IDENTIFYING UNKNOWN VARIABLES INFLUENCING CONTRACT COSTS AT MST

University of Twente

J.V.D. Straus

JANUARY 2025

UNIVERSITY OF TWENTE.

First Supervisor Dr. D. Guericke

Second Supervisor Dr. M. Machado

Company Supervisor M. Koenderink Msc.

COLOPHON

MANAGEMENT
Department
BMS

DATE
16-1-2024

REFERENCE
Reference

VERSION
1

STATUS
Status

PROJECT
Bachelor's Thesis

PROJECT NUMBER
N.A.

AUTHOR(S)
J.V.D Straus

EMAIL
j.v.d.straus@student.utwente.nl

POSTAL ADDRESS
P.O. Box 217
7500 AE Enschede

WEBSITE
www.utwente.nl

FILENAME
Thesis Justus Straus_S2636662_16-1-25

Management Summary

We identify unknown variables influencing contract costs within the Contracts & Process Management subdepartment at Medisch Spectrum Twente (MST). MST, a top clinical hospital in the Netherlands, manages several external contracts critical to operational efficiency and cost control. The subdepartment experiences significant cost fluctuations, complicating resource allocation.

The aim of this research is to uncover unknown variables that cause these fluctuations using data-driven methods, particularly Machine Learning models, to ultimately improve forecasting accuracy and contract monitoring. This research adopts the CRISP-ML, using anomaly detection and regression techniques to identify cause-and-effect relationships between variables.

Due to their financial impact, we focus on two contracts—clinical chemistry and laundry. Anomaly detection and deletion were performed on normalised costs to exclude anomalies in contract costs before performing several regression methods. Linear regression, multiple linear regression, and statistical two-way analysis of variance were used to identify relationships between variables. Shapley additive explanations, Shapley additive explanations with an adjustment factor, and random forest Regression were used to quantify variables' impacts on contract costs. Random forest regression was used to determine the accuracy when incorporating the statistically significant variables into the model.

Tables 0-1 and 0-2 summarise the key findings of this research, presenting the relative impact and direction of influence of the top five variables of the variables found to statistically significantly impact contract costs. The random forest model predicted the costs of the laundry contract with a promising result (R^2 = 0.5926). In contrast, the prediction for the clinical chemistry contract showed a negative R^2 value (R^2 = -0.2955), highlighting a need for significant improvement in the model's performance for this context.

Variable	Relative impact	Direction of influence
Number of clinical admissions	34.15%	Positive
Number of diagnoses treatment combinations	29.63%	Positive
Number of operations	17.51%	Positive
Number of visits to the outpatient clinic	8.61%	Positive
Number of day hospitalisations	8.33%	Negative

Table I Top five relative impacts laundry contract

Variable	Relative impact	Direction of influence
Number of clinical admissions	41.04%	Positive
Number of operations	21.06%	Negative
Number of diagnoses treatment combinations	12.54%	Positive
Number of day hospitalisations	10.62%	Negative
Number of visits to the outpatient clinic	7.62%	Positive

Table II Top five relative impacts clinical chemistry contract

Integrating these variables into forecasting models can help MST optimise resource allocation and improve operational efficiency. Incorporating periodic anomaly detection into contract management practices will further enhance monitoring capabilities.

This methodology offers a replicable approach for analysing other contracts at MST. The methodology can be shared with other healthcare institutions to drive improvements across the Dutch healthcare system. By utilising these insights, MST can address cost fluctuations, improve decision-making, and achieve greater control over contract management.

Acknowledgements

I want to express my gratitude to all who have supported me throughout conducting this research and writing this thesis.

First, I thank my supervisors, Daniela Guericke, Marcos Machado, and Marcel Koenderink, for their guidance and support throughout this thesis. Their quick and constructive feedback and ability to challenge me to push my boundaries have been valuable to the end product and my learning.

Furthermore, I would like to thank my colleagues from the Contracts & Process subdepartment at MST. You made me feel like a valued team member, creating a welcoming environment. The daily darting matches were a highlight of my time with you.

Additionally, I am grateful to all MST employees who assisted in various ways during this project. Your collaboration and input were essential in achieving my research objectives.

Lastly, I want to thank my family for their support throughout this project. A special thanks to my girlfriend, Nienke, for her constant encouragement and for providing valuable feedback by reviewing my thesis multiple times. I also extend my gratitude to my buddies, Ruben and Roy, for their assistance in checking the understandability of my work and offering thoughtful advice.

Thank you all for your contributions and guidance over the past months. I hope you enjoy reading this thesis as much as I enjoyed working on it.

Contents

List of Figures

List of Tables

List of Equations

Abbreviations

ANOVA	Analysis of Variance
CRISP-ML	Cross-Industry Standard Process for Machine Learning
DTC	Diagnosis Treatment Combination
iForest	Isolation Forest
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MST	Medisch Spectrum Twente
RF	Random Forest
SHAP	Shapley Additive Explanations
SLR	Systematic Literature Review
VIF	Variance Inflation Factor
XAI	Explainable AI

1. Introduction

Contract management is vital in ensuring operational efficiency and cost control within large organisations (Khan & Mir, 2021). Medisch Spectrum Twente (MST), one of the largest top clinical hospitals in the Netherlands, manages several external contracts worth millions of euros annually. However, MST's Contracts & Process subdepartment struggles with significant fluctuations in contract costs. These contract costs often exceed expectations, making it challenging to allocate resources and accurately predict future contract performance. The inability to accurately forecast due to unknown variables in the contract lifecycle creates difficulties for organisations (Brunet & César, 2019).

The core problem is that not all variables influencing contract costs are known. These unknown variables contribute to greater-than-expected variability, impacting forecasting accuracy and operational decision-making. The absence of knowledge of these variables limits forecasting ability.

The aim of this research is to uncover these unknown variables by applying data-driven models. By identifying the hidden variables influencing contract performance, the research seeks to improve forecasting accuracy and enhance contract monitoring within the Contracts & Process subdepartment at MST. Ultimately, the objective is to provide MST with the tools to predict contract outcomes better, optimise resource allocation, and ensure more effective contract management. This improvement ultimately leads to better cost control, operational efficiency, and patient service delivery.

1.1 Organisational Context

MST is a Dutch hospital in the region of Twente, with its most significant location in Enschede. It is one of the largest top clinical hospitals in the Netherlands, having approximately 3,500 employees (Medisch Spectrum Twente, 2024).

This thesis project conducts its research at MST, particularly in its Contracts & Process Management subdepartment within the Hospitality & Logistics Management department. This subdepartment facilitates all external contracts within MST and manages critical business processes to ensure efficiency, compliance, and alignment with organisational goals (MST, personal communication, September 5, 2024).

Contract management refers to the process companies use to negotiate, execute, monitor, modify, and end contracts with customers, vendors, distributors, contractors, and employees (Gutterman, 2023). Contract management in healthcare is diverse in stakeholders, types of contractual relationships, and purposes, such as waste-processing and food contracts within MST. It is crucial to view contracting as a strategic approach, as it can significantly impact the hospital's overall performance and the quality of health providers. Effective contract management is essential for maintaining efficiency, compliance, and operational efficiency (Khan & Mir, 2021).

1.2 Key Definitions

Contract costs in this research refer to the total costs of products or services acquired through a particular researched contract under contractual obligation.

Machine learning (ML) is a branch of Artificial Intelligence (AI) focused on creating analytical models. These models allow machines to adjust independently to new situations, allowing software to predict outcomes and respond well when models are applied to data (França, Monteiro, Arthur, & Iano, 2021). ML enables computers to identify patterns in large datasets and use those insights to tackle related tasks and challenges effectively.

Explainable AI (XAI) refers to a set of technical approaches to making AI systems understandable to users, covering everything from model interpretability and transparency to insights into data, performance, and uncertainty to support a holistic understanding of the AI's behaviour (Liao & Varshney, 2022).

1.3 Problem Context

The Contracts & Process subdepartment at MST faces challenges in forecasting contract costs. Historical data revealed that actual costs often exceed forecasts. These cost fluctuations result in resource management challenges, as unforeseen contract cost changes lead to inefficient resource allocation.

Ideally, contract costs would remain within expected boundaries, with known factors explaining any deviations. The subdepartment aims to uncover unknown variables to improve forecasting accuracy, enhance monitoring capabilities, and ultimately improve contract management. By identifying and understanding the key variables influencing contract performance, the subdepartment can better control costs and ensure that healthcare services are delivered efficiently.

The core problem of this research is determined by identifying the furthest actionable cause of the action problem. The action problem highlights the discrepancy between the perceived norm and reality by the problem owner. The reality should move in the direction of the norm. An action problem involves implementing change (Heerkens, Van den Winden, & Tjooitink, 2017). We define the action problem as "Larger than desired fluctuations in contract costs". Figure 1-1 shows the problem cluster, highlighting the action problem that led to identifying the core problem. The white squares represent contextual issues, the blue square highlights the core problem, and the green square denotes the action problem.

As a result of the analysis of the problem cluster, this research addresses the core problem: "Not all variables affecting contract costs are known". Identifying these variables will enable more effective contract management and forecasting at MST. The two problems that precede the core problem are not influenceable by this research and, therefore, have not been selected as the core problem.

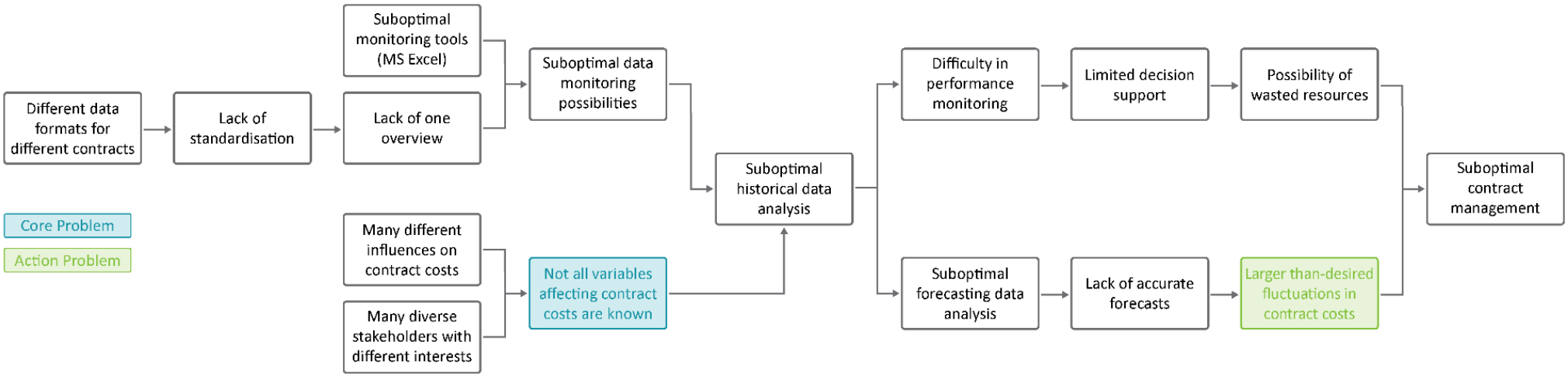

Figure 1-1 Problem Cluster

1.4 Research Questions

This chapter outlines the main research question and subquestions that guide this research, employing a data-driven approach to uncover unknown variables influencing contract costs. The following main research question emerges from the problem context:

"Which unknown variables influencing contract costs can be uncovered using data-driven models?"

The research is divided into subquestions to address the research question systematically. Subquestions one and two address the assessment of the current situation at the subdepartment. They also determine which data should be analysed.

1) "What challenges does the MST Contracts & Process Management subdepartment face in forecasting and monitoring contract costs?"
2) "What types of data are currently analysed and delivered for the contracts, and what additional data should be analysed in the research?"

After assessing the current situation, the research will focus on selecting the most suitable data-driven anomaly detection and regression models by addressing subquestions three and four.

3) "What are the most suitable data-driven anomaly detection models for analysing contract cost data in contract management in time series analysis?"
4) "Which regression models are most suitable for analysing cause-and-effect between variables in the context of this research?"

After selecting the most suitable models, they will be developed and applied to the data, addressing subquestion five.

5) "What unknown variables affecting contract costs can be identified through data-driven techniques?"

Finally, after identifying the relevant variables, their impact will be evaluated, addressing subquestions six and seven.

6) "To what extent and direction does each identified variable influence contract cost predictions?"
7) "When incorporating the identified variables, how well does the model perform, as measured by accuracy and other relevant evaluation metrics?"

1.5 Scope

The focus of this research is on MST's Contracts & Process Management subdepartment. Specifically, this study is restricted to this subdepartment and does not extend to other external or MST areas. The main objective is to identify unknown variables affecting contract costs to improve forecasting accuracy. However, developing new forecasting techniques is not part of this research.

The research analyses historical data for two contracts managed by the subdepartment, chosen for their significance to MST. The data spans from January 2019 to July 2024, based on availability. The research adopts the Cross-Industry Standard Process for Machine Learning (CRISP-ML) Reference model and integrates qualitative and quantitative methodologies.

The research does not explicity aim to enhance MST's operational efficiency; instead, it considers good operational efficiency a prerequisite for effectively managing the factors that impact contract performance.

1.6 Contribution to Knowledge

This research contributes to the knowledge of applying ML techniques and XAI methods in contract management, particularly within the healthcare sector. By identifying previously unknown variables that affect contract costs, this research advances the understanding of the cause-and-effect relationships in hospital contract management and operations.

Additionally, the research provides practical insights for MST in implementing these models to improve contract forecasting, monitoring, and overall management. This research also generates valuable insights that other hospitals can use to enhance their contract management practices. Beyond MST, this research offers a replicable framework that other healthcare institutions can adopt to tackle similar challenges.

2. Context Analysis

This chapter outlines the operational and organisational context of the subdepartment. The chapter begins by describing the two key contracts researched—clinical chemistry and laundry—and their relevance to MST's operations. Then, we identify the variables to be analysed during the research.

2.1 Researched Contracts

We focus on two of the six contracts managed by the subdepartment — a medical and facilitation services contract.

2.1.1 Clinical Chemistry Contract

Clinical chemistry is a specialised field within the hospital focused on quantitatively analysing bodily fluids for diagnostics and therapeutic purposes (Lloyd, 2023). Clinical chemistry involves specific analytical procedures that enable precise concentration measurements within the body, helping to evaluate and monitor various bodily functions.

Within MST, this contract encompasses all assays conducted in clinical chemistry (MST, personal communication, November 7, 2024). Additionally, it covers all blood testing assays within the region performed by the external party contracted by MST. The costs associated with the contract comprise a collection component and an analysis component for the required testing substance. Additionally, the contract encompasses the costs of advising clinical chemists of departments within MST where an advisor occasionally joins a medical assembly or is asked for advice.

The subdepartment responsible for these contracts meets annually with the top ten medical professional groups that incur the highest yearly costs to discuss anticipated costs and assay changes (MST, personal communication, November 7, 2024). Additionally, a standard procedure exists for requesting new clinical chemistry procedures, allowing the subdepartment to review and approve these requests, which include an estimated annual cost. This information is combined with the previous year's information to generate a yearly forecast of expected expenses. This forecast is updated monthly using the same data sources.

Each Diagnosis Treatment Combination (DTC) comprises various components essential for specific treatment, including clinical chemistry assays. Suppose any cost changes or a new clinical chemistry procedure are introduced, the DTC pricing should be adjusted to reflect the higher costs associated with this procedure (MST, personal communication, November 7, 2024). Since these new procedures incur additional expenses, they require additional reimbursement from the insurers, which is negotiated at the end of the year for the following year. To negotiate the correct reimbursements, the expected costs for next year should be accurate, underscoring the importance of a correct forecast.

Figure 2-1 shows the monthly normalised costs of the clinical chemistry and laundry contracts on a scale from zero to one. The blue line represents the clinical chemistry contract, while the orange line represents the laundry contract.

2.1.2 Laundry Contract

The laundry contract at MST covers all laundry needs within MST, ensuring a consistent supply of clean items necessary for daily hospital operations (MST, personal communication, November 18, 2024). The contract's scope can be divided into two categories: workwear, which includes uniforms for doctors and nurses, and general laundry, covering items such as bed sheets, towels, and similar essentials.

Currently, a forecast for the year is based on the costs of the past year and indexation. This approach assumes that the historical data of the preceding year provides the best estimate for upcoming costs. The forecast is initially established at the beginning of the year. It is subsequently adjusted monthly based on incoming data and gains or setbacks during the year.

Figure 2-1 shows the monthly normalised costs of the clinical chemistry and laundry contracts on a scale from zero to one. The blue line represents the clinical chemistry contract, while the orange line represents the laundry contract.

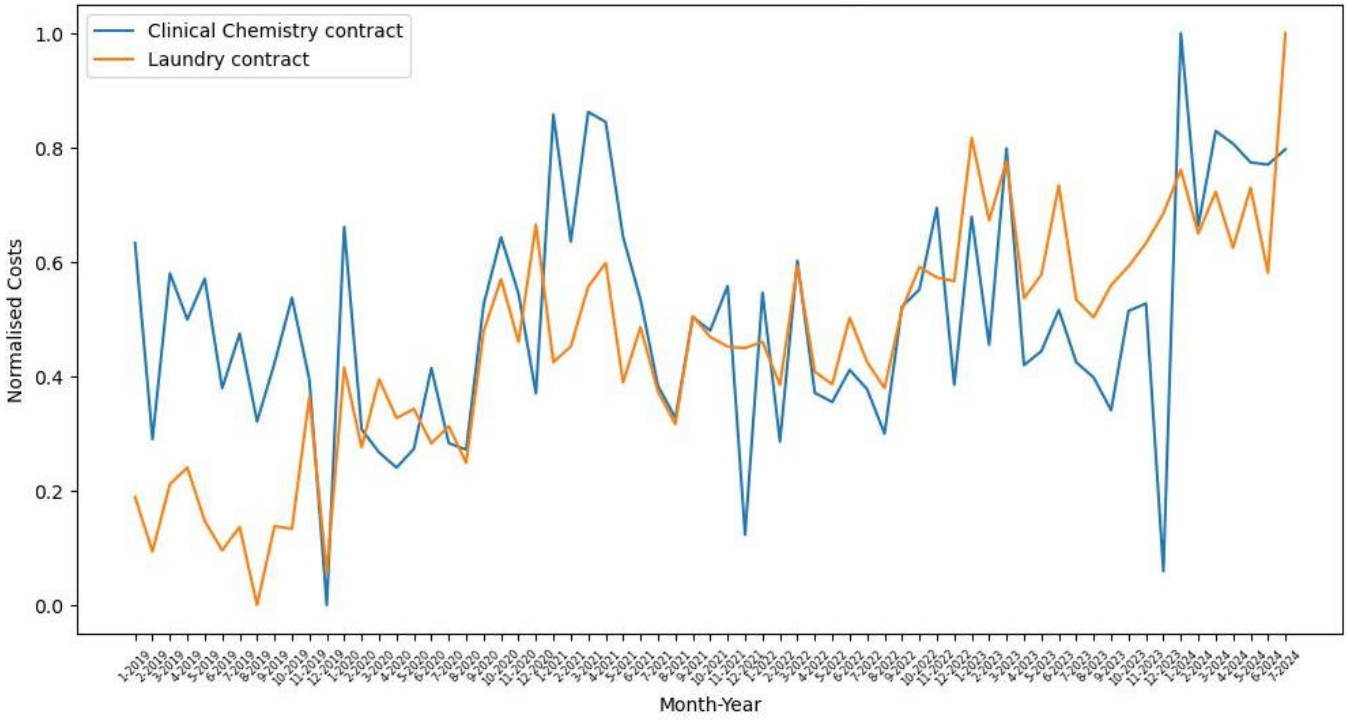

Figure 2-1 Normalised monthly costs of clinical chemistry contract on a 0-1 scale

2.2 Analysed Data

The variables for analysis were identified through stakeholder interviews and exploratory data analysis of contract data. These variables were selected because of their potential effect on contract costs. Other variables emerged during interviews with stakeholders besides the analysed variables. These variables and the rationale behind their exclusion can be found in Appendix A The variables which were analysed for both contracts include:

1) **Number of Dutch influenza cases**: The number of influenza cases in the Netherlands;
2) **Number of day hospitalisations**: Captures the number of patients admitted for procedures or treatments that only require a day stay, not overnight;
3) **Number of hospital employee shifts (eight-hour shifts)**: The total number of eight-hour shifts hospital staff work indicating operational activity levels;
4) **Number of visits at the outpatient clinic**: Captures the total outpatient clinic visits—both first-time and recurring;
5) **Number of DTCs**: This represents the number of DTCs processed. A DTC is composed of various components essential for specific treatment. Most treatments use one DTC;
6) **Number of clinical admissions**: This refers to the number of patients admitted to the hospital for overnight or extended stays for treatment or monitoring;

7) **Number of intensive care admissions**: This refers to the number of people admitted to intensive care;
8) **Number of operations**: This captures the total number of operations performed in the hospital;
9) **Seasonal effect**: Identifying whether there is a consistent seasonal pattern in contract costs. In this research, the seasons represent the quarters.

2.3 Chapter Conclusion

In this chapter, we explored the operational and organisational context of the research subdepartment at MST, identifying the challenges faced by monitoring and forecasting contract costs.

Addressing subquestion one, “What challenges does the MST Contracts & Process Management subdepartment face in forecasting and monitoring contract costs?” the chapter outlined the key inefficiencies, including unexplained cost fluctuations.

Furthermore, by addressing subquestion two, “What types of data are currently analysed and delivered for the contracts, and what additional data should be analysed in the research?” we identified the currently analysed data and highlighted the variables to be researched.

The next chapter reviews the existing anomaly detection and regression analysis literature, building on the insights presented in this chapter. It evaluates methodologies relevant to addressing the identified challenges and provides the theoretical foundation for this research.

3. Literature Review

This chapter outlines two systematic literature reviews (SLR) to select the most suitable models for anomaly detection and regression analysis in healthcare contract cost management for this research while ensuring a systematic selection of studies. We follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for selecting studies (Page et al., 2021). The PRISMA guidelines consist of four phases:

1) **Identification**: Relevant articles are found using keywords in database searches;
2) **Screening**: Studies are screened;
3) **Eligibility**: Studies are reviewed to ensure they meet the inclusion criteria;
4) **Inclusion**: The remaining studies are evaluated.

3.1 Anomaly Detection Methods

The subquestion addressed by this SLR to identify the most suitable anomaly detection method is:

"What are the most suitable data-driven anomaly detection models for analysing contract cost data in contract management in time series analysis?"

The SLR protocol can be found in Appendix B1.

We employ time series analysis to identify the most suitable data-driven anomaly detection models for analysing contract cost data. The aim is to identify unknown variables influencing unexpected deviations or fluctuations. Anomaly detection is finding data points that stand out because they significantly differ from the rest of the dataset (Niu, Lu, & Zhang, 2009). Time series anomaly detection focuses on identifying irregularities in data over time, considering the timing and sequence of the data points. This approach is essential for identifying outliers in contract cost data. Within the four included studies, the following anomaly detection methods—evaluated in the context of contract management, hospital operations, or with transferable insights from other sectors— were considered:

(1) Isolation Forest (iForest) algorithm;
(2) Support Vector Machine (SVM) algorithm;
(3) K-Nearest Neighbour (KNN) algorithm;
(4) Random Forest (RF);
(5) Autoencoders;
(6) iForest combined with autoencoders;
(7) Deep learning models;
(8) Vision Transformers.

