Politecnico di Torino

Politecnico di Torino

Determining the Key Factors and Estimation of Fuel Consumption in Cold Chain Logistics: A Machine Learning Approach

Master's Thesis
Department of Engineering Management & Production (DIGEP)
Graduation Session July-2024

Thesis Candidate: Tezcan Aral Tezcan
Student ID: S289829

Thesis Advisor: Giovanni Zenezini
Department of Engineering Management & Production (DIGEP)

Table of Content:

Abstract

This study investigates the key factors influencing fuel consumption in cold chain logistics (CCL) and presents a machine learning approach to estimate and optimize fuel usage. By analyzing data from various sources, the research identifies significant variables affecting fuel consumption, including vehicle age, maintenance frequency, temperature control settings, route characteristics, and load management. The findings highlight the importance of leveraging advanced technologies and machine learning models to enhance fuel efficiency, reduce costs, and improve environmental sustainability in CCL operations. Various linear regression models were tested to identify the best predictive solution, ensuring accurate and reliable estimates of fuel consumption under different conditions. This rigorous testing process helps identify the most effective strategies for minimizing fuel use. This approach paves the way for more sustainable and efficient logistics operations, ensuring adaptability and competitiveness in a rapidly evolving market.

Keywords: *Cold Chain Logistics (CCL), Transportation Efficiency, Food Logistics, Fuel Consumption, Machine Learning*

1. Introduction

1.1. Definition of Cold Chain Logistics and its Importance

Cold chain logistics (CCL) refers to the process of management and transportation of temperature-sensitive products, like perishable food items, through a supply chain with controlled temperatures (Wang et al., 2018). That may also be defined as a low-temperature supply chain system combining the refrigeration industry and logistics. The list of products that need cold chain logistics ranges from perishable goods like fresh produce, dairy products, meat, and seafood to special products like vaccines and medications (Han et al., 2021). Most of these products are highly perishable and demand strict temperature ranges to maintain their quality and make them fit for use or consumption. This same process is necessary to maintaining the quality, safety, and shelf life of these same perishable goods and in reducing food losses during the entire distribution process (Capo, 2021). In this respect, it may be said that CCL is crucial for making safe and high-quality products available to customers. It requires strict temperature control, monitoring, and special kinds of equipment and infrastructure. CCL is a complex and challenging field that needs a lot of factors to come together in order to be effective and sustainable. This ranges from energy consumption to fuel efficiency and environmental impact, involving processes at different stages: production, storage, and transportation down to consumption. In this regard, innovative technologies can be implemented, such as the Internet of Things (IoT) and Machine Learning (ML) tools targeted at enhancing efficiency and productivity in those processes. (Capo, 2021).

Preserved quality and safety are thus significant components of cold chain logistics. If the products that are perishable are subjected to higher temperatures or other unregulated adverse climactic conditions during transport or storage, their quality rapidly deteriorates, rendering markets unsafe for consumption (Capo, 2021). Loss of food as a consequence of poor management in the cold chain adds up to the food waste and has economic and environmental consequences associated with it (IPCC, 2019). Temperature control in CCL is very critical because small deviations from the optimum temperature window could mean this product will go to spoilage, reduction in quality, and sometimes even cause harm to consumers. However, maintaining the required temperature range can be energy-intensive, with refrigeration accounting for a significant portion of the energy consumed in CCL (Han et al., 2021). As such, identifying and prioritizing energy-efficient measures for cold chains, including the estimation of fuel consumption rates, is essential for mitigating these negative impacts and promoting sustainable food supply practices (Marchi et al., 2022b). As a result, it has made

fuel consumption efficiency one of the key factors toward the optimization of CCL operations. Besides, reducing this fuel consumption helps logistics companies in environmental sustainability, which reduces costs and improves the profitability of the firm (Wang et al., 2018).

In developing green supply chains, CCL's environmental performance is critically associated with the travel economy, vehicle loads, and fuel consumption rates of transport vehicles; hence, the study of their relationship is required (Rahman et al., 2022). The estimation methods would include doing regression analysis based on data statistics to provide a linear expression of fuel consumption per unit distance (Ning et al., 2023). Considering the parameters, for better fuel consumption rate estimation and promoting sustainable CCL practices, the application of machine learning techniques may be applied (Wang et al., 2018).

1.2. Global Trends and Market Growth of Cold Chain Logistics

The demand for cold chain logistics is growing rapidly, driven by factors such as globalization, urbanization, and increasing consumer awareness of food safety and quality (Chandran et al., 2022). The concept of CCL has been around since ancient times, with early civilizations using ice and snow to preserve food and other perishable items. The definition of CCL has evolved over the years, with the development of technology and the increasing demand for fresh and quality products. Over time, technological advancements have allowed for more sophisticated and reliable cold chain systems, enabling the growth of industries such as pharmaceuticals, food, and biotechnology. In recent years, CCL, especially for agricultural products, has seen significant development which aims to keep agricultural products fresh before they arrive at the designated locations (Han et al., 2021). Besides, the scale demand for CCL is growing due to the serious global loss of perishable food, causing a significant environmental burden.

A recent report by Grand View Research, Inc. projects today, cold chain logistics is a rapidly expanding industry, with a global market size estimated at USD 233.2 billion in 2022 and projected to reach $271 billion in 2023.Also the global CCL market forecasted to reach USD 892.3 billion by 2030, rising at a CAGR of 18.6% from 2023 to 2030 (Grand View Research , 2022). The market is expanding due to growing investment in cold chain development, rising Information Technology (IT) spending in cold storage logistics, and rising demand for high-quality goods. Currently in 2023, the Asia Pacific region, particularly China, is a major contributor to the cold chain market due to factors such as technological advancements in the packaging, processing, and storage of seafood products, rising demand, and growing cold chain infrastructure development (Grand View Research , 2022). Furthermore, the COVID-19 pandemic has increased the growth in e-commerce sales, which is driving the demand for cold chain solutions. The pandemic has led to a significant rise in the number of e-commerce purchases, including the purchase of perishable products, which must be kept in cold storage warehouses and distributed with thermally insulated packaging through refrigerated vehicles. Additionally, The importance of CCL has been further highlighted during the COVID-19 pandemic, where it played a critical role in the transport of vaccines (Grand View Research , 2022). This trend highlights the need for the food value chain to transition toward a cold-chain system that maintains perishable goods for extended durations.

Various trends and emerging markets are shaping the landscape of CCL. One such trend is the increasing use of data analytics in decision-making for CCL. This approach allows companies to better understand the complex factors influencing their logistics operations, and in turn, optimize their processes to minimize costs and energy consumption (Chaudhuri et al.,

2018). Additionally, the rise of bi-objective mathematical models for CCL networks takes into account economic, social, and environmental factors, providing a more comprehensive understanding of the industry (Z. Wang et al., 2020).

Technological advancements are playing a crucial role in improving the efficiency and sustainability of CCL. One such advancement is the use of RFID-based sensing for real-time monitoring of perishable cargo, which allows for improved cold chain management and reduced waste (Emenike et al., 2016). Moreover, the estimation of efficient fossil fuel prices that reflect supply costs, environmental costs, and general consumer taxes is another important aspect of sustainable CCL (International Monetary Fund, 2023). Briefly, the growth in CCL can be attributed to technological advancements and the growing need to ensure shipment integrity, efficiency, and safety (Grand View Research , 2022).

1.3. Problem Statement and Objectives

One of the primary challenges related to fuel consumption in CCL is the need to balance the travel economy with vehicle loads, in addition to considering the environmental impact of greenhouse gas emissions (Rahman et al., 2022). CCL networks are complex systems that require careful consideration of multiple factors, including economic, social, and environmental concerns (Z. Wang et al., 2020). As a result, it is crucial for researchers and practitioners alike to identify key supply-chain dependencies and explore ways to foster domestic production and export of low-carbon technologies (LCT) products through technology and innovation (International Monetary Fund, 2023).

The need for improved fuel efficiency and emission reduction is a pressing issue in the CCL sector. Given that the industry is responsible for a significant portion of global greenhouse gas emissions, it is imperative to explore strategies to reduce fuel consumption and improve overall environmental performance (Z. Wang et al., 2020). Some potential avenues for achieving these goals include optimizing the use of air conditioning systems, implementing thermal load management strategies, providing improvements in operational processes, and encouraging the adoption of new generation vehicle types. By focusing on these areas, the CCL sector can contribute to important reductions in greenhouse gas emissions and improvements in fuel economy.

The objectives of this research study are to provide a comprehensive literature review on the determination of key factors of fuel consumption in CCL, identify the main problems and challenges associated with fuel consumption and emissions in the industry using machine learning tool, and explore the relationship between fuel consumption and the factors related to vehicle selection, temperature control, route planning, load planning and driver behaviour. By examining the various factors that contribute to fuel consumption in CCL, the study aims to develop a mathematical model that considers vehicle, environment and temperature, route, load, driver related concerns. Evaluate the potential for fuel consumption reduction through improved air conditioning, thermal load management, and alternative fuel technologies. Provide valuable insights and recommendations for both academia and industry practitioners to promote more sustainable and efficient CCL operations. Ultimately, the study seeks to provide valuable insights and recommendations for both academia and industry practitioners, with the goal of promoting more sustainable and efficient CCL operations.

1.4. Scope, Limitations, and Significance of the Research

The scope of this study revolves around 5 main key factors: vehicle related, refrigeration related, route related, load related and driving related which have a significant impact on fuel consumption in CCL. The research methodology employed in this study is a machine learning approach, which will include several models to achieve and compare different solutions. This method allows the researchers to analyse large datasets and identify underlying themes and

patterns that can help in understanding the factors affecting fuel consumption in the CCL industry.

There are certain limitations that need to be addressed and offer opportunities for future research. One notable limitation is the need for validation of the findings through further empirical research and modelling. Additionally, the development of more comprehensive methods and datasets is crucial for effectively utilizing the potential of modern computing in this field. Future research could also focus on exploring other factors that may have an impact on fuel consumption in CCL, as well as the potential implementation of alternative fuel sources and technologies to reduce overall energy consumption.

The significance of this research lies in its potential to contribute to the decision-making processes in CCL by using data analytics. By identifying the key factors that influence fuel consumption and developing estimation models, industry practitioners can make more informed decisions regarding vehicle loads, route planning, and energy management. Moreover, this research can assist in the development of new fuel economy applications for the CCL industry. Ultimately, the findings of this study have the potential to lead to more efficient and environmentally sustainable practices in the CCL industry, benefiting both businesses and society as a whole.

2. Literature Review

2.1. Challenges in Cold Chain Logistics

Cold chain logistics presents numerous challenges in storage, packaging, and transportation (Al-Wakkal, 2020). Overcoming these challenges requires a strong understanding of the specific requirements of temperature-sensitive goods, as well as the implementation of effective monitoring systems and visibility into the supply chain (Marchi et al., 2022b). By addressing these challenges, logistics managers can ensure that temperature-sensitive products reach their final destination in optimal condition, maintaining their quality and safety throughout the entire cold chain process (Fan et al., 2021).

One of the primary challenges in CCL is maintaining the appropriate temperature conditions for temperature-sensitive products during storage (Mercier et al., 2017). Inadequate insulation, outdated equipment, and insufficient temperature monitoring and maintenance systems contribute to the inefficiencies in cold storage operations. These inefficiencies can result in substantial economic losses and increased greenhouse gas emissions, further highlighting the need for improved storage practices and facilities (Ashok et al., 2017). This requires a continuous monitoring system that can track and record temperature fluctuations. Inadequate temperature control can lead to product spoilage, decreased shelf life, and potential health risks (Ren et al., 2022). Additionally, managing the storage capacity to accommodate varying product volumes while ensuring that the temperature requirements are met can also be a challenge. A lack of supply chain visibility can further exacerbate these issues, making it difficult for logistics managers to make informed decisions about storage conditions and capacity (Ren et al., 2022).

Proper packaging plays a vital role in preserving the quality and safety of temperature-sensitive products. Poor packaging and insulation can result in temperature changes within the packaging, leading to product spoilage or damage. The use of inappropriately ventilated packaging can result in the deterioration of product quality. The challenge in packaging lies in selecting the right materials and insulation that can withstand the rigors of transportation while effectively maintaining the required temperature range (T. Ren et al., 2022). This often involves a delicate balance between cost and effectiveness, as more advanced packaging materials can be expensive. Furthermore, the packaging must be designed to accommodate

various product sizes and shapes, ensuring that they remain protected throughout the transportation process (Ren et al., 2022).

Transportation is another critical aspect of CCL, as it directly impacts the quality and safety of temperature-sensitive goods. Some challenges in transportation include selecting the appropriate transportation equipment that can maintain the required temperature range, carefully choosing transportation routes to minimize transit time and exposure to extreme temperature fluctuations, and ensuring perfect timing to coordinate the delivery of products to their final destination (Xu et al., 2023). Studies propose optimization models to minimize unit costs of product freshness, as well as carbon trading mechanisms to reduce the environmental impact of transportation. Visibility into the transportation process is essential to effectively manage these challenges, as it allows logistics managers to monitor and make real-time adjustments to transportation conditions and routes (Xu et al., 2023).

2.2. Fuel Consumption in Cold Chain Logistics and Environmental Impact

In the cold chain, there is a temperature-controlled section of the supply chain required for those products that need a series of temperature-controlled environments uninterrupted all the way from production to delivery (Capo, 2021). The inability to monitor the temperature and maintain it with proper systems can cause temperature excursions or equipment breakdowns, which remains a challenge to the safety and potency guarantees regarding perishable products during transportation and storage (Ashok et al., 2017).

The other major factor of CCL in the supply chain is associated with infrastructural and technological difficulties. It has been analyzed that location and routing are optimized in such a way that carbon imprinting and transport efficiency are considered for diminishing the challenges in CCL (Wang et al., 2018). This optimization will help the means towards diminishing the pressure on the environment and operational hazards by maintaining the cold chain process. Moreover, big data technology integrated with cloud logistics has been considered a viable way of improving CCL management (Xie & Zhao, 2016). Poor data collection and limited accountability structures can hinder accurate and regular inventory updates, leading to challenges in understanding the current status of cold chain equipment (Ashok et al., 2017).

Operational efficiency and logistics optimization in the view of CCL are the most important, for they would ensure that fuel consumption is kept at a minimum level and thus reduce its impact on environmental degradation. This proposed method could ensure the optimization of CCL through machine learning approaches in the estimation for fuel consumption of heavy-duty vehicles (Katreddi, 2023). In CCL operational efficiency optimization, it is most important to determine and understand the amount of energy used during all the different stages (Tan et al., 2021).

Factors influencing fuel consumption of cold chain vehicles again support the high cost and low environmental sustainability of such logistic operations (J. Zhang et al., 2019). The research in the field of fuel-consuming features of a vehicle, like engine efficiency, vehicle weight, and aerodynamics, becomes very crucial for finding improvement scopes (Katreddi, 2023). This can also be affected by external factors, which involve driving behavior, route optimization, and weather conditions (Rahman et al., 2022). A company can understand these factors to establish strategies and best practices to reduce fuel consumption and efficient CCL.

Fuel consumption in CCL has a direct impact on cost and environmental sustainability (Rahman et al., 2022). High fuel consumption can lead to increased operational costs, which can negatively affect a company's profitability and competitiveness in the market (Fares et al., 2023). Additionally, higher fuel consumption contributes to higher greenhouse gas emissions,

resulting in adverse environmental impacts (Al-Wakkal, 2020). The energy efficiency of cold warehousing is a significant contributor to sustainability impacts due to energy costs and greenhouse gas emissions (Al-Wakkal, 2020). Therefore, reducing fuel consumption in CCL is essential for achieving cost savings and minimizing the environmental footprint of these operations.

In addition to the challenges mentioned above, fuel consumption in CCL also faces environmental impacts, high energy consumption, and carbon emissions. CCL operations contribute significantly to carbon emissions, ozone depletion, and a decline in air quality, and depletion of non-renewable resources, further exacerbating climate change and environmental degradation. (J. Chen et al., 2021). By increasing greenhouse gas emissions, particularly carbon dioxide, it is a major contributor to global warming and climate change (Leng et al, 2020). Therefore, improving fuel efficiency in CCL can help to reduce greenhouse gas emissions and mitigate the impact of logistics operations on climate change (Jia, 2022). Moreover, the carbon emissions from CCL are relatively high, contributing to environmental degradation which negatively impact ecosystems and human health (Chandran et al., 2022). Since the refrigerated transport industry is believed to be responsible for 15% of all fossil fuel energy consumed worldwide, there has been an increase in interest in recent decades in optimizing these systems to decrease their environmental impact. In road refrigerated transport, vapor compression refrigeration units, which are often driven by diesel engines, are the most widely utilized systems (Maiorino et al., 2021). Greenhouse gas (GHG) emissions from CCL are particularly concerning, with the sector accounting for nearly 2.5% of global GHG emissions (Chandran et al., 2022). This includes both direct and indirect effects of the industry's activities, highlighting the significant environmental impact of CCL on a global scale. Therefore, it is crucial to focus on fuel efficiency and sustainability in the supply chain. This requires a holistic approach that considers factors such as logistics costs, energy consumption, and carbon emissions, aiming to minimize total costs and total carbon emissions (Xu et al., 2023). By addressing these challenges and implementing efficient and sustainable practices, CCL can continue to play a vital role in maintaining the quality and safety of fresh agro-products while contributing to a more sustainable future (Rahman et al., 2022).

Additionally, fuel consumption also contributes to air pollution, which can have detrimental effects on human health and the environment. Therefore, reducing fuel consumption can also contribute to improved air quality. To mitigate these environmental risks, it is essential to adopt strategies that reduce fuel consumption and promote eco-friendly practices in CCL operations (Rahman et al., 2022). By improving fuel consumption efficiency, companies can reduce their air pollution and contribute to a more sustainable future (Rahman et al., 2022).

Several fuel reduction strategies and best practices have been identified to help reduce fuel consumption in CCL. Some of these strategies include:

- Route optimization: By planning the most efficient routes, companies can minimize travel distances and reduce fuel consumption (Capo, 2021).
- Vehicle maintenance: Ensuring that vehicles are well-maintained, including regularly checking tire pressure and engine performance, can improve fuel efficiency (Katreddi, 2023).
- Driver training: Educating drivers on fuel-efficient driving behaviors, such as maintaining a steady speed, can help reduce fuel consumption (Fares et al., 2023).
- Load optimization: Properly distributing and minimizing vehicle loads can improve fuel efficiency (Rahman et al., 2022).

- Use of advanced technologies: Implementing machine learning and artificial intelligence algorithms for route planning, load optimization, and vehicle maintenance can further reduce fuel consumption (Chen, 2020).

By adopting these fuel reduction strategies and best practices, companies can improve the efficiency of CCL, reduce operational costs, and contribute to environmental sustainability.

2.3. Factors Affecting Fuel Consumption

Economic, environmental, and social effects are the most dominating issues in CCL (Leng et al, 2020). Considering those factors that influence or affect fuel consumption can turn around reducing the organization's chain on the environment, decreasing the cost of transportation, and enhancing the overall efficiency through SC. Literature review is done for some literature related to these variables of fuel consumption in CCL to present a deeper understanding of their extent and implications. Such consideration is important in the factors affecting fuel consumption in CCL, given the possible economic, environmental, and social impacts (Leng et al, 2020). With regard to the factors affecting fuel consumption, a business would ensure the optimization of its operations in CCL to minimize the environmental impacts while maintaining customer satisfaction and ensuring overall efficiency in the supply chain.

