Aalto University
School of Business

Analysing and Predicting Invoice Payment in Finnish Road Freight Transportation Industry

Bachelor's Thesis
Emmi Turunen
Aalto University School of Business
Information and Service Management
Fall 2022

Author Emmi Turunen

Title of thesis Analysing and Predicting Invoice Payment in Finnish Road Freight Transportation Industry

Degree Bachelor's degree

Degree programme Information and Service Management

Thesis advisor Riitta Hekkala

Year of approval 2022 **Number of pages** 41 **Language** English

Abstract

This thesis examines the relationship between diesel prices, 12-month Euribor interest rates, and bankruptcies of road freight transportation companies and their effects on invoice payment delays in a sample of road freight transportation companies. This investigation is relevant to the road freight transportation industry, as the changes in diesel prices and interest rates can significantly impact the transportation companies' cost structures. The sample was collected from a heavy equipment spare parts retailer and contained information on over 70 000 invoices.

The work consist of a literature review and two quantitative methods, time series regression model and binary classification models. The time series regression model looks at 92 months from 2015 to 2022. The independent variables are monthly average diesel prices, 12-month Euribor interest rates and bankruptcies, and the dependent variable is the average monthly payment delay in days. Binary classification models predict whether an invoice will be paid on time or late. These models include logistic regression and k-nearest neighbors.

Overall, the time series regression model is statistically significant. From the individual coefficients, diesel prices were the most significant predictor, suggesting that they impact payment delays the most. Euribor interest rates and bankruptcies were not found significant at the 0.05 level. Furthermore, the analysis revealed an unexpected relationship where a decrease in diesel prices leads to longer delays in payments. Additionally, the binary classification results show that the k-nearest neighbors' method with a small hyperparameter k value and the logistic regression model were effective in predicting late payments.

Keywords payment delays, time series regression, binary classification

Table of Contents

List of Figures

List of Tables

1 Introduction

The logistic system in Finland is a collaborative effort between stakeholders, including the government, the business sector, and consumers. It has a central role in maintaining the security of supply through transporting goods within the country. According to data from Statistics Finland, in Q2 2022, heavy-duty vehicles transported 54 million tons of cargo on domestic highways (Statistics Finland, 2022 a). Recent crises have highlighted the importance of efficient logistics in society and its ability to deliver goods at the right place, time, and condition.

According to SKAL, the Finnish Transport and Logistics trustee, most of Finland's road freight transportation companies are family-owned and operate with just one to three vehicles (Kujala et al., 2022). These companies are facing challenges related to payments and financial management due to the changes in their operating cost structures. According to a European Payment Report by Intrum, a credit management company, more than half of all Finnish companies expect payment delays and other payment-related issues to increase in the coming year. The main reasons for these payment challenges include rising inflation, higher interest rates, and increased regulation, which pressure companies and make it harder to manage finances and maintain a healthy cash flow (Intrum, 2022).

The literature suggests that payment delays in business-to-business (B2B) transactions can significantly impact firms' survival. A study by Connell (2014) found that late payments can increase company exit rates, potentially due to the financial strain caused by such delays. This research also found that negative economic trends and business cycles can affect exit rates, with downturns often leading to higher rates of business closures (Connell, 2014). Research suggests that rising energy and raw material prices can negatively impact firms' production costs, resulting in weaker profitability (Intrum, 2022). These factors can all contribute to the failure of financially constrained firms, resulting in negative implications for the overall economy.

Late payments can be a significant obstacle in the company's accounts receivable management process, increasing the time it takes to convert invoices to cash. Companies may benefit from improving their forecasts of future economic conditions to prepare adequate working capital investments during economic downturns. These include, for example, examining the customer-payment process to determine why

customers pay when they pay, decreasing excess receivables, and offering a reference point when receivables continue to rise due to economic uncertainty (Enqvist et al., 2014). Increasing adoption of information technology in various domains and industries, particularly machine learning techniques, has not eliminated the reliance on manual processes for managing internal operations in many businesses.

1.1 Research objectives and research questions

This thesis investigates payment delays and the potential benefits of using predictive modeling on invoice payments in the case of a small Finnish company that supplies heavy-duty vehicle replacement parts for road freight transportation companies. The company's invoice data comprises over 70,000 transactions from its customers. As the company manually handles its accounts receivable process, the investigation of supervised machine learning models for predicting invoice payments is relevant. In addition, this research aims to analyze the factors contributing to late payments among the case company's customers.

Therefore, the effects of diesel prices, Euribor interest rates, and bankruptcies on payment delays and using predictive modeling techniques to improve account receivable collection and predict customers likely to pay late are topics of interest. Thus, the objectives of research that guide this study are:

Research Question 1: What are the effects of diesel prices, interest rates and bankruptcies on payment delays?

Research Question 2: How could the company detect customers who pay late and optimize debt collection?

To achieve these objectives, my perspective in this paper can be stated through the following hypotheses:

Hypothesis 1: An increase in diesel prices leads to an increase in payment delays.

Hypothesis 2: An increase in Euribor 12-month interest rates leads to an increase in payment delays.

Hypothesis 3: An increase in bankruptcies within transportation and warehousing industry leads to increased payment delays.

Hypothesis 4: The company can use binary classification techniques to predict late payments based on specific customer invoice characteristics and optimize debt collection efforts.

1.2 Scope of research

This research will focus on the effects of fuel prices, interest rates, and bankruptcies on payment delays in the Finnish road freight transportation industry. The research will use data from a specific company and its customers' invoice data to provide insights on improving accounts receivable collection and predicting customers who are likely to pay their invoices late. The analysis will be limited to three modelling techniques. Importantly, it will not consider other factors affecting payment delays, such as technological advancements in logistics, salaries, companies' financial state, size, or access to alternative fuel sources.

Furthermore, the study will only examine the performance of the chosen modelling techniques on the data and will not explore other potential modelling approaches. Moreover, the research will not address the effects of payment delays on other stakeholders in the logistics system, such as consumers or the government.

1.3 Methodology

To address the research questions outlined above, this case study proposes to conduct a mixed-methods analysis that incorporates both qualitative and quantitative approaches. This involves conducting a literature review to identify and analyse relevant studies and findings on the factors influencing the likelihood of late payments on invoices and how the prior research has predicted invoice payment. A time series regression analysis and two supervised machine learning algorithms for classification problems will be applied for the quantitative part. The analysis seeks to uncover relationships among variables and forecast the likelihood of late invoice payment. Furthermore, the machine learning approach applied can potentially automate the process of reducing the rate of late payments.

1.4 Structure of the research

The following sections outline the structure of the rest of the thesis. In Chapter 2, a review of the existing literature on late payments prediction, the impact of fuel prices and interest rates on transportation, and the effects of the crisis on small companies are provided, thereby establishing the theoretical foundation for the current investigation. Chapter 3 discusses the methodologies, data, statistical model formations, and machine learning models employed in this work. Chapter 4 highlights the results, and Chapter 5 discusses the findings and their implications for practice.

2 Theoretical background

The ultimate drivers of changes in the operational environment faced by road freight transportation companies, such as rising gas prices and higher interest rates, can be traced back to recent global crises. These crises have inevitably impacted businesses and the economy. Although past literature discussing the effects of diesel prices and interest rates, specifically on road freight transportation companies, is limited, these topics can be analyzed in a more general context, which applies to understanding the effects on transportation companies. Fuel costs, particularly diesel prices, are a significant expense for road freight transportation companies; therefore, the extent to which diesel price increases affect their payment behavior is of interest.

In addition, as inflation and rising interest rates increase companies' debt servicing expenses, the impact of higher interest rates on payment delays requires investigation. As companies struggle to cope with these challenges, they may be more vulnerable to financial difficulties and are at an increased risk of bankruptcy. Therefore, this literature review aims to provide a comprehensive overview of previous research and other sources of information, such as government announcements, related news, and barometers focusing on late payment prediction and the factors influencing road freight transportation companies' payment behavior.

2.1 Past studies on late payment prediction

There is a growing body of work on applying predictive analysis to improve accounts receivable management. Research has been conducted on the invoice-to-cash process,

part of the order-to-cash process in commercial transactions, which incorporates both insolvency modelling and late payment prediction. Examples of such studies include those by Zheng et al. (2008), Appel et al. (2019, 2020), Kim & Kang (2016), and Shao et al. (2007). In addition to the invoice-to-cash process, statistical and machine learning modelling have been utilized in the field of finance for tasks such as credit risk assessment (Galindo & Tamayo, 2000; Khandani et al., 2011), credit rating analysis (Huang et al., 2004), and bankruptcy prediction (Atiya, 2001). These techniques have proven effective in accurately assessing financial risk and predicting outcomes in these domains.

