Siiri Ojala

Predictive Modelling and AI Integration for Enhanced Analysis of Warranty and Notification Data

A Case Study in Manufacturing Data Analysis

School of Technology and Innovation
Master's Thesis in Industrial Management

Vaasa 2024

UNIVERSITY OF VAASA
School of Technology and Innovation
Author: Siiri Ojala
Title of the Thesis: Predictive Modelling and AI Integration for Enhanced Analysis of Warranty and Notification Data : A Case Study in Manufacturing Data Analysis
Degree: Master of Science
Programme: Industrial Management
Supervisor: Rayko Toshev
Year: 2024 **Pages:** 72

ABSTRACT:

This thesis explores the application of predictive modelling and AI integration in the analysis of warranty and notification data within a manufacturing company context. The research aims to develop a Power BI tool to analyze and visualize the relationship between warranty and notification data while utilizing predictive analytics. Design science methodology, specifically the Design Science Research Methodology (DSRM), is employed in the research process.

The research objectives are threefold: (1) to investigate how predictive models can be utilized to analyze warranty and notification data, (2) to identify factors in the warranty and notification data connected to warranty claims, and (3) to explore potential future applications of AI in analyzing warranty and notification data in the next 10 years.

Methodologically, logistic regression and time series analysis are employed to develop predictive models. Logistic regression is utilized to predict product claims, while time series analysis is used to visualize trends and offer forecasting options within Power BI. Additionally, AI-powered tools such as the key influencers in Power BI are utilized to analyze factors affecting product claims.

Key findings indicate a significant positive relationship between notification count and the probability of a product claim, validating the original hypothesis. Recommendations for the case company include investing in augmented analytics, democratizing AI within the organization, and prioritizing clear communication to ensure effective and ethical use of AI technologies.

While the research demonstrates the effectiveness of predictive modelling and AI integration in warranty and notification data analysis, several limitations exist. These include the inability to fully integrate the logistic regression model into Power BI and the focus primarily on the probability of product claims, leaving other potential areas unexplored.

In conclusion, this thesis marks a step towards transforming warranty data into a tool for risk management and product improvement. Future research directions include refining predictive models, exploring advanced AI techniques, and increasing the generalizability of findings across different industries.

KEYWORDS: Business intelligence, predictive modelling, machine learning, logistic regression

VAASAN YLIOPISTO
School of Technology and Innovation

Tekijä:	Siiri Ojala
Tutkielman nimi:	Predictive Modelling and AI Integration for Enhanced Analysis of Warranty and Notification Data : A Case Study in Manufacturing Data Analysis
Tutkinto:	Kauppatieteiden maisteri
Oppiaine:	Industrial Management
Työn ohjaaja:	Rayko Toshev
Valmistumisvuosi:	2024 **Sivumäärä:** 72

TIIVISTELMÄ:

Tässä tutkielmassa syvennytään ennustemallintamisen ja tekoälyn integroinnin vaikutuksiin takuu- ja notifikaatiodatan analysoinnissa valmistavassa yrityksessä. Tutkielman päätavoitteena on kehittää ennustemallintamiseen perustuva Power BI -työkalu, jonka avulla voidaan analysoida ja visualisoida takuu- ja notifikaatiodatan välistä suhdetta. Tutkimuksessa käytetään suunnittelutieteellistä metodologiaa, joka mahdollistaa järjestelmällisen lähestymistavan ongelmanratkaisuun ja uusien ratkaisujen kehittämiseen.

Tutkielman tavoitteita on kolme: (1) tutkia ennustemallien soveltamista takuu- ja notifikaatiodatan analysoinnissa, (2) tunnistaa takuu- ja notifikaatiodatasta tekijöitä, jotka vaikuttavat reklamaatioihin, sekä (3) tutkia mahdollisia tapoja hyödyntää tekoälyä takuu- ja notifikaatiodatan analyysiin seuraavan vuosikymmenen aikana.

Tutkielmassa ennustemallien luonnissa hyödynnetään logistista regressiota ja aikasarja-analyysiä. Logistista regressiota käytetään ennustamaan reklamaatioita, ja aikasarja-analyysiä hyödynnetään tulevan kehityksen ennustamiseen ja visualisointiin. Lisäksi työssä käytetään tekoälypohjaisia työkaluja, kuten Power BI:n tärkeimmät vaikuttajat -työkalua.

Tutkielmassa löydettiin merkittävä positiivinen suhde notifikaatiomäärän ja reklamaatiotodennäköisyyden välillä. Näiden tulosten perusteella case-yritystä suositellaan panostamaan entistä enemmän tekoälypohjaiseen analytiikkaan sekä lisäämään ja laajentamaan tekoälyosaamista organisaatiossa. Lisäksi selkeä ja avoin kommunikaatio on keskeisen tärkeää, jotta tekoälyn tehokas ja eettinen käyttö voidaan varmistaa.

Vaikka tutkimuksen tulokset osoittavat ennustemallintamisen ja tekoälypohjaisten työkalujen tehokkuuden takuu- ja notifikaatiodatan analysoinnissa, tutkimuksen rajoituksiin kuuluu logistisen regressiomallin epätäydellinen integrointi kehitettyyn työkaluun sekä keskittyminen pääasiassa tuotereklamaatioihin. Tulevaisuudessa tutkimusta voidaan laajentaa tarkastelemalla muita ennustemalleja ja tekoälytyökaluja sekä tutkimalla aihetta eri toimialoilla ja konteksteissa.

AVAINSANAT: Business intelligence, ennustemallinnus, koneoppiminen, logistinen regression

Contents

Figures

Tables

Algorithms

1 Introduction

This thesis investigates predictive modelling of quality data and examines the potential use of AI within a specific case company. In today's business landscape, organizations increasingly rely on business intelligence (BI) solutions to make informed decisions and enhance their operations. This involves optimizing data utilization to identify trends and patterns. In line with this trend, the case company is concentrating on refining its analysis of manufacturing and product quality data to gain deeper insights into issues related to product claims. With a focus on predictive modelling and the expanding role of artificial intelligence (AI) in BI, this thesis aims to address key questions about leveraging these technologies to improve decision-making and enhance product quality and operational efficiency within the company.

1.1 Background

Following the developments in data analysis and business intelligence, the case company of this thesis has invested significantly in improving their analysis and business intelligence know-how. One significant area of data analysis for the case company is the manufacturing and product quality data. The case company has utilized their notification and warranty data separately for many years, but they have never been combined in one analysis tool. Warranty data includes data about product claims and products under warranty. Notification data includes notifications opened to order lines in the ERP system. It consists of ZP notifications opened during the manufacturing process and Z2 notifications used for order changes. Significant benefits can be achieved by combining the two data sources, such as improved knowledge and awareness of the factors affecting product claims. This will give the case company opportunities to improve their processes and quality.

1.2 Case Company

The case company is a central player in the power and automation domain, showcasing a rich history of innovation and adaptation. With a global presence and a diverse portfolio of products and services, the company has positioned itself as a key contributor to the evolution of industrial processes and energy management. This thesis is done for one manufacturing unit of the company.

1.3 Purpose

The purpose of this thesis is to develop the utilization of warranty and notification data within the case company by creating a new Power BI tool for internal use. This tool will empower the organization to conduct more comprehensive analyses of warranty and notification data, facilitating the identification of patterns and trends crucial for enhancing product quality, reducing warranty costs, and ultimately, elevating customer satisfaction. With a primary focus on the development of predictive models and the exploration of AI applications over the next decade, this thesis aims to pave the way for more sophisticated data-driven decision-making processes within the organization.

To achieve these objectives, a pragmatic approach is adopted, integrating elements of design science research methodology (DSRM). DSRM, as conceptualized by Hevner et al. (2004), provides a structured framework for addressing practical challenges through the creation and evaluation of IT artifacts. In this study, DSRM key stages, from identifying the organizational challenge of inefficient warranty and notification data utilization to developing effective solutions are used. Drawing insights from interdisciplinary fields such as computer science, data science, management, and engineering, the methodology leverages literature in manufacturing analytics, business intelligence, predictive modeling, machine learning, and AI integration. The aim is to enhance operational efficiency and decision-making processes within the organization.

In this study, an effort is made to contribute to the ongoing evolution of BI and AI integration, addressing practical challenges encountered by organizations in optimizing their data assets. Through the adoption of a pragmatic approach and the application of design science research methodology, the aim is to achieve tangible improvements in product quality, warranty management, and customer satisfaction within the case company.

1.4 Research Questions and Limitations

The case company has a need for a decision support tool that can assist the quality department in directing their efforts towards areas of focus and pinpointing root causes of issues, thereby facilitating quality improvement actions. Therefore, the primary goal of this thesis is to develop a Power BI tool, which will provide the case company insights into the relationship between warranty and notification data. As the company has not previously combined the two data sources, this thesis will provide new information on the relationship of notifications and product claims therefore enhancing the understanding of their relationship and enabling more informed decision-making processes.

Additionally, the second aim of this thesis is to introduce and implement predictive modeling techniques within the case company's operational framework, which has not been previously explored. Despite the absence of prior utilization, there is a clear interest within the case company to delve into this domain. By integrating predictive modeling methodologies into the analysis of warranty and notification data, this research seeks to capitalize the potential for enhancing decision-making processes and improving operational efficiency.

To fulfill these objectives, the aim is to address the following three research questions:

> *RQ1. How can predictive models be used to analyze warranty and notification data?*
> *RQ2. What factors in the warranty and notification data are connected to product claims?*
> *RQ3. How can the case company utilize AI in analyzing warranty and notification data in the next 10 years?*

Connected to the second research question, *what factors in the warranty and notification data are connected to product claims*, the case company had initially speculated that the quantity of notifications associated with an order line might influence the likelihood of a warranty claim. Based on this speculation, hypotheses were formulated to formally test this relationship. The Null Hypothesis (H0) posits that there is no statistically significant association between the quantity of notifications for an order line and the probability of a warranty claim, while the Alternative Hypothesis (H1) proposes the presence of a statistically significant relationship between these variables:

> *Null Hypothesis (H0): There is no statistically significant association between the quantity of notifications for an order line and the probability of a warranty claim.*
>
> *H1: There is a statistically significant relationship between the quantity of notifications for an order line and the probability of a warranty claim.*

The answers to the research questions are provided through the research objectives which are a comprehensive literature review on the topics of BI, AI and predictive modelling, and data collection and analysis utilizing predictive analysis.

This study includes several limitations. The research is confined to warranty data from the case company's CRM system, and ERP data including production, order change, and quality notification data from 2020 to 2023. The limited data sources and their availability may impact the generalizability of findings to different contexts or time periods. The accuracy and completeness of the data collected depend on the recording practices of the case company, potentially introducing biases or gaps. Furthermore, the study is specific to the industry and operational context of the case company, which limits the applicability of some of the findings to industries with similar characteristics.

The predictive models utilized are subject to various constraints, including input data quality and algorithm performance. Additionally, the examination of AI utilization in the future is inherently speculative, with findings influenced by evolving technological landscapes and regulatory changes. Despite these limitations, the research contributes valuable insights, providing a foundation for future investigations and practical applications in the field of warranty and notification data analysis.

1.5 Structure

This introduction chapter is followed by six chapters described below, each offering a detailed exploration of key aspects related to the research topic.

Chapter 2 provides a comprehensive literature review centered around business intelligence and the incorporation of Artificial Intelligence within this field. This chapter delves into subjects like machine learning, the significance of AI in building trust, and the future directions of AI in BI. Moreover, it scrutinizes predictive modeling, encompassing data preprocessing and logistic regression.

Chapter 3 outlines the research methodology, detailing the approach taken to address the research questions.

Chapter 4 begins with a discussion on Power BI, explaining its components, workflow, and functionalities. Subsequently, the chapter transitions into the thesis's tool development process, encompassing data collection, analysis, and visualization methodologies.

Chapter 5 delves deeper into predictive modeling, providing a focused discussion on its application within the research context.

Chapter 6 presents the results, which include predictive modeling outcomes, factors associated with product claims, and future AI utilization possibilities.

Chapter 7 engages in a reflective discussion and draws conclusions based on the research findings. It reflects on the obtained results, discusses managerial implications, and evaluates encountered limitations during the study.

2 Literature Review

In this chapter, the key concepts of the thesis are discussed utilizing prior academic research. These concepts include business intelligence, augmented analytics, and their developments as well as machine learning. The possibilities of utilizing augmented analytics in a manufacturing company context are also discussed through existing case studies.

2.1 Business Intelligence and Augmented Analytics

Business intelligence as a term has been used especially since the 1990s, but its history lies in the information sharing and decision support systems (DSSs) developed in the 1960s and 1970s (Gurcan et al., 2023; Sharda et al., 2018). The main objective of these systems has remained relatively unchanged throughout the decades and can be summarized as “to enable interactive access to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analyses” (Sharda et al., 2018, p. 16). Today, BI systems have become a crucial part of data analysis and decision making in companies, and a “cornerstone of enterprise decision support” (Tamang et al., 2021, p. 64). According to Tavera et al. (2021), BI is one of the central factors for company development along with big data and IoT. BI is also considered a core technology of Industry 4.0.

According to Tripathi et al. (2023, p. 1) "the primary goal of business intelligence (BI) software is to help users quickly find and analyze data that is critical to making informed business choices in real time". In conclusion, BI enables users to use data efficiently, provide answers to questions (Nakhal et al., 2021, p. 14), and “generate useful information from high-dimensional data that may support making better informed decisions” (Khan, W. A. et al., 2020a, p. 165). Its benefits include “faster, better and much more accurate planning and analysis, improved data quality, oppositional efficiency, employee and customer satisfaction, reduced costs, increased revenues and much better business decisions“ (Tamang et al., 2021, p. 65). However, many companies are failing to capture these benefits and value from their BI systems (Ain et al., 2019).