Xiong et al. (2022) evaluated multiple anomaly detection methods for high-dimensional energy data, specifically focusing on iForest, SVM, autoencoders, and iForest combined with autoencoders. iForest, a traditional ML algorithm tailored for anomaly detection, demonstrated high efficiency in isolating outliers with minimal splits. Combining iForest and autoencoders outperformed other techniques, proving the most effective in handling high-dimensional datasets. An autoencoder is a kind of neural network that learns to recreate its input as closely as possible by comparing its output to the original input (Xiong, Zhu, Liu, He, & Zhao, 2022). Its primary purpose is to reduce the difference between the two, and it can be used for various tasks, such as detecting anomalies.

In addition, Sihabuddin et al. (2023) proposed combining the iForest algorithm with Decision Tree Regression (DTR) to enhance the identification of anomalies in regression-based forecasting in the

context of the two air quality datasets. Sihabuddin et al. (2023) demonstrated that applying the IForest algorithm before Decision Tree Regression improved all evaluated performance metrics.

Moreover, Schirmer and Mporas (2024) explored the application of deep learning models, such as convolutional neural networks and long short-term memory networks, for various time series tasks, including anomaly detection. They found these methods significantly improved the evaluated performance metrics, such as mean absolute and normalised mean squared errors, compared to traditional ML methods, such as RFs. RF, KNN and SVM were evaluated and performed well in non-DL approaches, with RF being the best-performing algorithm.

Finally, Sana et al. (2024) explored the use of Vision Transformers (ViT) for detecting anomalies in Internet of Things (IoT) networks. Sana et al. (2024) explained that Vision Transformers are a specialised adaptation of the Transformer model tailored for image analysis. They utilise an encoder and a self-attention mechanism to process image segments, known as patches. This design allows the model to discern relationships between different image components, concentrate on critical details, and effectively manage complex patterns. The study concluded that ViTs demonstrate high accuracy and frequently outperform traditional models, particularly in complex environments like IoT networks.

This SLR aimed to identify the most suitable anomaly detection models for detecting anomalies in this research, specifically focusing on contract costs. The evaluation of four articles and various methods highlighted that while some deep learning models achieve high accuracy, their complexity makes them unsuitable for this study. As a result, these models are excluded from this research (Schirmer & Mporas, 2024; Xiong et al., 2022).

Additionally, we evaluated the use of ViT for detecting anomalies (Sana et al., 2024). While this method is highly effective for image-based analysis, it is unsuitable for this research, as no image data is analysed. Consequently, ViT is excluded as a potential method in this study.

We evaluated the suitability of the ML approaches RF, KNN, SVM, and iForest (Xiong et al. 2022; Schirmer and Mporas, 2024). RF and iForest were the best-performing algorithms in the two researched studies. iForest is designed explicitly for anomaly detection, whereas RF can also perform this function similarly. iForest's specialised design makes it more suitable for identifying anomalies effectively.

In conclusion, this SLR identifies iForest as the most suitable model for anomaly detection in this research. iForest's unique architecture, explicitly designed for detecting anomalies, ensures robust and efficient performance, particularly in high-dimensional datasets. Additionally, its ability to isolate anomalies effectively while complementing regression methods enhances its suitability for this research.

3.2 Regression Methods

The subquestion addressed by this second SLR to identify the most suitable regression models is:

"Which regression models are most suitable for analysing the relationship between variables in time series analysis in the context of this research?"

The SLR protocol can be found in Appendix B2.

The basic idea of regression analysis is that predictor variables explain an outcome variable (Zapf, Wiessner, & König, 2024). In the context of this research, regression analysis is adopted to explain the causes of cost changes. Within the five included studies, the following regression methods —evaluated

in the context of contract management, hospital operations, or with transferable insights from other sectors— were considered:

1) Univariate Logistic regression;
2) Adjusted logistic regression;
3) RF regression;
4) Support vector regression
5) Linear regression;
6) Multiple Linear Regression (MLR);
7) K-nearest neighbour regression;
8) L-1-regularised logistic regression;
9) Classificational regression.

Luo et al. (2021) proposed using univariate logistic regression, adjusted logistic regression, or RF in a hospital setting. These methods are used in the context of the cost of asthma treatment. The article stated that RF mainly improves the prediction accuracy of the tested regression methods. The key outcome variable in the research by Luo et al. (2021) is whether a patient is classified as "high-cost" or "low-cost". These variables are categorical, meaning a binary classification is given to the variables. This classification is done through the K-means algorithm, assigning costs above a certain threshold as high and costs below a certain threshold as low.

Additionally, Guo et al. (2022) proposed adopting support vector regression (SVR) in the context of cost prediction of maintenance and operation costs. This article concluded that it is highly effective in handling small datasets and minimises prediction errors using a non-linear kernel. Moreover, this article also compared linear regression and SVR in prediction. It concluded that SVR is more effective than linear regression. However, this article does not emphasise SVR's capability to uncover or interpret correlations between variables. Instead, SVR is utilised to develop a model that produces accurate predictions based on historical data without offering insight into the direct relationships or dependencies between the predictor variable and the target outcome.

Furthermore, Maryati et al. (2023) proposed several non-regression methods to forecast gold prices, such as exponential smoothing. This article also suggested using linear regression to forecast gold prices over the last 70 years. This article stated that linear regression severely underperforms exponential smoothing in predicting gold prices and struggles with accuracy in non-linear, volatile contexts. Specifically, as the article highlighted, gold prices exhibit significant short-term fluctuations, which are better captured by non-linear methods. However, linear regression remains reliable for comparing linear and stable variables over time. The article's findings suggested that the limitations of linear regression are primarily observed in highly volatile and complex data. In scenarios where volatility and nonlinearity are less frequent, linear regression can still be a suitable method. Thus, the article does not contradict the use of linear regression in other settings where the data exhibits more linear patterns, and a correlation should be observed.

In contrast, using time series analysis, Ampadu et al. (2024) proposed using MLR alongside the SARIMA model to predict the average annual cost of crashes on the US-16 Wyoming Downgrade. This study concluded that MLR is a robust method for forecasting crash costs by simultaneously accounting for the influence of multiple variables. MLR allows for modelling the effects of factors such as alcohol involvement, weather conditions, road surface types, and heavy truck presence on crash outcomes. By using MLR, the study estimated how each predictor contributes to changes in crash costs while controlling for the other variables. This approach is well-suited to understanding the interactions that drive costs. Since we focus on discovering multiple variables affecting contract costs, MLR suits this

research since it allows for modelling the effects of various factors. This approach is well-suited to understanding the correlation of independent and dependent variables.

Lastly, Nghiem et al. (2023) proposed several ML regression methods to predict high health-cost users among people with cardiovascular disease. The proposed methods are RF regression, K-nearest neighbour regression, L1-regularised logistic regression, and classification regression. These ML approaches, mainly RF and L1-regularised logistic regression, provided good predictive performance. The study concluded that ML offers a valuable tool for predicting and identifying risk factors in health economics, potentially aiding in planning health services and improving preventive measures. The combination of ML methods and regression can be seen as promising for this research.

This SLR aimed to identify the most suitable regression models for analysing cause-and-effect relationships between variables in the context of this research, specifically focusing on contract costs. The analysis of five regression methods revealed that some models are effective for prediction but less suitable for capturing direct relationships between variables. For instance, univariate and adjusted logistic regression, as well as K-nearest neighbour regression, are typically used for classification and prediction tasks in healthcare cost settings (Luo et al., 2021; Nghiem et al., 2023), making them less suitable for correlation analysis in this context. Similarly, SVR is recognised for its predictive accuracy, especially with smaller datasets (Guo, Wang, Zheng, & Ding, 2022). However, SVR cannot effectively uncover direct, interpretable correlations between variables. This limitation makes SVR less suitable for analysing cause-and-effect relationships, which is the primary focus of this research.

Additionally, RF regression can not effectively uncover direct correlations between variables. However, as highlighted by the articles, RF provides advantages when dealing with more complex and non-continuous data structures and excels in this. It can adapt to handling various data types, including categorical data. Although primarily used for prediction, as shown by Luo et al. (2021) and Nghiem et al. (2023), RF Regression can complement other regression techniques by offering robust predictive performance. However, its strength lies more in uncovering complex patterns than explaining direct cause-and-effect relationships. Therefore, while RF regression is valuable for improving prediction accuracy, it should be used alongside other regression or statistical methods to capture correlations.

Additionally, linear regression, despite its limitations in highly volatile or non-linear datasets (Maryati, Christian, & Paramita, 2023), remains reliable for data with linear patterns, which aligns with this research's objective of identifying correlations in a more stable context. The most suitable method identified is MLR, as it allows for the simultaneous modelling of several variables and captures the interactions between independent and dependent variables (Ampadu, Ker, Wulff, & Ksaibati, 2024). This ability makes MLR especially suitable for this research, as it seeks to uncover multiple variables affecting contract costs. However, linear regression is also adopted when analysing a single variable's effect on the dependent variable.

In conclusion, we select MLR through this SLR as the most suitable model for this research because it can model relationships between multiple independent variables and their outcome, offering insights into the cause-and-effect relationships impacting contract costs. However, before performing MLR, the simple relationship between one dependent and independent variable should be determined to include the variable in the MLR. We use linear regression to determine whether a variable should be included. RF regression is proposed for scenarios involving non-continuous data, where its strength in handling complex, non-linear patterns can complement the analysis. However, a suitable statistical method should complement this regression method to analyse the cause-and-effect between variables.

3.3 Gaps in Literature

The existing anomaly detection and regression analysis literature has provided valuable insights across various domains. However, several critical gaps in the literature remain, particularly in the context of healthcare contract cost management. We address the following gaps observed in the evaluated literature.

Firstly, multicollinearity among variables is a common challenge in MLR. However, the reviewed literature did not provide suitable solutions to address this issue. We address this gap by combining variables based on SHAP importance.

Secondly, The reviewed studies lacked the application of XAI techniques. We fill this gap using Shapley Additive exPlanations (SHAP) analysis. Furthermore, an adjustment factor ensures accurate SHAP values in a multicollinear environment.

Thirdly, the reviewed studies do not compare adjusted SHAP values to the original SHAP values to validate the accuracy and reliability of the adjustments made for multicollinearity. We address this gap by evaluating the correctness of the adjusted SHAP values, ensuring their alignment with expected differences and values.

Lastly, the observed literature did not validate the found XAI values with field experts. We address this gap by validating the obtained XAI values through consultation with experts from the subdepartment where the study was conducted.

3.4 Chapter Conclusion

In this chapter, we conducted two SLRs to identify the most suitable models for anomaly detection and regression analysis in healthcare contract cost management. Following the PRISMA guidelines, relevant studies were systematically identified, screened, and evaluated.

To address subquestion three, “What are the most suitable data-driven anomaly detection models for analysing contract cost data in contract management in time series analysis?” we identified iForest as the most suitable method.

Subquestion four asks, “Which regression models are most suitable for analysing the relationship between variables in time series analysis in the context of this research?”. We selected MLR as the primary method to model relationships between independent and dependent variables. We selected linear regression and RF to complement MLR. We also identified critical gaps in the literature, which this research addresses.

The insights from this chapter provide a solid methodological foundation for the research. The next chapter details the methodology employed to implement the selected models and address the identified research gaps.

4. Methodology

This chapter outlines the methodology employed to address the main research question. The research follows a structured and systematic approach using the CRISP-ML framework.

The selected ML models, including iForest, linear regression, MLR, and RF Regression, are discussed in detail. Furthermore, in this chapter, a novel approach to address multicollinearity by combining variables based on SHAP values is introduced. A complementary statistical technique, two-way ANOVA, is incorporated to identify and evaluate stable seasonal effects within contract costs.

Finally, this chapter outlines the validation metrics used to evaluate model performance. Figure 4-1 provides a global overview of the model applied in this research, showing the sequential steps from anomaly detection to the calculation of validation measures.

Figure 4-1 Model overview

4.1 Cross-Industry Standard Process for Machine Learning Reference Model

We adopt the CRISP-ML Reference Model, which provides a structured and iterative approach to developing ML solutions (Studer et al., 2021). This model supports a dynamic process that evolves with the data, continuously revisiting the problem definition and research questions as new insights emerge. By guiding the transformation from raw data to an ML model and eventually to actionable insights, CRISP-ML ensures that the research adapts to emerging findings, making it well-suited for this study's focus on identifying unknown variables through ML models. Figure 4-2 depicts the iterative model (Abonyi, Kummer, & Hanzeliek, 2022).

This iterative CRISP-ML model consists of the following six phases:

(1) **Business & Data Understanding**: Align business objectives, data, and context to ensure the available data effectively addresses the project's goals while considering the broader operational context. This phase helps in selecting the most appropriate methods for model development.
(2) **Data Preparation**: Clean the dataset, select relevant features, and standardise formats, ensuring data is on a consistent scale.
(3) **Modelling**: Select appropriate ML models based on insights from literature reviews. Choose the most relevant and effective models for the problem at hand.
(4) **Evaluation**: Assess the model's performance using various tests on the time series to ensure validation. This phase is iterative, possibly revisiting earlier stages if further refinement is needed.
(5) **Deployment**: Implement the model on real-world data.
(6) **Monitoring and maintenance**: Continuously monitor and update the model to address data shifts, preventing performance degradation and ensuring its performance remains optimal over time.

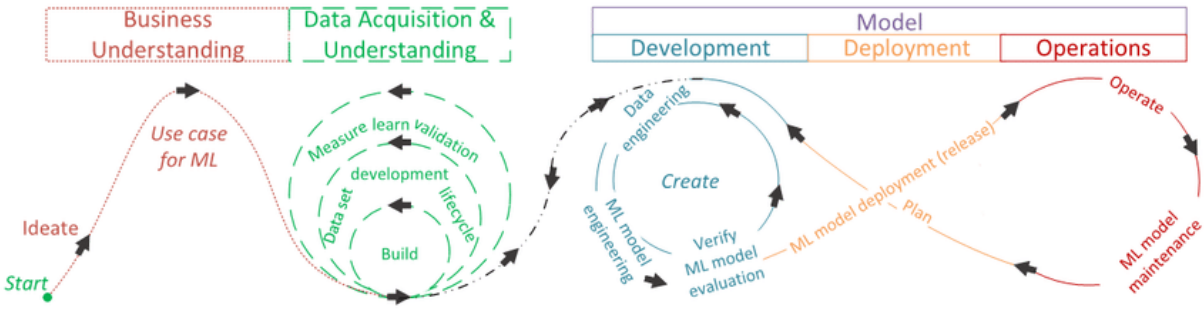

Figure 4-2 CRISP-ML Process

4.2 Machine Learning Models

The ML field distinguishes two types of ML models: supervised and unsupervised (Jordan & Mitchell, 2015). Supervised ML refers to systems where the model is trained using labelled data, meaning the input comes with the correct output. This method is used to make predictions or classifications. In contrast, unsupervised ML involves learning patterns from data without labelled responses. This approach is used to find structure in data, such as clustering. In this research, IForests and RFs are used as unsupervised approaches. At the same time, linear regression and MLR are applied as supervised approaches. Additionally, SHAP is incorporated as an interpretation method to provide insights into feature contributions across these models. SHAP does not perform predictions but is used to explain the influence of each feature on the model's predictions, enhancing the interpretability of supervised and unsupervised models.

4.2.1 Decision Tree

A decision tree is a method that breaks down the data step by step (Rokach & Maimon, 2005). At each step, the data is split into smaller groups based on the values of specific input features, continuing this process until it has been grouped into specific categories. A decision tree consists of internal nodes (which perform tests on attributes), branches (which represent the outcomes of the tests), and leaf nodes (which indicate the final classification or decision). The process involves selecting the attribute at each node to split the data. This theory is foundational for several ML models later explained in this chapter, which build upon the decision tree structure. Figure 4-3 depicts an example of a decision tree where the circles show internal nodes, the lines show branches, and the triangles show leaf nodes (Rokach & Maimon, 2005).

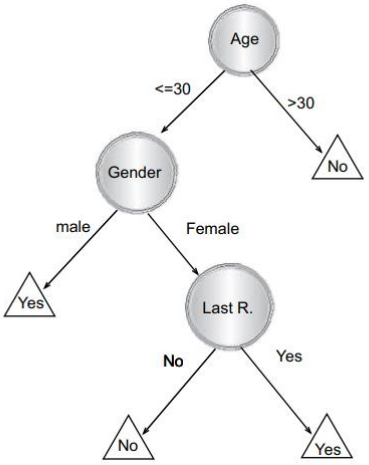

Figure 4-3 Decision tree presenting response to Direct Mailing

4.2.2 Isolation Forest

The iForest algorithm is grounded in the concept that anomalies are typically few and possess distinctive attributes compared to regular data points (Liu, Ting, & Zhou, 2009). These characteristics make anomalies more susceptible to isolation. The fundamental idea is that the path length required to isolate an instance in a tree structure can be used to indicate its anomaly status and, based on this value, can be flagged as an anomaly. iForest creates different Isolation Trees (iTrees) for a data set. iforest selects anomalies by looking for data points with shorter average path lengths across iTrees, meaning these data points are the easiest to isolate through iTrees.

An iTree is a binary tree structure that builds upon the decision tree theory and is designed to split data into smaller subsets at each branch (Liu, Ting, & Zhou, 2009). An iTree is looking to isolate data points from the dataset. Each node in an iTree represents a decision point where the dataset is split based on random attribute values, and a split value is chosen within that range of the attribute. This random partitioning helps create diverse trees that can efficiently isolate anomalies. The root node's path length indicates how easily a data point can be isolated. Generally, fewer and more distinct anomalies are isolated closer to the tree's root, resulting in shorter path lengths.

In contrast, regular instances require more splits to isolate, leading to longer path lengths. The smaller the iTree value, the easier the data point is isolated. The anomaly score is derived from its average path length across multiple iTrees. This score is normalised to fall between minus one and one. A lower score suggests that a data point is more likely to be an anomaly, as it suggests a shorter path length and easier isolation.

In this research, anomaly detection is carried out on normalised costs over the years, considering the annual indexation of the contracts managed by the subdepartment. This normalisation process ensures that genuine anomalies are detected rather than highlighting the years with the lowest and highest values. This normalisation was performed using Z-Score normalisation because this normalisation transforms the data into the standard normal distribution, enabling comparisons by highlighting deviations from the mean (Al-Faiz, Hadi, & Ibrahim, 2019). This normalisation technique suits time series with yearly indexations by isolating values based on the mean costs, making them stand out as anomalies.

Figure 4-4 illustrates an isolation tree where each node represents a decision point based on a specific feature or value. The figure illustrates an example of anomaly detection performed on the laundry contract. The branches indicate how the data points are separated based on those features. The nodes at the top are more general than those further down the trees because these focus on more specific splits. After being isolated, the model assigns the value, which is the length of the node, the anomaly value.

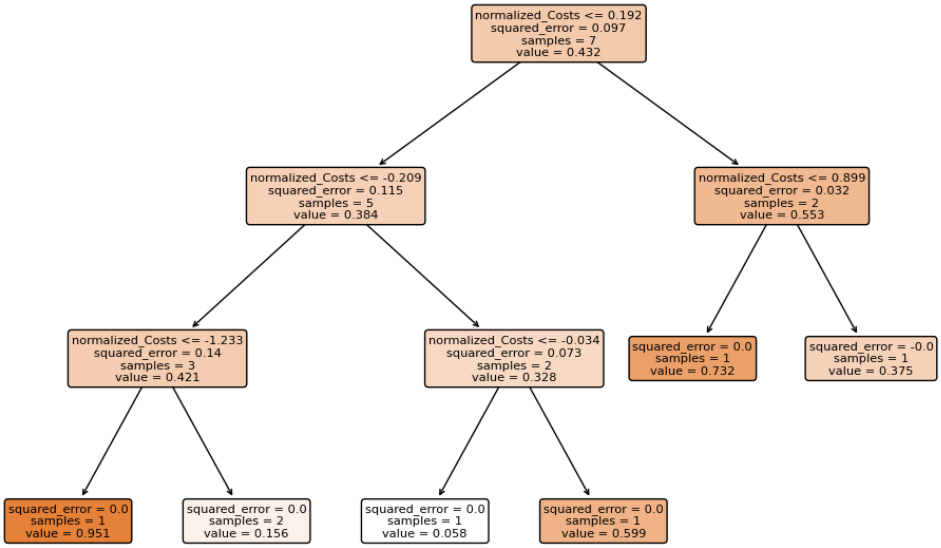

Figure 4-4 Isolation Tree example laundry contract

4.2.3 Linear Regression

Regression is a test to understand the relationship between two variables. Linear regression is a straightforward and widely used machine-learning algorithm that applies mathematical techniques to predict outcomes based on input data (Maulud & Abdulazeez, 2020). Linear regression is a model with a single independent variable.

Equation (1) expresses linear regression (Maulud & Abdulazeez, 2020).

$$y = k1X1 + c \quad (1)$$

Where:

- y denotes the dependent variable (outcome).
- $X1$ denotes the independent variable (predictor).
- $k1$ denotes the coefficient of the independent variable.
- c denotes the constant intercept.

4.2.4 Multiple Linear Regression

MLR is a statistical method used to examine the relationship between multiple independent variables (predictors) and a single dependent variable (outcome) (Marill, 2004). MLR extends the concept of simple linear regression, which involves only one predictor, by incorporating two or more predictors to provide a more comprehensive model.

Equation (2) expresses MLR (Marill, 2004).

$$y = k1X1 + k2X2 + \cdots + c \quad (2)$$

Where:

- y denotes the dependent variable (outcome);
- $X1\ and\ X2$ denote the independent variables (predictors);
- $k1\ and\ k2$ denote the coefficients that show the impact of each predictor on the outcome;
- c denotes the constant intercept.

MLR relies on the following assumptions:

- The relationship between the predictors and the outcome is linear;
- The differences between observed and predicted values' residuals should be independent;
- The residuals exhibit homoscedasticity (constant variance);
- The residuals are normally distributed.

The best-fitting plane for the data is determined through the least-squares technique. This method identifies the plane that reduces the total squared differences (residuals) between the actual y-values and the predicted y-values based on the regression plane.

One risk of employing MLR is where two or more independent variables are highly linearly correlated. This high correlation can significantly impact the estimated coefficients of the regression model. This phenomenon is called multicollinearity (Shrestha, 2020). The impacts of multicollinearity are that coefficients become unstable, and the standard errors of the coefficients increase. Multicollinearity should be assessed before employing the model to prevent the model from shifting. Shrestha (2020) proposed different methods to detect multicollinearity, of which two are used in this research.

Firstly, a pairwise scatterplot with Pearson's coefficient can be used to determine multicollinearity between variables. This coefficient quantifies the strength of these relationships, with values close to ±0.8 indicating possible collinearity. The coefficient is calculated through Equation (3) (Shrestha, 2020).

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{[n\Sigma X2 - (\Sigma X)2][n\Sigma Y2 - (\Sigma Y)2]}} \quad (3)$$

Where:

- r denotes the correlation coefficient;
- n denotes the number of data points;
- X denotes the first variable in the context;
- Y denotes the second variable in the context.

Secondly, the Variance Inflation Factor (VIF) quantifies how much a regression coefficient's variance increases as a result of multicollinearity. Table 4-1 shows the interpretation of each VIF value. A VIF value below five is seen as acceptable in this research (Shrestha, 2020).

VIF Value Range	Interpretation
VIF = 1	Variables are not correlated
1 < VIF ≤ 5	Variables are moderately correlated
5 < VIF ≤ 10	Multicollinearity exists among predictors in the regression model
VIF > 10	High multicollinearity: regression coefficients become unreliable in multicollinearity since high correlations among predictor variables make it difficult to distinguish their contributions to the model

Table 4-1 VIF values interpretation

Standardising variables before calculating the VIF helps ensure that multicollinearity is accurately assessed without being skewed by differing variable scales (Salmerón, García, & García, 2018). Therefore, we calculate VIF on scaled variables from minus one to one. Equation (4) expresses the calculation of VIF (Shrestha, 2020).

$$\text{VIF} = \frac{1}{1 - R^2} = \frac{1}{\text{Tolerance}} \quad (4)$$

Where:

- Tolerance denotes the inverse of VIF;
- R^2 denotes the coefficient of determination obtained by regressing the predictor on all other predictors;
- $1 - R^2$ indicates the proportion of variance in the predictor variable that remains unexplained by the other predictors.