Although predicting late payments is crucial for operations, there has been relatively little focus on this aspect of insolvency modelling, despite its widespread implementation and extensive study in various domains. Zheng et al. (2008) assess the predictive analysis to enhance the invoice-to-cash collection and minimize delays in accounts receivable collections. In addition, Appel et al. (2019, 2020) created a machine-learning model that can predict the likelihood of an invoice becoming overdue regarding accounts receivable (AR) processes. Their binary classification model is helpful for companies that handle large volumes of invoices, as it allows collectors to rank customers based on the likelihood of their invoices becoming overdue. In their 2019 and 2020 studies, Appel et al. focused on detecting overdue ARs with a high likelihood of being paid. The machine-learning model they developed generated a customized, ranked list of consumers to contact based on a collector's client information and payment histories. This approach helps companies to prioritize their AR collecting efforts and focus on collecting the invoices with the highest financial return.

Kim and Kang (2016) conducted a study evaluating four classification algorithms and a hybrid approach for predicting late payments. They also created customer scoring rules to predict the likelihood and number of late payments for each customer with outstanding debt. Their research mainly aimed to identify customers likely to pay soon and provide a list of those customers for debt collection call centres to contact. In 2007, Shao et al. published research on the order-to-cash domain that used a model to estimate collection volumes on customer accounts based on connections among known variables learned through the model.

In addition to the efforts of Kim and Kang (2016), Shao et al. (2007), Zheng et al. (2008), and Appel et al. (2019, 2020) to address the problem of late payments through research, the European Union has addressed this issue through legislation. The Late

Payment Directive (LPD) was introduced in 2011 to establish a timely payment culture and protect businesses' rights concerning payment practices (European Commission, 2022). The LPD requires debtors to pay interest on overdue invoices and sets deadlines for payment, with shorter deadlines for government agencies. This legal framework has been instrumental in improving the financial stability of small and medium-sized enterprises (SMEs), which are often disproportionately affected by late payments.

While legislative measures, such as the Late Payment Directive (LPD), have been instrumental in improving the financial stability of SMEs within the EU, payment terms on business-to-business transactions vary greatly. The case company typically offers a 14-day payment term to its customers, but in some exceptional cases, it may provide a 30-day payment term. Wilson et al. (2002) found that small companies may be hesitant to apply penalty interest rates on late payments because they worry about damaging customer relationships. Besides, the potential adverse effects of using penalty interest rates may make the company hesitant even if their customers consistently pay late. This can create challenges for the company in managing its accounts receivable and maintaining its cash flow. Enqvist et al. (2014) conducted a study that found a strong negative relationship between the length of time a company defers paying its accounts payable and its EBITDA. This conclusion suggests that shorter accounts payable cycles can improve profitability and that successful Finnish companies often use cash discounts on payables as a source of financing instead of relying on accounts payable trade credit (Enqvist et al., 2014).

These findings are relevant to the current research as they provide a foundation for understanding the complex dynamics involved in payment behaviors and can inform the development of effective strategies for predicting and managing late payments.

2.2 Past studies on fuel prices effects on transportation

According to a study by Delsaut (2014), changes in fuel prices can significantly impact the demand for road and rail travel. Specifically, when fuel prices increase, there is generally expected to be a decrease in the amount of road traffic and an increase in rail traffic. These effects tend to become more noticeable over time. Another study (Gohari et al., 2018) found that fuel price increases significantly affect the cost of transporting containers by ship, with a less significant but notable impact on containers by train or truck. They also noted that the main challenge in transportation costs is often oil price, which affects the retail price of diesel fuel.

Transportation accounts for 16% of total energy consumption in Finland (Statistics Finland, 2022 b), and road freight alone consumes nearly 1.3 billion litres of diesel fuel per year, which is half of the national road transport diesel consumption (Autoalan Tiedotuskeskus, 2022 b). The cost structure of transportation companies includes several components, such as salaries, fuel, and other expenses, including repairs and insurance. Fuel costs make up a significant portion (20 - 30% of total costs) for transportation companies, being the second largest cost after salaries (Kujala et al., 2022, Lintilä, 2022).

Energy prices in the euro area have recently increased significantly, mainly due to the ongoing conflict in Ukraine (Vilmi et al., 2022). This has negatively affected businesses that are now facing higher energy costs. Intuitively, higher fuel prices will affect the costs that transportation companies will have to bear. There has been a growing concern about the effect energy price increases have on the financial performance of transportation companies. According to Statistics Finland, transportation expenses for trucks and vans have been increasing by about 15% per year, putting pressure on these companies' profitability. A case study by SKAL found that transport volume and profitability are decreasing in comparison to the preceding year, and over two-thirds of companies are considering delaying investments due to the situation. Thus, the increase in transportation costs is a significant issue for the transportation industry, Finnish trade, and the industry as a whole (SKAL, 2022).

In order to alleviate some of the financial strain caused by the increase in fuel prices, the Finnish government has implemented a temporary fuel subsidy for transportation and heavy equipment businesses. This subsidy, which covers three months, is intended to compensate for the unexpected fuel price surge and helps companies offset some of the increased costs they face. It is granted retroactively for February-April 2022, compensating 5% of the gasoline expenditures (Valtiokonttori, 2022). While the subsidy is only in place for a short period, it is expected that companies will be able to pass on the higher costs to their customers (HE 92/2022 vp, 2022). Thus, the high fuel prices end up contributing to the price increase of other goods. Transportation companies often include clauses in their contracts with customers that allow for the adjustment of transport prices concerning changes in fuel costs. These clauses are typically reviewed periodically, but if fuel price increases are sudden, the clause review period may not be able to cover the additional fuel costs fully.

2.3 Past studies on interest rates effects on small companies

Several studies have discovered an association between small company failure and interest rates (Everett and Watson, 1998; Hall and Yong, 1991; Hudson, 1986). This relationship may be attributed to higher interest rates which makes borrowing more expensive for small businesses, reducing their ability to access the capital needed to sustain operations and growth. Small businesses may also be more vulnerable to economic downturns and other external factors that impact their ability to generate revenue and maintain a favorable financial position, making them more sensitive to changes in interest rates.

A study by Petersen and Rajan (1994) on small companies found that a longer relationship with a bank can improve the availability of funding for a small business. However, smaller companies may face financial challenges due to a lack of credibility and audited financial accounts that can be shared with external finance providers. Furthermore, small businesses often have unique capital structure concerns that differ from those of large corporations. These can include the entanglement of the entrepreneur's personal finances with those of the company and a reliance on private equity and loan markets rather than public markets for external finance (Berger and Udell, 1998).

In times of economic and financial crisis, the availability of external funding can become significantly constrained. Lending practices may become more cautious, with banks and other financial institutions becoming more selective about which businesses and households they extend credit to. This can make it more difficult for small businesses to obtain the capital they need to expand and invest, which can negatively impact the economy (Connell, 2014).

2.4 Crisis effects on small companies

Crises are typically defined as extreme, unexpected events that require an urgent response from organizations. The most used definition of a crisis in business and management research is a "low-probability, high-impact situation that critical stakeholders perceive to threaten the organization's viability" (Doern & Vorley, 2018). According to Kern et al., crises are not "self-apparent occurrences" but rather events that must be understood. When these crises occur, and the existing policies are

understood to be incapable of handling the new issues, a new policy paradigm replaces the old (Kern et al., 2014).

Economic and financial crises can have a particularly severe impact on loans for small firms. These firms may not have the same financial resources or creditworthiness level as larger companies, making them more vulnerable to funding shortages. As a result, small businesses may need to be particularly proactive in seeking out alternative funding sources and building strong relationships with financial institutions to ensure they have access to the capital they need to survive and thrive (Connell, 2014). Small businesses, in general, have constrained access to public financial markets and are more sensitive to cash flow crises caused by income fluctuations (Giunipero et al., 2022). The freight transport industry in Finland experienced a turnover drop from roughly € 6.5 billion in 2020 to € 6.1 billion as a result of the Covid-19 epidemic (HE 92/2022 vp, 2022). Thus, the epidemic has significantly impacted the Finnish transportation industry, and around one-fifth of Finnish transportation entrepreneurs believed bankruptcy was unavoidable by 2020 (SKAL, 2020 a).

In response, the Finnish Parliament temporarily amended bankruptcy legislation in 2020 to prevent enterprises impacted by the crisis from declaring bankruptcy. This legislation stated that a debtor's insolvency would not be inferred solely based on their failure to pay obligations within a week of a creditor's payment request (HE 46/2020 Vp, 2020). This legislative change has been credited with helping to prevent a wave of corona bankruptcies in the country (SKAL, 2020 b). To conclude, the growth of companies can be impacted by various factors, including inflation and interest rates, the availability of raw resources, and geographical hazards. For example, Russia's intervention in Ukraine has caused economic instability, disrupting supply chains and the availability of manufacturing components (Intrum, 2022). Hence, addressing potential challenges is critical for businesses and governments.