Central functionalities in BI systems include a data warehouse, online analytical processing (OLAP) and dashboards (Ain et al., 2019). Other key functionalities include data mining, ad hoc queries, and scorecards. Today, AI-enabled BI tools are readily accessible, offering capabilities such as predictive modelling.

Due to its system-based nature, BI is a combination of several branches. It provides computer science-based decision-making for businesses with a combination of business, decision science and computer science (Purnomo et al., 2021; Gurcan et al., 2023).

BI can be divided into three generations. The first generation of BI systems supported relational data formatting and was limited to batch loading once a day. The second generation added analytics as a central part of BI systems and supported ad hoc queries as well as self-service analytics. Today, third-generation BI systems are used. They are centered around the idea of AI meeting BI and enable “various AI-powered analytics for BI-enabled decision making”. (Bulusu & Abellera, 2021, p. 3-4).

As defined by Bulusu and Abellera (2021), AI-powered analytics are a part of the third-generation BI systems in use today. Therefore, understanding AI, its technologies, and possibilities for use in the BI context is crucial. The Confederation of Finnish Industries (EK) has recently conducted a survey regarding data use and AI in their member companies (EK, 2023). The results indicate that AI is seen as a technology with the greatest potential for business in the near future. The most important benefits for the use of AI based solutions included increased operational efficiency, cost savings, improved quality, competitive advantage, and reduced personnel workload. While companies were extremely interested in AI applications, only 42 % were planning on investing on them, while 59 % were either not planning on investing or were unsure about it. The main requirements for increasing the utilization of data were the need for more insight into the utilization of data, new skills, and the need for more experts on the subject. The survey clearly indicates that the awareness for AI technologies for data utilization is relatively high, but companies would require more information and skilled employees to be willing to invest in the solutions.

There are several different definitions for AI, and a consensus regarding its definition has not been reached. Kaplan and Haenlein (2019, p. 15) define AI as “a system’s ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation”. Rahman (2020, p. 20) defines AI as “a computer-based system that simulates characteristics of the human brain in order to perform activities that could otherwise only be performed by humans”. According to Kananen (2019, p. 27), AI is at its core programming, mathematics, and statistics, based on matrixes, vectors, derivation and statistical probabilities. According to McCarthy (2007), AI “is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.”.

The third generation BI with AI features has brought many benefits to companies, such as additional AI powered computing tools and data analytics in a more advanced level. First and second-generation BI systems provided mainly descriptive analytics, whereas third generation BI systems have introduced predictive and prescriptive analytics to BI system users. Figure 1 presents the different levels of business analytics.

The combination of BI and AI is often called augmented analytics (AA). According to Alghamdi and Al-Baity (2022), “AA enhances analysis, reduces time, and supports data preparation, visualization, modelling, and generation of insights“. According to Prat (2019), AA can automate the entire analytics cycle from problem identification to analysis, model building and decision-making, as AI can be applied to all these tasks. For example, in the case of AI enabled predictive analytics, the advanced AI technology has been proven to improve the intensity of BI prediction analysis (Chen & Wang, 2022). Although AA can improve BI analytics significantly in many ways, Alghamdi and Al-Baity (2022) highlight, that "AI-driven analytics cannot fully replace human decision-making, as most business problems cannot be solved purely by machines. Human interaction and perspectives are essential, and decision-makers still play an important role in sharing and operationalizing findings.".

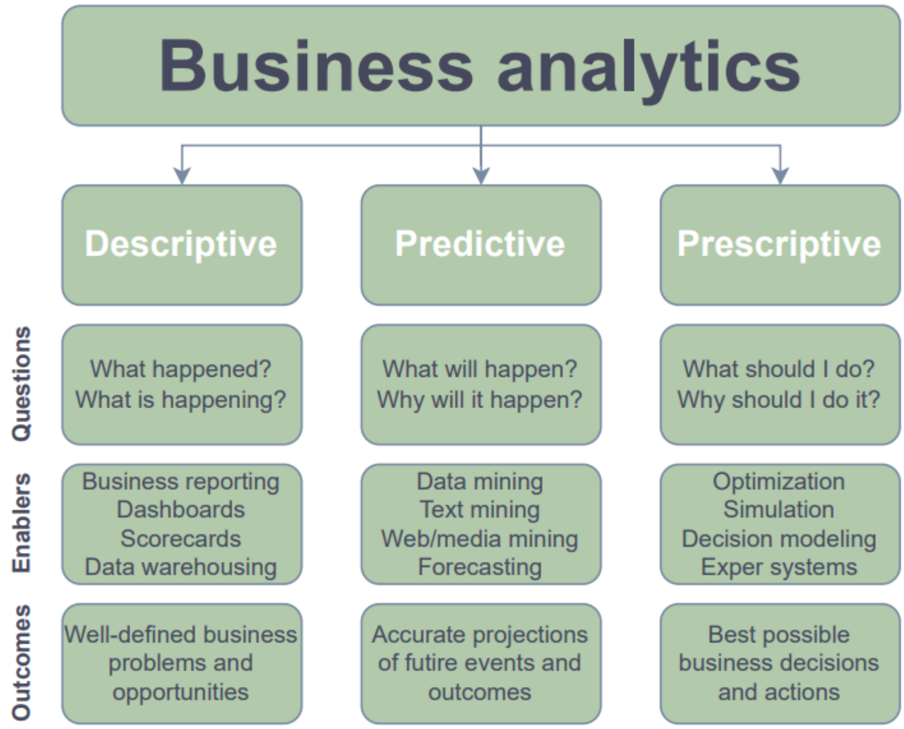

Business analytics
Descriptive
Predictive
Prescriptive
Questions
Enablers
Outcomes
What happened? What is happening?
What will happen? Why will it happen?
What should I do? Why should I do it?
Business reporting Dashboards Scorecards Data warehousing
Data mining Text mining Web/media mining Forecasting
Optimization Simulation Decision modeling Exper systems
Well-defined business problems and opportunities
Accurate projections of futire events and outcomes
Best possible business decisions and actions

Figure 1. Levels of business analytics (Sharda et al., 2018).

2.1.1 Augmented Analytics Implementation and Utilization

Today, the amount of data, especially complex data, is increasing rapidly in organizations. This has led to increased importance of augmented analytics and its data analysis abilities. Augmented analytics systems can handle large and complex data and consider factors which were previously too complex to consider (Rahman, 2020). According to M. A. Khan et al. (2020, p. 116013), modern BI “has a pivotal role in articulating a strategy and taking correct measures based on data”.

Due to its benefits and potential, companies should focus on effectively utilizing the augmented analytics systems. Figalist et al. (2022) have recognized that the utilization of AI in BI context is often left at a prototypical stage and not used for decision-making. In their article, the authors found five drivers for not using AI efficiently or at all in decision-making: low data quality, inability to cross cultural gap, ineffective prototypical analysis, inability to provide value, and lack of priority, time, and resources. They also give solutions to the problems in seven categories: data storage & model definition, data validation, collaboration, planning, explanation, translation, and costs. Additionally, they recommend choosing and prioritizing the different solution categories based on the stakeholders’ priorities.

When first considering implementing AI to a business analytics task, it should be considered whether the task at hand is suitable for prototyping. Akerkar (2019) recommends that the first task should allow for repetition, high volume, a pattern, and low cost of mistakes. Ensuring that the task is suitable for testing AI implementation will positively affect how future AI implementation tasks will be viewed in the organization.

To help companies evaluate the need for AI utilization for a business problem, Kananen (2019) has developed an AI Business Model Canvas (Figure 2). The canvas consists of a section for defining the goals and objectives as well as a section for evaluating the AI operation. The canvas is useful for evaluating whether the problem should be solved using AI or not as well as for documenting what exactly is wanted from the project.

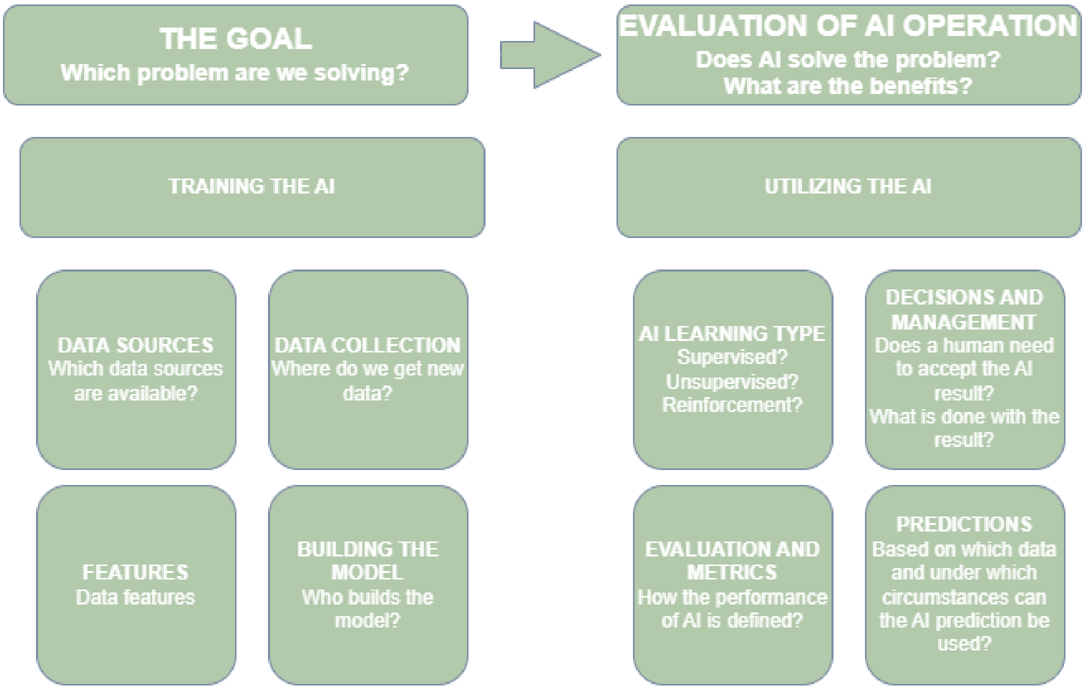

THE GOAL
Which problem are we solving?
EVALUATION OF AI OPERATION
Does AI solve the problem?
What are the benefits?
TRAINING THE AI
UTILIZING THE AI
DATA SOURCES
Which data sources are available?
DATA COLLECTION
Where do we get new data?
AI LEARNING TYPE
Supervised?
Unsupervised?
Reinforcement?
DECISIONS AND MANAGEMENT
Does a human need to accept the AI result?
What is done with the result?
FEATURES
Data features
BUILDING THE MODEL
Who builds the model?
EVALUATION AND METRICS
How the performance of AI is defined?
PREDICTIONS
Based on which data and under which circumstances can the AI prediction be used?

Figure 2. AI Business Model Canvas (Kananen, 2019).

Although the potential of using augmented analytics solutions is high, it is still recommended to not use it in every analysis. Big data analysis is challenging using traditional BI tools, but traditional tools are still applicable to smaller datasets. Additionally, AI tools can be utilized only for a certain task in the analysis such as visualization, while other parts can be done more efficiently utilizing traditional BI. (Alghamdi & Al-Baity, 2022).

Research in the use of AA in companies have shown that AA can be beneficial in analyzing many types of data. M. A. Khan et al. (2020) utilized AA in the context of demand forecasting for businesses. The focus was on predicting future demand using ML techniques applied to sales data. The results indicated that AA is beneficial for efficient management planning and operational excellence. while also improving efficiency, reducing costs, and predicting market trends.

Saadat et al. (2022) applied AA in the manufacturing domain, particularly in improving quality control processes. Their study emphasized predicting test failure and associated error messages using ML techniques applied to product configuration data. The outcomes underscored the significant benefits of AA in enhancing manufacturing operations, including more accurate quality control, streamlined management planning, and operational excellence. Additionally, the approach offered insights for reducing costs and anticipating market trends, contributing to improved efficiency and product reliability in manufacturing processes.

Hamzehi and Hosseini (2022) researched the implementation of AA techniques in a pharmaceutical company's operations, focusing on enhancing sales and customer segmentation strategies. They applied RFM and LRFM models along with clustering algorithms like K-means, DBSCAN, and Optics to analyze transaction data. The study demonstrated the efficacy of machine learning algorithms in optimizing sales performance and improving product distribution. By leveraging business intelligence insights, tailored strategies for each customer cluster were devised, leading to increased profitability and market responsiveness.

2.1.2 Trust and Transparency in Augmented Analytics

When AI is utilized in real-life scenarios, such as in the analysis of manufacturing data through augmented analytics, a critical concern arises: can AI and its outcomes be trusted? In traditional task completion, individuals possess a clear understanding of the decision-making processes that lead to specific results. However, when engaging with AI, these processes often remain opaque, leading to skepticism. This phenomenon, known as the black-box problem, has been underscored by numerous instances of AI delivering suboptimal results. For instance, biased recruiting systems have perpetuated exclusionary practices, resulting from biased training data and inadequate mitigation strategies. Such incidents have decreased trust in AI systems, highlighting the urgent need for transparency and accountability in AI-driven processes.

The APA Dictionary of Psychology defines trust as “reliance on or confidence in the dependability of someone or something” (American Psychological Association, n.d.). Trust in AI systems consists of several factors, which should be considered when using augmented analytics systems. As AI is becoming more and more implemented in different areas of society, organizations such as International Organization for Standardization (ISO) and the European Union have developed standards and guidelines on AI trust and trustworthiness. While AI trust refers to the confidence which users, stakeholders, and the public have in the capabilities, reliability, and overall performance of an AI system, AI trustworthiness refers to the qualities and characteristics of an AI system that make it worthy of trust. The terms are closely interlinked, and both should be prioritized to achieve either or both.