Infrastructures and regulations generally play a crucial role in determining the fuel consumption of CCL (Han et al., 2021). Most infrastructures of cold chains are in a substandard state, with lack of sufficient standardization that increases fuel consumption and other associated logistical problems (Han et al., 2021). Some of the factors that have generally plagued the cold chain management include the high installation and refrigeration systems cost, lack of finance, lack of government support, and inadequate infrastructure. National policy and financial intervention in countries like China are expected to be the main driving forces behind renovating infrastructure and improving CCL efficiency (Han et al., 2021). Not just the infrastructure itself, but the rules and regulations regarding fuel efficiency and emissions are slated to play a huge role in fuel usage in CCL (S. Wang, 2022).

Vehicle-related factors have important roles in fuel consumption within CCL (Rahman et al., 2022). Specifically, vehicle type, aerodynamics, and default fuel consumption all have essential roles in measuring overall fuel efficiency for CCL (Kirby et al., 2000). A number of vehicle types exist with different kinds of applied measures for aerodynamics, tires, and powertrain configurations, which are all contributors to fuel consumption (NHTSA, 2010). Proper vehicle selection with enhanced aerodynamics and fuel-efficient configurations can ensure considerable saving of fuel and a lesser impact on the environment in CCL (Smith et al., 2007). The specification of vehicles and their technologies have a large impact on fuel consumption for CCL (Leng et al, 2020). There exist many technologies and techniques developed for improving fuel efficiency in medium and heavy-duty vehicles, including those that optimize engine characteristics, advanced transmission systems, and energy-efficient tires (NHTSA, 2010). Therefore, the use of sophisticated vehicle technologies and following the right specifications can help reduce overall logistics costs, which involve fuel consumption and associated environmental impacts (Leng et al, 2020). Other factors that influence fuel consumption in CCL incorporate the age and maintenance status of a vehicle. According to Kirby et al. 2000, some of the vehicle factors that influence fuel efficiency include a vehicle's age, tire features, and engine features including size and horsepower (Kirby et al., 2000). For example, old motor vehicles may not have upgraded technologies of engines, hence consuming more fuel than newer vehicles with more efficient engines. Moreover, a poorly managed fleet may raise fuel consumption due to factors such as poor tire conditions, higher wear of the engines, and decreased efficiency in the running of components. As such,

optimization of fuel consumption and reduction of environmental impact may be attained through regular fleet maintenance and consideration of fleet age while choosing fleets for CCL operations.

Other relevant factors in fuel consumption for CCL would include refrigeration factors, projected by Capo in 2021 (Capo, 2021). These are container types and capacity of reefer units, which involve great discrepancies. Analysis of the proper selection of reefers will noticeably impact energy consumption and therefore GHC emission, as pointed out by Maiorino et al. 2021 (Maiorino et al., 2021). Understanding the involved trade-offs and complexities of reefer logistics is of prime importance to improve the efficiency of these units. The proper selection of the type of reefer unit and ensuring its efficiency will go a long way in reducing logistic firms' fuel consumption and, therefore, their environmental impact (Maiorino et al., 2021). One of the key factors related to the enclosures of the vehicles used for refrigerated transportation is thermal insulation. This is an important factor not only for cold chain quality but also for saving fuel. To some extent, temperature stability and saving fuel can be done by application of cold-chain insulated containers with phase-change materials (Capo, 2021). Another core factor that influences fuel consumption of CCL involves cargo pre-cooling and temperature control. Objectively, the proper pre-cooling of cargo and maintaining a constant temperature range throughout the transportation process would play a vital role in reducing fuel consumption to a large extent (Behdani et al., 2019). Additionally, the efficient management of the refrigeration system, such as temperature monitoring and control, can contribute to fuel efficiency in CCL (Z. Wang et al., 2020). Comprehensive monitoring and management of refrigeration temperature, selection of efficient refrigeration systems, and implementing new technologies are necessary for an efficient cold chain system.

Next is the environment and weather-related key factor group because it deals with fuel consumption in CCL. Temperature and humidity are the essential and important environmental factors that greatly influence fuel consumption in CCL (Chandran et al., 2022). Additionally, extreme temperatures and high humidity could increase energy usage during the process of transportation for temperature-sensitive products. This is because the energy required to keep the desired level of temperature and humidity in the refrigerated vehicles raises fuel consumption. For instance, it has been estimated that cold chain transportation-in particular, refrigerated semi trailer trucks use 20% more fuel than other modes of transportation (Chandran et al., 2022). The other variables, which raise fuel consumption in CCL, are road types and terrains (Kirby et al., 2000). Where there are steep inclines or high-altitude areas, the vehicle will need more power to attain the required speed and acceleration, hence consuming much fuel (Mills, 2019). Besides, uneven terrain may result in variation in engine load and affect its efficiency (Kirby et al., 2000). Another environmental factor that will affect fuel consumption in CCL is wind and, more generally, air resistance (NHTSA, 2010). Another environmental factor that will affect fuel consumption in CCL is wind and, more generally, air resistance, according to NHTSA, 2010. Because vehicles create some form of air resistance when on the move, the car needs more energy to be waste-resistant; hence, more fuel is used up. Wind direction and speed can still result in fuel inefficiency with headwinds offering more resistance and tailwinds providing a boost in momentum. The National Highway Traffic Safety Administration (NHTSA) has proposed the development of wind-averaged coefficient of drag values using computational fluid dynamics, coast-down, and constant-speed test procedures to better understand how wind and air resistance influence fuel consumption (NHTSA, 2010).

The next in line are route-related aspects. (Wang et al., 2018). Route planning and optimization are very important to the effectiveness of CCL and ultimately to fuel

consumption (Qin et al., 2019). Provided that the route planning was perfectly planned and optimized, it would indeed reduce the distribution cost and carbon emissions of logistics enterprises, thereby promoting their sustainability and eco-friendliness of operations. (J. Chen et al., 2021). In the cold-chain logistics vehicle-routing problem under time window constraints, controlling and limiting carbon emissions are directly related to fuel consumption (Wang et al., 2018). Reducing the total distances in route optimization will significantly reduce fuel consumption (Z. Wang et al., 2020). Customer satisfaction, transportation cost, energy consumption, and time should be put into consideration in the process of making an optimal distribution plan (Xu et al., 2023). Besides, a well-planned route may contribute to avoiding traffic congestion and hence further reduce energy consumption and carbon emissions (Guo et al., 2022). Zhao et al. have pointed out that the traffic pattern and how to avoid the formation of traffic congestion are also large factors affecting fuel consumption in CCL (Zhao et al., 2020). Long driving times imply higher energy use, augmented emission of CO2 equivalent emissions, and a potential loss of food safety due to the prolonged exposure of food to variable temperatures (S. Wang, 2022). Several studies have dealt with traffic patterns and novel strategies for avoiding congestion, which shall also be investigated here to help logistics companies enhance the general operational efficiency of their activities and cut down on fuel use (NHTSA, 2010). Fuel storage and handling practices are also important in the overall fuel consumption of CCL (Rahman et al., 2022). Proper storage and handling can reduce losses and ensure optimal vehicle performance, which may in turn reduce energy consumption and carbon emissions (Rahman et al., 2022). Furthermore, other ways in which transportation is undertaken by the company, like the mode of transport, in intermodal freight transport (IFT), will impact fuel consumption. While IFT may lead to extented transport times, it allows, in most cases, for lower fuel consumption compared to conventional road transport. (Fan et al., 2021).

The load carried by the vehicles also affects fuel consumption; therefore, efficiently managing and distributing cargo loads can lead to fuel savings. The weight and distribution of cargo play a significant role in determining fuel consumption in CCL (Rahman et al., 2022). Proper resource allocation and planning can help reduce the comprehensive cost of cold chain transportation (Xu et al., 2023). The distribution of cargo should be carefully planned to maintain balance, as uneven loads can lead to increased fuel consumption and handling difficulties. Furthermore, optimizing the weight of the cargo can help maintain temperature integrity and reduce energy consumption (Jia, 2022). Multi-stop strategies and consolidation can also affect fuel consumption in CCL, as the frequency and duration of door openings during transportation impact temperature control and energy usage (Tassou et al., 2009). Door opening frequency can lead to temperature fluctuations, which can compromise the quality of perishable goods and increase the need for additional cooling. To mitigate these effects, CCL enterprises should employ efficient routing and delivery strategies, which may include consolidating shipments (Behdani et al., 2019). These approaches can help reduce door opening frequency and duration, thereby minimizing temperature fluctuations and conserving energy (Maiorino et al., 2021). Pallet stacking and utilization are additional factors that can influence fuel consumption in CCL (Fan et al., 2021). Proper pallet stacking can lead to more efficient use of space within the refrigerated container, reducing the need for additional or larger vehicles. This, in turn, can result in lower fuel consumption and reduced carbon emissions (Jia, 2022). Additionally, appropriate pallet stacking can help maintain the integrity of perishable goods by promoting even temperature distribution throughout the container (Mercier et al., 2017). By addressing these factors, CCL enterprises can improve fuel efficiency, reduce costs, and contribute to a more sustainable supply chain (Chandran et al., 2022).

Driving behavior and techniques are other key elements of fuel consumption within CCL (Smith et al., 2007). It is envisaged that increasing drivers' awareness of the factors of fuel-saving techniques, such as maintaining optimal speeds and proper acceleration, will help lower fuel consumption (C. Zhang et al., 2021). A questionnaire-based study indicated that increasing drivers' knowledge with regard to fuel-saving techniques could certainly play a very important role in order to save fuel for CCL as a whole (C. Zhang et al., 2021). One study measured the driving pattern for 15 drivers along a 22-mile stretch of road, matching their driving practices to their fuel use (Mills, 2019). The finding revealed that acceleration, deceleration, and idle time directly impact fuel use. Another factor affecting fuel use by CCL is driving technology (Mills, 2019). Digitalization of the concerned industrial machines, in this case, transport vehicles, can make the vehicle achieve fuel efficiency and hence lead to the solving of challenges in the logistics sector (Al-Wakkal, 2020). Advanced technologies such as telematics systems can monitor and even give real-time feedback on driving behavior, route optimization, and vehicle performance to help reduce fuel consumption. Besides, the adoption of eco-driving technologies in cars can make drivers alter their driving habits to become more fuel-efficient. Driver education and awareness form part of the strategies for tackling fuel consumption issues in CCL (Smith et al., 2007). Proper training can guide drivers on how driving behavior impacts fuel usage and environmental issues, the reason it remains essential to incorporate proper driving practices for fuel efficiency as part of its features (Smith et al., 2007). Besides, awareness can be raised with drivers, so that drivers adopt more environmentally friendly driving habits and reduce fuel consumption, hence being more green in CCL (C. Zhang et al., 2021). They should develop eco-friendly driving habits wherever possible, with techniques proven to help reduce fuel consumption. These include smooth acceleration and deceleration, reducing idle time by turning off the engine during a long stop, constant speed, cruise control whenever possible, and route planning in order to avoid congestion and minimize extra miles of travel. These good practices will help drivers reduce fuel consumption considerably and offer a hand towards a more sustainable CCL process.

Literature reviews with regard to the same factors that affect fuel consumption in CCL come up with a few major factors contributing towards the overall environmental impact of supply chain operations. Economic, environmental, and social effects are among the most dominating issues in CCL, particularly fuel consumption, which has emerged as huge research hotspots with very optimistic prospects. Indeed, the factors that will affect fuel consumption in CCL can only be effectively measured and analyzed using a data-driven approach. Several studies focused on data-driven fuel consumption prediction models, classifying and summarizing data relevant to fuel consumption (D. Zhao et al., 2023). Besides, decision-making in CCL using data analytics has been the focus of recent literature (Chaudhuri et al., 2018). his data-driven approach enables a better comprehension of the relationships between several factors affecting fuel consumption.

Despite advancements in understanding and measuring factors affecting fuel consumption in CCL, several limitations and gaps still exist in the current literature. Most studies have applied factor models to estimate the amount of fuel consumption and carbon emissions, but either those models are not related to CCL or there are not enough fields of data to analyse. Moreover, whereas new ways, such as the location-routing problem-based low-carbon cold chain (LRPLCCC), are proposed, innovation in this line still requires further research and development (Leng et al, 2020). Therefore, future research should seek to address these limitations and gaps in an attempt to improve on the understanding and management of fuel consumption in CCL.

2.4. *Traditional Approaches for Fuel Consumption Estimation*

The traditional methods for fuel consumption estimation in CCL normally include physical models and analytical methods (Rahman et al., 2022). For example, Wang et al. proposed a vehicle routing problem in cold-chain logistics with time window constraints, taking into consideration the controlling and limitation of carbon emissions (Wang et al., 2018). In this respect, various mathematical models were applied in order to study the interrelationship of intelligent logistics, cold chain shipping, and fuel consumption efficiency: from qualitative data analysis to optimization techniques. Furthermore, a comprehensive review of literature on vehicle routing problem (VRP) in CCL reveals a substantial body of research dedicated to understanding and improving fuel consumption in this context (Qin et al., 2019). In the research of Zhao et al., there has been a comprehensive review of the data-driven fuel consumption prediction models by classifying and summarizing the relevant data that affects fuel consumption (D. Zhao et al., 2023). Another example includes the development of a model of fuel consumption prediction using the back-propagation training algorithm for artificial neural networks by Katreddi (Katreddi, 2023). Through all these techniques of data analysis, several researchers learn various factors affecting fuel consumption and thus be in a position to invent more accurate estimation models.

Literature confirms key findings in previous studies that bring out many aspects of fuel consumption efficiency in CCL. For example, the study on national policy and financial intervention in CCL recognized the need to optimize cold chain length for fresh produce to achieve fuel efficiency (Han et al., 2021). Further, studies on the interactions between energy savings and product quality have illustrated the potential for adjustment of vessel speed to maintain a balance in these two conflicting objectives in CCL (Fan et al., 2021). These findings underscore the potential requirement for additional research and innovation in terms of further fuel consumption mitigation in the CCL sector.

For as much as traditional approaches have been potent in fuel consumption estimation in CCL, there are inherent limitations. Among the most significant is the fact that these earlier methods do not exhibit the same levels of continuous improvement, as can be seen with more contemporary methods that now include significant use of Machine Learning technologies (Capo, 2021). Additionally, traditional methods may not be able to account for the wide range of factors that influence fuel consumption in CCL, such as cross-relationship of the criteria and the complex nature of CCL. Besides, basing results on historical data in statistical modeling and its regression analysis methods may limit the correctness in predicting complications in the realistic changeable environment.

2.5. *Machine Learning for Fuel Consumption Prediction*

Machine learning benefits fuel consumption prediction for CCL in a variety of ways. The core feature of modeling and predicting fuel consumption is crucial for enhancing the fuel economy of vehicles and also to detect fraudulent activities in fleet management (Wickramanayake & Bandara, 2016). Machine learning algorithms can analyze large data sets, identify patterns, and make predictions from historical data, thereby helping in increasing the accuracy of fuel consumption estimation that would increase the effectiveness in CCL (Capo, 2021). Machine learning is also able to account for such complex relationships which exist between these variables defining travel economy, vehicle loads, and fuel consumption (Rahman et al., 2022). Some of the advantages of fuel consumption prediction through the use of machine learning include the improvement in the accuracy of the predictions and continuous improvement for each and every data that is uploaded to the dataset to train the model. Predictive analytics is an advanced technique that makes predictions with respect to future events with the utilization of machine learning algorithms, along with historical data. Such requirements of demand necessitate CCL companies to

heavily rely on efficiency, punctuality, and accuracy. An intelligent cold chain management is the one that provides appropriate temperature, vibration, light, and humidity monitoring and control of perishable food during the cold chain (Kale & Patil, 2020).

Various machine learning algorithms have been used for fuel consumption prediction in CCL. Among the popular algorithms are Support Vector Machine (SVM) , regression models, and neural networks (Hamed et al., 2021). Algorithms used in these studies forecast fuel consumption for real-world scenarios with regard to economic travel, vehicle loads, and environmental conditions. Comparing the performance of different algorithms will help these researchers to find out which is most preferably suitable for an application. Thus, it provides more accurate and reliable predictions in cases such as fuel consumption predictions (Wickramanayake & Bandara, 2016). Popular machine learning algorithms for fuel consumption prediction include, Support Vector Machine (SVM), Regression models, Neural networks

According to earlier research, there is a probability that machine learning would enhance fuel consumption estimation in CCL. For instance, another study proposed a machine learning model based on the Support Vector Machine algorithm for vehicle fuel consumption estimation, including some determinants of travel economy and vehicle loads (Hamed et al., 2021). Another study dealt with machine-learning-based predictions that were designed to determine vehicle features with the most relevant effect on fuel consumption in heavy-duty vehicles with real data (Katreddi, 2023). Moreover, the optimization of cold-chain integrated inventory routing problems has also considered carbon emissions, in which attaining accurate fuel consumption prediction could play a significant role in sustaining the environmental performance of CCL (Li et al., 2019). These findings underscore the value of employing advanced machine learning methods to address the challenges of fuel consumption estimation and optimization in CCL.

2.6. Existing Research Gaps and Challenges

One of the most important gaps exists in the comprehensive application of machine learning and datasets for the training of models in CCL research. Efficiency and usefulness of machine learning models depend to large extends on the amount and better quality of data available for the training and validation purposes. Much of the researchvv that has been done is concentrated on a few selected factors or limited datasets, and this might not be the case in the real world due to the diverse factors surrounding fuel consumption. In view of this, it is necessary for the development of more extended and diversity-filled datasets that enable researchers to generate more accurate and robust machine learning models. In case of availability of such datasets, it would allow the research community to investigate the interdependencies between and interactions among multiple factors affecting fuel consumption in CCL (Chaudhuri et al., 2018).

Another research gap in this area is estimating fuel consumption in the CCL operational cycle by considering real-time data and external contingencies in the machine learning models developed. Current efforts are tending towards repeating historical data or making theoretical assumptions that do not always adapt to the dynamic character of the CCL operational cycle (Chen, 2020). For instance, traffic triggers, weather conditions, vehicle maintenance timings, are all very influential in fuel consumption and must all be included when coming up with models for such analysis (Katreddi, 2023). Additionally, with the rapid development in technology, specifically big data analytics, this limitation can be conquered using real-time monitoring logistics to build the model that will prove to be the most influential factors behind the delay or inefficiencies (Chen, 2020). If real-time data and

external factors could be integrated into the machine learning models, it may increase the precision and accord practical applicability to a significant level in industry.

Lastly, addressing the interpretability and explainability of machine learning models is another challenge that needs to be overcome in the field of fuel consumption estimation in CCL. While machine learning models have demonstrated their potential in providing accurate predictions, their complexity often makes it difficult for practitioners to understand and trust their results (Capo, 2021). This lack of transparency may lead to hesitance in adopting these models in real-world applications, despite their potential benefits. Therefore, research should focus on developing machine learning models that are not only accurate but also interpretable and explainable, enabling users to gain insights into the underlying relationships between various factors affecting fuel consumption (Chen, 2020). This would ultimately contribute to the successful implementation of machine learning models in CCL and help the industry optimize fuel consumption and reduce its environmental impact.