Kim (2019) proposed combining multi-collinear variables based on a hierarchal value, enabling the calculation of each variable's coefficient. We will combine variables using SHAP values, as detailed in Chapter 4.2.7.

P values are commonly used to measure the statistical significance of each predictor in the model, indicating whether the independent variable has a meaningful impact on the dependent variable. The *P*-value measures the probability of whether the null hypothesis of no relationship between the variables is true (Bonovas & Piovani, 2023). In MLR, as Uyanık and Güler (2013) mentioned, *P*-values help determine whether an independent variable contributes statistically significantly to the dependent variable. By standard practice, a *P*-value below the threshold of 0.05 suggests that the independent variable significantly affects the dependent variable. On the other hand, a higher *P*-value indicates a lack of sufficient evidence to reject the null hypothesis.

4.2.5 Random Forest Regression

RF, introduced by Breiman, is a tree-based ensemble with multiple decision trees to improve predictive performance and robustness (Breiman, 2001).

RF builds upon the concept of decision trees but introduces two elements of randomness, making it a more robust and versatile model for regression tasks (Breiman, 2001). The two main elements of randomness are Bootstrap sampling (Bagging) and Random feature selection. Bagging involves creating multiple random subsets of the original dataset through a process called sampling with replacement. This replacement means that each time a sample is drawn, the model places it back into the original dataset, allowing it to be selected again. As a result, some data points may appear more than once in a given subset, while others may not be included at all. Bootstrapping avoids fitting too closely to any particular set of data points and ensures that each tree learns different patterns through the variation. This process leads to the combined model being more robust and accurate.

As a result of random feature selection, only a random subset of features is considered at each split in the decision tree. This random selection increases diversity among the trees, reducing the correlation between the trees and, thereby, improving the model's overall robustness.

An RF regression model begins by applying bagging to create multiple trees. Each tree is trained on a slightly different dataset version, with random feature selection occurring at each node. By having random trees which are not too similar, RF regression prevents overfitting when more trees are added. Overfitting occurs in supervised ML when a model fits the training data too perfectly, leading to poor performance on unseen data (Ying, 2019). Overfitting typically results from small or noisy training sets, overly complex models, or algorithms that fail to generalise well.

Once all trees are built, the prediction process begins. In RF regression, each tree makes a prediction, and the model calculates the average of all tree predictions to generate the final result.

Categorical features are categorised using one-hot encoding in this research, where each category is a separate binary column; one value indicates the presence of that category, while the remaining columns are set to zero (Qiu, Liu, Zhou, & Huang, 2022). Seasonality is a categorical feature, with each quarter being a category. Consequently, all quarters are analysed separately in the RF regression.

4.2.6 Shapley Additive Explanations

Understanding why and how a model makes a particular prediction is often just as important as its accuracy (Lundberg & Lee, 2017). This understanding improves the interpretability of models, making them easier to explain and trust. Lundberg and Lee (2017) introduced SHAP values as a unified approach to measuring feature importance. SHAP assigns an importance value to each variable, explaining how each contributes to the model's predictions.

Building on the origins of SHAP values in cooperative game theory, this approach derives from the Shapley value, a concept in cooperative game theory that allocates the total value of a coalition to individual players based on their contributions (Nisan, Roughgarden, Tardos, & Vazirani, 2007). The Shapley value ensures fairness by considering all possible orders in which players – or features in the context of SHAP – can join the coalition. Each player's contribution is measured as the average marginal impact they bring to every possible subset of the coalition they join. In SHAP, features are treated as "players" in a cooperative game; the "value" is the prediction made by the model. The SHAP value for a feature represents its contribution to the prediction, averaged over all possible subsets of other features.

SHAP values ($\emptyset_i$) are unique because they meet three key requirements – local accuracy, missingness, and consistency – in explaining the predictions.

Local accuracy ensures that the explanation provided by SHAP aligns perfectly with the original model's output for a specific prediction. Essentially, the sum of the SHAP values for all variables in the explanations must exactly match the model's prediction $f(x)$ for a given input x. This requirement, expressed in Equation (5), makes SHAP reliable for individual predictions (Lundberg & Lee, 2017).

$$f(x) = g(x') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x_i' \quad (5)$$

Where:

- $f(x)$ denotes the model's original output, which refers to the predicted value or output generated by the machine learning model when all features are included;
- M denotes the number of features;
- $g(x')$ denotes the explanation output for the simplified input x';
- $\emptyset_i$ values denote the feature attributions for each feature i.

Missingness ensures that "missing" features do not impact the prediction. If a feature x'=0 (which means it is absent), then its contribution $\emptyset_i$ should be zero in the explanation model. This property guarantees that only the features influencing a prediction contribute to the output. Equation (6) expresses this property mathematically (Lundberg & Lee, 2017).

$$x'_i = 0 \rightarrow \emptyset_i = 0 \quad (6)$$

Consistency means that if a feature becomes more important to a prediction (or stays equally important) while everything else remains the same, its SHAP value should not decrease. In other words,

if a model relies more on a particular value for a prediction, the explanation should reflect that change. Equation (7) expresses this mathematically (Lundberg & Lee, 2017).

$$\text{If } f'_x(z') - f'_x(z'\backslash i) \geq f_x(z') - f_x(z'\backslash i) \text{ for all inputs } z' \in \{0,1\}^M, \text{ then } \emptyset_i(f', x) \geq \emptyset_i(f, x) \quad (7)$$

Where:

- $f(x)$ denotes the model's original output;
- $f_x(z'\backslash i)$ denotes the model's output for the subset z' excluding feature i;
- f'_x denotes the model's modified or new version;
- z' denotes whether the feature is included or excluded, where one indicates the feature is included and zero indicates it is excluded;
- $\emptyset_i$ values denote the feature attributions for each feature i.

On the condition of satisfying these requirements, the SHAP values are calculated by considering each feature's impact across all possible combinations of features. For each subset of features, the model's output is calculated with and without the feature in question, representing the marginal contribution. The difference between this output is weighted based on the subset's size. The SHAP value is the average of the marginal contributions, reflecting the overall importance while accounting for interactions with other features (Nisan, Roughgarden, Tardos, & Vazirani, 2007). Equation (8) expresses this mathematically (Lundberg & Lee, 2017).

$$\emptyset_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|! \cdot (M - |\, z \,| - 1)!\,|}{M!} (f_x(z') - f_x(z' \setminus i)) \quad (8)$$

Where:

- $\emptyset_i$ values denote the feature attributions for each feature i;
- $|z'|$ is the number of non-zero entries in z';
- M is the number of features;
- $z' \subseteq x'$ represents all z' subsets where the non-zero entries are a subset of the non-zero entries in x'
- $f(x)$ denotes the model's original output;
- $f_x(z'\backslash i)$ denotes the model's output for the subset z' excluding feature i.

While SHAP values provide a robust and consistent framework for feature importance, the method has one notable limitation. When independent variables are correlated, their marginal contributions overlap, leading to ambiguity in assigning their individual effects (Mishra, 2016). Consequently, this can lead to misleading SHAP values because the algorithm can not distinguish individual effects, leading to over- or underestimation of scores. Therefore, careful interpretation of the SHAP values is necessary.

A multicollinearity correction method has been proposed by Basu and Maji (2022) to address multicollinearity in SHAP calculation. This method adjusts SHAP calculations by introducing an Adjustment Factor, which removes the correlation effect of one feature on others during the calculation. Consequently, features are uncorrelated, allowing their contributions to be evaluated independently.

For an individual feature X_j, the adjustment factor AF_k ensures that the correlation between X_j and X_k is nullified. Equation (9) expresses the adjustment factor (Basu & Maji, 2022).

$$AF_k = -\frac{cov(X_j, X_k)}{var(X_j)} X_j \quad (9)$$

Where:

- $cov(X_j, X_k)$ denotes the Covariance between X_j and X_k, measuring their linear relationship;
- $var(X_j)$ denotes the variance of X_j, indicating its spread;
- X_j denotes the feature causing the correlation.

Equation (10) expresses the corrected SHAP value (Basu & Maji, 2022).

$$Corrected\ \emptyset_i = \emptyset_i + \sum_{k \neq i}^{M} AF_k \quad (10)$$

Where:

- M denotes the number of features.

A matrix approach can be used when multiple features are present. This approach ensures that SHAP values accurately reflect feature importance by nullifying multicollinearity (Basu & Maji, 2022). However, it requires much computational power, becoming more significant with larger datasets. This limitation is not an issue for this research since the data size is not too large.

We will apply the corrected and non-corrected SHAP values to compare their outputs and evaluate whether the expected differences are observed.

4.2.7 Combining Variables Through Shapley Additive Explanations

This section introduces a unique combination of methods for addressing multicollinearity in regression modelling. This method combines variables through SHAP importance with normalisation techniques to create a representative variable while maintaining interpretation.

As discussed earlier, SHAP values quantify the contribution of each variable to the model's output, providing insights into the relative importance of the independent variables. In this context, SHAP values are utilised to combine highly correlated variables to address multicollinearity.

Suppose the Pearson correlation coefficients indicate a strong relationship between two or more variables. Then, they are combined into a single weighted normalised variable. This approach ensures that the combined variable accurately reflects the relative importance of its components, as weighed by SHAP values. Furthermore, normalisation ensures that the combined variable is on the same scale as its components.

An RF regression is performed using the multicollinear variables as independent variables as predictors of the dependent variable. RFs are not affected by multicollinearity because, at each split, the algorithm is restricted from considering a random subset of the predictors (Breiman, 2001). This restriction prevents any one predictor from dominating the decision-making process in every tree, especially when there is a strong predictor in the data (James, Witten, Hastie, & Tibshirani, 2021). If the algorithm allowed all independent variables to be considered, trees would become overly similar, leading to correlated prediction. Limiting each split to a random subset of predictors makes the trees uncorrelated, making the resulting average more reliable and less variable, thus reducing the impact of multicollinearity.

The results of the RF Regression model are then used to compute the SHAP values for each independent variable, following Equation (8).

Subsequently, to calculate a single combined variable while maintaining the correct pattern, the independent variables are normalised to a standard scale by adjusting their values to fall within the same numerical range. Benhar, Idri, and Fernández-Alemán (2020) suggested using min-max normalisation to achieve this while preserving the linear relationship between the variables. Equation (11) expresses the formula for min-max normalisation (Benhar, Idri, & Fernández-Alemán, 2020).

$$x'_i(j) = \frac{x_i(j) - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{11}$$

Where:

- $x'_i(j)$ denotes the normalised value x_i at data point j;
- $\min(x_i)$ denotes the minimum value of x_i;
- $\max(x_i)$ denotes the maximum value of x_i;
- $x_i(j)$ denotes the value of x_i at data point j.

After calculating the combined normalised variable, the mean SHAP value for each independent variable is computed to determine its average contribution to the model's result. This calculation is done by summing the absolute SHAP values and multiplying the sum by the reciprocal number of SHAP values. Equation (12) expresses the mean SHAP value formula.

$$\overline{\emptyset_j} = \frac{1}{b}\sum_{j=1}^{b}|\emptyset_j| \tag{12}$$

Where:

- $\overline{\emptyset_j}$ denotes the average SHAP value of variable j;
- $|\emptyset_j|$ denotes the absolute value of $\emptyset_j$, representing the value's distance from zero;
- b denotes the number of SHAP values from this variable.

After calculating the average SHAP values, the relative average SHAP value is calculated to determine the relative contribution of each independent variable. The relative mean SHAP value is calculated by dividing the average SHAP value by the sum of all the average SHAP values. Equation (13) expresses the relative mean SHAP value formula.

$$Relative(\overline{\emptyset_\iota}) = \frac{\overline{\emptyset_\iota}}{\sum_{i=1}^{n}\overline{\emptyset_\iota}} \tag{13}$$

Where:

- $Relative(\overline{\emptyset_\iota})$ denotes the relative average SHAP value of variable j;
- $\overline{\emptyset_\iota}$ denotes the average SHAP value of variable j;
- n denotes the total number of variables.

Once the relative average SHAP values are computed, the combined variable for each data point is calculated by multiplying the normalised variables by their respective relative SHAP values. Equation (14) expresses the formula for the combined variable.

$$x_{combined}(j) = Relative(\overline{\emptyset_1}) * x'_1(j) + Relative(\overline{\emptyset_2}) * x'_2(j) \quad (14)$$

Where:

- $x'_1(j)$ and $x'_2(j)$ denote the correlated normalised variables at data point j;
- $Relative(\overline{\emptyset_1})$ and $Relative(\overline{\emptyset_2})$ are their respected relative average SHAP values;
- $x_{combined}(j)$ is the combined variable at data point j.

The separate independent variables' coefficients can be calculated after performing MLR with the combined independent variable. The following formula calculates the coefficient by multiplying the combined coefficient by the relative average SHAP value, then dividing by the scaling factor used in the normalisation formula. This scaling factor is essential to adjust the coefficients to their actual scale, reflecting their true size compared to their normalised values. Equation (15) expresses the formula for recalculating the individual coefficients.

$$coef_i = \frac{coef_{combined} * Relative(\overline{\emptyset_i})}{\max(x_i) - \min(x_i)} \quad (15)$$

Where:

- $coef_i$ denotes the individual coefficient of variable i scaled to the original non-normalised scale;
- $coef_{combined}$ denotes the coefficient of the combined variable;
- $Relative(\overline{\emptyset_i})$ denotes the relative average SHAP value of variable j;
- $\max(x_i) - \min(x_i)$ denote the maximum and minimum values of x_i, which is the scaling factor of the normalising formula used to combine the independent variables.

One could argue that RF regression's built-in feature importance – Gini Importance – is more suitable than SHAP importance values. Gini Importance is calculated during the training phase of the decision tree splitting process (Menze et al., 2009). At each decision tree node, the Random Forest algorithm selects the feature a threshold that most effectively reduces the Gini Impurity – a measure of how well the split separates the data into distinct classes. These reductions are accumulated for each feature across all nodes and all trees in the forest, resulting in a total importance score. This score is then normalised to provide a relative ranking of feature relevance within the model. Since the measure is computed during the training process, it is efficient and provides a quick way to identify which feature the RF considers most important.

However, as Menze et al. (2009) noted, Gini's importance has notable limitations. It has notable biases as it tends to overestimate the importance of features with higher variability or more unique categories, leading to misleading results in some datasets. SHAP uses game theory to assign fair contributions to variables to avoid this bias, regardless of their variability or number of categories. Furthermore, the Random sampling and feature selection process in RF can result in variability in Gini importance scores across different runs. SHAP values do not change among different runs because of their consistency property, offering consistent and reproducible explanations.

Therefore, using SHAP values as feature importance values was deemed more suitable than the built-in Gini importance values of the RF algorithm.

While the proposed method effectively combines multicollinear variables using SHAP values, it does not directly eliminate the challenges of multicollinearity. As discussed in Chapter 4.2.6, in the case

independent variables are correlated, as in this approach, their marginal contributions overlap, leading to ambiguity in assigning their individual effects (Mishra, 2016). Consequently, this can lead to misleading SHAP values because the algorithm can not distinguish individual effects, leading to over- or underestimation of scores. Therefore, careful interpretation of the SHAP values is necessary. However, in the context of this research, where the primary goal is to combine variables for use in MLR, achieving a perfect importance value is not essential. Instead, the focus is on creating a representative combined variable that reflects the shared contribution of correlated variables while maintaining a simplified and interpretable model structure. The computationally intensive solution in Chapter 4.2.6 is unnecessary here, as the goal is to create a representative combined variable capturing the shared contribution of correlated variables rather than achieving perfect variable importance.

Integrating SHAP values into this methodology provides a balanced trade-off between simplifying the model and maintaining interpretability. Despite its limitations, the method provides an option for more interpretable and consistent regression models in multicollinear environments while incorporating all multicollinear variables.

4.3 Statistical Two-way Analysis of Variance

Given seasonal data, a seasonal pattern is a pattern of ups or downs over seasons (Sclove & Wang, 2014). If the pattern is consistent over the years, it is called a stable seasonal pattern. Proposed by Sclove and Wang (2014), we adopt a two-way ANOVA approach to complement the regression methods. This approach helps detect stable seasonal patterns.

In a two-way ANOVA setup, the model considers two factors:

- Period effects represent the typical variations within each period across all years. For instance, Q4 may consistently show higher values compared to Q1;
- Yearly effects capture the variations across different years, providing insight into annual trends.

The total variability in the data is broken down into:

- Between-period variability, which accounts for differences between the means of each period;
- Between-year variability, which accounts for differences across years.

By analysing these components, ANOVA helps determine if the observed seasonal effects are stable. Stability in this context means the effects of specific periods are consistent from year to year. A stable seasonal effect is evaluated based on the *P*-value with a threshold of 0.05.

4.4 Validation Metrics

4.4.1 Coefficient of Determination

The coefficient of determination, or R^2, indicates the proportion of variation in the dependent variable that can be explained or predicted by the independent variable(s) (Chicco, Warrens, & Jurman, 2021). R^2 values range between zero and one, where higher values indicate that the model better explains the variability in the dependent variable. An R^2 value of one means a perfect explanation or prediction. In contrast, an R^2 value of zero indicates that the model indicates that the model does not capture any of the variability in the response data relative to the mean. Equation (16) expresses the coefficient of determination (Chicco, Warrens, & Jurman, 2021).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - Y_i)^2} \quad (16)$$

Where:

- X_i denotes the predicted i^{th} value;
- Y_i denotes the actual i^{th} value;
- n denotes the number of data points;
- and $\bar{Y}$ denotes the mean of the actual values. Equation (17) expresses the calculation of the mean of the actual values.

$$\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i \quad (17)$$

4.4.2 Mean Squared Error

The Mean Squared Error (MSE) measures the accuracy of a model by averaging the squared deviations between predicted values and actual observations (Chicco, Warrens, & Jurman, 2021). Suppose the model's MSE includes a single bad prediction. In that case, the squaring of the error increases its impact, making it more influential on the overall metric. The values of MSE range from zero to infinity, with zero representing the best outcome and higher values indicating worse model performance. Equation (18) expresses the MSE (Chicco, Warrens, & Jurman, 2021).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \quad (18)$$

Where:

- X_i denotes the predicted i^{th} value;
- Y_i denotes the actual i^{th} value;
- n denotes the number of data points.

4.4.3 Mean Absolute Error

Mean Absolute Error (MAE) reflects the mean of the absolute deviations between the predicted and actual values, indicating the model's accuracy (Chicco, Warrens, & Jurman, 2021). MAE is less sensitive to outliers than metrics like MSE. MAE provides a more generalised performance measure that smooths out the impact of outliers. The values of MSE range from zero to infinity, with zero representing the best outcome and higher values indicating worse model performance. Equation (19) expresses the MAE (Chicco, Warrens, & Jurman, 2021).

$$MAE = \frac{1}{m}\sum_{i=1}^{n}|X_i - Y_i| \quad (19)$$

Where:

- X_i denotes the predicted i^{th} value;
- Y_i deotes the actual i^{th} value;
- n denotes the number of data points.

4.5 Model

Figure 4-4 provides a flowchart illustrating the sequence of steps and decision points in the model, offering a visual overview of its structure and processes. All the theories discussed in this chapter are incorporated into this model.

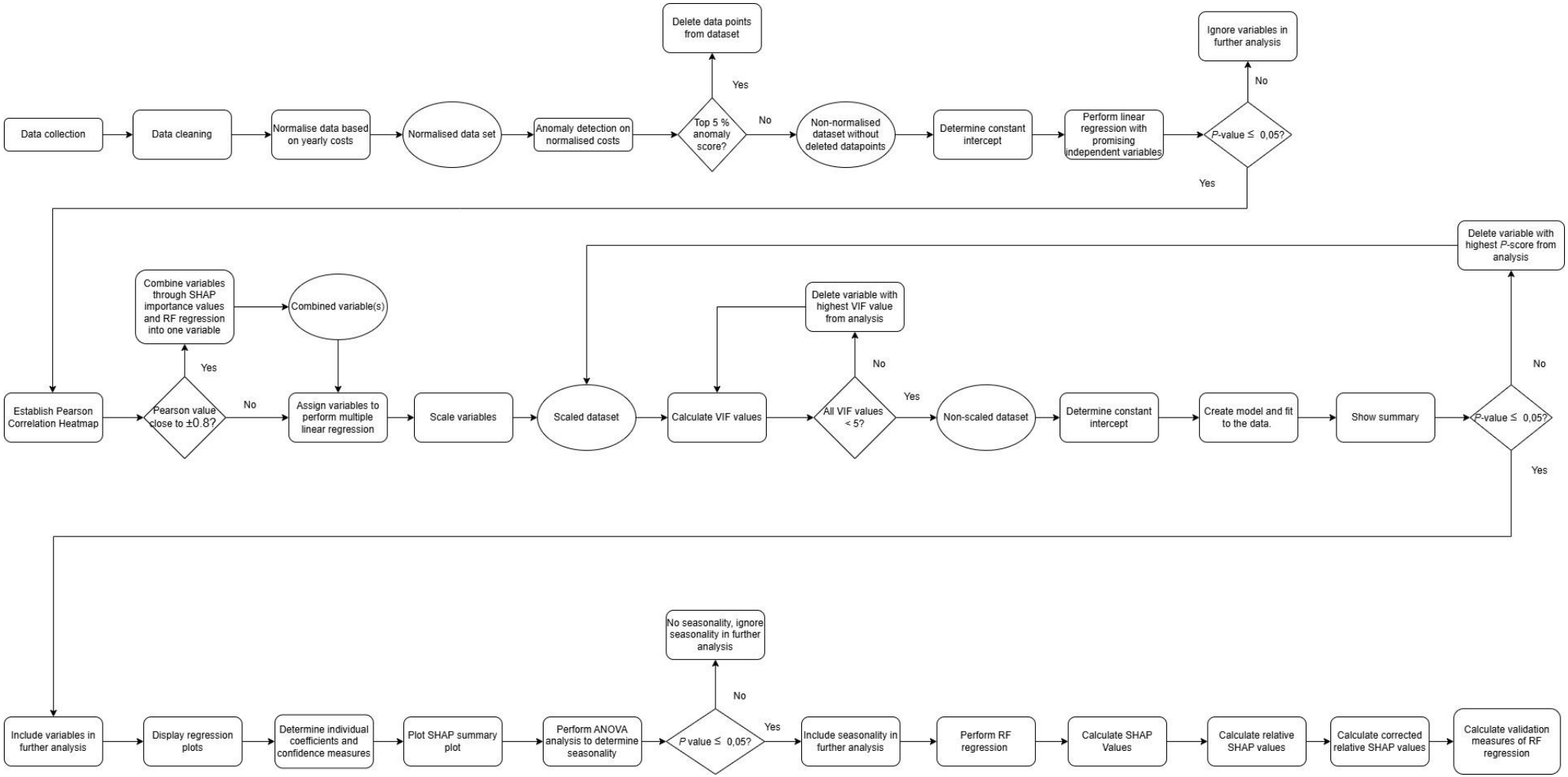

Figure 4-5 Model flowchart

4.6 Chapter Conclusion

Following the CRISP-ML Framework, this chapter outlined the methodology employed to address the main research question. The chosen approach integrates iForest for anomaly detection, MLR for identifying cause-and-effect relationships, and RF regression for handling complex patterns in the data. The chapter outlined how variables will be combined using SHAP values to address multicollinearity and how adjusted SHAP values will be calculated to ensure accurate results. Additionally, ANOVA was explained to evaluate stable seasonal effects in contract costs. The chapter also detailed the validation metrics used to validate the RF regression.