The European Commission has released Ethics Guidelines for Trustworthy AI (2019), which discusses the dimensions and requirements for trustworthy AI systems. According to the report, trustworthy AI should be lawful, ethical, and robust. ISO and International Electrotechnical Commission (IEC) have also developed their technical report on AI trust and trustworthiness, ISO/IEC TR 24028:2020 Artificial intelligence - Overview of trustworthiness in artificial intelligence (2020), where trust is said to be established by fairness, transparency, accountability, and controllability of the system.

Based on existing research and the EU and ISO guidelines on trustworthy AI, Kaur et al. (2023) have developed a framework for trustworthy AI (Figure 3). The baseline for the framework is on the three guidelines of the EU Ethics Guidelines for Trustworthy AI (2019): lawful, ethical, and robust. This is further divided into four categories: respect for human control, prevention of harm, fairness, and explicability. The framework can be utilized both during the development and use of AI systems. To operationalize the framework, companies should periodically assess their AI systems against these dimensions, prioritizing ongoing education for practitioners, implementing regular audits, and incorporating feedback mechanisms for continuous improvement. By adhering to these guidelines, companies can reinforce the ethical foundations of their AI applications and increase transparency, accountability, and user trust in an era increasingly defined by artificial intelligence.

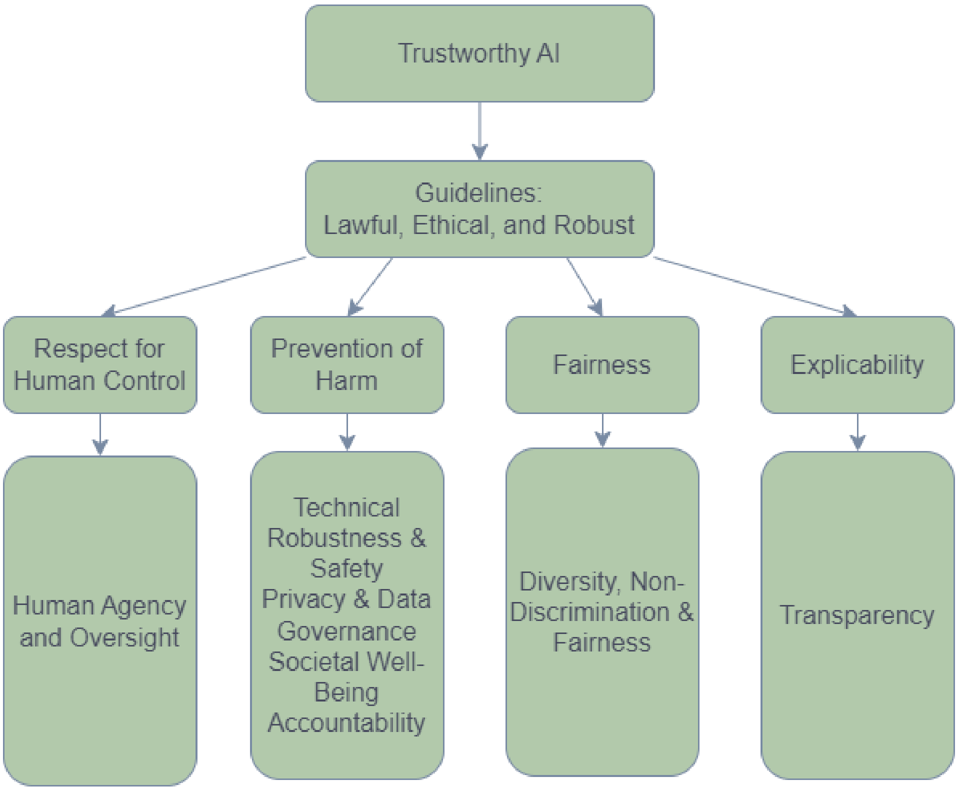

Trustworthy AI
Guidelines: Lawful, Ethical, and Robust
Respect for Human Control
Prevention of Harm
Fairness
Explicability
Human Agency and Oversight
Technical Robustness & Safety
Privacy & Data Governance
Societal Well-Being
Accountability
Diversity, Non-Discrimination & Fairness
Transparency

Figure 3. Trustworthy AI framework (Kaur et al., 2023).

Hasija and Esper (2022) have made qualitative research on AI technology acceptance in the field of supply chain management focusing on trust. Based on the results, they give several recommendations for AI adoption and implementation. They recommend, that managers should pay attention to pre-deployment communication and how AI systems are promoted and introduced internally to ensure a trusting environment, which will also improve the probability of correct use of the system. In addition, they recommend additional retraining for upkeeping the trust and ensuring successful use of the system in the future.

In conclusion, trust and transparency are foundational principles that underpin the successful adoption and implementation of AI technologies in augmented analytics. By adhering to ethical guidelines and fostering a culture of transparency, organizations can build trust among stakeholders and enhance the effectiveness of AI systems, such AA systems. Moving forward, prioritizing trust and transparency will be essential in navigating the evolving landscape of augmented analytics and harnessing the full potential of AI technologies.

2.1.3 Future Trends and Prospects in Augmented Analytics

Since the introduction of third generation BI systems in the 21st century, the AI aspect of BI has become more and more employed. Industry 4.0 tools are becoming more widespread, and among them the interest for utilizing AI also in the BI context is increasing. However, there are still a limited number of AI features available, and their adoption varies highly between organizations and industries. There is demand for the development of more AI tools for BI systems.

Tavera Romero and Rios Prado (2021, p. 8) assert that BI is not only a crucial business function but also "a requirement for later phases of industrialization". Moreover, a high level of interest in data analysis using AI has been demonstrated, as shown in the survey by the Confederation of Finnish Industries (2023). For example, the demand and interest towards predictive analytics are increasing. BI and data analytics are seen as a development pillar for companies due to their supporting role in decision-making, forecasting, and economy (Tavera Romero & Rios Prado, 2021).

As mentioned previously, the black-box problem affects trust for AI systems and therefore the implementation rate of AI can be affected negatively. One solution for the problem is explainable AI (XAI), which is being researched and developed at an increasing pace. According to Barredo Arrieta et al. (2020, p. 85), XAI is “one that produces details or reasons to make its functioning clear or easy to understand”. In an article by Meske et al. (2022) the objectives of XAI are generalized into 5 areas: explainability to evaluate AI, explainability to improve AI, explainability to learn from AI, explainability to justify AI, and explainability to manage AI (figure 4).

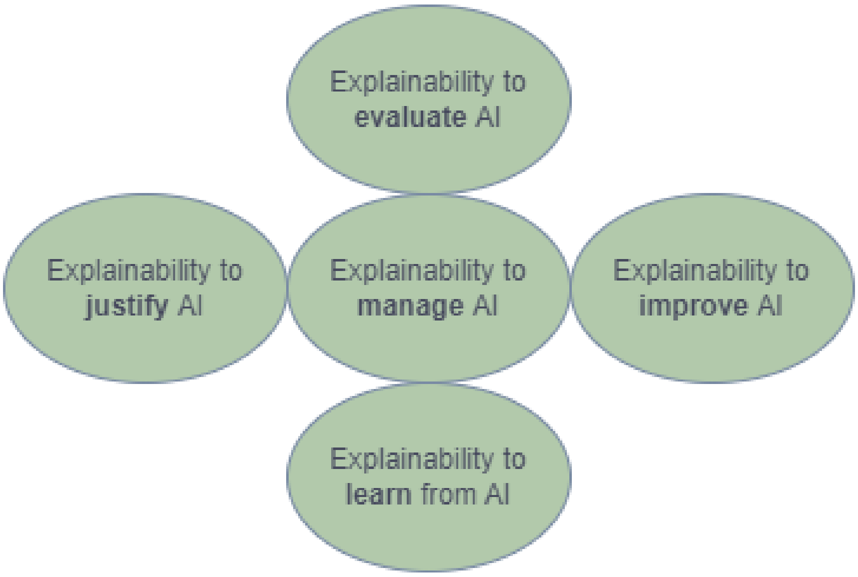

Explainability to evaluate AI
Explainability to justify AI
Explainability to manage AI
Explainability to improve AI
Explainability to learn from AI

Figure 4. Objectives of XAI (Meske et al., 2022).

XAI is central in the context of augmented analytics as it provides transparency into the decision-making processes of complex AI models, enabling users to understand and trust the insights generated. According to Li & Gregor (2011) "tools, which have enhanced explanatory facilities and provide justifications at the end of the consultation process, lead to improved decision-process satisfaction and decision advice transparency, subsequently leading to empowering effects like a higher sense of control and a lower perceived power distance”.

XAI plays a part in the future of AI-enhanced BI, projected to be AI-powered cognitive computing. This cognitive computing, a branch of AI, utilizes ML algorithms to mimic human-like cognitive functions. In third-generation BI, AI finds the right results, while cognitive computing prioritizes achieving optimal results through evidence-based reasoning. (Bulusu & Abellera, 2021, p. 210-211; Bousdekis et al., 2023).

Cognitive computing within BI will revolutionize user experience. Through natural language processing implementation, it ensures an agile interface. Additionally, it prioritizes trustworthiness and transparency through XAI. By utilizing XAI methodologies, cognitive computing enhances trust and transparency. Users understand decisions and trust insights provided. (Bulusu & Abellera, 2021, p. 210-211).

This holistic approach to AI-enhanced BI can incorporate AI-powered Chatbots. These Chatbots complement BI systems by providing natural language interfaces. They assist users in formulating queries, offering personalized insights, and automating routine tasks. Continuously improving through machine learning techniques, they streamline operations and empower users with a deeper understanding of their data. (Azmi et al., 2023).

Moreover, this integration unleashes the potential of advanced analytics. Leveraging deep learning, neural networks, and intelligent data mining, it extracts meaningful insights and patterns. Overall, the integration enhances the sophistication and utility of analytics, making BI systems more intelligent, responsive, and aligned with organizations' evolving needs. (Bulusu & Abellera, 2021, p. 210-211; Prat, 2019; Bousdekis et al., 2023).

2.2 Machine Learning in Augmented Analytics

The term machine learning is often used interchangeably with AI in day-to-day conversations. However, ML is in fact a subset of AI, which also includes other tools such as neural networks and computer vision. As with the definition for AI, a consensus regarding the definition for ML has not been reached either and depends on the context it is being discussed in. According to Lee (2019, p. 1), "Machine learning is a collection of algorithms and techniques used to design systems that learn from data. These systems are then able to perform predictions or deduce patterns from the supplied data". ML has also been defined as “the ability for a machine to improve what it does over time based on past results” (Rahman, 2020, p. 72), as “computational methods using the experience to improve the performance or to make accurate predictions” (Akerkar, 2019, p. 19), and as “a series of mathematical manipulations performed on important data in order to gain valuable insights” (Akerkar, 2019, p. 19).

ML systems are designed to use one of three learning types: supervised learning, unsupervised learning, or reinforcement learning. In supervised learning, the system is given the correct answers on the training dataset, which will help the system to give the correct answers when the system is given new data in real use after the training. In unsupervised learning, correct answers are not given to the system, but the system looks for patterns and other relevant information in the data and tries to answer the questions based on the observations. Reinforcement learning (or semi-supervised learning) doesn’t focus on the wrong or right answers for individual data points, but the system tries to find the overall result for a larger set of data. (Rahman, 2020, p. 74-80).

In augmented analytics, ML is used for decision making processes and for creating models on data to gain valuable knowledge. The popularity and development of ML technologies has led to increased intensity of BI prediction analysis as the predictions have become more effective and precise (Chen et al., 2022).

Predictive modeling stands out as one of the most prominent topics in augmented analytics. It involves using statistical algorithms and ML techniques to analyze historical data and make predictions about future outcomes or trends. Since 2014, predictive modelling has become a highly researched topic within BI research, and the trend is predicted to continue for the next few years (Chen & Wang, 2022). Although predictive analytics and modelling are often powered by AI, it isn’t always the case. When AI is used in the creation of predictive models, it is often called intelligent analytics. AI enables efficient, smart, and modern predictive modelling. Without AI, predictive modelling is an extremely slow process (Akerkar, 2019).

In day-to-day conversations, the words prediction and forecast are used interchangeably. However, there is a significant difference in the meaning of the two words in analytics. A prediction is a statement about a future event or outcome based on observation, experience, or analysis. A forecast is a specific type of prediction that is often associated with quantitative or data-driven analysis, where a time-dependent variables has been used in their creation. Predictive modelling is a more general term that can be applied to a wide range of prediction tasks, while forecasting specifically refers to predicting future values of time-dependent variables. Both involve the use of data and statistical methods, but the application and focus can differ based on the specific context and goals of the analysis. (Rahman, 2020).

Predictive modeling encompasses three key types: classification, regression, and time series analysis. In classification, algorithms like decision trees, neural networks, logistic regression, and naïve Bayes are employed to categorize data into predefined classes or labels. Regression analysis utilizes algorithms such as linear regression, and regression trees to predict continuous outcomes based on input variables. In contrast, time series prediction focuses on forecasting future values based on historical data patterns, using methods like averaging and exponential smoothing. It's noteworthy that most predictive modeling algorithms operate within the framework of supervised learning, where models are trained on labeled data with known outcomes to make predictions on new data points. (Sharda et al., 2018).

Although AI brings significant benefits to predictive modelling, such as enhanced productivity, challenges such as data quality, biases, and the need for human judgement in decision making remain. These challenges present research opportunities for improving the veracity, transparency, and governance of predictive models within the context of AI-powered analytics. (Prat, 2019).

2.2.1 Data Preprocessing

Data preprocessing holds a critical position in the realm of machine learning (ML) analysis, especially within the scope of augmented analytics. It serves as a fundamental preparatory phase, contributing significantly to the accuracy and efficacy of the subsequent analysis. This phase involves the meticulous preparation and transformation of raw data to address various challenges such as missing values, outliers, and noise, ensuring that the dataset is primed for analysis. By cleaning, standardizing, and organizing the data, the objective is to establish a refined foundation upon which ML models can extract meaningful patterns and insights, thereby enabling more robust and reliable predictions. Essentially, data preprocessing serves as the cornerstone for building accurate and efficient ML models by optimizing the quality and relevance of the input data (Sharda et al., 2018; Khan, W. A. et al., 2020b).