3. Methodology

3.1. Emprical Analysis

A typical day in CCL, for the delivery of commodities from a morning warehouse to the destination and back to the evening warehouse, undergoes a well-orchestrated process just to keep temperature-sensitive commodities within an environment that is controlled. Loading the goods into the refrigerated trucks with great care in the morning is necessary, where storing and tying is essential. These are accompanied along the way by temperature monitoring devices and data loggers, which record conditions to ensure the cold chain is maintained. On arrival at destination, efficient offloading and unloading procedures minimize exposure to ambient temperatures. Deliveries of goods to locations while documentation, if any, is handled efficiently. In the evening, empty trucks return to a cleansing process at the warehouse, after which they are sent out again the following morning. Here also, like in every step, number one is cold chain compliance—bringing the shipment to its destination with an optimum level of fuel consumption, freshness, and safety. Fuel consumption becomes very critical here, typically measured by liter of diesel fuel. Adding to this, in regard to fuel consumption, metrics are taken into consideration. They pose a great impact on the operation. In general, they involve factors related to Vehicle, Refrigeration and Temperature, Route, Load, and Driving behaviour. Among those numerous metrics, some are considered due to data tracking, measurability, and other limitations. The 22 factors affecting diesel consumption that need to be accounted for are given below with explanation:

3.1.1. Vehicle Related Factors

- ***Default Fuel Consumption (1):*** Default fuel consumption refers to the baseline fuel consumption rate of a vehicle which specified by the manufacturer. Various elements contribute to this consumption, including engine size, aerodynamic drag, and the presence of outdated technology in older vehicles. The reason why this factor is taken as default is that varying factors such as vehicle brand, model, aerodynamic effects and engine size are gathered under one roof. It represents the amount of fuel consumed per unit distance traveled and is typically measured in liters per kilometer (L/100km).
- ***Vehicle Age (2):*** The production year of a vehicle indicates its age and technological advancements. Older vehicles tend to have higher fuel consumption due to wear and tear, which can affect the vehicle's efficiency. As vehicles age, their engines may lose efficiency, and the vehicle's aerodynamics may degrade, leading to increased air resistance and higher fuel consumption. Additionally, newer vehicles often have more

efficient engine technologies, resulting in lower fuel consumption compared to older models.. The production year is typically measured in years (yr).

- ***Vehicle Mileage (3):*** Vehicle mileage refers to the total distance that a vehicle has traveled over its lifetime. Higher mileage can lead to increased wear and tear, potentially affecting fuel efficiency. Vehicle mileage is typically measured in kilometers (km).
- ***Vehicle Maintenance (4):*** Vehicle maintenance encompasses regular upkeep and servicing of vehicles. Proper maintenance ensures optimal performance and fuel efficiency by addressing factors such as tire conditions, engine wear, and component and coolant operation. Vehicle maintenance is crucial for minimizing fuel consumption and is measured based on last maintenance schedules and records in terms of weeks(w).
- ***Vehicle Volume (5):*** Vehicle volume refers to the physical capacity or size of the vehicle's cargo space. It is an important factor in determining the load capacity and load utilization in CCL. Vehicle volume is measured in cubic meters (m^3)

3.1.2. Refrigeration and Temperature Related Factors

- ***Ambient (Max) Temperature (7):*** Ambiento temperature is the temperature of the outside environment. It might have an impact, generally, on the energy demand by CCL refrigeration systems. At very high or low temperatures, the energy required to maintain the exact level of the temperature in refrigerated vehicles increases. It is measured in degrees Celsius ℃.
- ***Morning (Min) Temperature (8):*** Morning temperature refers to the minimum temperature of the day of delivery, usually measured at the start of a work day in the morning. This is particularly important in the case of early start-up of the cooler where goods are being transported either at +4 ℃ or at -18 ℃. The morning temperature will be measured in degrees Celsius ℃.
- ***Frozen Cabin (9):*** Cabin Frozen is a parameter depicting the temperature setting inside the vehicle. "0-1" Variable transformations for Cabin Frozen had been done prior to sampling. If it takes the value 1, that would be -18°C for frozen goods to really ensure product preservation that requires ultra-low temperature storage. And 0 is for +4°C, which is for chilled products not requiring frozen conditions but still needing controlled cooling. This Frozen Cabin feature must make available an estimate of energy consumption with respect to the quality and safety preservation of temperature-sensitive goods while in transit.
- ***Double Cabin (10):*** A vehicle double cabin is an additional inner insulated enclosure that should ideally maintain a constant cold temperature throughout the delivery process. This feature of the double cabin adds to the overall vehicle weight and thus affects its fuel efficiency. However, the sealed cabin door ensures less loss of cold temperatures. It might find a mention as one of the critical parameters defining the fuel efficiency and load-carrying capacity of the vehicle. In the case of "Double Cabin", "0-1" variable transformation was done before the data collection phase. The vehicles having that feature were assigned with the value "1" and vehicles without that feature were assigned with a value of "0".
- ***Cooling Stem Time (11):*** Cooling stem time refers to the duration required for the refrigeration system to reach and stabilize at the desired temperature range before loading the temperature-sensitive goods. It is an important factor in ensuring that the products are properly loaded into cooled environment before transportation. Cooling stem time is measured in minutes (min).

3.1.3. Route Related Factors

- ***Route Distance Forward (12):*** The total distance covered from the warehouse in the morning to the last destination of trònsporUED. This factor is very key to fuel consumption and gives an insight into how efficient the route logistics is. It is measured in kilometers (km).
- ***Route Distance Return (13):*** This is the total distance covered from the destination back to the warehouse at the close of a distribution process, accounting for miles covered where there is no coolant activity during the return journey. It is, however considered in fuel consumption and in calculating optimal logistics routes. Route distance return is measured in kilometers (km).
- ***Route Time Forward (14):*** The time consumed from the warehouse in the morning to the final destination. It is very critical information in planning and scheduling logistics operations and may impact fuel consumption. It is measured in minutes (min).
- ***Route Time Return (15):*** Route time return is the duration taken to back tract and reach the warehouse at the end of the distribution process. It represents time spent during the return journey without any coolant activity, stipulated in logistic planning and fuel consumption computations. Route time return is expressed in minutes (min).

3.1.4. Load Related Factors

- ***Load Weight (Payload) (16):*** The load weight or simply the payload represents the total weight of temperature-critical cargo to be moved. It is one of the principal factors when estimating vehicle capacity and fuel economy. This is normally given in kilograms (kg).
- ***Delivery Points (17):*** These are the total number of locations or, said differently, a number of stops that the vehicle has to make during the process of transportation. Each stop adds additional time, distance to the route, with an added opening coolant door that potentially impacts fuel consumption. The number of stops is measured as a numerical value (#).
- ***Load & Unload Time (18):*** This is the time consumed for the loading and offloading of the temperature-critical shipments from the truck. Effective load and unload processes waste less idle time and contribute to the overall efficiency of CCL. This may generally be measured in minutes (min).

3.1.5. Driving Related Factors

- ***Average Speed (19):*** Average speed refers to the average rate at which a vehicle travels during the process of transportation. It is very important in fuel consumption since high speeds would generally result in a high aerodynamic drag and high requirements for energy. Average speed is normally measured in kilometers per hour (km/h).
- ***Max Speed (20):*** This is the top speed that can be built up by the vehicle. It states that it is a very important parameter not only for drivers to safely deliver their goods but also not to be an inconvenience. Overspending too much may cost a waste of fuel and can be hazardous as well. Max speed is measured in kilometers per hour (km/h).
- ***Idle Time (engine off) (21):*** Idle time with engine off refers to the period of time when the vehicle is still standing with its engine turned off in the course of transportation. Since the air conditioner used in automobiles is engine driven, the air conditioner will not run during Idle Time with engine off period and will not consume fuel. However, since this is going to raise the temperature of the cabin, this period is kept short not to allow overshooting beyond the limit temperature. Any type of break or parking with an engine-off condition during working time is considered as Idle Time (engine off). The time for idle time with the engine off is measured in minutes (min).

- ***Idle Time (engine on) (22):*** Idle time with the engine on refers to the standstill time of the vehicle during transport that has an engine running. Since the vehicle air conditioner is driven by the engine, the air conditioner will be running and consuming energy as long as the engine is on. Too much idling leads to fuel waste and unnecessary emissions. Because reducing idle time under the engine increases fuel economy. Idle time under the engine is recorded in minutes (min).

3.2. Data Collection and Sampling Methods

Data were gathered over a span of 10 weeks (from weeks 5 to 14) from the 14 subcontractor dealership vans in Izmir, Turkey. The period of time is applicable for a relevant data analysis, because the period is not affected by some extraordinary events, such as bank holidays, and bad weather. Moreover, the stability of seasonal daily density during this time period serves to minimize the impact of external factors on the collected data.

The mentioned vans only deliver frozen food products such as frozen pizzas or potatoes or chilled nourishment, such as dairy products like cheese or yoghurt and cold cuts. Each week 2 to 4 days, products from different vendors arrive to the dealership. If the products are not going to be delivered same day, they are taken into the temperature controlled cool room or freezers. Usually, each truck has two employees who are in charge of managing, driving, and providing delivery assistance. When there is a lack of presence, an additional worker is hired, or else a single person completes the work to keep the firm running. As a result, business processes continued to operate normally over the monitoring time.

Initially, driver surveys and personal interviews with drivers and logistics managers were conducted to obtain some heuristic insights into the operational procedures and probable inefficiencies. These realizations have been very basic in the interpretation of quantitative data, correlating the understanding of the bigger picture with respect to the fuel consumption trends within the framework of CCL. Data collection from different sensors and monitoring devices mounted on every vehicle followed, recording variables such as fuel consumption, ambient temperature, route distance, vehicle speed, and others for a given period.

3.3. Statistical Tools Used for Analysis

- ***Python:*** (Programming Language for training, testing, and validation). Python is an extremely versatile and powerful programming language heavily used in data science and machine learning; it owes its flexibility and usability to the collection of libraries. This eases a nice running of several ML models where it gives support to carrying out the implementation of ML techniques that are possible to define and train in a way that the code is clear and easy to understand. Python also allows the easier validation and testing of the models. Such cross-validated approaches through model assessment and division of data into training and test sets make model outputs robust and easily transferable to fresh data. The handling of data by Python is also made easy, mainly through libraries such as Pandas and Numpy, which go a long way in the effective handling of large datasets. This is mainly by performing operations like cleaning and transforming data, data aggregation, and other vital processes that require data preparation before it is fed into a machine learning model.
- ***JupyterLab:*** (Text editor for Python). JupyterLab is a next-generation web-based user interface for Project Jupyter and is ideally an integrated development environment online which can provide support to data science, machine learning work. JupyterLab is a high-quality non-prose writing environment for Python code. Execution can then be completed to the nearest tick of any single cell in order to support a test and debug environment. It also supports rich text formatting and even has a feature for more advanced visualizations, which is used to display workflows in diagrams and results.

This also can be useful in order to guarantee analysis is reproducible when working with Jupyter Notebooks. Following from data preprocessing to model evaluation everything can be saved and shared with others in order to get the most out of collaboration and transparency.

- ***Pandas:*** (Data processing library). Pandas is an open source Softwares library, developed to be used in Python for data analysis and manipulation. Handling of missing data, removal of redundant information to save memory, rectifying inconsistencies in the data so that the dataset be clean and reliable before getting it fed into the machine learning models. Going from one data format to another, combining different datasets, and transforming data are common tasks in any analysis or modelling process. Pandas also allows one to get basic statistics and perform data exploration of the dataset, which gives pre-insight into how the dataset is patterned. It helps to recognize the significant factors impacting fuel consumption in cold chain logistics using this initial analysis.
- ***Numpy:*** (Numerical calculation library). Numpy is a base package for scientific computing with Python. It supports large, multidimensional arrays and matrices, along with mathematical functions that operate on them, including effective statistical computations for mean, median, minimum, maximum, and standard deviation. Not only that, it also houses a series of mathematical functions to perform various mathematical functions, such as trigonometric, statistical, and algebraic, to carry out complex calculations with data. Moreover, Numpy will ensure top performance so that you gain quick computation even with bulky datasets; this is pretty much necessary in big data processing.
- ***SciPy:*** (Scientific and technical computing library). SciPy works in parallel with Numpy as an extension for scientific and technical computing. It contains modules on optimization, integration, interpolation, and innumerable other complex mathematical functions in module form. Among its modules are tools for optimization, integration, and interpolation, which help in modeling and refining the models, whereas for any complex mathematical operation one requires in the analysis, it provides the same too. It also has modules for signal processing, which will, in turn, get help for time series, and statistical tools, along with hypothesis testing and other kinds of statistical analysis, which will help you validate your findings.
- ***ScikitLearn:*** (Machine learning library). ScikitLearn is a Python machine learning library with tools for data mining and data analysis on a general-purpose foundation formed by NumPy, SciPy, and Matplotlib. ScikitLearn provides a host of machine learning algorithms, among which is linear regression, whose application in fuel consumption estimation will be discussed. Secondly, it also provides a consistent interface for using these machine learning algorithms, which will allow the user to easily change between several models. ScikitLearn provides the ability to create machine learning pipelines. This capability makes the workflow from data preprocessing to model training and assessing very smooth, with the assurance that each of the steps is consistent and repeatable. Further, ScikitLearn provides some metrics for the model performance. Such metrics include mean squared error, R-squared, and different cross-validated scores, which help to assess the accuracy and generalizability of the models.
- ***Matplotlib:*** (Data visualization library including seaborn). It's a plotting library for the Python programming language. Matplotlib offers its users total control over plots they create, and these include static, animated, and interactive visualizations. One can use Matplotlib to produce a wide range of plots, from line graphs and histograms to bar charts and scatter plots. They help in exploring and giving better understanding of the data, and hence one can communicate the findings effectively. Tools for visualization

help in understanding the result of models. For example, one could plot the prediction versus the real fuel consumption in order to see how the regression models are doing. Visualization before and after processing gives an insight into the distribution of data, trends, and anomalies in the data. Seaborn is a library based on Matplotlib and is closely integrated with Pandas data structures. It provides a high-level interface for drawing informative and attractive statistical graphics.

3.4. Overview of Machine Learning Models

The consumption of fuel is analyzed for cold chain logistics; several linear regression models are considered for the selection of significant factors and the estimation of fuel consumption. Linear regression models are some primary tools in the field of statistical modeling and machine learning that model the relationship between a dependent variable and one or several independent variables. Below is explained in detail how each linear regression model has been adapted; after that, the equation and the terms involved for each one are found accordingly.

3.4.1. Simple Linear Regression:

Linear regression model is the simplest machine learning and statistical technique on the topic. That describes the relationship between one dependent variable and one or more independent variables. The process is the definition of a linear relationship for the predicted dependent variable with respect to independent variables. It is used so generally for reasons of its simplicity, interpretability, and because it gets tedious for many practical applications. The linear regression model can be best denoted using the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(Eq. 1)}$$

- y is the dependent variable (response variable). y
- $x_1, x_2, \ldots, x_p$ are the independent variables (predictors).
- β_0 the intercept term, representing the expected value of y when all independent variables are zero.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (weights) associated with the independent variables, indicating the change in y for a one-unit change in the corresponding x.
- ϵ is the error term, capturing the variability in y that cannot be explained by the linear relationship with the independent variables.

The coefficients $\beta_1, \beta_2, \ldots, \beta_p$ in the linear regression model are typically estimated using the method of least squares. This method minimizes the sum of the squared differences between the observed values and the values predicted by the model. The objective function to be minimized is:

$$\text{Minimize: } \{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p)^2\} \qquad \text{(Eq. 2)}$$

- n is the number of observations
- y_i is the observed value of the dependent variable for the i-th observation
- $\hat{y}_i$ is the predicted value from the linear regression model.

Assumptions: For linear regression to produce reliable and meaningful results, several assumptions must be met:

1. Linearity: The relationship between the dependent variable and the independent variables is linear.

2. Independence: The observations are independent of each other.
3. Homoscedasticity: The residuals (errors) have constant variance at all levels of x.
4. Normality: The residuals of the model are normally distributed.
5. No Multicollinearity: The independent variables are not highly correlated with each other.

Violations of these assumptions can lead to biased or inefficient estimates, reducing the reliability of the model's predictions.

3.4.2. PCR (Principal Component Regression):

Principal Component Regression (PCR) is a hybrid technique that combines the principles of Principal Component Analysis (PCA) and linear regression. It was specifically intended to address multicollinearity (when variables are highly correlated with each other) in datasets and reduce dimensionality, thereby stabilizing and increasing the interpretability of the regression models. PCR is particularly helpful when dealing with large datasets having many predictors, sometimes highly correlated. PCR avoids the pitfalls of multicollinearity and at the same time maximizes the predictive performance of the regression model by transforming the original predictors to a subset of uncorrelated components. Accordingly, the number of the predictors in the model is decreased due to suppression of the multicollinearity effect. Some of the benefits acquired from using PCR include effective multicollinearity management, effective reduction in the number of predictors, ease in the model, and an increase in ease in computation. However, shrinking many original predictors into few components makes it hard to interpret regression coefficients, and having too few components might result in losing important information. The two major steps in a PCR are PCA followed by linear regression.

a. *Principal Component Analysis (PCA):* PCA transforms the original set of predictors into a new set of uncorrelated variables called principal components. These components are linear combinations of the original predictors and capture the maximum variance in the data.

Let X be an $n \times p$ matrix representing the n observations and p predictors. PCA decomposes Xinto a set of principal components Z:

$$Z = XW \qquad \textbf{(Eq. 3)}$$

- W is a $p \times p$ matrix of eigenvectors of the covariance matrix $X^T X$
- Z is a $n \times p$ matrix of principal components.

a. *Linear Regression on Principal Components:* Once the principal components are obtained, linear regression is performed using these components as predictors instead of the original variables. The linear regression model can be expressed as:

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k + \epsilon \qquad \text{(Eq. 4)}$$

- $Z_1, Z_2, \ldots, Z_k$ are the first k principal components.

3.4.3. PLS (Partial Least Square Regression):

Partial Least Squares Regression (PLS) is a technique that combines the characteristics of Principal Component Analysis and multiple linear regression at the same time. The PLS method is custom-made for datasets in where there are many, possibly colinear, predictors. It looks for new components that explain the predictors at the same time as it predicts the

response variable well. PLS greatly helps when predictors are strongly collinear with one another or when the amount of predictors in a dataset largely surpasses the number of instances. PLS decomposes the predictor and response variable into latent structures that maximize the covariance between them. In fact, it works quite well on the multi-collinearity aspect, the predictors' number is reduced since it simplifies the model, and hence it is good for computational efficiency. But most importantly, the predictions are quite robust. However, as in PCR, the regression coefficients are very difficult to interpret, due to the construction of the latent variables, and the number of components requires a bit of judgment. The main steps in PLS are as follows:

a. *Latent Variables*: PLS finds latent variables (components) that are linear combinations of the original predictors and have the highest covariance with the response variable.

Let X be an $n \times p$ matrix representing the n observations and p predictors. PCA decomposes Xinto a set of principal components Z:

$$X = TP\,T\, + E \quad \textbf{(Eq. 5)}$$

$$y = UqT\, + F \quad \textbf{(Eq. 6)}$$

- X is a $n \times p$ matrix of predictors and y is an $n \times 1$ vector of responses
- T is a $n \times k$ matrix of scores for predictors.
- P is a $p \times k$ matrix of loadings for predictors.
- E is the matrix of residuals for predictors.
- U is a $n \times k$ matrix of scores for the response.
- q is a vector of loadings for the response.
- F is the vector of residuals for the response.
- k is the number of latent variables.

a. *Regression on Latent Variables*: Once the latent variables are obtained, a linear regression model is fit using these components:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i T_i + \epsilon \quad \textbf{(Eq. 7)}$$

- T_i are the latent variables obtained from PLS.

