decomposition [14]. When trying to establish the cointegration equation, we would thus choose variables that are able to capture the underlying trend, cycle, and seasonality in the I(1) time series. This leads us to step 4 in the unit root test procedure.

Finally, we perform an appropriate unit root test on the residuals from the cointegration regression to check for cointegration. If confirmed, we incorporate the estimated cointegration vectors as an additional predictor in the original regression and then re-estimate the regression following Kennedy's guide [20]. This inclusion helps achieve stationarity and should be executed cautiously to prevent spurious outcomes. Alternatively, differencing the data can be considered. In such cases, reversing the differencing process after conducting the forecast is crucial. This is necessary because the differenced data capture the changes, e.g., the invoiced amount, rather than the actual values. The last observation from the original data set is retrieved to reverse the differencing, and the differenced forecast values are added to it. This procedure is repeated for subsequent differenced values, ensuring each addition is made to the previous reconstructed value. Following this approach, the forecast data can be obtained in its original scale.

Table 4.2 shows the results of the ADF test applied on the invoiced data (i.e., sales) time series for each Norwegian customer in the provided data set. It shows that many of the customers' sales are stationary, implying that we can utilize sales in their original form when forecasting. Customers with insufficient data samples were not included in the test and subsequently removed from the data set. In line with our brief presentation of unit root test above, customers with non-stationary sales had sales that exhibited p-values larger than the significance level of 0.05 in the ADF test. In these cases, further processing, as described above, is necessary as non-stationary time series exhibit patterns that change over time, making it challenging to model their behavior accurately.

Table 4.2.: Counts of stationary sales among Norwegian customers in our data set. Almost 50% of the customers were not checked for stationarity due to lack of data and the amount of customers with stationary sales are slightly higher than customers with non-stationary sales.

Category	Count
Customers with stationary sales	915
Customers with non-stationary sales	755
Not checked customers	1590

4.2. Forecasting

Forecasting concerns making predictions or estimates about the future based on historical data and patterns, as explained in Section 2.3.1. Given the data presented in Section 4.1.1, we argue that time series regressions are the most suitable techniques to forecast sales. Time series regressions are designed to capture the time-dependent patterns in the data, provides coefficients that yields insights on the impact of the predictors and accounts for autocorrelation, which is relevant for our forecasting task. However, machine learning techniques such as LightGBM (LGBM), ElasticNet Regression (ER) and Extra Trees Regression (ERF) can also be used, particularly if there are complex and non-linear relationships between the predictors and the target variable and for handling high-dimensional data. We still chose to point our focus at time series regression, for simplicity and interpretability in the thesis, as well as due to our limited data availability. Referring to Appendix C.1, the data set contains different customers (and thus several independent variables) and several explanatory variables.

More of these variables can be used as explanatory variables and therefore, multivariate multiple regression can be utilized to make a model for sales. Multivariate multiple regression is a statistical modeling technique used to analyze the relationship between several dependent variables and multiple independent variables [56]. Multivariate multiple regression can be described by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon, \tag{4.1}$$

where Y are the dependent variables to be forecast, the $X's$ are the independent variables that are used in the forecast, β_0 is the intercept term, $\beta_1 - \beta_n$ are the regression coefficients and ε is the error term.

Before delving into the forecasting task, it is important to clarify that, for the purpose of illustration, we have chosen to utilize data from a single selected company, referred to as Customer X. Our analysis and forecasting are conducted using this specific data set as an example.

4.2.1. Forecasting Sales

In this forecasting task, the dependent variables to be forecast are the invoiced amount, i.e., sales.

Exploratory data analysis

To begin the forecasting task, we performed an exploratory data analysis to better understand the underlying patterns in the data.

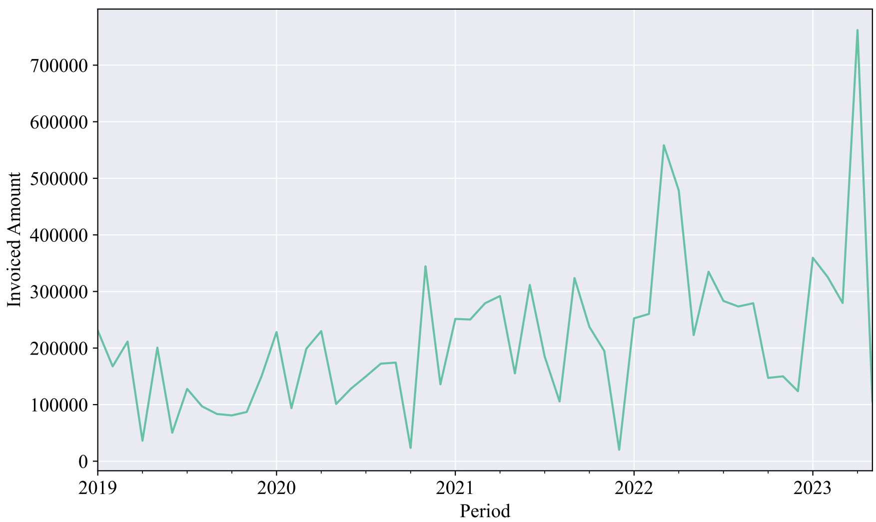


Figure 4.1.: Time plot showing the invoiced amount, i.e., the variable to be forecast, for Customer X.

We start by constructing a line plot of the time series of the invoiced amount for the selected company to get an overview of the data. Figure 4.1 shows the invoiced amount for the company over a five-year period. The plot shows an increasing trend. One can investigate the data further by examining the yearly data.

Figure 4.2 shows the yearly invoiced amount for the company. By visual inspection of Figure 4.2, there is a slight indication of seasonality, as the invoiced amount appears to be higher in some repeating periods, which can indicate Easter and Christmas effects.

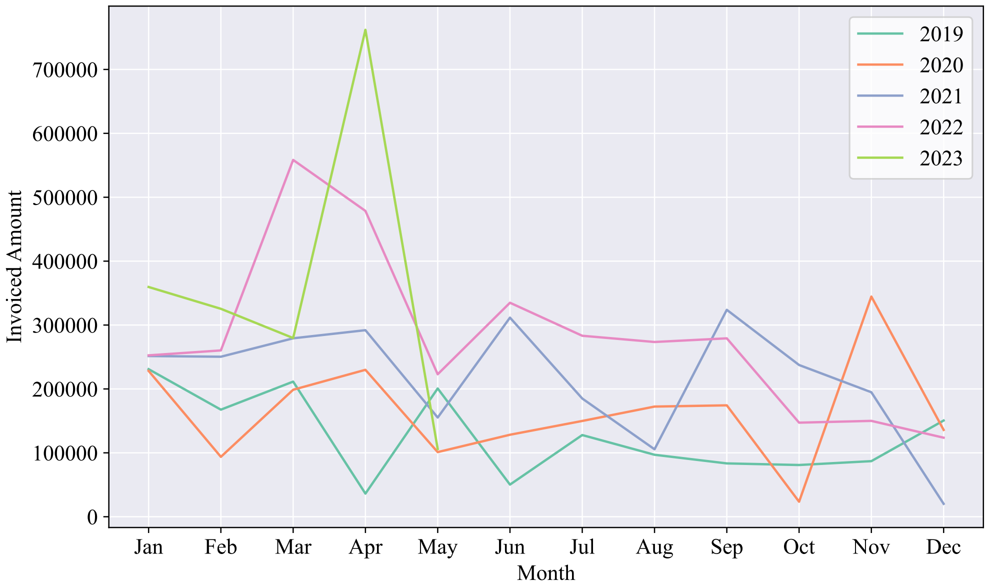


Figure 4.2.: Seasonal time plot showing the yearly invoiced amount for Customer X, indicating seasonality in the data.

However, to gain a deeper insight into the data before choosing the modeling techniques, a time series decomposition, explained in Section 2.3.1, is performed. We used additive decomposition as there is an indication of linear trend in our data. The decomposition is performed using the `statsmodels` function `seasonal_decompose` in Python. The result is shown in Figure 4.3.

The figure consists of five subplots, demonstrating the different components of the additive decomposition. The trend component displays the overall pattern of the data, excluding seasonal and cyclical fluctuations. Note that the trend, cycle, and residual components are displayed for a shorter time period than the original series. This occurs because of the smoothing and averaging procedures employed during the decomposition process to capture underlying patterns. Similarly, the residual component is also shorter as it represents the remaining variation after accounting for the trend and seasonal components. The trend suggests that the data is increasing with time. The seasonal component represents the periodic fluctuations in the data, occurring at regular intervals, yearly in this case, confirming the observations in Figure 4.2. On the other hand, the cycle component represents the long-term fluctuations or patterns in the data that repeats over a period longer that the seasonal cycle, and may represent underlying economic, social or environmental factors that influence the long-term behavior of the data. We see that the pattern changes after 2022, which may be related to the decline of the COVID-19 pandemic. Finally, the residual component indicates the random fluctuations in the data that are not accounted for by the other components. These fluctuations can be considered as noise or error. It is noteworthy that the residual component oscillates around zero. The noise might be a sequence of random variables that are uncorrelated and have constant variance, i.e., white noise [23]. This can be checked by performing the Ljung-Box test on the residual data. The Ljung-Box test (B.2.9) examines the autocorrelation of the residuals at different lags. It tests the null hypothesis that the autocorrelations up to a specified lag are zero, indicating no significant autocorrelation in the residuals. The test yielded a p-value of 0.308, which is greater than the chosen

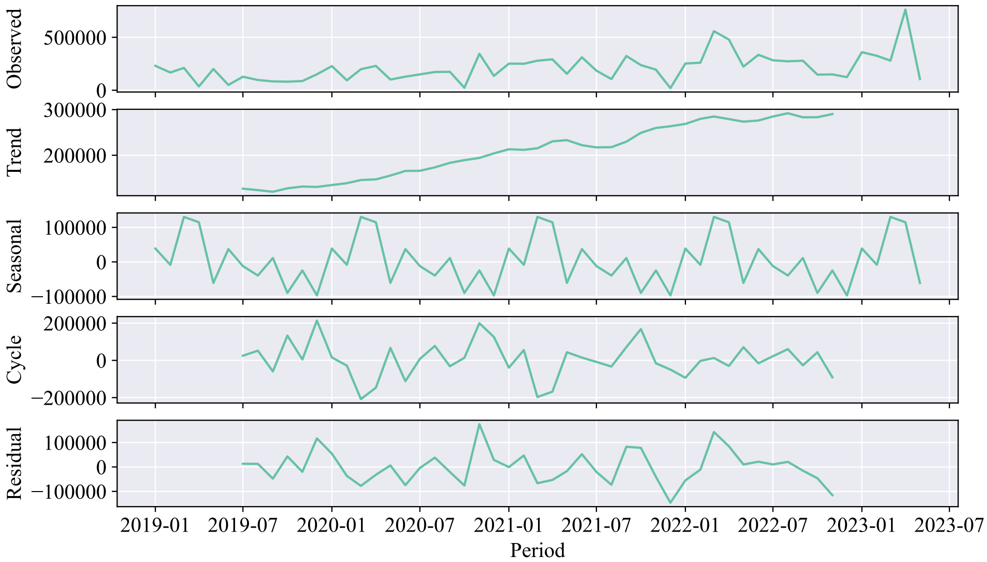


Figure 4.3.: Additive decomposition of the invoiced amount time series for Customer X, showing each component of the time series separately.

significance level 0.05. This suggests that the residuals resemble white noise.

In addition to identifying patterns in the time series, the time plot can also contribute to assessing the necessity of adjustments such as mathematical transformations or calendar adjustments. Such adjustments will ideally stabilize the variance in the time series, where rapid changes in parts of the time plot can affect the accuracy of the forecasting [15]. To assess changing variance over time, a the Breusch-Pagan test (B.2.10) is conducted. This statistical test checks for heteroscedasticity, indicating non-constant variance in the data. The null hypothesis assumes constant variance. If there is evidence of changing variance over time, it may be necessary to apply a transformation to stabilize the variance. The results of the Breusch-Pagan test can be found in Table 4.3.

Table 4.3.: Results of formal tests for heteroscedasticity with the Breusch-Pagan test showing that the variance in the data is not constant over time.

Test	Statistic	p-value
LM-Test	5.977	0.014
F-Test	6.482	0.0140

Table 4.3 presents the results of our heteroscedasticity tests. Both tests demonstrate p-values below the significance threshold of 0.05, suggesting the rejection of the null hypothesis. Consequently, this indicates the presence of heteroscedasticity in our data, implying a non-constant variance over time. To address this, we implement a seasonal adjustment to the data to eliminate the seasonal component. Post-adjustment, the Breusch-Pagan test reports p-values surpassing the significance level, confirming a stable variance in the adjusted data and thereby indicating its readiness for subsequent modeling.

Implementation

Initially, we need to check whether the times series are stationary. The ADF test is performed to determine if the data set for Customer X is stationary and contains a unit root. The results from the ADF test is shown in Table 4.4.

Table 4.4.: Results from the ADF test on the invoiced amount data for Customer X, indicating that the data is stationary.

Test statistic	p-value	Critical values
-5.645	6.165e-07	1%: -3.563 5%: -2.919 10%: -2.597

The test statistics are more negative than the critical values. This provide strong evidence against the the null hypothesis. Hence, the times series does not contain a unit root and we can thus conclude that the data is stationary with 99% confidence. This is also confirmed by the p-value which is significantly less than the significance level, i.e., 0.05.

Stationarity can can also be observed in an autocorrelation plot, as described Section 2.3.1, and the autocorrelation plot for our data set is shown in Figure 4.4. We see that the spikes mainly fall in the marked region, meaning that there is no correlation between the series and the lags of itself. The autocorrelation plot also confirms the seasonality of the time series, as we see an increase after 12 months. We want to include lagged values of invoiced amount as explanatory variables in the forecasting task, and we want to choose a lag with high autocorrelation as it is more likely to contribute to accurate predictions. Therefore, we choose a lag of 12 in the forecasting task.

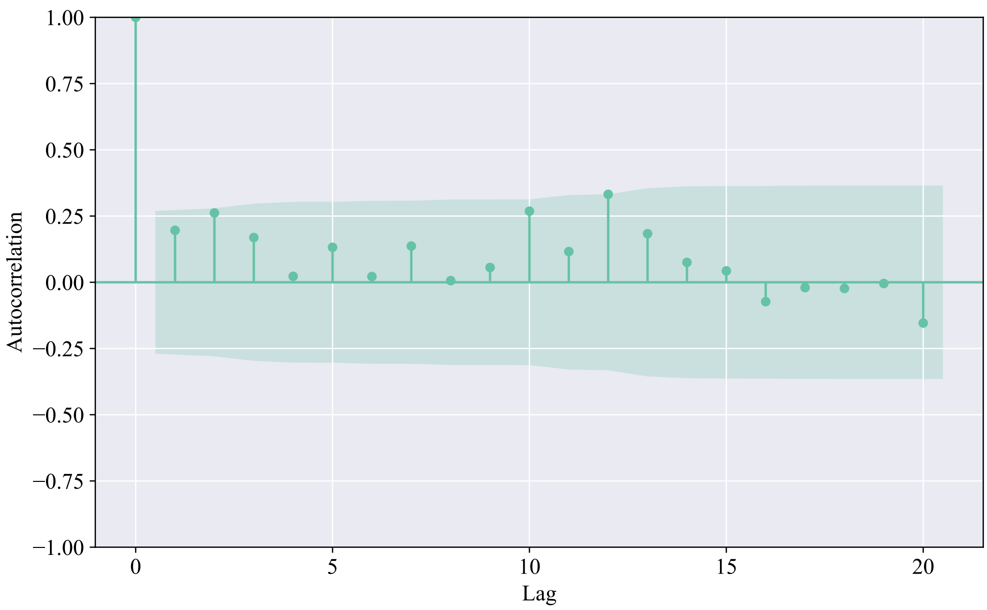


Figure 4.4.: Autocorrelation plot with 20 lags for the invoiced amount data for Customer X, suggesting that the data is stationary and that there is little correlation between the series and the lags of itself.

When forecasting time series, the trade-off between simple, understandable models and more complex black-box models are considered.

Regression is a simple, but valuable technique for time series forecasting as it considers trend, cycle, and seasonality, as well as permitting the inclusion of external factors. To forecast using regression, we identify the relevant explanatory variables for forecasting sales. It seems plausible to assume that sales depend on past sales as well as other variables. It is crucial to strike a balance between including data from our data set and incorporating external factors without making the model overly complex, confer the bias-variance trade-off. In our specific model, we incorporate historical data on the invoiced amount for each company as the variable of interest for forecasting.

To account for the rising trend in the underlying time series, we introduce a variable that increases linearly over time. By incorporating this linear trend into our forecasting model, we aimed to capture the rising population over time and leverage it as an influential factor in forecsating future outcomes.