The methods presented in this chapter provide a structured foundation for analysing the data and answering the research questions. The following chapter presents the results obtained from these methods.

5. Analysis, Results, and Discussion

This chapter presents the results obtained from the methodology described in the previous chapter. The results are structured to reflect the sequential approach of the methodology and provide a clear progression from data preparation to key findings.

The chapter begins by describing the data collection and cleaning steps, outlining the processes undertaken to ensure the dataset's quality. Next, an overview of the Python libraries used is provided.

Following this, the results are systematically presented and discussed. The outcomes from each method, including anomaly detection, regression analyses, and SHAP-based evaluations, are explained in detail. The chapter concludes by summarising the key findings.

5.1 Data Collection

This study collected data from several departments. The central departments from which this study collected data are the Business Intelligence Department and the Contracts and Hospitality Department of MST. The data was aggregated monthly from January 2019 until July 2024, which was the data range available at MST. The data was made available through CSV files.

On top of the data delivered by MST, the influenza data was collected from the Dutch National Institute for Public Health website (National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport, 2019, 2020, 2024). The influenza data represents the number of detections per 100,000 inhabitants in the Netherlands. The influenza data was missing from May 2020 until September 2020 due to the COVID-19 pandemic. During this period, COVID-19 restrictions significantly curtailed the circulation of respiratory viruses, making other respiratory infections negligible (National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport, 2021).

Given these circumstances, we assumed zero influenza cases for the missing period. While this assumption could theoretically underestimate influenza variability if the virus circulated undetected, such a scenario is unlikely given the limited transmission of respiratory viruses during that time. Nevertheless, if this assumption introduces underestimation, it could marginally skew the identified seasonal patterns, potentially reducing the analysis's accuracy.

We aggregated the data monthly, whereas the original influenza data was aggregated weekly. While the weekly influenza data was summed to match the monthly format, it may not perfectly reflect the precise distribution of cases across each month.

5.2 Data Cleaning

The research objective is to identify variables influencing contract costs over the time series of data availability. Regression analysis was employed to uncover patterns and relationships within the data to achieve the objective.

We have deleted cost posts to avoid disrupting the analysed pattern of laundry contract costs. Firstly, the plannable curtain washing, which occurs once every three months, has been removed from the laundry contract. Plannable curtain washing has significant costs and breaks the pattern because it occurs once every three months and is not affected by any external variables. Secondly, we excluded isolation jackets from the laundry contract since these jackets were introduced in 2020. Due to this later introduction, the data would cause a deviation from the pattern.

The data format in this phase was set as an XLSX file, with data aggregated monthly.

5.3 Python Libraries

The model was coded in Python 3.12.4 for both contracts using the following libraries' default settings:

- **Pandas** was used for data manipulations, providing data structures like DataFrames;
- **Numpy** was used for numerical operations, such as performing mathematical functions;
- **Matplotlib.pyplot** was used for creating visualisations;
- **Seaborn** is based on Matplotlib and was used for statistical graphics;
- **Statsmodels.API** was used for statistical models, including linear regression and multiple linear regression.
- **Sklearn.tree** was used for visualising decision trees;
- **SHAP** was used for XAI helping to interpret model predictions;
- **Scirpy** provided tools for statistical functions and tests;
- **Shap.initjs()** initialised JavaScript for SHAP visualisations;
- **Sklearn.preprocessing (StandardScaler, MinMaxScaler)** was used for data scaling and normalisation;
- **Sklearn.ensemble (RandomForestRegressor, IsolationForest)** was used to build ML models, such as the RF for regression and the IForest for anomaly detection;
- **Sklearn.metrics (r2_score, mean_squared_error, mean_absolute_error)** provided metrics for evaluating model performance.
- **statsmodels.stats.anova (anova lm)** was used for two-way ANOVA.

5.4 Results for Laundry Contract

The following subsections present a detailed, step-by-step explanation of the model's components and functionality integrated with the results and a discussion of the method of the laundry contract.

Understanding the input data and the pattern of the input data is essential before examining results and selecting independent variables as predictors of the dependent variable. The input data ranges from January 2019 until July 2024, meaning there are 67 data points to be analysed. In this report, all cost values are normalised on a scale from zero to one to reveal patterns while maintaining anonymity. Figure 5-1 shows a bar chart of the normalised monthly costs of the laundry contract. The chart highlights noticeable variations across months, with some months exhibiting significantly higher or lower average costs. For instance, Month eight stands out with markedly lower costs. Despite these month-to-month fluctuations, the costs remain relatively consistent, as most months have normalised values ranging between 0.4 and 0.5.

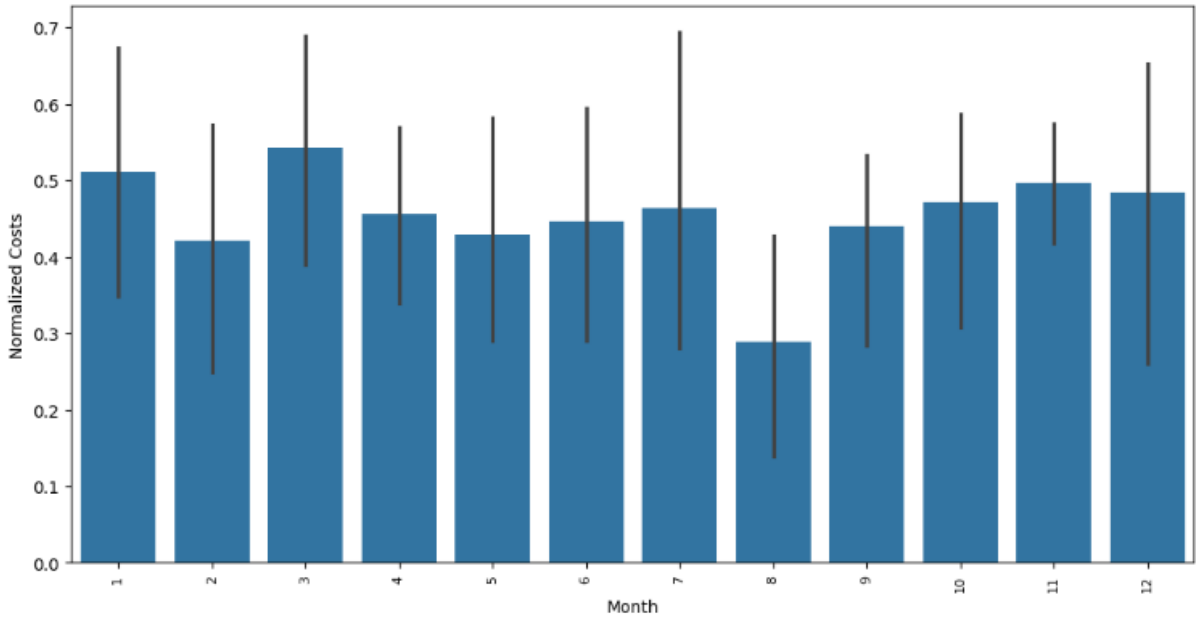

Figure 5-1 Average monthly normalised costs laundry contract

Appendix C presents the observed patterns of the independent variables analysed in the model.

5.4.1 Anomaly Detection

Anomaly detection and deletion were applied to eliminate the dataset's top 5% anomalies.

The five per cent rate was chosen due to an expectation of a low anomaly count, balancing the model's sensitivity with a reduced risk of false positives. Given the dataset's size, a higher contamination rate to avoid overfitting is unnecessary, as the dataset is not large enough to benefit from higher contamination. Testing confirmed that this assumption—expecting few anomalies—was correct, with observed anomalies aligning well with this assumption.

Table 5-1 presents the three anomalies found within the laundry contract. Figure 5-2 depicts anomaly detection in the costs of the laundry contract. The blue line represents the normalised costs, and the red crosses depict the found anomalies based on a five per cent contamination rate. Three anomalies were identified, corresponding to November 2019, December 2020, and July 2024. These anomalies show significantly higher costs than the normalised trend across other months.

Month	Year	Anomaly score
11	2019	-0.214
12	2020	-0.173
7	2024	-0.187

Table 5-1 Anomalies laundry costs (5%)

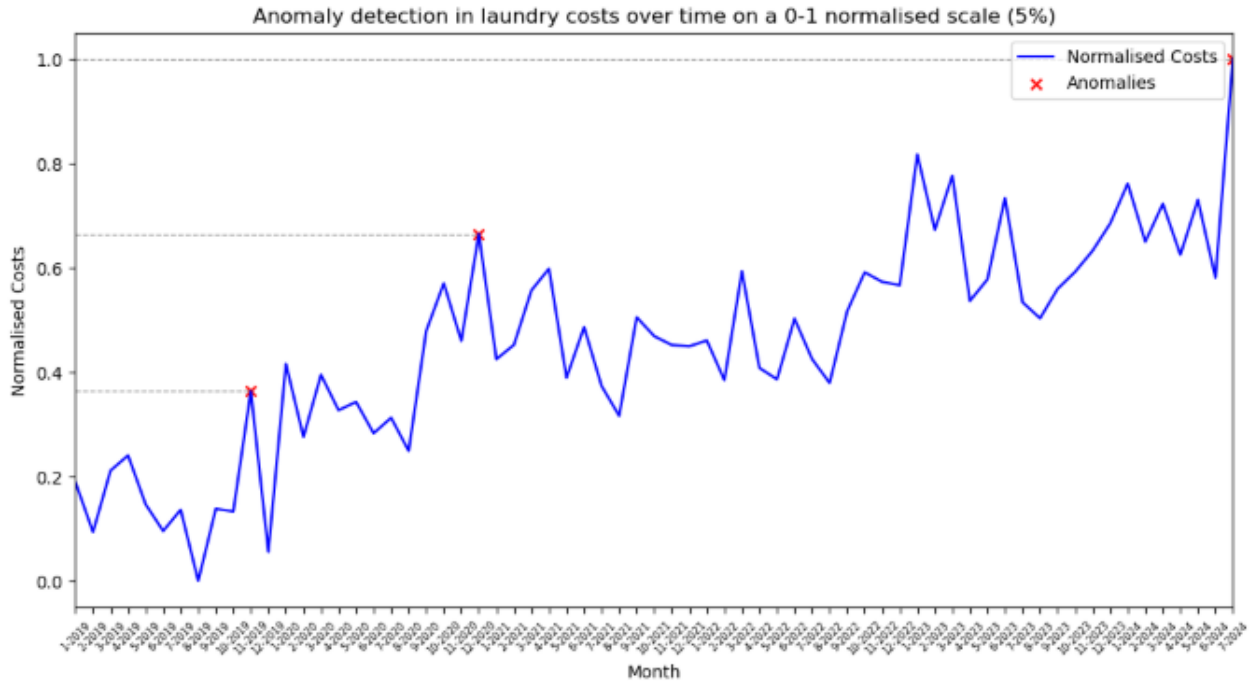

Figure 5-2 Anomaly detection laundry costs

5.4.2 Linear Regression

In this analysis, linear regression was conducted on the dependent variable, costs, to examine the impact of each promising independent variable. Variables below the *P*-value threshold were retained for the MLR. A constant was added to the model before running the linear regression.

Table 5-2 shows the linear regression results for the laundry contract.

Independent variable	**P>|t|**
Dutch Influenza cases	$1.62*10^{-11}$
Number of day hospitalisations	$9.11*10^{-40}$
Number of hospital employee shifts (eight-hour shifts)	$1.21*10^{-63}$
Number of visits to the outpatient clinic	$6.70*10^{-56}$
Number of DTCs	$7.08*10^{-59}$
Number of clinical admissions	$4.48*10^{-56}$
Number of operations	$7.18*10^{-50}$
Number of intensive care admissions	$1.83*10^{-36}$

Table 5-2 Regression results laundry contract

All analysed variables have a *P*-value lower than 0.05. Consequently, each of these independent variables were included in further analysis.

The consistently low *p*-values across both contracts' regression models confirm statistically significant individual correlations between each independent variable and costs. While the statistical significance of each variable is established, it is essential to consider the practical relevance of this result. Firstly, while the regression results indicated correlations between the dependent and independent variables, reflecting on the theory behind linear regression is essential. Linear regression models are designed to identify correlations between two variables; a detailed explanation of the method and its assumptions can be found in Section 4.2.3. This design means that linear regression can sometimes detect

correlations that may not be meaningful or true if certain factors are affected by each other or external factors.

In a hospital environment where variables may affect each other, it is crucial to evaluate their effects collectively through a complex regression model where multiple factors may influence each other. Interactions between these variables could lead to different and more trustworthy results because the interaction between variables is taken into account. By considering interactions, the model can produce more reliable results.

We addressed this issue by implementing an MLR model, which allows for the simultaneous assessment of multiple variables' effects on the dependent variable. Furthermore, multicollinearity is evaluated to account for intercorrelating effects.

5.4.3 Pearson Correlation Matrix

A Pearson correlation matrix was established to determine the multicollinearity of individual independent variables. If Pearson's value is close to ±0.8, multicollinearity exists between the independent variables.

The analysed variables are the same for both contracts. Figure 5-3 shows the Pearson correlation matrix for both contracts.

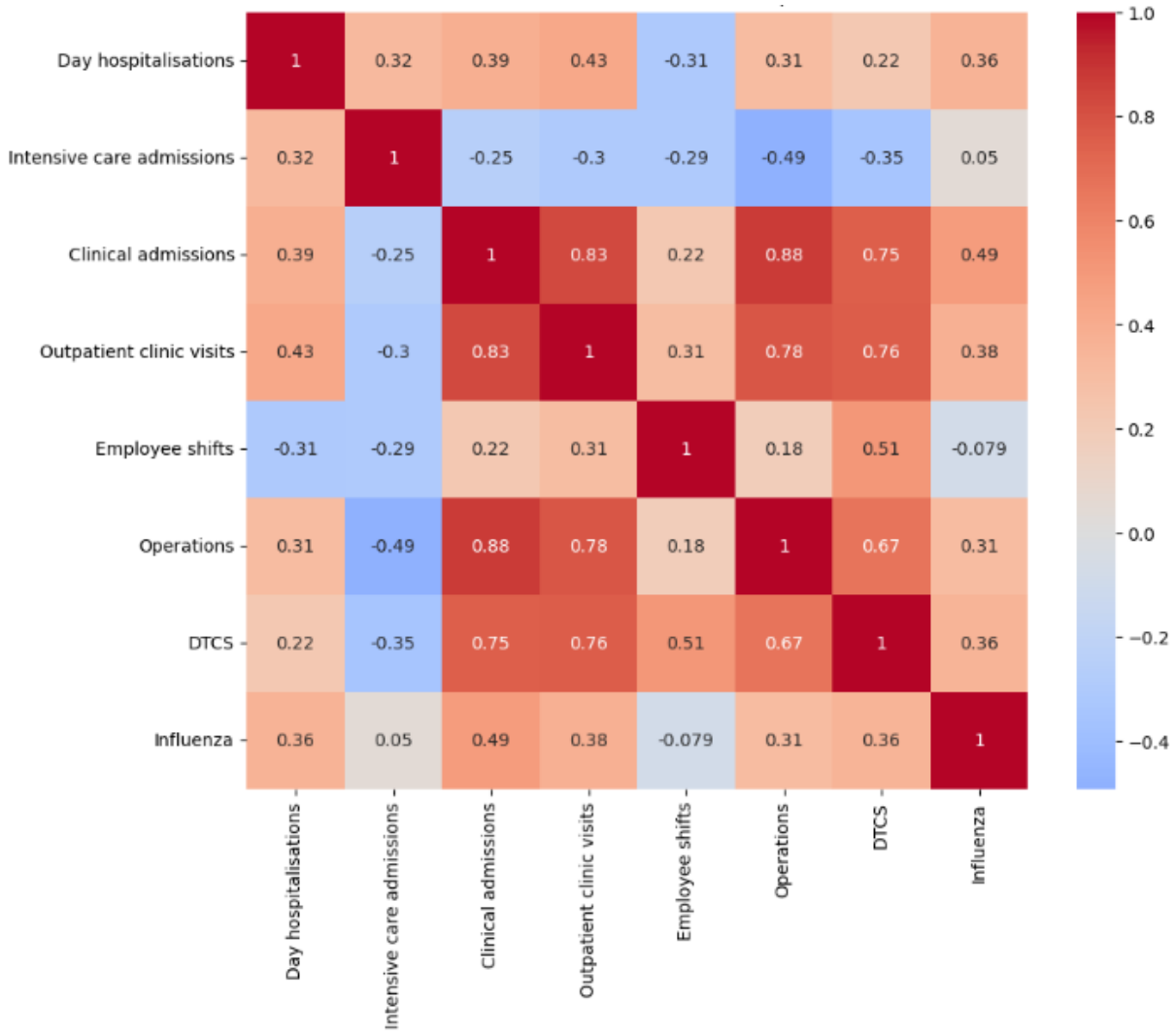

Figure 5-3 Pearson correlation matrix

One limitation of the Pearson correlation matrix is that it evaluates only linear correlation, which is sufficient since MLR assumes that the relationship between each variable is linear. Consequently, only evaluating linear correlations addresses multicollinearity sufficiently in this linear context.

Some variables have a Pearson's correlation value close to 0.5, indicating the presence of multicollinearity. However, it is not substantial enough to warrant combining the variables. A VIF evaluation was conducted later to assess multicollinearity and prevent it from being present when performing the MLR.

This Pearson correlation matrix shows the number of day hospitalisations, number of visits to the outpatient clinic, number of DTCs, and number of operations to be intercorrelated. Consequently, these four independent variables were combined into one independent normalised variable using relative SHAP values.

Table 5-3 shows the relative SHAP importance sorted from highest to lowest for the laundry contract.

Independent Variable	Relative Impact
Number of DTCs	49.89%
Number of visits to the outpatient clinic	21.21%
Number of clinical admissions	16.01%
Number of operations	12.89%

Table 5-3 Relative SHAP importance laundry contract

Figure 5-4 shows the pattern of the combined variable, showing the normalised combined costs derived from the correlated variables, later addressed as “Combined Variable“.

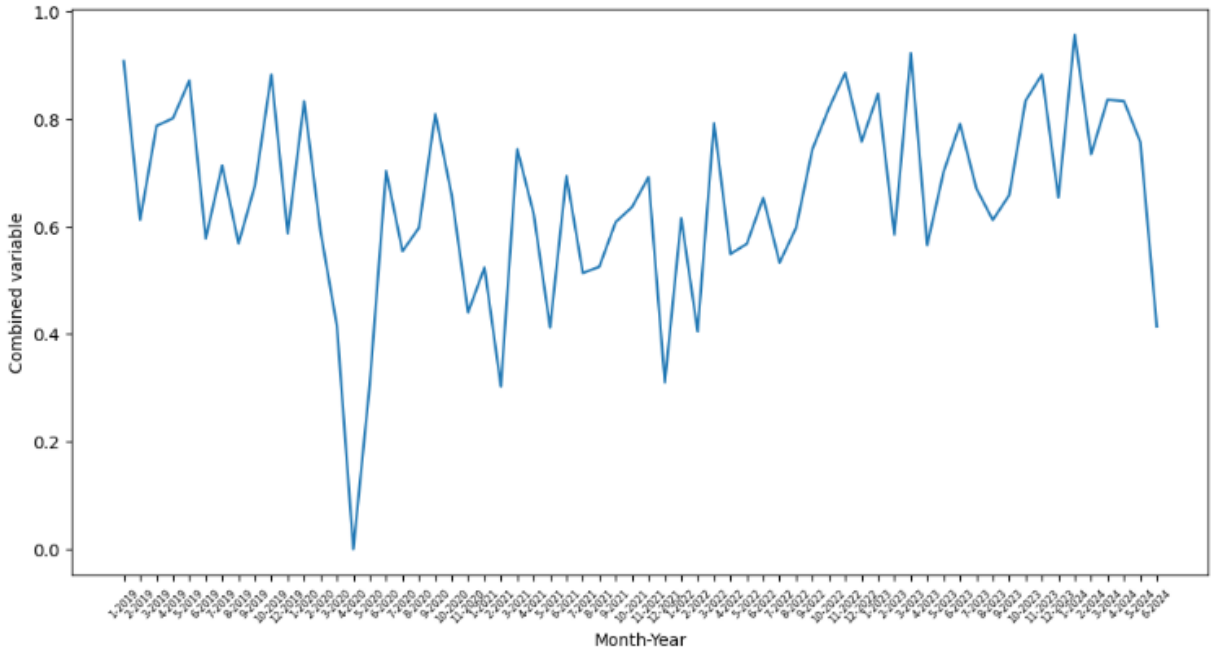

Figure 5-4 Combined variable’s pattern laundry contract

5.4.4 Variance Inflation Factor and Multiple Linear Regression

Since the calculation of the VIF value and MLR occur within a single recurring cycle, their results are merged in this chapter.

VIF does not capture the non-linear correlation between variables, which could distort the MLR. Furthermore, the cut-off point to accept or reject a VIF value is not universal. Common cut-off values are five and ten, but these numbers are not absolute. In this research, a VIF cut-off of five was chosen to identify and minimise multicollinearity. Some studies see this cut-off as low, potentially excluding some influential variables. However, by adopting this cut-off, we prioritised model clarity while acknowledging that flexibility in VIF interpretation might be necessary in different research contexts.

MLR provides a straightforward means of understanding the linear relationship between independent variables and the dependent variable. However, MLR also has limitations. The model assumes a linear relationship, which may not always align with input data. Using two-way ANOVA, seasonality was evaluated to replenish the linear model. However, other non-linear relationships were not evaluated. Future research could use more complex methods to capture and identify non-linear relationships.

Furthermore, a limitation of MLR is that outliers can significantly skew the results. Anomaly deletion was performed to minimise the impact of outliers on the dependent variable. However, no similar anomaly deletion was conducted for the independent variables, which could still affect the model's

accuracy. While the decision to leave the independent variables untouched was made to retain as much data as possible, this approach could introduce noise and reduce the overall robustness of the model.

Despite these limitations, MLR suits this research because of its explanatory power. By assessing the significance of each variable with a *P*-value threshold, this method selects independent variables that contribute to contract costs. Nevertheless, if the assumptions outlined in Chapter Four are violated, the reliability of the model's outcome is violated.

Table 5-4 presents the result of the first cycle. All VIF values are below the accepted threshold of five, so the *P*-value was calculated. The number of intensive care admissions shows the highest *P*-value, much higher than 0.05, showing weak to no correlation. Consequently, it was excluded from further analysis.

Independent variable	**VIF Value**	**P>|t| (*P*-value)**
Constant	N.A.	0.000
Dutch Influenza cases	1.38	0.187
Number of day hospitalisations	1.05	0.001
Number of hospital employee shifts (eight-hour shifts)	1.76	0.062
Number of intensive care admissions	1.62	0.738
Combined variable	2.71	0.004

Table 5-4 Cycle one VIF and MLR laundry contract

Table 5-5 presents the result of the second cycle. All VIF values are below the accepted threshold of five, so the *P*-value is calculated. The number of Dutch Influenza cases has the highest *P*-value and is higher than 0.05. Although this low *P*-value indicates some correlation, it is not statistically significant. Consequently, it was excluded from further analysis.

Figure 5-5 presents the SHAP summary plot, where a colour gradient indicates variable levels: red represents high values, while blue represents low values. As shown in Figure 5-4, the high values of the Dutch Influenza cases significantly affect costs. In contrast, the low number has a much smaller effect. One can question whether an MLR, where the continuous value of Dutch Influenza cases is compared to costs, is the correct method. This question arises because MLR assumes a linear relationship between variables, which may not hold if the effect of influenza on costs is non-linear or if other factors influence the costs. A more appropriate approach could be to assess whether the Dutch Influenza season consistently affects costs. Therefore, two-way ANOVA was used to determine whether the Influenza season statistically impacts laundry costs, as it can more effectively test for group differences without assuming a linear relationship. This evaluation was performed for both researched contracts. This variable is later addressed as "Influenza season".

