LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
School of Business and Management
Strategic Finance and Business Analytics

Filloreta Karavidaj

MODELING CUSTOMER PAYMENT BEHAVIOR AND
FORECASTING RECEIVABLES (Company case)

Master’s thesis
2019

Supervisor: Post-Doctoral Researcher, Azzurra Morreale
ABSTRACT

Author: Filloreta Karavidaj

Title: Modeling customer payment behavior and forecasting receivables

University: Lappeenranta University of Technology

Faculty: School of Business and Management

Major: Strategic Finance and Business Analytics

Thesis content: 66 pages, 16 figures, 2 tables

Examiner: Azzurra Morreale

Keywords: sales receivable, working capital, predictive modeling, SARIMAX, Kaplan-Meier

Nowadays it has become crucial for companies to closely monitor their working capital levels because of the critical role it has in business operations. Sales receivable are a key component of working capital and their estimation is equally important. In this thesis two statistical models are built, based on 3-years historical company data. One of the models is Kaplan-Meier, a customer-based model which models customer payment behavior and outputs the probability that the customer will pay the amount owed, if it has not paid until today. Modeling customer payment behavior enhances collections management on decision-making and it also assists sales department in making timely delivery stops. The second model is a SARIMAX model that predicts outstanding sales receivable amount at a certain date in the future. Reliable estimations of sales receivable from this model result in more accurate working capital estimation which in turn helps in determining the necessity for short-term external financing.
Throughout the writing of this thesis I have received great support and assistance. I owe my deepest gratitude to the case company for enabling me with such an opportunity. Many thanks go to my supervisor in the company Xiaoyang Wu, who made sure I had all the necessary resources for conducting my research. I would also like to thank Ville Könönen who willingly shared his expertise and expanded my knowledge in the research area.

I thank my academic supervisor Azzurra Morreale for her valuable feedback and the flexibility she offered me to make my own choices regarding the research topic and methodology. Finally, I would like to thank all my friends who were of great support in deliberating over problems arising along the journey as well as providing happy distraction outside of my research.
# TABLE OF CONTENTS

ABSTRACT .................................................................................................................. 2

ACKNOWLEDGEMENTS ............................................................................................. 3

TABLE OF CONTENTS ............................................................................................... 4

LIST OF FIGURES ......................................................................................................... 6

LIST OF ABBREVIATIONS ............................................................................................ 7

1. INTRODUCTION ........................................................................................................ 9
   1.1 Motivation and objective of the study ................................................................... 9
   1.2 Research questions ............................................................................................... 11
   1.3 Research methodology ......................................................................................... 12
   1.4 Thesis structure .................................................................................................... 14

2. THEORETICAL BACKGROUND ............................................................................... 15
   2.1 Trade credit ......................................................................................................... 15
      2.1.1 Theories .......................................................................................................... 18
      2.1.2 Supplier’s perspective ..................................................................................... 21
      2.1.3 Customer’s perspective .................................................................................. 22
   2.2 Definition of working capital ................................................................................ 23
      2.2.1 Working capital management ......................................................................... 25
      2.2.1 Determinants of working capital .................................................................... 27
      2.2.2 Measures of working capital management .................................................... 28
   2.3 Comparison against previous research ................................................................... 30

3. CASE COMPANY ...................................................................................................... 31
   3.1 Trade credit in the case company ......................................................................... 31
   3.2 Working capital in the case company ..................................................................... 32

4. PREDICTIVE MODELING ....................................................................................... 33
   4.1 Time series analysis .............................................................................................. 33
      4.1.1 Data processing and filtering .......................................................................... 34
   4.2 ARIMA modeling .................................................................................................. 36
   4.3 Kaplan – Meier model ......................................................................................... 40

5. DEVELOPMENT OF SARIMAX MODEL ............................................................ 42
   5.1 Data description and feature selection ................................................................... 42
5.2 Data cleaning and preprocessing ................................................................. 42
5.3 Model hyperparameter selection ................................................................. 42

6. RESULTS AND CONCLUSIONS ................................................................. 44
6.1 Customer behavior analysis with Kaplan-Meier ......................................... 44
6.2 Sales receivable prediction with Kaplan-Meier ........................................... 47
6.3 Sales receivable prediction with SARIMAX ............................................... 48
6.4 Summary .................................................................................................... 52
6.5 Conclusion ............................................................................................... 54
6.6 Limitations and suggestions for further research ....................................... 55

REFERENCES ............................................................................................ 56

APPENDIX
LIST OF FIGURES

Figure 1. Benefits of analyzing customer payment behavior. .................................................. 10
Figure 2. Impact of sales receivable analysis ........................................................................ 10
Figure 3. Thesis structure ........................................................................................................ 14
Figure 4. Trade credit relationship (Adapted from Petersen & Rajan 1997). ......................... 16
Figure 5. Typical balance sheet ............................................................................................. 24
Figure 6. Example of stationary time series (Adapted from Brooks 2014) ......................... 35
Figure 7. Example of non-stationary time series (Adapted from Brooks 2014) .................. 35
Figure 8. Payment behavior over all the customers ............................................................... 44
Figure 9. Regular customer payment behavior ...................................................................... 45
Figure 10. Irregular customer payment behavior ................................................................. 46
Figure 11. Sales receivable predictions based on customer behavior .................................... 47
Figure 12. Prediction error plotted against prediction horizon ............................................ 50
Figure 13. Actual and predicted sales receivable daily SARIMAX model prediction horizon 30 days. ................................................................................................................. 51
Figure 14. Actual and predicted sales receivable daily SARIMAX model prediction horizon 1 day ........................................................... .......................................................... 64
Figure 15. Actual and predicted sales receivable daily SARIMAX model prediction horizon 60 days. ..................................................................................................................... 65
Figure 16. Actual and predicted sales receivable daily SARIMAX model prediction horizon 90 days. ..................................................................................................................... 66
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>Auto-correlation Function</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike’s Information Criterion</td>
</tr>
<tr>
<td>AP</td>
<td>Accounts Payable</td>
</tr>
<tr>
<td>AR</td>
<td>Accounts Receivable</td>
</tr>
<tr>
<td>ARIMA</td>
<td>Autoregressive Integrated Moving Average</td>
</tr>
<tr>
<td>ARMA</td>
<td>Autoregressive Moving Average</td>
</tr>
<tr>
<td>B2B</td>
<td>Business-to-Business</td>
</tr>
<tr>
<td>CCC</td>
<td>Cash Conversion Cycle</td>
</tr>
<tr>
<td>DPO</td>
<td>Days Payables Outstanding</td>
</tr>
<tr>
<td>DSI</td>
<td>Days Sales Inventory</td>
</tr>
<tr>
<td>DSO</td>
<td>Days Sales Outstanding</td>
</tr>
<tr>
<td>ERP</td>
<td>Enterprise Resource Planning</td>
</tr>
<tr>
<td>MA</td>
<td>Moving Average</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>PACF</td>
<td>Partial Autocorrelation Function</td>
</tr>
<tr>
<td>SARIMA</td>
<td>Seasonal Autoregressive Integrated Moving Average</td>
</tr>
<tr>
<td>SARIMAX</td>
<td>Seasonal Autoregressive Integrated Moving Average with External Variables</td>
</tr>
<tr>
<td>WC</td>
<td>Working Capital</td>
</tr>
</tbody>
</table>
WCM Working Capital Management
1. INTRODUCTION

1.1 Motivation and objective of the study

Sales receivable management as part of working capital management has become an important topic for companies to monitor. Significance of proper working capital management for companies is twofold. On the one hand there is liquidity which should be kept at a certain level and on the other hand there is profitability which can be affected if too much money is tied up in working capital (Enqvist et al., 2014).

Vital to the financial health of a company, working capital and its components should be constantly monitored, hence, the decision to study sales receivable. Sales receivable and accounts receivable terms are used interchangeably in this report. Due to the nature of the business of the company of interest, individual invoices tend to be sizeable. Therefore, proper monitoring of accounts receivable is essential. To improve the monitoring, a model that exhibits customer payment behavior was built. The model is shortly introduced in the research methodology and presented later in the empirical part of the report. Additionally, to have a more proactive approach in working capital management a model to forecast outstanding accounts receivable was built. Forecasting with the second model is based on historical receivables and future sales estimates. Similarly, the model is shortly introduced in section 1.3 and presented in chapter 4. We study the importance of working capital and accounts receivable as one of its key components.