3 Methodology

3.1 Data

A sample dataset consisting of invoices issued by a company that sells heavy-duty vehicle spare parts mainly to road freight transportation companies provides the foundation for the analysis. This data was collected directly from the company's records to answer the research question and includes information on invoice payment dates, due dates, and active bills for nearly seven years, from 2015 to 2022. In addition to the primary invoice data, secondary data on monthly diesel prices in Finland (Autoalan Tiedotuskeskus, 2022 a), official statistics of Euribor 12-month interest rates (Suomen Pankki, 2022) and statistics of the monthly number of bankruptcies within transportation and warehousing industry in Finland (Statistics Finland, 2022 c) are used in the analysis. Despite being collected for different purposes, these datasets help answer the research questions of this study. Specifically, diesel prices, interest rates, and bankruptcies are used as explanatory variables to predict their impact on payment delays. Before building the model, it is important to consider the data's potential strengths and limitations.

One of the main advantages of the primary data is its large sample size, which can help in improving the accuracy of the analysis. In this case, the data collected from the case company includes 76 882 data points, including data points from each day invoice has been registered as paid to the company. The data shows how many days the payment was late or how early it was paid. On the other hand, the secondary data contains monthly averages of diesel prices, Euribor-12-month interest rates, and bankruptcies are measured as a total per month. This can be considered a limiting aspect because the monthly variables do not capture fluctuations or changes within each given month.

Furthermore, the information in the invoice data is limited. It does not include important details, such as the invoice amounts, the size and financial stability of the company respective to the invoice, the type of services it provides, the availability of alternative fuel sources or cost-saving measures, or the bankruptcies specifically from the transportation industry. As the models used in this work require different data types, the data has been modified accordingly. In the following sections, these modifications are discussed in further detail.

3.1.1 Data in time series model

To perform time series regression on the combined datasets, the daily invoice data is aggregated to the monthly level. This process may obscure significant fluctuations within each month. The analysis is based on a sample of 92 data points representing the monthly average payment delays in days (the difference between invoice payment date and due date), monthly average diesel prices, monthly average Euribor 12-month interest rates, and monthly bankruptcies in the transportation and warehousing industry. The natural logarithm (log) function is applied to the variables to normalize diesel prices, delays, and bankruptcies. The Euribor interest rates, expressed in percentages, are left in their standard form.

While analysing time series data, it is important to acknowledge that past events can influence future outcomes, but the future does not affect the past. Thus, the actual realization of the stochastic process is determined by the circumstances when the process begins, and subsequent events cannot alter these circumstances (Wooldridge, 2015, p. 311). Therefore, time series data represents a sample of random variables that meets the criteria for random sampling. Depending on the case at hand, the set of independent variables $(x_{t2}, \ldots, x_{tk})$ may contain contemporaneous explanatory variables z_t(static model), lagged explanatory variables $z_{t-1}, \ldots, z_{t-k}$(distributed lag model), or lagged dependent variables $y_{t-1}, \ldots, y_{t-k}$ (autoregressive model).

Hence, before model definition, the data is inspected for trends and seasonality, as these characteristics can weaken the reliability of the model's predictions if an unsuitable model is chosen. The accompanying visualizations in Figures 1 and 2 show the results of this inspection, revealing no apparent trends. Furthermore, the Dickey-Fuller test will be applied to confirm the stationary of residuals in Chapter 4.1.

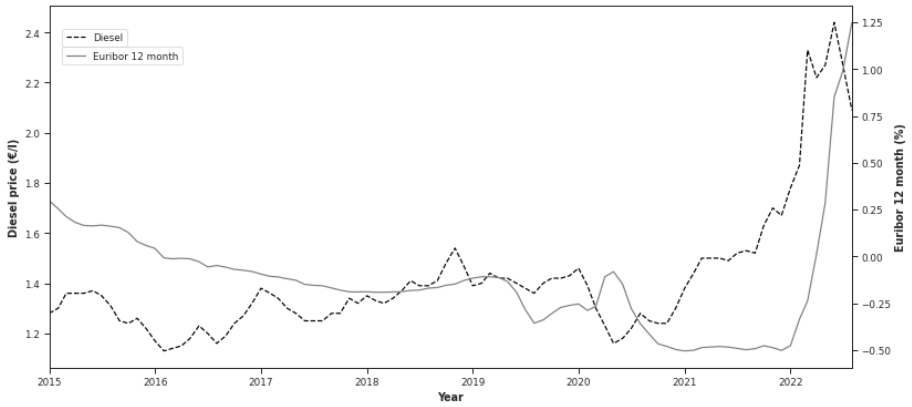

Figure 1. Diesel prices and Euribor-12-month interest rates 2015-2022

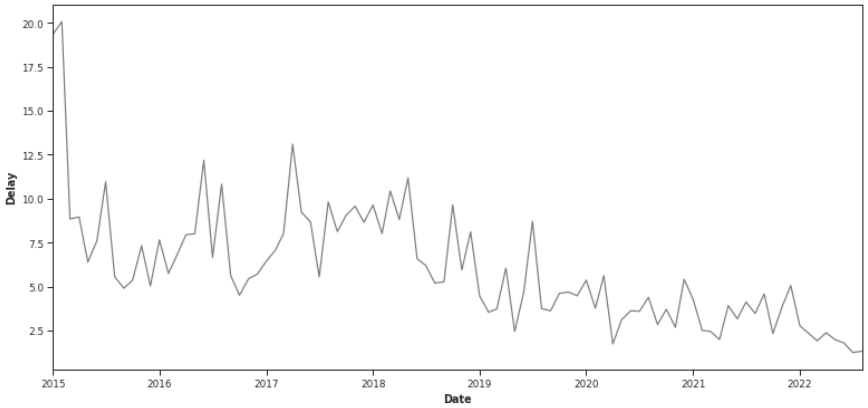

Figure 2. Payment delays in days 2015–2022

3.1.2 Data in binary classification models

All 76 882 data points in the primary data are utilized for binary classification purposes. The monthly average values are duplicated for each corresponding data point in the invoice dataset. This allows the integration of the secondary data into the analysis using the primary dataset as a foundation. Additionally, data augmentation techniques can create new, artificial data for the model instead of collecting more raw data. This synthetic data can help improve the model's performance by providing additional information and reducing overfitting (Jung, 2022, p. 162). Thus, three new features are created from the original set: the weekday of the invoice due date, days until the end of the month from the invoice due date, and the weekday of the invoice payment date. The term "label" is used in machine learning to refer to the output variable (the dependent or target variable). In this context, the label is a high-level property of the data points representing the quantity of interest and whether the

payment is made on time. Table 1 shows a comprehensive overview of the features employed in the model.

Table 1. Description of features

No.	Feature	Description
1	Late	Target variable that indicates whether the invoice is paid late or not late. Value 1 stands for late, while value 0 is not late.
2	DueDate_Weekday	Categorical variable that specifies the day of week on which the bill is due.
3	DaysToEndOfMonth	Numerical variable, that specifies the days remaining until the end of each month in which the bill is due.
4	Delay	Numerical variable, that describes the length invoice is past due in days, or a negative sign if the invoice was paid before the due date.
5	PaymentDate_Weekday	Categorical variable, that specifies the day of the week on which the bill was paid.
6	OpenBill	Numerical variable, that specifies the amount of the invoice that remains unpaid.
7	Diesel	Numerical variable, that represents the monthly diesel price.
8	12month	Numerical variable, that represents the monthly Euribor-12-month interest rates.

Correlation analysis is a statistical technique used to examine the relationship between variables. When applied to a dataset, it can help confirm if features strongly correlate with the target feature. This is helpful in the feature selection process. In this case, the correlation matrix shows that none of the intended features are strongly correlated, which suggests that these features are suitable for the model. The correlation matrix in Table 2 shows the relationships between features.

Table 2. Correlation matrix of features

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8
Feature 1	1							
Feature 2	0.045	1						
Feature 3	-0.046	-0.1	1					
Feature 4	0.19	-0.0029	0.00082	1				
Feature 5	-0.28	0.21	0.006	-0.033	1			
Feature 6	0.02	0.031	-0.011	0.021	0.0069	1		
Feature 7	0.19	-0.002	0.011	0.015	0.076	-0.028	1	
Feature 8	-0.19	0.0068	0.0098	-0.038	0.083	0.057	0.26	1

Data is then split into training and testing sets, where the training set has 70% of the data points, and the testing and validation set has 30% of the data. Training and testing set data split details are in detail in Table 3. After the split, the data is scaled with the sklearn standard scaler (Scikit-learn, n.d.). Balanced data has the proportions of different classes roughly equal, and this balance helps with improving the accuracy and generalizability of the model. In this model, the data appears to be balanced. The "on time = 0" percentage is near 51 % in both training and testing sets. Furthermore, the "late = 1" portion is around 49 in both sets. This means that there is a roughly equal amount of data in each of the two categories, with slightly more observations falling into the "not late" category.

