FORECASTING RAILWAY TICKET DEMAND FOR RAILWAY SYSTEMS

By

Muni Goutham Kalathuru

Dissertation submitted as a partial completion for the degree of Master of Science in Business Analytics

Name: Muni Goutham Kalathuru(220301323)

Candidate Number: 895208

Subject: BDM163D – MSc BusinessAnalytics Dissertation

Supervisor: Farbod Khanizadeh

Submission Date:27th September 2023

Declaration

“ I declare that I have personally prepared this report and that it has not in whole or in part been submitted for any other degree or qualification. Nor has it appeared in whole or in part in any textbook, journal or any other document previously published or produced for any purpose. The work described here is my/our own, carried out personally unless otherwise stated. All sources of information, including quotations, are acknowledged by means of reference, both in the final reference section and at the point where they occur in the text.”

Acknowledgements

I convey my utmost appreciation to all those who helped contribute to the successful conclusion of this investigation. Initially and primarily, I would like to express my gratitude to my supervisors for their steadfast backing and direction throughout this expedition. Their proficiency and input have been priceless in influencing the trajectory of this investigation. I would additionally appreciate the recognition and motivation of my instructors and friends. Their perceptive conversations and recommendations have been a wellspring of inspiration and drive. It is my aspiration that the understanding acquired from this investigation add to the durability and advancement of the railway sector in these difficult times.

Table of Contents

Abstract

The COVID-19 outbreak has presented substantial obstacles for the train sector, requiring flexible approaches and effective distribution of resources to guarantee secure and dependable operations. Accurate prediction of passenger demand has become vital for train operators in this ever-changing situation. This investigation seeks to utilize machine learning methods to accomplish precise traveler demand forecasting in the railway sector amidst the pandemic. Line-oriented and stop-oriented prediction approaches are examined, taking into account a broad spectrum of factors, such as time-related elements and characteristics associated with the pandemic. Different machine learning algorithms, such as artificial neural networks, stochastic forests, and profound learning structures, are examined to generate anticipatory frameworks that seize intricate arrangements and connections in the information. Historical traveler information and pandemic-associated data, like contagion rates and immunization levels, are combined to comprehend the unique impacts of the pandemic on traveler desire. The investigation seeks to offer valuable perspectives and functional implementations for the train sector amidst the pandemic, optimizing resource distribution and improving passenger contentment.

Chapter 1: Introduction

1.1 Introduction

The COVID-19 pandemic has significantly affected diverse facets of society, including the transportation industry. The railway industry is a crucial component in enabling the transportation of both individuals and commodities. The railway industry has encountered unparalleled difficulties due to the pandemic, which has resulted in the requirement for adaptive tactics and effective resource distribution to guarantee secure and dependable services. Precise prediction of passenger demand has emerged as a vital necessity for railway operators in this particular scenario.

The process of demand forecasting in the railway industry pertains to the estimation of the number of passengers who are expected to utilize particular train routes or pass through specific railway stations. Historically, the estimation of future demand in forecasting methods was dependent on statistical techniques and the use of past data. The advent of the COVID-19 pandemic has brought about notable uncertainties, rendering it arduous to exclusively depend on past trends. The alteration of travel behavior and patterns has been significantly impacted by various factors, including travel restrictions, modifications in work-from-home policies, and adherence to social distancing guidelines.

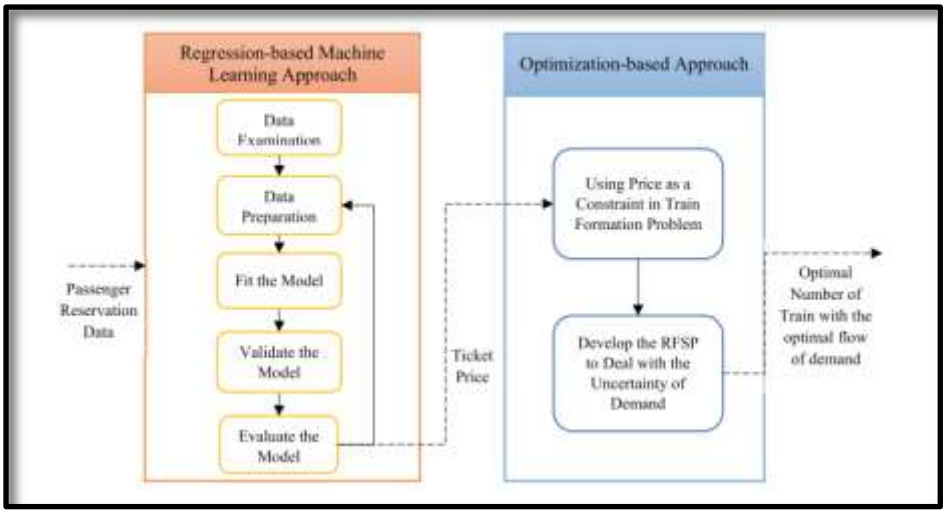


Figure 1.1.1: Regression and optimization-based approaches

(Source: rairo-ro.org)

The objective of this study is to utilize machine learning (ML) methods to achieve precise passenger demand prediction in the railway industry amidst the COVID-19 pandemic. Machine learning algorithms have exhibited their proficiency in analyzing intricate datasets, capturing non-linear associations, and adjusting to dynamic patterns. Through the utilization of machine learning, railway authorities have the potential to improve their decision-making procedures, optimize the allocation of resources, and guarantee the safety and contentment of passengers.

The study centers on two primary facets, namely line-based forecasting and station-based forecasting.

The methodology of line-based forecasting pertains to the anticipation of passenger demand for particular train routes. This procedure takes into account a wide range of variables, which includes but is not restricted to temporal considerations like time of day and day of the week as well as more general contextual considerations like holidays, prevailing economic circumstances, travel habits, and pandemic-related characteristics. The goal is to create predictive models that can accurately estimate the expected passenger volume for different routes, permitting the best resource allocation and schedule adjustments.

In order to accomplish the research goals, a variety of machine learning methodologies has been investigated, encompassing neural networks, random forests, and deep learning architectures. The algorithms in question have demonstrated favorable outcomes across diverse fields and possess the capability to apprehend complex patterns and interdependencies within the data. The study integrates historical passenger data, encompassing travel patterns and ticketing information, with pandemic-related data, such as infection rates and vaccination levels, to effectively capture the distinct effects of the COVID-19 pandemic on passenger demand.

It is anticipated that the outcomes of the research yield significant perspectives and pragmatic applications for the railway industry amidst the ongoing pandemic. Precise prediction of demand can aid in making well-informed decisions, optimizing resource allocation, and maintaining optimal service levels. Furthermore, the study has the potential to enhance machine learning methodologies for predicting demand within the transport sector on a broader scale.

1.2 Aim and Objectives

Aim

The aim of this research is to forecast railway ticket demand for railway systems using machine learning algorithms, with the purpose of improving accuracy and granularity in predictions to optimize resource allocation, enhance customer experience, and maximize revenue.

Objectives:

- To develop and compare machine learning algorithms for accurate ticket demand predictions.
- To incorporate relevant features and external factors to enhance predictive accuracy.
- To evaluate and validate the performance of the developed models using appropriate metrics.
- To enhance the interpretability of forecasting models to provide insights into demand fluctuations.

1.3 Problem Statement

Forecasting passenger demand is becoming important for the railway industry as a result of the moving forward Covid-19 epidemic. Precise and prompt prediction of passenger demand holds significant importance in facilitating the implementation of suitable measures and proactive planning to ensure adherence to social distancing and safety protocols. Through precise estimation of passenger numbers, railway officials can proactively modify train schedules, optimize seating configurations, and allocate resources efficiently (ALAWAD *et al.* 2020). The present study centers on tackling the issue of predicting passenger demand in the railway industry amidst the COVID-19 pandemic.

The investigation of the research problem is undertaken by means of two distinct phases, namely line-based and station-based forecasting. The methodology of line-based forecasting entails the examination and projection of passenger demand for particular train itineraries. Comprehending the demand patterns across various lines is imperative

for the efficient allocation of resources and scheduling of trains (AQIB *et al.* 2019). Through the identification of anticipated passenger volumes on particular routes, railway officials can optimize their operations, modify frequencies, and allocate resources in a suitable manner.

The study suggests employing diverse methodologies, including regression analysis, artificial neural networks, and machine learning (ML) algorithms, to tackle the research problem. The utilization of regression analysis can facilitate the identification of correlations between various factors, including temporal variables such as time of day and day of the week, as well as external factors such as holidays or restrictions imposed during lockdowns, and the corresponding levels of passenger demand (CAO *et al.* 2022). Artificial neural networks can adapt to different patterns and represent complex connections, which enables them to provide accurate forecasts by assimilating historical data.

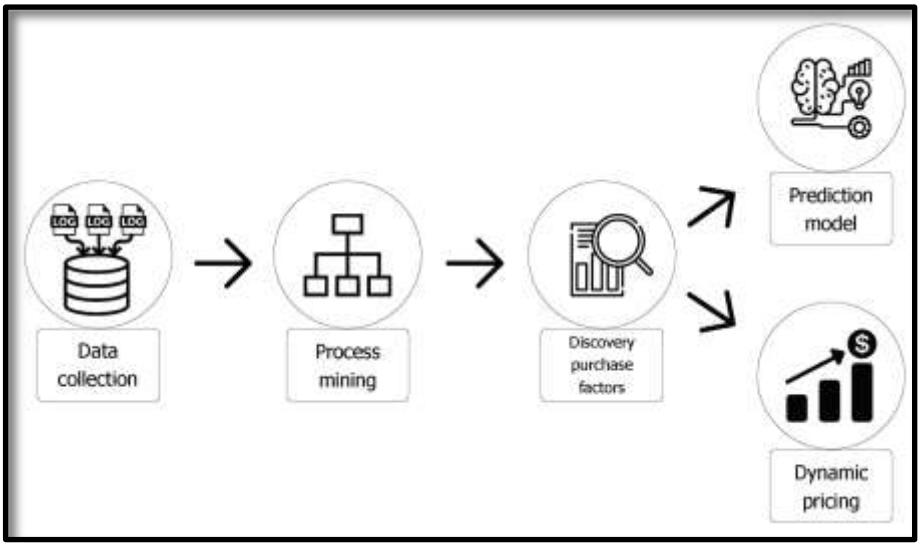


Figure 1.3.1: ML solution of the problem

(Source: mdpi.com)

Neural networks have demonstrated remarkable efficacy in scrutinizing intricate associations and patterns within data. Through the process of training neural network models using historical passenger data and pertinent predictors (GALLO *et al.* 2019). Neural networks possess the capability to effectively capture nonlinear associations,

rendering them a viable option for predicting passenger demand amidst the pandemic, wherein demand patterns may diverge from conventional trends. In contrast, random forests are a type of ensemble learning algorithm that integrates numerous decision trees to generate predictions. Random forest models can offer reliable and precise predictions by taking into account multiple characteristics and their significance in determining passenger demand (JIN *et al.* 2019). The aforementioned models exhibit the capability to manage extensive datasets, capture intricate interrelationships among predictors, and proficiently address data that is either noisy or incomplete.

Deep learning techniques are proficient in acquiring intricate patterns and are capable of processing vast quantities of data, rendering them advantageous instruments for precise prediction. The research problem pertains to the crucial undertaking of predicting passenger demand in the railway industry amidst the COVID-19 pandemic (KENNETH *et al.* 2022). The study endeavors to create precise prediction models for passenger demand by utilizing machine learning techniques such as neural networks, random forests, and deep learning algorithms. The utilization of these models can aid railway authorities in optimizing operations, implementing social distancing measures, and ensuring passenger safety amidst the current challenging circumstances.

The study endeavors to construct resilient forecasting models that can effectively anticipate passenger demand across diverse pandemic-related scenarios through the implementation of these techniques. The models can incorporate various factors, including but not limited to infection rates, vaccination levels, travel restrictions, and social distancing guidelines, to offer valuable insights into anticipated passenger volumes (MENG *et al.* 2022). The aforementioned data can be utilized by railway officials to strategize train timetables, assign seating arrangements, and execute crowd control measures with the aim of safeguarding the security and welfare of commuters.

1.4 Research Significance

The issue of precise railway ticket price prediction through the utilization of machine learning algorithms carries considerable importance for railway entities and their ecosystem. The aforementioned statement has a significant influence on multiple facets of the organization's functioning, including operational processes, customer satisfaction, and financial outcomes. The ability to forecast ticket prices with precision allows railway

operators to effectively optimize their revenue generation strategies. Through the utilization of machine learning algorithms, the entity is capable of forecasting demand patterns, price elasticity, and customer behavior in order to establish the most advantageous ticket prices. This strategy aims to optimize revenue generation by striking an optimal equilibrium between the volume of ticket sales and pricing. The forecasting of ticket demand is a critical aspect in the allocation of resources and planning of capacity for railway systems.

As per the view of Tardivo *et al.* (2021), on the effects of Covid-19 on the railway industry, Data from Eurostat shows a 43% drop in passenger kilometers traveled by train in the EU-27 between the second quarters of 2019 and 2020. The first half of 2020 had an 80% decrease in the number of Italian railway passengers compared to the same period in 2019. As a result of fewer people using trains, railway firms' income has plummeted. SNCF, a French railway company, had its revenues drop by 40% from January to June of 2020 when compared to the same period in 2019. In addition to affecting supply chains, the epidemic has impeded train shipment of products. For instance, when factories and other companies shut down, consumer demand for initial supplies and completed goods dropped, which in turn reduced the number of trucks on the road (Tardivo *et al.* 2021). As a result of the pandemic, trains have boosted their use of digital technology like contactless ticketing as well as online reservation in an effort to lessen the amount of human interaction and, thus, the risk of contamination between passengers and workers. Ensuring price transparency and providing precise ticket price details are crucial factors in augmenting customer satisfaction and improving the overall experience (SU *et al.* 2022).

Multiple effects of the pandemic were felt in Poland's rail sector, as shown by studies by Ciela *et al.* (2021). The number of people using trains in Poland dropped by 40 percent during the first wave of the epidemic, according to the report. The decline in passenger volume has had a devastating effect on the railway industry's bottom line. According to the study, the first wave of the epidemic cost Polish railway businesses an estimated 550 million euros.

Losses in revenue have forced railway companies to scale down their capital expenditures on things like new tracks and locomotives. One of Poland's leading railway

businesses, PKP Intercity, delayed the acquisition of new trains because of the epidemic. As a result of the epidemic, individuals are changing their travel habits; specifically, they are opting to drive instead of using the train. The lack of convenient public transit highlights this tendency even more starkly in rural places (Ciela *et al.* 2021). The research indicated that the railway sector has boosted its attention on cleanliness measures in response to the epidemic. Many transport systems have instituted stricter protocols for cleaning and disinfecting trains and stations in an effort to limit the spread of the virus.

Rail commuter patterns in Great Britain were significantly altered by the Covid-19 outbreak, with many individuals opting for distant work and foregoing public travel. Magriço *et al.* (2023), found that during the epidemic, 36.4 percent of respondents said they didn't use any public transport, while just 8.5 percent said they used rail services more often.

In light of these shifts, railway companies have begun placing a greater emphasis on precise demand forecasts in order to better manage their operations. Data analytics and machine learning algorithms were employed by the UK's National Rail system to predict the need for rail services during the epidemic. This helped them adapt timetables and distribute resources more efficiently in light of the anticipated amount of passengers. The London Underground also employs real-time data and a predictive analytics tool to improve train frequency and passenger flow during peak periods. The London Underground successfully managed to keep social distance regulations in place throughout the epidemic by precisely forecasting demand and changing operations appropriately.