3.4.4. Ridge Regression

Ridge regression, better known as Tikhonov regularization, is the method that uses multicollinearity-skewed multiple regression data. Existence of highly-correlated independent variables may make the OLS model very sensitive to small variations and hence undergo a large variance. Ridge regression fits a model and adds some bias to give a more definite reduction in standard errors. Ridge regression modifies the least squares objective function by attaching a penalty term that is proportional to the square of the magnitude of the coefficients. Ridge regression is particularly good at combating multicollinearity and enhances the prediction accuracy and stability because it adds a regularization term. However, Ridge regression adds bias into the model and does not perform feature selection since it does not put any coefficients exactly to zero. The ridge regression model is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad \textbf{(Eq. 8)}$$

$$\text{Minimize: } \{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 \} \quad \textbf{(Eq. 9)}$$

- y_i is the observed value of the dependent variable.
- x_{ij} are the observed values of the independent variables.
- λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients.

The strength of the penalty is determined by the degree controlled by λ. When λ=0, the case of ridge regression reduces to ordinary least squares regression. If λ is large, then the magnitude of the coefficients shrinks toward zero but does not reach exactly zero. By adding this penalty term in the cost function, ridge regression shrinks the coefficients, and hence it reduces their variance and mitigates effects from multicollinearity.

3.4.5. Lasso Regression:

Lasso regression, which stands for Least Absolute Shrinkage and Selection Operator, is a linear regression form containing L1 regularization. Lasso helps in shrinking the features to zero to help reduce variance and is also instrumental in feature selection by shrinking some coefficients right to zero. Lasso regression has many applications when we work with high-dimensional data, which would require feature selection. In performing lasso regression, it selects automatically in-built features and trumps multicollinearity by driving some of the coefficient estimates right to zero. It can, therefore, simplify the model and enhance the prediction accuracy. On the other hand, this regularization introduces bias that sometimes results in inconsistent variable selection, especially for predictors that are highly correlated. The lasso regression model can be put in the regular linear regression form, and the lasso regression to be minimized in the cost function contains the L1 penalty term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad \textbf{(Eq. 10)}$$

$$\text{Minimize: } \{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \} \quad \textbf{(Eq. 11)}$$

- y_i is the observed value of the dependent variable.
- x_{ij} are the observed values of the independent variables.
- λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients.

The L1 penalty in the cost function causes some of the coefficients to be exactly zero, which implements feature selection automatically and simplifies the model. This characteristic of Lasso allows shrinking of the coefficients to zero, and this is very useful in high dimensions. It retains only important predictors that help in identifying them. By introducing bias through regularization, lasso reduces the variance of the model that might improve generalizability and, therefore, performance on unseen data.

3.4.6. ElasticNet Regression:

ElasticNet regression is a type of regularized regression technique that improves ridge regression and lasso regression. It overcomes the drawbacks of both approaches, so it is most effective in situations where the datasets are characterized by having several relevant features that present a high level of multicollinearity or when the number of predictors is vastly greater than the number of observations. ElasticNet does a pretty good job of putting together

variable selection and regularization in one step, thus enhancing the prediction accuracy and interpretability of the model. ElasticNet applies both the L1 and the L2 penalty as a linear combination of Lasso and Ridge, respectively. In other words, ElasticNet is a very good approach to handle multicollinearity and do feature selection at the same time. The mix of penalties in ElasticNet is more effective in handling multicollinearity and in automating the feature selection but increases the complexity of the problem by tuning two hyperparameters, which turns out to be computationally intensive. The cost function for the optimization problem is written as the sum of the penalties of ridge regression and lasso regression. The ElasticNet model can be expressed by the linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(Eq. 12)}$$

$$\text{Minimize: } \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \} \qquad \textbf{(Eq. 13)}$$

- y_i is the observed value of the dependent variable.
- x_{ij} are the observed values of the independent variables.
- λ_1 is the regularization parameter for the ridge component (L2 penalty).
- λ_2 is the regularization parameter for the lasso component (L1 penalty).

3.4.7. Lars Regression:

Least Angle Regression (LARS) is an algorithm designed for regression against high-dimensional data with multicollinearity among predictors. It will be efficient under the condition: the number of predictors should be much larger than the number of observations. LARS is an algorithm that works iteratively over the predictors. It also provides a further less-greedy version of forward stepwise regression, which is computationally efficient and gives a full solution path with the variation of the regularization parameter. In this manner, LARS can build a linear model iteratively by selecting the most correlated predictors with the response and then adjusting the coefficients. LARS will provide a full solution path for dataset size, and a model can be determined in this way to study predictor inclusion and further model selection. This feature of the LARS model usually makes it harder to implement or interpret than the simpler regression methods. Of course, the general linear model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(Eq. 12)}$$

However, the algorithm for determining the coefficients β_j is unique to LARS. The algorithm proceeds as follows:

1. Initialization: Start with all coefficients $\beta_j = 0$
2. Iteration: At each step, identify the predictor most correlated with the residual from the previous step. Increase the coefficient of this predictor in the direction that reduces the residual sum of squares.
3. Pathwise Solution: Continue this process, moving the coefficients in the direction of the identified predictor until another predictor becomes equally correlated with the residual. Then, move in the direction equiangular between the predictors.

The process can be summarized by the following equations and steps:

$$\hat{y} = X\hat{\beta} \quad \textbf{(Eq. 13)}$$

- X is the matrix of predictors.
- $\hat{\beta}$ is the vector of estimated coefficients.
- $\hat{y}$ is the vector of predicted values.

At each step, the algorithm adjusts the coefficient β_j such that:

$$\beta_j^{(k+1)} = \beta_j^k + \gamma \quad \textbf{(Eq. 14)}$$

where γ is the step size determined based on the correlation of the predictors with the current residuals.

LARS is computationally efficient, especially for high-dimensional data, where the number of predictors is large. It provides a full piecewise linear solution path, which enables one to look at the whole sequence of feasible models as the regularization parameter changes. It can easily be modified to give solutions comparable to the lasso and forward stagewise regression with modifications to step size and selection criteria.

3.4.8. LassoLars Regression:

LassoLars stands for Least Angle Regression with Lasso, which is a refitted version of Lasso regression, combined with the efficiency of the LARS algorithm. It is designed to provide the benefits of the two: variable selection and regularization same as Lasso and being computationally efficient while providing a full path of solutions like LARS. This technique is very useful in high-dimensional data; that is, when the number of predictors in the data is very large. This ensures that some coefficients will be exactly equal to zero, making model simplification and interpretation possible. However, this implies very careful parameter-tuning considerations which, in turn, could be computationally intensive. The basic linear regression model that LassoLars follows is as shown:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad \textbf{(Eq. 15)}$$

$$\text{Minimize: } \{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \} \quad \textbf{(Eq. 16)}$$

LassoLars is a modification of the LARS algorithm to introduce some of the coefficients to be exactly zero due to the L1 penalty. From Lasso regression, it has the nature of the L1 penalty, which shrinks some coefficients toward zero for variable selection. LARS Algorithm: This is very useful and efficient in high dimensions computationally. It provides the complete path of the solution, which allows examination of a model for various levels of regularization and, more generally, in an environment for model selection. It will return piecewise linear coefficient paths so that the process by which predictors are added into the model may be further explained.

3.4.9. LassoLarsIC Regression:

LassoLarsIC extends the LassoLars to include the automatic selection of the regularization parameter using information criteria, such as AIC or BIC. The procedure thus integrates within a single tool the variable selection, the regularization properties of the

Lasso, the computational efficiency of LARS, and the automatic selection of the model complexity driven by the data. This automatically selects the regularization parameter using information criteria to avoid manual tuning or cross-validation and make the model more interpretable and computationally efficient. It extends LassoLarsIC along the same linear regression model of LassoLars because the cost function contains the Lasso regression L1-term:

LassoLarsIC selects the regularization parameter λ based on minimizing an information criterion, such as AIC or BIC:

1. Akaike Information Criterion (AIC): AIC balances the goodness of fit of the model with the complexity of the model. It is defined as:

$$AIC = 2k - 2ln(L) \quad \textbf{(Eq. 17)}$$

where k is the number of parameters in the model and L is the likelihood of the model.

2. Bayesian Information Criterion (BIC): BIC is similar to AIC but includes a stronger penalty for models with more parameters. It is defined as:

$$BIC = kln(n) - 2ln(L) \quad \textbf{(Eq. 18)}$$

where n is the number of observations.

The LassoLarsIC chooses the value of which minimizes one of the information criteria AIC or BIC. It basically searches for a model with good balance between complexity. This would mean that one is able to avoid the manual tuning implicit in using cross-validation. Includes the L1 penalty, which enables variable selection through shrinking some of the coefficients to zero. It applies the LARS algorithm; thus it is computationally efficient and suitable for high-dimensional data. Uses information criteria to balance model fit and complexity, which leads to more interpretable models.

3.4.10. BayesianRidge Regression

Bayesian Ridge Regression is a linear regression technique that adopts a Bayesian inference approach in estimating the distribution of the model parameters. It is a probabilistic regression approach that allows one to estimate the uncertainty in the model coefficients and regularizes to avoid overfitting. Further, Bayesian Ridge Regression helps in dealing with multicollinearity in the correct manner and makes the model robust. Bayesian Ridge Regression further introduces regularization via specification of a prior on both the parameters and the noise term, thereby making it better in terms of understanding and estimating uncertainty while obtaining a less sensitive model to multicollinearity and overfitting. In what follows, we derive Bayesian Ridge Regression as an extension of the standard linear regression model with priors placed on the model coefficients and the noise term:

In Bayesian Ridge Regression, priors are placed on the coefficients βj and the variance of the error term σ^2

$$\beta_j \sim N(0, \lambda^2) \quad , \quad \sigma^2 \sim InverseGamma(\alpha, \beta) \quad \textbf{(Eq. 19)}$$

where λ is the precision of the prior distribution for β_j (inverse variance). α and β are the hyperparameters of the Inverse Gamma distribution controlling the prior on σ^2.
The likelihood of the observed data is assumed to be Gaussian:

$$y_i \sim N(X_i\beta, \sigma^2) \quad \textbf{(Eq. 20)}$$

where X_i is the i-th row of the design matrix X. Incorporates prior distributions for model parameters, allowing for probabilistic interpretation and uncertainty estimation. The priors on the coefficients introduce regularization, helping to prevent overfitting and handle multicollinearity. Bayesian Ridge Regression automatically estimates the regularization parameters from the data, improving model robustness.

3.4.11. ARDRegression Regression:

ARDRegression is a Bayesian linear regression technique that tries to make a decision on the importance of predictors by putting a different type of prior distribution on the coefficients. Such models are very useful for high-dimensional data when it is important to retain only the most relevant predictors. ARDRegression prevents overfitting by adapting the precision parameters to shrink the coefficients of less relevant variables toward zero. Inclusion of priors in relevance determination through ARDRegression increases the robustness of the model, particularly in complex datasets. The ARD model extends the standard linear regression model by priors for the regression coefficients.

In ARDRegression, the coefficients β_j are treated as random variables with their own prior distributions. Typically, Gaussian priors are used:

$$\beta_j \sim N(0, \alpha_j^{-1}) \quad \textbf{(Eq. 21)}$$

Where α is the precision (inverse variance) of the prior distribution for β_j. A high value of α_j indicates that β_j is likely to be close to zero, thus deeming the corresponding predictor less relevant.

The likelihood of the observed data is assumed to be Gaussian:

$$y_i \sim N(X_i\beta, \sigma^2) \quad \textbf{(Eq. 22)}$$

where Xi is the i-th row of the design matrix X, and σ^2 is the variance of the error term. ARDRegression uses a Bayesian framework to estimate the coefficients, incorporating prior knowledge about their distribution. The model automatically adjusts the precision parameters α_j to identify and shrink the coefficients of less relevant predictors towards zero.

3.5. Model Evaluations (Success Metrics)

Once such regression model coefficients are estimated, the performance of the model has to be checked for validity and reliability. The model evaluation metrics provide some quantitative guidance to the states about how well the model fits the data and how well it probably generalizes to new data. Among the most common evaluation metrics for linear regression models are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). One cannot consider any of the three measures superior to the other since all three provide related information on different aspects of the performance of a model.

Mean Squared Error (MSE)

The Mean Squared Error (MSE) is one of the base measurements that is utilized to approximate the competency of a regression model that computes the mean of the squares of the amount of error—the difference of the observed true outcome—and the outcome that the model has estimated. Mathematically, MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \textbf{(Eq. 23)}$$

where n is the number of observations, y_i is the observed actual, $\hat{y}_i$ is the predicted value for the i-th observation. MSE generally shows how far the prediction errors deviate and, because it involves squaring, gives more weight to the larger errors. Therefore, a smaller MSE value indicates a closer fit of the model to data since the predicted values are much closer to the true value.

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is another commonly used metric that is derived from MSE. RMSE is the square root of the Mean Squared Error and provides an error metric in the same units as the dependent variable, making it more interpretable. The formula for RMSE is:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \textbf{(Eq. 24)}$$

RMSE, like MSE, reflects the average magnitude of the prediction errors, but it is more interpretable because it is in the same units as the dependent variable. A lower RMSE indicates better predictive accuracy of the model.

R-squared (R^2)

R-squared (R^2), also known as the coefficient of determination, is a metric that quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides an indication of the goodness of fit of the model. The formula for R-squared is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad \textbf{(Eq. 24)}$$

where $\bar{y}$ is the mean of the observed actual values.

The R^2 values are from 0 to 1. An R^2 value equal to 1 states the model explains all the variability of the response data 1 and the mean, thus perfectly fitting the model. An R^2 value of 0 states that the model does not explain any of the variability in the response data. Taking it more generally, the better the model fits, the more significant the R^2 value will be. One disadvantage of R^2 is that it does not adjust for the number of predictors within the model; therefore, it may suggest a too-optimistic fit in cases where too many predictors have been included in the model. Adjusted R^2 overcomes this weakness by controlling for the number of predictors in the model relative to the number of observations.

4. Data Analysis of Machine Learning Model for Fuel Consumption Estimation

4.1. Data Preparation and Descriptive Statistics

The primary step in the data preparation is to load the dataset. This is done using the pandas library, which is very efficient for the manipulation and analysis of data and designed for Python. The next step is to read a subset of data in columns 2 to 25 from the CSV file. The next 700 rows are selected, then read into a DataFrame for further analysis.

An overview of the data from df.info() command can help us to know what data structures and types exist in the data and how many non-null entries exist in the DataFrame. This overview helps in understanding the structure of the dataset, identifying the types of data present (numerical, categorical, etc.), and detecting any missing values. In this way, we could detect if some variables need to be transformed or reduced. However, since data transformation has been completed before the data collection, there were no need to transform or reduce the data. Example of data transformation can be given for cabin type and frozen cabin data. While these variables were nominal, the data has been collected as binary values (0-1).