The complexity of data has been escalating in tandem with technological advancements, underscoring the growing importance of data preprocessing (Khan, W. A. et al., 2020b). In the context of augmented analytics, where data integration and analysis are augmented by machine learning and AI capabilities, data preprocessing assumes even greater significance. In most scenarios, data preprocessing can consume a substantial portion of the total time allocated for creating ML models, owing to the inherent incompleteness, noise, and inconsistency of real-world data (Sharda et al., 2018).

Data preprocessing typically comprises four crucial steps, the successful completion of which results in well-structured data (Sharda et al., 2018). Figure 5 illustrates these steps along with their key tasks. Inadequate understanding or improper utilization of preprocessing and transformation techniques can lead ML algorithms to draw inaccurate conclusions (Khan, W. A. et al., 2020b).

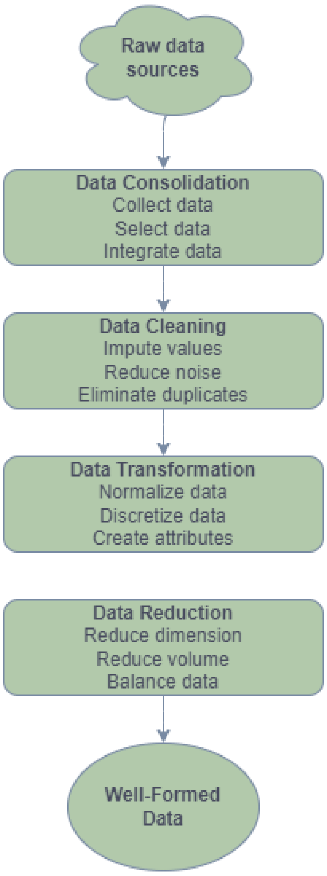


Figure 5. Data preprocessing steps (Sharda et al., 2018).

In the data consolidation step, relevant data is gathered, necessary variables are selected, and the data is integrated. Clarity regarding the objectives of the modeling process is essential before commencing consolidation to ensure the collection and integration of appropriate data (Sharda et al., 2018).

Data cleaning involves addressing erroneous or missing values and rectifying inconsistencies within the dataset to enhance its accuracy and reliability. This process encompasses tasks such as handling duplicates, imputing missing data, and correcting errors to prepare the dataset for analysis and modeling. Furthermore, noise in the data is mitigated by identifying and smoothing out outliers. Various methods such as statistics-based outlier detection, distance-based outlier algorithms, and density-based outlier detection can be employed for this purpose (Alexandropoulos et al., 2019; Sharda et al., 2018).

The data transformation stage entails applying diverse algorithms and techniques to convert the data into suitable formats for ML. This may involve processes like encoding or standardization. Normalization of data, which includes scaling and standardizing numerical features within a dataset to a consistent range, is also performed to prevent certain variables from disproportionately influencing ML models. Common methods include min–max normalization, z-score normalization, and discretization, where continuous data or variables are converted into discrete intervals or categories to simplify analysis or facilitate the application of certain algorithms (Alexandropoulos et al., 2019; Sharda et al., 2018).

In the final stage of data reduction, the dimensions and volume of data are reduced, and the data is balanced to simplify the dataset while retaining essential information and patterns. This aids in making the dataset more manageable and efficient for analysis, particularly beneficial when dealing with large datasets, as it reduces processing time and resource requirements without compromising the integrity of the analysis. Techniques such as over-sampling the underrepresented class, under-sampling the overrepresented class, or combining both over-sampling and under-sampling can be considered to balance the reduced data (Alexandropoulos et al., 2019; Sharda et al., 2018).

Given the complexity of data preprocessing tasks in ML modeling, automated data preprocessing tools have been developed, such as Atlantic (Santos & Ferreira, 2023). These tools streamline and automate preprocessing, simultaneously enhancing efficiency and effectiveness, thereby improving the performance of ML models applied to the data later. Atlantic, an open-source Python package, aims to autonomously determine the optimal combination of preprocessing mechanisms tailored to a given dataset, with the overarching goal of enhancing the performance of subsequent ML models applied to that data.

2.2.2 Logistic Regression

Logistic regression is one of the classification methods used in predictive modelling and ML predictive analytics. It is a statistical method used to model relationships between dependent variable (target variable) and independent variables (predictor variables). The model is created by fitting a logistic curve to the observed data. The logistic curve or Sigmoidal curve has an S-shape, where the curve starts with a gentle upward slope, then increases more steeply in the middle, and finally levels off as it approaches an asymptote. Logistic regression is suitable for use in cases, where the dependent variable is of binary nature, such as a does the patient have complications or not within 24 hours of the surgery. The goal of logistic regression is to predict the probability of an event belonging to one of two classes. (Pampel, 2021).

The core of logistic regression is the logistic function (sigmoid function). The logistic function transforms independent variable values X into probabilities through a nonlinear transformation, where the values X are mapped into a value between 0 and 1. These values can be interpreted as the probability of the event occurring. (Pampel, 2021).

The logistic function is defined as

$$P(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}, \qquad (1)$$

where

$P(Y = 1|X)$ is the probability of the event Y occurring given the input X,

$X_1, X_2, \ldots, X_n$ are the independent variables,

β_0 is the intercept when $X = 0$,

$\beta_1, \beta_2, \ldots \beta_n$ are the coefficients representing the effects of the independent variables on the outcome,

e is the base of the natural logarithm.

In logistic regression, the coefficients and intercept are calculated using optimization techniques such as gradient descent or maximum likelihood estimation (MLE). The goal is to find the values of coefficients that best fit the observed data and maximize the likelihood of observing the actual outcomes given the predictor variables. These coefficients are then used to make predictions on new data and calculate the probability of the analyzed event happening. Coefficients in logistic regression can range from negative to positive infinity. The interpretation of coefficients is based on their impact on the log odds of the event occurring, which is transformed into probabilities using the logistic function. A positive coefficient indicates that an increase in the value of the independent variable is associated with an increase in the log-odds of the event occurring, while a negative coefficient indicates the opposite. (Pampel, 2021).

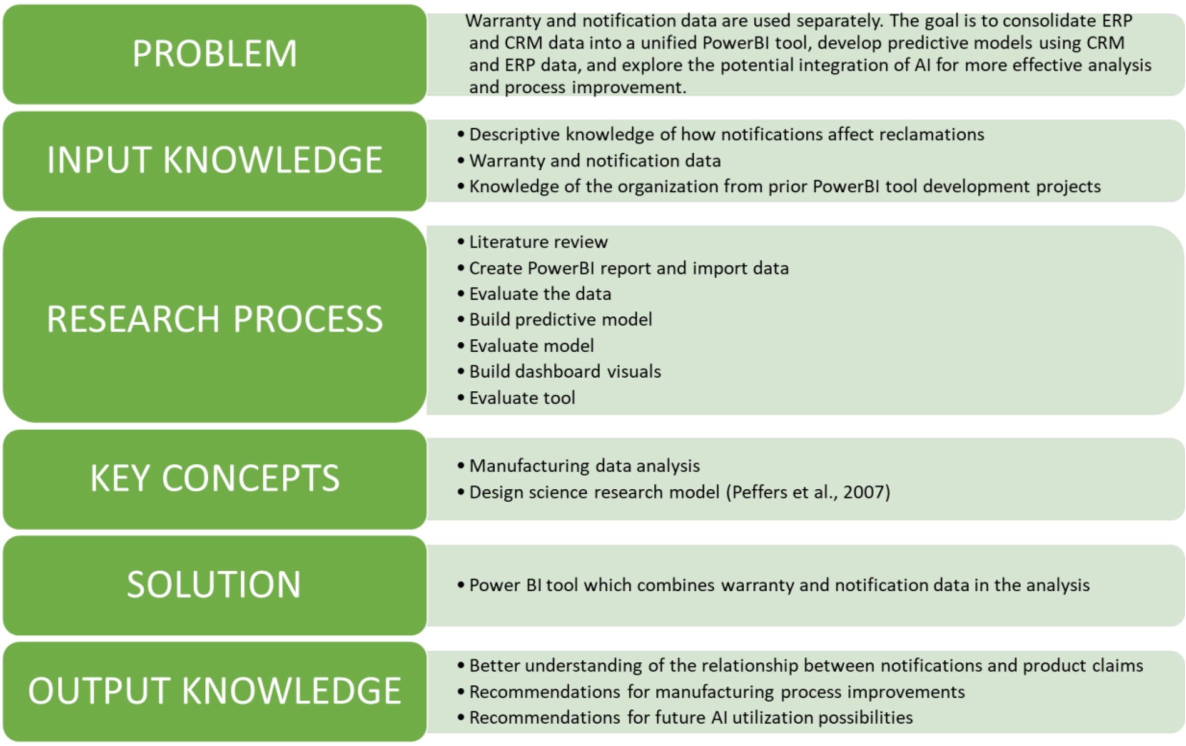
3 Research Method

Since the aim of the thesis is to develop a Power BI tool and research the relationships of warranty and notification data utilizing predictive analytics, design science (DS) was chosen as the research method. Design science "creates and evaluates IT artifacts intended to solve identified organizational problems" (Hevner et al., 2004, p. 77). Peffers et al. (2007), have combined research of DS in IT into a DS research methodology (DSRM), which was used in this study. The DSRM consists of six activities discussed in this study's context in the following paragraphs.

The first activity in the DSRM is problem identification and motivation. In this part, the problem was identified as inefficient and separate use of warranty and notification data. The resolution to the issue was determined to be a Power BI report, which could improve the data use, and the understanding of the two data sources and their relationship as well as offer insights to process improvements. To map out the research problem and process more, the DSR grid developed by vom Brocke et al. (2020) was utilized (Figure 6).

In the second activity of the DSRM framework, the objectives for the solution are defined. The objectives were defined as follows. Firstly, to combine notification and warranty data in one Power BI file; and subsequently, to construct a Power BI report which would serve as a platform for data analysis. The data analysis should utilize predictive modeling methods. The solution should be user-friendly, provide valuable information on product claims and notifications, and be adaptable for future improvements following the needs of the organization.

After this, a literature review was conducted to gather recent and relevant scientific information on the research topic. The literature review was conducted using the Tritonia Finna database. Relevant keywords such as manufacturing analytics, business intelligence, predictive modelling, machine learning, and AI integration were used to search for articles and other literature. The literature used in the review were published between years 2004 and 2023.



PROBLEM	Warranty and notification data are used separately. The goal is to consolidate ERP and CRM data into a unified PowerBI tool, develop predictive models using CRM and ERP data, and explore the potential integration of AI for more effective analysis and process improvement.
INPUT KNOWLEDGE	• Descriptive knowledge of how notifications affect reclamations • Warranty and notification data • Knowledge of the organization from prior PowerBI tool development projects
RESEARCH PROCESS	• Literature review • Create PowerBI report and import data • Evaluate the data • Build predictive model • Evaluate model • Build dashboard visuals • Evaluate tool
KEY CONCEPTS	• Manufacturing data analysis • Design science research model (Peffers et al., 2007)
SOLUTION	• Power BI tool which combines warranty and notification data in the analysis
OUTPUT KNOWLEDGE	• Better understanding of the relationship between notifications and product claims • Recommendations for manufacturing process improvements • Recommendations for future AI utilization possibilities

Figure 6. DSR grid

After the literature review, the third activity of the DSRM process, design and development, was done applying the knowledge gained from the literature review. The Power BI tool was developed by collecting and uploading the data, establishing the relationships between the datasets, and conducting analysis using Python and Power BI. Analysis was visualized using graphs, mainly line charts with time series forecasts. Based on the data analysis results about the influence of the notification count on product claims, it was decided that it should be investigated further using predictive modelling.

The predictive analysis was conducted separately by exporting a binary yes/no column for whether the product has been claimed, and columns with counts for two different types of notifications (Z2 and ZP) into an Excel file. The table had one row for each product under warranty at the time of the analysis. Then a Python code (Appendix 1) utilizing logistic regression from the Scikit-learn package was ran. The results were uploaded back to Power BI and visualized in a dashboard.

Following the fourth activity, demonstration, the predictive model accuracy was tested to ensure that the results of the tool can be trusted. This included checking the model accuracy measure to evaluate the performance of the model on the dataset. In addition to the model accuracy value, a confusion matrix was used to demonstrate and evaluate model performance.

In the fifth activity, evaluation, the report was tested by several future users of the report to ensure its effectiveness to solve the research problem. Following their comments, the report usability and visuals were improved to improve user experience.

In the last activity, communication, the tool, its background, and functionalities were presented to all possible future users in the case company.

4 Tool Development

In this section, the development process of the main tangible outcome of this thesis, the Power BI tool, is discussed. The process consisted of three parts: data collection, analysis, and visualization, which are discussed in detail in the following subsections after Power BI introduction.

4.1 Microsoft Power BI

Power BI is a business intelligence tool released in 2015 by Microsoft. The tool enables users to transform raw business data into information which can be used in decision making. According to Microsoft, Power BI can turn "unrelated sources of data into coherent, visually immersive, and interactive insights" (Microsoft, 2023). The most central feature of the software is the powerful visualization possibilities it offers for the reports. Power BI allows users to connect to various data sources, clean and transform data, and create interactive reports and dashboards.

4.1.1 Power BI Components

According to Sinha (2021) Power BI consists of seven key components. Each component or feature plays a distinct role, collectively empowering users to connect, model, visualize, and share insights seamlessly across the Power BI ecosystem.

First and most central component of Power BI is the Power BI desktop application. It is the primary authoring tool in the Power BI environment, facilitating data connection, shaping, code creation, and visualization.

Second component is power query, which serves as a data connectivity and preparation tool essential for dashboard creation within the Power BI desktop. Users can connect, consolidate, and transform data from numerous sources, accessible through the Get Data option in the Power BI desktop ribbon.