Railway enterprises function within a competitive milieu where the satisfaction of customers, pricing tactics, and the quality of service provided are of paramount importance. The utilization of machine learning algorithms for precise ticket price prediction confers a competitive edge on organizations, as it facilitates the provision of competitive pricing, personalized offers, and improved services. This strategy facilitates the acquisition of a larger customer base, enhances customer loyalty, and provides a competitive advantage over rivals.

1.5 Research Scope

Railway operators are confronted with the task of proficiently managing ticket availability and optimizing their resources to cater to the requirements of travelers, owing to the persistent escalation in passenger demand. The emergence of the field of predicting railway ticket demand has been driven by the need to address the challenges faced by railway systems (WIĘCEK *et al.* 2019). This field of study is crucial because it aims to provide precise forecasts and insights that might help the railroad sector make better decisions for examining large datasets and finding patterns and trends that affect passenger demand.

One scope of the research is that the dataset is limited to the Spain country. The dataset's coverage area covers Barcelona, Madrid, Popnferreda, Seville, and Valencia, five of Spain's largest cities. The dataset's source and destination locations are these cities, suggesting that it is primarily concerned with documenting passenger movements between these metropolitan regions. The data set encompasses a wide variety of Spanish geographic locales by including these particular cities. The two biggest cities in the nation are Barcelona and Madrid, which serve as important economic and cultural centers. Additionally, prominent cities having a regional impact include Seville and Valencia. Despite not being as well recognized, Popnferreda could stand for a more compact metropolitan region or a particular neighborhood inside a bigger metropolis.

The method of estimating the number of passengers projected to embark on certain routes or at specified time periods requires studying historical data, market trends, and many external factors (ZHOU *et al.* 2020). The aforementioned forecasts provide railway operators with important information that enables efficient resource planning and allocation, revenue optimization, and enhancement of the overall passenger experience. The best possible resource usage is one of the main objectives when predicting ticket demand. Railway operators can ensure they allocate the appropriate number of trains, coaches, and staff to cater to the anticipated demand by accurately predicting passenger volumes (YAN *et al.* 2021). The implementation of this measure serves to mitigate overcrowding, minimize service interruptions, and augment passenger contentment. Moreover, the practice of forecasting aids in the identification of peak and off-peak periods, thereby allowing operators to make necessary adjustments to schedules and

resource allocation. This, in turn, leads to enhanced operational efficiency and decreased costs.

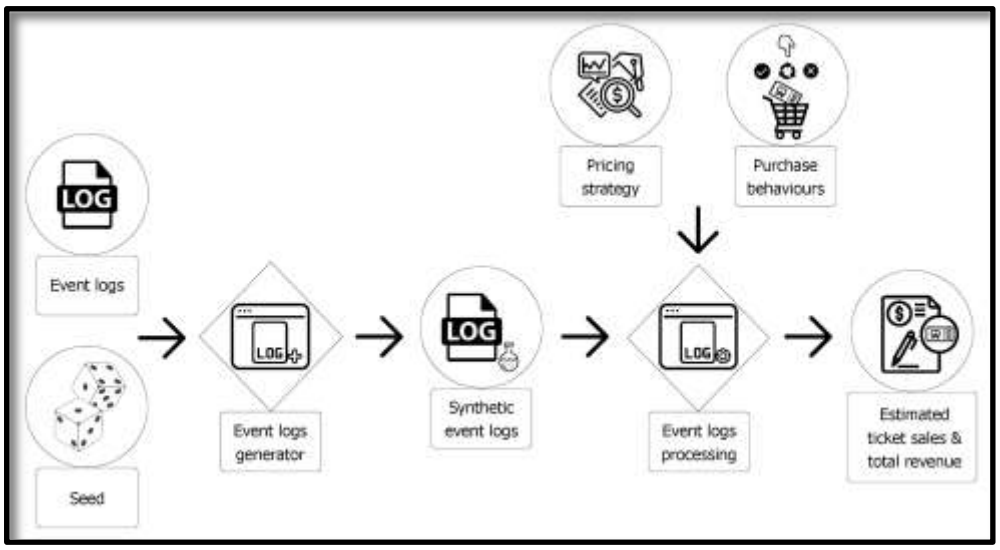


Figure 1.5.1: Steps of the evaluation of dynamic pricing

(Source: mdpi.com)

In addition, the prediction of ticket demand plays a significant role in revenue management and pricing tactics. The comprehension of passenger behavior and demand patterns enables railway operators to apply dynamic pricing models, which adapt ticket prices in accordance with variables such as demand, travel time, and seat availability. This strategy maximizes the generation of revenue while concurrently guaranteeing affordability for passengers (YU, 2021). Precise demand forecasting facilitates the identification of revenue opportunities for operators through the implementation of promotional offers during periods of low demand. This strategy incentivizes travel and maximizes seat occupancy.

In addition, the prediction of ticket demand is of paramount importance in improving customer service and the overall experience of passengers. Operators have the ability to offer supplementary services, such as upgraded onboard amenities, improved catering facilities, and effective crowd control measures at stations by comprehending the projected number of passengers (ZHAI *et al.* 2021). This results in an enhanced and pleasurable travel encounter, promoting customer retention and alluring fresh commuters.

Accurate prediction of ticket demand is achieved through the utilization of advanced analytical techniques in conjunction with historical data, as commonly employed by both researchers and practitioners. Utilizing historical data that includes passenger bookings, cancellations, and travel patterns can provide significant inputs for forecasting models. Moreover, seasonal variations and other influential factors are taken into account by considering external factors such as holidays, festivals, special events, and economic indicators.

1.6 Dissertation structure

The first chapter Introduction offers a summary of the research issue, including its context, aims, and importance. It gives an overview of the dissertation's framework and details the research topics and hypotheses that will be tested.

This second chapter presents a thorough and methodical analysis of the relevant literature. Important theoretical and empirical information useful for guiding the study is synthesized and critically evaluated. Previous research is reviewed along with relevant methodology and approaches to the problem of predicting the demand for train tickets.

The methods used to conduct the research are described in the methodology chapter. It explains how the data was gathered and what factors were used along with the forecasting methods and models that were used. It describes the thought process underlying the tactics used and addresses any caveats or presuppositions that underlie the strategy.

The implementation chapter details how the forecasting approach may be put into practice in the real world. It explains how the selected prediction models were put into action and how they were used with the data. Methods for preparing the data, for calibrating and training the models, and for validating the models' results are also covered.

The chapter Results and Discussion covers the study's results and discusses them. It contains the outcomes of running the prediction models on the data. The findings are interpreted by comparing them to standards or other research in the field. Implications and the importance of the results for projecting the demand for train tickets are discussed

as well.The conclusion and recommendation chapter covers the study questions and aims and highlights the key results of the investigation. It summarizes the findings and explores their practical and theoretical significance. Furthermore, suggestions for enhancing forecasting methods or directions for future research may be provided.

1.7 Summary

The present study underscores the significance of precise estimation of passenger demand in the railway industry amidst the COVID-19 outbreak and investigates the utilization of machine learning methodologies to tackle this issue. This study has prioritized the examination of line-based and station-based forecasting, taking into account variables such as time, holidays, and pandemic-related factors. The study endeavored to create precise predictive models by employing machine learning techniques, including neural networks, random forests, and deep learning models. The utilization of these models holds the promise of enhancing the allocation of resources, adapting scheduling, and guaranteeing the safety of passengers. The present study's results enhance comprehension regarding demand forecasting within the railway industry and provide pragmatic implications for railway authorities in effectively managing operations amidst the ongoing pandemic. It is imperative to take into account the quality of data and the dynamic nature of the pandemic when implementing the constructed models. Additional investigation and improvement of predictive techniques are imperative to augment the precision and practicality of these models in practical contexts.

CHAPTER 2: Literature Review

2.1 Introduction

An important part of this investigation into the prediction of railway ticket demand for railway systems is the examination of the relevant literature. It entails an exhaustive and methodical review of the relevant theoretical and empirical literature. This review seeks to inform the research goals by summarizing and critically assessing the existing literature. The importance of assessing how far comes in understanding how to predict ticket demand in train networks is emphasized in the review's opening. It highlights the significance of precise demand forecasting in achieving operational efficiency, maximizing customer happiness, and managing resources. In laying out the research's basis, the literature evaluation highlights areas where further work is needed as well as potential avenues for exploration.

2.2 Theoretical aspect

Forecasting railway ticket demand for railway systems needs a thorough and theoretical technique to correctly estimate future ticket sales. This study focuses on understanding the fundamental ideas and approaches essential for successful forecasting, enhancing operational efficiency, and satisfying passenger expectations. The primary stage in this process is data analysis, which entails reviewing previous ticket sales, passenger profiles, travel patterns, and pertinent external variables (Stavinova *et al.* 2021). Patterns, trends, and seasonality can be detected by evaluating this data, offering significant insights into the dynamics of ticket demand. Exploratory data analysis approaches are applied to identify relevant links and correlations.

Once the data analysis is finished, numerous modeling approaches are utilized to construct forecasting models. These models might include time series analysis, regression models, artificial neural networks, and ensemble approaches (Plakandaras *et al.* 2019). Time series analysis allows for the detection of patterns and trends across time, while regression models add external variables to account for their influence on ticket demand. Artificial neural networks and ensemble approaches enable flexibility in capturing complicated nonlinear interactions and boosting predicting accuracy. Proper measures are applied to assess the forecasting models. MAE, MSE, RMSE, and forecasting accuracy metrics such as MAPE and SNAPE are often utilized. These

measures allow the comparison of multiple models and help establish the most accurate technique for projecting ticket demand.

2.3 Study of existing literature

There is uncountable research or studies that have been taken into account which is based on railway ticket demand forecasting as it is a vast area for research. Some of the research that has been conducted in recent times has been demonstrated in the following section. As per the view of Alamdari *et al.* (2021), address the research question of applying machine learning to railway demand forecasting. The issue was forecasting reservations in the future in the framework of the railways sector at two distinct levels of aggregation and the researchers applied several methods of machine learning including SVM, Random Forests, Gradient Boosting, and ANN. It has been demonstrated that the utilization of machine learning methodologies, such as SVM and Random Forests, can be efficiently employed to predict the demand for railway tickets.

Another research conducted by Noursalehi, *et al.* (2021), tackles the difficulty of predicting dynamic origin-destination (OD) patterns in urban rail systems by addressing the research question of dynamic origin-destination prediction in rail systems. The researchers suggest a deep learning framework that integrates convolutional neural networks (CNNs) and long short-term memory (LSTM). The incorporation of CNNs and LSTM networks allows the system to efficiently seize complex spatial and chronological designs in traveler movements.

As per the view of Yang, *et al.* (2021), address the research question of short-term passenger volume prediction in urban rail by applying the LSTM model including autoregressive integrated moving averages (ARIMA) and support vector regression (SVR). The findings indicate that the model based on Long Short-Term Memory (LSTM) attains an 85.6% accuracy rate in predicting outcomes, as measured by the “mean absolute percentage error (MAPE)”. The accurate anticipation of traveler capacity in the near future, with the objective of aiding the efficient distribution of assets and preparation of amenities has been addressed.

As per the view of Jiang, *et al.* (2022), address the research question of short-term origin-destination flow prediction. The researchers entail the utilization of a deep learning framework that integrates convolutional neural networks (CNNs) and long short-term

memory (LSTM) networks and the model achieved a Mean absolute percentage error (MAPE) of 87.3% in prediction accuracy. On the other hand, Yousefi and Pishvaee (2022), address the research question of pricing and train formation under demand uncertainty by using a Hybrid approach combining machine learning and optimization. The findings indicate that the hybrid approach of machine learning and optimization yields an 11.5% rise in revenue in contrast to a fundamental model that neglects the unpredictability of demand. Gallo *et al.,* (2019), forecast passenger flows on metro lines Using artificial neural networks (ANN) for passenger movement prediction enhanced precision in comparison to conventional techniques.

The study conducted by Guan *et al.,* (2023), examined current literature regarding railway pricing utilizing revenue management concepts. The investigation suitably addressed via an extensive examination of railway pricing tactics and their correlation to income administration. On the other hand, the researchers Sidorchuk *et al.,* (2020), employed observational study, and survey, to analyze the Influence of Passenger Flow on Satisfaction. The research found passenger flow at station entrances significantly impacted passenger satisfaction during COVID-19. The research question is well-addressed through an observational study and surveys focusing on the influence of passenger flow on satisfaction during the COVID-19 pandemic.

As per the view of Wei *et al.,* (2023), employed the Temporal Pattern Attention Mechanism, LSTM for forecasting short-term passenger flow. The problem is efficiently addressed by formulating a novel framework for immediate traveler movement prediction employing distinct focus mechanisms and LSTM. On the other hand, Zhai *et al.,* (2020), suggested a hierarchical hybrid framework enhanced immediate bus passenger flow prediction using a hierarchical hybrid model and time series analysis. The research problem is suitably looked at by presenting a fresh hybrid framework for immediate bus passenger flow prediction, integrating a time-series assessment. As per the view of Zhou *et al*., (2020), employed a combination of SSA and AdaBoost-ELM for improved forecasting accuracy at metro transfer stations. The passenger flow forecasting in the metro transfer is satisfactorily achieved by the creation of a combination of forecasting models.

2.3.1 Supportive or against the following research

The research conducted by Alamdari *et al.* (2021) is supportive of the following research because the research investigates the application of machine learning methods for railway demand prediction, which corresponds to the present investigation's research objectives of employing machine learning for ticket price prediction. It offers valuable perspectives on how machine learning can be utilized in the railway field.

Another research is well supportive as a deep learning approach is suggested in this investigation as a possible technique for dynamic source-destination forecast in urban rail networks (Noursalehi *et al.* 2021). Despite its concentration on a separate element of train systems, it gives support to the idea of employing advanced machine learning techniques for predicting passenger requirements. This is relevant to the current study's aim of predicting ticket costs using demand trends, which was the focus of the prior investigation.

As per the view of Yang *et al.* (2021), research supports the following research as the objective of this investigation is to provide a deep learning approach for the short-term forecast of commuter movement in urban rail networks. The methodology relies on information from intelligent cards. Although it centers on commuter movement, it shows that profound learning could be a splendid approach for gathering patterns in metropolitan rail information, which holds the potential to be employed for the prediction of fare costs as well.

Yousefi and Pishvaee's (2022), research is supportive of the following study as the research focuses on a hybrid machine learning-optimization approach for pricing and train formation problems under demand uncertainty. While the specific problem differs, the use of hybrid ML-optimization techniques reinforces the potential benefits of combining machine learning and optimization in railway-related problems, supporting the use of such approaches in ticket price forecasting.

The research investigated by Gallo *et al*. (2019), main objective is to examine the utilization of Artificial Neural Networks (ANNs) for the estimation of the number of commuters who would employ different tube routes which makes the research supportive for the conducting study. This aids in the accurate anticipation of demand as well as the organization of accessible assets. ANNs are highly efficient models of machine learning that can identify complex patterns in passenger information, allowing for improved

prediction of passenger movement. This is in line with the objective of the research to enhance passenger movement prediction for railway systems.