Figure 1**: Overview of the Data**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Data columns (total 22 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   default_fc    700 non-null    float64
 1   age_van       700 non-null    int64
 2   mileage       700 non-null    int64
 3   maint_wk      700 non-null    int64
 4   volume        700 non-null    int64
 5   doubl_cab     700 non-null    int64
 6   frozn_cab     700 non-null    int64
 7   max_temp      700 non-null    int64
 8   min_temp      700 non-null    int64
 9   cool_stem     689 non-null    float64
 10  delivery_pts  689 non-null    float64
 11  forw_dist     689 non-null    float64
 12  retn_dist     689 non-null    float64
 13  forw_time     688 non-null    float64
 14  retn_time     688 non-null    float64
 15  kg_load       686 non-null    float64
 16  unload_time   686 non-null    float64
 17  avg_spd       688 non-null    float64
 18  max_spd       688 non-null    float64
 19  off_eng       686 non-null    float64
 20  on_eng        686 non-null    float64
 21  fuelc_100km   688 non-null    float64
dtypes: float64(14), int64(8)
memory usage: 120.4 KB
```

The CSV file is loaded into a DataFrame with rows 0 to 699 and columns 2 to 24. Also, from this DataFrame, the rows with an empty cell at the end of the row are deleted. Finally, the remaining missing values in the DataFrame are imputed using the average value of corresponding columns. The process is necessary so that machine learning and regression analysis work correctly with the dataset because the dataset can prototype false model prediction results and can give the biased result by not having all values. Integrity is maintained in the filling of missing values with means of columns, thereby giving consistency to the aggregated inputs to the algorithms, which makes the model reliable and of good

performance. Then, for the entire DataFrame, we check for the presence of null values using df.isnull().values.any().

The command df.describe().T will produce summary statistics that describe the dataset in terms of central tendency, dispersion, and shape of the data distribution. Afterwards, the .T method is used in order to view the output in columns since it makes the statistics more readable. The summary contains count, mean, standard deviation, minimum and maximum, and three quartile values (25%, 50%, 75%). The descriptive statistics are very primitive in explaining the characteristics of the data and allow one to identify potential misuse or possible outliers.

Figure 2**: Descriptive Statistics of the Data**

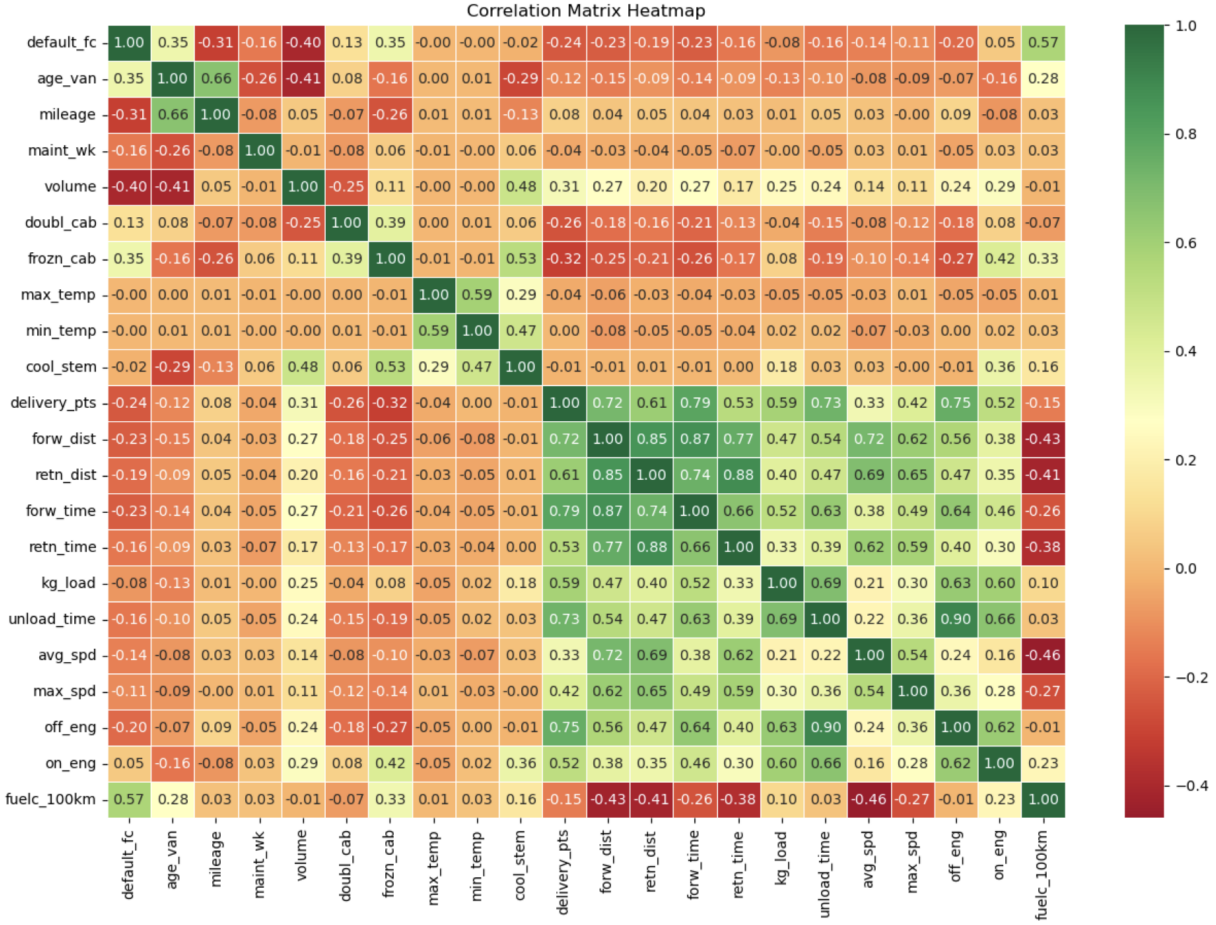
	count	mean	std	min	25%	50%	75%	max
default_fc	688.0	7.752907	0.544623	7.10	7.10	7.800000	8.20	8.7
age_van	688.0	8.274709	2.351586	4.00	7.00	8.000000	10.00	12.0
mileage	688.0	284092.023256	131297.332679	80243.00	155418.75	296330.500000	391751.50	536750.0
maint_wk	688.0	24.446221	13.818327	0.00	15.00	23.000000	37.00	51.0
volume	688.0	14.438953	1.594859	13.00	13.00	15.000000	15.00	17.0
doubl_cab	688.0	0.213663	0.410190	0.00	0.00	0.000000	0.00	1.0
frozn_cab	688.0	0.642442	0.479630	0.00	0.00	1.000000	1.00	1.0
max_temp	688.0	17.885174	4.393314	8.00	15.00	18.000000	20.00	29.0
min_temp	688.0	5.399709	2.855848	-2.00	4.00	6.000000	8.00	11.0
cool_stem	688.0	5.095930	4.062049	0.00	0.00	5.000000	8.00	17.0
delivery_pts	688.0	35.258721	7.431389	14.00	30.00	35.000000	40.00	57.0
forw_dist	688.0	114.627907	20.245608	59.00	102.00	114.000000	128.00	173.0
retn_dist	688.0	30.828488	8.857086	5.00	25.00	31.000000	37.00	56.0
forw_time	688.0	209.621543	25.727765	143.00	191.00	209.000000	226.00	292.0
retn_time	688.0	33.503639	8.459832	10.00	28.00	33.000000	39.00	60.0
kg_load	688.0	706.510949	118.221668	340.00	630.00	710.000000	780.00	1040.0
unload_time	688.0	113.354745	15.579388	62.00	104.00	113.000000	124.00	163.0
avg_spd	688.0	35.206696	3.495193	24.00	33.00	35.603348	37.00	44.0
max_spd	688.0	91.179039	6.159569	69.00	87.00	91.000000	96.00	108.0
off_eng	688.0	173.684672	20.610122	103.00	160.00	173.000000	188.00	237.0
on_eng	688.0	48.710949	6.925762	26.00	44.00	49.000000	53.25	69.0
fuelc_100km	688.0	15.278285	1.020721	12.78	14.51	15.175000	15.94	19.2

4.2. Correlation Analysis

The purpose of correlation analysis is to find out the connection between variables in the dataset. It measures the magnitude of association between two factors. It further aids in the detection of strongly related variables that may bring about an effect on the dependent variable, which in the current study is fuel consumption. This study further evaluates correlation with the help of graphical tools, for example, seaborn, where heat maps or pair plots are created. Seaborn is a powerful Python visualization library that provides a high-level interface to produce informative and beautiful datasets. We can thus plot the correlation matrix and check what set of variables have high correlation coefficients. If close to +1 or -1,

it is an indicator that the variables are in a high positive or high negative relationship. If close to 0, it is an indicator that the relationship is weak or there is none.

Figure 3**: Correlation Matrix Heatmap**



The correlation analysis counts for significant relationships between these various factors that are taken into consideration in influencing fuel consumption in cold chain logistics. The correlation between default fuel consumption and vehicle age (default_fc and age_van) amounts to 0.35, meaning that there is a moderate positive correlation. Moderate positive correlation means that an increase in the age of the vehicle should result in increased fuel usage since new generation vehicles include modern technological additions, leading to relative inefficiency of the older vehicles. The correlation between vehicle age (age_van) and vehicle mileage (mileage) is placed at 0.66, meaning that it is a strong positive correlation. Mileage is positively correlated with age, which is very intuitive since a vehicle that is older will have been placed into service for a greater period and thus would have traveled more distance. That vehicle might then be indicative of potential wear and tear, which implies that older vehicles with higher mileage are likely to experience difficulties with regard to maintaining fuel efficiency. The relationship between vehicle volume (volume) and cooling stem time (cool_stem) is 0.46, implying a moderate positive relationship. Bigger vehicles with higher cargo volume take more time to reach and maintain the desired temperature before they are loaded with cargoes. This, of course, is obvious: when a larger vehicle has more volume of storage, this correlates to more space within the cooling stem. This relationship then points to the relevance of an effective cooling system within a larger vehicle to hold this extended time and energy needed to maintain the right temperature.

The correlation between the presence of a double cabin (doubl_cab) and the use of a frozen cabin (frozn_cab) is 0.39, indicating a moderate positive relationship. This suggests

that vehicles equipped with a frozen cabin, which is necessary for transporting goods at -18°C, often have a double cabin. The double cabin provides better insulation and temperature control, which is crucial for maintaining the extremely low temperatures required for frozen goods. This setup helps in minimizing temperature fluctuations and ensuring the quality and safety of the transported goods. The relationship of using a frozen cabin (frozn_cab) and cooling stem time (cool_stem) is relatively strong at 0.53. It takes much stem time in getting the frozen cabin down to -18 °C before fruits will be loaded into the cabin. More cooling chamber time is needed to get the chamber's temperature down to the lowest set point for the fruits to remain frozen throughout the journey. In this regard, this indicates a significant role and practice of effective pre-cooling processing in vehicles meant to transport frozen products. The correlation between the use of a frozen cabin (frozn_cab) and idle time with the engine on (on_eng) is 0.42, showing a moderate positive relationship. It means that vehicles which are equipped with a frozen cabin are supposed to have the engine on more hours than with a cool cabin, which is necessary for the power supply of the refrigeration unit and the keeping of the air temperature inside the cabin at -18°C. The relationship shows that fuel consumption and operation costs increase in the case of maintaining very low temperatures during transport.

The relationship between max_temp and min_temp is strong and positive, at 0.59. This was expected, as generally, lower morning temperatures lead to lower maximum temperatures. The general weather patterns show that most often, cooler days can be forecast from cool mornings. Knowledge of this relationship will lead to planned adjustments of cooling requirements according to expected daily temperature profiles. The relationship of min_temp with cool_stem is measured at 0.47, which shows a moderate relationship. Most drivers have tended to set the cooling process according to ambient temperature prevailing in the morning. In the event of a high morning temperature, it will take a longer cooling stem time to reach the desired temperature and come to rest. This shows the importance of ambient conditions in planning the pre-cooling phase to ensure that temperatures are maintained throughout the trip.

The correlation analysis of delivery points (delivery_pts), forward distance (forw_dist), and forward time (forw_time) stands at a high positive correlation of 0.72 and 0.79 respectively. This means that an increased number of delivery points is directly proportional to the increase in the forward distance covered as well as the time taken to deliver the goods to those points. As more delivery points are serviced, the route becomes larger, and hence, the vehicle has to cover larger distances, which in turn increases the time taken to deliver the goods. The above relationship shows the underlying logistics of dealing with multiple delivery points efficiently, as it has a direct impact on the overall time and distance covered in delivering to different points. The correlation between delivery points (delivery_pts) and return distance (retn_dist) and return time (retn_time) is slightly less, but still high at 0.61 and 0.53 respectively. Though the correlation is not as high as in the previous case of forward distance and time, the correlation still shows that an increased number of delivery points generally increases the return journey distance and time as well. As deliveries are made and the vehicle has to turn back, the more the number of delivery points, the more the distance and time required to return. This shows the underlying cumulative effect of having multiple stops in the overall route planning, including the journey back to the starting point.

There is a positive 0.59 correlation between delivery points (delivery_pts) and kg_load and a positive 0.73 correlation between delivery points and unload time (unload_time). This means that more delivery points generally lead to a higher total load being delivered, while more time is consumed in unloading products at each stop. This relationship is intuitive because as the delivery points increase, the amount of cargo and the handling and unloading

time at each point also increase. The correlation of off_eng and on_eng with unload time (unload_time) equals 0.90 and 0.66, respectively. These are robust correlations. This is because the vehicle will be forced to stop and wait for the unloading to take place, which means the engine will either be running or will be switched off. The very high correlation with off-time indicates that engines are often shut off—possibly for fuel savings—when unloading. The moderate correlation with on-time suggests that, for significant portions of the unloading period, the engine remains on, likely to maintain cabin temperatures or for other operational reasons. This relationship underscores the impact of unloading activities on fuel consumption and vehicle idling practices.

The correlation of distance with speed is high for forw_dist and retn_dist in both avg_speed and max_speed, meaning that the farther the distance covered, the higher the average and maximum speed could likely be. This can best be explained, meaning longer routes can have more highways factored into them with higher velocities. At the same time, shorter distances to more delivery points will take more time on local streets, finding a place to park, and making more frequent stops, which makes the average speed lower. Such a relation develops different driving conditions and speed profiles depending on route length and delivery density.

The correlation analysis of the correlation of fuel consumption per 100 km with default fuel consumption leads to a high level of correlation value of 0.57. This means that the default fuel consumption, as expressed by the maker at the starting point, is highly related to the actual fuel consumption that occurs over a distance. In other words, the more default fuel consumption a given vehicle has, the more fuel consumption per 100 km will be, thereby ultimately and directly relating the level of importance that the overall level of fuel efficiency built into a vehicle is defined. The correlation between fuelc_100km and age_van is low and positive, r = 0.28. In more straightforward words, it means that a 100 km distance will likely burn more fuel with the car's age increasing. This may be because of wearing out of the engine parts and general inefficiency with the aging of the cars and degradation of the vehicle system. However, it does not go solid on correlation, and hence, there exist other factors that have a vast determination of the fuel consumption. On the other hand, fuelc_100km computes near zero correlations with vehicle mileage (mileage), vehicle maintenance (maint_wk), vehicle volume (volume), and whether the vehicle has a double cabin (doubl_cab). This means that about the other variables, these other variables do not directly affect the quantity of fuel consumed in 100 kilometers. This implies that the effect is so tiny that there is no visible effect at all, for if not, there have been apparent indicators that show that their impact is overwhelmed or overridden by more dominating factors such as fuel consumption default and vehicle age.

The correlation between Fuelc_100km and frozn_cab is positive and set at 0.33, hence a low to moderate positive correlation. This implies that a car at an ultra-low temperature of -18°C will expend much more energy to cool down things in the cabin than it will at its average temperature of +4°C and, therefore, also spend more fuel. Hence, it calls for ultra-low temperatures to be set to maintain the temperature in the cabin. It justifies that the vehicle will consume more fuel in that situation. The correlations of the fuelc_100km values at the maximum temperature of the stalk are pretty wrong. This suggests that the ambient temperature conditions and initial cooling time do not seem to influence fuel consumption per 100 km. These might be optimally taken care of in the operational structure so as not to influence fuel consumption.

The delivery points (delivery_pts) have a weakly negative correlation of -0.15 with fuel consumption per 100 kilometers (fuelc_100km). This means that as the number of delivery points increases, the fuel used per delivery slightly decreases. This reduction may be due to

economies of scale, where the fuel cost per delivery point drops as the number of stops increases, possibly because of optimized routing and load distribution. Fuel consumption per 100 kilometers also has a low to moderate negative correlation (ranging from -0.2 to -0.45) with forward distance (forw_dist), forward time (forw_time), return distance (retn_dist), and return time (retn_time). These negative correlations indicate that longer travel distances and times are associated with lower fuel consumption per 100 kilometers. This might be because longer routes usually involve more highway driving, which generally results in better fuel economy compared to shorter, stop-and-go trips in urban areas.

The relationship between fuel consumption per 100 kilometers (fuelc_100km) and both load weight (kg_load) and unload time (unload_time) is slightly positive. This means that heavier loads and longer unloading times do increase fuel consumption, but their effect is not as significant compared to other factors. The extra fuel used for carrying heavier loads and the time spent idling during unloading do matter, but they have a smaller impact.

Fuel consumption per 100 kilometers (fuelc_100km) has a low to moderate negative correlation with average speed (avg_spd) and maximum speed (max_speed), with values of -0.46 and -0.27 respectively. This suggests that higher speeds generally result in better fuel efficiency per 100 kilometers. Vehicles use less fuel per distance traveled when they go faster, especially on highways, because they maintain steady speeds and operate more efficiently.

Lastly, fuel consumption per 100 kilometers (fuelc_100km) has a near-zero correlation with idle time when the engine is off (off_eng) at -0.01 and a positive correlation with idle time when the engine is on (on_eng) at 0.23. When the engine is off, it doesn't use any fuel, which explains the near-zero correlation. However, when the engine is on, idling consumes more fuel, especially when running the air conditioning to keep the cabin cool.

4.3. Variance Inflation Factor (VIF)

In addition to correlation analysis, another measure called the Variance Inflation Factor (VIF) is used to check how much the variance of a regression coefficient is increased because of correlations with other predictors. When the VIF values are high, it means the predictors are highly correlated, which can make the regression results unstable and unreliable. If a VIF value is greater than 10, it indicates a significant problem with multicollinearity, meaning the predictors are very similar to each other. To fix this, we can identify the variables with high VIF values and consider removing or combining them to improve the model's reliability. In our study, factors like default fuel consumption (default_fc), age of the van (age_van), volume, delivery points (delivery_pts), forward distance (forw_dist), forward time (forw_time), unload time, average speed (avg_spd), maximum speed (max_spd), off engine time (off_eng), and on engine time (on_eng) have VIF values much higher than 10. This shows severe multicollinearity, making it difficult to determine the individual impact of each predictor on the dependent variable. This high multicollinearity increases the standard errors and makes the statistical conclusions less reliable. For example, the variable default_fc has the highest VIF value of 501.26, meaning it is almost perfectly correlated with other predictors in the model, with an R^2 value of 0.998005. Similarly, forw_time, forw_dist, and off_eng also have very high VIF values, indicating they are highly correlated with other variables. These high VIF values can distort the true relationships between predictors and the outcome, making it hard to understand which factors are really important.

Figure 4**: Variance Inflation Factor (VIF)**

	Factors	VIF_Value	R2		Factors	VIF_Value	R2
0	default_fc	501.256973	0.998005	10	delivery_pts	102.336277	0.990228
1	age_van	91.156933	0.989030	11	forw_dist	444.233982	0.997749
2	mileage	30.051703	0.966724	12	retn_dist	97.364963	0.989729
3	maint_wk	4.938279	0.797500	13	forw_time	527.055419	0.998103
4	volume	171.645790	0.994174	14	retn_time	78.214472	0.987215
5	doubl_cab	1.713973	0.416560	15	kg_load	85.248492	0.988270
6	frozn_cab	14.418581	0.930645	16	unload_time	385.108330	0.997403
7	max_temp	27.670763	0.963861	17	avg_spd	376.767837	0.997346
8	min_temp	10.194136	0.901904	18	max_spd	355.126397	0.997184
9	cool_stem	8.016844	0.875263	19	off_eng	481.909040	0.997925
				20	on_eng	273.106503	0.996338

To deal with multicollinearity problems, you may want to remove some high correlation predictors in the model with very high VIF over the threshold. Or, better yet, you one could reduce those correlated predictors into a handful of uncorrelated components through methods like Principal Component Analysis (PCA). Regularization methods such as Ridge Regression or Lasso Regression could also be employed to mitigate the effects of multicollinearity by adding a penalty to the size of the coefficients, thereby stabilizing the estimates. Similarly, these steps can enhance the reliability and interpretability of the model which in turn will provide precise and actionable intelligence to determine the underlying factors in consumption of fuel in case of cold chain logistics.

It's crucial to understand these relationships before applying machine learning models because multicollinearity, where independent variables are highly correlated, can distort the model's interpretation and predictions. This affects the reproducibility of the model and its interpretation because multicollinearity inflates the variance of our coefficient estimates and can lead to a highly sensitive model to a perfectly valid dataset.

4.4. Regression Analysis of Key Factors

Regression analysis is used to understand how the dependent variable (fuel consumption) relates to various independent variables (key factors). The process involves several steps to make sure the models are reliable and accurate.

Train-test split:

We start by dividing the dataset into training and testing sets before creating and testing the regression models. This is important to evaluate how well the model generalizes and performs on unseen data. A regular split ratio of 80% for training and 20% for testing is commonly used. The training set is used to fit the model, while the testing set assesses the model's performance.

Simple Linear Regression with OLS and Analysis:

Ordinary Least Squares (OLS) refers to a family of methods used for estimating the coefficients of the linear regression model. In simple terms, it fits a linear model of the response variable with a single independent variable. The resulting OLS estimator minimizes the sum of the squared errors, which is the difference between the observed values of the

dependent variable and the predicted values from the regression model. Performing an OLS provides us information about the strength and direction of the association between the variables as well as the importance of the predictor variable.

Figure 5**: OLS Regression Results**

OLS Regression Results

Dep. Variable:	fuelc_100km	**R-squared:**	0.757
Model:	OLS	**Adj. R-squared:**	0.748
Method:	Least Squares	**F-statistic:**	78.54
Date:	Thu, 23 May 2024	**Prob (F-statistic):**	2.88e-147
Time:	22:12:29	**Log-Likelihood:**	-398.27
No. Observations:	550	**AIC:**	840.5
Df Residuals:	528	**BIC:**	935.4
Df Model:	21		
Covariance Type:	nonrobust		

Linear Regression with Scikit-learn:

A widely used machine learning library in Python, scikit-learn, contains implementations of many algorithms useful for linear regression, that can be easily used to fit the model. While this approach may not be optimal for our current problem, scikit-learn's linear model is capable of handling multiple predictors (features) in the linear regression model easily. Therefore it is straightforward to extend the simple one-parameter linear regression to examine the influence of multiple features on the fuel consumption. This not only simplifies the model fitting but also provides tools such as regularization methods (e.g. ridge and lasso) and cross-validation schemes that are helpful in enhancing the performance and robustness of the model. We will discuss these topics in separate articles in future.

Significance of Independent Variables

The results of the regression analysis provided evidence that many of these adopted independent variables are statistically significant predictors of fuel consumption in cold chain logistics. These variables include defaul fuel consumption, vehicle mileage, last maintenance weeks, volume, existence of a double cabin, forward and return travel distance, forward travel time, load weight , unloading time, engine on idle time. Each of these variables' p-values are less than 0.05, indicating that all of them are very strongly related to fuel consumption. For instance, the p-value for default fuel consumption is 0, indicating that it is a very critical factor for fuel usage. Similarly, mileage of the vehicle, maintenance frequency, volume of the vehicle, presence of a double cabin, distances travelled -- both forward and return, the time taken during forward travel, the weight of the load, time taken for unloading, and the time for which the engine is on are all significant and, therefore, show their importance in the fuel consumption model.

On the other hand, most the variables were statistically insignificant predictors of the dependent variable, for the simple reason that they do not really relate strongly and/or consistently to fuel consumption within the particular dataset and model being used. Variables such as the age of the van, presence of a frozen cabin, ambient temperatures, cooling stem time, number of delivery points, return travel time, average and maximum speed, and engine off idle time might not significantly impact fuel consumption due to several

reasons. These variables could have too little variability within the data, their effects on fuel consumption may be indirect or perhaps mediated through other variables, or their impact may show only in certain contexts and therefore apply variably to different operating conditions or vehicle types. In addition, multicollinearity ,many of the independent variables being strongly correlated among each other, may further deflate the apparent significance of individual predictors. Thus, although these variables may intuitively seem relevant, the fact that they are not statistically significant suggests that they do not contribute to fuel consumption independently in this particular model. All of these variables have a p-value greater than 0.05, indicating that their effect on fuel consumption is statistically not significant in this model. For example, the p-value for the age of the van is 0.422, indicating that vans in all age categories do not significantly affect fuel consumption. The significance of independent variables is presented in Figure 6.

Figure 6**: Significance of Independent Variables**

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.7411	0.912	1.908	0.057	-0.051	3.533
default_fc	1.3163	0.078	16.829	0.000	1.163	1.470
age_van	-0.0204	0.025	-0.804	0.422	-0.070	0.029
mileage	2.391e-06	4.13e-07	5.789	0.000	1.58e-06	3.2e-06
maint_wk	0.0099	0.002	5.751	0.000	0.006	0.013
volume	0.1354	0.022	6.201	0.000	0.093	0.178
doubl_cab	-0.4100	0.064	-6.449	0.000	-0.535	-0.285
frozn_cab	-0.0553	0.103	-0.535	0.593	-0.259	0.148
max_temp	0.0039	0.006	0.628	0.531	-0.008	0.016
min_temp	-0.0111	0.012	-0.943	0.346	-0.034	0.012
cool_stem	0.0047	0.010	0.479	0.632	-0.014	0.024
delivery_pts	-0.0041	0.006	-0.648	0.517	-0.016	0.008
forw_dist	-0.0285	0.004	-6.467	0.000	-0.037	-0.020
retn_dist	-0.0237	0.007	-3.519	0.000	-0.037	-0.010
forw_time	0.0093	0.003	3.418	0.001	0.004	0.015
retn_time	0.0062	0.005	1.130	0.259	-0.005	0.017
kg_load	0.0012	0.000	4.218	0.000	0.001	0.002
unload_time	0.0089	0.004	2.388	0.017	0.002	0.016
avg_spd	-0.0195	0.014	-1.387	0.166	-0.047	0.008
max_spd	0.0025	0.005	0.510	0.610	-0.007	0.012
off_eng	-0.0007	0.003	-0.238	0.812	-0.006	0.005
on_eng	0.0295	0.007	4.011	0.000	0.015	0.044

To sum up, the regression analysis reveals that default fuel use consumption according to miles, maintenance frequency of vehicle, volume of vehicle, double cabin, forward and return distances, forward travel time, load weight, and time spent offloading and engine on-time are factors which predict fuel consumption in cold chain logistics significantly. The above results emphasize frequent service of vehicles and monitoring the usage of vehicles and optimize operational factors which improve fuel efficiency in transportation. The non-significance of variables such as age, van, multiple temperature measures, and speed measures themselves indicates that other operational variables are of more pivotal operation in determining fuel

consumption. These insights can help drive strategic decisions in fleet management toward better fuel efficiency and lower operational costs in cold chain logistics.

Considering that in machine learning, all independent variables are used, even if some are not significant in the regression analysis. Machine learning models can pick up on complicated relationships and interactions that traditional regression modeling might overlook. Variables that seem insignificant in a linear regression will still add to information when considered with other variables to enable better predictive accuracy. Further, regularization machine learning techniques, such as Lasso and Ridge regression, may automatically handle the irrelevant or less important variables with the redundant dimensionality by reducing their influence on the model to prevent overfitting. Cross-validation techniques ensure the model generalizes well to unseen data, hence confirming the overall robustness of the approach. By including all variables, it allows the model to fully see the data and pick up any subtle patterns and interactions that may be essential to predict fuel consumption.

Comparison of Linear Regression Models without Model Tuning:

After fitting the models using both OLS and Scikit-learn, it is important to compare their performance without any hyperparameter tuning. This first comparison can be useful in establishing the performance baseline of the models, and to discover if any have notably superior predictive capability. To evaluate the models, we use metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). Using these metrics, we can assess which model gives better predictions and how much variability the model can catch. These above actions of regression analysis are required at the time of constructing good predictive framework. Train-test split to validate the model performance on unseen data, OLS to get a basic understanding of the variable relationships, and Scikit-learn to make the model building and evaluation easy. Fine-tuning models will also allow easy comparisons with different models, showing how each type of model performs in its own right, which aids in making decisions for further model development.

Figure 7**: Model Comparison without Model Tuning**

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	LassoLarsIC	0.293078	0.541367	0.589917	0.730546	0.598533	2.200640
1	Ridge	0.293289	0.541562	0.595469	0.730352	0.591286	1.933583
2	LinearRegression	0.293642	0.541887	0.596244	0.730028	0.587223	1.741070
3	ARDRegression	0.317228	0.563230	0.597102	0.708343	0.597118	6.595681
4	BayesianRidge	0.292970	0.541267	0.732725	0.730645	0.427529	2.318382
5	ElasticNet	0.692866	0.832386	0.791546	0.362984	0.332554	15.034707
6	LassoLars	0.735005	0.857324	0.823007	0.324242	0.289038	15.616580
7	Lasso	0.735005	0.857324	0.823007	0.324242	0.289037	15.616580
8	Lars	0.299025	0.546832	8.365999	0.725078	-153.487619	1.762277

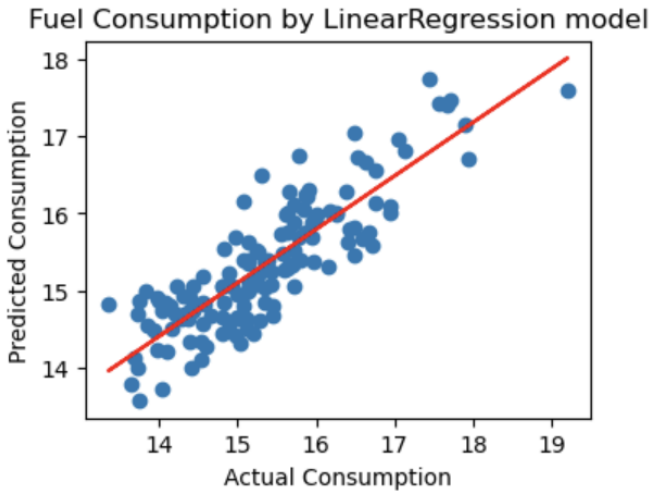
4.5. Model Tuning

Model tuning is basically optimizing the performance of machine learning models. It involves adjusting hyperparameters to improve the model's predictive accuracy and generalization capability. Each model has specific tuning procedures that enhance its ability

to handle the given data and improve success metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Figure 8**: Simple Linear Regression without Tuning**

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	LinearRegression	0.293642	0.541887	0.596244	0.730028	0.587223	1.74107



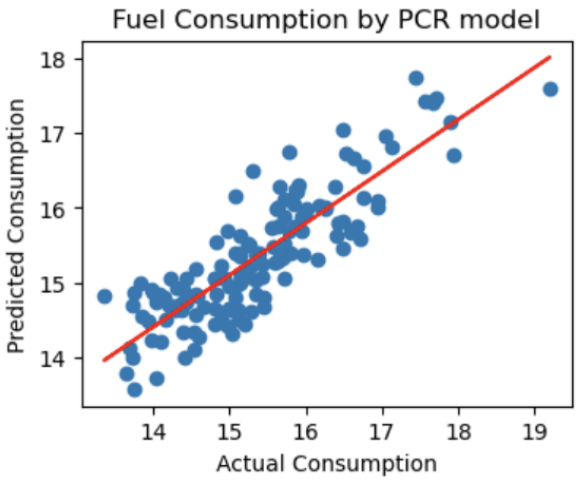
PCR Model Tuning

Principal Component Regression (PCR) tuning involves selecting the optimal number of principal components. The goal is to balance between dimensionality reduction and retaining sufficient variance in the data. By tuning the number of components, PCR reduces overfitting and improves model accuracy by retaining the most informative components.

Figure 9**: Principal Component Regression (PCR) Model Tuning**