Furthermore, the output gap is included as an explanatory variable. The output gap is defined as the ratio between the logarithm of the actual mainland gross domestic product (GDP) and the logarithm of the trend mainland GDP, and is an indication of the current state of the economy. As this data is obtained from Statistics Norway and only concerns Norway, our forecasting model can be applied to Norwegian companies only. Hence, we removed companies from other countries than Norway by filtering out non-Norwegian companies from our data set.

When the data was obtained from Statistics Norway's statistics bank [57], the output gap was calculated by applying a Hodrick-Prescott (HP) filter to the GDP data. The HP filter is a data-smoothing technique that removes short-term fluctuations [58]. It separates the trend from the cycle and is defined as the solution to the following optimization problem

$$y_t = \tau_t + c_t$$

$$min_{\{\tau_t\}} \left\{ \sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} \left[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1}) \right]^2 \right\} \tag{4.2}$$

where y_t is the original series, τ_t is the trend component and c_t is the cyclical component [59]. The deviation of the original series from the trend and the curvature of the estimated trend is minimized, and the trade-off between the two parameters are governed by the smoothing parameter λ [60]. A smoothing parameter $\lambda = 100$ is used based on recommendations for yearly data in from [59]. The HP filter provides an estimated GDP trend based on the actual GDP data.

Figure 4.5 illustrates the actual mainland GDP and the estimated trend mainland GDP, where the latter was calculated based on a HP filter. The logarithmic difference between these two represents the output gap. The periods where actual mainland GDP is significantly higher than the trend mainland GDP are periods with a boom, while when the opposite is the case we have a slump. A falling output gap indicates a business downturn and a rising output gap indicates a business upturn. To forecast sales, we extended the forecast period by adding macroeconomic forecasts from Statistics Norway as well as the estimated trend growth in mainland GDP. To achieve this, we calculated the median trend growth based on the last 10 years using our estimated trend mainland GDP data. We chose to extend the time series beyond the forecasting period, taking the sensitivity of the HP filter to endpoints into consideration. Adding more periods will stabilize the filter. Furthermore, the information was utilized to generate mainland

Figure 4.5.: The output gap is the ratio between the actual mainland GDP and the estimated mainland GDP trend, obtained by applying an HP filter to the data.

GDP, trend mainland GDP, and the output gap for an additional year, thus enabling us to forecast sales for a 3-year period beyond 2023.

Customer X's invoiced amount, synthesized population data, and our calculated output gap based on data from Statistics Norway were combined into a single data set with a monthly frequency, covering the available data period and three years ahead. The data from this data set is shown in Figure 4.6. Although our data set includes data from 2019, the forecasting process begins from 2020 onwards. This is because we utilize lagged values in our analysis, which necessitates excluding the data from the first year since we do not have lagged values for invoiced amount available for that period. To improve the accuracy of forecasts, the annual population and output gap data were interpolated to obtain monthly values. An alternative to this method is the DIMAS (Daily, Intraday, Monthly, Annual, Seasonal) method, which is a categorization framework used to classify and analyze time series data based on different frequencies and can provide insights into patterns and variations that occur at various time scales [61]. Possibly, all data could have been used in a annual format, but our data was not sufficient for these methods. The stationarity of the interpolated data sets was verified before proceeding with the forecasting task. The results of the stationarity check are presented in Table 4.5.

Table 4.5 shows the results from the ADF test performed on the two time series. Both are stationary.

We also calculated the Pearson correlation coefficients between the variables, confer Figure 4.7.

The correlation matrix depicts the empirical Pearson correlation coefficients between pairs of variables. The correlation coefficients range from -1 to 1, where values close to -1 indicate a strong negative correlation, values near 0 denote no correlation, and values close to 1 signify a strong positive correlation. In the context of our study, the analysis reveals a strong positive correlation between the invoiced amount and population, suggesting that rising population cause higher sales, which seems plausible. However, the

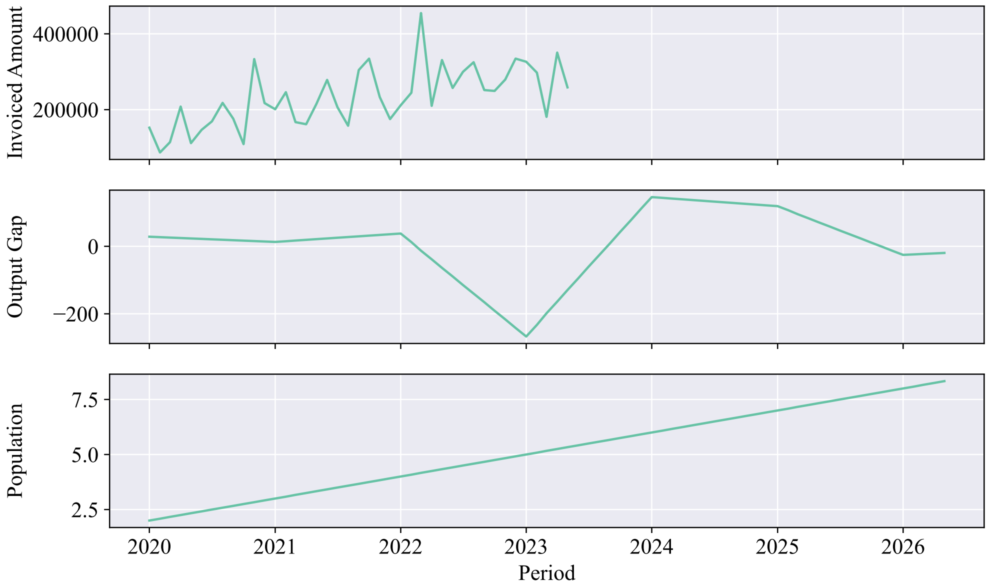


Figure 4.6.: Plot showing invoiced amount for Customer X, output gap and population. These are the variables that are used in the forecasting task. Invoiced amount is the variable to be forecast and is therefore not existing for the entire period.

Table 4.5.: Results from the ADF-test applied on the macroeconomic data from Statistics Norway that are used in the forecasts.

	Test statistic	**p-value**	**Critical values**
Population	-4.184	0.0007	1%: -3.512 5%: -2.897 10%: -2.586
Output gap	-3.553	0.007	1%: -3.508 5%: -2.895 10%: -2.586

correlation between the invoiced amount and output gap exhibits a moderate negative association, indicating a comparatively opposite impact of the output gap on invoiced amount fluctuations. Nevertheless, it is essential to acknowledge that correlation coefficients solely capture linear relationships and may not fully capture e.g., more intricate and non-linear associations between variables or even common third factors. In short, correlation does not imply causation.

We can check for causality with the Granger causality test[62], which is a statistical test used to determine whether one time series is useful in forecasting another. The test is further explained in Appendix B.2.11. The results of the test are presented in Table 4.6.

The causality test is performed at different lag lengths to account for potential time delays in the causal effect. The p-values indicate the statistical significance of the causal relationship at each lag. The null hypothesis is that the lagged values of the potential causal variable do not have a significant effect on the variable being tested. This means that lower p-values suggest stronger evidence against the null hypothesis and indicate a

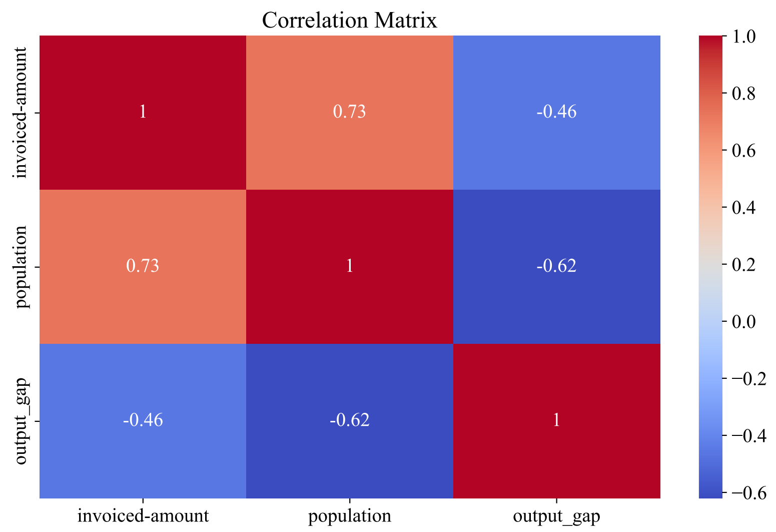


Figure 4.7.: Heatmap showing the correlation between the variables. We see that the invoiced amount is correlated to population.

Table 4.6.: Results from the Granger causality test assessing the causal relationship between the dependent variable and the explanatory variables. The test suggest a causal relationship between invoiced amount and population.

	p-value[Population]	**p-value[Output gap]**
Lag 1	0.000001	0.076124
Lag 2	0.000031	0.095381
Lag 3	0.001201	0.006107
Lag 4	0.047252	0.0037235
Lag 3	0.266637	0.064169

more significant causal relationship. We see that for population, the p-values at the first lags are below the significance level of 0.05, indicating a statistically significant causal relationship between population and invoiced amount. However, when the lag increases, the p-value increases, and the relationship becomes weaker. For the output gap, the p-values are higher, indicating that there is no evidence for a causal relationship between the output gap and the invoiced amount, which are supportive of the results from the correlation matrix in Figure 4.7.

After processing the data, we proceed with the forecasting task by employing various modeling techniques and comparing their performance. The data is divided into a training set and a test set, with the training set accounting for approximately 80% of the data, which is in accordance with what leading practitioners normally use (70-90%). The test set is used to evaluate and select the most suitable model. We explore different regression methods, including linear regression, random forest regression, and XGBoost regression. Linear regression is suitable for forecasting a continuous target variable when the relationships with predictors are approximately linear, offering a simple modeling approach. Random forest and XGBoost regression excel at capturing complex patterns and interactions, even in cases where predictors exhibit weak correlations with the target variable. Additionally, we employ the Autoregressive Integrated Moving Average (ARIMA) model, which is adept at handling stationary time series data and can effectively capture autocorrelation and seasonality. By considering the distinct strengths and limitations of these techniques, we compare and evaluate their performance to identify the most effective model for our forecasting objective.

Model 1: Linear regression For this model, the function `LinearRegression()` from the `scikit -learn` library in Python is employed to execute linear regression modeling and generate forecasts for the identical time period as the test set. The coefficients of the linear regression model are computed using the training data, optimizing the best-fit line that minimizes the sum of squared residuals between the forecast values and the actual values of the invoiced amount. Subsequently, these coefficients are applied to the test data, and the resulting forecasts are illustrated in Figure 4.8a.

We see that the forecast data follows the trend of the historical data. However, there is a significant difference between the range of the forecast data and the actual test data. The coefficients are shown in Table 4.7.

Table 4.7.: Coefficients from the linear regression model that forecasts invoiced amount.

Variable	Value
Output gap	-260.587
Lagged values of invoiced amount	-0.700
Population	94119.085

The population variable has a notable effect on the forecast values of the invoiced amount. The negative coefficient for the output gap implies that a decrease in the output gap is associated with a decrease in the forecast value of the invoiced amount, all other variables held constant. This implies that when the output gap e.g., is strongly positive and we thus have a boom, sales will be lower. That does not make sense, which points to the fact that we have not found the proper underlying drivers of sales. Conversely, the positive coefficient for population indicates that an increase in population corresponds to an increase in the forecast value of the dependent variable. This makes sense, as companies' sales normally increase with a larger population, ceteris paribus. On the other hand, historic sales have a small but positive impact on future sales. This also makes sense, as companies' sales tend to oscillate around a normal level. The substantial population coefficient explains the pronounced upward trend observed in Figure 4.8a.

Model 2: Random Forest regression The function `RandomForestRegressor()` from the `scikit -learn` library in Python is used to perform random forest regression modeling and forecast invoiced amount for the same period as the test set. The method builds an ensemble of decision trees where each tree is trained on a different random subset of the training data. Further, the trees work together to make forecasts by averaging the forecasts made by the individual trees. The trained random forest regression model yields the results shown in Figure 4.8b.

By visual inspection, we see that the forecast data follow the test data very poorly. It does not follow the training data well either, as it is more flattening.

Model 3: XGBoost regression This method utilizes the function `XGBRegressor()` from the `XGBoost` library in Python for regression modeling and forecasting. XGBoost is an ensemble learning algorithm that combines the forecasts of multiple weak decision trees to make accurate forecasts. The model is fitted by optimizing the model parameters to minimize the specified loss function. The trained XGBoost regression model is applied to the test data, and the generated forecasts for the invoiced amount are shown in Figure 4.8c.

The forecast data follows the fluctuations to a certain degree, but does not fit well with the actual test data, indicating that this model does not perform very well.

Model 4: ARIMA The ARIMA model leverages the autoregressive, differencing, and moving average components to capture temporal dependencies and make forecasts for time series data, as described in Section 2.3.1. It estimates the model parameters by fitting the autoregressive and moving average components to the differenced data, and once the model parameters are determined, it can generate forecasts by combining the estimated parameters with the past values of the target variable and forecast errors. It calculates the forecast values by recursively using the model equation and updating the forecast errors based on the observed values. To determine the optimal model parameters, we utilized the auto ARIMA function, which automatically selects the most suitable values for the order (p, d, q) and seasonal components (P, D, Q). The chosen parameters for the ARIMA model were (0, 1, 1) for the non-seasonal order and (0, 1, 1, 12) for the seasonal order, with an Akaike Information Criterion (AIC) value of 417. The results of the ARIMA model with these parameters are illustrated in Figure 4.8d.

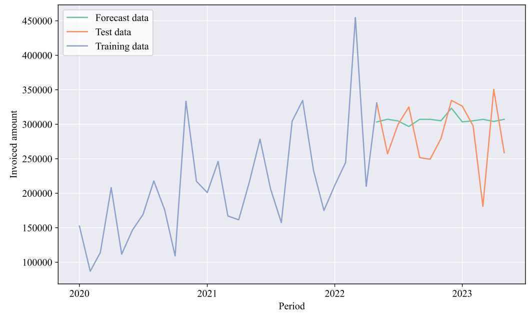
The forecasts on both the training and testing sets demonstrate a satisfactory fit for the data. While the ARIMA model may not capture every single spike in the data, it generally follows the underlying patterns and trends effectively.

Choosing a model for forecasting Based on Figure 4.8, we see that XGBoost is the best-performing regression model, and ARIMA also performs well. To provide further evidence and substantiate these observations, it is essential to evaluate the models using appropriate evaluation metrics.

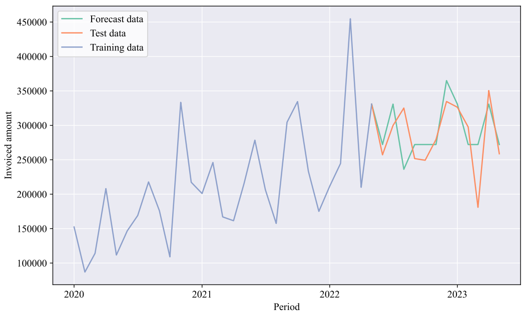
Normally, the F1 score and accuracy score are used to assess models in classification tasks. However, these are more appropriate for tasks where the target variable represents discrete classes. As our data is a continuous time series, R^2, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are more appropriate methods to validate our models. R^2 is the correlation coefficient between the actual and forecast values and thus indicates how well the independent variables (in our case, population and output gap) explain the variation in the dependent variable (in our case, sales). MSE and RMSE provide quantitative measures of how well the model's forecasts align with the actual value. We have also included the results from the Durbin-Watson test as an evaluation metric. The Durbin-Watson test is used to detect the presence of autocorrelation in the residuals of the models [63]. The test statistic ranges between 0 and 4, where 2 is the desired value, indicating no autocorrelation. The scores are shown in Table 4.8.

The regression models evaluated in this study exhibit poor overall performance, with the linear regression model performing the worst. Its high values of MSE and RMSE, along with a negative R^2 score and a low Durbin-Watson test statistic, indicate a weak fit to the data and negative autocorrelation in the model. The random forest model shows relatively better performance but still falls short, with lower evaluation metric values and an improved, yet negative, R^2 score. In contrast, the XGBoost regression model outperforms the other regression models, achieving the lowest values of MSE and RMSE, and a R^2 score of 0.2423, demonstrating that the model captures 24.23% of

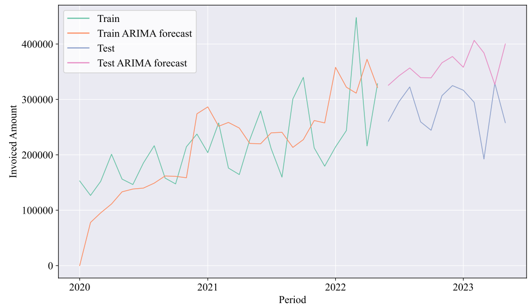
(a) Linear regression



(b) Random forest regression



(c) XGBoost regression



(d) ARIMA

Figure 4.8.: Comparison of forecast sales using different regression models and ARIMA. XGBoost and ARIMA provides the best models for our data.

the variance. The ARIMA model also stands out with a R^2 value of 0.266, slightly higher than XGBoost. However, it performs worse than all models in terms of MSE. Considering these metrics, XGBoost performs relatively better compared to the other models. It has the highest R^2 value among the non-negative values, the lowest MSE and RMSE, making it the preferred choice for the sales forecast. The Durbin-Watson test also indicates some autocorrelation in the XGBoost model, more than for random forest and ARIMA, which means that the residuals of the model are correlated with each other over time. The low overall performance of the models suggests the need for further exploration and modification of the modeling approach, the feature set, the choice of models, and feature engineering (including hyperparameter tuning).