This thesis has two objectives. One of the objectives is to improve operating working capital, which is aimed at achieving by modeling customer payment behavior. Knowing payment behavior and a probability that the customer will pay the open invoices gives proper guidance to our collections management team to take timely actions with the correct customers. This in turn results in lower outstanding receivables and improves operating working capital. This process is depicted in figure 1.
The second objective is to provide insight on future working capital position to better support management decision making. To achieve this, a model that forecasts sales receivable level was built. The model was built with varying prediction horizons. Figure 2 presents the process.

**Figure 1.** Benefits of analyzing customer payment behavior.

**Figure 2.** Impact of sales receivable analysis.
1.2 Research questions

Based on the research objectives described above the following research questions were formulated:

1. Can customer payment behavior modeling provide useful information for collections management?
2. How accurately autoregressive model can forecast sales receivable?
3. How much autoregressive model forecasting accuracy improves when sales estimations are added as an external variable?
4. How does sales receivable forecasting accuracy based on customer behavior modeling compare to the autoregressive and autoregressive with external variables approach?

How these questions will be tackled, will be explained in the next subchapter.
1.3 Research methodology

To get a better understanding on accounts receivable, research on trade credit was conducted. In addition, based on the significant role that accounts receivable have on working capital and the interlink between the two, research was also conducted on working capital. Furthermore, short theoretical background on predictive modeling is provided. More specifically time series analysis, ARIMA modeling and Kaplan-Meier model which were used in the empirical analysis were studied. The purpose of the literature review on predictive modeling was to deepen the knowledge on the topic and familiarize oneself with these models’ applications in practice.

In addition, interview was conducted regarding current state of operating working capital and generalities of trade credit policy of the case company. Case company will onwards in the report be referred to as company. Moreover, data on company’s accounts receivable was collected. The data was collected from the company’s ERP, and in agreement with the company the data cannot be published with this thesis report. The empirical research was conducted in two parts.

In the first part, customer payment behavior was analyzed by utilizing survival analysis. Survival analysis is a statistical package, which is used when one is interested to know how long time it would take for an event to happen. In other words, the analysis studies “time to event” data, which means, study the data until the event occurrence. (Jager et al., 2008)

Event of interest in this survival analysis study is bill payment and it is measured in days. Survival analysis was conducted by fitting Kaplan-Meier model on customer payment data. Lifeline library of python was used to carry out the survival analysis. Lifeline is a python implementation of survival analysis. (Davidson-Pilon et al., 2019) Subsequently, customer payment behavior model together with sales estimates was used to create a model for predicting sales receivable. More on survival analysis with Kaplan-Meier and its implementation on customer payment behavior model will follow in 4.3 where the model is explained in more detail and 6.1 where results of this analysis are presented.
In the second part, more sophisticated models were fitted on sales receivable data for predicting future sales receivable level. Variations of ARIMA class models were utilized for this purpose. Autoregressive Integrated Moving Average models use past values of variable of interest and current and past own error terms to make predictions. SARIMA is a model that belongs to the ARIMA modeling group. Main difference between the two is that SARIMA accounts for natural seasonality in prediction, therefore, letter S added to the name.

Initially, sales receivable data was fitted to SARIMA model. After further testing, exogenous variable (sales estimates) was included in the model resulting in the usage of the SARIMAX model. Adding the exogenous variable (X) to the model resulted in better prediction accuracy. Both models were tested, and their results were compared and analyzed. In the end, based on result analysis and comparison, conclusions were drawn. Additional description on the models and result comparisons will follow in 4.2 and 6.3 respectively.

For SARIMA and SARIMAX functions statsmodels python library was utilized. Statmodels is a python statistical toolkit that is composed of many models that assist on different statistical analysis. Results of analysis are tested and compared using all existing models. (Seabold et al., 2010) More on SARIMAX model development will be provided in chapter 5.
1.4 Thesis structure

This thesis report is composed of six chapters in total. Chapter 1 as seen above, gives an introduction. In the introduction, motives and objectives of the study were laid out. In addition, research questions and research methodology were also presented in chapter 1.

Subsequently, in chapter 2, theoretical background on trade credit and working capital is presented. Most studied theories on trade credit and its usage are summarized and trade credit usage from supplier and customer perspectives are considered. Working capital definition, measurements of working capital and working capital determinants are also discussed in this chapter.

Chapter 3 shortly introduces the company. It also presents and summarizes definitions of trade credit and working capital from company’s perspective. In Chapter 4, short introduction to predictive modelling is presented. Additionally, theoretical background on time series analysis is presented. Furthermore, ARIMA and Kaplan-Meier models that are used in the empirical analysis are included and explained in this chapter.

In chapter 5, steps to SARIMAX model development such as: data description, feature selection, model selection and evaluation are described. In chapter 6 results of the models are described separately, then, compared and analyzed. Based on results’ comparison and analysis, research questions are answered, and conclusions are drawn. Limitations of the study and suggestions for further research are also indicated in chapter 6. For illustrative purposes on thesis structure, see figure 3.

![Thesis structure diagram](image)

**Figure 3.** Thesis structure.
2. THEORETICAL BACKGROUND

2.1 Trade credit

Trade credit is a B2B agreement, where a firm purchases goods agreeing with the supplier to pay for the purchase at later date. It is considered as an alternative source of external financing. (Klapper et al., 2012)

The aim of this sub-chapter is to present research study conducted on trade credit and give answers to questions such as: *Why trade credit is more advantageous compared to traditional borrowing? Who offers credit? Who receives credit?* The topic will be elaborated from two different viewpoints, from the supplier’s perspective (impact on its receivables) and from the buyer’s perspective (impact on its payables). Firms should be careful to not tie too much money in AR because it can be costly for the company, since it affects working capital decision, which in turn affects profitability and liquidity of the firm (for more details on this, see below in section 2.2.1 working capital management). However, firms should also be careful not to put too tight of a trade policy in place since that means foregoing future sales opportunities (Michalski, 2007). For a better understanding of trade credit relationship, see figure 4.

Due to its wide-spread usage, trade credit has been actively researched for more than 40 years and from different perspectives (Seifert et al., 2013). In the beginning researchers were eager to know why suppliers are willing to lend even to customers that are not able to get financing from financial institutions, and they did come up with conclusions. Firstly, because suppliers are interested in capturing possible future business opportunities from these customers. Secondly, it is cheap for suppliers to monitor their customers. Thirdly, they can easily rely on the possibility to repossess and resell goods in case of customer default. (Petersen & Rajan, 1996)
It has also been proved that trade credit is generally more expensive than credit from banks (Petersen & Rajan, 1996), (Bougheas et al., 2008). Trade credit is associated with higher implicit interest rate (rate derived from comparison of cash discounted payment amount versus the base payment amount at invoice due date) because it accounts for a default and insurance premium. Default premium to count for the fact that supplier is lending even if banks are not, and insurance premium to count for possibility of future needs for liquidity. (Cunat, 2007)

(Ng et al., 1999), (Bougheas et al., 2008), (Seifert et al., 2013) claim that usage of trade credit and its terms differ across industries, countries, sometimes even across customers within the same industry. They proved that terms differ more across industries but less within an industry. In Finland for example, average maturity of trade credit receivables and payables is 30 days (Ferrando & Mulier, 2013). Several studies have concluded that trade credit is used more in developing countries where legal systems are weaker and financial markets are less developed (Demirguc-Kunt & Maksimovic, 2002), (Frank & Maksimovic, 2005), (Giannetti et al., 2011) and (Ferrando & Mulier, 2013).

Figure 4. Trade credit relationship (Adapted from Petersen & Rajan 1997).
In addition to using more traditional financing means available, firms sometimes choose to be financed by their suppliers in the form of trade credit. Trade credit is usually utilized for short term borrowing, but frequently, when credit is not obtainable from financial institutions, or firms are running out of bank credit, it is used as finance of last resort for medium term (Petersen & Rajan, 1996). Similar findings are presented in (Rajan & Zingales, 1995), (Giannetti, 2003), (Fisman & Love, 2002), (Ferrando & Mulier, 2013) and (Lawrenz & Obendorfer, 2018) which emphasize smaller-sized firms in using trade credit. In a study conducted by (Rajan & Zingales, 1995) trade credit appeared as one of the most important sources of short-term external financing in United States.