Independent variable	**VIF Value**	**P>|t| (*P*-value)**
Constant	N.A.	0.000
Dutch Influenza cases	1.35	0.164
Number of day hospitalisations	1.57	0.000
Number of hospital employee shifts (eight-hour shifts)	1.72	0.051
Combined variable	1.96	0.000

Table 5-5 Cycle two VIF and MLR laundry contract

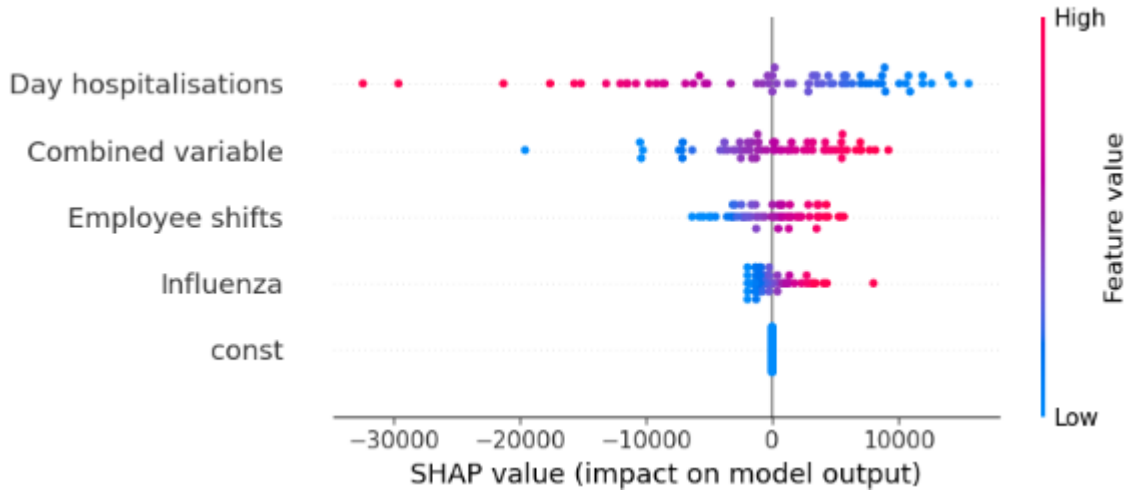

Figure 5-5 SHAP summary plot laundry contract MLR

Table 5-6 presents the result of the third cycle. All VIF values are below the accepted threshold of five, so the *P*-value was calculated. The *P*-value for the number of hospital employee shifts did not meet the significance threshold of 0.05 and was therefore excluded from further analysis.

Independent variable	**VIF Value**	**P>|t| (*P*-value)**
Constant	N.A.	0.000
Number of day hospitalisations	1.54	0.000
Number of hospital employee shifts (eight-hour shifts)	1.65	0.089
Combined variable	1.68	0.000

Table 5-6 Cycle three VIF and MLR laundry contract

Table 5-7 presents the result of the fourth cycle. All VIF values are below the accepted threshold of five, so the *P*-value was calculated. All *P*-values equal 0.000, so this combination has a statistically significant correlation. Therefore, this combination of continuous variables significantly influences contract costs. The combined variable shows a very high coefficient, which results from the normalisation of the variable on a scale from zero to one. The scale of this variable is significantly smaller compared to the other variables. The independent variables' coefficients were calculated in a later step.

Independent variable	**VIF Value**	**P>|t| (*P*-value)**	**Coefficient**	**95% Confidence Interval**	**Standard Error**
Constant	N.A.	0.000	N.A.	N.A.	N.A.
Number of day hospitalisations	1.12	0.000	-37.69	[-45.91, -29.46]	4.11
Combined variable	1.12	0.000	$4.32*10^{04}$	[$2.92*10^{04}$, 5.72^{04}]	6994.57

Table 5-7 Cycle four VIF and MLR laundry contract

The partial regression plots in Figure 5-6 show the adjusted relationships between the independent and dependent variables while accounting for other predictors' influence. Each plot features blue data points representing individual adjusted data values for the variable of interest after controlling for the effects of the other variables in the regression model. The trend line in each plot indicates the direction of the relationship.

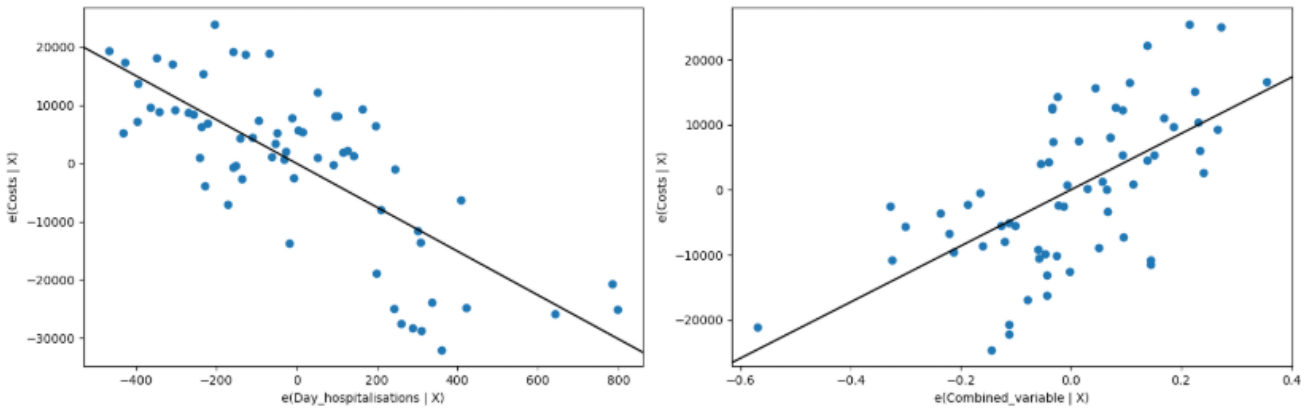

Figure 5-6 Regression plots laundry contract

5.4.5 Calculating the Individual Coefficient of Combined Variable

The individual coefficients of the combined variables were calculated after determining whether the combined variable had a statistically significant effect on contract costs.

This approach, however, has a drawback. Combining the correlated variables results in the loss of their patterns. While the variables are strongly correlated, they are not perfectly collinear, leading to a different overall pattern when combined. This pattern allows for the determination of correlation between the variables. However, the individual coefficients of the combined variables do not reflect the exact coefficients of the original variables. These coefficients provide a general indication of the direction and magnitude of the effect of each variable within the combination.

Table 5-8 shows the calculation of the individual coefficients.

Variable	**Coefficient**	**95% Confidence Interval**	**Standard Error**
Number of DTCs	2.03	[1.05, 3.00]	0.48
Number of visits to the outpatient clinic	0.49	[0.26, 0.73]	0.11
Number of clinical admissions	6.24	[3.24, 9.26]	1.50
Number of operations	4.36	[2.26, 6.46]	1.05

Table 5-8 Recalculated coefficients combined variable laundry contract

5.4.6 Two-way Analysis of Variance

Two-way ANOVA determined seasonality after determining the statistically significant continuous variables affecting contract costs. Figure 5-7 shows the seasonal pattern, with Season 1 representing January to March, Season 2 April to June, Season 3 July to September, and Season 4 October to December. There is a noticeable seasonal difference, with season one showing the highest average costs and season three the lowest average costs.

Additionally, two-way ANOVA was used to evaluate whether a consistent pattern exists during the influenza season. Whether a month can be labelled as influenza season was based on a threshold of 232 monthly influenza cases per 100000 inhabitants. This threshold was determined based on the weekly threshold of 58 influenza cases per 100000 inhabitants, as defined by the Dutch National Institute for Public Health and the Environment (National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport, 2024). By multiplying the weekly threshold by four, corresponding to the approximate number of weeks in a month, 232 cases per month were established

for identifying influenza season months. This aggregation resulted in eight months of influenza season in the past six years. Excluding 2020 because of the COVID-19 pandemic equates to an average of two annual influenza season months.

This method enabled categorising months as part of the influenza season to align aggression with the dataset. However, it does not directly reflect the measurement methodology traditionally used in the Netherlands, typically based on weekly influenza cases rather than monthly cases. This adaptation was necessary to align the data with the monthly aggregation of the available dataset. However, it introduces a potential discrepancy between this research's approach and standard measuring methods.

Table 5-9 presents the two-way ANOVA test results evaluating a consistent seasonal and influenza season effect for the laundry contract over the years.

While the results for both contracts indicate a consistent effect of the influenza season over the years, one could argue that a test of eight months of influenza seasonality does not provide sufficient evidence to establish a robust relationship. Moreover, within the six-year data range, two years did not include an influenza season. Nevertheless, the influenza season variable will be included in further analysis to assess the magnitude of its effect on contract cost.

Independent variable	**P>|t| (*P*-value)**
Season	$6.103323*10^{-03}$
Influenza season	$1.913611*10^{-02}$

Table 5-9 Two-way ANOVA results laundry contract

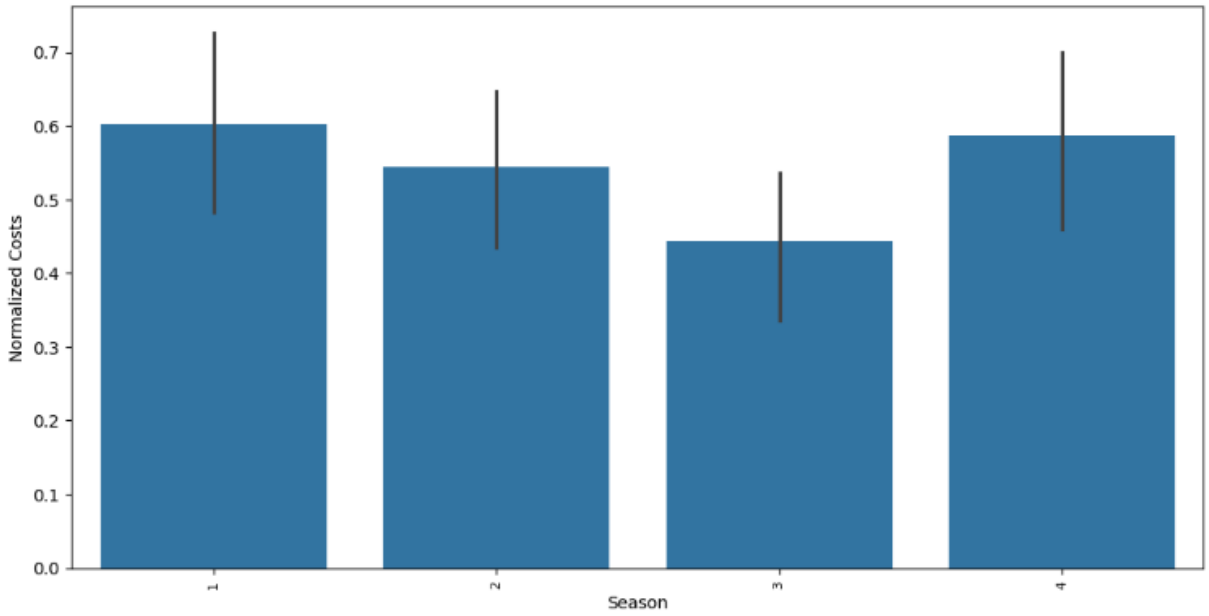

Figure 5-7 Seasonal laundry costs normalised

5.4.7 Shapley Additive Explanations Importance

After determining all significant variables that affect contract costs, the importance and effect of each variable were determined by combining an RF regression with SHAP analysis.

Table 5-10 provides each variable's relative impact and corrected relative impact on the model's output. Additionally, the table provides the variables' direction of influence concluded from Figure 5-

8, with a colour gradient indicating variable levels: red represents high values, while blue represents low values. A negative direction of influence implies an inverse relationship, meaning that as the value of the variable increases, the associated costs decrease. Conversely, a positive direction of influence signifies a direct relationship, where an increase in the variable value leads to higher costs. A neutral direction of influence indicates that a variable overall impact is balanced and not strongly directional. The direction of influence was only concluded for continuous variables since categorical variables do not possess a direction of influence.

Table 5-10 shows a clear difference between the relative impact of SHAP values and the relative impact of the corrected SHAP values. When looking at the theory of SHAP and the theory of the corrected SHAP values, this difference was expected. The non-corrected SHAP values can over- and underestimate certain variables due to multicollinearity, which is present in this dataset. This effect is visible, with one dominating variable, while the corrected SHAP values are much more balanced. Therefore, the corrected SHAP values provide a more accurate measure of variable impact in this context. Consequently, the corrected relative impact was concluded as the appropriate measure for interpreting the drivers of laundry costs in this analysis.

We discussed these results with the subdepartment that manages and is seen as an expert on the subject. The main explanation for the results was that day hospitalisations require fewer laundry costs than the other variables except for the number of DTCs. The rising number of DTCs indicates that the hospital is busier, leading to more laundry costs. Furthermore, the subdepartment recognised the seasonal variables. Overall, the results seemed well-founded to the subdepartment.

Variable	**Relative Impact**	**Corrected relative impact**	**Direction of influence**
Number of day hospitalisations	50.15%	8.33%	Negative
Number of operations	23.85%	17.51%	Positive
Number of visits to the outpatient clinic	9.77%	8.61%	Positive
Number of DTCs	6.49%	29.63%	Positive
Number of clinical admissions	4.09%	34.15%	Positive
1st quarter	3.20%	0.42%	N.A.
3rd quarter	0.90%	0.28%	N.A.
4th quarter	0.89%	0.24%	N.A.
2nd quarter	0.57%	0.35%	N.A.
Influenza season	0.20%	0.46%	N.A.

Table 5-10 Relative impacts laundry contract

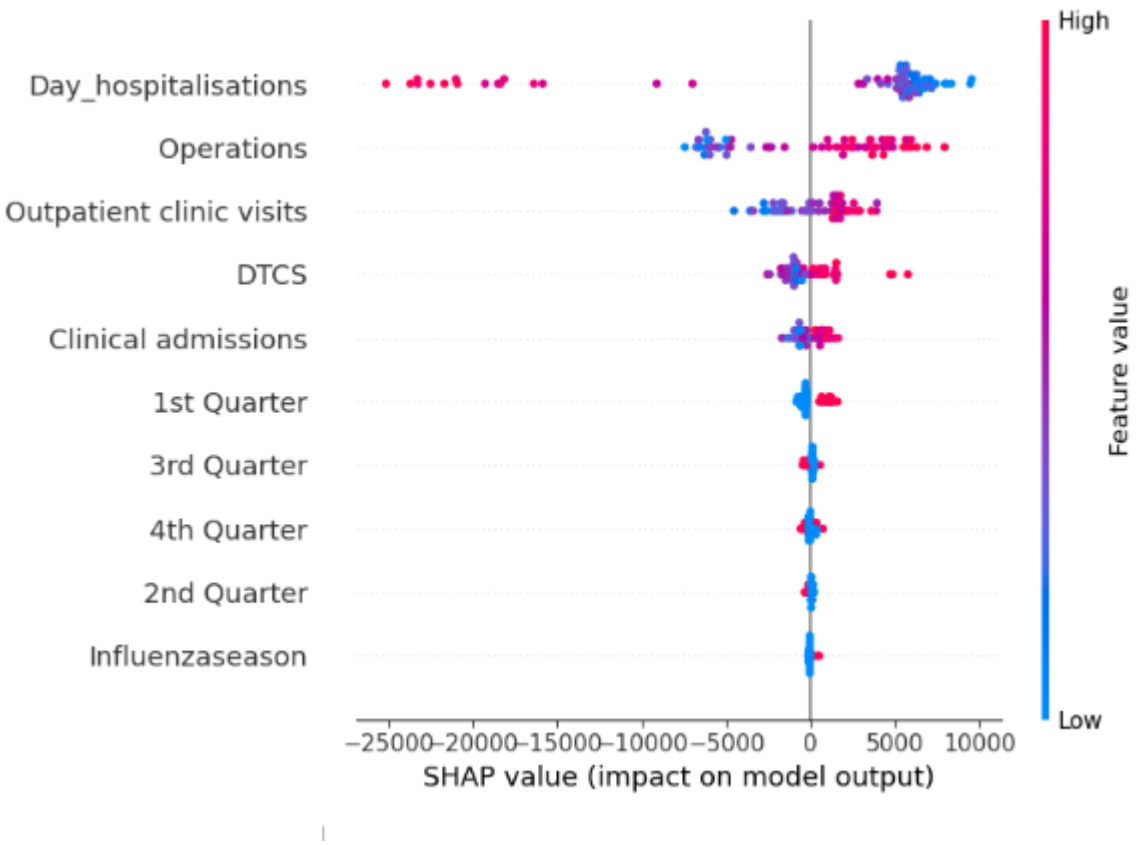

Figure 5-8 SHAP values laundry contract

5.4.8 Random Forest Prediction

The final step involved assessing the accuracy of the RF regression model's predictions. We used the data from January 2019 to July 2024 to train the RF regression model. Predictions were then made for the four months from August 2024 to November 2024.

Figure 5-9 illustrates the RF regression results for the analysed time series of the laundry contract, presented on a normalised scale ranging from zero to one for the actual values with the predicted values presented on the same scale. Additionally, Table 5-11 presents the validation metrics of the laundry contract on the same normalised scale. An R^2 of 0.5926 indicates that the predicted values explain approximately 59% of the variance in the actual values in the model. Furthermore, an MAE of 0.1939 is relatively low. Figure 5-9 shows the most significant deviation in October, where the significant peak was not predicted. These values show the variable's potential to be included in forecasting models.

However, the unexplained variance underscores the potential influence of other factors, such as economic trends and indexation adjustments, that were not included in the current model. Incorporating these variables in future iterations may improve predictive accuracy.

Furthermore, it is important to approach these results cautiously, as the predictions cover only four months. Predictions should be made over a longer time frame to obtain a more comprehensive assessment of the model's performance, capture more varied conditions, and validate the results further. These results have some limitations. The ratio between them is relatively high by utilising a limited dataset size (n=4) and a relatively extensive variable size (n=10 for the laundry contract and n=11 for the clinical chemistry contract). A high ratio between the data size and variables makes it harder for RF regression to distinguish relevant from relevant variables (Han & Kim, 2021). Additionally,

with limited data size, the diversity between trees is lower, decreasing the performance of RF regression. Additionally, using a small data size increases the risk of overfitting, which results in the regression's performance not generalising well to other data (Ying, 2019).

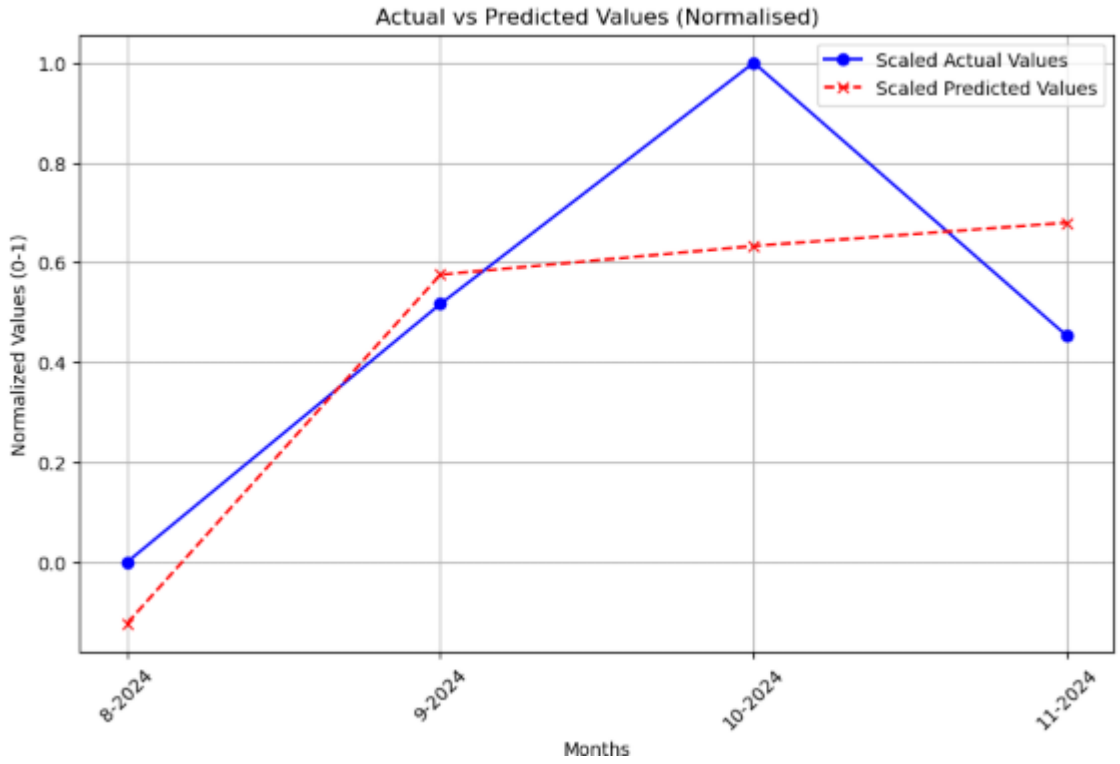

Figure 5-9 Comparison actual and prediction values laundry contract

Validation metric	**Outcome**
R^2 Score	0.5926
MSE	0.0511
MAE	0.1939

Table 5-11 Validation metrics laundry contract

5.5 Results for Clinical Chemistry Contract

This section presents the results of the Clinical Chemistry Contract. Figure 5-10 shows the monthly costs of the contract on a normalised scale in a bar chart. The chart shows significant variations, with Month one having the highest costs, around 0.7, and Month twelve the lowest, below 0.2. Costs in Month three are also relatively high, followed by an evident decline in the following months.

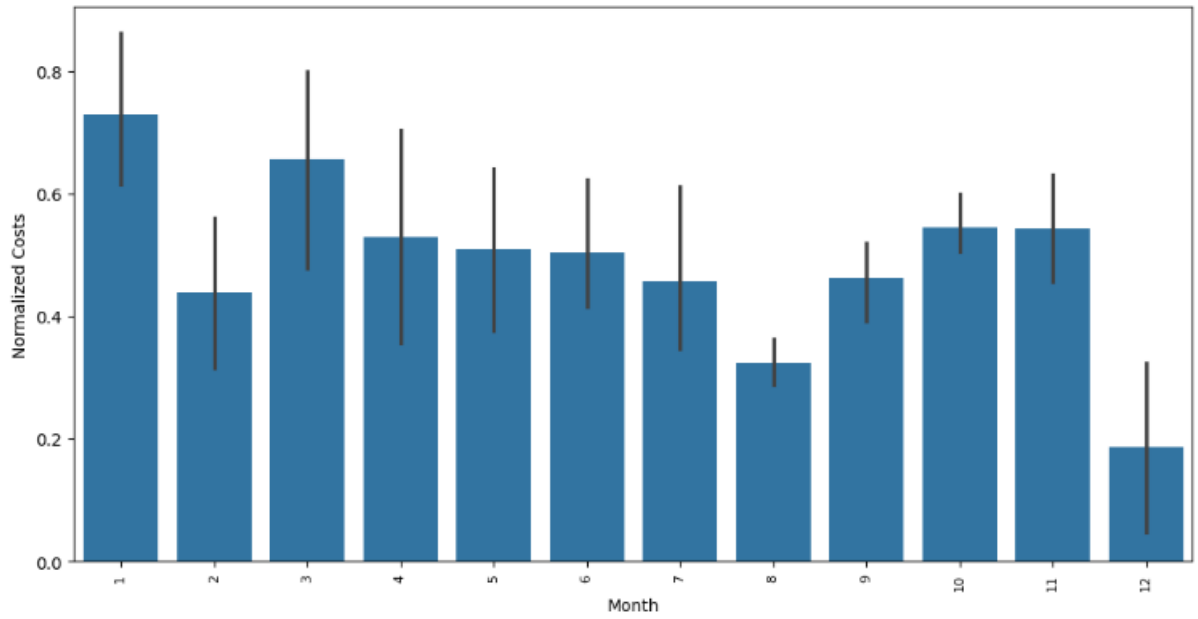

Figure 5-10 Normalised costs clinical chemistry contract

Appendix D presents the pattern of the analysed independent variables.

