UNIVERSITY OF TWENTE.

Faculty of Behavioural, Management & Social Sciences

Enhancing Baggage Handling Duration Predictions for KLM: A Data-driven and Machine Learning Approach Using Camera and Sensor Data

Marco Luis Ochoa Barnuevo

BSc Final Thesis

in collaboration with

KLM

University Supervisors:
Dr. Dennis Prak - University of Twente
Prof.Dr.Ir. Martijn Mes - University of Twente

KLM Supervisors:
Tomas Pippia (Data Scientist, KLM Royal Dutch Airlines)
Joyce Hu (Data Scientist, KLM Royal Dutch Airlines)

Faculty of Behavioural,
Management & Social Sciences
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Acknowledgements

I am immensely grateful for the opportunity I had to pursue my bachelor's degree. It began with my dedication and hard work, which led to a scholarship from Ecuador. Coming to the Netherlands and discovering the beauty of the country and the University of Twente was truly remarkable. During my time as a student, I had the privilege of visiting more than 20 countries, where I had the pleasure of meeting incredible people and gaining valuable knowledge and experiences. These encounters have broadened my horizons and enriched my personal growth. I would like to sincerely express my gratitude to the government of Ecuador for the scholarship that made this transformative journey possible.

First and foremost, I would like to express my heartfelt gratitude to God for granting me the opportunity to embark on this incredible journey. I also extend my deepest appreciation to my professors Dennis Prak and Martijn Mes, whose guidance, expertise, and continuous support have been invaluable throughout this research journey. Their insightful feedback, constructive criticism, and dedication to academic excellence have shaped the trajectory of my work and broadened my horizons.

I would also like to extend my gratitude to my supervisors at KLM Joyce Hu and Tomas Pippia for their constant encouragement and cooperation. Their willingness to share knowledge, engage in meaningful discussions, and provide valuable insights have undoubtedly enriched my understanding of the subject matter.

To my dearest friends, thank you for your support, encouragement, and understanding throughout my bachelor's degree. Your presence has been a source of motivation, and your belief in my abilities has given me the confidence to persevere and made this journey more enjoyable and memorable.

Lastly, but most importantly, I want to express my deepest gratitude to my family, especially to my mother and my girlfriend. Their unconditional love and belief in my abilities have been the driving force behind my success. Their sacrifices, patience, and encouragement have provided me with the strength and determination to overcome obstacles and pursue my goals and dreams abroad.

I hope this acknowledgment serves as a small token of my gratitude to all those who have supported and guided me throughout this endeavor. Without their contributions, this journey would not have been possible. Thank you from the bottom of my heart.

Management Summary

Accurate baggage loading/unloading estimation is crucial for KLM's efficiency at Schiphol Airport. However, their current estimation tool lacks data validation and needs improvement, leading to inaccuracies and inefficiencies turnaround process. In this project, we aim to address the challenges faced in KLM's baggage handling process, which impact operational efficiency and customer satisfaction. The research questions were designed to investigate how data-driven and Machine Learning methods using camera and sensor data can enhance the accuracy of baggage duration predictions at Schiphol Airport. For this, a comprehensive analysis of the current system, empirical data, and modeling techniques were covered.

Research Findings: Our research involved analyzing camera data and performing thorough data transformation and preparation to accurately determine separate unloading and loading durations. We developed four Machine Learning (ML) models (Random Forest, XGBoost, Artificial Neural Network, and Support Vector Regression) using additional three advanced feature selection algorithms. Key features included the number of bags, aircraft group, continents, hour of the day, and day of the week. The Neural Network slightly surpassed the other models and achieved a Root Mean Square Error (RMSE) of 6.19 for unloading, and Random Forest performed slightly better in the loading duration with an RMSE of 7.43.

When comparing our models with KLM's current tool and two data-driven methods, the data-driven method based on average durations by aircraft type and groups of bags slightly outperformed our ML models in RMSE in the overall and subsets of data. Moreover, it's important to note that the current tool approximation may have advantages in rare instances of large outlier durations. Nevertheless, the data-driven current tool, which is based on data-based loading speeds, significantly outperformed the approximation method that relies on intuitive loading speeds.