Before getting into further elaboration on trade credit, it is essential that we define trade credit terms. Credit terms are parameters that are associated with purchases on account, and are as the following: trade credit period (number of days the buyer can delay the payment, starting from the delivery date), discount rate (the rate of cash discount offered by the supplier) and discount period (number of days within which the buyer can utilize the offered discount) (Michalski, 2007), (Giannetti et al., 2011) and (Cunat, 2007). Length of trade credit period is the fundamental parameter of credit terms decision (Seifert et al., 2013). Standard form of a trade credit contract is for example: “2-10 net 30”, which means that the customer is offered 2% cash discount if it chooses to pay within the discount period, which is 10 days, or else it can pay within 30 days after the delivery has happened (Cunat, 2007).

Trade credit terms are important decisions that are dependent on factors like: market competition, type of the good supplied, demand elasticity, price and type of the customer. Furthermore, to make an informed decision, suppliers are advised to acquire information on customer’s inventory conversion cycle and receivables conversion cycle as these items form customer’s operating cycle, and the shorter this cycle the shorter the collection period for the supplier (Michalski, 2007).
2.1.1 Theories

According to available literature, some of the most studied theories of trade credit usage are financing advantage theory, price discrimination theory, transaction cost theory, quality guarantee theory, customized product theory.

*Financing advantage theory* suggests that there are more advantages that a supplier possesses in offering credit to a customer, rather than financial institutions. Firstly, there is an advantage in information acquisition. While financial institutions gather information about the customer, the supplier will get the information faster and with lower cost because information gathering is done on the course of the business (Petersen & Rajan, 1996) & (Mian & Smith, 1992). A main source of information on customer’s financial position usually is the time and size of the order. In addition, more knowledge is gained through ability or inability of the customer to use the offered discount (Petersen & Rajan, 1996) and (Fisman & Love, 2002).

Secondly, supplier has more control power over the customer than a bank has. Supplier is powerful when it comes to taking actions that have an immediate effect on operations of a customer, by cutting off the supply. This is a valid threat if there are not many alternative suppliers available for the goods supplied. However, even if there are any, switching costs are usually higher, therefore, customer is not willing to break up with the supplier. While a bank may be constrained by bankruptcy laws if it wishes to withdraw past finance from a customer, the supplier does not necessarily have to follow the same procedures. (Petersen & Rajan, 1996), (Cunat, 2007) and (Giannetti et al., 2011)
Thirdly, supplier has a comparative advantage over the bank in liquidation. Numerous researchers suggest that due to their established network of customers, it is faster and cheaper for the supplier to salvage the goods from the customer and resell them in case of customer’s default. It is important to mention that the resale is dependent on the durability and the level of transformation of the goods. The more durable and the less transformed the goods are, the better the collateral they provide. While a bank could also resell the goods, it would be costlier and time consuming for it to find an alternative buyer. (Mian & Smith, 1992) and (Petersen & Rajan, 1996)

*Price discrimination theory* suggests that trade credit takes place even if a supplier does not have any of the financial advantages listed above compared to financial institutions. Trade credit can therefore occur when suppliers want to price discriminate towards certain customers, since this is considered illegal practice if done directly. (Mian & Smith, 1992), (Petersen & Rajan, 1996), (Brennan et al., 1998), and (Frank & Maksimovic, 2005)

Suppliers tend to price discriminate towards risky customers for various reasons. (Fisman & Love, 2002) assumes that one reason could be low competition in a market where the demand elasticity of risky customers is higher than that of more creditworthy customers, or if there is adverse selection in the credit market. Furthermore, suppliers may price discriminate because they are certain that risky customers will still borrow from them, since they might be the cheapest source of funding available. Creditworthy customers usually pay as soon as possible once they realize that trade credit is overpriced.

*Customized product theory* is another theory that supports trade credit extension. Theory suggests that trade credit is advantageous especially if a supplier is producing some tailor-made product for the customer, since this way the customer is less willing to delay payment (Fisman & Love, 2002). Same argument is used by (Giannetti et al., 2011) who associate trade credit with the nature of the product being supplied. They argue that suppliers of differentiated goods extend more trade credit than those of standardized goods. Being a supplier of a customized product, a supplier holds an advantage over banks since in case of customer default, the supplier would easier deal with the resale as it is part of the business.
*Transaction cost theory* implies that trade credit is also used for cost minimization reasons. (Ferris, 1981) was amongst the first ones to mention this in his paper. According to him, firms prefer trade credit since they will be able to save on paying bills. Instead of having to transact with each delivery, a firm could accumulate bills and pay them monthly or quarterly, depending on the credit terms. However, (Frank & Maksimovic, 2005) counter argues the previous point claiming that by now we should have experienced a huge decrease in trade credit usage because of improvements in transaction technologies, but evidently no such thing has happened.

Inventory related cost is another item related to this theory. Suppliers need to carefully choose customers (time and location wise) that they want to serve, to better manage their inventories. Proper inventory management can result in lower warehousing costs (Emery, 1987). (Petersen & Rajan, 1996) holds the same view, suggesting that firms that sell seasonal products can maintain smooth production cycles only if they wisely manage their inventories. This theory was complemented with a study conducted by (Bougheas et al., 2008), who studied trade credit, putting extra emphasis on inventory. They also agree that offering trade credit can help firms reduce inventory costs. Nevertheless, they suggest that production decisions play a key role in reducing inventory costs. When production exceeds sales, inventory levels increase and thus firms are more willing to offer trade credit. Nonetheless, they find evidence that inventory costs differ notably across industries and firms. Inventory costs tend to reduce with firm size, therefore, as the firm enlarges in size, inventory costs have little impact in trade credit extension decision.

*Quality guarantee theory* proposes that suppliers extend trade credit as a signal that their product is qualitative, by giving their customers enough time to test their product. (Lee & Stowe, 1993), (Emery & Nayar, 1998), (Fisman & Love, 2002) and (Frank & Maksimovic, 2005)
2.1.2 Supplier’s perspective

Decision of whether to extend trade credit or not, is a compromise between lowering the risk of payment delays from risky customers and increasing sales by gaining new customers through a more liberal trade policy (Michalski, 2007). Suppliers that have better access to credit from financial institutions appear to extend more trade credit to their customers. Additionally, willingness to extend trade credit increases with firm size and age. Larger and older firms extend and borrow more through credit than smaller and younger ones. Another determinant of credit extension for the suppliers can be the change in their sales. Firms that are living in good financial times where sales are growing, tend to give more credit to their customers. Firms that have stable sales extend less credit compared to the previous. On the other hand, firms experiencing sales decline offer more than firms with stable sales, and the explanation behind this is that, they extend more credit when facing hard financial times, to be able to stay in the business. Suppliers also prefer to give credit to high credit quality customers, but these customers use less trade credit because they usually have access to credit from banks, and because it is cheaper, they utilize that. (Petersen & Rajan, 1996)

To complement the research in the area, (Giannetti et al., 2011) proved that willingness of the supplier to give credit depends on the nature of the goods being supplied. Their results showed that suppliers of differentiated products offer more credit than those of standardized products. Quality issues associated with differentiated products suggest that customers will need time to inspect the goods, therefore, suppliers will extend credit to them (Smith, 1987). Sometimes suppliers might extend credit even to risky customers, because they do not want to lose customers in the long run, and they believe that the value of their relationship is higher than the cost of helping them. Therefore, they are inclined to help their customers when they are facing financial difficulties (Wilner, 2000) and (Cunat, 2007). Occasionally, when the demand is uncertain, suppliers prefer to extend credit even to risky customers rather than have their inventories increase (Bougheas et al., 2008). According to (Giannetti et al., 2011), large and high credit quality firms are thought to have more bargaining power and are offered more credit on better credit terms.
2.1.3 Customer’s perspective

In their study (Petersen & Rajan, 1996) list a few determining factors about the customers who qualify for receiving credit. One of the main factors is credit quality of the customer. In addition, relationship of the customer with financial institutions is perceived important too. Although, the relationship on its own does not affect the volume offered to the customer, but if the customer has recently been denied credit from bank, the credit offered from the supplier will be lower. On the other hand, firms that do not have any relationships with banks receive more credit (own higher accounts payables) (Fisman & Love, 2002). Most importantly of all, relationship with the supplier is the one that leads the credit decision.