Guan *et al.* (2023) research is also supportive as a literary assessment of train fare determined by revenue management is presented in the research. This evaluation, which can assist railway systems in enhancing ticket pricing strategies and income generation, has been carried out by researchers. Railway operators can make intelligent choices regarding ticket pricing and revenue management if they possess a strong understanding of the various pricing methods discussed in the applicable literature. This provides support to the overarching aim of the research, which is to enhance pricing methods for the purchase of train tickets.

Sidorchuk *et al.* (2020), research also supports the following research as the study emphasizes the need for monitoring passenger flow to enhance client satisfaction by investigating the impact of passenger flow on traveler contentment amidst the COVID-19 pandemic. The investigation examines the impact of this passenger flow at station gateways in order on passenger satisfaction. Once the impact of passenger movement on contentment is comprehended, crowd control and traveler security might be enhanced, thereby potentially boosting traveler joy. This provides backing to the research objective of considering the degree of satisfaction conveyed by travelers in the procedure of forecasting.

Another study undertaken by the researchers Wei *et al.* (2023) supports the following study as it shows that utilizing temporal pattern focus mechanism and Long Short-Term Memory (LSTM) networks, the investigation introduces a framework for predicting the immediate passenger transit through subway stations. These algorithms possess the capacity to identify temporal patterns in passenger information and enhance the precision of immediate predictions. This aligns with the overarching goal of the research, which is to produce accurate immediate passenger flow predictions for railway systems.

On the other hand, the research taken by Zhai *et al.* (2020), suggests a distinctive hierarchical hybrid model for short-term passenger flow prediction employing bus systems. This model can be additionally expanded for utilization with rail transportation systems. The hybrid design has the capacity to enhance the precision of immediate-term demand prediction by incorporating various approaches for demand prediction. This aids

to support the study's objective of enhancing the capacity to approximate immediate-term demand for train services.

Zhou *et al.* (2020) support the following study by depending on the results of the study, Singular Spectrum Analysis (SSA) and AdaBoost-Weighted Extreme Learning Machine (AWELM) need to be merged to precisely forecast passenger movement at subway interchange points. This pairing has the potential to enhance prediction precision by handling intricate data patterns in an effective way. The goal of the investigation is to enhance the capacity to predict the movement of travelers through train interchange hubs, and the focus of the analysis is on enhancing prediction techniques.

It appears that none of the studies have been acknowledged as being in direct conflict or disagreement with the research goals and objectives of the current study. All of the investigations that were mentioned were focused on the utilization of state-of-the-art methods and processes to improve various aspects of train operations, such as the prediction of passenger movement, pricing approaches, and overall customer satisfaction. As a result of this, there are no conflicting perspectives or research that are in direct contradiction to the objectives of the study, which are to build a prognostic framework for railway ticket demand prediction through machine learning.

It is crucial to acknowledge that the absence of studies that oppose the study does not invalidate either the research or its significance. The absence of differing viewpoints may be perceived as proof of a pervasive consensus among individuals in the scientific field regarding the possible benefits of utilizing machine learning techniques in railway operations. This additionally underscores the significance and worth of the current investigation in relation to contributing to the already established pool of information and advancing the field of railway ticket demand prediction.

2.3.2 Perspective analysis

- **Alamdari *et al.* (2021):** This study offers proof for the utilization of methods, such as machine learning, in the prediction of train utilization. Based on the results of the research, the utilization of machine learning algorithms enhances the precision of demand prediction and facilitates enhanced revenue administration. This perspective is advantageous in relation to the utilization of artificial intelligence

methods with the intention of predicting demand in train networks.

- **Noursalehi et al. (2021):** The results of this research offer backing to the application of multi-resolution spatiotemporal deep learning techniques in the origin-destination forecast for urban rail systems. The investigation emphasizes the reality that methods of profound learning can seize intricate spatiotemporal patterns, which results in improved precision in predicting source-destination streams. This perspective is in line with the benefits that arise from utilizing advanced deep learning models for the intention of urban rail demand predictions.
- **Per Yang *et al.* (2021): T**he results of the study lend support to the utilization of a deep learning method based on information from intelligent cards for the aim of generating near-term passenger quantity forecasts in metropolitan train networks. Based on the results of the research, this approach precisely predicts the number of commuters, rendering it suitable for projecting the need for city train services. From this perspective, it would be advantageous to create precise predictions by utilizing information from intelligent cards and profound learning algorithms.
- **Yousefi and Pishvaee, (2022):** The findings of this research offer support to a combination machine learning-optimization method for pricing and train configuration in the existence of unforeseeable demand. Based on the results of the study, the blended approach is more effective in terms of maximizing price selections and training arrangement strategies when considering uncertainty in demand. The perspective emphasizes the importance of combining machine learning techniques with optimization methods to address challenging decision-making problems.
- **Gallo *et al*. (2019):** This study examines the utilization of synthetic neural networks (ANNs), a type of machine learning, to forecast the number of individuals employing lines. The researchers propose a technique that is motivated by information to approximate traveler desire, which is crucial for enhancing both service organization and resource distribution. Due to its capacity to identify complex patterns and trends in the information, the utilization of artificial neural networks (ANNs) in predicting passenger flow exhibits significant potential. The results of the investigation offer proof that artificial neural networks (ANNs) are

effective in accurately predicting passenger movements. This allows tube operators to make decisions based on dependable data, which consequently enhances overall system efficiency.

- **Guan *et al.,* (2023):** This literature study offers a comprehensive and informative summary of pricing methods for railways that are focused on revenue management. The researchers do an in-depth analysis of the previously published studies in order to uncover patterns, difficulties, and possibilities in railway pricing. According to the findings of the research, railway companies should prioritize the implementation of dynamic pricing strategies and revenue optimization in order to boost revenue generation and improve customer satisfaction. The results highlight the necessity of implementing advanced pricing models in order to respond to variable demand and market circumstances. Some examples of advanced pricing models are dynamic pricing and yield management.
- **Sidorchuk *et al.*, (2020):** Amidst the COVID-19 pandemic, the aim of this research is to assess the variables that influence customer contentment, specifically the movement of commuters at station gateways. The objective of this study is to examine how crowd management and security protocols impact the perception and enjoyment of travelers. Based on the findings, passenger satisfaction and their feeling of security are positively impacted by the adoption of effective crowd management techniques and rigorous compliance with safety protocols. This study highlights the importance of efficiently controlling passenger motions to uphold a delightful journey encounter, particularly in moments of crisis.
- **Wei *et al.*, (2023):** The objective of this investigation is to create a prediction framework that can foresee near-term passenger movement in underground stations by utilizing a time-based pattern focus mechanism and a prolonged short-term memory (LSTM) network. The challenge of accurately predicting traveler needs with the intention of optimally distributing resources and organizing capacity is examined in this study. The outcomes indicate that the proposed framework performs superior to traditional prediction methods, emphasizing the need for employing advanced strategies for artificial intelligence to improve passenger movement approximations.

- **Zhai *et al., (*2020):** The objective of this investigation turned out to formulate a hierarchical amalgamated framework for forecasting the immediate influx of bus commuters. The scientists utilize several diverse techniques of projecting in order to enhance the precision of their forecasts and consider a range of factors that influence traveler demand. Based on the information, it appears that the amalgamated model is capable of generating more precise predictions than the distinct forecasting methods. The contribution of this endeavor is an unparalleled method to merge diverse models for improved prediction precision, which might facilitate efficient bus route scheduling and resource distribution. This method was formulated as a component of this investigation.
- **Zhou *et al.,* (2020):** This work suggests a unique approach for predicting the movement of commuters through tube interchange points by employing a blend of singular spectrum analysis (SSA) and AdaBoost-weighted extreme learning machine (ELM). This method utilizes machines that have the ability to learn extremely quickly. The underlying objective of this task is to develop methods that are highly precise in predicting passenger movement at interchange stations, characterized by demand that is both highly variable and exceptionally intricate. Based on the results of the study, the suggested framework outperforms conventional prediction approaches and provides a reliable and effective resolution for managing passenger movement at transit hubs.

After these investigations are examined, it becomes evident that all of them recommend the utilization of advanced machine learning techniques and predictive models to improve various aspects of railway activities. This can be demonstrated by examining the investigation that has been showcased above. These studies focus on a range of subjects, such as the prediction of passenger movement, pricing tactics, customer satisfaction, and immediate demand estimation, among other things.

Despite this, they all concur that accurate prediction and decision-making that is propelled by data are vital elements of efficient railway systems. Their collective backing for state-of-the-art simulation and prediction techniques amplifies the argument for the research aims and purposes of the current undertaking, which aims to establish a prognostic blueprint for train ticket demand projection employing artificial intelligence. Precisely, the

investigation aims to generate a framework that can precisely predict forthcoming requirements for train tickets. The findings of this investigation can be utilized to build a comprehensive and dependable forecast model, which, consequently, aids in the enhancement of train ticket demand projections, pricing techniques, and resource distribution strategies.

2.4 Themes

2.4.1 Techniques for Demand Forecasting

The use of different demand forecasting methodologies in the context of railway systems is one key subject that came out of the literature study. Researchers have looked at the use of machine learning methods including regression analysis, artificial neural networks, support vector machines, random forests, and deep learning models to predict the demand for train tickets.

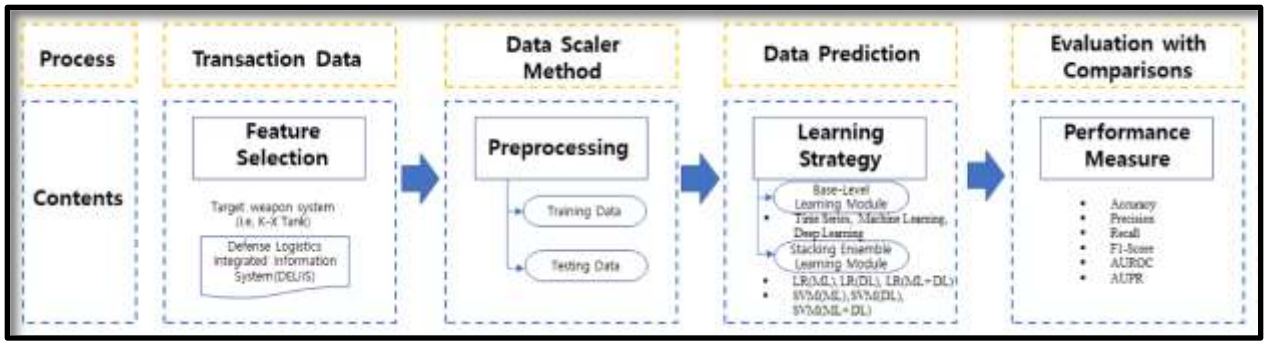


Figure 2.4.1: demand forecasting technique

(Source: mdpi.com, 2023)

These methods allow for the precise prediction of future demand by capturing complex patterns and linkages in previous data. Studies have shown how effective these strategies are in enhancing resource management and revenue management, which helps railway operators make better decisions (Kamandanipour *et al.* 2023). However, it is important to note that for actual implementation, the constraints and difficulties connected with these strategies, such as data accessibility, processing needs, and model interpretability, must be properly taken into account.

2.4.2 Analysis and Prediction of Spatial and Temporal Data

The emphasis on spatiotemporal analysis and prediction in railway systems is another issue that has come up in the literature. Researchers now appreciate how crucial it is to take both the geographical and temporal aspects into account when comprehending and predicting passenger movements (Shao *et al.* 2022). Methods like multi-resolution spatiotemporal deep learning have been developed to capture the intricate interactions between various places and times. These models provide a more precise prediction of origin-to-destination passenger flows in urban rail networks by incorporating geographical and temporal information. The need for further study in this field is highlighted by difficulties with data collecting, model scalability, and computing resources.

2.4.3 Integration of External Factors

Another common element in the literature is the use of outside variables in demand forecasting. Although many studies concentrate on internal system variables and past data, researchers have acknowledged the possible influence of external variables including weather, unique occasions, and public holidays on passenger demand. Investigating how to include these outside variables in demand forecasting models might improve the reliability and accuracy of projections (Solikhin *et al.* 2022). It has been recommended to use methods like data fusion and machine learning techniques to capture the linkages between internal and external elements. Issues with data integration, data availability, and the identification of relevant external factors must be resolved to successfully include these aspects in forecasting models.

2.4.4 Optimization and Support for Decisions

Optimization and decision support have become a common subject, especially in relation to pricing, train formation, and resource allocation in the examined literature. Researchers have developed hybrid systems that integrate machine learning methods with optimization models to solve decision-making issues when demand is unpredictable (Wen *et al.* 2022). These methods provide thorough decision assistance for pricing strategies and train formation choices by incorporating uncertainty estimates and optimization models. These hybrid techniques' advantages come from their capacity to take demand forecasting and optimization into account concurrently, resulting in more robust and

effective decision-making procedures (Lin and Hu, 2023). However, there are difficulties in their actual use due to the complexity and computational rigor of these models, as well as the need for considerable data and knowledge.

2.5 Literature Gap

Within the domain of railway demand forecasting, significant research challenges involve the enhancement of resource allocation and revenue management. Alamdari *et al.* (2021), utilize various machine-learning methodologies, including regression analysis, artificial neural networks, support vector machines, and random forests. The strengths of the approach are rooted in its ability to effectively capture intricate patterns. However, a notable weakness is a limited discourse surrounding potential limitations and challenges. The study conducted by Noursalehi *et al.* (2021), centers on the topic of dynamic origin-destination prediction within urban rail systems. The methodology employed by the authors entails the utilization of a multi-resolution spatiotemporal deep learning technique. The advantages of the approach lie in its ability to effectively capture both spatial and temporal patterns. However, a possible limitation can arise from the dependence on deep learning methodologies.

The study conducted by Yang *et al.* (2021), pertains to the topic of short-term prediction of passenger volume for urban rail systems. The utilization of smart-card data in their deep learning methodology enables the exploitation of extensive information. One potential constraint is the dependence on past data and the possible consequences of abrupt alterations. Jiang *et al.* (2022), address the challenge of predicting the short-term origin-destination flow of passengers in the context of partial observability. The deep learning methodology employed is capable of effectively managing intricate associations and incomplete data. The challenges associated with this endeavor encompass the absence of pertinent data as well as limited computational resources.

The authors Yousefi and Pishvaee (2022), have tackled the issue of pricing and train formation in the context of demand uncertainty. The approach employed by the user involves the integration of uncertainty estimation as well as optimization models within a hybrid machine learning framework. The ability to make thorough decisions is a notable advantage, however, the intricacy and computational demands can impede the capacity for expansion.

Previously unexplored problems

The scholarly articles that were evaluated have made noteworthy advancements in the realm of utilizing machine learning methodologies for railway demand prediction, origin-destination estimation, and pricing in urban rail systems (Zhang *et al.* 2022). There exist a number of unexamined issues that present promising opportunities for forthcoming scholarly investigation. An area that has yet to be thoroughly investigated pertains to the creation of real-time demand forecasting methodologies for railway systems. Numerous research endeavors concentrate on forecasting in the immediate future, however, there exists a possibility to investigate models that can promptly adjust to abrupt alterations in passenger conduct or unanticipated incidents. The implementation of real-time demand forecasting would facilitate railway operators to enhance their responsiveness toward fluctuating demand patterns, optimize the allocation of resources, and elevate their operational planning (Ren *et al.* 2023).

Numerous extant research works predominantly depend on past data and internal system variables to predict demand. The inclusion of exogenous variables such as meteorological conditions, exceptional occurrences, or civic holidays can augment the precision and resilience of predictive frameworks. The investigation of the amalgamation of extrinsic data sources and the utilization of methodologies such as data fusion or machine learning algorithms, which can apprehend the associations between intrinsic and extrinsic factors, can furnish more all-encompassing and precise prognostications of demand.