Table 3. The data is split with 70:30 ratio into training and testing sets

Dataset	Invoices	Late	On time	Scaled Invoices	Scaled Late	Scaled On time
Train	53 817	48.76%	51.24%	53 795	48.88%	51.12%
Test	23 065	49.07%	50.92%	23 055	48.85%	51.15%

3.2 Time series regression model

Linear regression is a statistical approach that fits a linear equation into observed data to model a relationship between a dependent variable and independent variables. A commonly used method for fitting the regression line is called Ordinary Least Squares (OLS). OLS method aims to determine the best line of fit that minimizes the sum of squared discrepancies between the predicted and observed variables. OLS method

estimates linear regression model parameters regardless of whether the data is cross-sectional or time series. Time series data differs from cross-sectional data as it is ordered chronologically and is a collection of observations on one or more variables over time (Wooldridge, 2015, p. 311). In this thesis, a linear regression model is fitted to a time series dataset consisting of monthly averages as observations of payment delays, diesel prices, Euribor-12-month interest rates, and monthly bankruptcies using the OLS method. The sample size is 92 months within the period from 1/2015 to 8/2022.

Generally, a time series model is used when changes in one variable at a given time are thought to affect the other variables immediately. Thus, the effect on the dependent is modelled as a function of the dependent variable *x*, assuming that there are no other influencing factors (Wooldridge, 2015, p. 313). The equation below represents a time series model with *n* independent variables where the dependent variable is a function of the independent variables:

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t, t = 1,2,\ldots,T$$

Before applying the OLS method to fit a time series regression model, it is necessary to evaluate its applicability. This includes determining if the Gauss-Markov assumptions for the model are valid. The Gauss-Markov assumptions are a set of conditions, required for the OLS estimator to be the best linear unbiased estimator. Failure to meet the assumptions may result in biased or unreliable estimates of the model parameters (Wooldridge, 2015, p. 316 - 322). The following discussion will describe the six finite sample properties of the time series assumptions for OLS estimator.

TS.1 Linearity in parameters. It is assumed that the true relationship between the dependent variable (y) and the independent variables (x) is linear. In other words, the expected value of the dependent variable is a linear function of the independent variables:

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$$

Where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, $x_{t1}, \ldots, x_{tk}$, y_t are the data realizations of variables and $u_t : t = 1, \ldots, T$ is the sequence of error. If the true relationship between y and x is not linear, the estimates of the model parameters may be biased or inefficient. In such cases, more complex models or estimation methods that can accommodate nonlinear relationships are applied.

TS.2 No perfect collinearity. Collinearity is present when two or more of the independent variables are correlated with one another. If the relationship between the independent variables is too strong, one can be perfectly predicted from the others. Therefore, collinearity in the model will make the regression coefficients' estimates unstable and inconsistent, and the conclusions about the relationships between the variables will be incorrect.

TS.3 Zero conditional mean. The expected value of the current time period's error term is zero, given the information available at the previous period. This means that the error term is not systematically related to the previous time period's error term or the independent variables. Mathematically this assumption is:

$$E\left(u_t|u_{t-1}\right), x_{t-1}=0$$

Where u_t is the error term at period t, and x_{t-1} is vector of independent variables at time t-1. This assumption implies that the error term represents random variation that is not explained by the model.

TS.4 Homoskedasticity. In a time series context, homoskedasticity assumes that the variance of error term is constant over time. Mathematically, the assumption can be represented as follows:

$$Var\left(u_t|X\right)=\sigma^2$$

where u_t is the error term at time *t*, and $\sigma^2$is the constant variance of the error term. In any of the periods, the volatility of the errors must not be related to the independent variables. When this assumption is violated, the errors become heteroscedastic.

TS.5 No serial correlation. Error terms in the model are not correlated with each other over time. If the error terms are correlated, it suggests that the model is omitting information about the error, which can lead to inaccurate predictions.

TS.6 Normality. The errors $u_t$are independent of X and are distributed independently and identically as normal with zero mean and variance equal to σ^2: $\left(0,\sigma^2\right)$.

Thus, time series multiple regression model that explains payment delays can be stated as follows:

$$\log\left(delays_t\right)=\beta_0+\beta_1\log\left(diesel_t\right)+\beta_2\log\left(euribor\,12_t\right)+\beta_3\log\left(bankruptcies_t\right)+u_t$$

Where $\log(delays_t)$ is the log transformed monthly average of payment delay in days at time t, $\log(diesel_t)$is the log transformed monthly average diesel price at time t, $euribor\,12_t$ is the 12-month Euribor rate at time t and $(bankruptcies_t)$ is the log transformed number of monthly bankruptcies in logistics and warehousing industry at time t. The intercept is measured by β_0, the slope of the monthly average of log diesel price in euros is measured by β_1, the slope of the monthly average of euribor-12-month interest rates is measured by β_2, the slope of monthly bankruptcies is measured by β_3, and $u_t$is the error term that represents the difference between observed value y_t and the expected value y_t. Therefore, this model aims to estimate the ceteris paribus effect of changes in payment delays (Paolella, 2019).

When time series assumptions 1 – 3 holds, the OLS is unbiased. Furthermore, when time series assumptions 1-5 holds, the OLS estimator is the Best Linear Unbiased estimator.

3.3 Binary classification models

There are two main approaches to data classification: assigning class labels 0 or 1 to an unknown data item or assigning class labels and providing a probability of class membership. Support vector machines (SVMs) are an example of the first approach. Moreover, logistic regression (LR), decision trees (DTs), k-nearest neighbors (kNN), and artificial neural networks (ANN) belong to the second approach. These methods can be effective in different situations, and the best approach depends on specific data and the classification goals (Dreiseitl et al., 2002). Predicting invoice payment is a common supervised learning binary classification problem. Thus, binary classification models can be applied to support the company in predicting their client's payment behavior.

Furthermore, predicting invoice payment provides a solution for better cash flow estimates, which is crucial in ensuring financial stability (Appel et al., 2019). While economic applications, such as time series regression, often involve estimating parameters that describe the relationship between y and x, it is worth noting that machine learning algorithms are not specifically designed for this purpose. Instead, supervised machine learning is generally used to predict one variable, y, based on x, and therefore, they can help find patterns in complex task structures (Mullainathan & Spiess, 2017).

According to Jung (2022, p. 70), machine learning generally involves three main components: data, model, and loss function. Data is used to train the model to make predictions, the model is a set of algorithms that can process the data, and the loss function measures how well the model can make predictions. As data has been discussed in Chapter 3.1.2, models and their loss functions are discussed in the following chapters.

3.3.1 Logistic regression for binary classification

Linear regression and logistic regression are both supervised machine learning methods. While linear regression is used to predict a continuous value, logistic regression is used for classification tasks to predict binary outcomes (such as whether an invoice is paid late or not). The aim is to construct a model to correctly predict the class label (i.e., 0 = Not Late, or 1 = Late) for a given input datapoint. Then, the model is supplied those datapoints (x) and trained to generate a prediction of the corresponding label ($\hat{y}$). A loss function is needed to minimize loss between estimate $\hat{y}$ and true y. While linear and logistic regression uses the same model, they are distinct in their loss functions. In linear regression, the loss function is least squared error, while logistic regression has a loss function called logistic loss, or binary cross-entropy, which is defined as follows:

$$\log loss = \sum_{i=0}^{n} -\left(y_i * \log\left(p_i\right) + \left(1 - y_i\right) * \log\left(1 - p_i\right)\right)$$

Where n is the sample size indexed by i, y_i is the true class for index i, and p_i is the model's prediction for the index i. Minimizing log-loss is equivalent to maximizing the log-likelihood. The weighted sum of inputs is passed through an activation function, that maps values between 0 and 1 and minimizes the loss function over all samples. This activation function is called sigmoid function:

$$s\left(z\right) = \frac{1}{1 + e^{-z}}$$

Where $s\left(z\right)$ is the output between 0 and 1, z is the function's input and e is the base of natural log. After training, the model's performance is evaluated on a separate testing dataset, in which true labels are known. Logistic regression, like linear regression, assumes that the data is normally distributed. This assumption is not always true in practice, so comparing the model's performance to another model on the same data can help determine the most effective.

3.3.2 K-Nearest Neighbors for binary classification

K-Nearest Neighbors is a non-parametric method for classification and regression tasks involving linear or non-linear data. It works by examining the data points closest to a given sample and classifying the sample based on the majority class of those nearest neighbors of each data point. The value of hyper parameter *k*, the distance metric used to calculate the similarity between data points, and the choice of predictor variables can all significantly impact the model's performance. By default, the *k*NN algorithm uses the Euclidean distance measure to determine the distance between data points (Scikit-learn. 2022). The Euclidean distance is the square root of the sum of the squares of the differences between the coordinates of the two points in n-dimensional space (Zhang, 2016):

$$D(p,q)=\sqrt{((p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2)}$$

Where p and q are subjects to be compared with n characteristics.

The hyper parameter *k* determines the number of neighboring points used to make predictions, and statistical assumptions or parameters do not determine its size. It is up to the user to define this parameter. A larger value of *k* can reduce random error in the model but may also cause the model to overlook essential patterns in the data. On the other hand, a smaller value of *k* may be more sensitive to these patterns but can also result in overfitting, and the model becomes too complex and performs poorly on new unseen data (Zhang, 2016).