Figure 4.9 illustrates the results of the forecasting task utilizing XGBoost with the entire available data set as the training data. The model forecasts the invoiced amount for a three-year period ahead. Comparing it to Figure 4.8c, we observe that the XGBoost forecast trained on the entire available data set delivers a worse forecast. There may be several reasons for this decrease in performance. Overfitting can occur if the model becomes too closely tailored to the characteristics of the training data and fails to generalize well to new data. Using the data set for hyperparameter tuning before modeling could also have led to data leakage and information from the data set can unintentionally have influenced the model. Therefore, it is generally recommended to separate the data set into training and test sets. However, due to lack of data over a longer period of time, we chose to use the entire data set as training data to perform future forecasts.

Table 4.8.: Evaluation of performance for the three regression models using R^2, MSE, RMSE and the Durbin-Watson test statistic as evaluation metrics. The XG-Boost and ARIMA models have the best overall performances among the models.

	$\mathbf{R^2}$	MSE	RMSE	Durbin-Watson
Linear regression	-1.4100	5.0329e+09	7.0943e+04	0.1120
Random forest regression	-0.1973	2.5002e+09	5.0002e+04	1.9446
XGBoost regression	0.2423	1.5822e+09	3.9777e+04	2.8084
ARIMA	0.266	1.2659e+10	1.1251e+05	2.1362

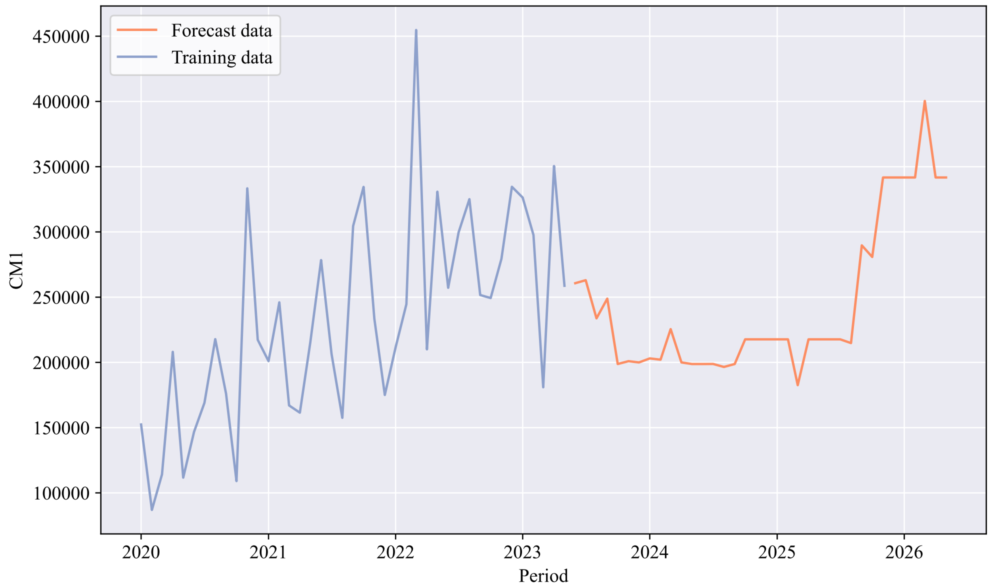


Figure 4.9.: Forecast of future invoiced amount using the XGBoost model. The model does not catch the fluctions well.

4.2.2. Forecasting Profitability

In this section, our objective is to forecast profitability using a similar approach to the one employed in Section 4.2.1 for sales forecasting. However, in this particular task, we focus on forecasting the contribution margin as the dependent variable. The contribution margin represents the revenue remaining after deducting the costs associated with producing a product or service [64]. This metric allows us to assess the extent to which each sale contributes to the company's overall profit. There are three different levels of contribution margin:

- **Contribution Margin 1 (CM1)**: CM1 represents the contribution margin, and it is calculated by subtracting the variable costs tied to production and delivery from sales.
- **Contribution Margin 2 (CM2)**: CM2 is equal to CM1 minus fixed costs which are indirectly associated with the production and delivery.
- **Contribution Margin 3 (CM3)**: CM3 represents the contribution margin after deducting all costs (both variable and fixed), including trouble and delay. CM3 provides a more accurate measure of the profitability.

With the available data with format as presented in Appendix C.1, CM1 will be used in this thesis, and it is calculated as following:

$$CM1 = \frac{Invoiced\ amount - Variable\ costs}{Invoiced\ amount} \tag{4.3}$$

Exploratory analysis

Initially, an exploratory data analysis is performed to understand the data pattern. CM1 for Customer X is calculated using Equation 4.3 and the time series is shown in Figure 4.10.

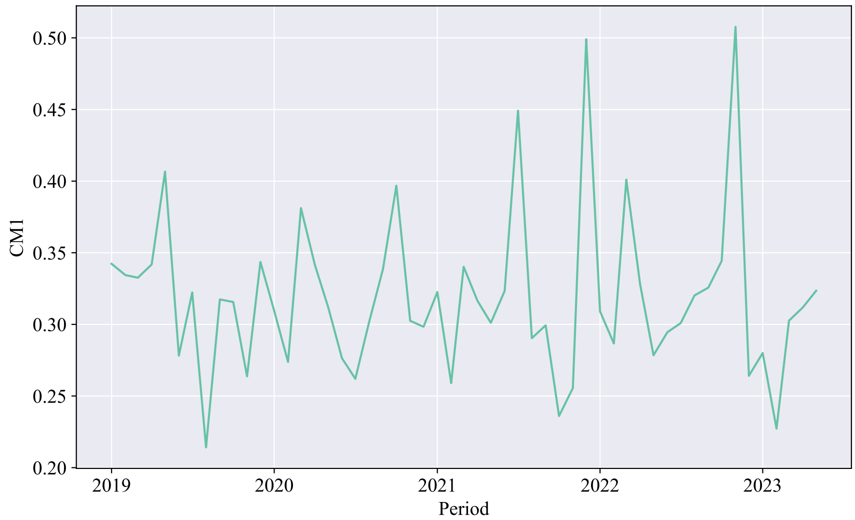


Figure 4.10.: Time plot showing the contibution margin, i.e., the variable to be forecast, for Customer X.

Figure 4.10 shows CM1 for Customer X over a five-year period. The plot reveals a modest upward trend, accompanied by fluctuations centered around the value of approximately 0.3. Figure 4.11 shows the yearly CM1 for the company, and we see that there is a tendency of seasonality at the beginning of the year; however, this seasonality does not persist consistently throughout the year.

As for sales, the Breusch-Pagan test is conducted on the time series to check for heteroscedasticity, and if there is need for adjustments. The null hypothesis assumes constant variance. The results of the Breusch-Pagan test can be found in Table 4.9.

Table 4.9.: Results of formal tests for heteroscedasticity with the Breusch-Pagan test suggesting that the variance in the data is constant over time.

Test	Statistic	p-value
LM-Test	1.465	0.226
F-Test	1.450	0.234

Table 4.9 shows that both the tests does not reject the null hypothesis with a significance level of 0.05, as both p-values are higher than 0.05. Thus, there is not evidence

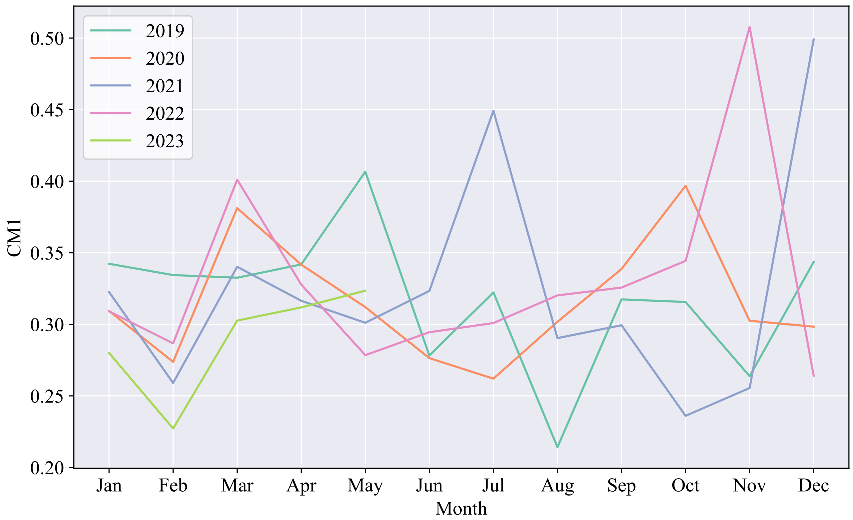


Figure 4.11.: Seasonal time plot showing the yearly contribution for Customer X, indicating a slight tendency of seasonality in the data.

of heteroscedasticity in the data, and the variance in the data is constant over time, meaning that we do not need to adjust the data.

Figure 4.12.: Additive decomposition of the contribution margin time series for Customer X, showing each component of the time series separately.

Figure 4.12 presents the additive decomposition of the CM1 time series, enabling a more comprehensive examination of its components: trend, seasonality, cycle and residuals. The decomposition confirms the presence of a detectable trend and some level of seasonality within the data. The analysis of the time series decomposition reveal that the cycle component and the residual component exhibit notable similarity. This finding implies that the decomposition model employed may not have adequately captured all the underlying patterns and variations present in the data. In particular, the contribution margin data, being more complex due to factors such as changing market dynamics, shifts in consumer behavior or external economic influences, can pose challenges in the forecasting task. However, the p-value from the Ljung-Box test on the residual data is 0.348, indicating that the residuals are white noise.

Implementation

For this particular forecasting task, the variables used for forecasting are population and output gap, which were also employed in Section 4.2.1. The CM1 data, displayed in Figure 4.10, serves as the target variable. It is essential to assess the stationarity of the data, and this is achieved through the ADF test. While the test results for population and output gap are already available in Table 4.5, the ADF test results for CM1 are presented in Table 4.10.

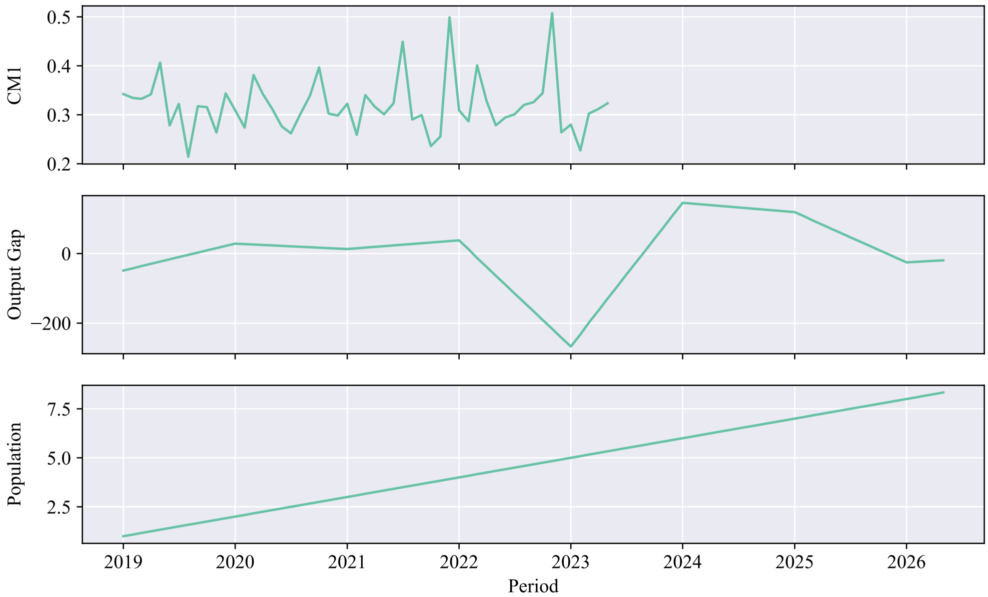


Figure 4.13.: Plot showing contribution margin for Customer X, output gap and population. These are the variables that are used in the forecasting task. Contribution margin is the variable to be forecast and is therefore not existing for the entire period.

Table 4.10.: Results from the ADF test on the contribution margin data for Customer X, indicating that the data is stationary.

Test statistic	p-value	Critical values
-5.342	4.469e-06	1%: -3.571
		5%: -2.922
		10%: -2.599

As is evident from Table 4.10, the contribution margin is stationary. The stationarity is also confirmed in the autocorrelation plot shown in Figure 4.14, where the spikes fall within in the marked area. However, there is not any seasonal pattern, making it more difficult to choose a lag of the contribution margin value to use in the forecasting task. We chose a lag of 12, as for sales, as the contribution margin is related to sales. We confirm that this assumption is reasonable by looking at the twelfth lag in the autocorrelation plot, which is one of the higher autocorrelations among the lags. However, this is negative due to the fluctuations in the data and must be kept in mind.

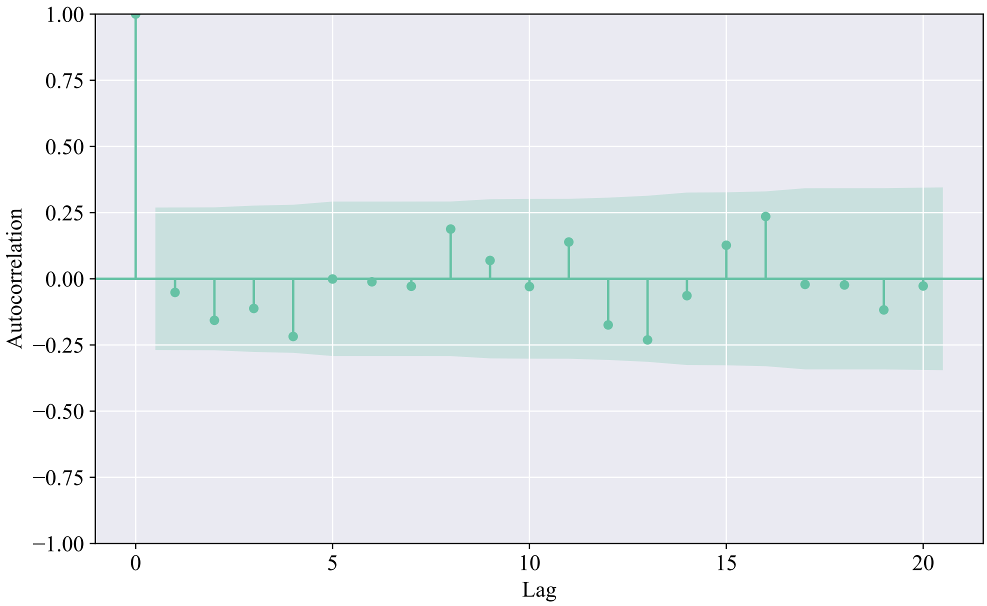


Figure 4.14.: Autocorrelation plot with 20 lags for the contribution margin data for Customer X, suggesting that the data is stationary and that there is little correlation between the series and the lags of itself.

To assess the correlation between the explanatory variables and the dependent variable, we also computed the correlation matrix, depicted in Figure 4.15.

The explanatory variables, population and output gap, exhibit weak correlations with CM1, indicating a lack of linear relationship between each variable and CM1. However, to assess the strength of these relationships more accurately, regression modeling is applied.

Similarly to the sales forecasting task discussed in Section 4.2.1, the CM1 data is divided into a training set and a testing set, and various regression models are employed to identify the most suitable model for the data. We employ the same models as we did when forecasting sales, namely linear regression, random forest regression, XGBoost regression, and ARIMA.

Model 1: Linear regression The performance of the linear regression model can be observed in Figure 4.16a. It is evident that the forecast data exhibits a declining trend, which contrasts with the actual test data. Moreover, the spikes in the actual data surpass the forecast values, indicating a suboptimal performance of the model. The coefficients of the linear regression model are provided in Table 4.11.