Third component is Power Pivot, the calculation engine in Power BI which models table relationships and executes calculations using Data Analysis Expressions (DAX). DAX is a formula language used in Power BI, Excel Power Pivot, and SQL Server Analysis Services for creating custom calculations and aggregations in data models.

The fourth component, Power View, is the visualization technology for creating the visuals of the report. These include different types of charts, tables, maps, and other visualizations.

Fifth component, The Power BI Service, is a cloud-based platform which enables the publishing of reports and datasets from Power BI desktop. It also enables data refreshing and collaborative use of the reports.

The sixth component, Power BI Report Server, operates within an organization's on-premises environment, functioning as a secure platform for publishing, managing, and accessing Power BI reports and datasets. It allows organizations to keep their data within their own firewall, providing a solution for businesses that prioritize on-premises data management and security.

Finally, the seventh component of the tool is the Power BI Mobile App, which allows users to easily view visualizations and data on their mobile devices. The mobile app has become increasingly important as the use of mobile devices in business has increased significantly.

4.1.2 Power BI Workflow

The common workflow of creating a Power BI report consists of three phases. First, the data is uploaded and connected in Power BI desktop using power query and power pivot. This creates a new report which is then built and developed in the desktop. The most central component during this phase is the power view, which enables users to create visuals such as tables and different types of charts for the report. In the third phase, the report is published to the Power BI service platform, where users can access and use the report from their computers or mobile devices.

4.2 Data Collection

As the first step of building the tool, a Power BI report was created in the Power BI desktop environment. Then, the required data was uploaded from the case company cloud storage service to the report. The goal was to integrate warranty and notification data, necessitating the retrieval of data from two distinct sources stored in the company's cloud storage.

Next, the data was consolidated and the relationships between different variables in the data sources were established utilizing power query. Due to differences in the processing level of the data in the different data sources, this step was essential to ensure effective analysis and trustworthy analysis results. The case company processes notification data at the order line level, while handling warranty data at the serial number level. Consequently, to establish relationships between data of two different levels, additional variables containing order line information had to be created to some of the tables.

4.3 Analysis

After the data was uploaded to the report and the relationship between different tables were established, the data was analyzed to discern noteworthy patterns and insights, paving the way for a more in-depth exploration. The hypothesis the case company had about the relationship between notifications and warranty claims was that as the number of notifications increased for an order line, the likelihood of a warranty claim also increases.

This was tested in Power BI first by creating a table which displayed the number of notifications for products which had a warranty claim in the company CRM system. The table was made more intuitive to read by sorting the claims based on their fault codes. The table indicated that the hypothesis could be correct, as the claimed products seemed to often have multiple notifications. Additionally, some descriptive statistics were calculated, such as how many notifications a claimed product had in average.

Secondly, the hypothesis was tested using the AI powered key influencers tool. The tool was used to analyze which factors in the combined warranty and notification data influenced the product to be claimed. As expected, the tool indicated that both types of notifications, ZP and Z2, influenced the probability of a claim strongly.

Following the validation of the initial hypothesis through the conducted analyses, a comprehensive examination of the relationship between the number of notifications and the probability of a claim was seen necessary. This was done by creating a predictive model using logistic regression algorithm in Python. Due to the way the data was stored and accessed in Power BI, running the Python model in Power BI was not a viable option. Instead, the code was run separately, and the results were imported back to Power BI.

4.4 Visualization

In the third part of the tool creation, the data and analysis results were visualized to provide a user-friendly tool for exploring the notification and warranty data as well as their relationship. The visualizations were designed to be easily readable and simple to enable a smooth user experience.

The tool visualizations include tables, bar charts, key influencer visualizations, and line charts, as well as a dashboard to visualize the results of the predictive modelling. Several visualizations incorporated filters to empower users with the ability to refine the displayed data according to their preferences. The number of graphs per page were kept to the minimum to enhance the readability of the report. This also makes finding the correct page and graph easier to the user.

The readability and interpretability of the report was also enhanced by adding captions and other information for the users to the report pages.

5 Predictive Modelling

In this chapter, the predictive modelling of the relationship between notifications and claim probability is explained in detail. As the initial analysis made in the Power BI indicated that the relationship might be positive and significant, the following hypothesis were made:

> ***Null Hypothesis (H0):*** *There is no statistically significant association between the quantity of notifications for an order line and the probability of a warranty claim.*
>
> ***H1:*** *There is a statistically significant relationship between the quantity of notifications for an order line and the probability of a warranty claim.*

Two possible outcomes for the analysis exists. Either a relationship is identified, indicating that as the number of notifications increases for an order line, the likelihood of a warranty claim also increases, or a significant relationship between the quantity of notifications and the probability of a warranty claim is not found.

Due to the binary nature of the predicted outcome (whether the product is claimed or not), logistic regression was deemed the most suitable method for modeling the relationship between notifications and claims, as well as the probability of a claim. Logistic regression is well-suited for modeling the probability of a binary outcome.

To validate or refute the null hypothesis, three variables were utilized in the model: a binary 0/1 column denoting whether a product had been claimed or not, and two columns representing counts of two types of notifications used by the case company in their order and manufacturing process, Z2 and ZP. The dataset was organized into an Excel file, with each row representing a single product. At the time of analysis, over 737,000 products were under warranty. All products under warranty were utilized in the model to enhance its accuracy.

Following data extraction from Power BI into an Excel file, the dataset was scrutinized to identify any rows with missing data—a critical step in ensuring the integrity and reliability of subsequent analyses. The Python logistic regression program was then executed, with the remaining chapter outlining the modeling process utilizing pseudo-code. The actual code is presented in appendix 1.

First, the program reads the data and extracts the features (notification counts) and target variable (claimed). Then the features are scaled. Scaling the data can improve convergence, interpretability, and consistency in modelling. In this case, the standardization ensures that each feature has a mean of 0 and a standard deviation of 1.

```
# Read Data
df = read_excel(file_path)

# Extract Features and Target Variable
X = df[['ZPCount', 'Z2Count']]
y = df['Claimed']

# Scale Numerical Features
X_scaled = standardize_features(X)
```

Algorithm 1. Read data, extract, and scale features.

After the initial steps, the data is split into train and test data. The model is trained using the training data and tested on separate test data to separately evaluate the performance of the trained data. The primary reason for this split is to assess how well the trained model generalizes to new, unseen data. In this case, 20 % of the data was used for testing, while the remaining 80 % was used for training. A random state parameter was set to ensure reproducibility if the code would be run more than once.

```
# Split Data
X_train, X_test, y_train, y_test = train_test_split(X_scaled,
y, test_size=0.2, random_state=42)
```

Algorithm 2. Split data.

Then, the model is trained using the train data. This involves adjusting the model's parameters to learn patterns in the training data, enabling it to predict the binary outcome of is the product claimed or not.

```
# Train Logistic Regression Model
model = train_logistic_regression(X_train, y_train)
```

Algorithm 3. Train the model.

After the training stage, the model is used to predict the variables of the test set. The predicted values are stored in the variable y_pred.

```
# Make Predictions on Test Set
y_pred = model.predict(X_test)
```

Algorithm 4. Predict test set variables.

Two new columns are added to the original data frame (df) containing predictions and probabilities. The model predicts on the scaled features (X_scaled) and the probabilities are extracted from the predicted probabilities of the positive class (claimed is 1, meaning the product is claimed). The updated data frame is printed on the Excel file on a new output tab.

```
# Update DataFrame with Predictions
df['Predicted'] = model.predict(X_scaled)
df['Probability'] = model.predict_proba(X_scaled)[:, 1]

# Write Updated DataFrame to Excel
write_to_excel(df, file_path, sheet_name='OUTPUT')
```

Algorithm 5. Update table with predictions.

Then, the probabilities for different combinations of notifications counts are calculated. A new data frame is created to present the different combinations of notification counts and their probabilities and printed on the Excel file on a new probabilities tab.

```
# Generate Combinations
ZPCount_values = generate_values()
Z2Count_values = generate_values()

# Calculate Probabilities for Combinations
probabilities = calculate_probabilities(model, ZPCount_val-
ues, Z2Count_values)

# Create Results DataFrame
result_df     =     create_results_dataframe(ZPCount_values,
Z2Count_values, probabilities)

# Write Results DataFrame to Excel
write_to_excel(result_df, file_path, sheet_name='Probabili-
ties')
```

Algorithm 6. Create and calculate table for claim probabilities.

The new output sheet is generated within the Excel file, featuring a table that includes the original data along with two additional columns: ‘Predicted and ‘Probability’. The binary ‘Predicted’ column holds the model prediction of the claim, and the probability column has the probability of the claim calculated using the coefficients the model has calculated for the two types of notifications. In addition to the predicted and probability columns, the program prints out the coefficients and intercept for the model features. These variables can be used to calculate the probability of claim in the model in different notification count combinations.

In addition to the output table containing the predictions and probabilities for the warranty data, the program calculated a claim probability table based on the model coefficients and intercept. This table was used to enhance the model’s comprehensibility and generalizability.

In addition to the code training and testing the logistic regression model and printing the new tables to Excel, the program prints out some key metrics of the model. These include the accuracy of the model ranging from 0 to 1, the model coefficients, and a flattened confusion matrix.

```
# Evaluate the model
calculate_accuracy(y_test, y_pred)
print_accuracy(accuracy)

# Print feature importance (coefficients in the case of lo-
gistic regression)
print('Coefficient for ZPcount:', model.coef_[0][0])
print('Coefficient for Z2count:', model.coef_[0][1])

# Calculate confusion matrix
calculate_confusion_matrix(y_test, y_pred)

# Flatten and print confusion matrix
flatten_and_print_confusion_matrix(cm)
```

Algorithm 7. Evaluate model performance.

6 Results

In this section, the outcomes of the research are presented, highlighting the achievements and key findings. The focus is on the development of a Power BI tool with predictive analytics capabilities using the Design Science Research Methodology (DSRM). The following subsections detail the results of the predictive modelling, factors connected to product claims, and potential future AI utilization possibilities.

6.1 Predictive Modelling

The first research question of the thesis *was how predictive models can be used to analyze warranty and notification data*. As stated in the literature review, predictive modelling encompasses three key types: classification, regression, and time series analysis. All three types of models have their use in warranty and notification data analysis. Classification models are useful to predict whether a product will be claimed or not, while regression models are suitable for understanding the relationship between various factors and outcomes, such as analyzing what type of claims different types of products typically experience, and how specific product characteristics may influence the frequency or severity of these claims. Finally, time series analysis can be utilized to predict future warranty claim frequency or to detect anomalies or unusual patterns in warranty and notification data that deviate from expected trends. In this thesis, two distinct types of predictive modelling methods were employed in the creation of the Power BI tool, encompassing both classification and time series methodologies.

6.1.1 Classification

The primary focus of this thesis revolves around utilizing logistic regression to forecast product claims, thereby clarifying the relationship between notifications and warranty claims. As described in the previous parts of this thesis in more detail, by combining notification counts for each product under warranty with information about whether the product has been claimed or not, a Python-based logistic regression model predicting warranty claims was built. In addition to classifying products into those which will and will not be claimed, the model yields valuable information about the effect of notifications to the claim probability. The process is relatively simple, and can be automated in Power BI, which enhances the attractivity of the solution further by reducing the requirements for manual work.

The process and results explained in the results for RQ2 illustrate that logistic regression is a valuable tool for the case company to analyze warranty and notification data. Most importantly, the tool can be utilized to predict which products will be claimed by the customer based on other information, such as the number of notifications. This information can highlight to the case company to which issues they should focus on to effectively reduce the number of product claims. The company can monitor the effectiveness of their actions aimed at reducing product claims by creating a subsequent model after implementing these measures. This new model would allow them to compare the coefficients obtained before and after the implementation, providing insights into the impact of their interventions on the predictive factors of warranty claims.

6.1.2 Time Series Analysis

In addition to classification methods, time series analysis was also utilized in the Power BI tool to analyze warranty and notification data. Power BI visualizations include a forecast option to line charts, which creates charts visualizing both the existing input data and the forecast. This is a useful tool for the case company as the charts are both easy to create and interpret, and suitable for visualizing basic data about warranty and notification data.

Figure 7 presents one of the charts utilizing time series forecasting in the developed Power BI tool. Similar charts were used to visualize the number of notifications. In the line chart, the red line represents the number of claimed serials in the past. The black line represents the predicted number of claimed serials in the next 12 months, while the grey area indicates the confidence interval for the same period. Additionally, the black dashed line represents the trend of the claimed serials increasing the readability of the report.