Furthermore, *k*NN – method does not have a single loss function, and either it employs a loss function in the traditional sense. Instead, the loss in *k*NN is the error between the predicted and true label of the input data point. In general, the loss can be defined as the sum of the distances between the input data point and its nearest neighbors. With this loss function, the *k*NN model is optimized by choosing the value of *k* that minimizes the loss. Appendix 4 explores different values for the hyper parameter *k* and their impact on model performance.

4 Results

4.1 Time series regression results

The aim is to understand the factors influencing payment delays in road freight transportation companies using time series regression. This method is well-suited for analysing data collected over time, making it an appropriate choice for the analysis. The Gauss-Markov assumptions were tested and found to be satisfactory to ensure the model's validity.

The following hypotheses were formulated to guide the analysis:

Hypothesis 1: An increase in diesel prices leads to an increase in payment delays.

Hypothesis 2: An increase in Euribor 12-month interest rates leads to an increase in payment delays.

Hypothesis 3: An increase in bankruptcies within transportation and warehousing industry leads to increased payment delays.

Parameters follow a linear pattern. The linearity of residuals is tested with the graphical and analytical methods. Histograms and Quantile-Quantile plots are graphical methods and calculating the Shapiro-Wilk-test score is analytical. Upon the first two exploratory analyses of the data, the assumption of linearity is not satisfied (Appendix 2). A log transformation is applied in variable delays, diesel prices, and bankruptcies to normalize the distribution, and the assumption of linearity is satisfied.

In Figure 3, the histogram of residuals shows that residuals variance is approximately normally distributed with a bell-shaped curve - a common characteristic of normal distribution. Additionally, the quantile-quantile plot of residuals is linear, and the error terms are assumed as normally distributed. Shapiro-Wilk test confirms the normality of the sample. The test statistic is 0.993 and has the corresponding p-value of 0.891. The p-value fails to reach the commonly used threshold of 0.05 for statistical significance, indicating that the data is consistent with a normal distribution.

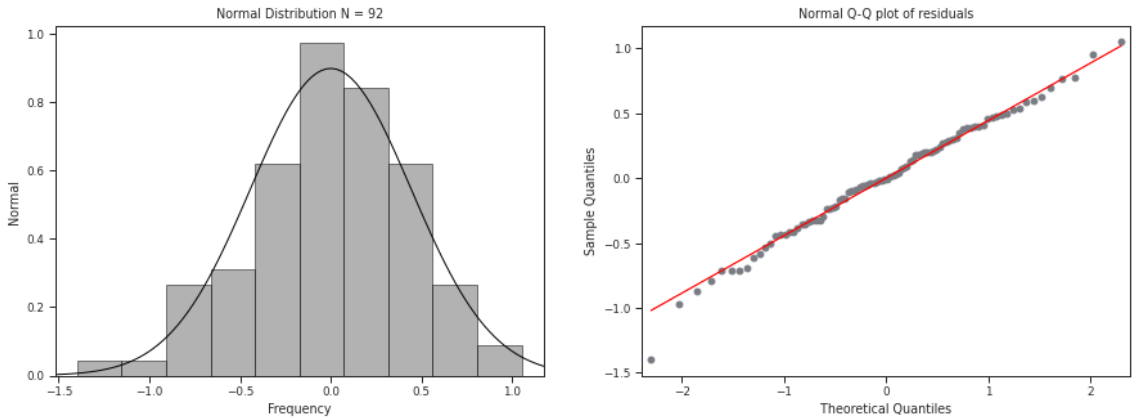

Figure 3. Histogram of residuals and Quantile-Quantile plot of residuals

Assumption of multicollinearity is not violated. Variance inflator factor (VIF) is a multicollinearity measure with severity rising as the value increases. A VIF score of 1 suggests that there is no multicollinearity. Values ranging from 1 to 5 indicate moderate multicollinearity, while values above 5 indicate strong multicollinearity. Based on the values in Table 4, all variables have an acceptable amount of multicollinearity, with "log_diesel" having the highest score (1.0859), suggesting that it has the strongest relationship with the other variables in the model.

Table 4. Variance Inflator factors of features

Feature	VIF
Log_diesel	1.085303
Euribor 12-month	1.069302
Log_bankruptcy	1.018906

Data satisfies the zero conditional mean assumption. The plots in Figure 4 indicate that residuals of fitted values scatter around the zero lines, and no clear pattern is detected. This is a good sign and indicates that the errors in the predictions are random.

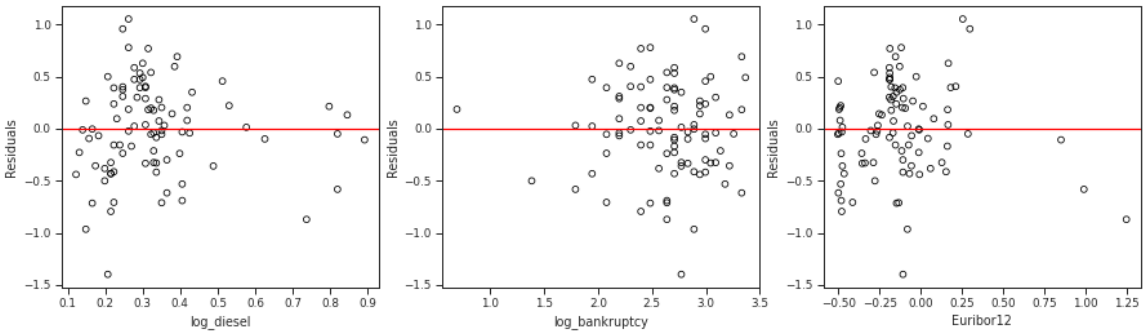

Figure 4. Zero conditional mean visualization: Residuals vs. fitted plots

The homoskedasticity assumption is satisfied. The Breusch-Pagan test assesses a regression model's homoscedasticity (constant variance). This test compares the model with constant variance (the null hypothesis) to a model with heteroscedastic variance (the alternative hypothesis). This test results with a p-value of 0.0932, which is compared to the 0.05 threshold. The conclusion is that the model is not heteroscedastic, and the variances are assumed as constant.

No serial correlation is detected. In Figure 5, the confidence interval is the grey-shaded region with a default significance level of 0.05. The vertical lines represent lag values, which are the numbers of previous values used to compute the correlation coefficient. Any vertical value inside the confidence interval represents the nonsignificant relationship between the current and most recent observed values. Based on the confidence intervals plotted in the graph, serial correlation is not a serious issue for this model.

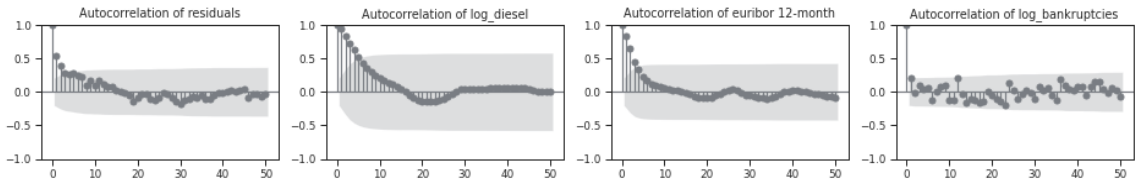

Figure 5. Visualisation for serial correlation

Lastly, the stationarity of the residuals is checked using the Augmented Dickey-Fuller (ADF) test, which has a test statistic of -3.94 and a p-value of 0.00175. These results show that the residuals are stationary. The estimates of the parameters in the multiple

linear regression model are unbiased, as the model satisfies the Gauss-Markov assumptions. The empirical results of the model are provided in Table 5.

Table 5. Empirical regression results in log-transformed delays, diesel price, and bankruptcies with Euribor 12-month interest rates as a non-transformed variable to identify factors influencing payment delays

Variable	Coefficient	Std.Error	T-Statistic	P-value
Constant	2.1749	0.298	7.306	0.000
Log_diesel	-2.2498	0.308	-7.293	0.000
Euribor-12-month	0.3014	0.165	1.830	0.071
Log_bankruptcy	0.0987	0.108	0.911	0.365
R-squared	0.377	No. Observations		92
Adj. R-squared	0.355	Omnibus		1.866
F-statistic	17.73	Prob (Omnibus)		0.393
Prob (F-statistic)	4.36e-09	Skew		-0.281
Durbin-Watson	0.837	Kurtosis		3.198
Jarque-Bera (JB)	1.362	Prob (JB)		0.506
Log-Likelihood	-55.782	Cond. No		21.0

F-statistic probability is an indicator of the model's overall significance level. It determines whether the model's independent variables (diesel prices, 12-month Euribor rates, and bankruptcies) are significantly related to the dependent variable (payment delays). The lower the p-value, the more significant the model. The model has a p-value of 4.36e-09, below the 0.05 threshold, indicating that the model is significant. The model's p-value being low indicates that it is unlikely that the relationship between the variables is a random occurrence.