Main Recommendations: Based on the positive outcomes of our research, we recommend that KLM focuses on further refining the models and data-driven methods by implementing the following strategies. First, collecting a year-round extended and diverse dataset that covers all seasons and months will provide a comprehensive understanding of the variations in baggage handling durations, enabling more accurate predictions. Additionally, gathering more data specifically for wide-body aircraft will enhance the statistical significance and generalizability of the ML models. To improve data quality, regular maintenance, calibration, improved network connectivity, and bug fixes should be implemented. Furthermore,

incorporating additional factors such as the number of workers, their state, or the state of belt loaders, and among others involved in the baggage handling process can further enhance task time estimation once more reliable data becomes available. Lastly, implementing prediction intervals using methods like bootstrapping or Bayesian inference will provide a range of values or confidence levels, effectively capturing uncertainty and improving the reliability of predictions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The aviation sector is an essential component of contemporary transportation, linking people and companies all over the world. The success of this business is contingent on the proper functioning of many processes and systems, including ground-handling activities. The aircraft turnaround process is the time between an aircraft's arrival at a terminal and its subsequent departure during which different procedures such as refueling, cleaning, luggage loading and unloading, and passenger boarding take place. This procedure is crucial to airline operations since it directly affects flight schedules, personnel allocation, and, ultimately, customer satisfaction. During aircraft turnaround, baggage handlers unload luggage and cargo from both the front and rear cargo holds, after which the bags are sorted and transported to their next destination through the baggage handling system.

Baggage handling efficiency is critical to the performance of the aircraft turnaround process because it influences the time necessary for the aircraft to be ready for departure. Delays, higher operating expenses, and decreased customer satisfaction can all result from inefficient baggage handling systems. Airlines use numerous methods to estimate the time necessary for luggage loading and unloading in order to improve the aircraft turnaround process. These technologies, however, are frequently based on manual or outdated methods, which can lead to mistakes and inefficiencies.

As a result, the aviation sector is reliant on efficient aircraft turnaround and luggage handling, and accurate estimations of luggage loading and unloading times are necessary. Data-driven technologies can improve the precision and dependability of these estimations by analyzing preprocessed data from various sources such as camera feeds and previous flight data. Prediction models can be developed using this data to forecast the time required for luggage loading and unloading more accurately. These precise estimations not only enhance operational efficiency and cost savings but also lead to customer satisfaction by reducing delays and disturbances, resulting in higher revenue for airlines and aviation-related organizations. Harnessing these technologies offers an opportunity to streamline operations and improve customer satisfaction, which is critical to the success of the aviation sector.

1.1 Background and Context for the Study

In this section, information will be provided about the companies and departments involved in this project, as well as the project itself and its relevance.

Company

KLM Royal Dutch Airlines, established in 1919, is the national airline of the Netherlands and the oldest operating airline in the world under its current name. It provides passenger and cargo flights to over 145 global destinations, serving more than 35 million passengers each year. KLM is a founding member of the SkyTeam airline alliance and shares the same holding company with Air France.

The Business Platform Ground of KLM's Data & Technology department is in charge of analyzing data from flights and planes in order to gain insights and improve operations. This division is responsible for activities that take place below the wing, such as luggage handling, aircraft refueling, and ground handling. In order to improve the efficiency and accuracy of ground operations, they provide prediction tools and optimizers based on data analytics and data-driven methodologies.

The Terra Team is a Data & Technology sub-department in charge of the Terra scheduling tool. During the day of operations, the Terra tool enhances the resource scheduling process for ground-handling workers. The Terra Team collaborates with different departments to guarantee that the TERRA tool is reliable, precise and satisfies the demands of the airline.

Schiphol Airport, located in Amsterdam, is the Netherlands' primary airport. It served over 71 million passengers in 2019, making it the third-busiest airport in Europe by passenger traffic. Schiphol Airport is the hub for KLM and its SkyTeam partners, as well as several other airlines. The airport is known for its efficient and innovative operations, winning many awards for sustainability and customer service initiatives.

Project

The project, developed for KLM's Business Platform Ground, aims to enhance the efficiency of their Schiphol Airport turnaround processes by predicting the duration of baggage loading and unloading. This will involve validating the current prediction tool, conducting an in-depth analysis of the data, creating multiple prediction models, and implementing them while comparing their performance to each other and to the current tool. Additionally, the project will consist of conducting a comprehensive literature review to gain a deeper understanding of the critical role played by baggage loading and unloading procedures within the aircraft turnaround process, as well as analyzing current state-of-the-art techniques used in luggage handling and turnaround procedures.

The data for the project is collected through the Schiphol DeepTurn initiative, which produces timestamp data for all the different processes that take place during a turnaround by making use of camera feeds on the ramp. Using such information, it will be feasible to de-

velop a prediction model to estimate the duration of luggage loading and unloading during turnarounds and lower the degree of uncertainty in these durations. The new prediction model will be used in some of KLM's internal optimization tools such as TERRA and it will be useful for The Terra Team for operations decision support.