Opposing supplier’s point of view, young firms need more credit since they have not established proper creditworthiness and reputation and do not have access to other financing forms, whereas creditworthy customers borrow less through trade credit (Cunat, 2007) and (Frank & Maksimovic, 2005). In terms of size and age (Ferrando & Mulier, 2013) found out that, larger and older firms with more collateral have easier access to trade credit. Even though they need it less, they do not hesitate to utilize it as a cash management tool. By stretching their accounts payable they are better able to manage their cash flow needs. They also show that firms that receive credit from their suppliers are more willing to do the same towards their customers. Nevertheless, despite firm’s wish to receive credit, it can happen that the wish does not meet supplier’s willingness to lend. (Cunat, 2007) suggests that trade credit grows as the relationship with the customers gets older and tighter.
2.2 Definition of working capital

(Brealey et al., 2001, 167-170) in line with most available literature, define working capital as *current assets* minus *current liabilities*. Items that belong to current assets include *accounts receivable (AR), inventory, cash* and *marketable securities*. Current liabilities on the other hand consist of *accounts payable (AP)* and other *short-term borrowing*. For a more detailed view, refer to figure 5, which displays a sample of a typical balance sheet.

A crucial portion of current assets is usually represented by inventory and accounts receivable, whilst accounts payable usually represent a big part of current liabilities (Deloof, 2003). Remaining items of current assets and liabilities do not directly affect operating working capital. Therefore, they are taken into consideration when financial decisions are at stake (Hawawini et al., 1986). The term “current” in this context represents a timeframe equal to one year or less. Conforming to the above definition, current assets are assets that are expected to be converted into cash within a year, whereas current liabilities are short-term liabilities that are due and expected to be cleared within a year (Raheman & Nasr, 2007).

Working capital is the capital employed by a company to finance daily operations (Atseye et al., 2015), (Enqvist et al., 2014). In agreement with the above definition of working capital, (Ding et al., 2013), (Fatimatzahra & Kusumastuti, 2016) and (Atseye et al., 2015) regard working capital as a measure of a company’s liquidity. (Atseye et al., 2015) refers to working capital as the “business wheel” or “circulating capital” which changes form (from cash to inventory, from inventory to receivables) during business days. Working capital and net working capital terms are used interchangeably in literature. However, there is difference between gross working capital and net working capital. Gross working capital indicates the total of a company’s current assets, while net working capital equals the difference between current assets and liabilities.
Different authors suggest different approaches on following working capital. According to (Knauer & Wöhrmann, 2013), the only advantage of using net working capital instead of gross is that current liabilities are moderately related to inventories and their transformation into accounts payables. Hence, if net working capital is used, the effect of accounts payable into working capital financing is not taken into consideration.

Despite its important role in financial decision making, research on working capital has been neglected because it concerns short term financing. Most studies on corporate finance have been focused on long term financing. (Garcia, 2011)

<table>
<thead>
<tr>
<th>ASSETS</th>
<th>LIABILITIES AND SHAREHOLDER'S EQUITY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current assets</strong></td>
<td><strong>Current liabilities</strong></td>
</tr>
<tr>
<td>Cash</td>
<td>Accounts payable</td>
</tr>
<tr>
<td>Marketable securities</td>
<td>Notes payable</td>
</tr>
<tr>
<td>Accounts receivable</td>
<td>Accrued expenses</td>
</tr>
<tr>
<td>Inventory</td>
<td>Other current liabilities</td>
</tr>
<tr>
<td>Prepaid expenses</td>
<td></td>
</tr>
<tr>
<td>Other current assets</td>
<td></td>
</tr>
<tr>
<td><strong>Property and equipment</strong></td>
<td><strong>Long-term financing</strong></td>
</tr>
<tr>
<td>Land</td>
<td>Long-term debt</td>
</tr>
<tr>
<td>Buildings</td>
<td>Equity financing</td>
</tr>
<tr>
<td>Equipment</td>
<td></td>
</tr>
<tr>
<td><strong>Other assets</strong></td>
<td>Intangible assets</td>
</tr>
<tr>
<td><strong>Total assets</strong></td>
<td><strong>Total debt and equity</strong></td>
</tr>
</tbody>
</table>

**Figure 5.** Typical balance sheet.
2.2.1 Working capital management

Different sources present slightly different definitions for working capital management. (Hofmann et al., 2011, 13) states that working capital management (WCM) attempts to address issues related to planning, movement and control of current assets and current liabilities and the relationship between the two. (Eljelly, 2004) adds to the previous and defines WCM as a process that involves planning of current assets and liabilities, taking into consideration the risk of failure to meet daily obligations and the possibility of over-investing in current assets and liabilities. Other authors such as: (Malmi & Ikaheimo, 2003), (Garcia, 2011), (Nimalathason, 2010), and (Mansoori & Muhammad, 2012) conform to this definition of WCM.

Studies have shown that efficient working capital management is vital to the financial health of any company. Management of working capital is important because it directly affects the profitability of a firm (Smith, 1987), (Shin & Soenen, 1998), (Deloof, 2003), (Ukaegbu, 2014), (Enqvist et al., 2014) as well as its liquidity (Smith, 1987), (Kim et al., 1998), (Ding et al., 2013), (Fatimatuzzahra & Kusumastuti, 2016).

(Knauer & Wöhrmann, 2013) further explain how WCM can affect both profitability and liquidity of a firm. They suggest that liquidity is affected if cash inflows from customers are delayed because of trade credit policy but on the other hand, payments to suppliers need to be made. Similarly, they explain that firm profitability is affected by WCM through sales and capital employed. Therefore, firms aim to achieve an optimal working capital, where they adequately balance between liquidity and firm profitability (Mansoori & Muhammad, 2012), (Enqvist et al., 2014), which according to (Zariyawati et al., 2009) is a hard task for managers to perform.

Closer attention should be paid to liquidity, since too much liquidity harms profitability of a firm, whereas too little of it means that a firm’s current liabilities exceed its current assets, which in turn can lead a firm towards bankruptcy (Smith, 1973) and (Kieschnick et al., 2012).
A study on WCM (Ukaegbu, 2014), lists three approaches that firms employ to manage their working capital management. The *conservative approach* where a company uses its current assets only in critical conditions, but mostly uses its fixed assets to finance its operations. The *aggressive approach* suggests in keeping smaller portion of current assets compared to total assets. The *moderate approach* suggests that current assets should be used to finance daily operations of a firm.

Firms are advised to preserve relevant levels of working capital. Exaggerated investment in working capital means that funds are locked up in cash and this has negative impact on profits (Deloof, 2003), whereas low investment in working capital can cause disruptions in production. From this we establish that deciding on levels of working capital comprises a trade-off between a firm’s profitability and its liquidity. (Banos-Caballero et al., 2014) indicates that increase in working capital goes hand in hand with firm’s performance until a certain point, and beyond this point, further increase in working capital will have a negative effect in firm performance.

Most of studies conducted on working capital have tried to capture relationship between working capital and profitability of the firm. Majority of these studies perform their analysis using panel data with a time frame between 4 – 20 years (Knauer & Wöhrmann, 2013). (Shin & Soenen, 1998) was amongst the first studies to investigate such interrelation. The study concluded that there existed a strong negative relationship between net trade cycle, which was used as a measure for management of working capital of firms and their profitability. This suggests that the shorter net trade cycle the higher the profitability.

In a similar manner (Deloof, 2003), studied the relationship between cash conversion cycle (CCC) used as a measure of working capital management and gross operating income which was used as measure of profitability of the firms. In line with the previous study, this one study found that there was a strong negative relationship between CCC and profitability.
Many other studies such as: (Lazaridis & Tryfonidis, 2006), (Raheman & Nasr, 2007), (Zariyawati et al., 2009), (Mathuva, 2009), and (Gill et al., 2010) came to similar conclusions, that there exists a negative relationship between a firm’s CCC and their profitability.

### 2.2.1 Determinants of working capital

(Banos-Caballero et al., 2010) and (Atseye et al., 2015) state that working capital is determined by internal firm-specific factors such as: size, age, profitability, growth opportunity, leverage, industry etc., and by external factors such as: GDP, interest rate and tax rate.

Growth opportunity affects working capital management because when companies foresee sales growth, they tend to increase inventories. Studies have shown that company size can be a determinant of working capital. (Banos-Caballero et al., 2010) explained that working capital requirement tends to increase with firm size. The study implies that smaller companies have less receivables and inventories since the costs are higher, whereas for bigger companies the opposite holds.