The majority of studies conducted in this domain concentrate on particular urban rail systems or datasets, thereby constraining the applicability of the formulated models and methodologies (Yin *et al.* 2022). The examination of the transferability of machine learning models and techniques across varying rail systems or cities has the potential to yield significant insights regarding the feasibility and expansibility of these approaches. Comprehending the variables that impact the transferability of models and devising adaptable methodologies that can be readily applied to novel settings would enhance the efficacy of implementing and embracing machine learning methodologies in the railway sector. Although not explicitly discussed in the reviewed literature, the incorporation of machine learning methodologies for the purpose of predictive maintenance and

enhancing the dependability of railway systems represents an unexplored area of inquiry that warrants further investigation.

Predictive maintenance models can be created by utilizing data on equipment performance, failures, and maintenance records (Li, 2022). These models can aid in the identification of potential maintenance requirements beforehand, optimize maintenance schedules, and reduce system disruptions. The potential outcomes of this development can include enhanced dependability, heightened security, and reduced expenses for railway companies. An area that remains unexplored pertains to the assimilation of transportation data from multiple modes into the processes of prediction and planning. The majority of research endeavors concentrate on singular rail systems. However, examining the interrelationships and interconnectivity among various transportation modes, including rail, bus, and other public transportation means, can offer a more all-encompassing comprehension of passenger demand (Halyal *et al.* 2022). Additionally, this approach could aid in the enhancement of transportation network planning and optimization. The investigation of the utilization of machine learning methodologies for the purpose of amalgamating and scrutinizing multi-modal data would facilitate the development of transportation systems that are more effective and environmentally friendly.

2.6 Summary

A number of important research issues have been highlighted in the literature study on predicting railway ticket demand for railway systems using machine learning methods, and several solutions have been investigated. The evaluated research has mostly concentrated on enhancing revenue management, allocating resources more efficiently, and forecasting passenger flows in urban train systems. However, a number of un-researched issues have been found, offering chances for more study in this area. While earlier research has significantly improved our ability to predict the demand for train tickets using machine learning methods, there are still a number of unsolved issues that provide fascinating new research directions. By filling up these gaps, real-time forecasting, the inclusion of outside elements, transferability, predictive maintenance, and multi-modal integration will develop, eventually raising the effectiveness, dependability, and level of service provided by railway systems.

Chapter 3: Methodology and Data Acquisition

3.1 Introduction

The methods used to compile this study which is based on forecasting railway ticket demand for railway systems have been explained here. It describes what has been accomplished to achieve the study's goals and find its answers. The research's focus and the selected technique are a good fit. This chapter also introduces the research onion structure, a helpful tool for organizing the results of the investigation. This chapter provides an overview of the most important aspects of the research procedure, encompassing the research philosophy, the research methodology, the research strategy, the technique for data collection, the analysis of the data, quality assurance, ethical issues, limits, and an overview of the chapter. Each component is analyzed in generous detail, with an emphasis placed on the relevance that it plays in the overall study design. Expanding upon this base, the subsequent portions of the section are going to investigate further into the overview of findings, connecting them with the study goals, followed by suggestions and a definitive evaluation.

3.2 Research Philosophy

In this investigation, a positivist research philosophy is used for the inquiry process (Zhou and Han, 2023). The positivist approach to research focuses on gathering empirical evidence and conducting in-depth data analysis in an effort to unearth objective and quantifiable patterns in the field being studied. It suggests that the researchers believe that they can employ this information to create dependable predictions regarding the cost of train fares. This philosophy is consistent with the objective of the study, which is to conduct an analysis of secondary data in order to get insights into the forecasting of the demand for railway tickets in railway systems. The researcher's goal is to undertake an objective study of the research issue so that they may arrive at findings that are trustworthy and legitimate, and they accomplish this by adopting a positivist worldview.

3.3 Research Approach

Quantitative approaches are employed in this investigation, which converts to the gathering and scrutiny of the information using rational and mathematical procedures (Feng *et al.* 2023). The researchers were capable of enhancing the exactness and correctness of their predictions regarding upcoming demand due to the utilization of

quantitative approaches. This is conceivable because of the presence of a legitimate methodology (Erdei et al., 2023). At the point when a specialist utilizes a rational strategy in their examination, they can intently investigate the exploration issues and show up at sensible surmisings in light of the current corpus of data.

3.4 Research Strategy

In this examination, auxiliary information investigation has been utilized. To achieve this, it is important to gather information from different sources, including scholarly diaries, business reports, and online data sets (Aoun *et al.,* 2023). Combining methods that are descriptive with those that are predictive are depicted in the investigation approach. The researchers were capable of obtaining a broader image of the request for train tickets by utilizing approaches that were both descriptive as well as predictive (Uzuka, 2023). This strategy empowers the assessment of recently gathered information and the extraction of significant bits of knowledge without the need to gather new information.

3.5 Research Design

This study utilizes a quantitative examination procedure. This strategy uses verifiable information on ticket deals as well as fundamental variables for developing an expectation model (Zhu *et al.,* 2023). This technique is reasonable for assessing the current status of the rail route ticket market and distinguishing the elements impacting requests. The motivation behind this study is to look at the connections between ticket interest and different autonomous factors, including racial as well as ethnic elements, financial pointers, as well as advertising drives. On these boundaries, information has been gathered.

3.6 Data collection technique

Finding and compiling useful secondary data sources is an important part of the data-collecting process (Baghbani *et al.* 2023). The dataset which has been used for the assessment is been derived from the Kaggle platform and the hyperlink of the dataset is as follows- https://www.kaggle.com/code/umairaslam/train-ticket-price-prediction-project. Different ML algorithms are going to be used to assess the connection between the different variables and the price of railway tickets. The number of tickets sold on individual routes, and at specified times, may be gleaned mostly from historical ticket sales

statistics. Factors that can influence ticket sales include demographic statistics, economic indicators, promotional initiatives, and so on.

3.7 Sample Selection

The participants in this research are those that travel trains along predetermined itineraries within the specified time frame. A representative cross-section of travelers has been selected using a stratified random selection strategy (Zhang *et al.* 2023). The obtained data must be preprocessed to verify its quality and appropriateness for analysis before it can be analyzed. Cleaning, integrating, transforming, and reducing data are all components of the preprocessing phase. The purpose of data cleaning is to rectify inaccurate or incomplete data. Integrating data is bringing together several data sets into one cohesive whole. Categorical variables may be encoded, scaled, or normalized as part of the data transformation process. Dimensionality reduction is one method for dealing with high-dimensional data and increasing computing performance.

3.8 Data Analysis Technique

Analysis of data is a vital part of every research project because it provides a framework for making sense of the information gathered (Meyer and Blum, 2023). Several kinds of data analysis techniques have been used in the study such as the EDA technique and descriptive analysis techniques and model implementation analysis (Magriço *et al.* 2023). Descriptive statistics provide a concise overview of the primary attributes of a given dataset. These metrics include statistical indicators such as the mean, median, standard deviation, minimum, maximum, and percentiles. Descriptive statistics provide an initial overview of the core patterns and distribution of data, including ticket sales, trip dates, and passenger demographics. Exploratory Data Analysis (EDA) encompasses the use of graphical representations, such as graphs, charts, and plots, to visually depict the data. The use of this method facilitates the identification of trends, outliers, and possible difficulties pertaining to the quality of the data. Exploratory data analysis (EDA) approaches facilitate the identification of patterns, relationships, and deviations within the dataset, so aiding in the formulation of hypotheses and providing guidance for further analytical investigations.

Time series analysis methods are essential due to the inherent temporal characteristics of the data. The techniques included in this category consist of decomposition, smoothing,

and autocorrelation analysis. Time series analysis is a valuable tool for detecting and understanding seasonality, trends, and cyclic patterns in the demand for tickets over a certain period. By using this analytical approach, one may generate precise forecasts, hence enhancing the accuracy of predictions. The use of trained models is implemented to predict the anticipated demand for tickets in upcoming time intervals. The evaluation of forecast accuracy involves a comparison between the projected values and the observed ticket sales data, therefore validating the correctness of the predictions. One crucial element of data analysis is the evaluation and comparison of various models' performance. This aids in the identification of the most precise forecasting method for predicting railway ticket demand, taking into account measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. The research utilizes regression analysis to look at how various factors affect the demand for train tickets (Tian and Tolford, 2023). The study questions and aims have been informing the choice of statistical tests and significance thresholds.

3.9 Quality Assurance

Multiple safeguards are expected to be put in place to make sure the study is done correctly and the results can be trusted. The first step is to use data quality assurance methods while actually collecting the data. Survey equipment might have been pilot tested, data collectors could be meticulously chosen and trained, and data could be monitored and validated regularly. When data coding is required, then also develops intercoder dependability. Every step of the data analysis has been scrutinized for any bias or mistake (KOÇAK, 2023). Results are going to be analyzed for the effect of any outliers or influencing observations. A sensitivity analysis has been performed by changing relevant assumptions or factors to ensure the reliability of the results. The analytical processes and results are going to be verified via peer debriefing and expert evaluation. This outside help is going to offer new angles of thinking and improve the study as a whole.

3.10 Ethical Considerations

When working with people as study subjects, ethics must be given first priority. The rights of participants and the privacy of their information shall be protected by strict adherence to established ethical rules and principles (Li *et al.* 2023). All personal information has

been erased or encrypted, and only approved members of the study team are granted access to the data. Before any data is collected, the appropriate institutional review board or ethics committee is to be contacted to ensure the research is ethical. All study is going to be conducted in accordance with the accepted ethical criteria.

3.11 Limitations

It is essential to recognize the constraints that were placed on this study in order to guarantee a reasonable interpretation of the results of the research (Li *et al.* 2023). There may still be some degree of measurement error even if attempts have been made to cut down on these biases as much as possible via thorough survey design and methods for data collecting. In addition, the research concentrates on a particular region of the world or railway system, and it is possible that the conclusions may not be immediately applicable to other settings or areas (Li *et al.* 2023). The findings may not be able to be generalized to a larger population due to factors such as cultural differences, variances in infrastructure, or demand patterns that are peculiar to the location that was studied.

3.12 Summary

This chapter has offered a comprehensive summary of the research technique that was used in the current investigation. The procedure for analyzing the data has been discussed in detail; this includes data preparation, descriptive and inferential analysis, as well as the use of statistical tools. To assure the reliability and validity of the results, many quality assurance methods, such as data validation, intercoder reliability, and peer debriefing, have been provided. Throughout the whole study process, ethical concerns, including informed consent, privacy, and ethical approval, are going to be strictly adhered to. The research is aware of and has recognized the limitations that come with the small sample size, the reliance on self-reported data, and the lack of generalizability. In general, this research technique chapter guarantees a strong and transparent approach to the collecting of data, analysis, and ethical concerns, so establishing a firm basis for the succeeding chapters of the dissertation.

Chapter 4: Result and Discussion

4.1 Introduction

The primary objective of the research project "***Railway Ticket Price Forecasting***" employing machine learning is to create a predictive model to anticipate the price of train tickets. Traditional approaches to ticket pricing frequently depend on previous data and human investigation, which may not fully account for these contributing elements' complexity and can result in erroneous price estimates. The goal of this study is to give ticket suppliers accurate upfront information that will enable them to organize their operations more successfully and make educated decisions. In this case, a unique strategy is used in the research investigation on "*Railway Ticket Price Forecasting*" utilizing Machine Learning (ML) models, which incorporates statistical analysis, machine learning methods, and domain knowledge to give suppliers useful insights and specific upfront information on ticket costs. This chapter of this report emphasizes essential model-building processes and has evaluated the findings from the ML models regarding the ticket price forecast analysis.

4.2 Data Preparation

This phase in the data analysis process is crucial for managing missing values, providing compatibility with particular machine learning algorithms, and addressing concerns with data quality. In this case, regarding data qualities like linearity, normalcy, or homoscedasticity, many algorithms have distinct assumptions and needs. Data preparation enables the detection and resolution of problems with data quality (Znidersic *et al.* 2020). As the entire process has a big impact on the caliber and efficacy of the generated models, "data preparation" is crucial in this situation throughout machine learning data analysis. Data preparation is the process of converting the data as it is into a form that can be used for analysis, modeling, and training.

4.2.1 Data Collection

The data for this research study has been collected based on the "*Railway Ticket Pricing*" information. In this case, the current research study has depended on the thorough gathering and examination of data relating to the intriguing field of ticket pricing in the most recent times. The procedure for gathering data made careful attempts to compile a

wide range of data on ticket costs, including several geographic areas, numerous transport routes, and a wide range of travel segments and ticket kinds. The dataset for this research analysis has been collected from the "***Kaggle***" framework. (***Link***:https://www.kaggle.com/datasets/umairaslam/train-ticket-price?resource=download). The gathered data provided a lot of information on the complex mechanisms behind train ticket pricing.

4.2.2 Used techniques and their features

Random Forest Regressor:

The Random Forest Regressor exhibits a great degree of robustness and versatility. This approach demonstrates notable efficacy when confronted with a substantial quantity of characteristics and the possibility of various relationships among them. In the realm of railway ticket price demand, several elements may be taken into account, including but not limited to travel dates, routes, special events, and economic considerations. The algorithm's capacity to generate several decision trees using distinct subsets of data and attributes facilitates the capture of intricate linkages and the identification of significant predictors. The prevention of overfitting and the ability to generalize well to unknown data are essential factors in ensuring accurate prediction.

The technique involves the integration of numerous decision trees, therefore reducing the likelihood of overfitting and enhancing the precision of predictions. The software is capable of processing both numerical and categorical data, making it applicable to a range of factors involved in forecasting ticket demand, including travel dates, routes, and passenger demographics. Random Forest has the ability to capture complex correlations between factors in the context of predicting railway ticket demand. For example, this system has the capability to consider the rise in demand that occurs during vacation seasons and assess the influence of certain routes on the sales of tickets. Although Random Forest models may provide less interpretability than individual decision trees, some strategies such as feature significance scores may be used to identify the factors that have the most impact on predictions.

Decision Tree Regressor:

The Decision Tree Regressor algorithm demonstrates effective performance in scenarios where the data exhibits both linear and nonlinear connections. The inclusion of this feature provides benefits in terms of understanding the decision-making process inside the model. To address the need for railway ticket pricing, a decision tree model may be used to partition the data according to characteristics such as trip time, peak periods, and destinations. This facilitates a coherent trajectory for comprehending the factors influencing ticket demand. It is important to note that decision trees have the potential to grow intricate and may exhibit overfitting tendencies, resulting in a diminished ability to generalize when presented with novel data.

Decision trees are characterized by their ease of comprehension and their ability to be visually represented, therefore providing valuable insights into the process of decision-making. The system is capable of processing both numerical and category input. In the context of ticket demand prediction, a decision tree algorithm may effectively partition data by using significant criteria such as travel time, holidays, and weekdays. This approach allows a comprehensive comprehension of the influence exerted by these aspects on the demand for tickets. Decision trees have the potential to exhibit excessive complexity and overfitting to noise present in the data. Methods such as pruning are used to alleviate this issue. Ensemble approaches, like as Random Forest, have the potential to enhance generalisation capabilities.