Additionally, Adjusted R-squared measures how well the statistical model fits the data. Its value ranges from 0 to 1, where a higher value indicates a better fit. This model explains 35.5% of the variance in payment delays, but most of the variance remains unexplained. This indicates that the model is not a perfect fit for the data.

The estimated relationships between diesel prices, Euribor-12-month rates, bankruptcies, and payment delays are quantified by the coefficients in the model.

Because payment delays, diesel prices, and bankruptcies are log-transformed, the meaning of their coefficients differs from the usual interpretation. Thus, when diesel prices fall by 2.25%, payment delays are expected to increase by 1%, assuming all other variables remain constant. Similarly, when bankruptcies increase by 0.1, the payment delays are expected to increase by 1%. Furthermore, when Euribor -12-month rates rise by 0.3014, the payment delays increase by 1% while all other variables remain constant. Despite having a p-value of 0.071, which is not statistically significant at the 0.05 level, Euribor-12-month is still statistically significant at the 0.1 level. While diesel prices have a statistically significant p-value of 0.000, bankruptcies have a p-value of 0.365, which is not statistically significant.

The following conclusions about the hypotheses are based on the results of the time series regression analysis:

Hypothesis 1: The results show strong evidence that a rise in diesel prices is linked to a decline in payment delays, with all other variables held constant. Therefore, the null hypothesis (H0: An increase in diesel prices may lead to an increase in payment delays) is rejected in favor of the alternative hypothesis (H1: An increase in diesel prices may lead to a decrease in payment delays).

Hypothesis 2: Based on the p-value of 0.071 for the Euribor variable, the null hypothesis is not rejected at the 0.05 level. Therefore, there is no strong evidence to support the alternative hypothesis (H1: An increase in Euribor's 12-month interest rates may lead to a decrease in payment delays) and the null hypothesis (H0: An increase in Euribor's 12-month interest rates may lead to an increase in payment delays) holds.

Hypothesis 3: The p-value for bankruptcies is 0.365, which is not considered statistically significant. This means that the null hypothesis (H0: An increase in bankruptcies within the transportation and warehousing industry may lead to an increase in payment delays) is not rejected.

4.2 Binary classification results

Hypothesis 4: The company can use binary classification techniques to predict late payments based on specific customer invoice characteristics and optimize debt collection efforts.

Several metrics are used to evaluate the performance of a machine learning model, including accuracy, precision, F1 score, R2 score, and loss. A model with high accuracy can make correct predictions more often, while a model with high precision can correctly identify true positives out of all positive predictions. The F1 score, a combination of precision and accuracy, indicates the model's ability to classify data correctly into positive and negative classes. A high F1 score (close to 1) indicates that the model predicts a low rate of false positives and negatives. Accuracy, precision, recall, and F1 score are evaluation metrics that can be calculated from the confusion matrix (Jung, 2022, p. 68). Hence, a confusion matrix, also known as an error matrix, is used to evaluate a classification mode's performance. The matrix shown in Figure 6 compares the true label values to those predicted by the models.

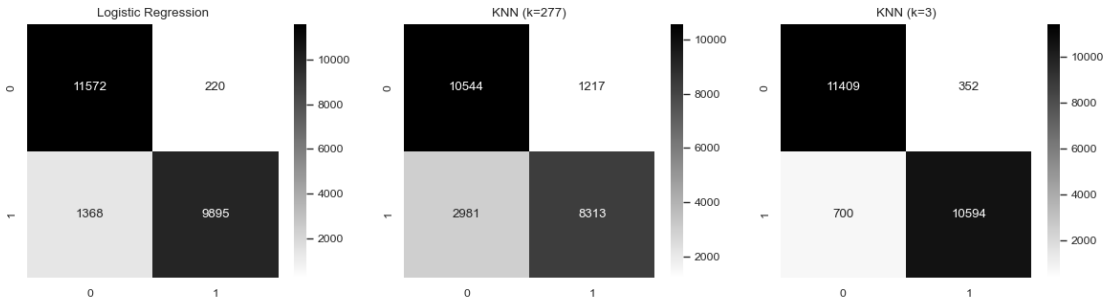

Figure 6. Comparison of Confusion Matrices for Logistic Regression and kNN with large and small hyperparameter k.

Receiver Operating Characteristic Area Under the Curve (ROC AUC) measures the area under the receiver operating characteristic (ROC) curve. It visually represents the model's capacity to differentiate between positive and negative classes. The area under the curve (AUC) varies between 0 and 1. AUC near 1 indicates that the model distinguishes well between the true positive and true negative classes, resulting in fewer false positive and false negative predictions. Figure 7 illustrates a visual comparison of the models' ROC curves.

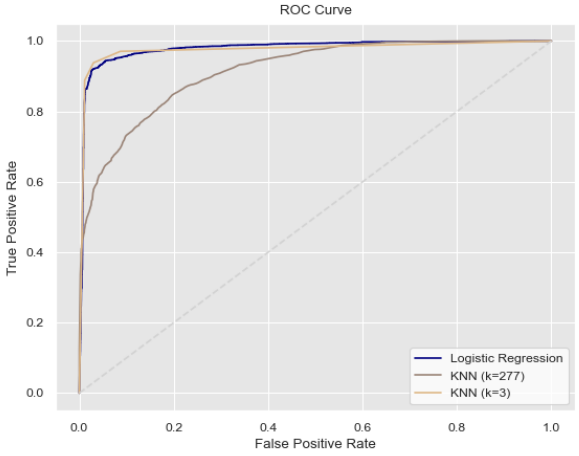

Figure 7. Comparison of ROC curves for Logistic Regression and k-nearest neighbor models

When a logistic regression model has a high R2 score, it indicates that the model's predictions closely agree with the actual data. Furthermore, log loss measures how well the model can predict the probability of an event occurring, and a low log loss indicates that the model can predict these probabilities accurately. As discussed earlier, the *k*NN model does not have a generally agreed loss function. The results of the performance comparisons among the different models are in Table 6.

Table 6. Results from classification models

	Logistic Regression	*k*NN (k=3)	*k*NN (k=277)
Accuracy	0.931	0.948	0.817
Precision	0.978	0.966	0.872
F1 score	0.926	0.946	0.798
Recall	0.747	0.927	0.736
ROC AUC	0.979	0.979	0.915
R2 score	0.747	-	-
Loss	0.257	-	-

Comparing the models, it appears that the *k*NN model with k=3 has the best overall performance, followed by the logistic regression model. The *k*NN model with k=277 has

the lowest performance among the three models. While the *k*NN model with k=3 may be more sensitive to patterns in the data and at risk of overfitting, it appears to be a trade-off worth making in this case, considering the model's overall performance. Therefore, these results suggest that carefully tuning the value of hyper parameter k in the *k*NN model can improve the model's performance.

5 Discussion and conclusions

Time series regression results provide mixed support for the initial hypothesis, where increased diesel prices, interest rates, and bankruptcies are assumed to lead to longer payment delays. While it is possible that these factors could influence payment behavior in some contexts, the analysis revealed an unexpected relationship between diesel prices and payment delays, with higher diesel prices associated with shorter payment delays.

Another key finding from the time series regression analysis is that higher interest rates are associated with longer payment delays, supporting the initial hypothesis. This suggests that changes in interest rates impact businesses' ability to meet their financial obligations on time. Furthermore, the results offer information to conclude that increased bankruptcies are associated with increased payment delays. Lastly, a significant finding from the binary classification application is that the late payments can be classified with k-nearest neighbors having a small hyper parameter k and a logistic regression model with more than 0.93 accuracies. This suggests that certain customer characteristics may be predictive of payment behavior and could be used to inform debt collection efforts.

5.1 Implications to research

Previous research has shown that changes in fuel prices can affect the demand for different modes of transportation (Delsaut, 2014) and the overall cost structure of transportation companies (Gohari et al., 2018, Kujala et al., 2022, Lintilä, 2022). Additionally, SKAL barometers discussing the fuel prices effects on transportation companies suggest that transportation companies' financial performance weakens as a result of higher fuel prices (SKAL, 2022). Therefore, the time series regression finding that a decrease in diesel prices leads to an increase in payment delays is surprising. This finding challenges the previously established knowledge that higher fuel

prices negatively impact the financial performance of freight transportation companies. It should be emphasized that this finding is conditional on all other variables remaining constant.

Findings from past literature support the finding that an increase in Euribor-12-month rates and an increase in bankruptcies are associated with increased payment delays. Higher interest rates can make borrowing more expensive for small companies and potentially reduce access to capital. This is consistent with the findings of Everett and Watson (1998), Hall and Yong (1991), and Hudson (1986), who all identified a relationship between small company failure and interest rates. However, the finding adds to the existing literature by demonstrating that the relationship between interest rates and payment delays is not necessarily linear. This suggests that small companies may be particularly sensitive to even small changes in interest rates.