In addition, we have observed that age of the company affects the possibility of financing. Therefore, it can influence working capital requirements (Chiou et al., 2006). Empirical research has shown that industry as well has an impact in working capital. (Filbeck & Krueger, 2005) concluded that different industries employ different trade credit policies and have different inventory requirements.
2.2.2 Measures of working capital management

Most widely used measure of working capital management is cash conversion cycle (CCC), which originally was introduced in 1980 by Richards & Laughlin. CCC measures the time from when the company pays its suppliers for raw materials until the moment when the company receives money from its customers by selling the products that it makes (Ding et al., 2013). This measure is usually used by companies when they want to compare their cycle with previous years, or when they want to compare themselves with competitors (Fatimatuzzahra & Kusumastuti, 2016).

CCC therefore, is expressed using measures of three components such as: accounts receivable conversion period, inventories conversion period and accounts payable deferral period (Enqvist et al., 2014), and its formula is as follows:

\[
CCC = DSO + DSI - DPO
\]  

(1)

(Hofman & Kotzab, 2010) listed definitions and formulas of CCC components:

Days sales outstanding (DSO) – measures the number of days from when a product is sold until the money for that product is collected, and it is expressed using the following:

\[
DSO = \frac{Accounts \ receivable}{Sales} \times 365
\]  

(2)

Days sales inventory (DSI) – measures the number of days it takes a company to convert its inventory (including work in progress) into product sales, and its formula is the following:

\[
DSI = \frac{Inventory}{Cost \ of \ goods \ sold} \times 365
\]  

(3)
Days payables outstanding (DPO) – measures how many days it takes a company to pay its suppliers after having purchased from them, and its formula is the following:

$$DPO = \frac{Accounts\ payable}{Cost\ of\ goods\ sold} \times 365$$

Earlier in 2.2.1 we noted that keeping sound levels of working capital within a company means being cautious about the liquidity and profitability trade-off. Therefore, companies aim at shortening their CCC as much as possible, which is reached by speeding up receivable collections, decreasing inventory conversion period and stretching payables as much as possible. (Enqvist et al., 2014)

However, companies should be careful when trying to reduce CCC. They could harm profitability because speeding up receivables collection could worsen relationship with customers and stretching payables could harm relationship with suppliers and damage their reputation. (Garcia, 2011)
2.3 Comparison against previous research

Similar research was conducted previously by (Zeng et al., 2007). In their study the authors built a model to forecast whether a new invoice will be paid or not. The objective of the study was reduction of outstanding receivables through improvements in collections methods. Authors used C4.5 decision tree induction machine learning algorithm. They used data sets that covered invoices for one year from four companies. Features that were used to build the model were invoice base amount, payment terms and invoice category. These features were the independent variables of the model while payment time was the dependent variable. The built model was able to predict if a newly created invoice will be paid on time or not, and if not, provide the length of the delay. With this model they were able to tailor collection strategy per customer with less manual effort. Subsequently, they were able to improve accuracy and reduce costs.

In this empirical study SARIMAX model was used to predict the total level of outstanding receivables and Kaplan-Meier analysis to predict payment outcome of individual customers. Data set used in the study was actual invoicing data from a company. When compared to the previously mentioned study, similarity was found in the objectives of the studies, even though the results are not comparable because they answer slightly different research questions. Objective of the previous study was to improve collections management through more advanced methods and subsequently reduce level of outstanding receivables. In this thesis work two aligned objectives were set. Firstly, the customer model Kaplan-Meier aimed at reducing outstanding receivables by improving collections management. Secondly, SARIMAX model aimed at providing better support for management decision making.

However, differences as mentioned previously can be noted on how the objectives were reached. In this study two models were built because two aligned objectives were at stake: predicting payment outcomes of customers individually but of the same interest was predicting the overall level of outstanding receivables. On other hand, the previous study’s focus was improving collections management.
3. CASE COMPANY

Company x (later in the report referred to as “company”) is a large forest industry company that manufactures pulp, paper and other wood-based products. It is a global company, headquartered in Helsinki. The company is composed of a few business units and the empirical research of this thesis concerns one of them. Most of the sales of the selected business unit come from large individual customers.

Individual invoices tend to be sizeable due to the nature of the business. Therefore, proper monitoring of accounts receivable is essential. In order to improve the monitoring a model that exhibits customer payment behavior was built. The model is presented in the empirical part of the report. Additionally, to have a more proactive approach in working capital management a model to forecast outstanding receivables was built. Forecasting with the model is based on historical receivables and future sales estimates.

3.1 Trade credit in the case company

The company conducts most of its business on credit. Out of total sales, 80% of transactions are conducted on credit. The majority of the remaining 20% is conducted through letter of credit, and a small portion is cash in advance transactions.

As suggested by literature, credit terms do differ across customers in the company. Credit terms are x amount of days on a certain point, for example, 45 days on bill of lading. Usually it is the number of days against bill of lading (proof that the goods have been delivered).

Trade credit period offered in contracts varies across customers and it can be as short as 7 days to as long as 120 days, and this variation depends on the customer. The most important customer related factor when negotiating credit terms, is their financial stability. Customers who are financially strong are offered longer credit period, better discount rate and longer discount period. On the other hand, customers with weaker financial stability will end up with less favorable credit terms.
Another key factor on contracted credit terms noted by the company was the payment term norm in the customer’s country of residence. In certain regions the norm might be 30 days in others 120 days. In Germany for example, most customers have credit period in a range of 30 - 60 days, and it is acceptable to the company because that is considered the norm of the country. On the other hand, in Italy, Spain or France, trade credit period of 120 days is acceptable.

Credit terms of customers with one-year contractual agreements are reviewed once a year, when the negotiation for next year’s contract takes place. Usually, credit terms do not change every year, however, the opportunity to change them is available once a year. For customers who have multiyear contracts, credit terms do not change annually but can be changed when the contract is updated.

3.2 Working capital in the case company

Company’s definition of working capital is in line with the literature presented in section 2 of chapter 2. Absolute OWC figure is calculated based on equation (5):

\[ \text{OWC} = \text{Total Inventories} + \text{Accounts Receivable} - \text{Accounts Payable} \] (5)

Working capital management is conducted in a similar fashion as in other companies in the same industry. Company manages WC by optimizing inventory levels, by speeding up receivables collection, and by stretching payment period of payables as much as possible. This definition conforms to the CCC equation. CCC is a formula expressed in days which measures working capital, and the objective of any company is to keep CCC as short as possible. For more details on this, see section 2.2.2 measures of working capital management. Since working capital position depends on receivables and payables as well as inventory, the company is always cautious when negotiating credit terms with their customers and suppliers.

Keeping sound levels of WC is recognized by the company management as an important objective. Therefore, to raise awareness and accountability across the whole organization, WC is linked to employees’ bonuses. This indicates that WC is listed amongst the most important KPIs.
4. PREDICTIVE MODELING

In this chapter there will be a short introduction to predictive modeling, a deeper review of time series analysis, specifically covering family of ARIMA models, and a short background on survival analysis model Kaplan – Meier.

(Kuhn & Johnson, 2013) define predictive modeling as the process of establishing a mathematical model or mechanism that makes accurate forecasts, with the possibility of interpreting and evaluating the model’s accuracy on these forecasts. In line with the above definition, two main objectives of predictive modeling are generating accurate forecasts and model interpretation.

It is of high importance to emphasize the trade/off one makes between complexity and accuracy when building a predictive model whose main goal is its performance accuracy. Aiming for higher model accuracy can make the model more complex and harder to understand. Another critical point to remember when embarking a journey in predictive modeling is understanding the distribution of the variable we are predicting. That is the first and most important step of the journey and an essential point when the separation of the data into training and testing sets is done. (Kuhn & Johnson, 2013, 4-11)

4.1 Time series analysis

A time series is a set of data points each being recorded at a specific time t. Time series can be discrete and continuous. Main difference between the two is that, in discrete time series the data points are recorded at fixed intervals, while in continuous time series data points are recorded continuously (over a time interval). (Brockwell & Davis, 1991, 1)

Time series models are a category of models that try to predict variables based on their past values and/or current and past values of their own error term. Time series analysis are used to observe trends and identify patterns and based on those, ultimately make predictions on what is going to happen in the future.
Based on this definition a distinction between time series analysis and multivariate models can be made. Multivariate models try to make predictions on a variable of interest using the effect of other explanatory variables on the variable of interest, whereas time series analysis as mentioned above, makes predictions based only on the series’ own past values. (Brooks, 2014, 251)

Most commonly used family of models in time series analysis is the autoregressive integrated moving average (ARIMA) models, which will be explained in the following. However, before analyzing ARIMA models, some important concepts will be presented, which are essential in understanding how to proceed with modeling using ARIMA.