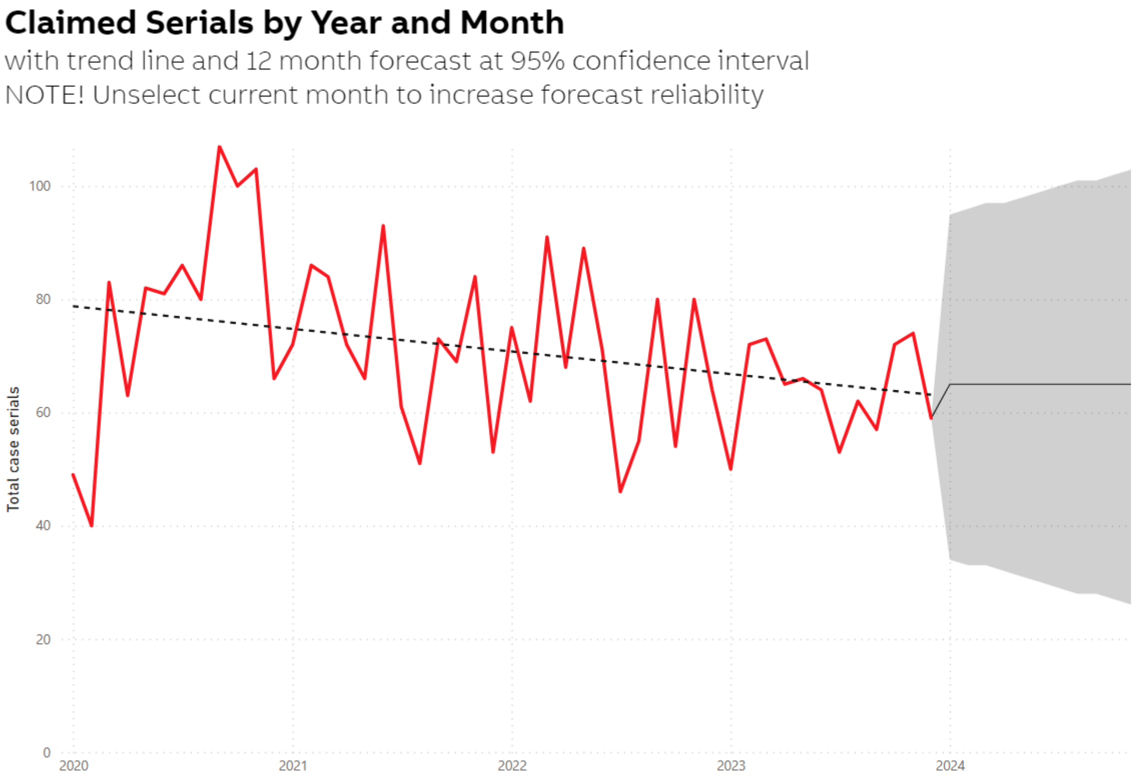


Figure 7. Time series forecast for claimed serials.

Forecasting in Power View utilizes different exponential smoothing methods such as seasonal and non-seasonal algorithm for creating the forecast. The most suitable method is selected for each chart based on the data qualities. When creating the chart, the user can customize the forecast horizon and confidence level according to their preferences. For the case company, utilizing time series analysis within Power BI offers a straightforward method to monitor the evolution and trends of warranty and notification data, all while presenting forecasts in a single chart. Time series forecasting is a valuable tool for predicting future trends and patterns in claim frequencies, notification counts, and other relevant metrics. This capability allows the case company to anticipate fluctuations in warranty claims and notification volumes, detect seasonal or cyclical patterns, and proactively address potential issues before they escalate. Moreover, time series forecasting can support the case company's long-term planning, resource allocation, and budgeting efforts by providing valuable insights into the future demand for warranty services and customer support.

6.2 Factors Connected to Product Claims

The second research question of this thesis is *what factors in the warranty and notification data are connected to product claims*. In this chapter, the factors connected to product claims are discussed in two parts. First, the results gained through predictive modelling are discussed, followed by the results gained using key influencers tool.

6.2.1 Predictive Modelling Results

Prior to this thesis, the case company was already under the impression that as the number of notifications increases for a product, so does the probability of a claim. However, this had not been statistically proven. The question was researched in this thesis by utilizing predictive modelling of existing warranty and notification data. The combined data was used to build a machine learning based predictive model utilizing logistic regression.

The first output of the logistic regression model was a table with the predicted outcome (will the product be claimed or not) for over 737,000 formerly or presently warrantied products and the probability of those products being claimed. 99.595 % of the rows were predicted to not be claimed, while 0.406 % of the rows were predicted to be claimed. When the percentage of predicted product claims is translated into PPM (parts per million), the predicted claims get a PPM value of 4,060.

Secondly, the model calculated the coefficients for the two types of notifications and their intercept. The coefficients for the notification counts indicate the change in the log-odds of the product claim probability for a one unit change in the corresponding notification count. In the model, the coefficient for ZP count is 1.565, while the coefficient for Z2 count is 1.576. A positive coefficient means that as the notification count increases, the probability of a product claim also increases. Therefore, it can be stated that both the ZP and Z2 count have a positive impact on the probability of a claim in the model when their value is increased by one unit. Additionally, the program calculated the intercept for the claim. The intercept represents the log-odds of the product claim when both notification counts are 0. In the model the intercept was -4.555. The negative intercept indicates, that when the notification counts are both 0, the product is more likely not to be claimed than claimed.

When the coefficients for the notification types and the intercept are analyzed, it is already clear that in the model the increase in notifications increases the probability of a claim, and that the probability of a claim is low when the number of both notification types is 0 for a product.

As the log-odds and intercept are relatively difficult to interpret in more detail due to their logarithmic nature, the coefficients and intercept were used to calculate simple probabilities for a product claim for different combinations of notifications. This was done directly in the Python code. The program harnessed the coefficient and intercept values to generate a table for claim probabilities. This table encompassed every conceivable combination of the two notification types, ranging from 0 to 5 instances each, along with the corresponding probability of a claim occurrence. This approach facilitates a broader understanding of the model's implications by offering a clear and generalized overview of the results.

Table 1 presents the claim probabilities for different combinations of the number of notifications. The probabilities are color coded so that claim probabilities under 0.5 are highlighted in green, probabilities between 0.5 and 0.95 are highlighted in yellow, and probabilities over 0.95 are highlighted in red.

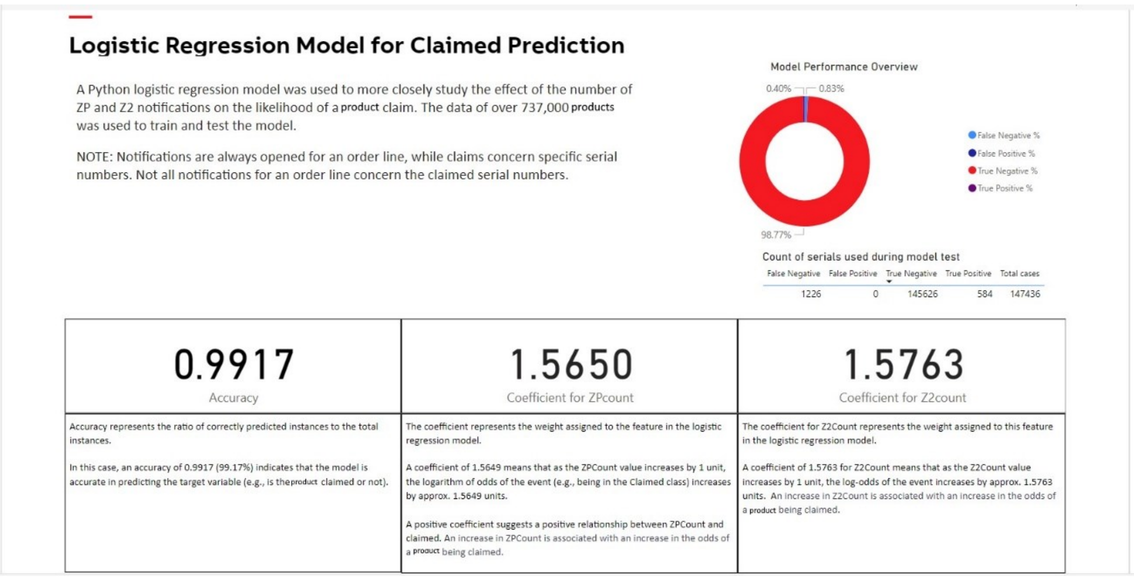
The probabilities in table 1 indicate the same as the coefficients and intercept, but in a significantly more easily interpretable form: when the number of notifications increases, the probability of a claim also increases. When both the notification types are 0, the claim probability is very low, only 0.01. The claim probability remains relatively low also when the total number of notifications is 1 or 2. However, the probability of the claim exceeds 0.50 when the total number of notifications is 3 or more. The highest claim probability of 0.99 is reached when both notification types have a value of 5. When the counts of notifications are increased over 5, the probability of claim is also increased, but as the logistic function used in logistic regression can never reach the value of 0 or 1, a point where claim probability is 1 can never be reached.

ZP count	Z2 count	Probability of claim
0	0	0,010405899
0	1	0,048400781
0	2	0,197445993
0	3	0,543380631
0	4	0,851984817
0	5	0,965328412
1	0	0,047881365
1	1	0,195655956
1	2	0,540566797
1	3	0,850549656
1	4	0,96494702
1	5	0,992545881
2	0	0,193878227
2	1	0,537750374
2	2	0,849103044
2	3	0,964561586
2	4	0,992461547
2	5	0,998432125
3	0	0,53493154
3	1	0,847644938
3	2	0,964172071
3	3	0,992376266
3	4	0,99841428
3	5	0,999671756
4	0	0,846175294
4	1	0,963778437
4	2	0,992290027
4	3	0,998396233
4	4	0,999668015
4	5	0,999931347
5	0	0,963380641
5	1	0,992202821
5	2	0,99837798
5	3	0,999664232
5	4	0,999930565
5	5	0,999985644

Table 1. Influence of the number of notifications on the probability of claim.

Furthermore, the model demonstrated a high accuracy rate of 0.9917, indicating its effectiveness in predicting whether a product would be claimed based on notification counts. This accuracy suggests that the model can reliably differentiate between products likely to be claimed and those not.

The model results were visualized in the Power BI tool in a one-page dashboard (Figure 8). The dashboard included general explanation of the model, its values including accuracy and coefficients, and model performance overview.



Logistic Regression Model for Claimed Prediction

A Python logistic regression model was used to more closely study the effect of the number of ZP and Z2 notifications on the likelihood of a product claim. The data of over 737,000 products was used to train and test the model.

NOTE: Notifications are always opened for an order line, while claims concern specific serial numbers. Not all notifications for an order line concern the claimed serial numbers.

Model Performance Overview

Count of serials used during model test

False Negative	False Positive	True Negative	True Positive	Total cases
1226	0	145626	584	147436

0.9917 Accuracy	1.5650 Coefficient for ZPcount	1.5763 Coefficient for Z2count
Accuracy represents the ratio of correctly predicted instances to the total instances. In this case, an accuracy of 0.9917 (99.17%) indicates that the model is accurate in predicting the target variable (e.g., is theproduct claimed or not).	The coefficient represents the weight assigned to the feature in the logistic regression model. A coefficient of 1.5649 means that as the ZPCount value increases by 1 unit, the logarithm of odds of the event (e.g., being in the Claimed class) increases by approx. 1.5649 units. A positive coefficient suggests a positive relationship between ZPCount and claimed. An increase in ZPCount is associated with an increase in the odds of a product being claimed.	The coefficient for Z2Count represents the weight assigned to this feature in the logistic regression model. A coefficient of 1.5763 for Z2Count means that as the Z2Count value increases by 1 unit, the log-odds of the event increases by approx. 1.5763 units. An increase in Z2Count is associated with an increase in the odds of a product being claimed.

Figure 8. Logistic regression model dashboard.

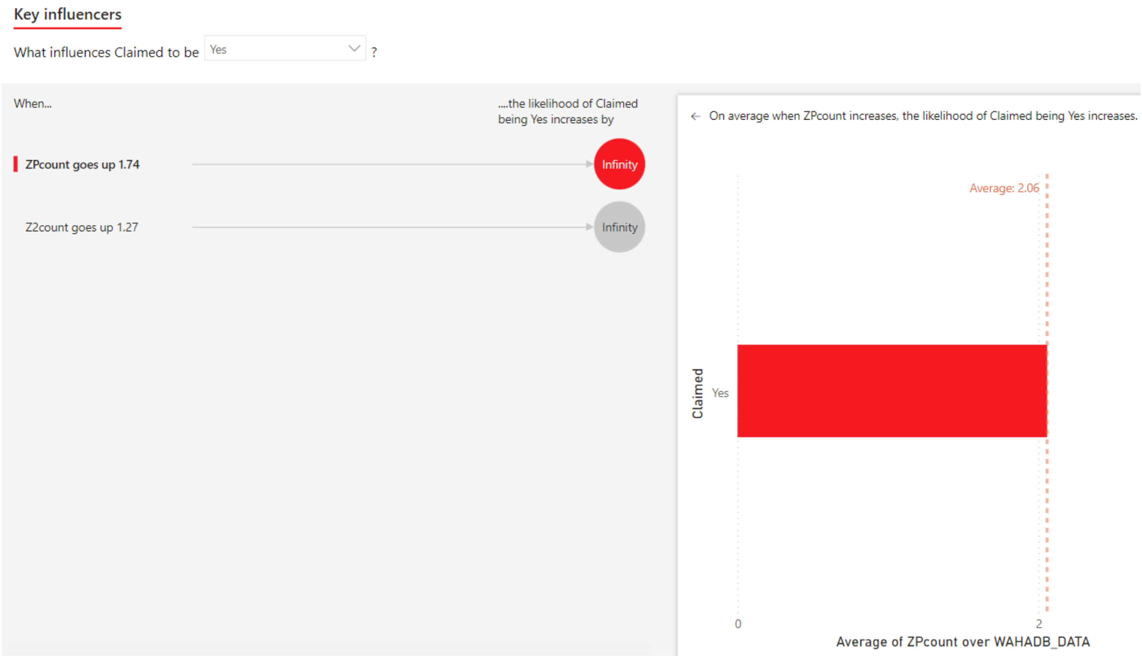
Based on the analysis results of the logistic regression model, H1 can be confirmed. The positive coefficients for both notification types indicate, that as the number of notifications is increased by one unit, the probability of the product claim is increased. Therefore, it can be stated that in the logistic regression model based on the case company warranty and notification data there is a statistically significant relationship between the quantity of notifications for an order line and the probability of a warranty claim of one or more delivered products of that order line. This relationship exists most likely because notifications are only opened for an order line when the product does not conform to what has been ordered, or when the customer wants to make changes to the order after it has been confirmed. This complicates the manufacturing process, which increases the risk of mistakes during it.

As the model was able to confirm the positive relationship of the number of notifications and the probability of a claim, and as it has high potential to indicate which of the delivered products are at a higher risk of warranty claim, the case company should consider investing effort to make the model run directly in Power BI. This would enable the company to fully utilize the model and its benefits. Additionally, the model would automatically be given new rows of data every time the Power BI report is refreshed. Additionally, implementing the model directly in Power BI enables the use of Python visualizations, such as Python visualizations, such as interactive heatmaps highlighting high-risk products and drill-down capabilities to explore specific categories or countries in more detail.