The output gap lagged CM1, and the population does not have significant impacts on the forecast values of CM1. The negative coefficient for historic CM1 indicates that positive historic CM1 leads to a decrease in future CM1, assuming all other variables are constant. Furthermore, the coefficients for population and output gap are quite small, indicating their limited explanatory power or perhaps a non-linear relation. However,

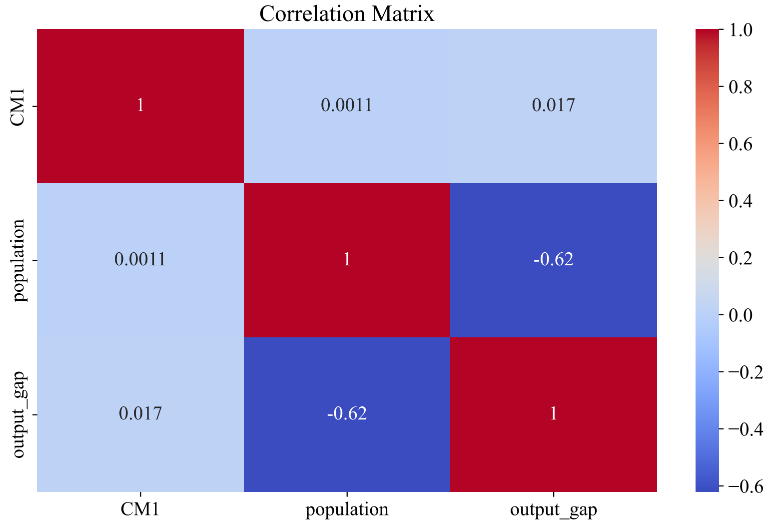


Figure 4.15.: Heatmap showing the correlation between the variables. We see that the invoiced amount is very weakly correlated to the other variables.

Table 4.11.: Coefficients from the linear regression model that forecasts profitability.

Variable	Value
Output gap	0.0002
Lagged values of CM1	-0.0754
Population	0.0083

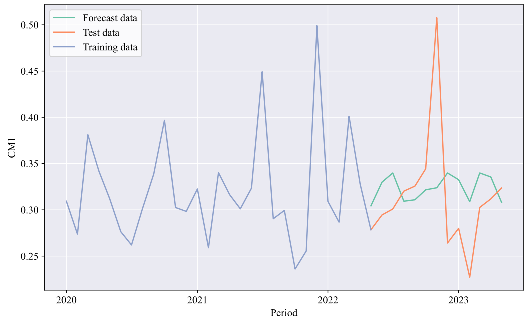
this is expected as we know that there are weak correlations between the dependent variable and the explanatory variables.

Model 2: Random Forest regression The results obtained from the trained random forest regression model are presented in Figure 4.16b. Upon visual examination, it is evident that the forecast data does not closely align with the test data, although it performs better than the linear regression model. The forecast trend appears to be relatively flat and fails to capture the increasing trend observed in the training data.

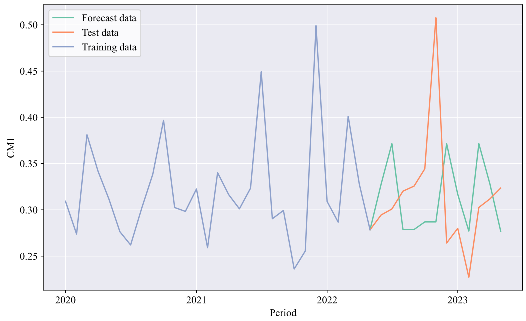
Model 3: XGBoost regression The generated forecasts for CM1 are shown in Figure 4.16c. The forecast data follows the fluctuations to a certain degree, and the height of the spikes is more similar to the actual data compared to the two other regression models. Consequently, the XGBoost regression model appears to outperform the other regression models in terms of capturing the patterns and variations in the CM1 data.

Model 4: ARIMA The Auto ARIMA algorithm determined the optimal order and seasonal order parameters for the ARIMA model as (0,2,2) and (0,1,0,12), respectively, with an AIC value of -14.051, indicating a good fit. The model's performance was evaluated using the training and test sets, and the results are illustrated in Figure 4.16d. Upon visual inspection, it is evident that the forecast values for both the training and test sets closely align with the corresponding actual data. The model successfully captures the spikes in the data and exhibits a consistent pattern throughout the time series.

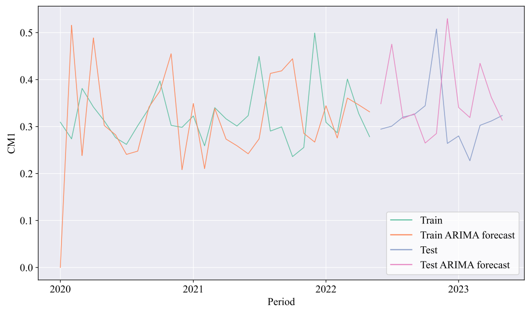
(a) Linear regression



(b) Random forest regression



(c) XGBoost regression



(d) ARIMA

Figure 4.16.: Comparison of forecast contribution margin with different regression models and ARIMA.

The overall performance of the regression models is poor. The linear regression model performs the worst, with high forecast errors and a weak fit to the data. The random forest regression model shows a slightly better performance but still exhibits poor results. The XGBoost regression model, despite visually appearing better, also demonstrates poor performance with higher forecast errors. Considering the evaluation metrics, the random forest model is selected as the most suitable choice among the regression models. However, it is important to note that all models exhibit poor performance, with low R^2 values and relatively high forecast errors. Non-linear regression methods could have been experimented with as they might would allow for more flexibility in capturing complex patterns. Due to the scope of this thesis and the limited data available, we chose to focus on the most commonly used methods.

Figure 4.17 presents the results of the forecasting task using random forest regression with the entire available data for training. The model forecasts CM1 values for the next three years. The model was tuned using `GridSearchCV` which performs an exhaustive search over a provided hyperparameter grid using cross-validation to evaluate the model's performance and provides the best hyperparameters from the search, which were used in our modeling.

However, the forecast CM1 values do not align well with the fluctuations observed in the training data. It is worth noting that this outcome is consistent with the evaluation metrics presented in Table 4.12, which indicates poor performance. Despite this, the forecast values do tend to cluster around the mean value of the observed fluctuations, which could explain the model's relatively better performance during evaluation.

Table 4.12.: Evaluation of performance for the three regression models using R^2, MSE, RMSE and the Durbin-Watson test statistic as evaluation metrics. The random forest regression model has the best overall performances among the models.

	$\mathbf{R^2}$	**MSE**	**RMSE**	**Durbin-Watson**
Linear regression	-0.0547	0.0042	0.0648	1.6319
Random forest regression	0.0110	0.0039	0.0628	1.7904
XGBoost regression	-0.4671	0.0058	0.0765	1.9127
ARIMA	-3.1801	0.0147	0.1216	2.2698

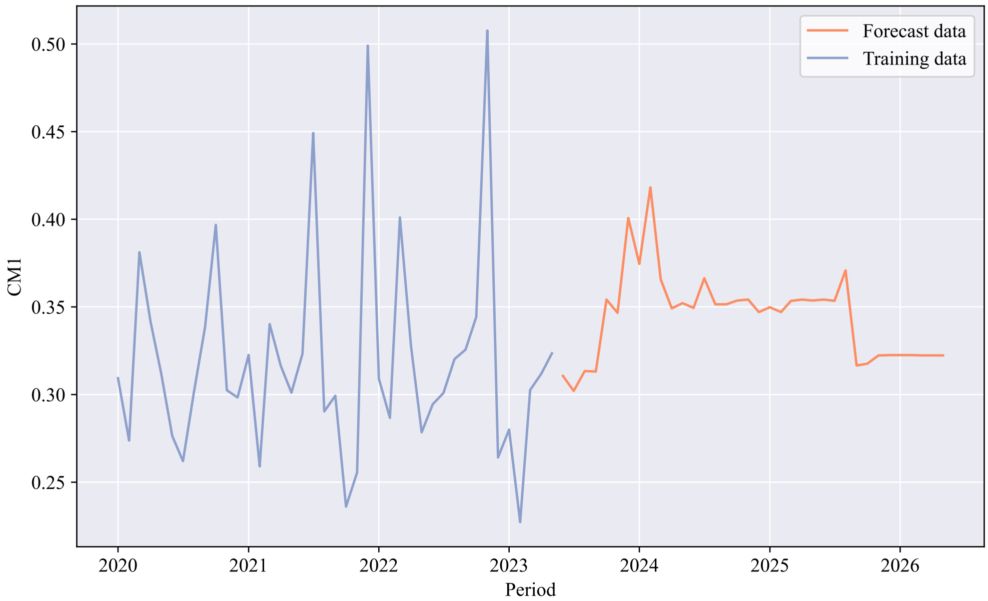


Figure 4.17.: Forecast of future contribution margin using the random forest regression model. The model does not catch the spikes in the data well.

4.2.3. Discussion

Initially, our plan was to conduct a multivariate multiple regression analysis, aiming to develop a comprehensive model trained on the entire data set, including all customers. However, this was challenging due to the diverse nature of the companies involved, representing different industries and having different underlying drivers. The factors describing the different companies often differed. In addition, there were differences when it came to stationarity and other statistical traits. As a result, implementing a single general model that encompassed all customers became a complex task. Nevertheless, for simplicity and ease of interpretation, we decided to employ different forecasting techniques for a specific company. This approach can be replicated for all customers by employing looping techniques and re-estimation procedures. Alternatively, performing a panel data analysis instead of a time series analysis may address this challenge, as it is more appropriate for data sets with observations on multiple entities, e.g., multiple customers, observed over time.

In retrospect, we acknowledge that our initial approach of solely relying on the Augmented Dickey-Fuller (ADF) test to assess stationarity in our time series data for the forecasting task may have been insufficient. While the ADF test provides valuable insights, we recognize the importance of conducting additional tests to confirm the stationarity of the data. The Phillips-Perron (PP) test, which has higher power in detecting autocorrelation in the underlying data, should have been included to check for the presence of autocorrelation. This would have provided a more comprehensive analysis and a better understanding of the data's characteristics. Additionally, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, with its different null hypothesis, could have been used to validate the results obtained from the ADF tests. Performing both the KPSS and ADF tests would have allowed us to obtain consistent results indicating stationarity or non-stationarity or divergent results pointing to either strictly stationarity or difference stationarity. Regrettably, we did not perform the KWSS and PP tests in our analysis due to time constraints and a lack of awareness regarding the potential benefits they could have provided. This limitation should be acknowledged, and in future research, we recommend incorporating these tests to ensure a more thorough assessment of stationarity in time series data.

Considering the limitations of the models and the relatively high forecast errors, it may be necessary to add internal and external data, measure data quality, explore alternative modeling techniques, and/or reconsider the choice of features and feature engineering to improve the models' forecast performance. Our chosen techniques of linear regression, random forest regression, XGBoost regression, and ARIMA performed sub-optimally in the forecasting tasks. The performance issues could be related to several factors. First, the linear regression models assumed a linear relationship between the predictors and the target variable, which might not accurately capture the complexities and non-linearities present in the data. Random forest regression and XGBoost regression, although more flexible, might not have effectively captured the specific dynamics and patterns within the time series data. Alternatively, we could have explored other techniques that could potentially yield better results. For instance, exponential smoothing methods, more complex ARIMA models, vector autoregression models (VAR), and machine learning methods. However, due to data and time limitations and other constraints, we had to narrow down the number of models to be used, leading us to choose linear regression, random forest regression, XGBoost regression, and ARIMA. Despite their limitations, these models were selected based on their availability, familiarity, and potential suitability for our forecasting task.

One interesting observation is the discrepancy between the random forest regression model and the linear regression and XGBoost regression models. Despite both random forest and XGBoost utilizing decision trees as base models, their performance differs. This discrepancy can be attributed to various factors, including their distinct training approaches, handling of gradients and residuals, hyperparameter tuning, and treatment of feature importance. While it would have been intriguing to investigate this anomaly further, time constraints necessitated accepting these differences as they are.

When forecasting, it is essential to recognize the inherent errors and uncertainty associated with the used estimators, as they, e.g., rely on data, the underlying data-generating process, assumptions, model specifications, and parameter estimates that may not capture the underlying dynamics of the data. Uncertainties can arise from missing or incomplete, or erroneous data, model specification choices, assumptions about the underlying data-generating process, and the unpredictable nature of future events. Therefore, it is important to interpret forecasting results carefully and be aware of the fact that they are subject to errors and uncertainties.

When it came to choosing the best model for our forecasting tasks, it was challenging as all four models, in general, did not perform well. Each model exhibited contrasting strengths and weaknesses, making it difficult to determine which aspect should be given the highest priority. While some models had high values on the evaluation metrics and on the R^2 score, others had low scores on both. The optimal, good-fitting model should have a high R^2 score and low errors. This conflicting nature of evaluation metrics, R^2 scores and Durbin-Watson statistics with our models posed a dilemma regarding which factors to prioritize in the model selection process. We needed to consider the forecasting objectives, our data, and the overall trade-offs between accuracy and model fit, resulting in choosing XGBoost for forecasting sales and random forest regression for forecasting profitability. We chose different models for the two forecasting tasks, despite the similarity between the two forecasting tasks. This can be due to the specific characteristics of each task, as the key distinction lies in the dependent variable, which was sales and CM1, respectively.

4.3. Exposure

In order to incorporate the concept of exposure into the calculation of customer lifetime value (CLV), it is essential to determine two types of exposures: gross exposure and net exposure.

Gross exposure represents the outstanding sales credit that a customer has accumulated over time, indicating the amount that has not yet been repaid. It serves as an estimate of the potential gross loss for the company in the event of customer default.

To calculate the gross exposure at default, we aggregate the adjusted lent amounts at a specific time (denoted as T). The credit score of the customer acts as a proxy for the probability of default. By considering the adjusted lent amounts and the probability of default, we can determine the gross exposure.

However, for a more accurate assessment of the company's risk, it is important to calculate the net exposure. Net exposure takes into account various risk-mitigating factors, such as bank guarantees, sureties, credit insurance, collaterals, mortgage security, and credit insurance premiums. By subtracting these factors from the gross exposure, we obtain an estimate of the net loss for the company in the event of customer default.

Given the limited availability of information regarding certain risk-mitigating factors, we rely on gross exposure as a proxy for assessing risk. It is important to acknowledge the limitations of this approach and the potential impact on the accuracy of the results.

In our analysis, we utilize the formula for gross exposure, which is calculated as follows:

$$\text{Gross exposure} = \text{forecast sales} \times \text{c}, \tag{4.4}$$

Here, the scale factor c, provided by Customer X, is equal to 0.25. This factor is used to estimate the gross exposure based on the forecast sales.

Figure 4.18 provides a visual representation of the gross exposure over the desired time frame for calculating CLV.

4.4. Survival Analysis

Survival analysis is a statistical technique that combines classification and regression to predict events such as customer default or termination of business with a company. The goal is to identify the variables that are most influential and hence can be indicators of customer default or attrition.

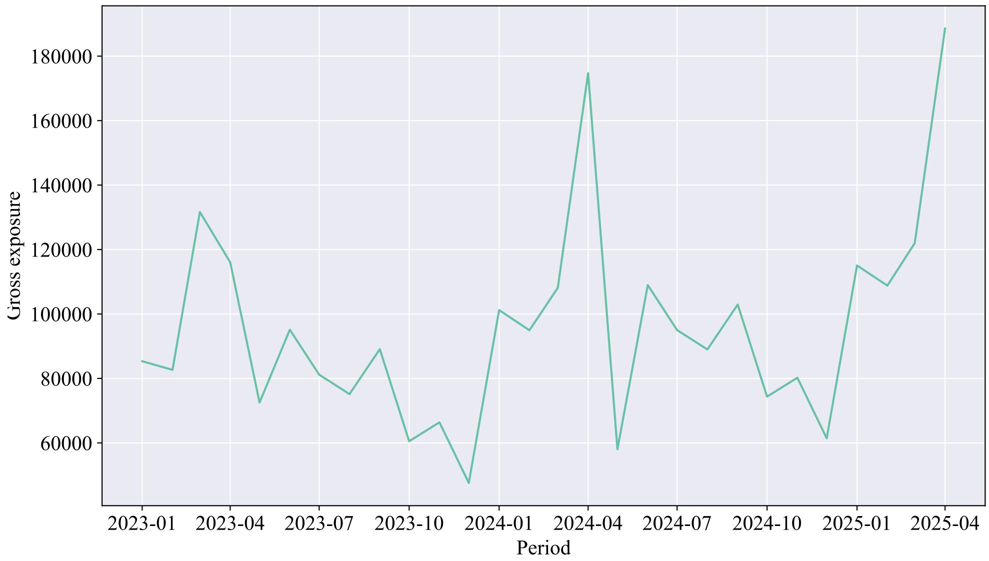


Figure 4.18.: Provided gross exposure for Customer X shown as a time plot.