4.1.1 Data processing and filtering

Major part in time series analysis involves processing data, by changing attributes of the series, getting rid of all the signals and being left with only noise, preparing the data for modeling.

Stationary process: a series is strictly stationary if the distribution of its values does not change over time. A weakly stationary process needs to have a constant mean, variance and autocovariance structure, in other words, it needs to satisfy the following equations:

\[ E(\gamma_t) = \mu \] \hspace{1cm} (6)

\[ E(\gamma_t - \mu)(\gamma_t - \mu) = \sigma^2 < \infty \] \hspace{1cm} (7)

\[ E(\gamma_{t1} - \mu)(\gamma_{t2} - \mu) = \gamma_{t2} - t1 \forall t1, t2 \] \hspace{1cm} (8)

If the series is not stationary, we can difference it to make it stationary. (Brooks, 2014, 252)

Examples of stationary and non-stationary time series are given in figures 6 and 7.
Seasonality: in time series, are events that happen at regular intervals, for example, same month if observations are monthly or same year if they are yearly. Naturally, seasonality in modeling should be counted for, therefore, it is subtracted from the series. (Brooks, 2014, 493)
**Autocorrelation**: reflects the influence one value has in its subsequent value in a time series. Autocorrelation is used to detect non-randomness in data and helps in choosing an appropriate model in time series if the data is not random. It is the same as correlation between two random variables, except that in time series correlation of a series is measured between the series itself at different lags (past periods). (Brooks, 2014, 680)

**White noise process**: a series can be white noise if all trends have been eliminated, seasonality and autocorrelation. It means that every observation has similar variance and 0 correlation with all other observations in the series. A white noise process satisfies the following equations:

\[ E(\gamma_t) = \mu \]  
\[ \text{var}(\gamma_t) = \sigma^2 \]  
\[ \gamma_{t-r} = \sigma^2 \text{ if } t = r; 0 \text{ otherwise} \]  
(Brooks, 2014, 696)

**Smoothing**: is another filter that can be used in time series, for clearer image of the data. It is generally done to reveal patterns better in the series. For example, if there is seasonality present, series can be smoothed, and the trend can be observed afterwards. Smoothing can be exponential and moving average. (Brooks, 2014, 283)

### 4.2 ARIMA modeling

An **Autoregressive (AR) model** is a model where the value of \( y \) that we are trying to predict, depends only on the past values that \( y \) took previously and an error term. The order of an autoregressive model depicts the number of preceding values in the series that are used to predict \( y \). Therefore, a second order autoregressive model AR (2) can be expressed as:

\[ y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + u_t \]  
where \( y_t \) is the value that is being predicted, \( y_{t-1}, y_{t-2} \) are the two previous lags, \( \phi_1, \phi_2 \) correspond to the weights or the importance of the lags respectively in predicting the new variable and \( u_t \) is a white noise term.
Stationarity condition is an important feature in AR models because if the series is non-stationary it means that the prediction is made assuming that previous value of the error term will have equal effect on the predicted value, which usually is not the case. (Brooks, 2014, 259-267)

A Moving Average (MA) model is one of the simplest of the ARIMA family. Unlike autoregressive models that use past values of the variable of interest for prediction, it uses error terms of current and past values to predict the future. The order of a moving average model depicts the number of past error terms used in forecasting. For example, a second order of moving average MA (2) can be expressed as follows:

\[ y_t = \mu + \theta_1 u_{t-1} + \theta_2 u_{t-2} \]  \hspace{1cm} (13)

where \( y_t \) is the value that is being predicted, \( u_t, u_{t-1}, u_{t-2} \) are the current and two previous lags of error terms, \( \theta_1, \theta_2 \) reflect the weights or the importance of the error terms’ lags respectively in predicting the new variable. Conditions that a moving average model needs to meet are: constant mean, constant variance and autocovariances which will be non-zero up to a lag and zero thereafter. (Brooks, 2014, 256)

An Autoregressive Moving average (ARMA) model is a combination of AR and MA models, that owns features of both AR and MA and has \((p,q)\) parameters. Therefore, to make predictions, this model uses both, past values of the variable of interest and its current and past own error terms. By combining AR (2) and MA (2) equations presented above we get an ARMA (2,2) like below:

\[ y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 u_{t-1} + \theta_2 u_{t-2} + u_t \]  \hspace{1cm} (14)
To determine the right order of ARMA model, understanding autocorrelation function (acf) and partial autocorrelation function (pacf) terms is essential. Acf and pacf help in determining the right order of the lags in moving average and autoregressive processes respectively. Acf shows how well a time series is related to its past values and it takes into consideration trends and seasonality. Pacf is a kind of autocorrelation but because it is conditional it is called partial. It is used to describe correlation between current values and values at different lags (periods ago), after controlling for intermediate lags. For example, pacf for lag 4 measures correlation between \( y_t \) and \( y_{t-4} \) after controlling intermediate lags like: \( y_{t-1}, y_{t-2}, y_{t-3} \). However, it is important to remember that acf and pacf can be plotted and observed only if the series is stationary. (Brooks, 2014, 268)

Another method for appropriate model parameter selection is Akaike Information Criterion (AIC), which is used in this study’s hyperparameter selection and it is expressed with the following formula:

\[
AIC = -2 \log(L) + 2m
\]  

(15)

where \( L \) is the likelihood of the data and \( m \) is the number of parameters. (Hyndman & Athanasopoulos, 2014, 232)

An Autoregressive Integrated Moving average (ARIMA) model is different from ARMA model because it includes differencing in it. Therefore, an ARIMA model is composed of \((p,d,q)\) parameters, where \( p \) is AR order, \( d \) is differencing order needed for stationarity and \( q \) is MA order. Differencing is employed in a series when the aim is to understand period to period change. (Brooks, 2014, 276)

A Seasonal Autoregressive Integrated Moving Average SARIMA model is another class of models that belongs to ARIMA family. The difference between simple ARIMA and SARIMA is the added letter S, which stands for seasonality and is attached to account for natural seasonality in the series. (Vagropoulos et al., 2016)
Differently from ARIMA, SARIMA models have two sets of parameters: \((p,d,q)(P,D,Q)s\) where \((p,d,q)\) is the non-seasonal part of the model and \((P,D,Q)s\) is the seasonal part of the model. What these parameters stand for was explained above in the ARIMA section. However, the additional parameter \(s\) represents the seasonal component. For example, SARIMA model with parameters \((1,1,1)(1,1,1)_4\) can be expressed with the following equation:

\[
(1 - \phi_1 B)(1 - \Phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 + \theta_1 B)(1 + \Theta_1 B^4)\varepsilon_t
\]

(16)

Where:

\((1-B) = y_t - y_{t-1}\): back-shifting time series by one period (non-seasonal differencing);

\((1-B^4) = y_t - y_{t-4}\): back-shifting time series by four periods (seasonal differencing);

\((1 - \phi_1 B)\): we account for time series one period ago in the prediction (non-seasonal AR term)

\((1 - \Phi_1 B^4)\): back-shifting time series by four periods (seasonal AR term)

\((1 + \theta_1 B)\): we account for error in time series one period ago in the prediction (non-seasonal MA term)

\((1 + \Theta_1 B^4)\): back-shifting error in time series by four periods (seasonal MA term)

(Hyndman & Athanasopoulos, 2014, 242)

A SARIMAX model is a supplement of SARIMA model explained above. Letter X added to the name stands for exogenous. Thus, a SARIMAX model is a multivariate model of SARIMA, which is capable of accounting for exogenous explanatory variables into modelling, to increase model’s predictive performance. (Vagropoulos et al., 2016)

Sales receivable display a periodic phenomenon; therefore, SARIMAX model is chosen for prediction purpose of this research study. In addition, SARIMAX is the only actively maintained and developed univariate autoregressive modelling function in Statsmodels library, which suggests that it should be used instead of AR function even when one wants to do a simple order one AR (1) prediction. In which case the choice of parameters would be \((1,0,0), (0,0,0)\).
4.3 Kaplan – Meier model