6.2.2 Key Influencers Tool Results

In addition to predictive modelling, the factors in the warranty and notification data affecting product claims were analyzed using the AI powered key influencers Power BI tool. The analysis was made in three parts to ensure easy readability of the report. First analysis included the manufacturing process related data, the second visualization included product related data, and the third visualization included sales related data.

The first analysis of the order and manufacturing process data revealed that both the number of ZP and Z2 notifications have a significant effect on the product claim probability (Figure 8). The visualization indicates that when the number of ZP notifications goes up to 1.74 the likelihood of the product being claimed is increased by infinity. For Z2 notifications, the visualization indicates that when the number of Z2 notifications goes up to 1.27 the likelihood of the product being claimed is increased by infinity. These results are in line with the results of the logistic regression model.



Key influencers

What influences Claimed to be Yes ?

When... ...the likelihood of Claimed being Yes increases by

ZPcount goes up 1.74 Infinity

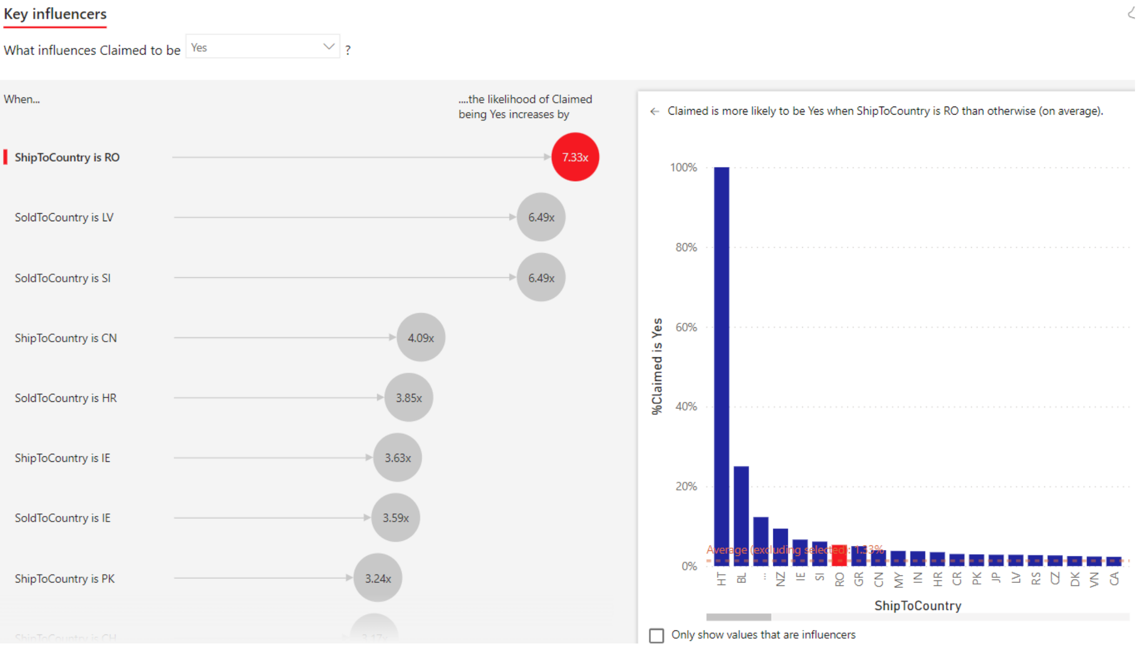
Z2count goes up 1.27 Infinity

On average when ZPcount increases, the likelihood of Claimed being Yes increases.

Figure 9. Manufacturing process key influencers.

The second key influencers analysis was made with product related data. The tool was able to recognize and rank different product types and attributes which affect the probability of a product claim. The tool was able to recognize many product types and attributes which have a clear effect on the claim probability. The products the tool highlighted as the most probable to be claimed included prototype products and other special low-volume products, but also regular products with lower reliability were highlighted.

In the third analysis, the tool analyzed key claim influencers from sales data. The tool indicated that some sold to and ship to countries were more problematic than others (Figure 9). For example, when Romania is the ship to country, the product claim probability is increased by over 7 times. The graph also shows that over 5 % of the products shipped to Romania are claimed. According to the visual, on average, 1.38 % of products are claimed.



Key influencers

What influences Claimed to be Yes ?

When...	...the likelihood of Claimed being Yes increases by
ShipToCountry is RO	7.33x
SoldToCountry is LV	6.49x
SoldToCountry is SI	6.49x
ShipToCountry is CN	4.09x
SoldToCountry is HR	3.85x
ShipToCountry is IE	3.63x
SoldToCountry is IE	3.59x
ShipToCountry is PK	3.24x
ShipToCountry is CH	3.17x

← Claimed is more likely to be Yes when ShipToCountry is RO than otherwise (on average).

☐ Only show values that are influencers

Figure 10. Sales key influencers

6.3 Future AI Utilization Possibilities

The third research question of the thesis is about *how the case company can utilize AI in analyzing warranty and notification data in the next 10 years.* As covered in the previous chapters of this thesis, AI was already utilized in the analysis of the warranty and notification data in the form of a ML based logistic regression model, time series forecasting, and by utilizing the AI powered key influencers tool in Power BI. All three analyses provided valuable information which the case company can benefit from and develop their operations and processes accordingly. However, as AI had not been utilized prior to this thesis in the analysis, the three analysis types can be viewed as the first steps. In the future, the case company has many ways to utilize AI in a more advanced level in the context of warranty and notification data.

As the first step towards utilizing AI in the analysis of warranty and notification data, the case company should improve the implementation of the AI powered analytics in their reporting. Due to the way data was imported to the Power BI report from the data warehouse, the logistic regression model could not be run directly in Power BI tool developed in this thesis. Power BI has enabled running Python code directly in the desktop environment, which enables seamless integration of Python code for advanced analytics such as logistic regression directly within Power BI. This allows for the creation of more sophisticated predictive models and insights, leveraging the power of AI and enhancing the company's ability to derive valuable insights from warranty and notification data. One of the key benefits of running the models directly in Power BI is the enhanced efficiency and convenience in analysis workflows, as it eliminates the need for separate software or platforms for model execution and visualization. Additionally, by integrating Python code seamlessly into Power BI, the company can leverage its existing infrastructure and expertise, streamlining the analytics process and facilitating quicker decision-making based on real-time insights.

After overcoming the first challenges of implementing AI analytics in their reporting more seamlessly, the case company has many opportunities to utilize AI in the analysis of the data. One of the opportunities is to widen the scope of AI-powered analysis to inspect other issues in the warranty and notification data, such as examining the typical types of claims associated with different product categories, and how specific product characteristics may influence the frequency or severity of these claims. This type of analysis is best supported by ML models such as decision trees, random forests, and support vector machines. Also, topics such as predicting future warranty claim frequency or detecting anomalies or unusual patterns in warranty and notification data could be analyzed using time series ML models such as autoregressive integrated moving average (ARIMA), seasonal decomposition of time series (STL), and exponential smoothing methods. These analyses could still be performed using Python, which can be seamlessly integrated into Power BI by running it directly within the platform.

In addition to widening their scope of AI powered analysis of warranty and notification data, including preprocessing and model optimization represents a significant opportunity for the case company. In preprocessing, the integration of AI algorithms for tasks like data cleaning and normalization streamlines operations and often outperforms human efforts, particularly with large datasets. This approach ensures higher data quality, laying a solid foundation for analysis of the data. Post-analysis, AI-driven model optimization techniques further enhance results by fine-tuning parameters and architectures, ultimately improving model performance and the quality of insights derived. By embracing AI across these stages, the company can boost operational efficiency and ensure more accurate decision-making processes, fostering innovation in warranty and notification data analysis.

While the case company currently has many opportunities to enhance their use of AI in data analysis and broaden their examination of warranty and notification data, the future of AI in BI holds numerous unpredictable possibilities. As discussed in the literature section, AI in BI is expected to undergo transformative trends, presenting promising prospects. Enhanced AI capabilities are positioned to provide BI systems with more precise and nuanced insights, while improvements in natural language processing are expected to facilitate smoother user interactions with BI platforms. Additionally, predictive analytics and prescriptive recommendations are anticipated to become more sophisticated, empowering organizations to forecast trends and make proactive, data-driven decisions. There is an expected emphasis on explainability and user-friendly interfaces, which should foster greater trust and adoption of AI in BI across diverse industries. As AI technologies continue to progress, the integration of AI-driven automation and augmented analytics is likely to streamline BI processes, enabling businesses to extract actionable intelligence more efficiently and maintain a competitive edge in the dynamic landscape of data-driven decision-making.

7 Discussion & Conclusions

In this chapter, the findings and conclusions drawn from the research on the utilization of predictive modelling and AI integration in the analysis of warranty and notification data within the context of a manufacturing company are discussed. The aim of this thesis was to develop a new Power BI report combining warranty and notification data, which the case company had not previously done. The report was to visualize the relationship between the two data sources and utilize predictive modelling. Additionally, the possibilities of utilizing AI in the future were to be researched. Based on the research process and its description in the preceding sections of this thesis, it is evident that the outlined objectives have been achieved. The last three sections of this thesis will include reflection on the research results, recommendations for the case company, and the evaluation and limitations of the research.

7.1 Reflection on the Results

This thesis answered to the following three research questions:

> *RQ1. How can predictive models be used to analyze warranty and notification data?*
> *RQ2. What factors in the warranty and notification data are connected to product claims?*
> *RQ3. How can the case company utilize AI in analyzing warranty and notification data in the next 10 years?*

Firstly, two predictive modelling techniques were successfully applied to analyze warranty and notification data of the case company. Logistic regression was used to accurately predict product claims, while time series analysis visualized trends and offered customizable forecasting options in Power BI. As the case company had not used predictive modelling prior to this thesis, the process of mapping what type of predictive modelling would most benefit the organization was one of the most challenging parts of the research process. Reflecting this on the feedback gotten from the case company of the results, the first trial of utilizing predictive modelling was successful. The easily readable forecast charts provide value to the case company as a simple way to follow the developments in notification and claim related issues. The first trial of using Python in predictive modelling also proved to be valuable, as the results from the model showed that predictive modelling can provide advanced information from warranty and notification data. This type of information is valuable and can motivate the case company to utilize predictive modelling in the future.

Although two predictive analytics methods were successfully used in this thesis, and valuable information was gained from both the time series and classification methods, harnessing the full potential of the logistic regression model for the case company to use was not successful. The results of the model were used to demonstrate that the number of notifications significantly affects the probability of a product claim, but the model output could not be utilized fully as the model could not be run directly on Power BI. If this could have been achieved, the case company could have visualized which of the delivered products are at the highest risk of being claimed due to warranty reasons. This would have added an additional layer of analysis that could have widened the scope of the Power BI tool.

Utilizing predictive modelling, we were able to provide answers to the second research question about the factors contributing to product claims. Both the logistic regression model and the Power BI key influencers analysis confirmed the H1 hypothesis and proved that the increased number of notifications for a product was significantly and positively affecting the probability of a product claim. This was the most important finding of the thesis due to the significant value for the case company. Now that the relationship has been analyzed, the company can take action to reduce product claims and warranty costs by focusing on the product at the highest risk of claim prior to their dispatch to customers. Additionally, the key influencers analysis indicated that certain products as well as ship to and sold to countries increase the risk of product claims. Although these results might not lead to any complex action at the case company, they are factors which the company is interested in being aware of.

Furthermore, the exploration of AI in analyzing warranty and notification data has laid the foundation for future advancements in predictive analytics within the case company. By successfully applying logistic regression and time series analysis, valuable insights were derived regarding the relationship between notifications and product claims. However, the full potential of AI in this context remains untapped due to certain limitations encountered during the research process.

Moving forward, the case company can capitalize on the insights gained from this thesis to enhance their predictive analytics capabilities. One avenue for improvement is to streamline the integration of predictive models, such as logistic regression, directly into the Power BI tool. This seamless integration would enable real-time analysis and visualization of risk factors, empowering decision-makers to proactively mitigate warranty claims.

Additionally, future efforts should focus on expanding the scope of AI-powered analysis to address other pertinent issues within warranty and notification data. For instance, implementing machine learning models like decision trees and random forests can facilitate the identification of typical claim patterns across different product categories. Similarly, leveraging time series models such as ARIMA and STL can aid in forecasting future warranty claim frequencies and detecting anomalies in data patterns.

Moreover, preprocessing and model optimization present significant opportunities for enhancing the accuracy and efficiency of predictive analytics. By employing AI algorithms for data cleaning and normalization, the case company can ensure higher data quality, thereby improving the reliability of predictive models. Furthermore, post-analysis optimization techniques can fine-tune model parameters and architectures, leading to improved performance and more actionable insights.

Looking ahead, the future of AI in BI holds immense potential for transformative advancements. As highlighted in the literature review, ongoing developments in AI technologies are expected to revolutionize BI systems, enabling organizations to derive more precise insights and make data-driven decisions with greater confidence. By embracing these emerging trends and continually refining their predictive analytics capabilities, the case company can position itself at the forefront of innovation in warranty and notification data analysis, driving operational excellence and enhancing customer satisfaction.

Altogether, leveraging logistic regression and time series analysis, this thesis explored the potential of AI in analyzing warranty and notification data. The research demonstrated the ability to predict product claims, identified key influencing factors, and paved the way for future implementations. While limitations exist, like refining the logistic regression model, this project signifies a crucial step towards transforming warranty data into a powerful tool for risk management, product improvement, and ultimately, enhancing customer satisfaction.