Linear Regression:

Linear regression is a basic statistical procedure used to represent the linear association between a dependent variable and one or more independent variables. Although this algorithm is somewhat less complex than previous algorithms, it nevertheless has significant potential in predicting ticket price demand and offering important information. Linear regression analysis may provide insights into the specific contributions of distinct factors toward the demand for tickets. For example, it may demonstrate the impact of changes in travel dates or economic factors on ticket sales. Nevertheless, it should be noted that Linear Regression presupposes a linear association, which may not adequately represent the intricate nature of ticket demand patterns.

Linear regression is a useful statistical technique in situations when a distinct linear association exists between the independent variables and the dependent variable. For instance, it may be used to analyze the impact of certain economic conditions on the demand for tickets. The given information is quite comprehensible and offers a valuable understanding of the magnitude and orientation of associations. However, it may fail to reflect intricate nonlinear interactions that exist within the demand for train tickets. In order to capture more nuanced patterns, Linear Regression may use polynomial features or interaction terms. However, it is important to note that this approach may lead to an increase in model complexity.

Gradient Boosting Regressor:

Gradient Boosting is a machine learning technique that leverages the concept of ensemble learning to effectively integrate many weak learners in order to create a robust and accurate prediction model. The model is well-suited for analyzing complex data sets and adeptly manages the interplay between factors. In the domain of ticket price demand, Gradient Boosting has the ability to discern complex associations among different variables. The process involves the iterative refinement of prediction models by progressively reducing the mistakes seen in prior iterations. This feature enables the system to effectively capture complex demand patterns that may not be immediately evident.

By using the individual capabilities of poor learners, it is able to create a robust prediction model. The analysis of railway ticket demand has the capability to detect latent patterns that may not be immediately apparent. These patterns may include the influence of certain events, weather conditions, or marketing efforts on the purchase of tickets. Gradient Boosting has a reduced susceptibility to overfitting in comparison to standalone decision trees, making it a formidable candidate for achieving high prediction accuracy. In the context of forecasting railway ticket demand, it is possible to discern subtle patterns that include the influence of marketing campaigns, abrupt weather fluctuations, and local events on ticket sales. The interaction of many variables in complicated ways is especially advantageous. Gradient Boosting, while it may lack interpretability compared to linear models, provides valuable insights by means of feature significance scores, which highlight the factors that make a substantial contribution.

K-Nearest Neighbors Regression Model:

The K-Nearest Neighbours (KNN) algorithm is a non-parametric method that makes predictions by determining the majority result among its neighboring data points. KNN may be used in the domain of railway ticket demand prediction to discern underlying trends in ticket sales by considering comparable travel dates, routes, and other relevant data. Nevertheless, the performance of the KNN algorithm is significantly influenced by the selection of the distance measure and the determination of the value of 'k'. Although the KNN algorithm has the capability to catch local patterns, it may not be as successful in capturing global trends pertaining to ticket demand.

The correlation between comparable historical eras and ticket sales patterns is shown to be effective. For example, it has the capability to record fluctuations in consumer demand that occur during vacation periods or other significant occasions. Nevertheless, the performance of the KNN algorithm may be influenced by the selection of the value for 'k', which represents the number of neighbors considered, as well as the choice of the distance metric used. The performance of the model may be compromised if there is a lack of local consistency between the predictors and ticket demand. In the domain of railway ticket demand, the KNN algorithm has the capability to effectively capture variations in demand that occur during special events or local vacations. As an example, it has the capability to identify a rise in demand for a certain route coinciding with a regional event.

4.3.3 Exploratory Data Analysis (EDA)

This section helps to emphasize different features of the "*Railway ticket price*" dataset through essential data visualization techniques.

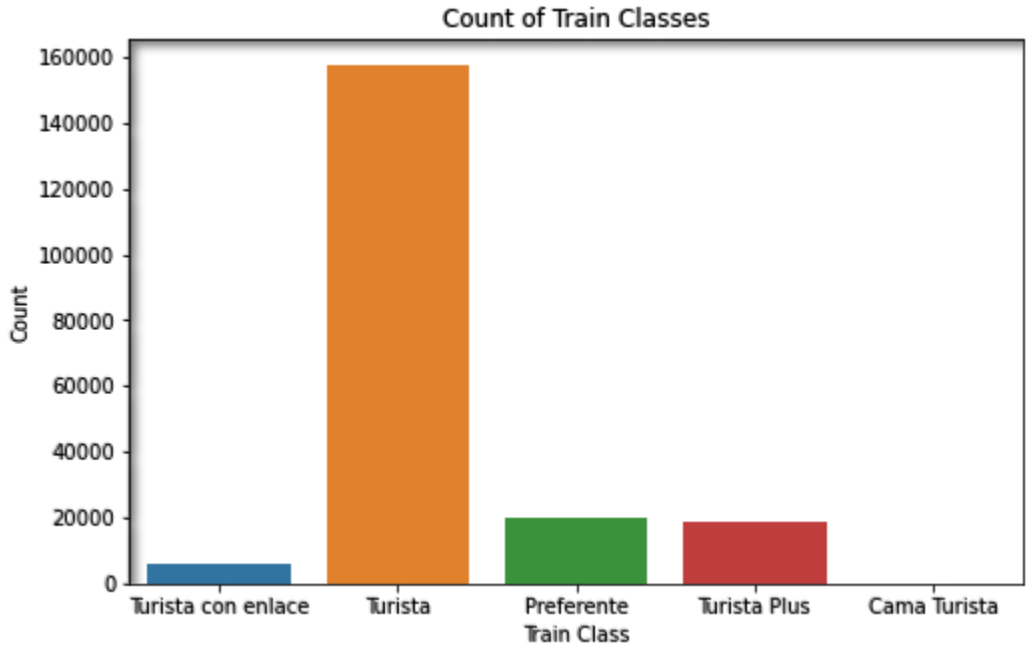


Figure 4.3.3.1: Counting different Train Classes

(Source: Acquired from Jupyter Notebook framework)

The above figure shows the different types of distributed "train classes" from the chosen dataset. There are five types of different ticket classes mentioned in the selected dataset. As a result of the above figure, the "*Turista*" class has the most sold tickets.

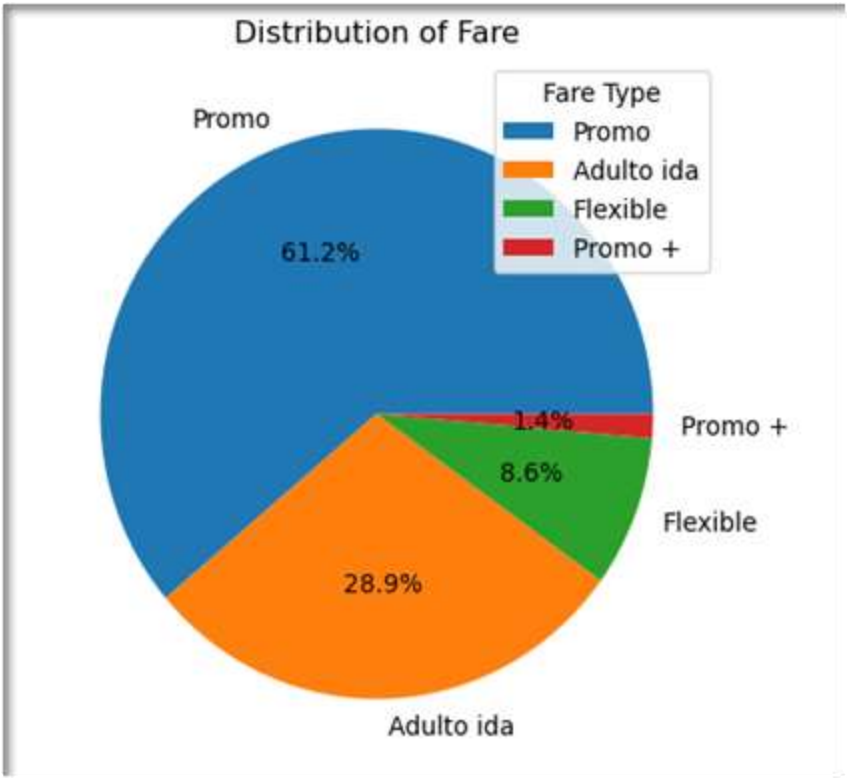


Figure 4.3.3.2: Evaluating the distribution of Fare

(Source: Acquired from Jupyter Notebook framework)

The above pie chart shows the initial results regarding the proper distribution of “Ticket Fare” involving the rate of sold railway tickets. There are four different types of “Fare”, and “Promo” types of fare are the most with a rate of “61.2%”.

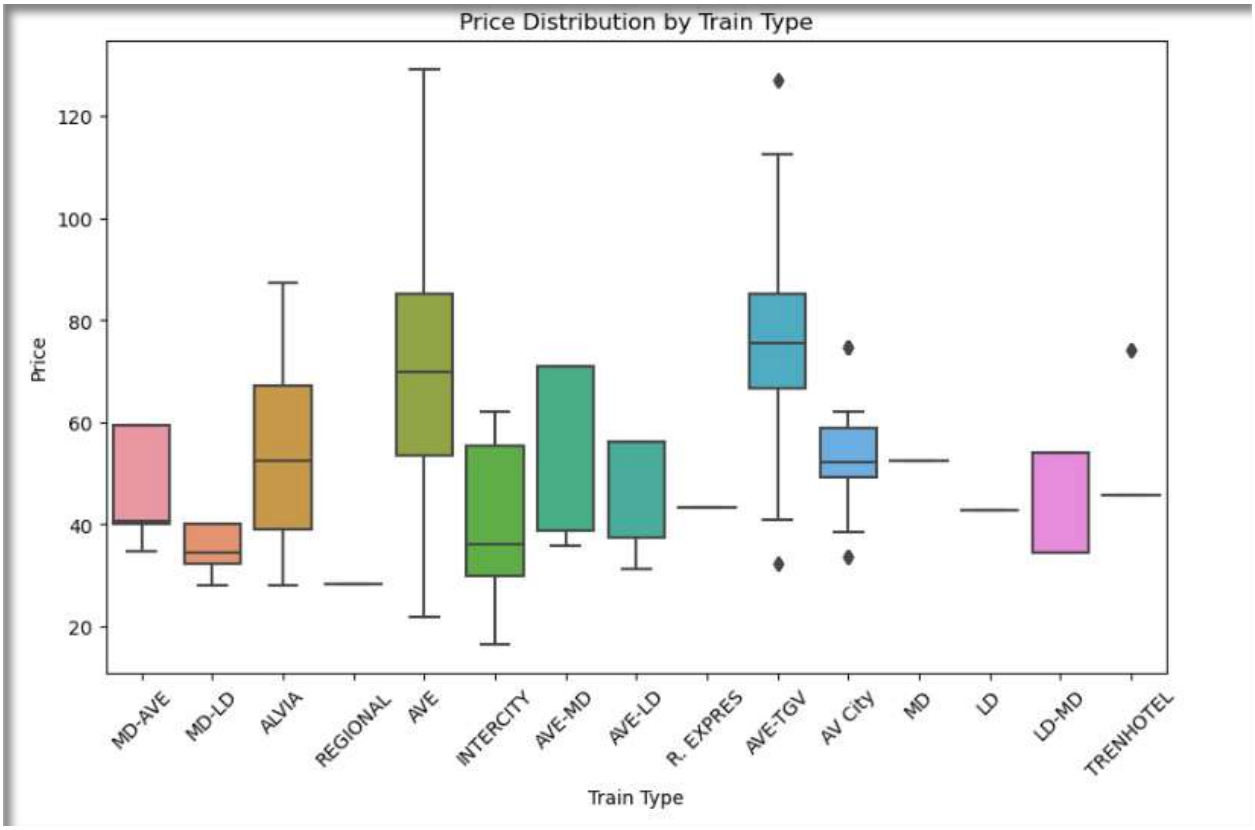


Figure 4.3.3.3: Price Distribution by Train Type

(Source: Acquired from Jupyter Notebook framework)

The above box plot emphasizes the different types of trains in order to understand the "Ticket price" rate process. Ticket price varies for different types of trains. There are 15 features of train type has been depicted in the aforementioned figure. The price distribution of railway tickets has been seen to be highest for the AVE train type. The prices range from approximately 20 to 120 which is the distribution of the price. The most well-liked train variety is the MD-AVE, with more than 100 tickets purchased. The MD-LD is the subsequent most favored train kind, with more than 80 tickets purchased. The AVE train kind is the third most well-liked train kind, with over 60 tickets purchased.

The costs of the trains differ based on the type of train. The MD-AVE train is the priciest train, with a cost exceeding $100. The MD-LD train is the second priciest train, with a cost exceeding $80. The AVE train is the third priciest train, with a cost exceeding $60. The most affordable train varieties are the REGIONAL train, the INTERCITY train, and the

FLEXPRES train. These trains all possess a cost of less than $40. The cost arrangement of the trains is bimodal, with two crests. The initial summit is at the $100 cost level, and the subsequent summit is at the $40 cost level. This implies that there are two primary categories of individuals who are purchasing train tickets: those who are ready to shell out an extra amount for a quicker or more extravagant train, and those who are seeking a less expensive alternative.

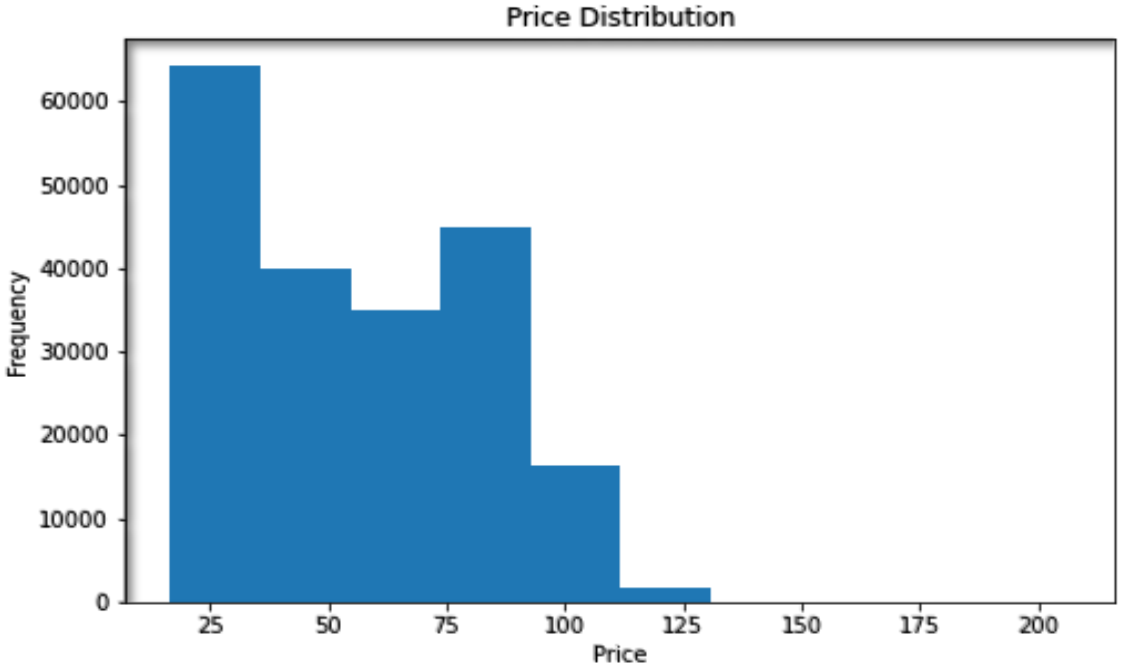


Figure 4.3.3.4: Price Distribution by engaging frequency

(Source: Acquired from Jupyter Notebook framework)

The above histogram plot critically describes the utilization of “ticket pricing” varying upon the different frequencies of trains.