The results of the binary classification model suggest that logistic regression and k-nearest neighbors effectively predict late invoices, as they achieved high accuracy rates. These findings align with past research, such as the study by Zheng et al. (2008), which found that logistic regression classification performed well among the other five models, with cost-sensitive learning having the highest accuracy. Appel et al. (2019, 2022) also found that logistic regression and kNN performed well in their classification problem, although not as well as XGBoost and Random Forest methods. These findings add to the existing literature on predicting payment delays and may be helpful for companies looking to improve their accounts receivable collection.

5.2 Implications to practice

There are various potential explanations for the disparity between the predicted and observed relationship between diesel prices and payment delays. One possibility is that the model does not account for daily data, and a more detailed examination may identify relationships that were not observable in the monthly data. Additionally, many transportation companies have clauses with their customers that allow for adjusting transport prices in response to changes in fuel costs. These clauses may mitigate the impact of higher fuel prices on the transportation company's financial performance by allowing them to pass on the increased costs to their customers. Therefore, the effect of higher fuel prices on payment delays may be less significant for transportation companies with such clauses. It's also highly likely that increased diesel prices may not

have been fully incorporated in the data, as they might have a delayed effect on payment behavior.

Further explanation could be that changes in account receivable collection regulations or improvements in payment processes within the company over the period 2015-2022 have contributed to a decrease in payment delays. In 2019, the company transitioned from traditional mailed invoices to electronic invoicing. Based on the findings in this thesis, there is evidence that the company's transition to electronic billing has successfully decreased payment delays. These internal factors may impact payment delays more than the external factors of diesel prices and interest rates. Therefore, it is important to underline that the analysis results are based on a specific dataset and may not be applicable in other contexts. Moreover, other factors not included in the model may also contribute to the relationships between the variables.

5.3 Limitations and future research

The analysis in this thesis was an exploratory investigation that analysed the relationship between fuel prices, interest rates, bankruptcies, and payment delays within the heavy-duty vehicle spare parts retailers' customer base over 1/2015-8/2022. The findings are limited to the specific company's customer base and period studied; therefore, further research is needed to confirm the generalizability of these results to other contexts. Future research could focus on expanding the sample of freight transportation companies and include a longer time in the analysis to understand further the relationship between diesel prices, interest rates, bankruptcies, and payment delays. This could help to identify any potential phenomena that may not have been apparent in this work.

Another potential aspect for future research are the internal factors influencing payment delays. This includes analysing the impact of changes in account receivable collection policies and companies payment processes. Additionally, future research should evaluate other classification methods to improve the model's accuracy for predicting payment delays and add different features to the model.

References

Appel, A. P., Malfatti, G. L., Cunha, R. L. de F., Lima, B., & de Paula, R. 2020, August 11. Predicting account receivables with machine learning. arXiv.org. [Last accessed: November 24, 2022]. Available at: https://arxiv.org/abs/2008.07363v1

Appel, A. P., Oliveira, V., Lima, B., Malfatti, G. L., de Santana, V. F., & de Paula, R. 2019, December 20. Optimize cash collection: Use machine learning to predicting invoice payment. arXiv.org. [Last accessed: November 24, 2022]. Available at: https://arxiv.org/abs/1912.10828

Atiya, A. F. 2001. Bankruptcy prediction for credit risk using Neural Networks: A survey and new results. IEEE Transactions on Neural Networks, 12(4), 929–935. https://doi.org/10.1109/72.935101

Autoalan Tiedotuskeskus. 2022 a. Bensiinin ja dieselin hintakehitys. [Last accessed: November 28, 2022]. Available at: https://www.aut.fi/tilastot/verotus_hintakehitys_ja_liikennemenot/bensiinin_ja_dieselin_hintakehitys

Autoalan Tiedotuskeskus. 2022 b. Energy use of road transport. [Last accessed November 23, 2022]. Available at: https://www.aut.fi/en/statistics/energy_use_in_transport_sector/energy_use_of_road_transport

Berger, A. N., & Udell, G. F. 1998. The Economics of Small Business Finance: The Roles of Private Equity and Debt Markets in the Financial Growth Cycle. SSRN Electronic Journal. [Last accessed November 23, 2022]. Available at: https://doi.org/10.2139/ssrn.137991

Connell, W. 2014. European Commission, Directorate-General for Economic and Financial Affairs, The economic impact of late payments. Publications Office. [Last accessed November 23, 2022]. Available at: https://data.europa.eu/doi/10.2765/71358

Delsaut, M. 2014. The Effect of Fuel Price on Demands for Road and Rail Travel: An Application to the French Case. Transportation Research Procedia, 1(1), 177–187. [Last accessed November 23, 2022]. Available at: https://doi.org/10.1016/j.trpro.2014.07.018

Doern, R., Williams, N., & Vorley, T. 2018. Special Issue on Entrepreneurship and Crises: Business as Usual? An Introduction and Review of the Literature, 31(5-6), 400-412. doi:10.1080/08985626.2018.1541590

Enqvist, J., Graham, M., & Nikkinen, J. 2014. The impact of working capital management on firm profitability in different business cycles: Evidence from Finland. Research in International Business and Finance, 32, 36–49. [Last accessed November 23, 2022]. Available at: https://doi.org/10.1016/j.ribaf.2014.03.005

European Commission. 2022. Study on building a responsible payment culture in the EU: improving the effectiveness of the Late Payment Directive (2011/7/EU). Publications Office of the European Union. [Last accessed November 23, 2022]. Available at: https://data.europa.eu/doi/10.2873/34185

Everett, Jim & Watson, John. 1998. Small Business Failure and External Risk Factors. Small Business Economics. 11. 371-90. 10.1023/A:1008065527282.

Galindo, J., & Tamayo, P. 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics, 15(1/2), 107-143. doi:10.1023/a:1008699112516

Giunipero, L. C., Denslow, D., & Rynarzewska, A. I. 2022. Small business survival and COVID-19 - An exploratory analysis of carriers. Research in Transportation Economics, 93, 101087. [Last accessed November 23, 2022]. Available at: https://doi.org/10.1016/j.retrec.2021.101087

Gohari, A. 2018. The Effect Of Fuel Price Increase On Transport Cost Of Container Transport Vehicles. International Journal Of Geomate, 15 (50). [Last accessed November 23, 2022]. Available at: https://doi.org/10.21660/2018.50.30814

Hall, Graham & Young, Barbara. 1991. Factors Associated with Insolvency Amongst Small Firms. International Small Business Journal - INT SMALL BUS J. 9. 54-63. 10.1177/026624269100900204.

HE 46/2020 vp. Hallituksen esitys. 2020. [Last accessed: November 23, 2022]. Available at: https://www.eduskunta.fi/FI/vaski/HallituksenEsitys/Sivut/HE_46+2020.aspx

Huang, Z., Chen, H., Hsu, C., Chen, W. & Wu, S. 2004. Credit rating analysis with support Vector Machines and neural networks: A market comparative study. Decision

Support Systems, 37(4), 543–558. [Last accessed November 23, 2022]. Available at: https://doi.org/10.1016/s0167-9236(03)00086-1

Hudson, John. 1989. The Birth and Death of Firms in England and Wales During the Inter-War Years. Business History. 31. 102-121. 10.1080/00076798900000067.

Intrum. 2022. European Payment Report 2022. [Last accessed November 23, 2022]. Available at: https://www.intrum.com/publications/european-payment-report/european-payment-report-2022/

Khandani, A. E., Lo, A. W., & Mukherjee, A. 2011. What happened to the quants in August 2007? Evidence from factors and covariance. Journal of Financial Markets, 2011, vol. 14, issue 1, 1-46.

Kern, F., Kuzemko, C., & Mitchell, C. 2014. Measuring and explaining policy paradigm change: the case of UK energy policy. Policy & Politics, 42(4), 513–530. [Last accessed November 23, 2022]. Available at: https://doi.org/10.1332/030557312x655765

Kim, J., & Kang, P. 2016. Late payment prediction models for fair allocation of customer contact lists to call center agents. Decision Support Systems, 85, 84-101. doi: 10.1016/j.dss.2016.03.002

Kujala, A., Herrala, A., & Murto, P. 2022. Ammattidiesel Käyttöön – SKAL Ry. [Last accessed November 23, 2022]. Available at: https://www.skal.fi/sites/default/files/sisaltosivujen_tiedostot/skal_ammattidiesel_kayttoon_raportti.pdf

Lintilä, M. 2022. FINLEX. Hallituksen esitykset: 2022, HE 92/2022 vp. [Last accessed: December 19, 2022]. Available at: https://www.finlex.fi/fi/esitykset/he/2022/20220092.pdf

Petersen, M. A., & Rajan, R. G. 1994. The Benefits of Lending Relationships: Evidence from Small Business Data. The Journal of Finance, 49(1), 3–37. [Last accessed November 23, 2022]. Available at: https://doi.org/10.1111/j.1540-6261.1994.tb04418.x

Paolella. 2019. Linear models and time-series analysis: regression, ANOVA, ARMA and GARCH (1st edition). Wiley.