Lastly, the results presented one key recommendation for the case company. The logistic regression model confirmed the company's hypothesis of the positive relationship of the number of notifications and the probability of a warranty claim. This relationship is most likely connected to the fact that notifications complicate the manufacturing process and expose the products to possible errors such as faulty design or missing parts. Due to high importance of notifications to the flow of information in the manufacturing organization, the answer to decreasing the number of product claims caused by these errors is not decreasing the number of notifications. However, as the case company already have a final inspection in use for specific cases and sample tests, the final inspection should be utilized for checking the high-risk products before their dispatch from the factory. As most products do not have any notifications, adding a mandatory final inspection to high claim risk products with many notifications does not increase the workload of the inspection team significantly. The probability of claim in the logistic regression exceeds 0,5 when the total number of notifications is 3, which could be the recommended threshold for the final inspection. During final inspection, some of the nonconformities could be caught and fixed before dispatching, which could also lower warranty costs.

7.2 Managerial Implications

Based on the research questions and the results of this thesis, several potential recommendations for companies considering utilizing augmented analytics exist. Firstly, the results of this thesis have proved that utilizing AI in the BI context is effective and can provide insights which could not be achieved with the same effort using traditional BI tools. Therefore, the companies should consider invest in augmented analytics and develop the skills of their employees to increase understanding and utilization of AI and AA.

Secondly, although AI holds immense potential to solve business problems and deliver valuable insights through AA, organizations must proceed with cautious optimism and careful planning to guarantee its effective and ethical use. As highlighted in the literature review, successful AI adoption hinges on prioritizing clear and comprehensive communication. Building trust with all stakeholders is crucial through transparent explanations of the technology's capabilities, limitations, and potential impacts. Otherwise, issues such as misaligned expectations and disengaged workforce might slow down effective AI adoption and hinder its ability to solve business problems.

Third, companies should prioritize democratizing augmented analytics by creating accessible and interpretable AI models. While investing in advanced technology and expert talent is crucial, unlocking the full potential of AA requires empowering a wider range of employees to understand and utilize its insights. To achieve this, the AA reports should be designed with clear user interfaces and be easily interpretable. The employees should be trained for understanding data analysis and encouraged to collaboratively work with AA projects. Cultivating a collaborative organization with a diverse skillset in augmented analytics can create a potent catalyst for truly innovative and data-driven decision-making, potentially leading to improved business performance.

7.3 Evaluation and Limitations

In this thesis, predictive modelling was successfully applied to the case company warranty and notification data. The results demonstrate the value of AI in warranty and notification data analysis using logistic regression and time series analysis. Answers to other research questions regarding key factors for product claims and future AI opportunities were also provided. The positive relationship between notification count and claim probability was identified, providing valuable insights for the case company. Additionally, the thesis outlines several specific ways the case company can leverage AI in the future. Finally, the thesis was able to fulfill its objective, as a functional Power BI tool that visualizes data and offers basic forecasting, showcasing the potential of AI-powered reporting was developed.

While the research findings offer valuable insights for the case company, their generalizability to other contexts needs careful consideration. This research utilized data specific to the case company's products, manufacturing processes, and warranty/notification system. These factors may differ significantly in other companies, limiting the direct applicability of findings. Though the sample size was notable with data of over 737,000 products, the sample size of products and claims analyzed is not large enough to represent the full spectrum of potential scenarios across other industries.

Although the thesis outcomes can be evaluated as successful, there are several limitations. While the developed logistic regression model offered valuable insights, its inability to run directly within the Power BI report limits its accessibility and automation potential. This means the model requires separate execution and cannot automatically update with new data, hindering its full integration into the company's reporting workflow. Second, the research primarily focused on analyzing the probability of product claims, leaving other potentially valuable areas unexplored. Investigating factors related to claim types, severity, and specific product characteristics could provide further insights for targeted interventions and process improvements.

While the employed logistic regression and time series analysis techniques proved effective, the research primarily utilized foundational AI methods. Exploring more advanced techniques like deep learning could unlock further potential for uncovering complex patterns and relationships within the warranty and notification data, leading to even more sophisticated insights.

Finally, While the ethical implications of AI integration were briefly touched upon, a more in-depth exploration of factors like data privacy, bias mitigation, and explainability of AI models would strengthen the overall analysis and demonstrate responsible AI implementation practices.

In the future, deepening the scope of the analysis could provide more valuable knowledge. The claim types and severity could be explored to identifying product characteristics, manufacturing processes, or customer usage patterns associated with more severe claims, enabling targeted interventions and risk mitigation strategies. Time series or deep learning models to predict future claim trends and estimate associated costs could also be developed.

Second potential future research direction could be expanding the generalizability of the findings. Investigating the applicability of similar AI-powered analyses across different manufacturing industries or companies with diverse product lines and warranty/notification systems could increase the generalizability of the results significantly. This could identify industry-wide trends and factors influencing the effectiveness of AI-based warranty analysis.

Overall, this thesis successfully demonstrated the potential of AI in analyzing warranty and notification data and provided valuable insights for the case company. However, further research is needed to address limitations and increase the generalizability of the findings.

References

Ain, N., Vaia, G., DeLone, W. H., & Waheed, M. (2019). Two decades of research on business intelligence system adoption, utilization and success – A systematic literature review. DECISION SUPPORT SYSTEMS, 125, 113113. https://doi.org/10.1016/j.dss.2019.113113

Akerkar, R. (2019). Artificial intelligence for business. Springer.

Alexandropoulos, S. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. Knowledge engineering review, 34, . https://doi.org/10.1017/S026988891800036X

Alghamdi, N. A., & Al-Baity, H. H. (2022). Augmented Analytics Driven by AI: A Digital Transformation beyond Business Intelligence. Sensors (Basel, Switzerland), 22(20), 8071. https://doi.org/10.3390/s22208071

American Psychological Association (n.d.). APA Dictionary of Psychology. Retrieved November 15, 2023, from https://dictionary.apa.org/trust

Azmi, M., Mansour, A., & Azmi, C. (2023). A Context-Aware Empowering Business with AI: Case of Chatbots in Business Intelligence Systems. https://doi.org/10.1016/j.procs.2023.09.068

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bousdekis, A., Lepenioti, K., Apostolou, D., & Mentzas, G. (2023). Data analytics in quality 4.0: Literature review and future research directions. International journal of computer integrated manufacturing, 36(5), 678-701. https://doi.org/10.1080/0951192X.2022.2128219

Bulusu, L., & Abellera, R. (2021). AI meets BI: Artificial intelligence and business intelligence. CRC Press.

Chen, Y., Li, C., & Wang, H. (2022). Big Data and Predictive Analytics for Business Intelligence: A Bibliographic Study (2000–2021). Forecasting, 4(4), 767-786. https://doi.org/10.3390/forecast4040042

EK. (2023). Data- ja tekoälykysely jäsenyrityksille 2023. Retreived 3032-11-14 from https://ek.fi/wp-content/uploads/2023/11/EK-yrityskysely_data_teknologiat_digiosaaminen2023_13_11_2023.pdf

European Commission. (2019). Ethics Guidelines for Trustworthy AI. European Commission. https://doi.org/10.2759/346720

Figalist, I., Elsner, C., Bosch, J., & Olsson, H. H. (2022). Breaking the vicious circle: A case study on why AI for software analytics and business intelligence does not take off in practice. The Journal of systems and software, 184, 111135. https://doi.org/10.1016/j.jss.2021.111135

Gurcan, F., Ayaz, A., Menekse Dalveren, G. G., & Derawi, M. (2023). Business Intelligence Strategies, Best Practices, and Latest Trends: Analysis of Scientometric Data from 2003 to 2023 Using Machine Learning. Sustainability (Basel, Switzerland), 15(13), 9854. https://doi.org/10.3390/su15139854

Hamzehi, M., & Hosseini, S. (2022). Business intelligence using machine learning algorithms. Multimedia tools and applications, 81(23), 33233-33251. https://doi.org/10.1007/s11042-022-13132-3

Hasija, A., & Esper, T. L. (2022). In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance. Journal of business logistics, 43(3), 388-412. https://doi.org/10.1111/jbl.12301

International Organization for Standardization. (2020). Artificial intelligence - Overview of trustworthiness in artificial intelligence (ISO/IEC TR 24028:2020). Retreived 2023-11-15 from https://www.iso.org/standard/77608.html

Kananen, H., & Puolitaival, H. (2019). Tekoäly: Bisneksen uudet työkalut. Alma Talent Oy.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business horizons, 62(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy Artificial Intelligence: A Review. ACM computing surveys, 55(2), 1-38. https://doi.org/10.1145/3491209

Khan, M. A., Saqib, S., Alyas, T., Ur Rehman, A., Saeed, Y., Zeb, A., . . . Mohamed, E. M. (2020). Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. IEEE access, 8, 116013-116023. https://doi.org/10.1109/ACCESS.2020.3003790

Khan, W. A., Chung, S., Awan, M. U., & Wen, X. (2020a). Machine learning facilitated business intelligence (Part I): Neural networks learning algorithms and applications. Industrial management + data systems, 120(1), 164-195. https://doi.org/10.1108/IMDS-07-2019-0361

Khan, W. A., Chung, S., Awan, M. U., & Wen, X. (2020b). Machine learning facilitated business intelligence (Part II): Neural networks optimization techniques and applications. Industrial management + data systems, 120(1), 128-163. https://doi.org/10.1108/IMDS-06-2019-0351

Lee, W. (2019). Python Machine Learning. John Wiley & Sons, Incorporated.

Li, M., & Gregor, S. (2011). Outcomes of effective explanations: Empowering citizens through online advice. Decision Support Systems, 52(1), 119-132. https://doi.org/10.1016/j.dss.2011.06.001

McCarthy, J. (2007). What is artificial intelligence?. Standford University. Retreived 2023-11-13 from https://www-formal.stanford.edu/jmc/whatisai.pdf

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. Information systems management, 39(1), 53-63. https://doi.org/10.1080/10580530.2020.1849465

Microsoft. (2023). What is Power BI?. Microsoft Learn. Retrieved 2024-1-13 from https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

Nakhal A, A., Patriarca, R., Di Gravio, G., Antonioni, G., & Paltrinieri, N. (2021). Investigating occupational and operational industrial safety data through Business Intelligence and Machine Learning. Journal of loss prevention in the process industries, 73, 104608. https://doi.org/10.1016/j.jlp.2021.104608

Pampel, F. (2021). Logistic regression: A primer. SAGE Publications, Inc., https://doi.org/10.4135/9781071878729

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. Journal of management information systems, 24(3), 45-77. https://doi.org/10.2753/MIS0742-1222240302

Prat, N. (2019). Augmented Analytics. Business & information systems engineering, 61(3), 375-380. https://doi.org/10.1007/s12599-019-00589-0

Purnomo, A., Firdaus, M., Sutiksno, D. U., Putra, R. S., & Hasanah, U. (2021). Mapping of Business Intelligence Research Themes: Four Decade Review. https://doi.org/10.1109/COMNETSAT53002.2021.9530790

Saadat, R., Syed-Mohamad, S. M., Azmi, A., & Keikhosrokiani, P. (2022). Enhancing manufacturing process by predicting component failures using machine learning. Neural computing & applications, 34(20), 18155-18169. https://doi.org/10.1007/s00521-022-07465-1

Santos, L., & Ferreira, L. (2023). Atlantic—Automated data preprocessing framework for supervised machine learning. Software impacts, 17, 100532. https://doi.org/10.1016/j.simpa.2023.100532

Sharda, R., Delen, D., & Turban, E. (2018). Business intelligence, analytics, and data science: A managerial perspective (Fourth edition.). Pearson.

Sinha, C. (2021). Mastering power BI: Build business intelligence applications powered with DAX calculations. insightful visualizations, advanced BI techniques, and loads of data sources. BPB Publications.

Tamang, M. D., Kumar Shukla, V., Anwar, S., & Punhani, R. (2021). Improving Business Intelligence through Machine Learning Algorithms. https://doi.org/10.1109/ICIEM51511.2021.9445344

Tavera Romero, C. A., Ortiz, J. H., Khalaf, O. I., & Rios Prado, A. (2021). Business Intelligence: Business Evolution after Industry 4.0. Sustainability (Basel, Switzerland), 13(18), 10026. https://doi.org/10.3390/su131810026

Tripathi, M. A., Madhavi, K., Kandi, V. P., Nassa, V. K., Mallik, B., & Chakravarthi, M. K. (2023). Machine learning models for evaluating the benefits of business intelligence systems. Journal of high technology management research, 34(2), 100470. https://doi.org/10.1016/j.hitech.2023.100470

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. https://doi.org/10.1007/978-3-030-46781-4_1

Was Rahman. (2020). AI and Machine Learning. Sage Publications Pvt. Ltd.

Appendix

Python Code for Logistic Regression Model Implementation

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

# Read data from Excel file
file_path = r'
df = pd.read_excel(file_path)

# Extract features and target variable
X = df[['ZPCount', 'Z2Count']]
y = df['Claimed']

# Scale numerical features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled,
y, test_size=0.2, random_state=42)

# Train the logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')

# Print feature importance (coefficients in the case of logistic regression)
print('Coefficient for ZPcount:', model.coef_[0][0])
print('Coefficient for Z2count:', model.coef_[0][1])

# Add new columns to the existing DataFrame
df['Predicted'] = model.predict(X_scaled)
df['Probability'] = model.predict_proba(X_scaled)[:, 1]
```

```
# Write the updated DataFrame back to the Excel file
with pd.ExcelWriter(file_path, engine='openpyxl', mode='a',
if_sheet_exists='replace') as writer:
    df.to_excel(writer, sheet_name='OUTPUT', index=False)

# Evaluate the model
y_pred = model.predict(X_test)

# Calculate confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Flatten and print confusion matrix
flat_cm = cm.flatten()
print('True Negative:', flat_cm[0])
print('False Positive:', flat_cm[1])
print('False Negative:', flat_cm[2])
print('True Positive:', flat_cm[3])
```