While machine learning methods can be used to forecast the time until customer default based on features or predictors like customer demographics, behavior, and transaction history, they cannot handle censored data effectively. Survival analysis, on the other hand, is specifically designed to handle censored data and covariates, making it more suitable for calculating the time until default for a customer.

4.4.1. Data Exploration

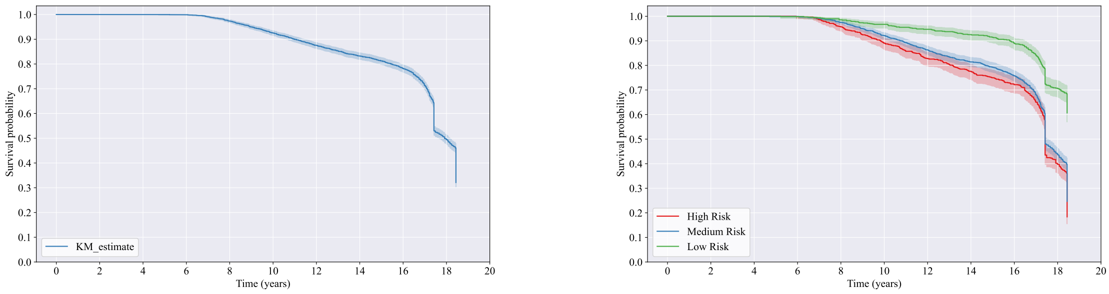
Before delving into the survival analysis, it is essential to explore the data set to gain insights into the default patterns and the nature of the data. The provided data set has 15848 unique customers, and the overall default rate is 42.7%, meaning that 57.3% of the data is censored. The data is transformed into the duration of the customer as well as if they are censored or not.

Exploratory variables, such as country, size, and risk segment, can enhance the estimation of customer duration in the survival analysis. The risk segment is a simplified grouping of the credit score, where a higher risk segment corresponds to a higher chance of default. The data set includes categorical variables such as country, risk segment, and invoice reminder degree. Customers with null values are removed from the data set. For the continent analysis, the data is consolidated into eight continents, with Africa and Oceania grouped as "Other." Scandinavia is treated separately from Europe due to its substantial representation. Table 4.13 provides an overview of the counts of countries by continent in the data set. It is worth noting that the number of countries varies significantly across continents due to the limited sample size.

The Kaplan-Meier curve from `lifelines` library in Python was utilized to get an indication of the survival probability. As depicted in Figure 4.19a, the curve demonstrates that beyond 14 years, approximately 70% of the customers remain active. The risk groups are plotted in Figure 4.19b, and as is assumed, the high-risk customer group has a faster decrease in survival probability than the medium and low-risk customers. While the Kaplan-Meier estimate provides a valuable indication of data survival rates, it is important to acknowledge its limitations. Specifically, the assumptions that subjects

Table 4.13.: Counts of countries by continent among the customers in the data set. The majority of customers are Scandinavian.

Continent	Defaults	Count
America	33	112
Asia	63	235
Europe	248	782
Scandinavia	6365	14503
Other	9	33



(a) [The Kaplan-Meier curve for survival based on the data set.

(b) The Kaplan-Meier curve for the three risk groups.

Figure 4.19.: Kaplan-Meier curves illustrating the overall survival and risk segments. The lower-risk segment demonstrates a higher probability of survival over an extended period compared to the high-risk segment.

who are censored have the same survival prospects as those who are still being monitored and that survival probability is uniform for all subjects restrict its ability to provide a precise estimation of the duration until default for credit customers. This is because the customer has several valuable features, both time-independent and time-dependent, that contribute to the survival rate, which we will account for in the Cox proportional hazard model.

4.4.2. Implementation

The data set is split into an in-sample training set (60%) and a post-out-of-sample test set (40%) [65]. The Cox proportional hazards (Cox-PH) model is estimated, and a Weibull accelerated failure time (WeibullAFT) model is also employed for comparison purposes. The Cox-PH model includes time-dependent variables such as mainland GDP to capture their impact on the hazard function. This model compares the current covariate values of an individual who experienced default to those of individuals at risk of default at that time [66].

From equation 2.15, coefficients $b_1, b_2, ..., b_n$ measure the impact of the covariates, and a change in these will either increase or decrease the baseline hazard. A positive sign means that the risk of an event is increasing, and thus the probability of an event occurring is higher. The magnitude determines if the coefficient will increase or decrease the hazard based on if they are larger or smaller than one.

Regularization The Cox proportional hazards models with varying penalizer and $l1$ ratio combinations provide valuable insights into their performance. The penalizer controls the level of regularization applied to the model coefficients, while the $l1$ determines the balance between Lasso and Ridge regularization. Lasso (L1 regularization) promotes sparsity by penalizing the absolute values of coefficients, whereas Ridge (L2 regularization) penalizes the squared magnitudes.

Applying Lasso and Ridge regularization to the Cox model improved its performance and behavior. Lasso helps identify important features by setting some coefficients to zero, enhancing interpretability, and preventing overfitting. Ridge regularization encourages smaller, evenly distributed coefficients, reducing the impact of individual features and addressing multicollinearity.

4.4.3. Testing

We use four different accuracy criteria to select the best model. The concordance index (C-index), ranging from 0 to 1, measures the accuracy of survival models, with 1 indicating a perfect model and 0 representing a completely random sample. The Akaike Information Criterion (AIC) compares the goodness of fit with lower AIC values indicating a better fit to the data. Additionally, the log-likelihood ratio and the accompanying LL test provide insights into the significance of the model.

Table 4.14.: Summary of the Cox-PH model results with different regularization settings, suggesting that a lower penalizer yields better model performance.

Penalizer	$l1$ Ratio	Model	C-index	AIC	LL-ratio	$-log2(p)$
0.01	0.5	Cox PH	0.75	16028.89	791.03 (14 df)	528.3
0.01	0.0	Cox PH	0.75	16036.31	783.61 (14 df)	523.04
0.1	0.0	Cox PH	0.72	16431.68	388.24 (14 df)	243.89
0.1	1.0	Cox PH	0.71	16735.27	84.64 (14 df)	37.92
-	-	WeibullAFT	0.75	21078.21	696.01 (14 df)	460.87

The results of the Cox-PH model with different regularization settings are summarized in Table 4.14. Models with a penalizer of 0.01 consistently exhibit good performance, as indicated by high concordance indices (0.75) reflecting accurate risk ranking. The log-likelihood ratio test also supports the significance of these models, showing significant improvements compared to the null model. On the other hand, models with a penalizer of 0.1 show slightly lower concordance indices (0.72-0.71) and weaker improvements in likelihood. This suggests that a lower penalizer (0.01) leads to better model performance.

Comparing the models with a penalizer of 0.01, the $l1$ ratio does not significantly impact the concordance index or likelihood improvements. Similarly, for models with a penalizer of 0.1, the $l1$ ratio does not substantially affect performance. However, both penalizer settings show higher partial AIC values, indicating a relatively worse model fit compared to the models with a penalizer of 0.01. This suggests that fine-tuning the regularization parameters could further improve the models. The WeibullAFT model shows a high C-index. However, it shows a bad fit from the AIC value. The superior performance of the Cox-PH model with regularization could be attributed to the properties of ℓ_2 regularization, such as its ability to effectively control for overfitting and handle multicollinearity among predictors, resulting in improved forecast accuracy and model fit.

The Cox time-varying model reveals that the GDP variable has a statistically significant impact on the event of interest. With a coefficient of -0.06, each unit increase in GDP corresponds to a 6% decrease in the hazard (risk) of the event. The precise estimation is indicated by the small standard error of 0.00. The confidence interval of the coefficient (-0.07 to -0.05) further supports the statistical significance of the effect. The hazard ratio, calculated as the exponentiated coefficient, is 0.94. The low-risk variable shows a significant negative relationship (-0.45, CI: -0.70 to -0.19), while the medium-risk variable shows a significant positive relationship (0.30, CI: 0.04 to 0.55). The high-risk variable is not statistically significant (0.16, CI: -0.10 to 0.42). Other predictors, such as invoice reminder degree and gross exposure, are not statistically significant as their confidence intervals include zero.

Figure 4.20 shows a plot of the confidence bands displaying the estimated coefficients and their associated uncertainty. Each vertical line represents a coefficient, and the width of the shaded region around it represents the confidence interval. A wider shaded region indicates higher uncertainty in the estimate. By examining the plot, we can assess the statistical significance of each coefficient based on whether the confidence interval includes zero.

The lower the partial log-likelihood, the better the fit, with a value of -56458.60 obtained in this model. The AIC value of 112919.20 without other covariates and 110174.72, including the covariates suggests that the covariates contribute to a better model. The log-likelihood ratio test confirms the significance of the model. With a test statistic of 82.24 on 1 degree of freedom and a p-value < 0.005, there is strong evidence of improved fit compared to the null model. The -log2(p) value of 62.85 further emphasizes the high level of statistical significance.

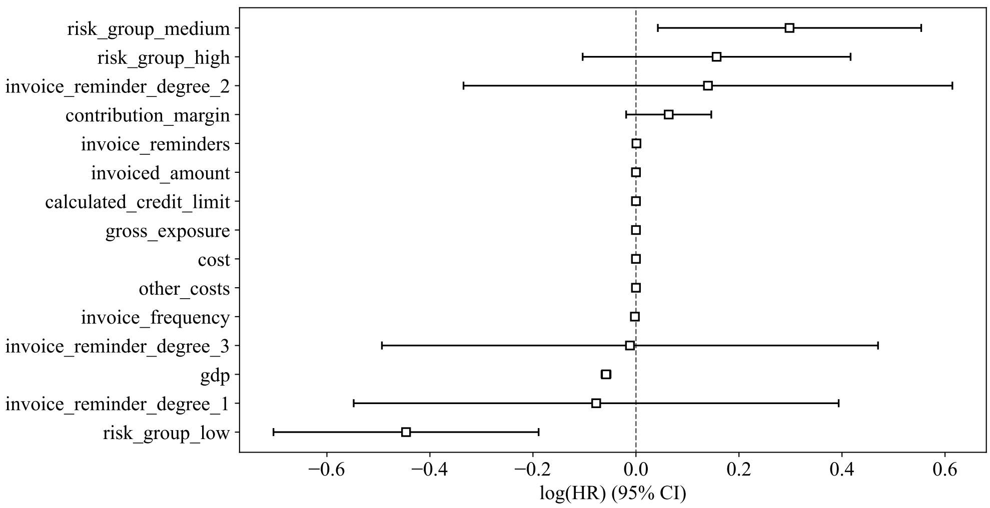


Figure 4.20.: Confidence bands displaying the estimated coefficients and their associated uncertainty, indicating a contribution to increased or decreased risk of default.

Figure 4.21 displays the hazard function for Customer X, comparing the Kaplan-Meier estimate and WeibullAFT estimate. The Cox Proportional Hazards (CPH) model closely follows the Kaplan-Meier curve, capturing the overall survival probability trend. However, Customer X exhibits a distinct pattern, with a higher survival probability than the general estimate until the 18-year mark. The WeibullAFT estimate shows an even higher survival probability between 8 and 14 years, but after 11 years, it decays exponentially. This disparity indicates that the WeibullAFT model fails to capture the nuanced patterns in the data and emphasizes the limitations of assuming an underlying Weibull distribution. These findings underscore the importance of individual customer characteristics in determining long-term survival probabilities. Additionally, it highlights the correlation between credit score and survival probability, as Customer X is categorized as a low-risk customer.

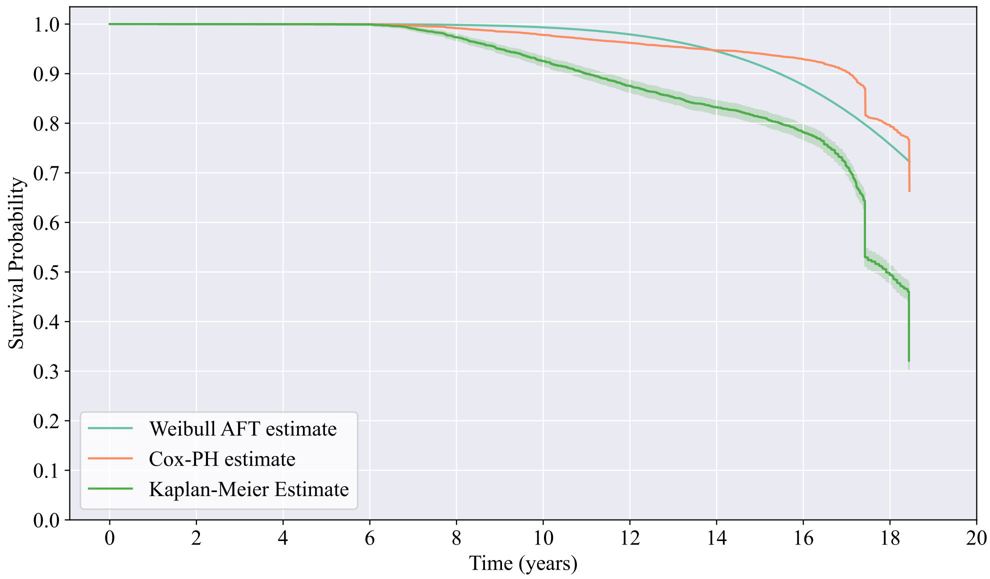


Figure 4.21.: Comparison of the survival probability for Customer X with three models: Kaplan-Meier, C-PH, and WeibullAFT. Customer X exhibits a significantly lower probability of default compared to the average survival rate captured by the models.

4.4.4. Discussion

The Cox proportional hazard (Cox-PH) model with regularization proves to be a suitable choice for forecasting the time until default for credit customers. The regularization technique effectively controls overfitting and handles multicollinearity among predictors, resulting in improved forecast accuracy and model fit. The time-independent model performed significantly better than the time-dependent model.

By analyzing the coefficients from the Cox-PH model with different regularization settings, we gain insights into the significance of covariates. The low-risk group and a higher degree of invoice reminders contribute to a decreased risk of default, while the medium-risk group and higher contribution margin contribute to an increased risk. The impact of credit score is also evident, as lower credit scores are associated with a higher risk of default. However, the coefficients for the high-risk group do not show significant influence, suggesting that other factors may play a more prominent role in determining default risk. Additionally, the location of companies emerges as an important factor influencing default risk. The analysis reveals that the location of the companies impacts their default risk, indicating the potential influence of regional economic conditions and market dynamics on the likelihood of default.

The inclusion of GDP mainland as a covariate in the survival analysis yields interesting results. Higher GDP is associated with a decreased risk of default, indicating that favorable economic conditions and a robust business environment contribute to lower default probabilities. The significant impact of GDP highlights the importance of considering macroeconomic factors when assessing credit risk. By incorporating GDP as a time-varying covariate, the model captures the dynamic nature of economic conditions and provides valuable insights for credit assessment and risk management.

Additionally, the inclusion of additional covariates, such as company size and years since inception, could further enhance the forecast capabilities of the model by capturing additional dimensions of credit risk.

In conclusion, the Cox-PH model with regularization, along with the incorporation of mainland GDP as a time-varying covariate, provides valuable insights into the time until default for credit customers. The findings highlight the importance of GDP and risk group in forecasting the hazard of default. These findings contribute to the calculation of forecast customer lifetime value and facilitate informed decision-making in credit assessment.

4.5. Risk

As addressed in Section 2.3.4, we need to establish a robust measure of the risk-adjusted discount rate tied to the customers of the enterprise.

Our modified risk-adjusted discount rate (RADR) approach

We argue that the relevant risks tied to the discount rate may be summarized into the following five components:

1. The risk-free rate plus a small risk premium (which we have set equal to the 3-month Nibor in this case): all companies must pay a risk premium on top of the central bank's key rate.

2. An idiosyncratic risk and liquidity premium for customer i at time t, which we for the sake of simplicity and to protect our collaborators from the private sector, have set to 2 percentage points in this thesis: different companies have different risks due to its profit and loss account, its balance sheet, etc.

3. A risk-adjustment of RADR for the credit score: if e.g., the customer's credit score is 100%, which implies that the customer will almost surely go bankrupt, the numerator will be 0 in an intertemporal perspective.

4. A risk-adjustment of RADR of the probability of customer churn (= 1-probability of customer retention). Clearly, if a customer leaves the company, the cash flow from that customer would drop to zero.

5. A risk adjustment of the shortfall or downfall risk. If the customer decides to buy less from the company, the cash flow would drop.

6. A risk-adjustment of the net exposure. If the customer has high net exposure and defaults, it will hit the company hard.