```
Optimum number of components: 21
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	PCR	0.293642	0.541887	0.596244	0.730028	0.587223	15.264036



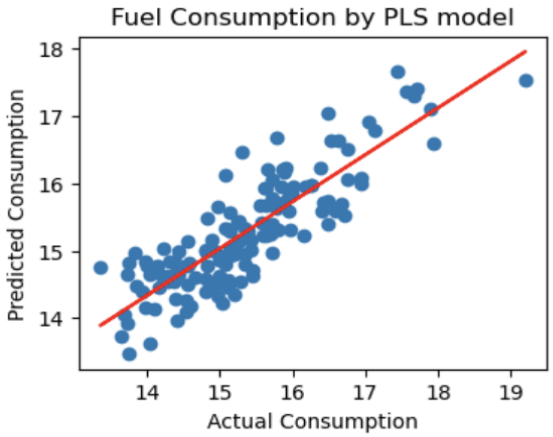
PLS Model Tuning

Partial Least Squares (PLS) regression tuning also involves selecting the number of components. PLS aims to find the components that maximize the covariance between the predictors and the response variable. Tuning helps in identifying the optimal number of components that explain the variance in the response variable effectively, leading to improved prediction performance.

Figure 10**: Partial Least Squares (PLS) Model Tuning**

```
Optimum number of components: 14
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	PLS	0.302715	0.550195	0.595501	0.721686	0.587846	15.264036



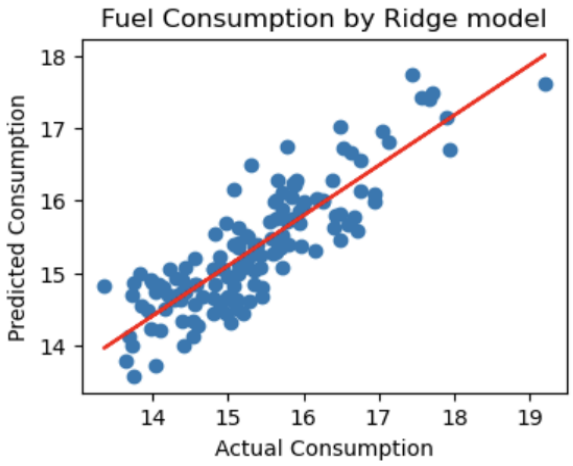
Ridge Model Tuning

Ridge regression tuning involves adjusting the regularization parameter (lambda). This parameter controls the degree of shrinkage applied to the regression coefficients. By tuning lambda, Ridge regression balances the trade-off between bias and variance, helping to mitigate multicollinearity and improve the stability and accuracy of the model.

Figure 11**: Ridge Model Tuning**

```
Optimum alpha:0.5748784976988678
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	Ridge	0.293421	0.541684	0.595396	0.730231	0.590219	1.852971



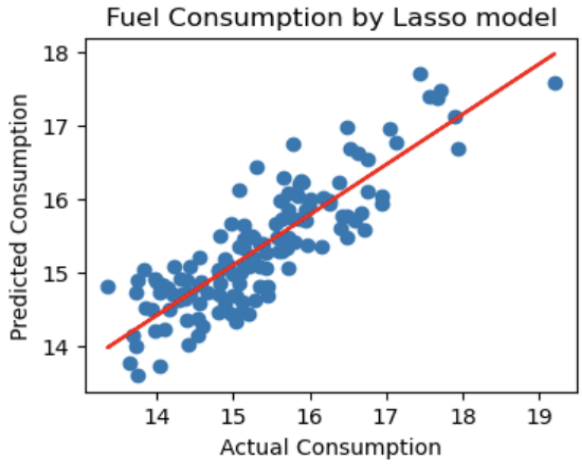
Lasso Model Tuning

Lasso regression tuning focuses on selecting the regularization parameter (lambda). This parameter determines the extent to which coefficients are shrunk towards zero. By tuning lambda, Lasso performs feature selection, removing irrelevant features and reducing model complexity, which enhances prediction accuracy and interpretability.

Figure 12**: Lasso Model Tuning**

```
Optimum alpha:0.006170889988321264
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	Lasso	0.293229	0.541506	0.59213	0.730407	0.595514	2.24309



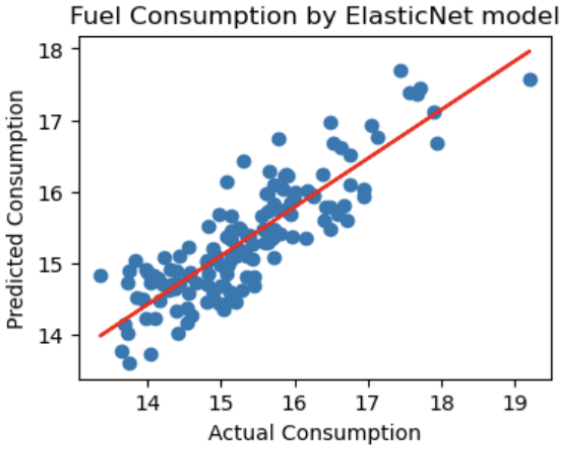
ElasticNet Model Tuning

ElasticNet tuning involves adjusting two regularization parameters: lambda1 (L1 penalty) and lambda2 (L2 penalty). Tuning these parameters allows ElasticNet to balance between Lasso and Ridge penalties, achieving both feature selection and coefficient shrinkage. This helps in handling multicollinearity and improving model robustness and accuracy.

Figure 13**: Elasticnet Model Tuning**

```
Optimum alpha:0.008706881789759882
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	ElasticNet	0.293863	0.542091	0.593791	0.729824	0.593808	2.448414



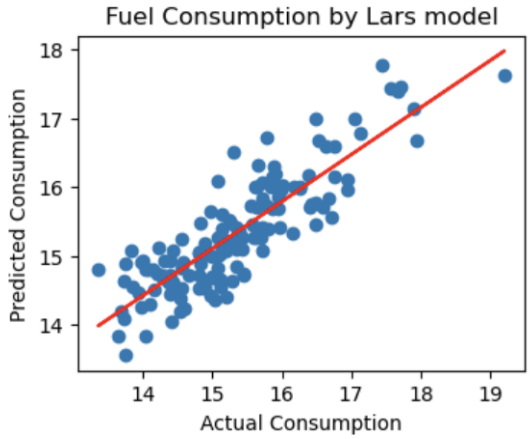
Lars Model Tuning

Least Angle Regression (LARS) tuning involves selecting the number of steps to include in the model. This determines how many predictors are included in the final model. Tuning helps in controlling the complexity of the model and avoiding overfitting, thereby enhancing predictive performance.

Figure 14**: Lars Model Tuning**

```
Optimum n_nonzero_coefs: 15
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	Lars	0.291978	0.540350	0.635221	0.731558	0.566466	15.264036



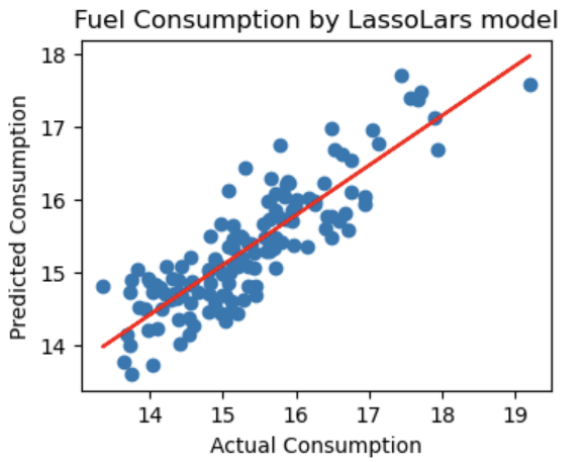
LassoLars Model Tuning

LassoLars tuning also involves adjusting the number of steps or the alpha parameter. This process ensures that the model includes only the most relevant predictors by applying L1 regularization in a stepwise manner. Tuning helps in achieving an optimal balance between model simplicity and accuracy.

Figure 15**: LassoLars Model Tuning**

```
Optimum alpha:0.0063477017739096
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	LassoLars	0.293266	0.54154	0.592012	0.730374	0.595752	2.251777

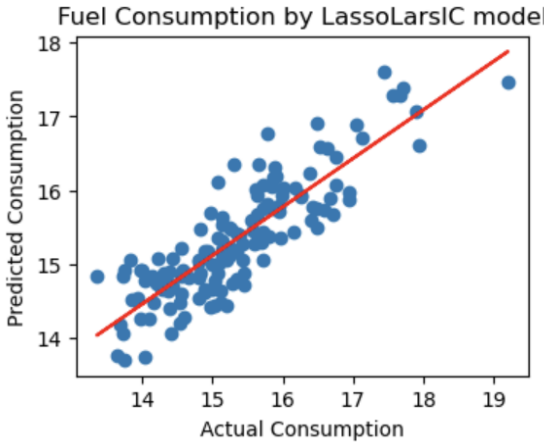


LassoLarsIC Model Tuning

LassoLarsIC tuning focuses on selecting the best model based on Information Criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). This tuning process helps in choosing a model that balances goodness of fit and model complexity, leading to improved generalization and predictive performance.

Figure 16**: LassoLarsIC Model Tuning**

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	LassoLarsIC	0.298513	0.546363	0.61131	0.72555	0.583021	2.963055



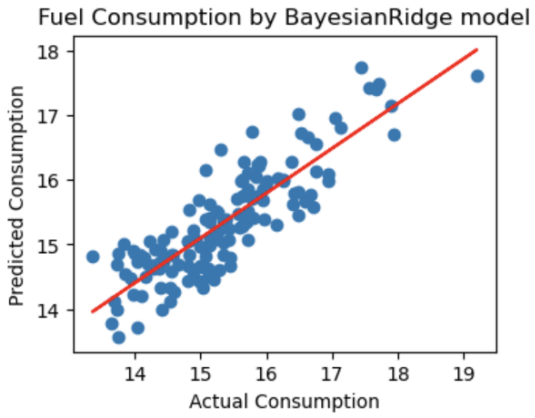
BayesianRidge Model Tuning

Bayesian Ridge Regression tuning involves adjusting hyperparameters related to the priors on the coefficients and the noise term. This process allows for the incorporation of prior knowledge and helps in regularizing the model. Tuning these parameters improves the model's ability to generalize to new data and provides more reliable uncertainty estimates for the predictions.