SKAL Kuljetusbarometri 2/2022. 2022. SKAL. [Last accessed: November 29, 2022]. Available at: https://www.skal.fi/fi/julkaisut/skal-kuljetusbarometri-22022-kustannusnousu-vaikuttanut-rajusti-kuljetusyritysten

SKAL Koronabarometri. 2020 a. SKAL. [Last accessed November 23, 2022]. Available at: https://www.skal.fi/fi/julkaisut/skal-koronabarometri-konkurssiaalto-uhkaa-kuljetusalaa

SKAL. 2020 b. Yrittäjät: Konkurssilain koronapoikkeusten jatko suuri helpotus monelle yrittäjälle. [Last accessed November 23, 2022]. Available at: https://www.skal.fi/fi/julkaisut/yrittajat-konkurssilain-koronapoikkeusten-jatko-suuri-helpotus-monelle-yrittajalle

Sklearn.neighbors.kneighborsclassifier. Scikit-learn. [Last accessed: December 23, 2022]. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Sklearn.preprocessing.StandardScaler. Scikit-learn. [Last accessed: December 23, 2022]. Available at: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Statistics Finland. 2022 a. Goods transport by road. Reference period: 2022, 2nd quarter. ISSN=2342-3617. [Last accessed November 28, 2022]. Available at: https://stat.fi/en/publication/ckt8l87io5g100b56pttx9t34

Statistics Finland. 2022 b. Energy in Finland 2022. Doria. [Last accessed: November 25, 2022]. Available at: https://doria.fi/handle/10024/185778

Statistics Findand. 2022 c. Konkurssit kuukausittain alueittain ja toimialoittain. PxWeb. [Last accessed: December 23, 2022]. Available at: https://pxdata.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin__kony/statfin_kony_pxt_13fe.px/

Suomen Pankki. 2022. Euriborkorot kuukausittain. [Last accessed: November 28, 2022]. Available at: https://www.suomenpankki.fi/fi/Tilastot/korot/kuviot/korot_kuviot/euriborkorot_kk_chrt_fi/

US 7919150B1. 2000. Enhancing delinquent debt collection using statistical models of debt historical information and account events. Fair Isaac Corporation, San Diego, CA (US). (Shao, M., Zoldi, S., Cameron, G., Martin, R., Drossu, R., Zhang, G., Shoham, D.) 09/607,747. 30.6.2000. 6 p.

Valtiokonttori. 2022. Polttoainetuki. [Last accessed November 24, 2022]. Available at: https://www.valtiokonttori.fi/palvelut/korvaus-ja-vahinkopalvelut/polttoainetuki/

Vilmi, L., Kortelainen, M., Nelimarkka, J., Kärkkäinen, S., & Simola, H. 2022. Bank of Finland Bulletin. [Last accessed November 23, 2022]. Available at: https://www.bofbulletin.fi/en/2022/3/energy-crisis-pushing-up-general-price-level-adverse-impact-on-economic-growth-still-to-come/

Wilson, N., & Summers, B. 2002. Trade credit terms offered by small firms: Survey evidence and empirical analysis. Journal of Business Finance & Accounting, 29(3&4), 317-351. doi:10.1111/1468-5957.00434

Wooldridge, J. M. 2015. Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning. ISBN 978-1-305-27010-7.

Zeng, S., Melville, P., Lang, C. A., Boier-Martin, I., & Murphy, C. 2008. Using predictive analysis to improve invoice-to-cash collection. Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08. https://doi.org/10.1145/1401890.1402014

Zhang Z. 2016. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11), 218. https://doi.org/10.21037/atm.2016.03.37

Appendices

Appendix 1: Exploratory Timeseries Multiple Linear Regression Models for Predicting Delays

In the below table, four different models that were developed prior the final model are explored. In model 1, the dependent variable is "delays," The independent variables are diesel, Euribor-12-month interest rates, bankruptcies, imports, and exports. The coefficients $\beta_1, \beta_2, \beta_3, \beta_4$and β_5 represent the model's parameters, and the error term u_t represents the random error in the model. In model 2, the dependent variable is the natural logarithm of delays, and the independent variables are the natural logarithms of diesel, bankruptcies, imports, and exports. Independent variable Euribor-12 month is in its standard form. Model 3 is like model 2 but does not include the log(exports) variable as an independent variable. Model 4 is similar to model 2 but does not include the log(imports) variable as an independent variable.

Table 7. Models for predicting payment delays

Model	Equation
Model 1	$delays_t = \beta_0 + \beta_1 diesel_t + \beta_2 euribor\,12_t + \beta_3 bankruptcies_t + \beta_4 imports_t + \beta_5 exports_t + u_t$
Model 2	$\log(delays_t) = \beta_0 + \beta_1 \log(diesel_t) + \beta_2 euribor\,12_t + \beta_3 \log(bankruptcies_t) + \beta_4 \log(imports_t) + \beta_5 \log(exports_t) + u_t$
Model 3	$\log(delays_t) = \beta_0 + \beta_1 \log(diesel_t) + \beta_2 euribor\,12_t + \beta_3 \log(bankruptcies_t) + \beta_4 \log(imports_t) + u_t$
Model 4	$\log(delays_t) = \beta_0 + \beta_1 \log(diesel_t) + \beta_2 euribor\,12_t + \beta_3 \log(bankruptcies_t) + \beta_5 \log(exports_t) + u_t$

Appendix 2: Assessment of Normality for Exploratory Time Series Multiple Regression Models

Shapiro-Wilk test is used to check whether a sample of data comes from a normal distribution. The p-value of Shapiro-Wilk test is compared to a chosen alpha level (in this case, α=0.05) to determine whether the sample is significantly different from a normal distribution.

Table 8. Shapiro-Wilk test results comparison

Model	p-value	alpha	Normality comment
1	0.000	0.05	Sample does not look Gaussian.
2	0.000	0.05	Sample does not look Gaussian.
3	0.797	0.05	Sample looks Gaussian.
4	0.677	0.05	Sample looks Gaussian.

Appendix 3: Assessment of Heteroscedasticity for Exploratory Time Series Multiple Regression Models

The Breusch-Pagan test is used to check whether the variance of a regression model is constant across all values of the independent variables (property known as homoscedasticity). The p-value of Breusch-Pagan test is compared to a chosen alpha level (in this case, α=0.05) to determine whether the variance is significantly different from constant.

Table 9. Breuch-Pagan test results comparison

Model	p-value	alpha	Heteroscedasticity comment
1	0.02	0.05	Sample does not look homoscedastic.
2	2.105e-80	0.05	Sample does not look homoscedastic.
3	2.105e-80	0.05	Sample does not look homoscedastic.
4	0.041	0.05	Sample does not look homoscedastic.

Appendix 4: Performance of k-Nearest Neighbours (*k*NN) Classifier with Different Values of *k*

Table 10. Impact of k on the performance of the model

	Accuracy	Precision	Recall	F1 score	ROC AUC	Kappa
KNN (k=1)	0.961	0.969	0.952	0.960	0.961	0.922
KNN (k=3)	0.954	0.968	0.938	0.953	0.976	0.908
KNN (k=15)	0.930	0.957	0.898	0.926	0.978	0.860
KNN (k=19)	0.921	0.950	0.886	0.917	0.977	0.843
KNN (k=43)	0.895	0.936	0.843	0.887	0.967	0.789
KNN (k=139)	0.848	0.895	0.781	0.834	0.937	0.695
KNN (k=277)	0.817	0.872	0.736	0.798	0.915	0.635

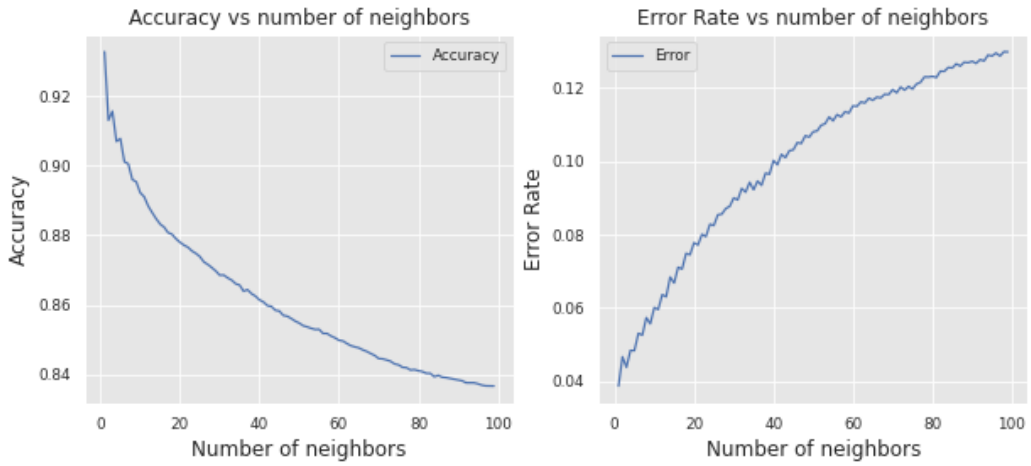

Figure 8. Impact of k on accuracy and error