Even though our micro version of Ross' [67] arbitrage pricing theory below also includes a risk adjustment of RADR for the credit score, the risk of customer churn, shortfall/downfall risk, and net exposure, we have decided to exclude the calculations tied to the customer churn, shortfall risk, and exposure to protect the dynamic solution of our private sector collaborator. Our model for the risk-adjusted discount rate for customer i at time t is as follows:

$$E\left[r_{it}\right] = r_f + p_{it} + \beta_{1it} \cdot \pi_{1it} \cdot 3MNibor + \beta_{2it} \cdot \pi_{2it} \cdot 3MNibor + \beta_{3it} \cdot \pi_{3it} \cdot 3MNibor + \beta_{4it} \cdot \pi_{4it} \cdot 3MNibor, \tag{4.5}$$

where $E\left[r_{it}\right]$ is the risk-adjusted discount rate for customer i at time t, r_f is the risk-free rate of return, p_{it} represents the idiosyncratic risk and liquidity premium for customer i at time t (set to 2 percentage points in this thesis), β_{1it} is the credit score beta factor for customer i at time t, π_{1it} is the credit score risk factor for customer i at time t, β_{2it} is the customer churn beta factor for customer i at time t, π_{2it} is the customer churn risk factor for customer i at time t, β_{3it} is the shortfall risk beta factor for customer i at time t, π_{3it} is the shortfall risk factor for customer i at time t, and β_{4it} is the net exposure beta factor for customer i at time t, while π_{4it} is the net exposure risk factor for customer i at time t.

We set $r_f + p_{it} = 3MNibor$ (3-month interbank offered rate in the Norwegian money market) in this thesis, which is constant, partly for the sake of simplicity and partly to protect the dynamic solution of our private sector collaborator.

In our calculation, we have used the following simplified version of our risk-adjusted discount rate for customer i at time t:

$$E\left[r_{it}\right] = 3MNibor + 2percentage\ points + \pi_{creditscore,i_t} \cdot (3MNibor). \tag{4.6}$$

We will not perform forecasting on the RADR at every time window for all companies, even though it is relatively straightforward to do so. In short, we have chosen to use the last credit score in our calculations of the RADR; see the simplified formula above. This implies that we use a constant RADR, but the principles underlying our CLV calculation still hold.

4.6. Calculation of CLV

The calculation of the Customer lifetime value (CLV) considers all its components as time-dependent entities. By formulating each component as a time series, we ensure a more dynamic representation of the CLV. This approach enables us to compute either the cumulative CLV across one or more customers or to determine the CLV at a specific point in time.

We have considered two scenarios for our analysis: the customer either stays with the company or the customer goes bankrupt. In the case where the customer remains active, we calculate the CLV at a chosen time T, using the following formula:

$$CLV_t = \sum_{t=0}^{E_t(T_t)} \frac{E_t\left[CF_t\right]}{(1+k_t)^t} \tag{4.7}$$

Empirical results were obtained by forecasting the CLV for Customer X in this thesis until 2025-04-01, which yielded a value of NOK 34,650,604.

In the case of customer default, assuming a scenario where default occurs at time 2025-05-01, the expected loss is incorporated into the calculation at the time of default T. When a customer defaults, the company faces the risk of not receiving the full amount owed, resulting in a loss. By subtracting the expected loss from the cash flows in the CLV calculation, the adjustment for this potential loss is made. The CLV is then determined using the following formula:

$$CLV_t = \sum_{t=0}^{E_t(T_t)} \frac{E_t\left[CF_t\right] - E_{T|default}\left[r \cdot D_T\right]}{(1+k_t)^t} \tag{4.8}$$

The calculation resulted in a value of NOK 31,704,238 for default at the time 2025-05-01. This means that the CLV of customer X is reduced by 2,946,366 when they default, reflecting the potential gross exposure. These results demonstrate the potential reduction in revenue caused by a customer default. To improve accuracy, the inclusion of trouble and delay would be interesting as it would reflect the time and resources spent on handling the customer. However, difficulties arise due to the lack of necessary data and measuring techniques among companies.

The pseudocode for the CLV algorithm is given as follows:

Algorithm 1: Calculation of customer lifetime value (CLV).

Require: CustomerData, SurvivalModel, SalesForecastingModel, ProfitabilityForecastingModel, NetExposureForecastingModel, RiskAdjustedDiscountFactor

Ensure: CLV

```
Load CustomerData from .csv file
Initialize CLV to 0
for each customer in CustomerData do
    Calculate survival_time using SurvivalModel
    Forecast future_sales for the customer using SalesForecastingModel
    Forecast future_profitability for the customer using ProfitabilityForecastingModel
    Calculate future_cash flow for the customer using the product of
    SalesForecastingModel and ProfitabilityForecastingModel
    Forecast net_exposure for the customer using NetExposureForecastingModel
    Calculate present_value of future cash flows (cashflow_forecast-net_exposure)
    discounted by (1 + RiskAdjustedDiscountFactor) raised to the power of
    survival_time
    Add present_value to CLV
end for
return CLV
```

4.7. Discussion

In this section, we will discuss the main findings, limitations, and challenges encountered during the forecasting of customer lifetime value (CLV) in our thesis. We will also reflect on the difficulties faced and provide recommendations for future research.

One of the main challenges we faced was the quality and composition of the data used for CLV calculation. The data we received did not align very well with our desired variables, and there were limitations in terms of data accessibility and timeliness. Noisy data made it difficult to accurately decompose the time series, which affected the accuracy of our model results. To improve future CLV forecasting, it is crucial to enhance data accessibility, quality, and enrichment, considering additional data sources such as Enin. Improving data accessibility and quality is crucial to enhance the accuracy and reliability of CLV calculations.

The selection of XGBoost demonstrates superior performance, as evidenced both visually and through statistical evaluation. While machine learning models can effectively forecast customer behavior, they are often considered black-box models, which may pose challenges when decision-making requires transparency. Furthermore, we aimed to include the measurement of trouble and delay as profitability factors, but practical measures for these factors were limited. Future research should focus on developing accurate measurement procedures and definitions in this area.

Accurately calculating CLV is challenging due to the inherent uncertainty and accumulated error in the models. Forecasting human behavior accurately is difficult, as humans are inherently unpredictable. To address this challenge, further investigation into covariance matrices is recommended to capture the uncertainty in customer behavior better. Additionally, integrating state space models, such as Kalman filters, into the CLV calculation can enhance accuracy by capturing the impact of booms and slumps and linking them to customer actions. Future research should explore the application of these models in CLV calculations.

Calculating CLV for the two scenarios highlighted the impact of gross exposure continues if the case of customer defaults. When a customer defaults, the expected loss reduces the cash flows and, consequently, the overall CLV. The magnitude of this reduction depends on factors such as the credit risk associated with the customer and the amount of gross exposure. Customers with higher credit risk and larger gross exposure are more likely to result in significant reductions in CLV when default occurs. We clearly see that incorporating measures that detect when a customer is headed towards default should feed back to credit limit reductions and more alerted follow-up. We hence stress CLV as a valuable metric in a dynamic credit management system. Incorporating automated credit limits will further enhance the value chain, which is an interesting direction for future work.

While our CLV calculation contributes valuable insights into credit management processes, it also has limitations. We used gross exposure instead of considering factors such as collateral, bank guarantees, and credit insurance that could mitigate losses in case of default. Incorporating additional data and advanced modeling techniques to estimate net exposure accurately would address this limitation. Furthermore, the static treatment of credit scores is another limitation, as they fluctuate over time. Future research should explore dynamic credit score modeling to improve the accuracy of CLV estimation.

4.8. Summary of Findings

XGBoost and random forest regression were used for forecasting sales and profitability, respectively, after testing with different models on our data. During the forecasting process, we faced challenges due to the diverse nature of the companies involved, making it difficult to develop a single comprehensive model. The chosen models for testing, including linear regression, random forest regression, XGBoost regression, and ARIMA, did not perform well, possibly due to their limitations in capturing the complexities of the data. Alternative modeling techniques could have been explored, but several constraints limited our selection. Choosing the best model was challenging, as each had strengths and weaknesses, and evaluation metrics, R^2 scores and Durbin-Watson statistics provided conflicting information. Based on the forecasting analysis conducted, it became evident that solely relying on only two features, output gap and population, might not be sufficient to capture the necessary patterns and trends from the data. As a result, the performance of the model was relatively poor. It is important to acknowledge that there are numerous methods and models available that can be applied to address this problem, and our choice is meant to provide a starting point and an illustrative example.

Regarding modeling customer duration, the Cox proportional hazards (Cox-PH) survival model outperformed the Weibull accelerated failure time (WeibullAFT) model due to its higher concordance index and ability to incorporate time-varying covariates. The analysis revealed significant findings, such as the association of the low-risk group and low amount of invoice reminders with a decreased risk of default. On the other hand, the medium-risk group and higher contribution margin were associated with an increased risk. Factors like credit score, location, and mainland GDP also played significant roles in default risk. However, incorporating additional variables such as company size, mean trouble and delay, and public affairs could further enhance the accuracy of the model.

The calculation of customer lifetime value (CLV) incorporated forecasting of sales and profitability, exposure, risk-adjusted discount factor, and survival analysis. The results highlighted the difference in CLV as a consequence of customer default, emphasizing the importance of incorporating risk and profitability measurements in credit management.

To improve the performance of the models, we suggest incorporating a multivariate multiple time-series forecast and utilizing a state space approach with a Kalman filter. Furthermore, having access to more comprehensive data, including company-specific information, competitor analysis, and customer segmentation, would enhance the accuracy of the forecasting results. Investing time in model development and refinement, such as model tuning, and conducting further research on accurate measurement procedures for trouble and delay, would provide valuable insights and improve the accuracy of CLV calculations.

5. Conclusions and Further Work

5.1. Conclusion

This thesis adopts a holistic, transdisciplinary, and metadisciplinary approach, to conduct an in-depth exploration of the current state of credit management in Norway and propose methods and models to improve it.

- The survey results provide a comprehensive analysis of the current state of credit management across industries in Norway. The findings reveal a significant weakness in credit and analytical maturity among most industries, except for the banking and finance sector and, to some extent, the transportation sector. Through empirical analysis, our initial hypothesis of no significant difference in credit and analytical maturity between banking and finance and other sectors is formally rejected. The banking and finance industry exhibits a notably higher degree of credit and analytical maturity compared to other industries. This conclusion is supported by the survey results and corroborated by in-depth interviews with leading industry experts, establishing the banking and finance sector as the benchmark in credit management..
- Manual processes and subjective decisions characterize many stages in the credit value chain. This was confirmed by our survey results as automated procedures for classifying customers emerged as a significant area for improvement. Various measures can classify credit customers, like loss given default (LGD) and value at risk (VaR), which captures the customer risk were presented; however, customer lifetime value (CLV) emerges as a valuable measure as it incorporates profitability of the customer as well as risk.
- Empirical results demonstrate the application of various statistical methods, including linear regression, random forest regression, XGBoost regression, and ARIMA modeling, as well as survival analysis for customer duration. These results provide a foundation for future improvements in forecasting customer classification measures. By calculating CLV, businesses can enhance their understanding of customer profitability and risk, enabling more informed decision-making and effective customer segmentation.

Although the modeling of customer duration performs well, the performance of the forecasting models falls short. In general, our models do not capture the intricacies of the data-generating process. To improve forecasting accuracy, additional factors beyond the population and output gap should be considered, such as company background variables, strategy, customer sentiment measured through natural language processing, trouble and delay indicators, and outstanding amounts. However, these factors were not accessible due to time and data limitations. Exploring advanced modeling techniques, such as state space modeling with a Kalman filter, could also contribute to enhanced results if supported by a larger and more reliable data set. Still, there are notable exceptions. The Cox-PH model provides acceptable results in forecasting customer duration, particularly when using time-independent variables. Due to data limitations, the discount factor is simplified, but it contributes to computing the discounted value that customers bring to the company in the future.

Nonetheless, the numerical calculation of CLV demonstrates the importance and significance of proper credit risk management and the prevention of customer bankruptcies, as default reduces the forecast CLV. Given the inherent challenge of measuring and forecasting human behavior, further accuracy improvements are necessary. Despite this, our research successfully addresses the research objectives.

This thesis sheds light on the current state of credit management in Norway and identifies specific areas for improvement. By utilizing a transdisciplinary approach, advanced modeling techniques, and feedback mechanisms, businesses can enhance their credit management practices, promote optimized resource allocation, and achieve effective credit risk management. The findings and recommendations of this thesis provide valuable insights for businesses seeking to enhance their credit management strategies and ultimately improve their financial performance, as well as predictability for their customers. Our contribution serves as a stepping stone toward creating digital twins for business processes and underscores the importance of data quality, availability, and transparency when aiming to model human behavior.

5.2. Evaluation and Assessment

With this thesis, we embarked on a journey to contribute to credit management by utilizing a systematic approach from cybernetics, recognizing the potential of feedback loops to enhance risk and profitability optimization. However, we encountered several tough challenges along the way, primarily stemming from the lack of analytical maturity in the field, which resulted in poor data availability and quality. We can clearly say that there is a need for several improvements and more data (both measurements and metrics) along the entire credit value chain. These limitations necessitated adjustments to our original plan and methodology, requiring us to adapt, adjust, and refine our approach.

The transdisciplinarity of this work has presented challenges along the way, with a large, demanding, and new field to familiarize ourselves with. Navigating a new field of study has been a challenging task for us, including learning and communicating with experienced individuals from the credit industry. The task of connecting with companies from different industries and persuading them to participate in our research, as well as gathering data, was also a demanding task. However, it has been a valuable learning experience as we have gained insights about the credit industry and into tools that will be useful in the future, and we have encountered individuals who generously shared their knowledge and industry insights with us.

Initially, we aimed to incorporate a broader range of elements within credit risk modeling, aiming for a more comprehensive exploration of the topic. The initial plan included several different aspects.

First, we had planned to identify and analyze potential leading, coincident, and lagging indicators to construct early warning systems, for the drivers of customer lifetime value (CLV), including sales, probability, customer survival, net exposure, and risk.

Second, we wanted to dive into customer classification, including factors such as customer status (i.e., potential, new, existing active, existing passive, lost, and former customers), customer segments (i.e., A, B, C, D, and E customers), customer categories (i.e., good customers, potentially good customers, bad customers, potentially bad customers, others), and customer portfolios.

Third, we wanted to build robust, automatic credit limit models.

Fourth, we wanted to address dynamic credit risk optimization by maximizing CLV while considering credit limits and credit scores.

Fifth, we wanted to address using data from the internet, by building e.g., measures like customer sentiment, to begin diving into actionable customer insights.

By including such themes, we aimed to develop a more holistic perspective on the credit management process and address the challenges identified through the survey in a more holistic and advanced manner. However, due to time constraints and limitations in data quality and accessibility, we had to limit the scope. Consequently, our thesis focused solely on conducting a survey capturing the state of the art regarding credit management in the Norwegian industry and addressing challenges by modeling CLV, which still serves as a solid foundation for future work, encompassing the aspects we were unable to incorporate within the scope of this project.

5.3. Further Work

There are several factors that can contribute to deepening our understanding of credit management and expanding the current modeling framework. By incorporating additional components and exploring alternative approaches, a more comprehensive and nuanced perspective can be achieved. The following can be further explored and developed to enhance the overall analysis and provide valuable insights for future research:

- Apply the components that were initially planned but not included in this thesis for a more holistic perspective, i.e., dynamic customer classification, dynamic credit risk optimization, automatic credit limits, and actionable customer insights.
- Enrich internal data and key performance indicators (KPIs) with more external data and ensure a higher data quality.
- Explore other modeling techniques beyond those used in the current analysis and perform fine-tuning and optimization of the models.
- Do more comprehensive research on external factors and drivers that can be included in the model and investigate the relationships and dependencies among these to obtain better models.
- Expand the model to a more general model, trained on more data from more companies, that is able to take into account the specific characteristics and drivers of individual companies that the model is applied to. Utilizing panel data analysis is a suggested method to obtain such a model.
- Establish a closer collaboration with industry experts to gain valuable insights into the specific needs, challenges, and emerging trends within the field, ensuring that the developed models and frameworks align effectively with practical requirements.

These avenues of further research can contribute to a more comprehensive understanding of credit management, enhance the modeling of CLV, and facilitate the incorporation of relevant drivers and contextual factors in a comprehensive framework.

Bibliography

[1] F. J. Muhlhorn, Customer profitability analysis: Measurement, concentration, and research directions, *Journal of Interactive Marketing*, 13:25–40 (1999).

[2] M. Crouhy, D. Galai, R. Mark, *Risk Management*, McGraw-Hill, 2001.

[3] V. Kumar, R. Venkatesan, T. Bohling, D. Beckmann, The power of clv: Managing customer lifetime value at ibm, *Marketing Science*, 27:585–599 (2008).

[4] T. Benzschawel, *Credit Risk Modelling—Facts, Theory and Applications*, Risk Books, 2012.