5.5.1 Anomaly Detection

Table 5-12 presents the three anomalies found within the laundry contract. Figure 5-11 presents the three anomalies over the time series, representing three instances where the costs were significantly lower than in other months in December 2019, 2021, and 2023. It is noteworthy that all three anomalies occur in December.

Month	Year	Anomaly score
12	2019	-0.252
12	2021	-0.197
12	2023	-0.229

Table 5-12 Anomalies clinical chemistry costs (5%)

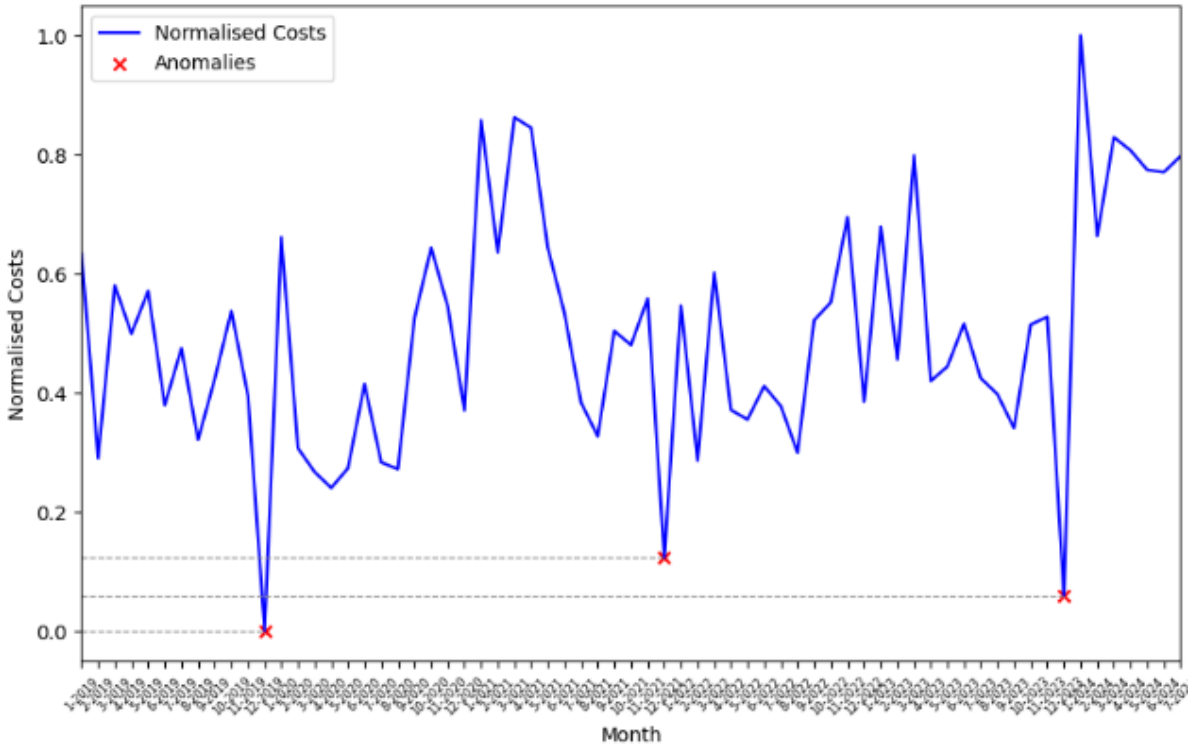

Figure 5-11 Anomaly detection clinical chemistry costs

5.5.2 Linear Regression

Table 5-12 shows the linear regression results for the laundry contract.

Independent variable	**P>|t| (*P*-value)**
Dutch Influenza cases	$1.81*10^{-11}$
Number of day hospitalisations	$7.81*10^{-50}$
Number of hospital employee shifts (eight-hour shifts)	$1.19*10^{-69}$
Number of visits to the outpatient clinic	$4.30*10^{-73}$
Number of DTCs	$9.52*10^{-78}$
Number of clinical admissions	$1.86*10^{-69}$
Number of operations	$1.94*10^{-56}$
Number of intensive care admissions	$7.17*10^{-44}$

Table 5-13 Regression results laundry contract

All analysed variables have a *P*-value lower than 0.05, indicating statistically significant relationships at the five per cent threshold. Consequently, each of these independent variables was included in further analysis.

5.5.3 Pearson Correlation Matrix

The analysed variables are the same for both contracts. The Pearson correlation matrix for the contracts is shown in Appendix C.

This Pearson correlation matrix shows the number of day hospitalisations, number of visits to the outpatient clinic, number of DTCs, and number of operations to be intercorrelated. Consequently, these four independent variables were combined into one independent normalised variable using relative SHAP values.

Table 5-14 shows the relative SHAP importance sorted from highest to lowest for the clinical chemistry contract.

Independent Variable	Relative Impact
Number of visits to the outpatient clinic	50.75%
Number of DTCs	34.21%
Number of operation	9.36%
Number of clinical admissions	5.66%

Table 5-14 Relative SHAP importance clinical chemistry contract

Figure 5-12 shows the combined variable's pattern, later addressed as "Combined Variable".

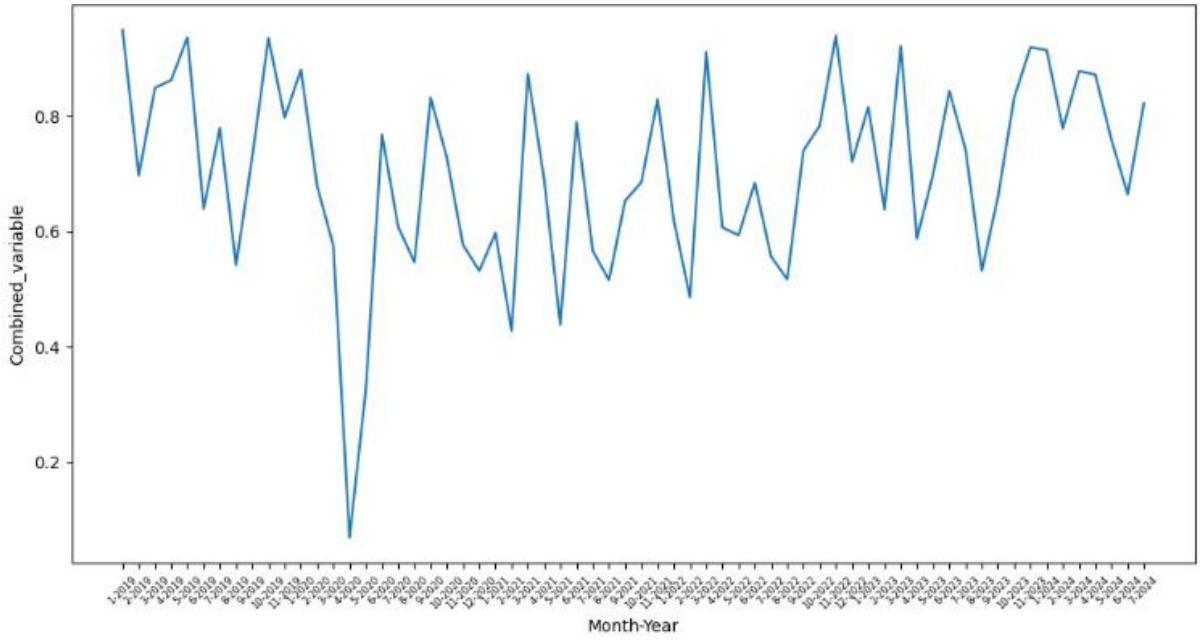

Figure 5-12 Combined variable's pattern clinical chemistry contract

5.5.4 Variance Inflation Factor and Multiple Linear Regression

Table 5-15 presents the result of the first cycle. All VIF values are below the accepted threshold of five, so the *P*-value was calculated. The number of Dutch Influenza cases has a very high *P*-value, indicating no linear correlation in this combination. Consequently, the variable was excluded from further analysis. However, as mentioned in Chapter 5.4.4, the Influenza seasonality was evaluated using two-way ANOVA.

Independent variable	VIF Value	P>|t| (*P*-value)
Constant	N.A.	0.000
Dutch influenza cases	1.29	0.951
Number of day hospitalisations	1.85	0.002
Number of hospital employee shifts (eight-hour shifts)	1.53	0.242
Number of intensive care admissions	1.55	0.014
Combined variable	1.91	0.000

Table 5-15 Cycle one VIF and MLR clinical chemistry contract

Table 5-16 presents the result of the second cycle. All VIF values are below the accepted threshold of five, so the *P*-value was calculated. The number of hospital employee shifts has a higher *P*-value than

0.05, indicating no statistically significant linear correlation in this combination. Consequently, the variable was excluded from further analysis.

Independent variable	VIF Value	P>|t| (*P*-value)
Constant	N.A.	0.000
Number of day hospitalisations	1.77	0.000
Number of hospital employee shifts (eight-hour shifts)	1.51	0.234
Number of intensive care admissions	1.54	0.013
Combined variable	1.75	0.000

Table 5-16 Cycle two VIF and MLR clinical chemistry contract

Table 5-17 presents the result of the third cycle. All VIF values are below the accepted threshold of five, so the *P*-value was calculated. All independent variables show a *P*-value below the threshold of 0.05, indicating a significant statistical correlation. Consequently, these continuous variables were concluded to affect contract costs.

Independent variable	VIF Value	P>|t| (*P*-value)	Coefficient	95% Confidence Interval	Std err
Constant	N.A.	0.000	N.A.	N.A.	N.A.
Number of day hospitalisations	1.49	0.002	-59.07	[-96.011, -22.121]	-22.121
Number of intensive care admissions	1.55	0.013	69.10	[15.42, 122.78]	26.84
Combined variable	1.38	0.000	$2.21*10^{05}$	[$1.62*10^{05}$, $2.80^{e}*10^{05}$]	$2.95*10^{04}$

Table 5-17 Cycle three VIF and MLR clinical chemistry contract

The partial regression plots in Figure 5-13 present the adjusted relationships between the independent and dependent variables while accounting for other independent variables' influence.

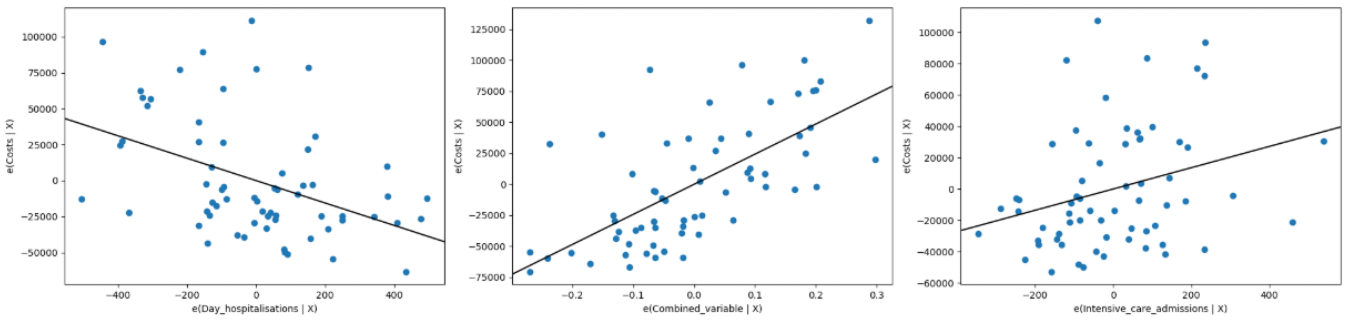

Figure 5-13 Partial regression plots clinical chemistry contract

5.5.5 Calculating the Individual Coefficient of Combined Variable

Table 5-18 shows the result of calculating the individual coefficients.

Variable	Coefficient	[0.025, 0.975]	Standard Error
Number of DTCs	8.79	[6.43, 6.43]	1.18
Number of visits to the outpatient clinic	7.28	[5.33, 9.23]	0.97
Number of clinical admissions	13.97	[10.23, 17.71]	1.87
Number of operations	20.04	[14.67, 25.40]	2.68

Table 5-18 Recalculated coefficients combined variable clinical chemistry contract

5.5.6 Two-way Analysis of Variance

Table 5-19 presents the two-way ANOVA test results evaluating a consistent seasonal and Influenza season effect for the clinical chemistry contract over the years. Figure 5-14 shows the pattern over the seasons, showing an evident decline from the first to the last.

Independent variable	P>|t| (*P*-value)
Season	$4.395055*10^{-03}$
Influenza season	$3.025114*10^{-02}$

Table 5-19 ANOVA results clinical chemistry contract

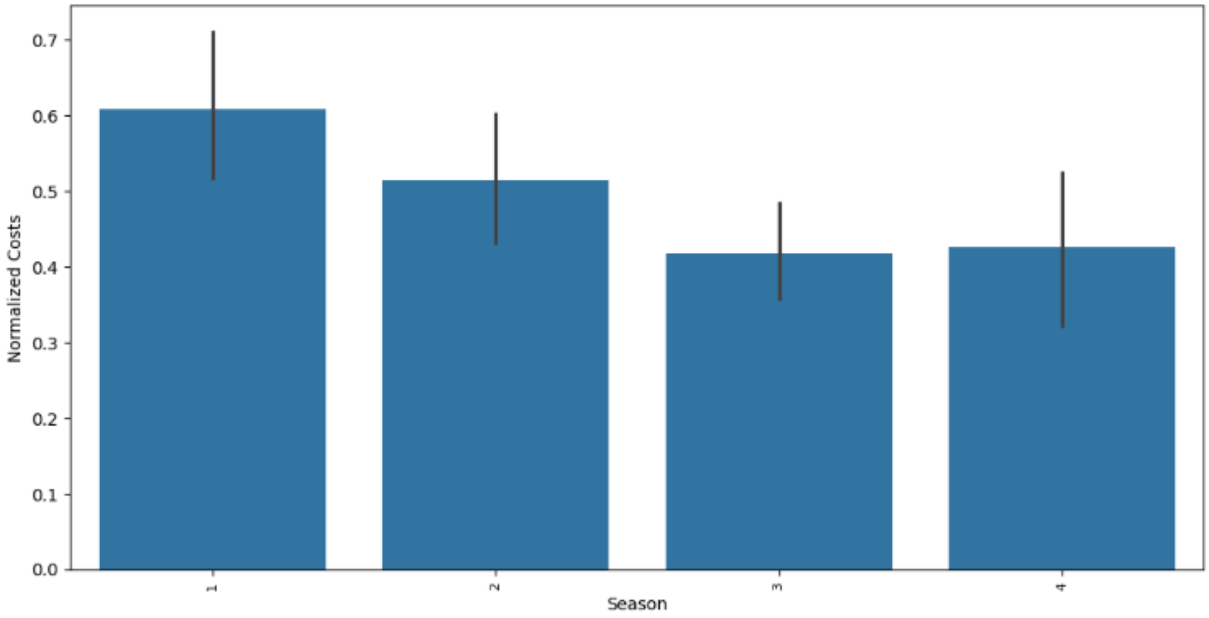

Figure 5-14 Seasonal costs clinical chemistry costs normalised

5.5.7 Shapley Additive Explanations Importance

Table 5-20 provides each variable's relative impact and corrected relative impact on the model's output and the direction of influence. Table 5-20 shows a clear difference between the relative impact of SHAP values and the relative impact of the corrected SHAP values. The difference is comparable to that observed in the laundry contract analysis. Consequently, the same conclusion was drawn: the corrected relative impact represents the variables' relative impact.

We discussed these results with the subdepartment that manages and is seen as an expert on the subject. The main explanation for the results was that clinical admissions require more expensive clinical chemistry procedures than day hospitalisations and operations. Hence, the positive direction of influence for clinical admissions and the negative direction of influence for the number of day hospitalisations and operations was explained. Furthermore, the positive direction of influence of the

number of DTCs was explained by the nature of every clinical chemistry assay, which requires a DTC request. Furthermore, the subdepartment recognised the seasonality aspects of the results. Overall, the results seemed well-founded to the subdepartment.

Additionally, Figure 5-15 displays the SHAP values for the laundry contract, with a colour gradient indicating variable levels: red represents high values, while blue represents low values. Most variables are balanced, with a high value in the variables positively influencing costs. However, the number of day hospitalisations negatively influences costs.

Variable	**Relative impact**	**Corrected relative impact**	**Direction of influence**
Number of visits to the outpatient clinic	48.94%	7.62%	Positive
Number of day hospitalisations	12.30%	10.62%	Negative
1st quarter	9.55%	0.49%	N.A.
Number of DTCs	6.93%	12.54%	Positive
Number of clinical admissions	6.13%	41.04%	Positive
Number of operations	5.39%	21.06%	Negative
Number of intensive care admissions	4.85%	4.57%	Neutral
4th quarter	3.13%	0.51%	N.A.
3rd quarter	1.05%	0.50%	N.A.
2nd quarter	0.71%	0.48%	N.A.
Influenza season	0.29%	0.57%	N.A.

Table 5-20 Relative impacts clinical chemistry contract

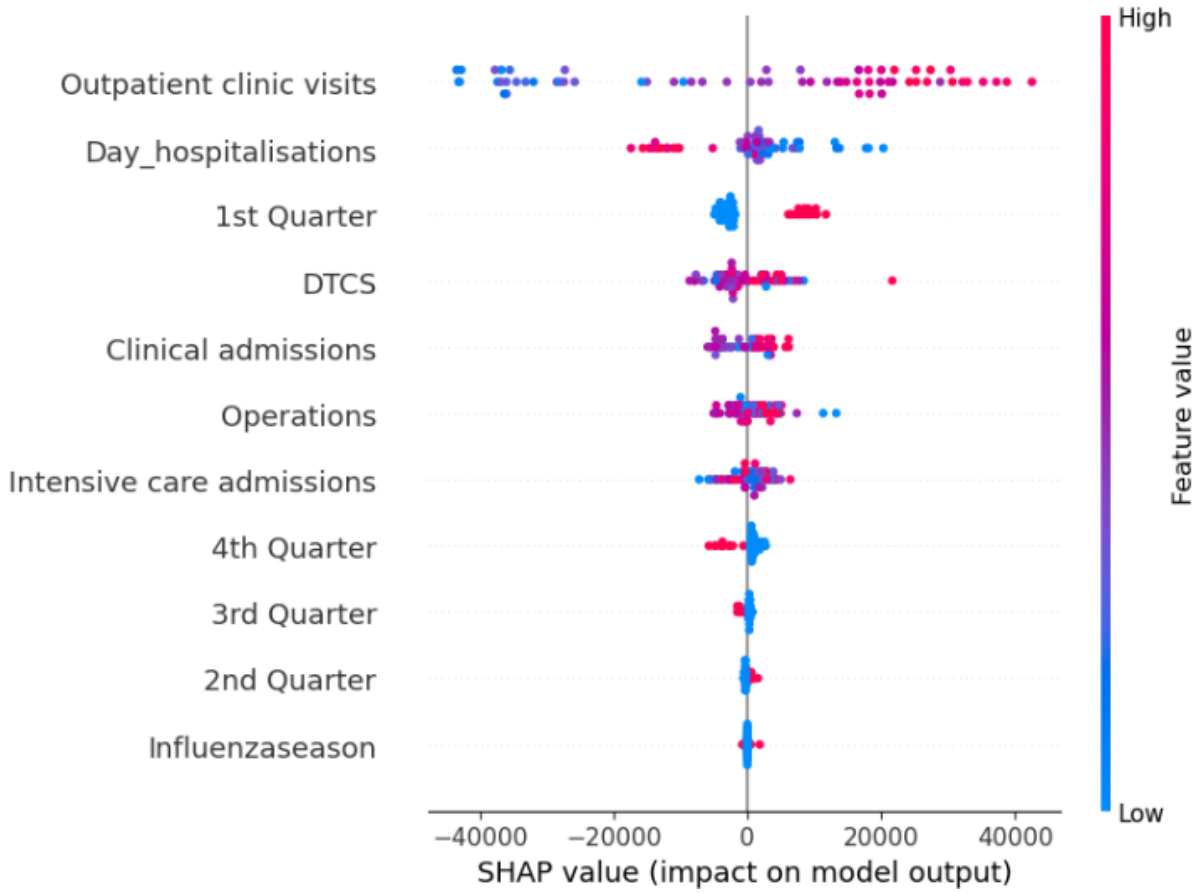

Figure 5-15 SHAP values clinical chemistry contract

5.5.8 Random Forest Prediction

Figure 5-16 illustrates the RF regression results for the analysed time series of the clinical chemistry contract, presented on a normalised scale ranging from zero to one for the actual values, while the predicted values are presented on the same scale. Additionally, Table 5-20 presents the validation metrics of the laundry contract on the same normalised scale. An R^2 of -0.2955 indicates that the model is a poor fit for the data, as it performs worse than using the mean of the actual data as a prediction. Figure 5-16 shows that the predicted values are consistently too low. This low prediction can result from no indexation values being accounted for in the prediction, as there has been an indexation number of eight per cent year on year from 2023 to 2024 (MST, Personal communication, January 14, 2025). However, as explained in Chapter 5.4.8, these results must be interpreted cautiously because of the small data size.

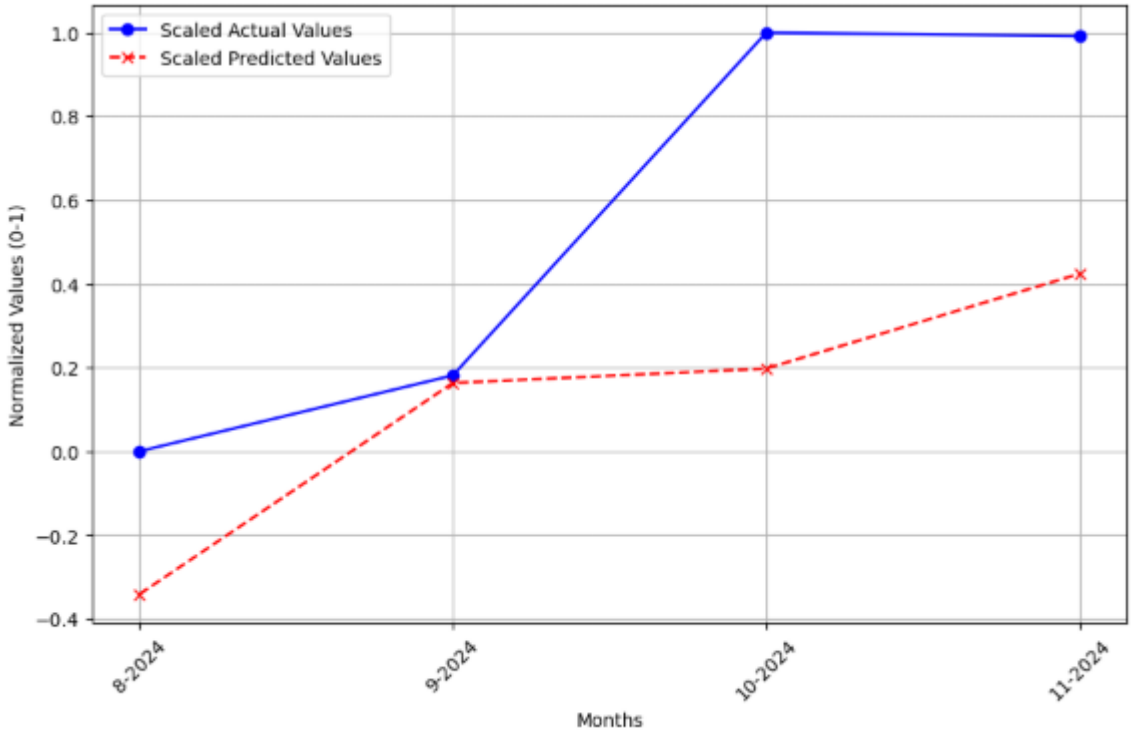

Figure 5-16 Comparison actual and prediction values clinical chemistry contract

Validation metric	Outcome
R^2 Score	-0.2955
MSE	0.2708
MAE	0.4324

Table 5-21 Validation metrics clinical chemistry contract

5.6 Key Findings

Tables 5-22 and 5-23 outline the variables statistically impacting the laundry and clinical chemistry contract costs and their direction of influence. The analysis reveals that operational variables significantly influence laundry and clinical chemistry contract costs. Notably, the number of clinical admissions is both contracts' most significant cost driver. Furthermore, the cost drivers for both contracts are similar, but the magnitudes of their effects differ. Seasonal factors and the influenza season also affect costs for both contracts, albeit to a much lesser extent.