Figure 17**: BayesianRidge Model Tuning**

```
Optimum:{lambda_2: 3.851, lambda_1: 8.471, alpha_2: 0.083, alpha_1: 0.009}
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept	defa
0	BayesianRidge	0.293281	0.541554	0.595697	0.730359	0.591255	1.938886	



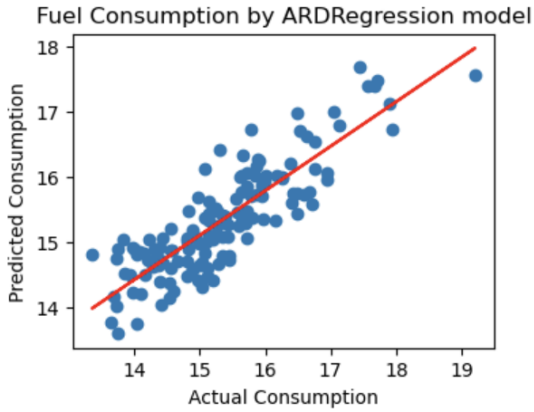
ARDRegression Model Tuning

Automatic Relevance Determination (ARD) Regression tuning involves adjusting hyperparameters related to the priors on the coefficients. This process helps in determining the relevance of each predictor, effectively performing feature selection and regularization. Tuning ARD regression improves model interpretability and predictive accuracy by focusing on the most relevant features.

Figure 18**: ARDRegression Model Tuning**