[5] F. Black, M. Scholes, The valuation of option contracts and a test of market efficiency, *The Journal of Finance*, 27:399 (1972).

[6] R. C. Merton, Theory of rational option pricing, *The Bell Journal of Economics and Management Science*, 4:141 (1973).

[7] H. Nobanee, H. Shanti, H. Aldhanhani, A. Alblooshi, E. Alali, Big data and credit risk assessment: A bibliometric review, current streams, and directions for future research, *Cogent Economics & Finance*, 10 (2022).

[8] R. J. Powell, D. E. Allen, Cvar and credit risk measurement, *School of Accounting, Finance and Economics & FEMARC Working Paper Series* (2009).

[9] H. Aeron, T. Bhaskar, R. Sundararajan, L. A. I. [Last Author's Last Name], A metric for customer lifetime value of credit card customers, *Journal of Database Marketing & Customer Strategy Management*, 15:153–168 (2008).

[10] V. Kumar, *Customer lifetime value: The path to profitability*, Now Publishers Inc., 2008.

[11] R. J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice*, OTexts, 2018.

[12] C. Chatfield, *Time-series forecasting*, CRC press, 2000.

[13] V. Cerqueira, L. Torgo, I. Mozetič, Evaluating time series forecasting models: An empirical study on performance estimation methods, *Machine Learning*, 109:1997–2028 (2020).

[14] H. Wold, A study in the analysis of stationary time series, Ph.D. thesis, Almqvist & Wiksell (1938).

[15] A. Zemkoho, A basic time series forecasting course with python, in *Operations Research Forum*, vol. 4, Springer, 2023, pp. 1–43.

[16] E. C. Nwogu, I. S. Iwueze, K. C. Dozie, H. I. Mbachu, Choice between mixed and multiplicative models in time series decomposition, *Int J Stat Appl Math*, 9:153–159 (2019).

[17] G. S. Maddala, I.-M. Kim, *Unit Roots, Cointegration, and Structural Change*, Cambridge University Press, 1998.

[18] A. Zemkoho, S. Ahipasaoglu, Forecasting, https://github.com/abzemkoho/forecasting/blob/main/Forecasting_Lecture_Notes.pdf, Accessed: May 9, 2023 (2022).

[19] G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

[20] P. Kennedy, *A guide to econometrics*, John Wiley & Sons, 2008.

[21] N. H. Chan, N. Terrin, Inference for unstable long-memory processes with applications to fractional unit root autoregressions, *The annals of Statistics*, 23:1662–1683 (1995).

[22] L. A. Gil-Alana, P. M. Robinson, Testing of unit root and other nonstationary hypotheses in macroeconomic time series, *Journal of Econometrics*, 80:241–268 (1997).

[23] F. X. Diebold, *Elements of forecasting*, Citeseer, 1998.

[24] S. Luo, X. Kong, T. Nie, Spline based survival model for credit risk modeling, *European Journal of Operational Research*, 254:249–258 (2016).

[25] R. P. Suárez, R. C. Abad, J. M. V. Fernández, Probability of default estimation in credit risk using a nonparametric approach (2019), Available online at http://dm.udc.es/preprint/Pelaez_Cao_Vilar_Probability_default_estimation_EJOR.pdf.

[26] P. K. Trivedi, *Microeconometrics: Methods and Applications*, Cambridge University Press, 2005.

[27] T. Bellotti, J. Crook, Credit scoring with macroeconomic variables using survival analysis, *Credit Research Centre* (2007).

[28] D. Cox, Regression models and life-tables, *Journal of the Royal Statistical Society, Series B*, 34:187–220 (1972).

[29] J. N. Crook, D. B. Edelman, L. C. Thomas, Recent developments in consumer credit risk assessment, *Journal of Banking & Finance*, 31:3848–3865 (2007).

[30] A. Gross, V. Clark, *Survival Distributions: Reliability Applications in the Biomedical Sciences*, John Wiley & Sons, Inc, New York, 1975.

[31] M. Sheikh, A. C. C. Coolen, Analysis of overfitting in the regularized cox model, *Journal of Physics A: Mathematical and Theoretical*, 52:384002 (2019).

[32] T. Gjesdal, Jostein & Johnsen, *Kravsetting, lønnsomhetsmåling og verdivurdering*, Cappelen Akademisk Forlag, 1999.

[33] D. Watson, A. Head, G. M. Mantovani, E. Rossi, et al., *CORPORATE FINANCE. Principles and practice in Europe*, Pearsons, 2016.

[34] J. L. Treynor, Market value, time, and risk, *Time, and Risk (August 8, 1961)* (1961).

[35] W. F. Sharpe, Capital asset prices: A theory of market equilibrium under conditions of risk, *The journal of finance*, 19:425–442 (1964).

[36] J. Lintner, The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets: A reply, *The Review of Economics and Statistics*, 51:222–224 (1969).

[37] J. Mossin, Equilibrium in a capital asset market, *Econometrica: Journal of the econometric society*, pp. 768–783 (1966).

[38] H. Markowitz, Portfolio selection, *The Journal of Finance*, 7:77–91 (1952).

[39] M. Q. Patton, *Qualitative research & evaluation methods: Integrating theory and practice*, Sage publications, 2014.

[40] A. Galloway, Non-probability sampling, 2005.

[41] Effect.IO, Effect.IO, https://effect.io/research/, Accessed on April 27, 2023 (2021).

[42] P. Hall, B. U. Park, R. J. Samworth, Choice of neighbor order in nearest-neighbor classification (2008).

[43] M. L. Waskom, seaborn: statistical data visualization, *Journal of Open Source Software*, 6:3021 (2021).

[44] N. Razali, Y. B. Wah, Power comparisons of shapiro–wilk, kolmogorov–smirnov, lilliefors and anderson–darling tests, *Journal of Statistical Modeling and Analytics*, 2:21–33 (2011), Available online at `https://web.archive.org/web/20150630110326/http://instatmy.org.my/downloads/e-jurnal%202/3.pdf`.

[45] K. Pearson, The problem of the random walk, *Nature*, 72:294–294 (1905).

[46] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, A. Keshri, Descriptive statistics and normality tests for statistical data, *Annals of Cardiac Anaesthesia*, 22:67–72 (2019).

[47] C. Z. Mooney, C. F. Mooney, R. D. Duval, C. L. Mooney, R. Duvall, *Bootstrapping: A nonparametric approach to statistical inference*, no. 95, sage, 1993.

[48] D. Kai-Ineman, A. Tversky, Prospect theory: An analysis of decision under risk, *Econometrica*, 47:363–391 (1979).

[49] K. J. Haarberg, The STO DELTA PAI model, Prognosesenteret.

[50] K. Król, D. Zdonek, Analytics maturity models: An overview, *Information*, 11:142 (2020).

[51] M. Aaberge, J. Johanraj, Digital twins for business processes: A research on customer loyalty management, 2022.

[52] T. H. Davenport, J. G. Harris, R. Morison, *Analytics at work: Smarter decisions, better results*, Harvard Business School Publishing Corporation, United States of America, 2010.

[53] SSB StatBank: Mainland gdp, by industry and expenditure, current prices, Statistics Norway (SSB) - StatBank, `https://www.ssb.no/statbank/table/12880/` (2023).

[54] F. Furlanetto, K. Hagelund, F. Hansen, O. Robstad, Norges bank output gap estimates: Forecasting properties, reliability and cyclical sensitivity (2020).

[55] P. D. Allison, *Missing Data*, Sage Publications, Beverly Hills, CA, 2002.

[56] J. Hartung, G. Knapp, Multivariate multiple regression, *Encyclopedia of statistics in behavioral science* (2005).

[57] SSB StatBank: Population and population changes, Statistics Norway (SSB) - StatBank, https://www.ssb.no/statbank/table/07459/ (2023).

[58] A. Harvey, T. Trimbur, Trend estimation and the hodrick-prescott filter, *Journal of the Japan Statistical Society*, 38:41–49 (2008).

[59] R. J. Hodrick, E. C. Prescott, Postwar us business cycles: an empirical investigation, *Journal of Money, credit, and Banking*, pp. 1–16 (1997).

[60] M. O. Ravn, H. Uhlig, On adjusting the hodrick-prescott filter for the frequency of observations, *Review of economics and statistics*, 84:371–376 (2002).

[61] Y. Chen, S. Liu, J. Yang, H. Jing, W. Zhao, G. Yang, A joint time-frequency domain transformer for multivariate time series forecasting (2023).

[62] J. D. Hamilton, *Time series analysis*, Princeton university press, 2020.

[63] W. Krämer, Durbin–watson test, in *International encyclopedia of statistical science*, Springer, 2011, pp. 408–409.

[64] P. W. Farris, N. Bendle, P. E. Pfeifer, D. Reibstein, *Marketing metrics: The definitive guide to measuring marketing performance*, Pearson Education, 2010.

[65] T. Belloti, J. Crook, Forecasting and stress testing credit card default using dynamic models, *International Journal of Forecasting*, 29:563–574 (2013).

[66] T. Therneau, C. Crowson, E. Atkinson, Using time dependent covariates and time dependent coefficients in the cox model, *Mayo Clinic* (2023), March 11, 2023.

[67] S. A. Ross, The arbitrage theory of capital asset pricing, *Journal of Economic Theory*, 13:341–360 (1976).

[68] D. P. Doane, L. E. Seward, Measuring skewness: a forgotten statistic?, *Journal of statistics education*, 19 (2011).

[69] P. H. Westfall, Kurtosis as peakedness, 1905 - 2014. r.i.p., *The American Statistician*, 68:191–195 (2014).

[70] S. A. Klugman, H. H. Panjer, G. E. Willmot, *Loss Models: From Data to Decisions*, 5th edition, Wiley, 2019.

[71] S. S. Shapiro, M. B. Wilk, An analysis of variance test for normality (complete samples), *Biometrika*, 52:591–611 (1965).

[72] R. D'Agostino, A. Belanger, R. J. D'Agostino, A suggestion for using powerful and informative tests of normality, *The American Statistician*, 44:316–321 (1990).

[73] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine*, 50:157–175 (1900).

[74] H. Mann, D. Whitney, On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, 18:50–60 (1947).

[75] W. Kruskal, W. Wallis, Use of ranks in one-criterion variance analysis, *Journal of the American statistical Association*, 47:583–621 (1952).

[76] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association*, 32:675–701 (1937).

[77] D. Dickey, W. Fuller, Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, 74:427–431 (1979).

[78] P. Phillips, P. Perron, Testing for a unit root in time series regression, *Biometrika*, 75:335–346 (1988).

[79] D. Kwiatkowski, P. Phillips, P. Schmidt, Y. Shin, Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics*, 54:159–178 (1992).

[80] S. Johansen, Statistical analysis of cointegration vectors, *Journal of Economic Dynamics and Control*, 12:231–254 (1988).

[81] J. Miles, R-squared, adjusted r-squared, *Encyclopedia of statistics in behavioral science* (2005).

[82] G. M. Ljung, G. E. P. Box, On a measure of lack of fit in time series models, *Biometrika*, 65:297–303 (1978).

[83] T. Breusch, A. Pagan, A simple test for heteroscedasticity and random coefficient variation, *Econometrica*, 47:1287–1294 (1979).

[84] C. Granger, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, 37:424–438 (1969).

A. Survey

A.1. Credit Management Survey

Below, we will address several different topics within credit management. Specifically, we have formulated statements where we would like you to indicate where your business is on a scale from 0 (totally disagree) to 10 (totally agree).

Credit strategy
The credit strategy of the business is normally laid down in the credit policy and forms the basis for how credit is handled. In what follows, we will ask some questions related to the business' credit strategy.

Our company has...

1. a credit policy that everyone follows fully
2. a credit policy with very clear goals, guidelines, incentives, mandates, and consequences for deviations from the credit policy
3. a highly competent credit committee that meets regularly and in the event of important, unforeseen events
4. leaders who base all their decisions on facts (i.e., data, analyses, and models)
5. access to very easily accessible, reliable, relevant, and enriched data of very high quality
6. automatic and very good methods and models to optimize credit risk given the customers' risk-adjusted profitability and credit limits as well as external factors

Classification and analyses of customers
In order to be able to treat customers in an optimal way, it is important for the business to have an overview of and be able to analyze, among other things, profitability, uncertainty, and potential for various customers, as well as the development of customer statuses (i.e., potential, new, existing active, existing passive, lost, and former customers), customer segments, customer categories (i.e., good, potentially good, bad, potentially bad, and other customers), and customer portfolios over time.

Our company has...

7. automatic and very good and easily accessible overview of customer statuses, customer segments, customer categories, and customer portfolios
8. automatic and very good methods and models for selecting and prioritizing customers

9. automatic and very good methods and models for calculating the customer's lifetime value (CLV)

10. automatic and very good methods and models for obtaining actionable insights wrt. customer behavior, customers' preferred content, customers' preferred channels, and phase in the customer value chain (KYC)

11. automatic and very good methods and models for uncovering money laundering (AML)

12. automatic and very good methods and models for relevant external factors (e.g., the business cycle, market development, and competition) wrt. the business's decisions, processes, and analyses

Establishing credit

Establishing credit involves several different processes, and the more of the establishment the business can automate, the better the customer service. In addition, it frees up time and resources to work on other issues.

Our company has...

13. very clear guidelines for establishing credit

14. no manual credit procedures

15. automatic credit limits

16. always full acceptance among all sellers regarding the refusal of credit

Monitoring and following up credit

Monitoring and following up on approved credits is important. In this context, it may be useful to establish leading, coincident, and lagging credit indicators so that the company can establish early warning systems regarding credit risk, etc., to save resources and prevent losses.

Our company has...

17. very good routines for handling payment delays

18. very good routines for identifying and following up on payment notices and collection notices

19. very good routines for identifying and avoiding potential losses

20. automatic and very good methods and models for calculating the effects of loss prevention activities

21. automatic and very good methods and models for calculating risk-adjusted profitability (including the costs associated with trouble and delay)

22. automatic and very good methods and models for calculating the customer sentiment

23. very clear guidelines for changes in credit and credit limits

Recovering losses

When the company loses money on customers, it is important that the company has procedures and competencies that allow it to recover losses as effectively as possible.

Our company has...

24. very good overview of which customers pose a risk of loss for the company
25. very good routines for following up on debt collection cases
26. very good collaterals for all customers (e.g., mortgage securities, sureties, bank guarantees, credit insurance, and overview of the net sale value of the customer's inventory)
27. very good methods and models for calculating the probability of loss recovery, default probability, expected loss given default, and net exposure at default.

A.2. Survey Plots

Hypothesis 1

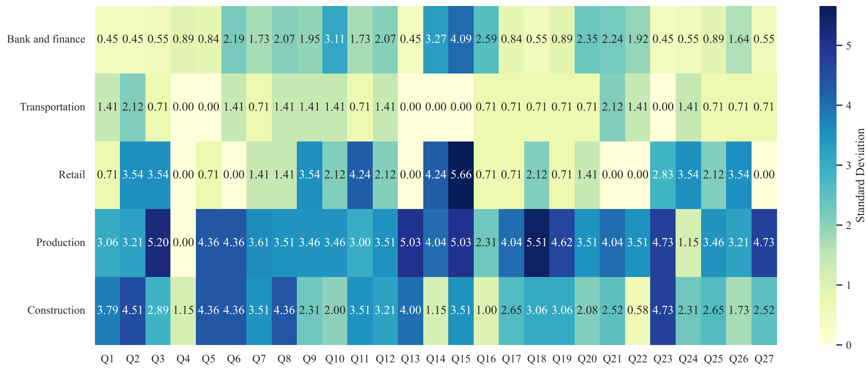


Figure A.1.: Heatmap showing the variance in each question for each industry, with standard deviations in each cell.

B. Statistics

B.1. Tests and Measures for Survey Analysis

B.1.1. Skewness and Kurtosis

Skewness is a measure of lack of symmetry of the normal distribution [46]. The formulas for skewness in terms of the second and third moments around the mean are:

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{B.1}$$

and

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3, \tag{B.2}$$

where n is the number of samples and $\bar{x}$ is the mean. It is common to use these in the Fisher-Pearson coefficient of skewness [68]:

$$skrewness = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}. \tag{B.3}$$

The skewness for a normal distribution is zero.

Kurtosis is related to the tails of the distribution [69]. Data sets with high kurtosis tend to have heavy tails or outliers, while data sets with low kurtosis tend to have lighter tails or lack outliers. The formula for kurtosis is:

$$kurtosis = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \tag{B.4}$$

where the exponent in Equation B.2 is changed from 3 to 4.

A normal distribution with $kurtosis \approx 3$ is called mesokurtic, a distribution with $kurtosis < 3$ is called platykurtic, and a distribution with $kurtosis < 3$ is leptokurtic.