Furthermore, with an R^2 score of 0.5926 for the laundry contract and -0.2955 for the clinical chemistry contract, the laundry contract model appears useful with further refinement. At the same time, the clinical chemistry predicted values that were too low for the four predicted months.

Variable	Relative impact	Direction of influence
Number of clinical admissions	34.15%	Positive
Number of DTCs	29.63%	Positive
Number of operations	17.51%	Positive
Number of visits to the outpatient clinic	8.61%	Positive
Number of day hospitalisations	8.33%	Negative
Influenza season	0.46%	N.A.
1st quarter	0.42%	N.A.
2nd quarter	0.35%	N.A.
3rd quarter	0.28%	N.A.
4th quarter	0.24%	N.A.

Table 5-22 Relative impacts laundry contract

Variable	Relative impact	Direction of influence
Number of clinical admissions	41.04%	Positive
Number of operations	21.06%	Negative
Number of DTCs	12.54%	Positive
Number of day hospitalisations	10.62%	Negative
Number of visits to the outpatient clinic	7.62%	Positive
Number of intensive care admissions	4.57%	Neutral
Influenza season	0.57%	N.A.
4th quarter	0.51%	N.A.
3rd quarter	0.50%	N.A.
1st quarter	0.49%	N.A.
2nd quarter	0.48%	N.A.

Table 5-23 Relative impacts clinical chemistry contract

5.7 Chapter Conclusion

This chapter outlined the results addressing subquestions five, six, and seven. It focused on identifying unknown variables influencing contract costs, evaluating their impacts, and assessing the explainable performance of models incorporating these variables.

In addressing subquestion five, “What unknown variables affecting contract costs can be identified through data-driven techniques?”, several key variables were identified for both contracts. These included clinical admissions, diagnoses treatment combinations, and outpatient visits.

Subquestion six, “To what extent and direction does each identified variable influence contract cost predictions?” was also addressed. This subquestion provided new insights into the newly identified variables by quantifying each identified variable’s relative impact and direction of influence by leveraging SHAP values.

Finally, subquestion seven, “When incorporating the identified variables, how well does the model perform, as measured by accuracy and other relevant evaluation metrics?” was answered. The results demonstrated mixed results of the RF prediction models.

The next chapter concludes the main research question. It highlights the research limitations and offers recommendations for handling these limitations in future research. Additionally, it offers recommendations for MST to apply these findings in practice.

6. Conclusion, Recommendations, and Limitations

6.1 Conclusion

We aimed to identify previously unknown variables influencing contract costs at MST, ultimately enhancing forecasting and monitoring capabilities within the Contracts & Process Management subdepartment. We addressed the main research question: "Which unknown variables influencing contract costs can be uncovered using data-driven models?". We uncovered previously unidentified key variables by applying data-driven methods, particularly ML techniques and SHAP interpretation. These findings contribute significantly to the scientific understanding of cost variability within hospital contract costs.

This research's core contribution is demonstrating the identification and quantification of variables that drive cost fluctuations through iForest anomaly detection, linear regression, MLR, and two-way ANOVA, combined with SHAP explanations of RF regression. This research advances theory and practice by integrating XAI methods into contract management.

The main theoretical contributions lie in filling the identified gaps in the literature outlined in Chapter 3.3. To our knowledge, no interpretable solution addressed multicollinearity for MLR using SHAP. Therefore, this research is the first to offer a replicable approach that combines different variables using SHAP values. Additionally, to our knowledge, this research is the first to apply XAI techniques in a hospital environment where multicollinearity between variables is present, leveraging SHAP with an adjustment factor. To validate the adjusted SHAP values, they were compared to the original ones and validated with the subdepartment's field experts, which we did not observe in the reviewed literature.

From a practical perspective, integrating ML techniques and SHAP analysis offers a replicable framework for analysing other contracts in healthcare institutions. Furthermore, the findings offer actionable insights for healthcare organisations, particularly MST, to improve cost forecasting accuracy.

This research is a step forward in combining ML methods with XAI to uncover hidden cost drivers in healthcare contracts. By offering both theoretical advancements and practical tools, it provides a foundation for future research and implementation.

6.2 Limitations

While this research established many insightful insights, several limitations impacting the outcomes should be considered when interpreting the findings.

1) **Exclusion of potential variables**: While this study attempted to include all relevant variables through interviews with key stakeholders and exploratory data analysis, there is a risk that crucial variables influencing contract costs will be left out. If important variables are excluded, the model lacks a complete understanding of the factors driving contract costs. Furthermore, if some variables are excluded, the model may disproportionately emphasise the importance of those included, creating an imbalanced or distorted view of the factors influencing contract costs. For example, the RF prediction models did not incorporate indexation data or other financial factors. This missingness could lead to – as demonstrated in the clinical chemistry prediction – lower predictions than actual values.
2) **Model limitations:** The models used in this research provided valuable insights into the variables influencing contract costs but were subject to certain limitations. MLR assumes linear relationships between variables, which may not fully capture the non-linear dynamics present

in real-world data. While some seasonal effects were addressed using two-way ANOVA, other complex patterns remained unexplored.
Additionally, outliers were removed from the dependent variable through anomaly detection to improve model robustness. However, anomalies in independent variables were not excluded, potentially introducing noise into the results. Noise from anomalies in independent variables can distort relationships, reduce accuracy, bias estimates, and lead to misleading insights or overfitting.

3) **Limited data range**: The included data ranges from January 2019 until July 2024, limited through data availability at MST, limiting the insights to that timeframe. This range may not fully capture long-term patterns. Furthermore, significant events within the data range, such as the COVID-19 pandemic, could introduce variability in contract costs that may not be representative of typical operations. Furthermore, overfitting can occur in RF regression through a limited data range (Ying, 2019). Overfitting would lead to the regression's results not being generalisable to other data.
4) **Data quality**: In terms of data quality, missing values—influenza data for parts of 2020—posed challenges for seasonal analyses. Additionally, adjustments to align data formats, such as converting weekly influenza reports to monthly aggregates, may have introduced discrepancies. While anomalies in independent variables were retained to preserve dataset completeness, they could contribute noise, potentially affecting model precision.
5) **Using Python libraries' default settings:** While the model's performance is up to standard with current validation measures, no other settings for the Python functions have been explored. For example, Han and Kim (2021) noted that using different settings in an RF regression could be beneficial.
6) **A limited number of analysed articles in the SLRs:** While two SLRs were employed, the number of reviewed studies included is limited to four and five, respectively. The articles have been rigorously chosen based on relevance and quality to ensure meaningful insights. However, a smaller pool of articles may not fully capture the diversity of approaches available in the field of research. Furthermore, critical gaps or new techniques might be missed with fewer evaluated studies.

Future research could broaden the scope of variables to ensure that no significant factors are overlooked, and the research should include financial factors in prediction models to evaluate the performance with these factors included. More advanced and complex models might also prove helpful in identifying non-linear patterns in the data. Expanding the dataset to include a more extended timeframe would help uncover long-term trends and minimise the influence of short-term fluctuations, such as those caused by the COVID-19 pandemic. Furthermore, with more data availability, the impact of some of the variables outlined in Appendix A on the contract costs can be evaluated. Additionally, different settings for the Python functions could be explored to explore the model's optimal settings. Lastly, future research could expand the SLR to include more studies as the field develops and explore different inclusion- and exclusion criteria to capture a more comprehensive selection of studies on the topic.

6.3 Recommendations

After researching and developing a model to determine cause-and-effect relationships within contract costs, several recommendations can be made to the MST Contracts & Process Management subdepartment to enhance their forecasting and monitoring capabilities. These recommendations align with the practical needs of the subdepartment at MST and are structured to ensure their applicability.

1) Incorporate identified variables into forecasting models

We identified variables such as clinical admissions, outpatient clinic visits, and DTCs as significant predictors. Currently, no forecasting models leveraging other variables are incorporated by the subdepartment. The subdepartment could establish a research project in collaboration with a university, engaging a student to investigate how to incorporate these variables into a tailored forecasting model. This approach would allow exploring the most effective methods to integrate the identified variables, ensuring practical applicability and alignment with the subdepartment's needs.

2) Integration of anomaly detection model

An anomaly detection model, such as iForest, should be introduced to identify outliers and deviations periodically in contract costs. The subdepartment can flag irregularities by periodically performing anomaly detection on contract costs, enabling it to address deviations before they escalate. The subdepartment can do this by regularly running a Python script with contract costs or investigating how to integrate anomaly detection in existing tools, such as Power BI.

3) Regularly review the identified variables

When integrating the identified variables into forecasting models, the variables should influence contract costs. When operations evolve, key cost drivers may change. The subdepartment can validate and refine the findings by periodically (e.g., yearly) running the Python script developed in this research using updated data. This iterative process ensures that the identified variables and their impacts remain relevant.

4) Use this replicable framework for other contracts

If time and resources allow, the subdepartment can leverage the developed Python model to identify key cost drivers for other contracts managed by the subdepartment beyond those investigated in this research. The subdepartment can uncover new insights into cost-influencing variables by applying the model to additional contracts, providing a broader understanding of cost drivers across its portfolio. However, we recommend doing this after incorporating the identified variables into forecasting models since the benefit of identifying the variables is then clearer.

5) Aligning budgets with seasonal patterns

The two-way ANOVA analysis revealed stable seasonal patterns in contract costs. The subdepartment can use this information to align its budgeting and operational planning processes with these seasonal trends. For example, during periods identified as cost-intensive, such as the first quarter, additional resources can be allocated proactively to avoid disruptions.

6) Share findings with other hospitals

The findings of this research, including the identified cost drivers and the methodological framework, can provide valuable insights to other hospitals facing similar challenges in contract management. Sharing these results can improve the healthcare sector's forecasting accuracy and cost control. For example, this sharing of information can be achieved by presenting the findings at a Santheon hospitals meeting, a network to which MST belongs.

Bibliography

Abonyi, J., Kummer, A., & Hanzeliek, P. P. (2022). Edge-Computing and Machine-Learning-Based Framework for Software Sensor Development. *Sensors*. doi:10.3390/s22114048

Al-Faiz, M. Z., Hadi, S., & Ibrahim, A. A. (2019). The effect of Z-Score standardization (normalization) on binary input due to the speed of learning in back-propagation neural network. *Iraqi Journal of Information & Communications Technology*, 42-48. doi:10.31987/ijict.1.3.41

Ampadu, V.-M. K., Ker, A. J., Wulff, S. S., & Ksaibati, K. (2024). Predicted Average Annual Cost of Crashes on the US-16 Wyoming Downgrade Using Time Series Analysis and Forecasting. *International Journal of Intelligent Transportation Systems Research*, 53-68. doi:10.1007/s13177-023-00378-w

Basu, I., & Maji, S. (2022). Multicollinearity Correction and Combined Feature Effect in Shapley Values. *AI 2021: Advances in Artificial Intelligence* (pp. 79–90). Springer.

Benhar, H., Idri, A., & Fernández-Alemán, J. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine, 195*. doi:10.1016/j.cmpb.2020.105635

Bonovas, S., & Piovani, D. (2023). On p-Values and Statistical Significance. *Journal of Clinical Medicine, 12*(3). doi:10.3390/jcm12030900

Breiman, L. (2001). Random forests. *Machine Learning*, 5-32. doi:10.1023/A:1010933404324

Brunet, A., & César, F. (2019). *Contract Management: Contractual performance, renegotiation, and claims: How to safeguard and increase profit margins.* Cham, Switzerland: Springer Nature Switzerland AG. doi:/10.1007/978-3-030-68076-3

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*. doi:10.7717/peerj-cs.623

França, R. P., Monteiro, A. C., Arthur, R., & Iano, Y. (2021). An overview of deep learning in big data, image, and signal processing in the modern digital age. In *Trends in Deep Learning Methodologies: Algorithms, Applications, and Systems* (pp. 63-87). Academic Press. doi:10.1016/B978-0-12-822226-3.00003-9

Guo, Y.-p., Wang, D.-f., Zheng, Y., & Ding, W.-b. (2022). Mathematical Methods for Maintenance and Operation Cost Prediction Based on Transfer Learning in State Grid. *Applied Mathematics Journal of Chinese Universities, 37*(4), 598-614. doi:10.1007/s11766-022-4319-7

Gutterman, A. S. (2023). Contract management. *SSRN*, 1-48.

Han, S., & Kim, H. (2021). Optimal Feature Set Size in Random Forest Regression. *Applied Sciences*. doi:10.3390/app11083428

Heerkens, H., Van den Winden, A., & Tjooitink, J.-W. (2017). *Solving Managerial Problems Systematically* (1st ed.). [Place of publication not identified]: Noordhoff Uitgevers : Noordhoff Uitgevers BV.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R.* New York: Springer.

Javaheri, S. H., Sepehrie, M. M., & Teimourpour, B. (2014). Response Modeling in Direct Marketing: A Data Mining-Based Approach for Target Selection. In Y. Zhao, & Y. Cen, *Data Mining Applications with R* (pp. 153-181). doi:10.1016/B978-0-12-411511-8.00006-2

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 255-260. doi:10.1126/science.aaa8415

Khan, A., & Mir, M. S. (2021). Contracting in healthcare. *Biomedical Journal of Scientific & Technical Research, 36*(3), 28503-28507. doi:10.26717/BJSTR.2021.36.005846

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology, 72*(6), 555-566. doi:10.4097/kja.19087

Liao, Q. V., & Varshney, K. R. (2022). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint*. doi:10.48550/arXiv.2110.10790

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2009). Isolation Forest. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. doi:10.1109/ICDM.2008.17

Lloyd, R. V. (2023). *Pathology: Historical and Contemporary.* Springer Nature Switzerland AG. doi:10.1007/978-3-031-39554-3

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017).* Long Beach, CA, USA.

Luo, L., Yu, X., Yong, Z., Li, C., & Gu, Y. (2021). Design Comorbidity Portfolios to Improve Treatment Cost Prediction of Asthma Using Machine Learning. *IEEE Journal of Biomedical and Health Informatics*, 2263-2272. doi:10.1109/JBHI.2020.3034092

Marill, K. A. (2004). Advanced statistics: linear regression, part II: multiple linear regression. *Academic Emergency Medicine, 11*(1), 94–102. doi:10.1197/j.aem.2003.09.006

Maryati, I., Christian, & Paramita, A. S. (2023). Gold Prices Time-Series Forecasting: Comparison of Statistical Techniques. *Journal of Applied Data Sciences, 4*(4), 372-381. doi:10.47738/jads.v4i4.135

Maulud, D. H., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends, 1*(4), 140-147. doi:10.38094/jastt1457

Medisch Spectrum Twente. (2024). *over-mst*. Retrieved from Medisch Spectrum Twente: https://www.mst.nl/over-mst/

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*. doi:10.1186/1471-2105-10-213

Mishra, S. K. (2016). Shapley Value Regression and the Resolution of Multicollinearity. *SSRN Electronic Journal*. doi:10.1016/j.artint.2021.103502

National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport. (2019). *Annual report surveillance of influenza and other respiratory infections in the Netherlands: Winter 2018/2019.* Bilthoven, Netherlands: National Institute for Public Health and the Environment.

National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport. (2020). *Annual report surveillance of influenza and other respiratory infections in the Netherlands: Winter 2019/2020.* Bilthoven, Netherlands: National Institute for Public Health and the Environment.

National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport. (2021). *COVID-19 dominated the 2020/2021 flu season.* Bilthoven, Netherlands: National Institute for Public Health and the Environment. Retrieved from https://www.rivm.nl/nieuws/covid-19-overheerste-griepseizoen-2020-2021

National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport. (2024, November 7). *Feiten en cijfers griep*. Retrieved from RIVM: https://www.rivm.nl/griep-griepprik/feiten-en-cijfers

National Institute for Public Health and the Environment, Ministry of Health, Welfare and Sport. (2024, November 15). *Griepepidemie in Nederland*. Retrieved from RIVM: https://www.rivm.nl/nieuws/griepepidemie-in-nederland-1

Nghiem, N., Atkinson, J., Nguyen, B. P., Tran-Duy, A., & Wilson, N. (2023). Predicting High Health-Cost Users Among People with Cardiovascular Disease Using Machine Learning and Nationwide Linked Social Administrative Datasets. *Health Economics Review, 13*. doi:10.1186/s13561-023-00422-1

Nisan, N., Roughgarden, T., Tardos, É., & Vazirani, V. V. (2007). *Algorithmic Game Theory.* Cambridge: Cambridge University Press.

Niu, L., Lu, J., & Zhang, G. (2009). 1.5.3 Data Mining. In L. Niu, J. Lu, & G. Zhang, *Cognition-Driven Decision Support for Business Intelligence: Models, Techniques, Systems, and Applications* (pp. 14-15). Berlin: Springer.

Page, M., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Hróbjartsso. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. doi:10.1136/bmj.n71

Qiu, S., Liu, Q., Zhou, S., & Huang, W. (2022). Adversarial Attack and Defense Technologies in Natural Language Processing: A Survey. *Neurocomputing, 492*, 278–307. doi:10.1016/j.neucom.2022.04.020

Rokach, L., & Maimon, O. (2005). Decision Trees. In *Data Mining and Knowledge Discovery Handbook* (pp. 165-192). Springer.

Salmerón, R., García, C. B., & García, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation, 88*(12), 2365–2384. doi:10.1080/00949655.2018.1463376

Sana, L., Nazir, M. M., Yang, J., Hussain, L., Chen, Y.-L., Ku, C. S., . . . Ypp Po, U. (2024). Securing the IoT Cyber Environment: Enhancing Intrusion Anomaly Detection with Vision Transformers. *IEEE Access*. doi:10.1109/ACCESS.2024.3404778

Schirmer, P. A., & Mporas, I. (2024). PyDTS: A Python Toolkit for Deep Learning Time Series Modelling. *Entropy*. doi:10.3390/e26040311

Sclove, S. L., & Wang, F. (2014). ANOVA-Based Tests for Stable Seasonal Pattern, with Applications to Analysis and Forecasting of Economic Data. (pp. 4061-4070). state Chicago: American Statistical Association.

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics, 8*(2), 39-42. doi:10.12691/ajams-8-2-1

Sihabuddin, A., Rokhman, N., & Wahyudi, E. E. (2024). A Machine Learning Approach on Outlier Removal for Decision Tree Regression Method. *IETA Journal*. doi:10.18280/isi.290414

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning & Knowledge Extraction*, 392–413. doi:10.3390/make3020020

Uyanık, G. K., & Güler, N. (2013). A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, 234–240. doi:10.1016/j.sbspro.2013.12.027

Xiong, Z., Zhu, D., Liu, D., He, S., & Zhao, L. (2022). Anomaly Detection of Metallurgical Energy Data Based on iForest-AE. *Applied Sciences*. doi:10.3390/app12199977

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*. doi:10.1088/1742-6596/1168/2/022022

Zapf, A., Wiessner, C., & König, I. R. (2024). Regression Analyses and Their Particularities in Observational Studies. *Deutsches Ärzteblatt International, 121*(4), 128-134. doi:10.3238/arztebl.m2023.0278

Appendix

A. Variables Not Included in Analysis

1. **Australian Influenza cases:** This variable was not analysed because no data was available from 2020 to 2021 due to the COVID-19 pandemic.
2. **Demographic data**: Several demographic variables were not analysed due to limitations in their analysis. Demographic data is typically more meaningful when analysed yearly than monthly. Since this research focused on monthly data from 2019 onward, incorporating demographic variables was not feasible. Furthermore, analysing demographic data based on only five years would not result in reliable or representative results.
3. **Employee holiday days**: This variable was not analysed because an increase in employee holiday days directly results in fewer eight-hour shifts. As a result, this variable accounts for all the reasons employees might not be working.
4. **Hospital in and outflow:** In MST, sensors document the in and outflow of persons. These sensors have been in place since July 2019. This variable was not analysed since the data start is seven months later than the start of the analysed data range, and the in- and outflow was almost non-existent during 2020 and 2021 due to the COVID-19 pandemic.

B. Systematic Literature Review Protocols

B.1 Systematic Literature Review Protocol Anomaly Detection Methods

This study uses four inclusion criteria and four exclusion criteria. A publication will be excluded from the research if it does not fulfil one of the specified criteria. Tables A-1 and A-2 show the inclusion and exclusion criteria.

Inclusion Criteria	**Motivation**
Contains anomaly detection in time series as a method	Relevant with the focus on anomaly detection in time series
Was researched in contract management, hospital operations, or transferrable insights from another sector	With the focus on contract management in the knowledge question but keeping the freedom to collect transferable insights
Is a peer-reviewed article	Ensures credibility and usability of sources
Utilises a quantitative data-driven anomaly detection method	Ensures relevance to the research

Table A-1 Inclusion Criteria anomaly detection methods

Exclusion Criteria	**Motivation**
Publications published before 2014 unless they represent foundational work in anomaly detection	This includes recent research reflecting advancements in anomaly detection over the last ten years.
Publications published in another language than English	Ensure accessibility, continuity with research, and comprehension of the literature. Furthermore, it ensures the author can read the materials.
Not available to the author in full-text format	Ensures the author can access the complete study for thorough review and analysis.

Table A-2 Exclusion Criteria anomaly detection methods

This systematic literature utilises two databases:

- **Scopus:** Chosen for its broad coverage of multidisciplinary topics, which helps to cover different perspectives on anomaly detection and contract management.
- **IEEE Xplore:** Chosen for its focus on science, technology, and engineering. Its specialisation in data analytics and computer science makes it suitable for researching anomaly detection and time series analysis advancements.

We adopted a structured approach to find the most suitable anomaly detection method. The key concepts from the knowledge question and the inclusion/exclusion criteria form the foundation of the following concepts.

Key Concepts	**Related Terms and Synonyms**
Anomaly detection	Anomaly model, Outlier detection, Deviation detection
Time series	Sequential data, chronological data, Temporal data
Cost forecasting, healthcare	Financial forecasting, Healthcare operations, Procurement

Table A-3 Key concepts anomaly detection methods

We adopted the following search string incorporating the key- and related concepts:

("Anomaly detection" OR "Anomaly model" OR "Outlier detection") AND ("Time series" OR "Sequential data" OR "Chronological data") AND ("Financial forecasting" OR "Healthcare operations" OR "Cost prediction")

The search string yielded 313 results, of which 256 records were excluded due to automation tools. When identifying potential studies, the non-eligible studies were first removed. This removal was done by applying Scopus and IEEE Xplore's automation tools according to the inclusion - and exclusion criteria. Some studies were excluded due to multiple criteria. Consequently, the automation tools do not add up exactly; the total number of excluded studies is those excluded by automation tools. After cross-examining the results of the databases, no duplicates were found and removed. These exclusions included one study due to language, 83 due to lack of open access restrictions, 162 due to not being a peer-reviewed article and being a wrong document type, and 21 due to the publication year being before 2014.