Kaplan-Meier is a non-parametric model and it is the most popular model used to perform survival analysis. It was first introduced in 1958 by Kaplan & Meier from where it also got the name. The purpose of survival analysis is analyzing and modeling “time to event” data, where time is an outcome variable until the occurrence of the event of interest. The event of interest can vary depending on the field of the study and time to event can be measured in days, weeks, months or years, again depending on the type of the study. This kind of analysis is most widely used in medical research, but its application has been useful in other fields such as economics, marketing, insurance etc. (Rich et al., 2010)

This study’s event of interest is bill payment and time to the event is measured in days. To perform survival analysis in this study, Kaplan-Meier estimate was used, which is expressed as in equation (17):

\[
\hat{S}(t) = \prod_{i=1}^{t} \left(1 - \frac{d_i}{n_i}\right)
\]

Where \(\hat{S}\) is the probability at time \(t\) that the customer will pay in the future if it has not paid until time \(t\), \(d_i\) is the number of paid invoices at time \(t\) and \(n_i\) is the number of unpaid invoices at time \(t\). (Cleves et al., 2008, 93)

In this thesis, Kaplan-Meier model’s use is twofold. First, the model is utilized to estimate how much is the probability that the customer will pay the amount owed if it has not paid until now. The procedure is repeated for all customers, using formula presented in equation 17. More on the practicality of the model and its result will follow in 6.1.

Secondly, with the help of the customer model and sales estimates combined together a model for forecasting receivables is built. This model can forecast overall behavior of the sales receivable well, but it lacks the ability to forecast trends and periodic fluctuations present in sales receivable. However, this model was presented to emphasize the importance of sales estimates into predicting sales receivables. More details on this will follow in 6.2. Mathematical representation of this model can be expressed as in equation (18):
\[ R(t) = \sum_{i \in O} P_{c(i)}(t)I_i + \sum_{c \in C} \sum_{i=0}^{t} P_c(t-1) E_c(i) \]

Where:

- \( O \) is the set of open invoices;
- \( C \) is the set of customers;
- \( R(t) \) is estimated receivables \( t \) days in future;
- \( P_{c(i)}(t) \) is the payment probability of the customer of the open invoice \( i \) \( t \) days in future;
- \( P_c(t) \) is the payment probability of the customer \( c \) \( t \) days in future;
- \( I_i \) is the monetary value of the open invoice \( i \);
- \( E_c(t) \) is the estimated day sales for the customer \( c \) \( t \) days in future;

Note that the customer payment probabilities \( P_c(t) \) in equation (18) can be expressed by using Kaplan – Meier model defined in equation (17):

\[ P(t) = \tilde{S}(t-1) - \tilde{S}(t) \]
5. DEVELOPMENT OF SARIMAX MODEL

5.1 Data description and feature selection

The original data set contains sales receivable amounts from January 4th 2016 until March 6th 2019. Data set has 21,882 rows and 183 columns. From 183 available features, in this analysis only 4 columns are used, namely: customer number, baseline date, payment date and normalized currency. Currency was normalized (changed to Euros) because the amounts of outstanding receivables were in different currencies. An additional feature was created for supportive purposes during analysis. The feature is calculated as (payment date – baseline date) and is used to express the current payment delay, as expressed in equation 19.

\[
\text{Current payment delay} = \text{Payment date} - \text{Baseline date}
\]

(19)

5.2 Data cleaning and preprocessing

In this step, selection of specific accounts and filters from the raw data set was performed. Fortunately, there was no missing data in the variables of interest, therefore, no interpolation action was needed. July 2018 was the division line between training and testing data sets. The split was done at the half year end and based on the nature of the data, the testing set was a representative sample of the data. In addition, the split was close to ideal recommendation 76/24, 76% used for training and 24% for testing purposes.

5.3 Model hyperparameter selection

Model was fitted to the training data in for loop, with a periodic length (seasonality) of 7 days and all other parameters were tested with values 0, 1 and 2 and the one with the lowest AIC score was picked. The set of values 0, 1 and 2 may sound small, but higher order autoregressive models are heavy and prone to overfit to the training data.

Model with the lowest AIC score had this set of parameters: \((p, d, q, P, D, Q, S) = (0, 1, 0) (1, 1, 1, 7)\). It is important to be noted that the same set of parameters were used in case of SARIMA and SARIMAX models. Parameters are explained in table 1 below.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>p = 0</td>
<td>Order of non – seasonal AR term</td>
</tr>
<tr>
<td>d = 1</td>
<td>Order of non – seasonal differencing</td>
</tr>
<tr>
<td>q = 0</td>
<td>Order of non – seasonal MA term</td>
</tr>
<tr>
<td>P = 1</td>
<td>Order of seasonal AR term</td>
</tr>
<tr>
<td>D = 1</td>
<td>Order of seasonal differencing</td>
</tr>
<tr>
<td>Q = 1</td>
<td>Order of seasonal MA term</td>
</tr>
<tr>
<td>S = 7</td>
<td>Seasonality term</td>
</tr>
</tbody>
</table>

Table 1. SARIMA and SARIMAX parameter description.
6. RESULTS AND CONCLUSIONS

6.1 Customer behavior analysis with Kaplan-Meier

With Kaplan-Meier time-to-event analysis it is possible to capture anomalies in customer payment behavior. In figure 8, average payment behavior of all customers is presented. Analysis is done with each invoice having the same weight, without taking into consideration the sales volume of individual invoices. X-axis in the graph represents days from baseline date of the invoice and Y-axis represents pay probability. Baseline date of the invoice is the date when the invoice is issued. The figure is a mixture of different payment terms and behaviors. From the figure it is derived that customers pay at different times from the baseline date. In addition, it is implied that it is rare to have payments after 100 days. This overall analysis is not useful if learning more about a specific customer’s payment behavior is at stake. Therefore, Kaplan-Meier analysis is more useful when utilized in a customer level. The analysis shows the normal behavior of the customer. Customer normal behavior can then be used to benchmark future payment behavior to detect anomalies.

![Payment behavior over all the customers](image)

**Figure 8.** Payment behavior over all the customers.
In figure 9 and 10 two examples of customer payment behavior are presented. Different shapes of the curves indicate different payment behaviors. Graph in figure 9 depicts a customer with a consistent payment behavior, where almost all invoices are paid around day 90, which is the due date based on this customer’s payment terms. Graph in figure 10 on the other hand depicts a customer whose payment behavior is less consistent. Less consistent because this customer has payment terms of 30 days and invoices are paid in a range of 10 – 60 days from baseline date.

**Figure 9.** Regular customer payment behavior.
Another difference among the graphs is the blue shaded area. Blue shaded area represents double-sided 95% confidence interval from the observed average. Confidence interval of 95% is described as a value range that has 95% probability that the range will contain the true unknown parameter (James et al., 2013, 66).

In graph in figure 9 confidence interval is visibly smaller compared to confidence interval in graph in figure 10, indicating that estimated average based on the observed data is more reliable. This kind of analysis is beneficial if anomalies in customer behavior are detected. It supports sales department in making timely delivery stops and taking a more proactive approach with the customer in question. Based on the past payment behavior, the probability that the customers will pay in the future (if they have not paid until now) can be calculated. This information will be available to credit management and will improve collections management process.
6.2 Sales receivable prediction with Kaplan-Meier

Based on the individual customer models that were presented in 6.1, this general daily Kaplan-Meier model for predicting receivables was built. The prediction model is a combination of sales estimates and customer payment behavior model. As it can be seen from figure 11, predicted sales receivable line does not follow the same path as actuals line. The predicted sales receivable’s line is an expected value of sales receivable where payment probabilities are computed from Kaplan-Meier estimate.

![Sales receivables prediction based on customer behavior](image)

**Figure 11.** Sales receivable predictions based on customer behavior.
In figure 11, predicted sales receivable curve is a forecast from 01.07.2018 onwards. In the beginning of the month, the current open invoices (green line) contribute the most to the sales receivable estimate, but then gradually new sales (red line) start to dominate. Intersection point between red and green line at the end of July means that at that point in time customer model and sales estimates have equal prediction power on sales receivable. However, after two months the customer model component starts to disappear whereas sales estimates component takes over the prediction power.

This model underperforms when compared to SARIMX daily model because it only considers past behavior of customers and it does not consider history of sales receivable, trend and seasonality. Nevertheless, it was presented here to highlight the importance of sales estimates in predicting sales receivable, which will further be supported by the SARIMAX model in the latter section. Mathematical representation of the model was presented with equation 18 in 4.3.