B.1.2. The Jarque-Bera Test

The Jarque-Bera test is a goodness-of-fit test that checks whether sample data have skewness and kurtosis matching a normal distribution. It's a type of non-parametric test.

The Jarque-Bera test statistic JB is calculated as:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{B.5}$$

where n is the number of observations or size of the sample, S is the sample skewness, and K is the sample kurtosis [70]. The test's null hypothesis is that the data is normally distributed, indicated by the skewness of zero and the excess kurtosis of zero.

If the JB statistic is sufficiently large, it would indicate the data does not have the skewness and kurtosis of a normal distribution, and the null hypothesis would be rejected.

B.1.3. The Shapiro-Wilk Test

The Shapiro-Wilk test is one commonly used normality test, and it is parametric. The null hypothesis is that the data is normally distributed, and to reject or accept the null hypothesis; a W statistic is calculated and compared to a critical value. The W statistic is given by

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{B.6}$$

where n is the sample size, $y_{(i)}$ is the i'th order value of the sample data, $\overline{y}$ is the sample mean, and a_i are coefficients depending on the sample size and the distributional parameters of the normal distribution [71].

B.1.4. The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test is a nonparametric test used to compare a sample distribution with a reference probability distribution or two sample distributions. It provides a method to test the goodness of fit or whether two empirical distributions differ significantly.

The test statistic is the maximum distance D between the cumulative distribution function (CDF) of the sample and the CDF of the reference distribution or between the CDFs of two samples. It is calculated as:

$$D = \sup_x |F_n(x) - F(x)| \tag{B.7}$$

where $F_n(x)$ is the empirical distribution function of the sample and $F(x)$ is the CDF of the reference distribution [70]. When comparing two samples, $F(x)$ would be replaced with the empirical distribution function of the second sample.

The null hypothesis of the K-S test is that the sample is drawn from the reference distribution in the case of a one-sample test or that two samples are drawn from the same distribution in the case of a two-sample test. The null hypothesis can be rejected if the D statistic is sufficiently large.

B.1.5. The Anderson Darling Test

The Anderson-Darling test is a statistical test commonly used to assess whether a given data sample follows a specific distribution, such as the normal distribution. It is a goodness-of-fit test that measures the discrepancy between the observed data and the expected distribution.

The null hypothesis in the Anderson-Darling test assumes that the data is drawn from a specified distribution, while the alternative hypothesis suggests otherwise. Therefore, the test calculates a test statistic based on the differences between the observed data and the expected distribution. This test statistic is then compared to critical values or p-values to determine the acceptance or rejection of the null hypothesis.

The Anderson-Darling test statistic is defined as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln(F(X_i)) + \ln(1 - F(X_{n+1-i}))\right] \tag{B.8}$$

where n is the sample size, X_i is the ith ordered value of the sample data, and $F(X_i)$ is the corresponding cumulative distribution function (CDF) value. The critical values or p-values associated with the test statistic are obtained from statistical tables or software packages [70].

B.1.6. The D'Agostino K-Squared Test

D'Agostino's K-squared test is a statistical test used to determine whether a data set is normally distributed [72]. Unlike the Shapiro-Wilk test, which also checks for normality, D'Agostino's K-squared test checks for skewness and kurtosis separately and combines these measures into a single test statistic.

The test statistic K^2 is computed as:

$$K^2 = Z_g^2 + Z_k^2 \tag{B.9}$$

where Z_g and Z_k are the test statistics for skewness and kurtosis, respectively. These statistics are computed based on the sample skewness g_1 and sample kurtosis g_2, and then standardized.

The null hypothesis of D'Agostino's K-squared test is that the data follows a normal distribution. If the computed K^2 value is sufficiently large, it suggests that the data distribution deviates from normality. Thus, the null hypothesis is rejected.

B.1.7. Pearson's Chi-Square Test

Pearson's Chi-Square Test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference arose by chance. It's non-parametric and used when the data is divided into categories rather than numerical values.

The test statistic for the Pearson's Chi-Square Test is denoted as χ^2 and is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{B.10}$$

Here, n is the number of categories, O_i is the observed frequency of category i, and E_i is the expected frequency of category i [73]. The null hypothesis is that there's no association between the categories (i.e., they're independent), and the test assesses whether the observed values significantly differ from the expected values.

B.1.8. The Mann-Whitney U Test

The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a nonparametric statistical significance test used to compare the distributions of two independent samples. It is used to test the null hypothesis that the distributions of two populations are the same. The Mann-Whitney U Test does not assume that the data are normally distributed.

The Mann-Whitney U statistic is calculated by first ranking all observations from both samples from smallest to largest, then summing the ranks for the observations that come from each sample. If there are n_1 observations from sample 1 and n_2 observations from sample 2, then U is given by:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{B.11}$$

where R_1 is the sum of the ranks in sample 1. An equivalent formula in terms of sample 2 can also be used [74].

The null hypothesis for the Mann-Whitney U Test is that the two samples come from the same population (i.e., the population distributions are equal). If the calculated U statistic is less than a critical value from the U distribution table, the null hypothesis is rejected. This would suggest that the distributions of the two populations are different.

B.1.9. The Kruskal-Willis Test

The Kruskal-Wallis test is a nonparametric test that extends the Mann-Whitney U test to more than two groups. It compares the medians of the groups to test the null hypothesis that all groups come from the same population (or, equivalently, from different populations with the same distribution).

The test statistic H is calculated as:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(n+1) \tag{B.12}$$

where n is the total number of observations, g is the number of groups, R_i is the sum of the ranks in the ith group, and n_i is the number of observations in the ith group [75]. The ranks are based on all observations, regardless of which group they are in.

If the calculated H statistic is sufficiently large (based on a chi-square distribution with $g - 1$ degrees of freedom), the null hypothesis can be rejected, suggesting that at least one group is different from the others.

B.1.10. Friedman Test

The Friedman test is a non-parametric statistical test used to determine whether there are any significant differences among multiple related groups or treatments. It is particularly useful when the data cannot be assumed to follow a normal distribution or when the data are measured on an ordinal scale.

The test calculates a test statistic called the Friedman chi-square (χ^2), which is based on the ranks of the observations within each group. The test statistic is given by the formula:

$$\chi^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \tag{B.13}$$

where n is the number of subjects or observations, k is the number of groups or treatments, and R_j represents the sum of ranks for the jth group [76]. The null hypothesis of the Friedman test is that there are no differences among the groups, indicating that the population medians of all groups are equal. If the Friedman chi-square statistic is sufficiently large, it suggests that there are significant differences among the groups, and the null hypothesis is rejected.

B.2. Tests and Measures for Applied Modeling

B.2.1. Augmented Dickey-Fuller (ADF) Test

The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a time series has a unit root, indicating non-stationarity. The ADF test estimates an autoregressive model of the time series and examines the significance of the coefficient associated with the lagged difference term. The test statistic is given by:

$$ADF = \frac{\hat{\rho}}{SE(\hat{\rho})} \tag{B.14}$$

where $\hat{\rho}$ is the estimated coefficient of the lagged difference term, and $SE(\hat{\rho})$ is the standard error of the estimated coefficient.

The null hypothesis of the ADF test is that the time series possesses a unit root, implying non-stationarity. Conversely, the alternative hypothesis suggests stationarity. Critical values for the ADF statistic are used to compare the test statistic and determine the outcome [77]. If the ADF statistic is smaller than the critical values, the null hypothesis of a unit root is rejected, indicating stationarity. On the other hand, if the ADF statistic is larger than the critical values, the null hypothesis cannot be rejected, suggesting non-stationarity.

B.2.2. Phillips-Perron (PP) Test

The Phillips-Perron (PP) test is a statistical test used to examine the presence of a unit root in a time series. It allows for serial correlation and heteroscedasticity in the error terms, providing robustness to certain violations of the standard assumptions [78].

The test estimates an autoregressive model of the time series and examines the significance of the coefficient associated with the lagged difference term. The test statistic is given by:

$$PP = \frac{\hat{\rho} - 1}{SE(\hat{\rho})} \tag{B.15}$$

where $\hat{\rho}$ is the estimated coefficient of the lagged difference term, and $SE(\hat{\rho})$ is the standard error of the estimated coefficient.

The null hypothesis of the PP test is that the time series possesses a unit root, indicating non-stationarity. Conversely, the alternative hypothesis suggests stationarity. Critical values for the PP statistic are used to compare the test statistic and determine the outcome. If the PP statistic is smaller than the critical values, the null hypothesis of a unit root is rejected, indicating stationarity. On the other hand, if the PP statistic is larger than the critical values, the null hypothesis cannot be rejected, suggesting non-stationarity.

B.2.3. Kwiatkowski–Phillips–Schmidt–Shin (KPSS) Test

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is a statistical test used to assess the stationarity of a time series. It provides an alternative approach to the ADF test.

The test compares the sum of squared deviations from a deterministic trend to the sum of squared residuals under the null hypothesis of stationarity [79]. The test statistic is calculated as follows:

$$KPSS = \frac{T \times \mathrm{SSE}_{\mathrm{trend}}}{\sigma^2} \tag{B.16}$$

where T is the number of observations, $\mathrm{SSE}_{\mathrm{trend}}$ is the sum of squared residuals from a regression on a deterministic trend, and σ^2 is the estimated variance of the residuals.

The null hypothesis of the KPSS test assumes that the time series is trend-stationary, meaning it has a constant mean and variance over time, allowing for a deterministic trend. The alternative hypothesis suggests the presence of a unit root, indicating non-stationarity. Critical values are used to compare the KPSS test statistic and determine the outcome. If the KPSS statistic is larger than the critical values, the null hypothesis of stationarity is rejected, indicating non-stationarity. Conversely, if the KPSS statistic is smaller than the critical values, the null hypothesis cannot be rejected, suggesting trend stationarity.

B.2.4. Johansen Test

The Johansen test is a procedure for testing the cointegration of several, say k, I(1) time series. This test permits more than one cointegrating relationship, unlike the Engle-Granger method, so it is suitable for systems of equations.

The null hypothesis is that the number of cointegrating vectors is r=r* with 0 <= r* < k under the alternative hypothesis that r=k. Two types of likelihood ratio tests are conducted, namely trace test and maximum eigenvalue test.

The trace test statistic is given as:

$$\lambda_{trace}(r*) = -T \sum_{i=r*+1}^{k} \ln\left(1 - \hat{\lambda}_i\right) \tag{B.17}$$

where $\hat{\lambda}_i$ are the estimated eigenvalues from the Johansen procedure.

The maximum eigenvalue test statistic is given as:

$$\lambda_{max}(r*, r*+1) = -T \ln\left(1 - \hat{\lambda}_{r*+1}\right) \tag{B.18}$$

where T is the number of usable observations.

If the calculated test statistic is greater than the critical value, the null hypothesis is rejected [80]. The trace test checks the null hypothesis that the number of cointegration vectors is less than or equal to r against an unrestricted alternative, while the maximum eigenvalue test checks the null hypothesis that the number of cointegration vectors is r against the alternative of r+1 cointegration vectors.

B.2.5. MSE and RMSE

Accuracy is important when it comes to forecasting, as the main purpose of forecasting is to make informed decisions based on future forecasts. The most commonly used error measures are mean Squared error (MSE), and root mean squared error (RMSE) [23]. These are defined as:

$$MSE = \frac{1}{n} \sum_{t=1}^{n} e_t^2, \tag{B.19}$$

and

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} e_t^2}, \tag{B.20}$$

where n is the number of observations in the data set, t is the time and e_t represents the residual at time t. RMSE is particularly useful because it penalizes larger errors more than smaller ones, making it an excellent choice for applications where larger deviations from the true values are undesirable.

B.2.6. R^2

Another statistical measure that is often used in regression models is R^2, which determines the proportion of variance in the dependent variable that can be explained by the independent variable, i.e., how well the estimated data fit the actual data regression model [81]. It is given by the following formula:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}, \tag{B.21}$$

where the nominator is the sum of squares of residuals and the denominator is the total sum of squares.

When assessing accuracy, it is crucial to compare the forecasts of two or more methods against the original time series. This involves examining the errors between each forecast time series and the original series.

B.2.7. Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a widely used measure of a statistical model's quality. It deals with the trade-off between the goodness of fit of the model and the complexity of the model. It is defined as:

$$AIC = 2k - 2\ln(L) \tag{B.22}$$

where k is the number of model parameters (the number of variables in the model plus the constant) and L is the maximized value of the likelihood function for the estimated model. Lower values of AIC indicate better fitting models [23]. AIC is a relative measure of model quality, and so it only has meaning if we compare the AIC for one model with the AIC for some alternative model.

B.2.8. Durbin-Watson Statistic

The Durbin-Watson statistic is a test statistic used to detect the presence of autocorrelation (a relationship between values separated from each other by a given time lag) in the residuals from a regression analysis.

It is given by:

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{B.23}$$

where e_t are the residuals associated with the time series. The statistic ranges from 0 to 4, with a value around 2 indicating no autocorrelation. A value towards 0 indicates positive autocorrelation and a value towards 4 indicates negative autocorrelation [23]. The Durbin-Watson statistic helps us understand the correlation in the residuals, which in turn can inform whether our model is appropriately specified.

B.2.9. Ljung-Box Test

The Ljung-Box test is a tool designed to detect autocorrelation in the residuals of a time series model [82]. Autocorrelation, or the lack of independence among residuals, can invalidate the underlying assumptions of many models.

The test statistic is computed as:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{B.24}$$

where n represents the number of observations, h is the number of lags being considered, and $\hat{\rho}_k$ is the estimated autocorrelation at lag k.

The null hypothesis for this test is that the residuals are independently distributed. If the p-value is significant (usually below 0.05), the null hypothesis is rejected, implying that the residuals exhibit autocorrelation. The Ljung-Box test is used on the residuals of a fitted model, assuming that the time series is stationary.

B.2.10. Breusch-Pagan Test

The Breusch-Pagan test is a statistical test used to examine the presence of heteroscedasticity in a regression model. The test is performed to determine whether the variance of the error term in the regression model is constant across different levels of the independent variables.

The test involves estimating the regression model and obtaining the residuals. It then examines the relationship between the squared residuals and the independent variables [83]. The null hypothesis assumes homoscedasticity, indicating that the squared residuals are not significantly related to the independent variables.

The test statistic is based on the F-statistic, comparing the explained variation in the squared residuals using the independent variables to the unexplained variation.

The Breusch-Pagan test statistic is calculated as follows:

$$LM = nR^2 \tag{B.25}$$

where n is the number of observations and R^2 is the coefficient of determination from the auxiliary regression of the squared residuals on the independent variables.

The test statistic follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the auxiliary regression.

Critical values are used to compare the test statistic and determine the significance level. If the test statistic is larger than the critical values, the null hypothesis of homoscedasticity is rejected, indicating the presence of heteroscedasticity. On the other hand, if the test statistic is smaller than the critical values, the null hypothesis cannot be rejected, suggesting the absence of heteroscedasticity.

B.2.11. Granger Causality Test

The Granger causality test discerns if past values of one time series contribute significant information for predicting a second-time series [84]. It assesses causal relationships by comparing the explained variation in the dependent variable with and without considering lagged values of the independent variable.

The test statistic, an F-statistic, is calculated as follows:

$$F = \frac{(RSS_{\text{restricted}} - RSS_{\text{unrestricted}})/k}{RSS_{\text{unrestricted}}/(n-p-1)} \tag{B.26}$$

where $RSS_{\text{restricted}}$ and $RSS_{\text{unrestricted}}$ represent the residual sum of squares from models without and with lagged values of the independent variable, respectively. k is the number of lagged values, n is the number of observations, and p is the number of parameters in the model.

The null hypothesis of no Granger causality is rejected if the test statistic exceeds the critical values, signifying a causal relationship. Otherwise, it's accepted, implying an absence of Granger causality.

C. Data Set Used in Applied Modelling

Table C.1.: Structure of the provided data set with illustrative customer data, used to model CLV in this thesis.

Obs period	Customer ID	Company Name	Postal Code	Country	Customer group
2023-03	10001	XX	8500	DK	109
2023-04	10002	XX	4800	Pl	109
2023-04	10003	XX	1234	NO	109
2013-06	10004	XX	9999	Other	102D

Industry	Start date	End date	Credit score	Credit limit	Invoice frequency
126	14.03.2020	NA	0.12	1000000	1
110	18.03.2012	17.04.2023	0.24	100000	5
94	14.04.2016	19.04.2023	0.75	100000	7
83	10.03.2022	12.08.2011	0.63	1000000	1

Invoiced amount	Variable costs	Fixed costs	Invoice reminders	Reminder degree
11528.31928	627.294513	-49.821745	23	2
2777.500583	912.482117	-83.692405	7	1
22444.46407	8725.929879	-972.402389	0	0
1528.319281	627.294513	-49.821745	23	2

NTNU

Norwegian University of
Science and Technology