After removing duplicates and by automation tools, the remaining records were screened by reading the abstract and title. Through this process, their application to this research was determined. Fifty studies were excluded during this phase. The primary reasons for exclusion were as follows: 35 studies did not employ anomaly detection as a method, fourteen were deemed not applicable to this research, and one study was invalid due to fraud.

Then, the remaining studies were assessed for eligibility by reading the full-text version. Three of the seven remaining reports assessed for eligibility were excluded after full-text review because, for two studies, their methods were non-applicable or, for one study, their context was non-transferable to a healthcare context. Ultimately, four studies were included in the review. This process is illustrated in the inclusion diagram in Figure A-1 (Page et al., 2021).

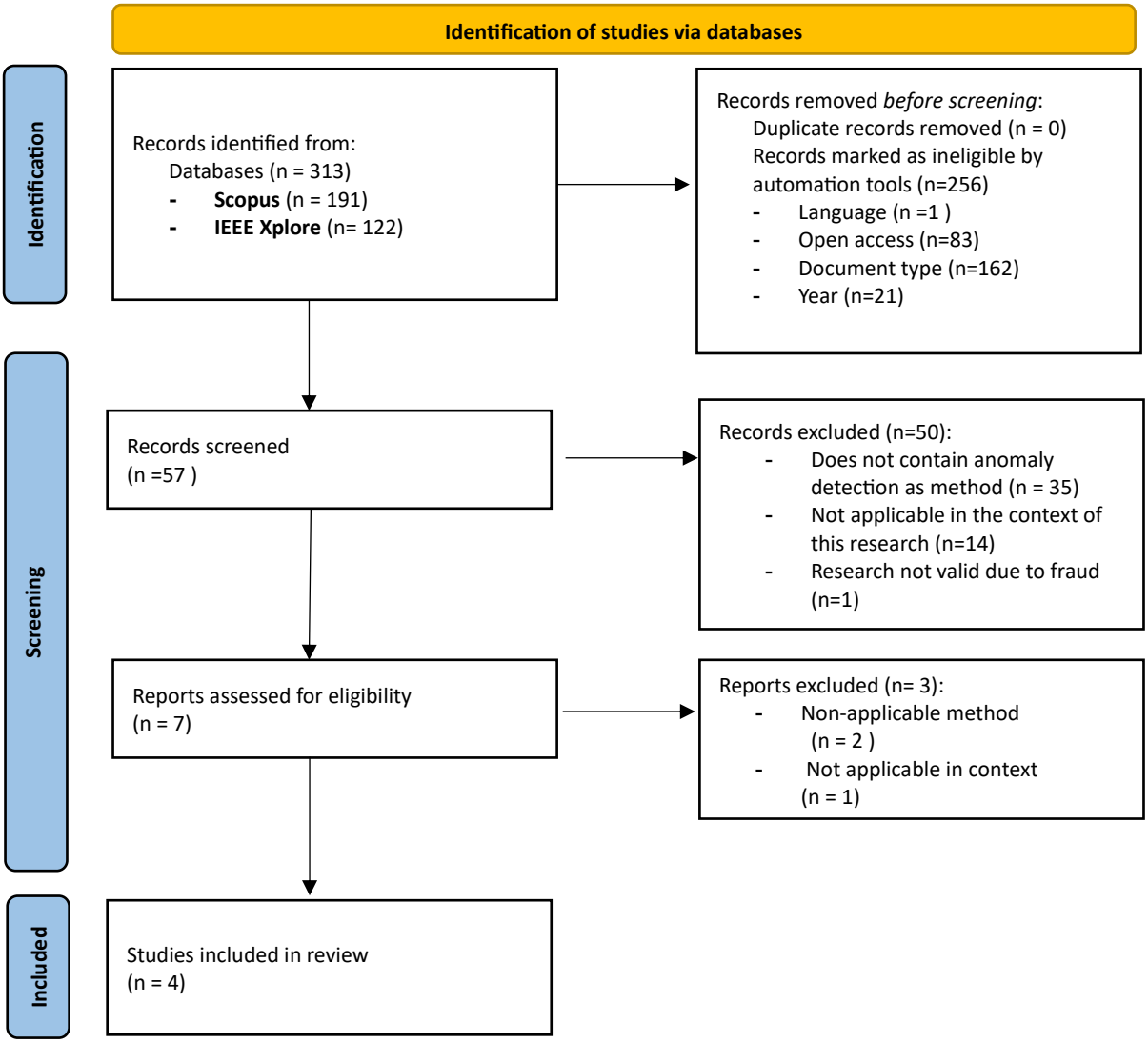

Figure A-1 Inclusion diagram anomaly detection method

Table A-4 shows the key concepts, models, and methods identified in the reviewed articles, offering a structured overview of the literature that informed the research.

Title	Author(s)	Year	Key algorithms/methods	Contribution
Anomaly Detection of Metallurgical Energy Data Based on iForest-AE	(Xiong, Zhu, Liu, He, & Zhao, 2022)	2022	iForest algorithm, Autoencoders Algorithm, iForest-AE Method	The iForest algorithm effectively detects anomalies by isolating outliers with fewer splits. The iForest-AE method further improves this by combining iForest with autoencoders, enhancing anomaly detection accuracy.
PyDTS: A Python Toolkit for Deep	(Schirmer & Mporas, 2024)	2024	Time series modelling, anomaly	Deep learning models, especially

Learning Time Series Modelling			detection, ML (RF, KNN, SVM), DL (CNN, LSTM)	CNN and LSTM, excel in detecting anomalies in time series data. RF, KNN, and SVM perform well in non-DL approaches, with RF being the best-performing algorithm.
An ML Approach on Outlier Removal for Decision Tree Regression Method	(Sihabuddin, Rokhman, & Wahyudi, 2024)	2023	IForest, Decision Tree Regression, Minimum Covariance Determinant-DTR, Local Outlier Factor-DTR, One Class SVM-DTR	IForest, combined with Decision Tree Regression, provides improved outlier detection and enhances the identification of anomalies in regression analysis.
Securing the IoT Cyber Environment: Enhancing Intrusion Anomaly Detection With Vision Transformers	(Sana et al., 2024)	2024	Vision Transformers, IoT, anomaly detection, deep learning	Vision Transformers provide superior accuracy in detecting anomalies within IoT networks, surpassing traditional models. However, traditional models like RF and ensemble bagged trees achieved more than 99.90% accuracy, making them strong contenders for specific use cases.

Table A-4 Concept matrix anomaly detection methods

B.2 Systematic Literature Review Protocol Regression Models

This study uses four inclusion criteria and three exclusion criteria. A publication will be excluded from the research if it does not fulfil one of the specified criteria. Tables A-5 and A-6 show the inclusion- and exclusion criteria.

Inclusion Criteria	**Motivation**
It contains regression in time series as a method with a single dependent variable	Relevant with the focus on regression in time series
Was researched in contract management, hospital operations, or transferrable insights from another sector	With the focus on contract management in the knowledge question but keeping the freedom to collect transferable insights
Is a peer-reviewed article	Ensures credibility and usability of sources
Utilises a quantitative data-driven regression detection method	Ensures relevance to the research

Table A-5 Inclusion criteria regression methods

Exclusion Criteria	Motivation
Publications published before 2014 unless they represent foundational work in anomaly detection	This includes recent research reflecting advancements in anomaly detection over the last ten years.
Publications published in another language than English	Ensure accessibility, continuity with research, and comprehension of the literature. Furthermore, it ensures the author can read the materials.
Not available to the author in full-text format	Ensures the author can access the complete study for thorough review and analysis.

Table A-6 Exclusion criteria regression methods

This systematic literature utilises three databases:

- **Scopus:** Chosen for its broad coverage of multidisciplinary topics, which helps cover different regression perspectives.
- **IEEE Xplore:** Chosen for its focus on science, technology, and engineering. Its specialisation in data analytics and computer science makes it suitable for researching regression and time series analysis advancements.
- **ACM Digital Library:** Chosen for its specialisation in computing and information technology, ACM provides access to research in areas such as ML, artificial intelligence, and decision-making support systems, making it particularly suitable for analysing regression methods.

To answer the knowledge question, "Which regression models are most suitable for analysing cause-and-effect between variables in the context of this research?

We adopted a structured approach. The key concepts from the knowledge question and the inclusion/exclusion criteria form the foundation of the concepts shown in Table A-7.

Key Concepts	Related Terms and Synonyms
Regression	Linear regression, Non-linear regression, Least squares estimation, Regression analysis, Regression modelling
Time series	Sequential data, chronological data, Temporal data
Cost forecasting	Financial forecasting, Procurement
Healthcare	Healthcare operations

Table A-7 Key concepts regression methods

After trial- and error, search among three databases to find the correct search string. In order to find suitable regression methods with the inclusion criteria, we adopted the following search string incorporating the key- and related concepts:

("Regression" OR "Linear regression" OR "Non-linear regression" OR "Least squares estimation" OR "Regression analysis") AND ("Time series" OR "Sequential data" OR "Chronological data") AND ("Financial forecasting" OR "Healthcare operations" OR "Cost prediction")

The search string yielded 256 results, of which 155 records were excluded using automation tools. These exclusions comprised four duplicate records and 126 studies that were not peer-reviewed articles.

After applying these automated exclusions, 97 studies remained and were systematically screened and assessed based on the predefined inclusion and exclusion criteria. During the screening phase, 90 studies were excluded for the following reasons: 64 did not employ regression as a method, 24 were deemed not applicable to a hospital environment, and two were inaccessible to the author.

Two of the seven remaining studies were excluded after a full-text review because their methods were unsuitable for application in a hospital setting. Consequently, five studies were ultimately included in the review. The inclusion diagram in Figure A-2 depicts the entire process (Page et al., 2021).

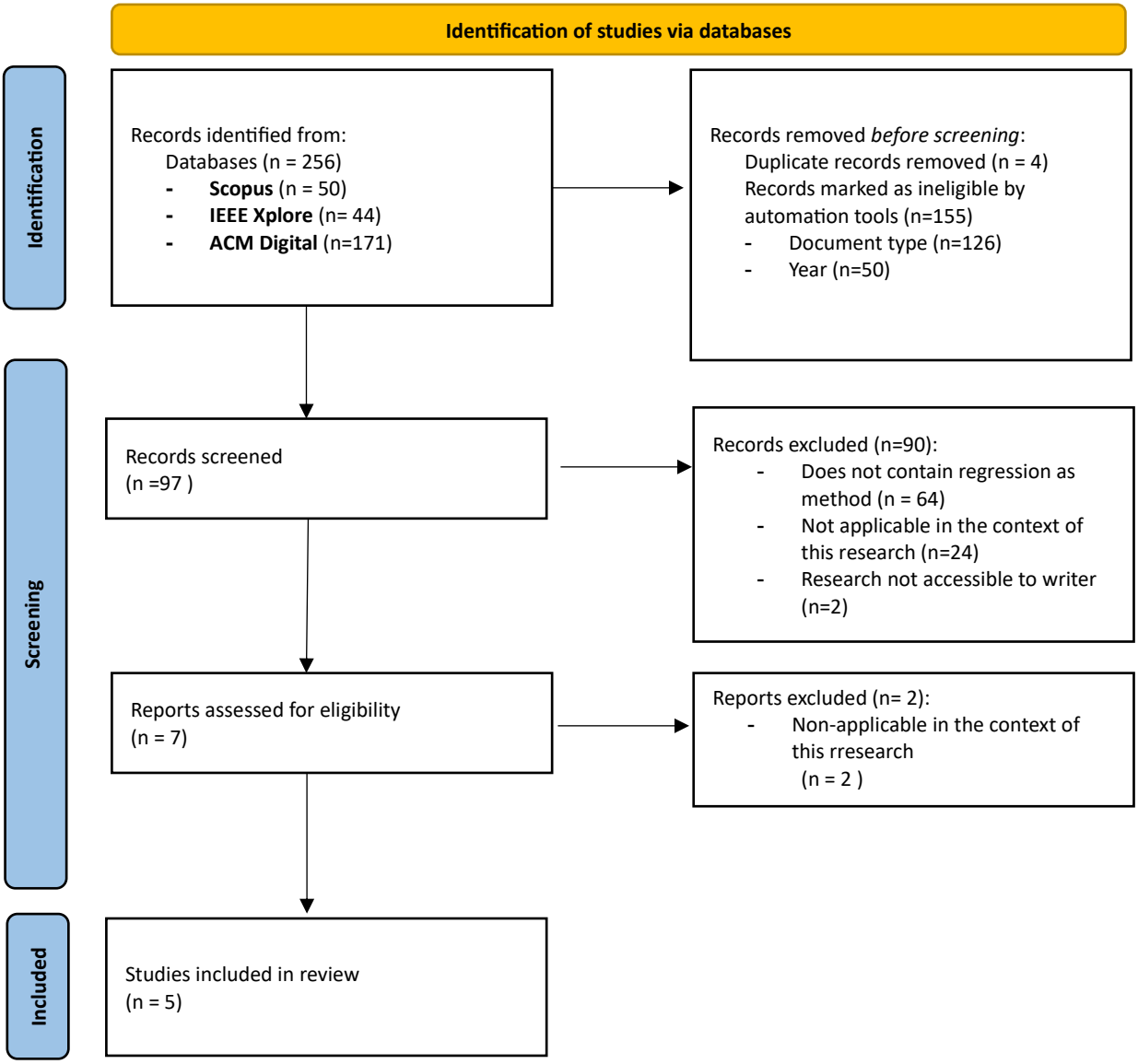

Figure A-2 Inclusion diagram regression methods

After excluding studies according to the set criteria, this study evaluates five.

Table A-8 outlines the key concepts, models, and methods identified in the reviewed articles, offering a structured overview of the literature that informed the research.

Title	Author(s)	Year	Key algorithms/methods	Contribution
Design comorbidity portfolios to improve treatment cost prediction of asthma using Machine Learning	(Luo, Yu, Yong, Li, & Gu, 2021)	2021	Univariate Logistic regression, adjusted logistic regression, RF	Logistic regression is suitable for healthcare for identifying cause-and-effect relationships when the outcome is categorical (high or low). RF does not directly provide cause-and-effect insights.
Mathematical Methods for Maintenance and Operation Cost Prediction Based on Transfer Learning in State Grid	(Guo, Wang, Zheng, & Ding, 2022)	2022	Support vector regression	Support vector regression is useful for cost prediction but less suitable for cause-and-effect analysis.
Gold Prices Time-Series Forecasting: Comparison of Statistical Techniques	(Maryati, Christian, & Paramita, 2023)	2023	Exponential smoothing, Linear regression	Linear regression directly applies to continuous variables, supporting its use in analysing cause-and-effect in contract costs.
Predicted Average Annual Cost of Crashes on the US-16 Wyoming Downgrade Using Time Series Analysis and Forecasting	(Ampadu, Ker, Wulff, & Ksaibati, 2024)	2024	MLR, Automated Integrated Moving Average (SARIMA)	It shows that MLR is suitable for continuous variables in analysing cause-and-effect if multiple independent variables are involved.
Predicting High Health-Cost Users Among People with Cardiovascular Disease Using Machine Learning and Nationwide Linked Social	(Nghiem, Atkinson, Nguyen, Tran-Duy, & Wilson, 2023)	2023	RF, K-Nearest Neighbour, L1-regularised logistic regression, Classificational regression, traditional regression scores.	All discussed methods are highly suited for categorical regression but not for cause-and-effect regression in this context.

Administrative Datasets				

Table A-8 Concept matrix regression methods

C. Analysed Variables

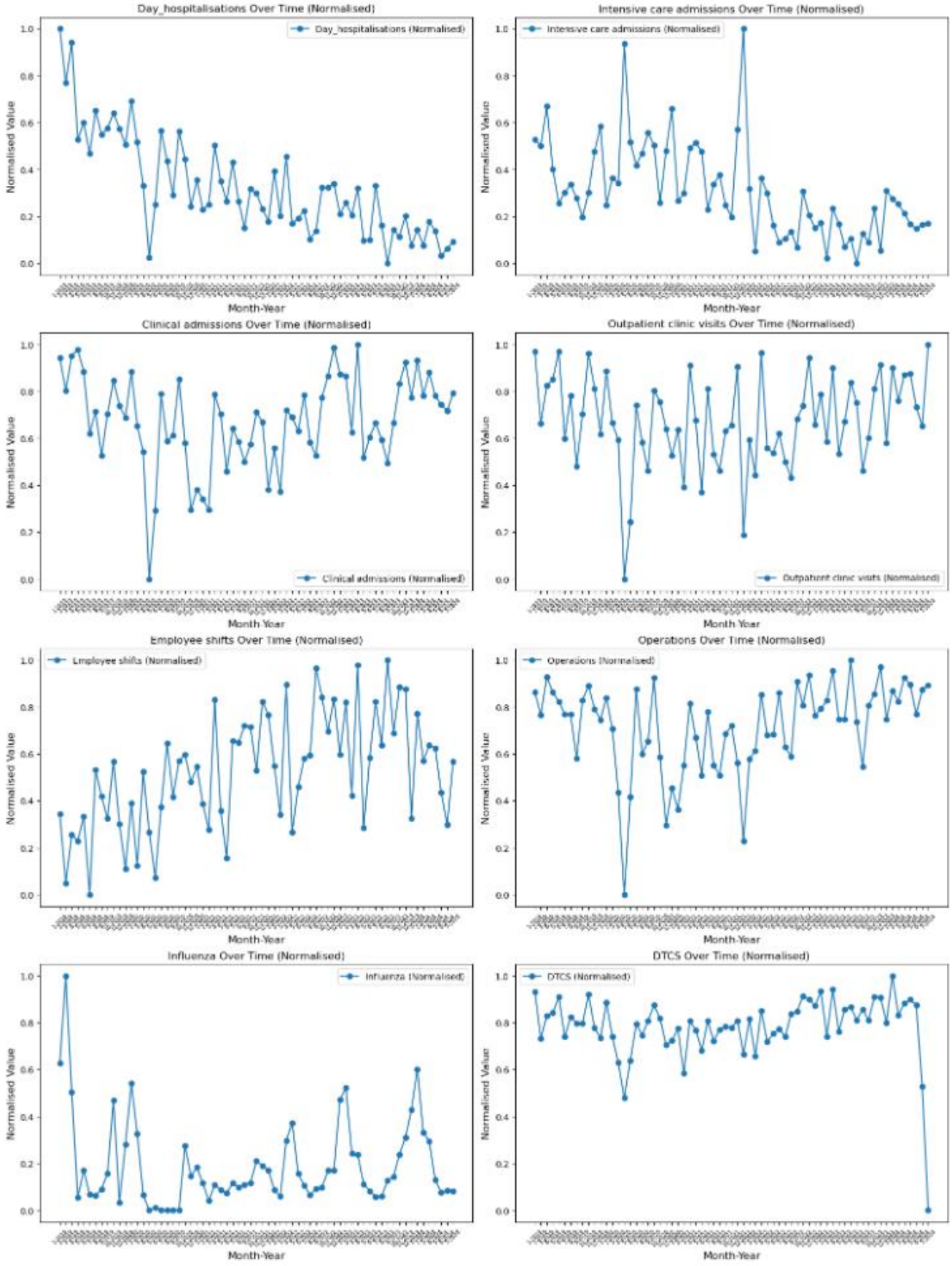

Figure A-3 Patterns independent variables normalised

D. Pearson Correlation Matrix

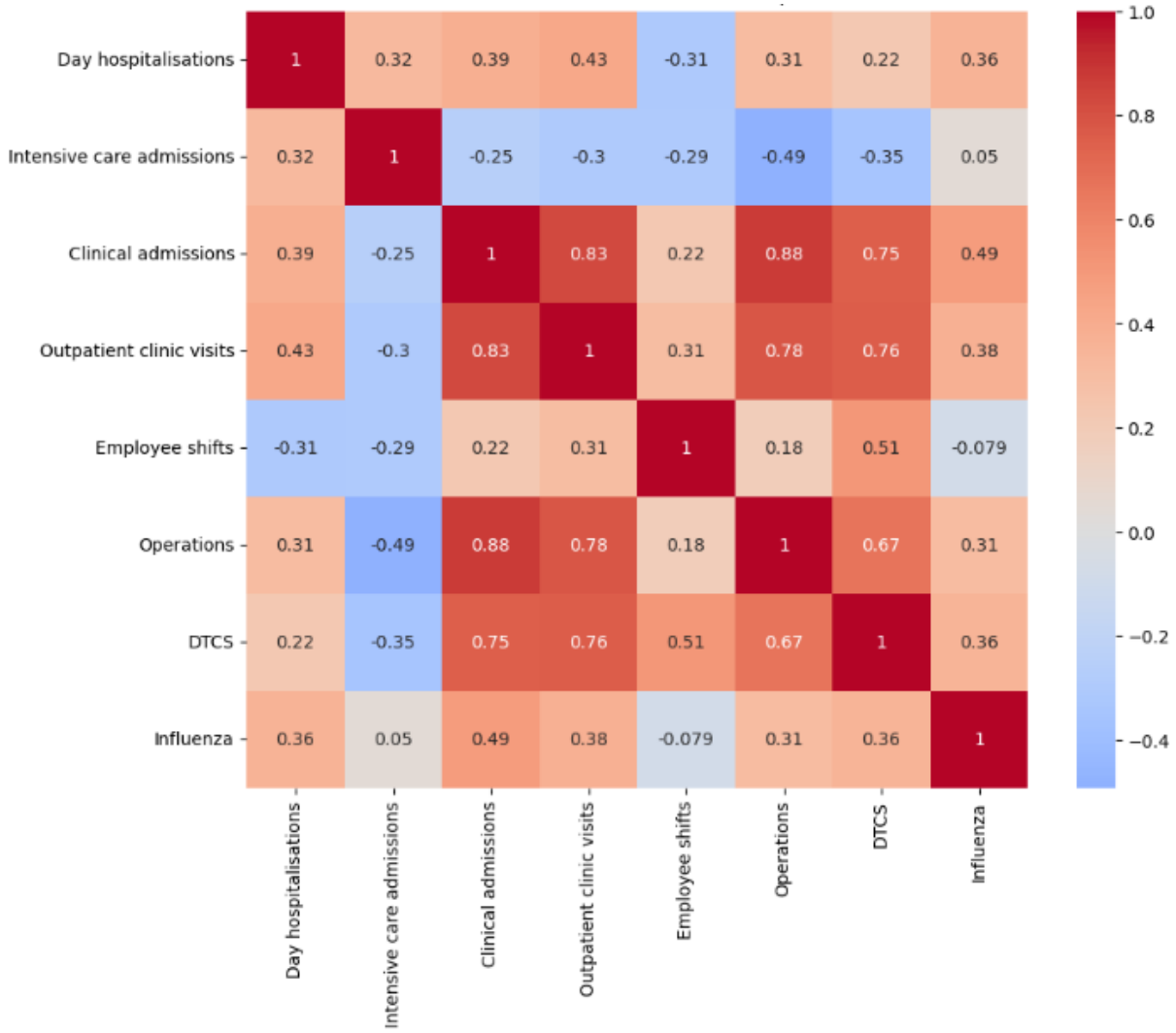

Figure A-4 Pearson correlation heatmap