### 6.3 Sales receivable prediction with SARIMAX

In this section another model built for predicting sales receivable is presented. The model is SARIMAX statistical regression and uses sales estimates as an external variable in addition to using historical sales receivable values to predict receivables. Initially, daily, weekly and monthly SARIMA models were built without taking sales estimates into account and their mean absolute errors were calculated. Mean absolute error measures deviation between actual and forecasted values. Therefore, the smaller MAE value the better the forecast. MAE is expressed with the following equation (20):

\[
\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|
\]

Where \( n \) is the number of observations, \( y_i \) is the actual value and \( \hat{y}_i \) is the predicted value (Cao & Tay, 2003).
As observed in section 6.2 in the Kaplan-Meier model estimation, sales estimates appeared to be a key component in predicting sales receivable level in the future.

Therefore, sales estimates as an external variable were added and three SARIMAX models were built and their mean absolute errors were calculated. In all cases predictions were more accurate (smaller MAE) when sales estimates were accounted for in the prediction. Reason is that SARIMAX models can model the periodic behavior of sales receivable whereas Kaplan-Meier model provides only an expected value that is a weighted average.

Among the three models, daily SARIMAX model was chosen to be used. The daily model because it captures weekly seasonality whereas weekly and monthly models can only capture month and quarter level seasonality. In addition, monthly model is not feasible for this purpose as the model does not have enough data points for that, it would require unrealistically long history of data.

<table>
<thead>
<tr>
<th>Prediction horizon</th>
<th>MAE without sales estimates</th>
<th>MAE with sales estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 day</td>
<td>4.37E+06</td>
<td>4.23E+06</td>
</tr>
<tr>
<td>30 days</td>
<td>2.75E+07</td>
<td>2.05E+07</td>
</tr>
<tr>
<td>60 days</td>
<td>4.19E+07</td>
<td>2.09E+07</td>
</tr>
<tr>
<td>90 days</td>
<td>9.77E+07</td>
<td>3.61E+07</td>
</tr>
</tbody>
</table>

Table 2. MAE of daily SARIMAX model with and without exogenous variable with different prediction horizons.
Figure 12 shows prediction error as a function of prediction horizon with and without sales estimates included in daily SARIMAX modeling. From the same figure it is observed that the further the prediction in the future the higher the mean absolute error, if sales estimates are not taken into consideration. However, when sales estimates are accounted for in the model, the mean absolute error remains static the further the prediction. Table 1 above provides mean absolute errors with and without sales estimates and for different time horizons. Therefore, with a static prediction error, it is important to highlight the fact that the better our sales estimates the better the receivables prediction will be.

![Figure 12. Prediction error plotted against prediction horizon.](image)
To support graph depicted in figure 12, daily SARIMAX model was used to predict sales receivable using different time horizons. Figure 14 below presents daily SARIMAX model with 30 days horizon, the model that produced the most accurate predictions amongst other models with time horizons of 1 day, 60 days and 90 days.

Figures 13, 15, 16 (see APPENDIX) represent daily SARIMAX models with 1 day, 60 days and 90 days horizon. All models use historical sales receivable levels as well as sales estimates as inputs. Estimates were then compared against actual sales receivable levels as shown in the graphs. It is clearly visible from the graphs that the shorter the time horizon the better the prediction.

![Graph of Actual and Predicted Sales Receivables](image)

**Figure 13.** Actual and predicted sales receivable daily SARIMAX model prediction horizon 30 days.
6.4 Summary

In this summary section, based on results, research questions laid out in 1.2 are answered.

1. *Can customer payment behavior modeling provide useful information for collections management?*

   Developing an understanding of past customer payment behavior can help collections management in many ways. Given customer payment terms they can see what “normal payment behavior” for a specific customer is, and when that specific customer appears to deviate from its normal behavior take timely action. Knowing in time if the customer is going to pay can help speed and tailor collection process. When customers start to show irregular payment behavior frequently, this triggers collections management to pay closer attention and detect risky open invoices before they become delinquent. Basically, modeling customer payment behavior helps collection management in cost reduction and increased efficiency in their process.

   A customer model that suits the above-mentioned needs has been built and it is expected that the model is going to support company’s collection management team. Results of this work were presented in 6.1.

2. *How accurately autoregressive model can forecast sales receivable?*

   By now it has been established that sales receivable is one of the key components of working capital and that their forecasting is important, and amongst other things helps in determining necessity for short-term external financing. In addition to studying customer payment behavior, it was perceived that forecasting sales receivable amount is as important. Autoregressive models are one of the most commonly used models when time series analysis is at stake. Furthermore, considering that only data available was history of sales receivable, ARIMA models sounded most convenient choice. The company did not have any forecasting model in place previously, therefore, no result benchmarking was possible. However, given the nature of the business the results were perceived as with reasonable accuracy.
3. **How much autoregressive model forecasting accuracy improves when sales estimations are added as an external variable?**

Autoregressive model forecasting accuracy improved significantly when sales estimates were added in the model as an external variable. Importance of sales receivable was first depicted in 6.2 in figure 11, where results of sales receivable prediction with Kaplan-Meier customer model were laid out. When it comes to prediction power, this model underperformed when compared to the SARIMAX because it did not consider items such as: history of sales receivable, trend or seasonality. However, it is a powerful graph because it clearly highlights the importance of sales estimates in prediction. In figure 12 prediction error without sales estimates and with sales estimates against prediction horizon was plotted. As observed, the difference between the two is huge and explains that the further the prediction in the future the poorer the prediction when sales estimates are not taken into consideration. On the other hand, when sales estimates are counted for, prediction error remains static.

4. **How does sales receivable forecasting accuracy based on customer behavior modeling compare to the autoregressive and autoregressive with external variables approach?**

When comparing accuracy of the model based on customer behavior to autoregressive and autoregressive with external variable models, higher sales forecasting accuracy is attached with autoregressive models. The reason as previously mentioned, is that sales receivable forecasting based on customer behavior (Kaplan-Meier) does not consider sales receivable history, trend or seasonality, it only considers past customer behavior. However, when comparing SARIMA and SARIMAX, higher forecasting accuracy resulted with SARIMAX, as supported by results presented in table 2 and figure 12.
6.5 Conclusion

Three models to forecast sales receivable were developed. Kaplan-Meier customer payment behavior analysis was the first model, which was able to detect anomalies in customer payment behavior. There was a total availability of 229 customers of which 122 had 20 or more invoices and were accepted for customer specifics modeling. Based on customer’s past payment behavior this model can calculate the probability that the customer will pay in the future the amount owed if it has not paid until today.

Using customer payment behavior models combined with sales estimates another model to predict total sales receivable was developed. The model derives rough sales receivable estimation because it does not contain autoregressive components or other variables. This model, as explained in 6.2, is not able to provide best sales receivable prediction as it does not consider sales receivable history, trend or seasonality. However, the model highlights the importance of sales estimates in predicting receivables.

An additional model to predict sales receivable using SARIMAX was developed. Daily, weekly and monthly SARIMAX models were developed and based on their mean absolute error calculations, it was concluded that daily model provided the most accurate sales receivable forecasts. This model too, used sales estimates as an exogenous variable in addition to its autoregressive component.

It was concluded that when an individual customer’s receivables prediction is at interest, Kaplan-Meier analysis is preferred. However, when total level of sales receivable prediction is at stake, daily SARIMAX model is preferred.
6.6 Limitations and suggestions for further research

Further development of the models is part of the next stage of this project. A cloud-based dashboard for the models will be developed. The dashboard intends to create value for respective stakeholders in the company. Dashboard will be compatible and integrable with the CRM platform of the company. This would enable sales managers effortless monitoring over customers and support in taking timely actions when needed.

Regarding model development, it is important to note that in this research study, SARIMAX captures only one kind of seasonality. For example, in daily model 7 days long period seasonality is used. However, in the case of daily model, it is interesting and very important to capture business month level behavior, quarter level behavior, half a year level behavior and yearly behavior. These could be added to the model by using suitable external variables.
REFERENCES


APPENDIX

Figure 14. Actual and predicted sales receivable daily SARIMAX model prediction horizon 1 day.
Figure 15. Actual and predicted sales receivable daily SARIMAX model prediction horizon 60 days.
Figure 16. Actual and predicted sales receivable daily SARIMAX model prediction horizon 90 days.