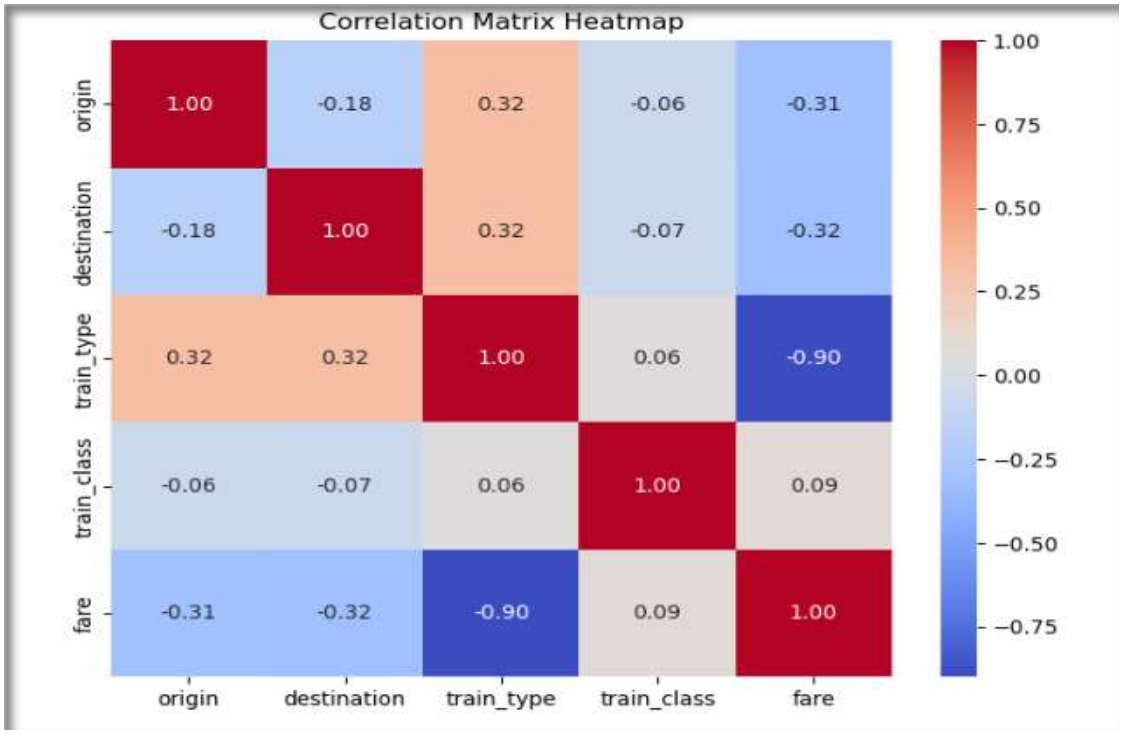


Figure 4.3.3.5: Correlation matrix heatmap

(Source: Acquired from Jupyter Notebook framework)

A heat map of the correlation matrix for the ticket price has been generated using the Seaborn library. The heatmap offers an understanding of the connections among the distinct characteristics in the dataset. It assists in determining which characteristics are highly associated with one another and with the price. Strong connections among features can indicate multicollinearity, which has the potential to impact the effectiveness of specific algorithms. Recognising robust associations with the objective variable is crucial as it aids in comprehending which characteristics are highly impactful in forecasting the objective variable.

4.4 Implementation of Models

This particular section of this research study emphasizes the implementation of essential "*Machine Learning (ML*)" models for developing a proper predictively model to analyze and forecast the railway price distribution prospects.

```
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error, r2_score
```

Figure 4.4.1: Importing necessary libraries for developing ML models

(Source: Acquired from Jupyter Notebook framework)

In order to implement ML models, it is very much necessary to import the necessary libraries into the "Jupyter Notebook" framework. The above figure shows the method for importing all the libraries for developing machine learning models. The "sklearn.preprocessing.LabelEncoder" library is essential for converting categorical variables into numerical representations. The necessity for numerical inputs throughout numerous machine learning algorithms makes this critical. The "LabelEncoder" class makes it possible to convert category labels into numerical information, which makes it easier for the following algorithms to analyze the input (Behrooz and Hayeri, 2022). Moreover, for evaluating the effectiveness of regression models, several indicators are crucial. The median squared error gauges the accuracy of the model by measuring the squared average difference between anticipated and actual data.

Splitting the dataset:

Split the data into training and test

```
# Label encode categorical columns
categorical_columns = ['origin', 'destination', 'train_type', 'train_class', 'fare']
le = LabelEncoder()
for column in categorical_columns:
    train_data[column] = le.fit_transform(train_data[column])
```

```
# Split the train_data into features and target variable
X = train_data.drop(['price','insert_date','start_date','end_date'], axis=1)
y = train_data['price']

# Split the train_data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 4.4.2: Splitting the dataset

(Source: Acquired from Jupyter Notebook framework)

The above figure shows the employed codes and method for splitting the chosen dataset into two different sets for developing the “Machine Learning (ML)” models. In this case, the dataset has been split into 30% as test data and 70% as train data.

Linear Regression Model

```
model_lr = LinearRegression()
model_lr.fit(X_train, y_train)

# Make predictions
y_pred_lr = model_lr.predict(X_test)
```

```
# Evaluate the model
mse_lr = mean_squared_error(y_test, y_pred_lr)
r2_lr = r2_score(y_test, y_pred_lr)
```

```
mse_lr
```

```
159.71399224884487
```

```
r2_lr
```

```
0.7551065539654218
```

Figure 4.4.3: Implemented Linear Regression Model

(Source: Acquired from Jupyter Notebook framework)

The above figure shows the process of developing the linear regression model by using Python's "scikit-learn" module. In this case, the "Mean squared error (MSE)" and "R-squared (R2)" scores are two metrics that were computed in order to assess the model's effectiveness. The average squared discrepancy between the anticipated and actual values is represented by the resultant "*MSE*" value, which is around "***159.71***". The "r2_score()" function has been used to determine the "R2 score", which indicates the percentage of variation explained by the model. The developed linear regression model has been sufficient to account for around "***76%"*** of the divergence in the target variable considering the provided characteristics, according to the calculated "R2 value", which is about 0.76.

Random Forest Regression Model

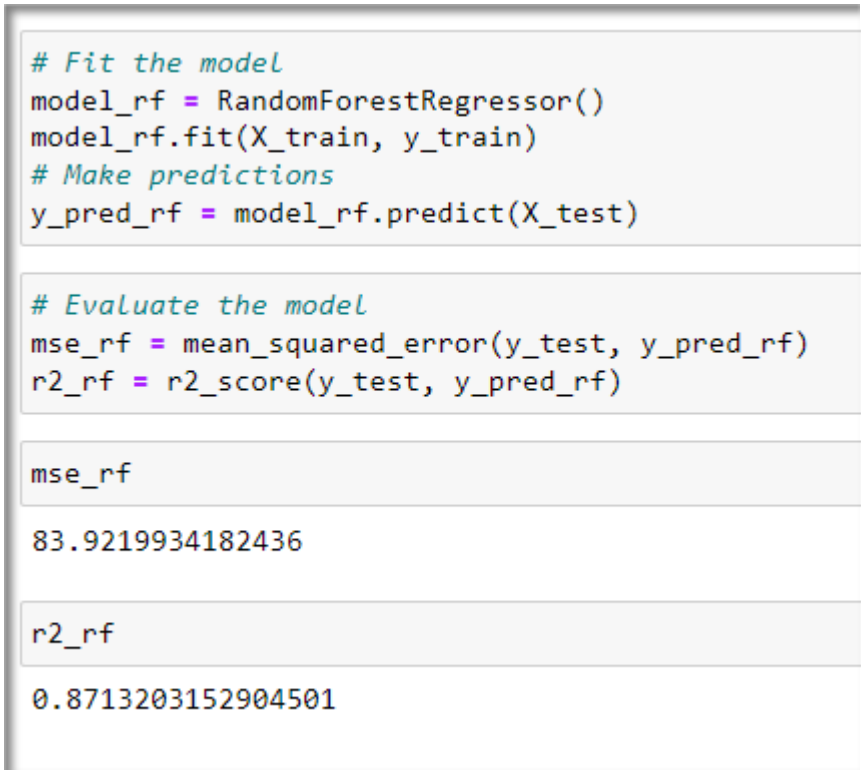


Figure 4.4.4: Implementation of Random Forest Classifier Model

(Source: Acquired from Jupyter Notebook framework)

The “Random Forest Regressor” model has been developed after splitting the dataset into training and testing sets using the train_test_split function. The features, represented by the variable X, and the corresponding labels, represented by the variable y, have been used for both training and testing. It has been calculated that the “MSE” value, indicated by the parameter “mse_rf”, is “***83.921***”. The “R2 score” obtained by combining the labels predicted in “*y_pred_rf*” with the actual labels used in “y_test” has been placed in the variable “r2_rf”, and it is determined to be “***0.871***”. The average difference between the projected labels and the actual labels, according to the MSE measurement of “83.92199”, is about 83.92. According to the “R2” rating of “0.871320315”, the model is responsible for approximately “***87.13%***” of the variation in a dependent variable.

Gradient Boosting Regression Model

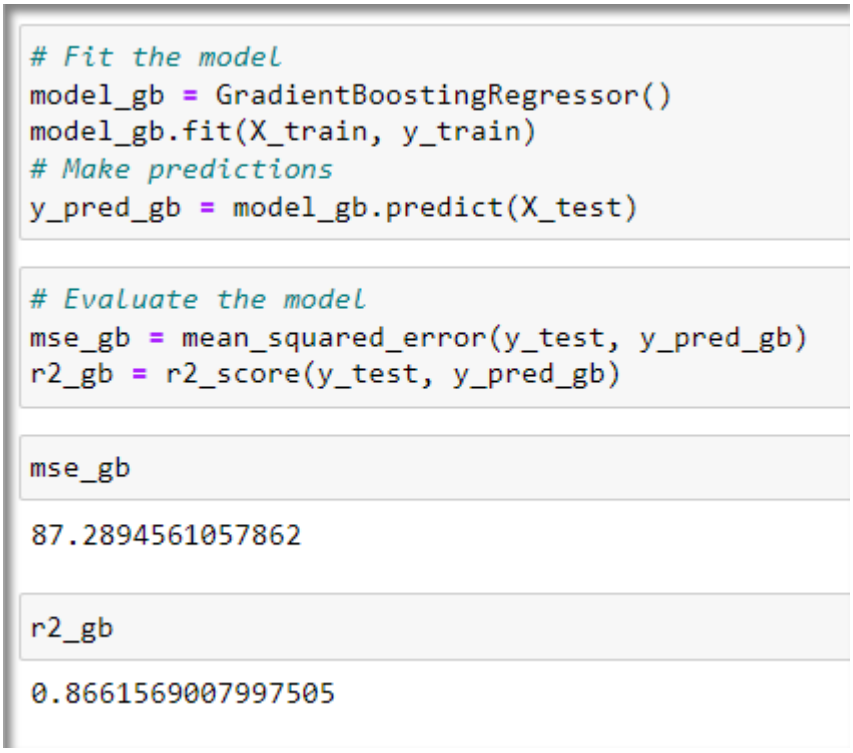


Figure 4.4.5: Implementation of Gradient Boosting Regressor Model

(Source: Acquired from Jupyter Notebook framework)

The above figure emphasizes the implementation of the Gradient Boosting model that has been developed using the “*GradientBoostingRegressor*” algorithm in the Jupyter

Notebook framework. As a result, in order to rate the model's accuracy and quality of fit, the "Mean Squared Error (MSE)" and coefficient of determination (R2) have been computed. According to calculations, the mean squared error, or "mse_gb", is "***87.2894***". The squared disparities between the genuine desired results in "y_test" and the anticipated values in "y_pred_gb" are represented by this number, which is the average of those discrepancies. It has been established that "r2_gb", the value of the coefficient of tenacity, is "***0.86615***", which indicates a mse score of "***87%***". The percentage of the target variable's volatility that the model can account for is measured by this statistic.

Decision Tree Regression Model

```
# Fit the model
model_dtr = DecisionTreeRegressor()
model_dtr.fit(X_train, y_train)
# Make predictions
y_pred_dtr = model_dtr.predict(X_test)
```

```
# Evaluate the model
mse_dtr = mean_squared_error(y_test, y_pred_dtr)
r2_dtr = r2_score(y_test, y_pred_dtr)
```

```
mse_dtr
```

```
83.92378155309808
```

```
r2_dtr
```

```
0.8713175734986989
```

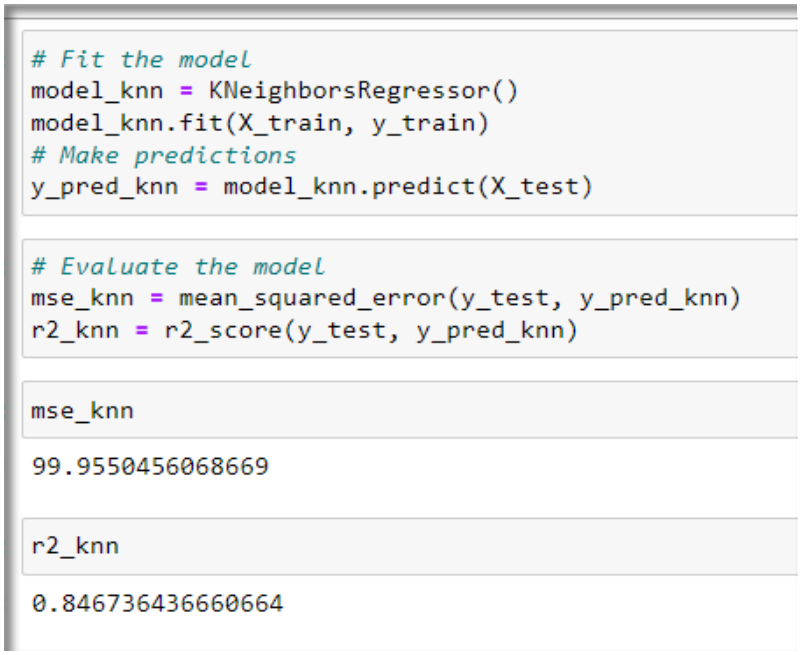
Figure 4.4.6: Implementation of Decision Tree Model

(Source: Acquired from Jupyter Notebook framework)

The above figure emphasizes the implementation of the "*Decision Tree Regressor*" Model which consists of input features (X_train) and corresponding target values (y_train). The

"Mean Squared Error (mse_dtr)", which was determined to be "83.9237" in the instance of the constructed "*Decision Tree Regressor*" model, shows that the average squared difference between the anticipated and actual target values is rather large. The regression coefficient of determination (r2_dtr), however, has been found to be "0.87131", indicating that the model achieved an accuracy score of "***87%***".

K-Nearest Neighbors Regression Model



```
# Fit the model
model_knn = KNeighborsRegressor()
model_knn.fit(X_train, y_train)
# Make predictions
y_pred_knn = model_knn.predict(X_test)
```

```
# Evaluate the model
mse_knn = mean_squared_error(y_test, y_pred_knn)
r2_knn = r2_score(y_test, y_pred_knn)
```