```
Optimum:{lambda_2: 0.000, lambda_1: 1.546, alpha_2: 0.000, alpha_1: 0.830}
```

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept	defi
0	ARDRegression	0.295815	0.543889	0.588536	0.72803	0.608237	1.954899	



Each of these tuning processes is applied to enhance the model's performance by:

- Reducing overfitting and improving generalization to unseen data.
- Selecting the most relevant features and reducing model complexity.
- Balancing the trade-off between bias and variance.
- Improving the accuracy, stability, and interpretability of the models.

By carefully tuning these models, the research ensures that the predictive models are robust, reliable, and capable of providing accurate estimates of fuel consumption in cold chain logistics. This ultimately leads to better decision-making and optimization of logistics operations.

5. Discussion of Results

5.1. Interpretation of Results

Evaluating the ability to predict fuel consumption using various machine learning models is an important process in cold chain logistics. We can compare various models and judge which one gives the best performance among particular success metrics. This consists of aggregating the outputs of different tuned models, ranking them on some performance metrics, and studying the results. Machine learning model comparison is used to determine how a number of regression models perform in the context of a particular set of data and a given metric. This score is measured in the primary metric against which we will compare the models, the Cross-Validation Root Mean Squared Error (Cross Val. RMSE), indicating its predictive performance and generalization on unseen data.

Figure 19**: *Comparison of Machine Learning Models***

	Model	MSE	RMSE	Cross Val.RMSE	R2	Cross Val.R2	Intercept
0	ARDRegression	0.295815	0.543889	0.588536	0.728030	0.608237	1.954899
1	LassoLars	0.293266	0.541540	0.592012	0.730374	0.595752	2.251777
2	Lasso	0.293229	0.541506	0.592130	0.730407	0.595514	2.243090
3	ElasticNet	0.293863	0.542091	0.593791	0.729824	0.593808	2.448414
4	Ridge	0.293421	0.541684	0.595396	0.730231	0.590219	1.852971
5	PLS	0.294156	0.542362	0.595501	0.729555	0.587846	15.264036
6	BayesianRidge	0.293281	0.541554	0.595697	0.730359	0.591255	1.938886
7	LinearRegression	0.293642	0.541887	0.596244	0.730028	0.587223	1.741070
8	PCR	0.293642	0.541887	0.596244	0.730028	0.587223	15.264036
9	LassoLarsIC	0.298513	0.546363	0.611310	0.725550	0.583021	2.963055
10	Lars	0.291978	0.540350	0.635221	0.731558	0.566466	15.264036

After evaluating and comparing the models based on Cross-Validation RMSE, the top three identified models are ARDRegression, LassoLarsIC, and LassoLars. ARDRegression achieved the lowest Cross-Validation RMSE, indicating it provides the most accurate fuel consumption predictions. This model uses a Bayesian framework that automatically determines the relevance of predictors, effectively using the most important variables while ignoring irrelevant ones. This leads to more precise estimates and better generalization to new data by applying separate priors to each coefficient, a form of regularization that prevents overfitting.

LassoLarsIC, which combines the LARS algorithm with Lasso regression and information criteria for model selection, also performed well, though it had a slightly higher RMSE than ARDRegression. It simplifies the model and improves interpretability by setting some coefficients to zero, balancing goodness of fit with model complexity through the use of information criteria like AIC or BIC. This enhances its predictive performance.

LassoLars, a variant of the LARS algorithm incorporating Lasso regularization, came third in the comparison. It is computationally efficient and capable of handling large datasets with many predictors, benefiting from Lasso's feature selection ability, which contributed to its high ranking.

The comparison highlights that models incorporating Bayesian techniques and information criteria, such as ARDRegression and LassoLarsIC, tend to perform better in predicting fuel consumption using research data on cold chain logistics. These models effectively balance model complexity and prediction accuracy, making them particularly well-suited for this type of analysis. ARDRegression's ability to automatically determine the relevance of each predictor, combined with its robust regularization approach, makes it the best-performing model in this study.

Figure 20: ***Model Comparison Chart by Cross Validated RMSE***

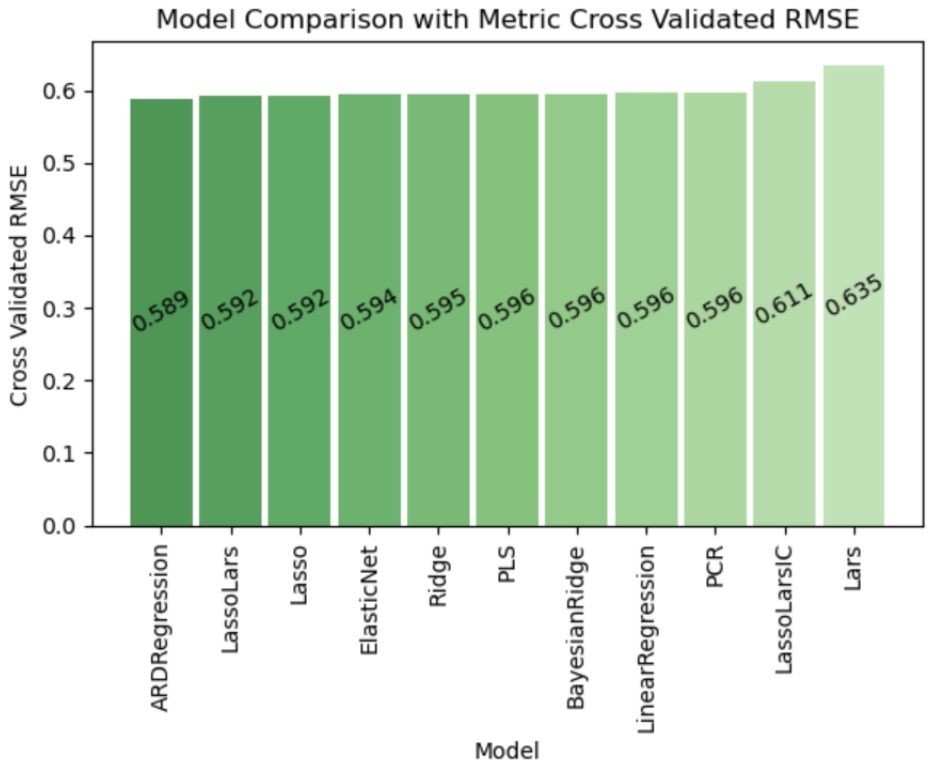
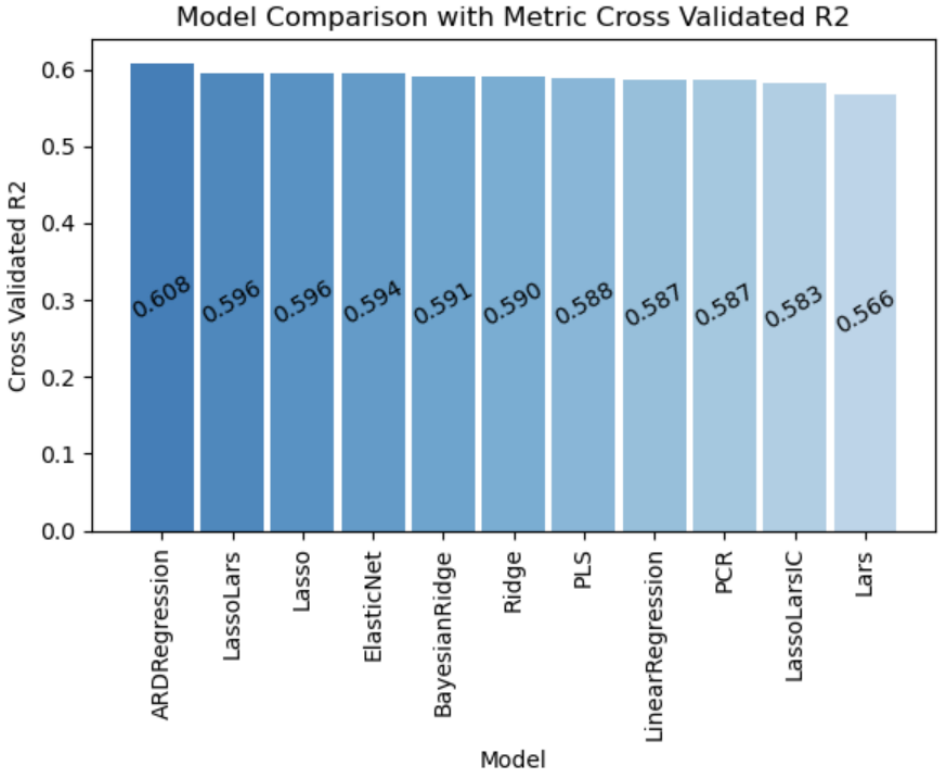


Figure 21: ***Model Comparison Chart by Cross Validated R2***



5.2. Key Factors in Fuel Consumption Estimation Model

Figure 22*: Coefficients Comparison of ML Models (Vehicle Related)*

	Model	Intercept	default_fc.coef	age_van.coef	mileage.coef	maint_wk.coef	volume.coef
0	ARDRegression	1.954899	1.268167	-0.005001	0.000002	0.009910	0.134865
1	LassoLars	2.251777	1.232818	0	0.000002	0.010008	0.136102
2	Lasso	2.243090	1.233701	0	0.000002	0.010013	0.136223
3	ElasticNet	2.448414	1.214627	0	0.000002	0.009856	0.133458
4	Ridge	1.852971	1.299464	-0.016329	0.000002	0.009878	0.135462
5	PLS	15.264036	0.719826	-0.054784	0.318238	0.137354	0.215727
6	BayesianRidge	1.938886	1.286567	-0.013236	0.000002	0.009888	0.135493
7	LinearRegression	1.741070	1.316280	-0.020362	0.000002	0.009864	0.135424
8	PCR	15.264036	2.5763e-07	-0.000908	-0.014693	0.016658	0.002490
9	LassoLarsIC	2.963055	1.162124	0	0.000002	0.009562	0.126397
10	Lars	15.264036	0.674448	0	0.262158	0.132864	0.218151

Key information is provided by the coefficients for vehicle-related factors in the fuel consumption estimation models that explain how these attributes affect fuel consumption. The default fuel consumption rate and the vehicle age are both consistently significant predictors across all models we analyzed. The default fuel consumption rate (default_fc.coef) is positive, meaning that greater baseline consumption increases the overall fuel consumption expectedly. Interestingly, the vehicle age (age_van.coef) generally has a negative coefficient in some models like ARDRegression and Ridge, suggesting that newer vehicles, which are typically more efficient, help reduce fuel consumption. However, in models like Lasso and ElasticNet, vehicle age coefficients are zero, indicating that these models either do not account for this factor or deem it less significant when penalizing less influential predictors. Vehicle maintenance (maint_wk.coef) consistently shows a positive impact, highlighting the importance of regular maintenance in mitigating increased fuel consumption due to wear and tear. The volume of the vehicle (volume.coef) also contributes positively across all models, suggesting that larger vehicles, which can carry more load, tend to consume more fuel, aligning with expectations. The double cabin (doubl_cab.coef) feature increases aerodynamic drag and weight, thus negatively impacting fuel consumption, except in the case of Ridge regression where it shows the highest negative impact.

Figure 23*: Coefficients Comparison of ML Models (Temperature Related)*

	Model	doubl_cab.coef	frozn_cab.coef	max_temp.coef	min_temp.coef	cool_stem.coef
0	ARDRegression	-0.385519	0.000078	0.001909	-0.005554	0.003505
1	LassoLars	-0.382827	0	0.002887	-0.008151	0.003363
2	Lasso	-0.383949	0	0.002916	-0.008195	0.003366
3	ElasticNet	-0.389034	0	0.003191	-0.009093	0.003675
4	Ridge	-0.410067	-0.049608	0.003892	-0.011065	0.004677
5	PLS	-0.168643	-0.025669	0.017705	-0.030689	0.018182
6	BayesianRidge	-0.410044	-0.045384	0.003881	-0.011053	0.004695
7	LinearRegression	-0.409975	-0.055336	0.003906	-0.011087	0.004661
8	PCR	0.027753	-0.037298	-0.019066	0.004482	-0.022586
9	LassoLarsIC	-0.293358	0	0.000608	-0.004667	0.003115
10	Lars	-0.168215	0	0.005539	-0.009719	0.006336

Temperature-related factors are important in fuel consumption, especially in cold chain logistics where temperatures must be controlled. The ambient maximum temperature (max_temp) coefficients vary slightly across models but are generally small, indicating a moderate direct effect on fuel consumption. The frozen cabin parameter (frozn_cab), showing whether goods are kept at freezing temperatures, has a positive coefficient in models like ARDRegression, reflecting higher energy needs for freezing conditions. However, models like Lasso and ElasticNet show negative or zero coefficients for this parameter, suggesting they minimize its importance with regularization. The morning minimum temperature (min_temp) coefficient is close to zero across models, indicating minimal direct impact on fuel consumption due to advancements in insulation and cooling technologies.

Figure 24*: Coefficients Comparison of ML Models (Route Related)*

	Model	delivery_pts.coef	forw_dist.coef	retn_dist.coef	forw_time.coef	retn_time.coef
0	ARDRegression	-0.001893	-0.029655	-0.020144	0.009667	0.002997
1	LassoLars	-0.002411	-0.028609	-0.022759	0.009349	0.005096
2	Lasso	-0.002441	-0.028593	-0.022790	0.009345	0.005122
3	ElasticNet	-0.002686	-0.028500	-0.023027	0.009271	0.005333
4	Ridge	-0.004003	-0.028418	-0.023704	0.009244	0.006140
5	PLS	-0.027377	-0.570451	-0.202320	0.228955	0.051173
6	BayesianRidge	-0.003965	-0.028380	-0.023730	0.009231	0.006121
7	LinearRegression	-0.004054	-0.028469	-0.023670	0.009261	0.006165
8	PCR	-0.020026	0.000333	0.027756	-0.023422	0.033292
9	LassoLarsIC	0	-0.029639	-0.020439	0.009509	0.003069
10	Lars	0	-0.468190	-0.145807	0.141516	0

Because route-related characteristics are directly correlated with trip time and distance, they play a critical role in determining fuel usage. All models have negative coefficients for

the forward and return route distances (forw_dist.coef and retn_dist.coef), indicating that longer routes tend to use less gasoline per kilometer, presumably as a result of faster speeds and fewer stops. Longer travel times, perhaps as a result of idling and fluctuating traffic circumstances, increase fuel consumption, but the amount of time spent on these routes (forw_time.coef and retn_time.coef) has a favorable impact on fuel consumption. This emphasizes how crucial it is to optimize routes for both distance and travel time in order to successfully reduce fuel usage.

Figure 25*: Coefficients Comparison of ML Models (Load & Driving Related)*

	Model	kg_load.coef	unload_time.coef	avg_spd.coef	max_spd.coef	off_eng.coef	on_eng.coef
0	ARDRegression	0.001171	0.008775	-0.013138	0.001546	0.000165	0.025122
1	LassoLars	0.001157	0.009009	-0.018522	0.002394	-0.000347	0.027074
2	Lasso	0.001157	0.009020	-0.018573	0.002404	-0.000352	0.027080
3	ElasticNet	0.001163	0.009056	-0.019113	0.002536	-0.000506	0.027673
4	Ridge	0.001165	0.008973	-0.019597	0.002578	-0.000694	0.029431
5	PLS	0.139533	0.136886	-0.072733	0.012281	-0.010945	0.203847
6	BayesianRidge	0.001167	0.008995	-0.019709	0.002612	-0.000717	0.029397
7	LinearRegression	0.001162	0.008942	-0.019450	0.002535	-0.000666	0.029484
8	PCR	0.111607	0.173119	0.317650	-0.700798	1.042336	-0.445308
9	LassoLarsIC	0.001171	0.008173	-0.014805	0.001601	0	0.026628
10	Lars	0.135211	0.120728	-0.109940	0	0	0.195261

Driving and load related practices have also impact on fuel consumption. The payload weight (kg_load.coef) is always positive, meaning that when loads are heavier, slightly more fuel is needed to carry them because it takes more energy. It is possible that lengthier loading/unloading times will result in slightly higher overall fuel usage because of the necessity for refrigeration and idling during this time. This is indicated by the positive unload time (unload_time.coef). In most models, the average speed (avg_spd.coef) have negative coefficients, while the maximum speed (max_spd.coef) has positive coefficients. This implies that traveling at a high speed, which increases fuel consumption owing to aerodynamic drag, is less fuel-efficient than maintaining a constant, moderate pace. Mixed coefficients are shown for engine off periods (off_eng.coef), with some models suggesting very little fuel use at these times. On the other hand, idle time with the engine running (on_eng.coef) has a substantial positive coefficient in all models, indicating that it continues to operate the engine and the refrigeration system, which significantly increases fuel consumption.

These analyses offer insightful information about the primary factors influencing fuel use in cold chain logistics, highlighting the significance of van selection, well prepared route planning, and effective load management as ways to reduce fuel consumption.

5.3. Interpretation of Fuel Consumption Patterns

Fuel consumption patterns in cold chain logistics entail a very complex study of the various factors that come into play to determine the amount of fuel used in the transportation of temperature-sensitive goods. Descriptive statistics, a correlation matrix, and the coefficients from the multiple machine learning models together give important insights into these patterns.

Looking at the descriptive statistics, we see high variability of the key factors influencing fuel consumption: default fuel consumption rate, vehicle age, mileage, maintenance frequency, and vehicle volume. For example, the default fuel consumption rate averages 7.75 liters per 100 kilometers, with a relatively small standard deviation, which is consistent with fuel consumption across different types of vehicles. Vehicle age and mileage show higher variability, indicative of very different conditions and usage patterns throughout the fleet; therefore, these could be very impacting factors for fuel consumption efficiency. Regular maintenance, measured by weeks since the last service, is essential in maintaining the efficiency of fuel consumption, with vehicle age and mileage, older and heavily used vehicles tend to use more fuel if not properly serviced. The correlation matrix further details the relationships between these variables. For instance, a positive relationship is evident between the default fuel consumption and vehicle age, which means that older vehicles tend to use more fuel. Likewise, vehicle maintenance is negatively correlated with fuel consumption, emphasizing the importance of regular maintenance for minimizing the use of fuel. Vehicle volume is also positively related to fuel consumption, consistent with the notion that large vehicles, though they can carry more tons of load, consume more fuel.

Coefficients from the different machine learning models, ARDRegression, Lasso, Ridge, and ElasticNet, indicate the degree to which each of the factors influences fuel consumption. Vehicle-related factors include vehicle default fuel consumption rate and vehicle age. The high value of the default fuel consumption rate for all four models shows that it is a critical factor in determining overall fuel consumption. Vehicle age shows a negative value in models such as Ridge, where the resulting coefficients would indicate that newer vehicles are more fuel-efficient. Temperature-related factors include ambient maximum temperature and the setting of the frozen cabin. Coefficients for ambient temperature are relatively small, thus having a moderate effect, but the setting of the frozen cabin has a more remarkable effect, with most models indicating increased fuel consumption to maintain very low temperatures. Route-related factors greatly impact fuel consumption patterns as reflected by coefficients for forward and return route distances and times. Generally, longer distances are correlated with reduced per-kilometer fuel consumption, possibly resulting from improved driving conditions over long hauls. On the other hand, the longer travel time increases fuel consumption because of factors such as idling and varying traffic conditions. Load and driving-related factors indicate that heavier payloads and longer unload times increase fuel consumption. Average speed has a negative coefficient, which suggests that maintaining a steady and moderate speed is more fuel-efficient. High maximum speeds, on the other hand, increase fuel consumption due to aerodynamic drag. Idle time, particularly with the engine running, significantly contributes to higher fuel consumption since the engine and the refrigeration system are on at all times.

In conclusion, this research demonstrates several aspects of fuel consumption patterns in cold chain logistics. It highlights the importance of optimizing vehicle maintenance, route planning, load management, and driving behavior to boost fuel efficiency. With knowledge about these factors, logistics companies can significantly reduce their fuel use, save costs, and lessen their environmental impact.

5.4. Implications and Recommendations for Cold Chain Logistics Efficiency

The implications of this research on cold chain logistics efficiency highlight the potential transformative power of advanced technologies—especially machine learning models—to optimize fuel consumption, reduce costs, and enhance environmental sustainability. By drawing on data-driven insight and predictive analytics, companies in the cold chain logistics sector may develop strategies that significantly improve operational efficiency. In this sense,

the environmental advantages of machine learning and other advanced technologies in cold chain logistics are huge, mainly because such technologies aid in fuel consumption optimization and so significantly reduce greenhouse gas emissions. This contributes to the larger objective of rendering supply chain operations sustainable. Companies are then in a position to align their practices with environmental regulations and standards, advancing their corporate social responsibility profiles in light of a growing consumer demand for logistics practices that are environmentally friendly.

Machine learning models offer robust tools for the analysis of huge amounts of data, which will allow discovery of patterns and correlations not immediately obvious by any other method. These models may be used in forecasting fuel consumption based on a vast variety of factors: vehicle type, weight of the load, route characteristic, and environmental conditions. The predictive ability of such models means that firms can optimize the logistics operations so that they could minimize the fuel used and the attendant costs. For instance, machine learning algorithms could predict the impact of specific variables on fuel consumption and thus enable logistics managers to adjust their strategy proactively. Integration with Internet of Things devices further increases the efficiency of the cold chain logistics. IoT devices allow for the real-time monitoring and data collection on various parameters related to temperature, humidity, and vehicle performance. This real-time data can be fed into machine learning models to update and refine the predictions continuously, ensuring the logistics operations dynamically adapt to changing conditions. Integration helps to keep the temperature for perishable goods at an optimal level, thus reducing the energy for refrigeration and subsequently the fuel consumption. Advanced technologies also help in improved resource allocation and planning. Machine learning models can help determine the most efficient use of refrigeration units, ensuring that they operate within optimal parameters. That helps in energy conservation, as well as in the increased longevity of the equipment, and hence cost savings in maintenance and replacement. In addition, predictive maintenance, powered by machine learning, can anticipate equipment failures before they ever take place and allow interventions that are on time and prevent costly downtimes and inefficient fuel consumption.

In sum, the fusion of advanced technologies and machine learning models heralds a new frontier for cold chain logistics. Being able to harness the power of predictive analytics and real-time data, companies are capable of huge improvements in fuel efficiency, cost efficiency, and environmental sustainability. This practice will not only optimize present operations but also set a runway for continuous improvement and innovation within the logistics sector to ensure resilience and competitiveness in a changing market.

6. Limitations and future research

Key limitations to the study include the scope of the data used for analysis. While the dataset used in this work is rich in information, increasing the size of the datasets, covering longer time periods and using larger numbers of vehicles, would make the findings more robust. Future studies should collect and analyze data across multiple seasons and years for a better understanding of fuel consumption patterns and factors that cause their variations. This will include all forms of operational conditions and variables.

This research is based mainly on conventional diesel vehicles. As the logistics industry evolves, taking into consideration the various emerging technologies, electric vehicles, hydrogen-powered vehicles, and hybrid vehicles, is likely to give additional insights into fuel efficiency and environmental impact. The operational characteristics and efficiencies of these technologies present unique features that are likely to change significantly in the dynamics of cold chain logistics. Future studies need to investigate the incorporation and performance of such alternative vehicles in cold chain logistics frameworks.

The study might also benefit from the consideration of a more extended list of variables that affect fuel consumption. Critical components in the determination of fuel efficiency, such as road quality, road gradient, and driver aggressiveness, were not comprehensively studied in this research. Further research can include such variables to come up with more exact and holistic models of fuel consumption in cold chain logistics. Inquiring about how shifting drivers' behaviors or different road conditions affect fuel utilization might lead to more accurate and actionable recommendations.

Geographical diversity is another ground for future research. Results are based on data from a particular region, which may not be wholly representative of global cold chain logistics operations. Replicating this study in various parts of the world would help to validate findings, as well as test their applicability across different climates, infrastructural qualities, and logistical challenges. This may be done through comparative studies in different regions.

Lastly, the different time windows used by a fleet could be a determinant in fuel consumption. Although this study is based mostly on an average operation time, different times of the day, such as early morning, afternoon, and evening, could affect fuel efficiency because of variables such as traffic, temperature changes, and driver fatigue. Other future studies should look at how such time-related variables affect fuel consumption so that logistics firms can optimize their operations based on time. In brief, though this study informs significantly on fuel consumption in cold chain logistics, further research will be necessary to enhance the understanding and optimization of cold chain logistics efficiency.

7. Conclusion

Analyzing the fuel consumption pattern in cold chain logistics shows a variety of factors that interact with fuel usage during the transportation of temperature-sensitive goods. The variability in default fuel consumption rate, vehicle age, mileage, maintenance frequency, and vehicle volume is attested to by the descriptive statistics. The correlation matrix shows how such variables are correlated with each other. For example, a positive correlation between the default fuel consumption and vehicle age indicates that older vehicles are more likely to consume more fuel; in contrast, the negative correlation between route distance and fuel consumption underscores the importance of route planning in minimizing fuel usage. Indeed, the coefficients of the machine learning models ARDRegression, Lasso, Ridge, and ElasticNet show in detail the quantitative influence of each factor on fuel consumption. Vehicle-related factors, such as default fuel consumption rate and vehicle age, are found to be strong predictors, with the effect of vehicle age suggesting better fuel efficiency of newer vehicles. Temperature-related factors, consisting of ambient maximum temperature and the setting of the frozen cabin, come moderately significant in consumption patterns. Route-related factors are found to have a significant effect on fuel consumption, based on the distance and time of both the forward and return routes, where the longer time taken to travel resulted in increased fuel consumption due to factors such as idling and varying traffic conditions. Load and driving-related factors demonstrate that heavier payloads and longer unload times increase fuel consumption, while keeping the speed steady and moderate tends to be the most fuel-efficient strategy, as opposed to driving at low or high speeds.

In conclusion, the study highlights that fuel consumption in cold chain logistics is multifaceted. It emphasizes the need for optimization of vehicle maintenance, route planning, and management of load and driving behavior for more effective fuel efficiency. In so doing, logistics companies could drastically cut down on fuel consumption, which results in cost savings and other environmental benefits. Integration with the machine learning model provides a robust framework to predict fuel consumption and detect optimization opportunities to help the logistics industry base its continuous improvement and innovation.

8. List of Figures:

9. References:

1. Capo, C. (2021) When physical experiments meet machine learning experiments for the understanding and prediction of the ageing of refrigerated transport vehicles. Mechanics [physics.med-ph]. Université de Lyon. Retrieved from https://theses.hal.science/tel-03358954v1/document
2. Wang, S., Tao, F., & Shi, Y. (2018). Optimization of Location–Routing Problem for Cold Chain Logistics Considering Carbon Footprint. International Journal of Environmental Research and Public Health, 15(1), 86. https://doi.org/10.3390/ijerph15010086
3. IPCC. (2019). Special Report on Climate Change and Land. Chapter 5 — Food Security. Retrieved from https://www.ipcc.ch/srccl/chapter/chapter-5/
4. Marchi, B., Bettoni, L., & Zanoni, S. (2022b). Assessment of energy efficiency measures in food cold Supply chains: A Dairy Industry case study. Energies, 15(19), 6901. https://doi.org/10.3390/en15196901
5. Rahman, M. H., Rahman, M. F., & Tseng, T. (2022). Estimation of fuel consumption and selection of the most carbon-efficient route for cold-chain logistics. International Journal of Systems Science: Operations & Logistics, 1–17. https://doi.org/10.1080/23302674.2022.2075043
6. Ning, T., Han, Y., & Fang, M. (2023). Research on cold chain logistics optimization model considering low-carbon emissions. International Journal of Low-carbon Technologies, 18, 354–366. https://doi.org/10.1093/ijlct/ctad021
7. Chandran, R., Hasanuzzaman, M., Arıcı, M., & Kumar, L. (2022). Energy, economic and environmental impact analysis of phase change materials for cold chain transportation in Malaysia. Journal of Energy Storage, 55. https://doi.org/10.1016/j.est.2022.105481
8. Han, J., Zuo, M., Zhu, W., Zuo, J., Lü, E., & Yang, X. (2021). A comprehensive review of cold chain logistics for fresh agricultural products: Current status, challenges, and future trends. Trends in Food Science and Technology, 109, 536–551. https://doi.org/10.1016/j.tifs.2021.01.066
9. Grand View Research. (2022). Cold Chain Market Size, Share & Trends Analysis Report By Type (Storage, Transportation, Packaging, Monitoring Components), By Temperature Range, By Application, By Region, And Segment Forecasts, 2023 - 2030. Retrieved from https://www.grandviewresearch.com/industry-analysis/cold-chain-market
10. Chaudhuri, A., Dukovska-Popovska, I., Subramanian, N., Chan, H. K., & Bai, R. (2018). Decision-making in cold chain logistics using data analytics: a literature review. The International Journal of Logistics Management, 29(3), 839–861. https://doi.org/10.1108/ijlm-03-2017-0059
11. Wang, Z., Leng, L., Wang, S., Li, G., & Zhao, Y. (2020). A Hyperheuristic Approach for Location-Routing Problem of Cold Chain Logistics considering Fuel Consumption. Computational Intelligence and Neuroscience, 2020, 1–17. https://doi.org/10.1155/2020/8395754
12. Emenike, C. C., Van Eyk, N. P., & Hoffman, A. J. (2016, November 1). Improving Cold Chain Logistics through RFID temperature sensing and Predictive Modelling. IEEE Xplore. https://doi.org/10.1109/ITSC.2016.7795932
13. International Monetary Fund. (2023). Data for a Greener World. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/Books/Issues/2023/04/04/Data-for-a-Greener-World-A-Guide-for-Practitioners-and-Policymakers-522462

14. Ren, Q., Fang, K., Yang, X., & Han, J. (2022). Ensuring the quality of meat in cold chain logistics: A comprehensive review. Trends in Food Science and Technology, 119, 133–151. https://doi.org/10.1016/j.tifs.2021.12.006
15. Ren, T., Ren, J., Matellini, D. B., & Ouyang, W. (2022). A comprehensive review of modern cold chain shipping solutions. Sustainability, 14(22), 14746. https://doi.org/10.3390/su142214746
16. Xu, B., Sun, J., Zhang, Z., & Gu, R. (2023). Research on Cold Chain Logistics Transportation Scheme under Complex Conditional Constraints. Sustainability, 15(10), 8431. https://doi.org/10.3390/su15108431
17. Zhang, X., Sun, Y., & Sun, Y. (2022). Research on Cold chain Logistics Traceability System of fresh agricultural products based on Blockchain. Computational Intelligence and Neuroscience, 2022, 1–13. https://doi.org/10.1155/2022/1957957
18. Ashok, A., Brison, M., & LeTallec, Y. (2017). Improving cold chain systems: Challenges and solutions. *Vaccine*, *35*(17), 2217–2223. https://doi.org/10.1016/j.vaccine.2016.08.045
19. Xie, T., & Zhao, M. (2016). Research on cold chain logistics joint distribution model based on cloud logistics. https://doi.org/10.1109/imcec.2016.7867320
20. Katreddi, S. (2023). Development of Machine Learning based approach to predict fuel consumption and maintenance cost of Heavy-Duty Vehicles using diesel and alternative fuels. https://doi.org/10.33915/etd.11780
21. Tan, D., Suvarna, M., Tan, Y. S., Li, J., & Wang, X. (2021). A three-step machine learning framework for energy profiling, activity state prediction and production estimation in smart process manufacturing. *Applied Energy*, *291*, 116808. https://doi.org/10.1016/j.apenergy.2021.116808
22. Zhang, J., Cao, W., & Park, M. (2019). Reliability analysis and optimization of cold chain distribution system for fresh agricultural products. Sustainability, 11(13), 3618. https://doi.org/10.3390/su11133618
23. Fares, N., Lloret, J., Kumar, V., Frederico, G. F., & Kamach, O. (2023). A hybrid framework for fleet management with quality concerns: a case for the food industry. International Journal of Quality & Reliability Management. https://doi.org/10.1108/ijqrm-08-2022-0241
24. Al-Wakkal, W. (2020). A framework for sustainable cold chain logistics in Over-The-Counter (OTC) drugs. Retrieved from https://hdl.handle.net/20.500.12380/300800
25. Chen, Y. (2020). Intelligent algorithms for cold chain logistics distribution optimization based on big data cloud computing analysis. Journal of Cloud Computing, 9(1). https://doi.org/10.1186/s13677-020-00174-x
26. Leng, L., Zhang, J., Zhang, C., Zhao, Y., Wang, W., & Li, G. (2020). A novel bi-objective model of cold chain logistics considering location-routing decision and environmental effects. PLOS ONE, 15(4), e0230867. https://doi.org/10.1371/journal.pone.0230867
27. Wang, S. (2022). Study on cold chain Logistics operation and risk Control of fresh e-Commerce products. *Advances in Multimedia*, *2022*, 1–11. https://doi.org/10.1155/2022/7272370
28. Kirby, H. R., Hutton, B., McQuaid, R. W., Raeside, R., & Zhang, X. (2000). Modelling the effects of transport policy levers on fuel efficiency and national fuel consumption.

Transportation Research Part D: Transport and Environment, 5(4), 265–282. https://doi.org/10.1016/s1361-9209(99)00037-1

29. NHTSA. (2010). Factors and Considerations for Establishing a Fuel Efficiency Regulatory Program for Commercial Medium- and Heavy-Duty Vehicles. United States Department of Transportation - (National Highway Traffic Safety Administration). CAFE Phase 1: NHTSA Study. Retrieved from https://www.nhtsa.gov/document/cafe-phase-1-nhtsa-study
30. Smith, C., Kent, J., & Roberts, C. (2007). An assessment of factors affecting fleet fuel performance. Journal of Transportation Management, 18(1), 65-88. Retrieved from https://digitalcommons.cwu.edu/cgi/viewcontent.cgi?article=1163&context=cobfac
31. Maiorino, A., Petruzziello, F., & Aprea, C. (2021). Refrigerated Transport: state of the art, technical issues, innovations and challenges for sustainability. Energies, 14(21), 7237. https://doi.org/10.3390/en14217237
32. Behdani, B., Fan, Y., & Bloemhof-Ruwaard, J. (2019). Cool chain and temperature-controlled transport: An overview of concepts, challenges, and technologies. In Elsevier eBooks (pp. 167–183). https://doi.org/10.1016/b978-0-12-813411-5.00012-0
33. Qin, G., Tao, F., & Li, L. (2019). A vehicle routing optimization problem for cold chain logistics considering customer satisfaction and carbon emissions. International Journal of Environmental Research and Public Health, 16(4), 576. https://doi.org/10.3390/ijerph16040576
34. Chen, J., Liao, W., & Yu, C. (2021). Route optimization for cold chain logistics of front warehouses based on traffic congestion and carbon emission. Computers & Industrial Engineering, 161, 107663. https://doi.org/10.1016/j.cie.2021.107663
35. Guo, X., Zhang, W., & Liu, B. (2022). Low-carbon routing for cold-chain logistics considering the time-dependent effects of traffic congestion. Transportation Research Part D: Transport and Environment, 113, 103502. https://doi.org/10.1016/j.trd.2022.103502
36. Zhao, Z., Li, X., & Zhou, X. (2020). Optimization of transportation routing problem for fresh food in time-varying road network: Considering both food safety reliability and temperature control. PLOS ONE, 15(7), e0235950. https://doi.org/10.1371/journal.pone.0235950
37. Fan, Y., De Kleuver, C., De Leeuw, S., & Behdani, B. (2021). Trading off cost, emission, and quality in cold chain design: A simulation approach. Computers & Industrial Engineering, 158, 107442. https://doi.org/10.1016/j.cie.2021.107442
38. Jia, X. (2022). Research on the Optimization of Cold chain Logistics Distribution Path of Agricultural Products E-Commerce in urban Ecosystem from the perspective of Carbon Neutrality. Frontiers in Ecology and Evolution, 10. https://doi.org/10.3389/fevo.2022.966111
39. Tassou, S., De-Lille, G., & Ge, Y. (2009). Food transport refrigeration – Approaches to reduce energy consumption and environmental impacts of road transport. Applied Thermal Engineering, 29(8–9), 1467–1477. https://doi.org/10.1016/j.applthermaleng.2008.06.027
40. Mercier, S., Villeneuve, S., Mondor, M., & Uysal, I. (2017). Time–Temperature Management along the food Cold Chain: A review of Recent developments. Comprehensive Reviews in Food Science and Food Safety, 16(4), 647–667. https://doi.org/10.1111/1541-4337.12269

41. Mills, S. A. (2019). Design of experiment and analysis techniques for fuel consumption data using heavy-duty diesel vehicles and on-road testing. https://doi.org/10.33915/etd.3815
42. Zhang, C., He, J., Bai, C., Yan, X., Gong, J., & Zhang, H. (2021). How to use Advanced fleet Management System to promote energy saving in transportation: A survey of Drivers' Awareness of Fuel-Saving Factors. Journal of Advanced Transportation, 2021, 1–19. https://doi.org/10.1155/2021/9987101
43. Zhao, D., Li, H., Hou, J., Gong, P., Zhong, Y., He, W., & Zhi-Jun, F. (2023). A review of the Data-Driven Prediction Method of Vehicle Fuel Consumption. Energies, 16(14), 5258. https://doi.org/10.3390/en16145258
44. Wickramanayake, S., & Bandara, H. M. N. D. (2016). Fuel consumption prediction of fleet vehicles using Machine Learning: A comparative study. Fuel Consumption Prediction of Fleet Vehicles Using Machine Learning: A Comparative Study. https://doi.org/10.1109/mercon.2016.7480121
45. Hamed, M. A., Khafagy, M. H., & Badry, R. M. (2021). Fuel Consumption Prediction Model using Machine Learning. International Journal of Advanced Computer Science and Applications, 12(11). https://doi.org/10.14569/ijacsa.2021.0121146
46. Kale, S. D., & Patil, S. D. (2020). Need for predictive data analytics in cold chain management. In Lecture notes in electrical engineering (pp. 115–129). https://doi.org/10.1007/978-981-15-6229-7_9
47. Li, L., Yang, Y., & Qin, G. (2019). Optimization of integrated inventory routing problem for cold chain logistics considering carbon footprint and carbon regulations. Sustainability, 11(17), 4628. https://doi.org/10.3390/su11174628