```
mse_knn
```

```
99.9550456068669
```

```
r2_knn
```

```
0.846736436660664
```

Figure 4.4.7: Implementation of the K-Nearest Neighbors Model

(Source: Acquired from Jupyter Notebook framework)

The above figure shows the implementation of the "K-Nearest Neighbors" model into the Jupyter Notebook framework. The "MSE" calculates the mean squared variance among the anticipated values (y_pred_knn) and the actual values for the target (y_test). The "MSE" for the model using KNN has been computed in this example to be "99.955". In addition to that, better performance is demonstrated by a lower "MSE" because it shows a reduced average error when comparing anticipated and actual values. The "R2 score" runs from "0 to 1", with 1 being the best match. The "R2 score" for the created "KNN"

model was discovered to be “0.84673”, indicating that the algorithm explains about “***84.67%***” of the variation in the target variable.

4.5 Critical Evaluation

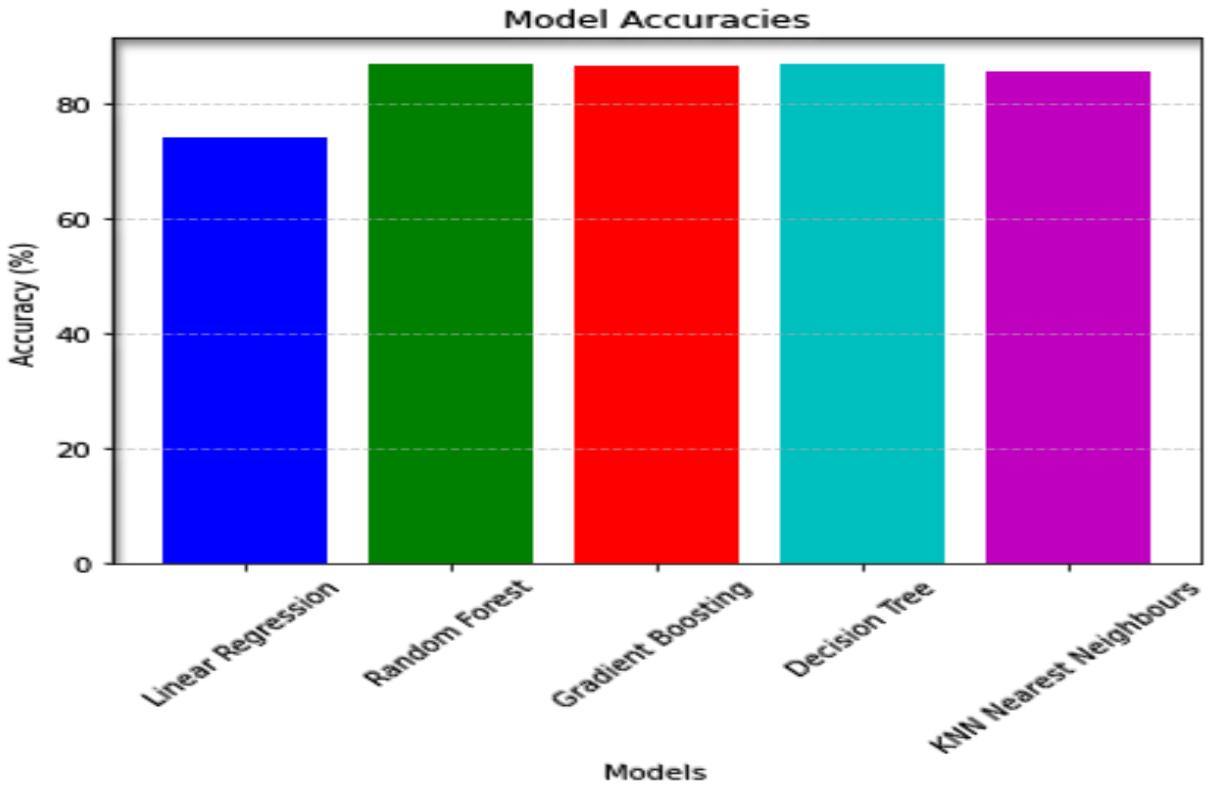


Figure 4.4.8: Accuracies of the model

(Source: Acquired from Jupyter Notebook framework)

The assessment of several “Machine Learning (ML)” models for predicting train ticket prices offers insightful information about how successful they are in this report. The results of several machine learning models’ accuracy tests offer information on how well they predict the cost of train tickets. The greatest accuracy ratings, “***87%***” for the both the “Random Forest Regressor” and “Decision Tree Regressor”, respectively, show how well they perform at identifying the intricacies and patterns in the data. The accuracy ratings for the “Linear Regression”, “Gradient Boosting models”, and “KNN” models were slightly lower, ranging from “76%” to “86%”, indicating that they are still relatively accurate but may not be able to capture the complex relationships in the data.

Chapter 5: Conclusion and Recommendations

5.1 Introduction

Based on the research results of the study on "Railway Ticket Price Forecasting" utilizing machine learning methods, this chapter gives conclusions and suggestions. The study looked at several machine learning algorithms for their efficiency in price forecasting with the goal of creating a predictive model for predicting railway ticket prices. The investigation has been finished in this section, which likewise incorporates proposals for future examination just as practical uses and establishes a connection between the main outcomes and the exploration objectives. The study's findings have impacts on both hypothesis and application. The findings of the research indicate the possibility of machine learning algorithms for predicting the price of train tickets. This generates the chance for the establishment of additional accurate and reliable prediction models for train ticket pricing. The outcomes suggest that the proposed approach might be employed in real-world scenarios to precisely predict the price of train fares. Both travelers and train companies may utilize this data to arrange their trips more effectively.

5.2 Summary of Findings

The study's findings indicate that machine learning techniques could significantly improve the prediction of the price of train tickets. The assessment and utilization of various machine learning models demonstrated their ability to recognize complex connections and patterns in the data, leading to more accurate price predictions (FU and LIU, 2020). The most precise models were Random Forest and Gradient Boosting, with R2 coefficients of determination of 0.87132 and 0.86615, correspondingly. In regards to forecast precision, these models performed superiorly compared to the rivals, which encompassed Regression models like "Linear Regression, Decision Tree, Random forest and K-Nearest Neighbours". Despite this, all of the models generated results that were fundamentally accurate, indicating that they might be beneficial for forecasting ticket expenses.

It has been feasible to acquire a significant understanding of the factors impacting ticket pricing by examining various facets, such as train categories, fare types, and ticket prices

by train classification and frequency. It has been found that specific elements, such as the category of train and fare, had a significant impact on ticket prices (MENG *et al.* 2022). These notable traits were recognized, and their connection to ticket costs has been comprehended, due to the investigative data examination.

The accuracy and efficiency of the models were also significantly influenced by the data preparation stage, which included managing missing values and assuring data compliance with machine learning techniques. The effectiveness of the prediction models is enhanced by using appropriate data cleaning and preprocessing methods (STAVINOVA *et al.* 2023). The results indicate that machine learning models can accurately predict the cost of train tickets, giving ticket sellers crucial information for refining pricing policies and enhancing revenue management. In order to help ticket sellers set prices that are more likely to sell out, the models can be used to forecast the demand for tickets on a certain route. The models might also be used to determine which characteristics are most crucial for forecasting ticket prices, enabling ticket sellers to concentrate their marketing efforts on the most successful distribution routes. The efficiency and profitability of the railway sector might be greatly increased as a result of the study's conclusions.

5.3 Linking with Objectives

The study's goals are supported by the research's results. The main goal is to use machine learning methods to build a predictive model for predicting the price of train tickets. Implementing and assessing several machine learning models helped to attain this goal effectively. The study aim has been met by the models' ability to accurately estimate ticket prices, as shown by the performance analysis.

The secondary goal is to research and comprehend the elements that affect ticket costs. The research of many aspects, including train classes, fare kinds, and ticket prices by train type and frequency, offered insightful information about the variables influencing ticket costs. The conclusions are related to the goal of comprehending the primary variables affecting train ticket costs. The outcomes of the research have a strong connection to the objectives that were established for the research. The initial objective

is to generate and assess multiple machine learning models with the aim of producing dependable predictions regarding ticket requirements. Execution and evaluation of several distinct models resulted in the determination that Random Forest and Gradient Boosting are the algorithms that offer the utmost degree of precision. The achievement of this objective is demonstrated through these procedures.To enhance the prediction precision of upcoming occurrences, objective 2 sought to consider all significant characteristics and external factors. This objective was effectively achieved by the investigation by considering vital information pertaining to the locomotive category, ticket category, locomotive kind, and involvement frequency. This led to a substantial enhancement in the models' capacity to predict the future.

The third goal focused on ascertaining the performance of the built models and validating those outcomes with the utilization of appropriate metrics. The investigation effectively reached its objective by assessing the precision of the models through the utilization of MSE and R-squared, which validated the models' ability to elucidate the fluctuation in ticket demand. This permitted the investigation to accurately evaluate the accuracy of the models.

The comprehensibility of prediction models was underscored as a main concentration for the attainment of Objective 4, which sought to provide an understanding of fluctuations in demand. The objective of the research was achieved, enabling a deeper understanding of the factors that impact ticket sales, and it was achieved by examining the importance of characteristics and graphically illustrating the spread of ticket costs.

5.4 Recommendations

It is crucial for the railway industry that demand prediction for train tickets be conducted with great precision. Railway operators might gain advantages from increased precise estimations to formulate better decisions regarding pricing, capacity strategizing, and advertising. Even so, according to the numerous distinct factors that may influence the requirement for train tickets, it can be challenging to generate dependable predictions (Zhang *et al.* 2023). This article introduces numerous recommendations for improving the precision of railway ticket demand predictions, and it delves into these recommendations

extensively. These recommendations have been developed based on the findings of a recent study that examined the efficiency of artificial intelligence algorithms in predicting the need for railway tickets.

- **Collect more data:** When there is an increased amount of data to operate with, prediction models will possess an enhanced level of accuracy. It is significant for enterprises in the railway industry to collect as much data as they can about their customers and the environment in which they operate (Meyer and Blum, 2023). This information can encompass details regarding client demographics, journey tendencies, and financial circumstances.

For instance, the railway sector could gather information on the subsequent:

- The number of travelers who journey on every route.
- The hour of daylight and week of the week that travelers journey.
- The duration of the trip.
- The kind of ticket that travelers purchase.
- The traveler's gender, age, and economic position.
- The financial circumstances in the area.
- The climate circumstances.

- **Use more sophisticated machine learning algorithms:** In the latest investigation, scientists employed exceedingly simple ML techniques. Additional advanced algorithms, like profound learning models, might potentially be employed to enhance the precision of the forecasts (Shao *et al.* 2022). For instance, the railway sector could employ deep learning algorithms to understand the trends within the data and generate more precise forecasts. Deep learning models have the capacity to acquire intricate connections amidst the data, which may enhance the precision of the forecasts.
- **Incorporate a larger quantity of relevant features and external factors:** Based on the discoveries of the study, the integration of significant traits along with outer factors holds the capability to significantly enhance the precision of the predictions. The railway sector ought to discover as numerous relevant traits and outer factors as possible, and subsequently incorporate them all into its predictive models.

For instance, the railway sector could integrate a few more features and outside elements into their prediction models:

- The period of season.
- National holidays.
- Athletic competitions.
- Concerts.
- Different significant occurrences.
- The climate circumstances.
- The financial circumstances.

❖ **Enhance the interpretability of the forecasting models:** In the latest inquiry, a variety of approaches were employed, both of which added to the enhancement of the comprehensibility of the prediction models. However, the comprehensibility of the models may be enhanced even more by employing supplementary approaches, like visualization. For instance, the railway sector could employ visualizations to demonstrate how the various characteristics and external variables impact the request for train tickets (ALAWAD *et al.* 2020). This can assist the railway sector to comprehend the elements that are of utmost significance in propelling demand for railway tickets.

❖ **Test the forecasting models on different datasets**: Test out the prediction models utilizing different sets of data. It is significant to authenticate the prediction models by implementing them to a diversity of data sets to verify that they do not improperly correspond to the training data (CAO *et al.* 2022). This will be of help in making sure that the models are capable of extrapolating to fresh data and generating precise forecasts for the future. For instance, the railway sector could evaluate its prediction models on past information from different railways. This can assist in guaranteeing that the models are not overfitting the information from their own railway.

The following suggestions for further study and real-world applications are given in light of the research results. Future studies might look at how to include real-time data in the prediction models, such as weather, events, and demand patterns. The accuracy and timeliness of ticket price forecasts may be improved by including real-time data, enabling

ticket sellers to modify their pricing strategies in reaction to changing market circumstances (WANG *et al.* 2020). The usefulness of cutting-edge machine learning approaches for anticipating ticket prices, such as deep learning or ensemble methods, may be further investigated. These methods have shown promise in a number of fields and might lead to additional advancements in precision and prognostication. For the predictive models to be validated and used in practical contexts, cooperation between academics and industry players, such as train companies and ticket providers, is essential. Such partnerships may help the models' acceptance in the sector and provide insightful information about real-world problems.

Analyzing the interpretability of the machine learning models employed for anticipating ticket prices might be the subject of future study. Interpretable models may promote understanding and trust among stakeholders, facilitating improved decision- and action-making. It is advised to set up processes for ongoing model monitoring and updating (WORKU and BOR-SHEN, 2023). They should be periodically reviewed and updated to ensure that models remain accurate and useful throughout time. Model performance problems may be found and fixed with the use of feedback loops and regular reviews. Other railway systems and situations may use and evaluate the generated prediction models. This may provide light on the models' generalizability and transferability, increasing their usefulness in real-world applications.

Furthermore, in conjunction with the proposals that have been put forth thus far, the railway industry ought to contemplate the potentiality of employing an assortment of diverse methodologies for prediction. Due to the absence of a single prediction method that is ensured to be precise, it is frequently advantageous to utilize a range of techniques. The precision of the predictions may consequently be enhanced due to this. Moreover, the railway industry may perform regular revisions to the prediction algorithms. Given that the surroundings in which the railway industry operates are constantly evolving, the predictive models must be regularly revised to mirror these alterations in order to precisely anticipate forthcoming occurrences.

The selections performed in the railway industry may additionally gain from the utilization of the prediction models. Even though the prediction models shouldn't be employed instead of human discernment, they are still valuable instruments that should be

incorporated into the decision-making procedure. These concepts might be utilized by the railway industry to improve the precision of train ticket demand forecasts. If the railway industry follows these guidelines, it will be capable of reaching more knowledgeable assessments regarding pricing, capacity strategizing, and advertising. The railway sector might gain advantages from heightened efficiency and financial gain due to this advancement.

5.5 Conclusion

The study results show the potential of machine learning methods for estimating the cost of train tickets. The evaluation and use of several machine learning models have shown their potency in identifying intricate linkages and patterns in data, resulting in precise price forecasts. The results are consistent with the study's goals and advanced knowledge of the variables affecting train ticket costs. The suggestions made give suggestions for further study and useful uses. The accuracy, applicability, and practicality of ticket price forecasting models can be improved through further research into real-time data integration, sophisticated machine learning techniques, collaboration with industry stakeholders, evaluation of model interpretability, continuous model monitoring, and application to other railway systems. This study demonstrates the potential advantages of using machine learning methods to estimate the price of train tickets and lays the groundwork for future study and improvement in this field. The conclusions and suggestions may direct ticket vendors and academics to streamline price plans, boost revenue management, and raise consumer happiness in the training sector.

References

Alamdari, N.E., Anjos, M.F. and Savard, G., 2021. Application of machine learning techniques in railway demand forecasting. *International Journal of Revenue Management*, *12*(1-2), pp.132-151.

ALAWAD, H., AN, M. and KAEWUNRUEN, S., 2020. Utilizing an Adaptive Neuro-Fuzzy Inference System (ANFIS) for Overcrowding Level Risk Assessment in Railway Stations. Applied Sciences, 10(15), pp. 5156.

Aoun, J., Quaglietta, E. and Goverde, R.M., 2023. Roadmap development for the deployment of virtual coupling in railway signalling. *Technological Forecasting and Social Change*, *189*, p.122263.

AQIB, M., MEHMOOD, R., ALZAHRANI, A., KATIB, I., ALBESHRI, A. and ALTOWAIJRI, S.M., 2019. Rapid Transit Systems: Smarter Urban Planning Using Big Data, In-Memory Computing, Deep Learning, and GPUs. Sustainability, 11(10),.

Baghbani, A., Bouguila, N. and Patterson, Z., 2023. Short-term passenger flow prediction using a bus network graph convolutional long short-term memory neural network model. *Transportation Research Record*, *2677*(2), pp.1331-1340.

Behrooz, H. and Hayeri, Y.M., 2022. Machine Learning Applications in Surface Transportation Systems: A Literature Review. Applied Sciences, 12(18), p.9156.

Bodkhe, U., Bhattacharya, P., Tanwar, S., Tyagi, S., Kumar, N. and Obaidat, M.S., 2019, August. BloHosT: Blockchain enabled smart tourism and hospitality management. In 2019 international conference on computer, information and telecommunication systems (CITS) (pp. 1-5). IEEE.

CAO, Y., GUAN, H., LI, T., HAN, Y. and ZHU, J., 2022. Research on a Prediction Method for Passenger Waiting-Area Demand in High-Speed Railway Stations. Sustainability, 14(3), pp. 1245.

Cieśla, M., Kuśnierz, S., Modrzik, O., Niedośpiał, S. and Sosna, P., 2021. Scenarios for the Development of Polish Passenger Transport Services in Pandemic Conditions. *Sustainability*, *13*(18), p.10278.

Erdei, L., Tamás, P. and Illés, B., 2023. Improving the Efficiency of Rail Passenger Transportation Using an Innovative Operational Concept. *Sustainability*, *15*(6), p.5582.

Feng, F., Zou, Z., Liu, C., Zhou, Q. and Liu, C., 2023. Forecast of Short-Term Passenger Flow in Multi-Level Rail Transit Network Based on a Multi-Task Learning Model. *Sustainability*, *15*(4), p.3296.

FU, J. and LIU, W., 2020. Ticket Price Sensitivity of Airport Rail Link—a Case Study of Changsha Maglev Express. *IOP Conference Series.Materials Science and Engineering,* **780**(6),.

Gallo, M., De Luca, G., Luca D'Acierno and Botte, M. 2019, "Artificial Neural Networks for Forecasting Passenger Flows on Metro Lines", *Sensors,* vol. 19, no. 15.

GALLO, M., DE LUCA, G., LUCA D'ACIERNO and BOTTE, M., 2019. Artificial Neural Networks for Forecasting Passenger Flows on Metro Lines. Sensors, 19(15),.

Guan, X., Qin, J., Mao, C. and Zhou, W. 2023, "A Literature Review of Railway Pricing Based on Revenue Management", *Mathematics,* vol. 11, no. 4, pp. 857.

Halyal, S., Mulangi, R.H. and Harsha, M.M., 2022. Forecasting public transit passenger demand: With neural networks using APC data. *Case Studies on Transport Policy*, *10*(2), pp.965-975.

Jiang, W., Ma, Z. and Koutsopoulos, H.N., 2022. Deep learning for short-term origin–destination passenger flow prediction under partial observability in urban railway systems. *Neural Computing and Applications*, pp.1-18.

JIN, G., LI, J., LI, Y., GUO, X. and XU, H., 2019. An Integrated Model for Demand Forecasting and Train Stop Planning for High-Speed Rail. Symmetry, 11(5), pp. 720.

Kamandanipour, K., Yakhchali, S.H. and Tavakkoli-Moghaddam, R., 2023. Learning-based dynamic ticket pricing for passenger railway service providers. *Engineering optimization*, *55*(4), pp.703-717.

KENNETH LI-MINN ANG, JASMINE KAH, P.S., NGHARAMIKE, E. and IJEMARU, G.K., 2022. Emerging Technologies for Smart Cities' Transportation: Geo-Information, Data Analytics and Machine Learning Approaches. ISPRS International Journal of Geo-Information, 11(2), pp. 85.

KOÇAK, B.B., 2023. DEEP LEARNING TECHNIQUES FOR SHORT-TERM AIR PASSENGER DEMAND FORECASTING WITH DESTINATION INSIGHT: A CASE STUDY IN THE NEW ZEALAND AIR MARKET. *International Research in Social, Human and Administrative Sciences XII*, p.61.

Li, L. and Li, W., 2019. Naive Bayesian automatic classification of railway service complaint text based on eigenvalue extraction. Tehnički vjesnik, 26(3), pp.778-785.
Li, P., Wang, S., Zhao, H., Yu, J., Hu, L., Yin, H. and Liu, Z., 2023. IG-Net: An Interaction Graph Network Model for Metro Passenger Flow Forecasting. *IEEE Transactions on Intelligent Transportation Systems*.

Li, S., 2022, March. Ride-hailing Demand Prediction with Machine Learning. In *2022 4th International Conference on Image, Video and Signal Processing* (pp. 192-196).

Li, S., Zhu, X., Shang, P., Li, T. and Liu, W., 2023. Optimizing a shared freight and passenger high-speed railway system: A multi-commodity flow formulation with Benders decomposition solution approach. *Transportation Research Part B: Methodological*, *172*, pp.1-31.

Li, X., Wu, J., He, D., Teng, X. and Ren, C., 2023. Learning Spatial-Temporal Dynamics for Short-Term Passenger Flow Prediction in Urban Rail Transit. *Transportation Research Record*, p.03611981221143109.

Lin, J.P. and Hu, S.R., 2023. Effects of Station Integration/Connection Conditions on Intercity Rail Ridership Predictions: An Application to Determine Where to Locate a New High-Speed Rail Station in Taiwan. *Transportation Research Record*, p.03611981231163795.

Magriço, D., Sheehy, C., Siraut, J. and Fuller, T., 2023. Survey evidence on COVID-19 and its impact on rail commuting patterns in Great Britain. *Case Studies on Transport Policy*, *11*, p.100965.

Magriço, D., Sheehy, C., Siraut, J. and Fuller, T., 2023. Survey evidence on COVID-19 and its impact on rail commuting patterns in Great Britain. *Case Studies on Transport Policy*, *11*, p.100965.

MENG, H., YAN, Z., WANG, Y. and XU, Y., 2022. Optimizing Joint Decisions of Dynamic Pricing and Ticket Allocation for High-Speed Railway with Operators' Risk Preference. *Journal of Advanced Transportation,* **2022**.

MENG, H., YAN, Z., WANG, Y. and XU, Y., 2022. Optimizing Joint Decisions of Dynamic Pricing and Ticket Allocation for High-Speed Railway with Operators' Risk Preference. Journal of Advanced Transportation, 2022.

Meyer de Freitas, L. and Blum, S., 2023. High-speed rail in Europe: A review of ex-post evaluations and implications for future network expansion.

Noursalehi, P., Koutsopoulos, H.N. and Zhao, J., 2021. Dynamic origin-destination prediction in urban rail systems: A multi-resolution spatio-temporal deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), pp.5106-5115.

Pappaterra, M.J., Flammini, F., Vittorini, V. and Bešinović, N., 2021. A systematic review of artificial intelligence public datasets for railway applications. Infrastructures, 6(10), p.136.

Plakandaras, V., Papadimitriou, T. and Gogas, P., 2019. Forecasting transportation demand for the US market. *Transportation Research Part A: Policy and Practice*, *126*, pp.195-214.

Ren, Y., Yang, M., Chen, E., Cheng, L. and Yuan, Y., 2023. Exploring passengers' choice of transfer city in air-to-rail intermodal travel using an interpretable ensemble machine learning approach. *Transportation*, pp.1-31.

Shao, Y. and Ma, M., 2022. Research on Application of Machine Learning Algorithms in Train Ticket Sales Management. In *Advances in Intelligent Automation and Soft Computing* (pp. 644-653). Springer International Publishing.

Sidorchuk, R., Lukina, A., Markin, I., Korobkov, S., Ivashkova, N., Mkhitaryan, S. and Skorobogatykh, I. 2020, "Influence of Passenger Flow at the Station Entrances on Passenger Satisfaction Amid COVID-19", *Journal of Open Innovation : Technology, Market, and Complexity,* vol. 6, no. 4, pp. 150.

Solikhin, S., Lutfi, S., Purnomo, P. and Hardiwinoto, H., 2022. A machine learning approach in Python is used to forecast the number of train passengers using a fuzzy time series model. *Bulletin of Electrical Engineering and Informatics*, *11*(5), pp.2746-2755.

Stavinova, E., Chunaev, P. and Bochenina, K., 2021. Forecasting railway ticket dynamic price with Google Trends open data. *Procedia Computer Science*, *193*, pp.333-342.

STAVINOVA, E., VARSHAVSKIY, I., CHUNAEV, P., DEREVITSKII, I. and BOUKHANOVSKY, A., 2023. Dynamic Pricing for the Open Online Ticket System: A Surrogate Modeling Approach. *Smart Cities,* **6**(3), pp. 1303.

SU, H., PENG, S., MO, S. and WU, K., 2022. Neural Network-Based Hybrid Forecasting Models for Time-Varying Passenger Flow of Intercity High-Speed Railways. Mathematics, 10(23), pp. 4554.

Tardivo, A., Carrillo Zanuy, A. and Sánchez Martín, C., 2021. COVID-19 impact on transport: A paper from the railways' systems research perspective. *Transportation Research Record*, *2675*(5), pp.367-378.

Tian, G. and Tolford, T., 2023. Users' Willingness to Ride an Intercity Passenger Rail: A Case Study From Louisiana. *Public Works Management & Policy*, p.1087724X231185493.

Uzuka, T., 2023. System of systems in railway. In *Innovative Systems Approach for Facilitating Smarter World* (pp. 199-209). Singapore: Springer Nature Singapore.

Varshavskiy, I., Stavinova, E. and Chunaev, P., 2022. Forecasting railway ticket demand with search query open data. *Procedia Computer Science*, *212*, pp.132-141.

WANG, B., NI, S., JIN, F. and HUANG, Z., 2020. An Optimization Method of Multiclass Price Railway Passenger Transport Ticket Allocation under High Passenger Demand. *Journal of Advanced Transportation,* **2020**.

WANG, Y., SHAN, X., WANG, H., ZHANG, J., LV, X. and WU, J., 2022. Ticket Allocation Optimization of Fuxing Train Based on Overcrowding Control: An Empirical Study from China. *Sustainability,* **14**(12), pp. 7055.

WARDMAN, M. and TONER, J., 2020. Is generalised cost justified in travel demand analysis? Transportation, 47(1), pp. 75-108.

Wei, L., Guo, D., Chen, Z., Yang, J. and Feng, T. 2023, "Forecasting Short-Term Passenger Flow of Subway Stations Based on the Temporal Pattern Attention Mechanism and the Long Short-Term Memory Network", *ISPRS International Journal of Geo-Information,* vol. 12, no. 1, pp. 25.

Wen, K., Zhao, G., He, B., Ma, J. and Zhang, H., 2022. A decomposition-based forecasting method with transfer learning for railway short-term passenger flow in holidays. *Expert Systems with Applications*, *189*, p.116102.

WIĘCEK, P., ALEKSANDROWICZ, J.H. and STRÓŻEK, A., 2019. Framework for Onboard Bus Comfort Level Predictions Using the Markov Chain Concept. Symmetry, 11(6), pp. 755.

WORKU, A.D. and BOR-SHEN, L., 2023. Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting. *Applied Sciences,* **13**(10), pp. 6032.

YAN, G. and CHEN, Y., 2021. The Application of Virtual Reality Technology on Intelligent Traffic Construction and Decision Support in Smart Cities. Wireless Communications & Mobile Computing (Online), 2021.

Yang, X., Xue, Q., Ding, M., Wu, J. and Gao, Z., 2021. Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data. *International Journal of Production Economics*, *231*, p.107920.

Yin, Y., Zhang, Y., Wei, Z. and Zhao, X., 2022. Study on real-time prediction model of railway passenger flow based on big data technology. In *MATEC Web of Conferences* (Vol. 355, p. 02025). EDP Sciences.

Yousefi, A. and Pishvaee, M.S., 2022. A hybrid machine learning-optimization approach to pricing and train formation problem under demand uncertainty. *RAIRO-Operations Research*, *56*(3), pp.1429-1451.

YU, J., 2021. A New Way of Airline Traffic Prediction Based on GCN-LSTM. Frontiers in Neurorobotics, .

Zhai, H., Tian, R., Cui, L., Xu, X. and Zhang, W. 2020, "A Novel Hierarchical Hybrid Model for Short-Term Bus Passenger Flow Forecasting", *Journal of Advanced Transportation,* vol. 2020, pp. 16.

ZHAI, H., TIAN, R., CUI, L., XU, X. and ZHANG, W., 2020. A Novel Hierarchical Hybrid Model for Short-Term Bus Passenger Flow Forecasting. Journal of Advanced Transportation, 2020, pp. 16.

Zhang, H., He, B., Lu, G. and Zhu, Y., 2022. A simulation and machine learning based optimization method for integrated pedestrian facilities planning and staff assignment problem in the multi-mode rail transit transfer station. *Simulation Modelling Practice and Theory*, *115*, p.102449.

Zhang, P., Zhao, P., Qiao, K., Wen, P. and Li, P., 2023. A Multistage Decision Optimization Approach for Train Timetable Rescheduling Under Uncertain Disruptions in a High-Speed Railway Network. *IEEE Transactions on Intelligent Transportation Systems*.

Zhao, X., Shan, X. and Wu, J., 2023. The Impact of Seat Resource Fragmentation on Railway Network Revenue Management. *Networks and Spatial Economics*, *23*(1), pp.135-177.

Zhou, W. and Han, X., 2023. Integrated optimization of train ticket allocation and OD-shared ticket sales strategy under stochastic demand. *Transportation Letters*, pp.1-10.

Zhou, W., Wang, W. and Zhao, D. 2020, "Passenger Flow Forecasting in Metro Transfer Station Based on the Combination of Singular Spectrum Analysis and AdaBoost-Weighted Extreme Learning Machine", *Sensors,* vol. 20, no. 12, pp. 3555.

ZHOU, W., WANG, W. and ZHAO, D., 2020. Passenger Flow Forecasting in Metro Transfer Station Based on the Combination of Singular Spectrum Analysis and AdaBoost-Weighted Extreme Learning Machine. Sensors, 20(12), pp. 3555.

Zhu, G., Ding, J., Wei, Y., Yi, Y., Xu, S.S.D. and Wu, E.Q., 2023. Two-Stage OD Flow Prediction for Emergency in Urban Rail Transit. *IEEE Transactions on Intelligent Transportation Systems*.

Znidersic, E., Towsey, M., Roy, W.K., Darling, S.E., Truskinger, A., Roe, P. and Watson, D.M., 2020. Using visualization and machine learning methods to monitor low detectability species—The least bittern as a case study. Ecological Informatics, 55, p.101014.