Relevance

This project has the potential to advance both academic and practical knowledge in the aviation industry, with implications for airlines worldwide.

From an academic perspective, this study has the potential to contribute to studies on data-driven decision-making in aviation operations. The outcomes of the study could also be used to inform future studies on enhancing passenger processes and customer service quality at airports. Furthermore, precise forecasting of the duration of various tasks during the turnaround process can be a helpful input for much more complicated operations planning and decision-making difficulties faced by airlines. The research can add to existing knowledge on predictive modeling in turnaround operations by reviewing and using cutting-edge methodologies.

From a practical perspective, this project has direct implications for KLM and other airlines, as it aims to increase the efficiency of KLM's planning process. Currently, excessive time allocation leads to inaccurate and prolonged task scheduling, resulting in inefficiency. Despite the existence of a standard slack time based on factors like minimum travel time and ground staff readiness, the inaccurate estimation of unloading and loading durations further increases staff slack time and hampers efficient resource allocation by planning tools. Therefore, by accurately predicting the duration of baggage loading and unloading, the project can improve ground operations planning, leading to more efficient turnaround operations, increased operational efficiency, and better support for planning and decision-making. These improvements can result in airline cost savings and improved customer satisfaction. Additionally, the use of machine learning and predictive modeling techniques can help automate decision-making processes and improve the accuracy of predictions, making turnaround operations more reliable and consistent.

1.2 Problem Definition

As a major player in the airline industry, KLM faces challenges related to its baggage loading and unloading process. These challenges have the potential to impact its operational efficiency, customer satisfaction, and overall competitiveness. Accurately estimating the time required for this process can be complex and difficult, and KLM is seeking to improve the accuracy of these estimates to enhance its overall performance.

1.2.1 Problem Identification

After generating a list of problems, the following interconnected issues were identified as the root causes of these inefficiencies in KLM's turnaround process:

The lack of proper data validation for the current estimation tool used to predict baggage loading and unloading duration during turnarounds is one of the primary problems. This creates uncertainty about the accuracy of the task duration estimates, leading to inexact turnaround times, delays in other processes, and ultimately higher operational costs, decreased customer satisfaction, and potential loss of revenue for KLM.

KLM may face a challenge in terms of its data collection and analysis methods, which could rely heavily on manual processes. Although these methods can be time-consuming and potentially prone to errors, they may not provide real-time insights into the turnaround process. This could potentially result in difficulties in collecting accurate data for the turnaround process, which highlights the importance of proper data validation. However, with the help of Schiphol's DeepTurnaround initiative and the data collection from the ramp, this process becomes now more reliable, providing the necessary validation for the prediction tool.

Additionally, the lack of visibility into the baggage loading and unloading process makes it challenging to identify bottlenecks and areas for improvement. This is compounded by task duration variability and reliance on manual data collection and analysis methods, which further hinder KLM's ability to optimize its turnaround process.

These problems have a cascading effect on other processes, leading to further operational and financial losses for KLM. Without real-time insights into the turnaround process, it is difficult to identify the root causes of inefficiencies or to optimize the process for future turnarounds. This perpetuates the cycle of inefficient turnarounds and inaccurate task duration estimates, leading to ongoing operational and financial losses for KLM.

Therefore, the core problem we are going to solve is the following:

1.2.2 Core Problem

"The accuracy of task time estimations for baggage loading and unloading in the KLM turnaround process is questioned due to inadequate data validation and the need for improved precision in the current approach"

1.2.3 Norm and Reality

Reality: The current estimation tool used by KLM lacks sufficient data validation, leading to concerns about the accuracy of the task time estimations and potentially resulting in inefficient turnaround processes.
Norm: Empirical assessment of the current formula's quality, detailed analysis of the data, and an ideal improved prediction model that leads to more efficient planning in Terra.

1.2.4 Scope of Research

This research project aims to analyze and enhance the accuracy of KLM's current estimation tool for baggage loading and unloading durations during turnarounds at Schiphol Airport. The study will be conducted from August to November 2022 and from February to March 2023, focusing on the prediction time frame of 3 to 1 hour before flight departure. By using data-driven and Machine Learning techniques, the project aims to identify patterns and significant features to develop a more precise prediction model. The performance of the new model will be compared to the existing tool or a suitable approximation if data is unavailable, to assess its effectiveness. Furthermore, an implementation plan for integrating the new model into the TERRA tool will be provided to KLM. This study will yield valuable insights into predicting luggage handling duration in the turnaround process, with the potential for adaptation by other airlines and airports after adjusting data gathering and preparation methods.

1.3 Research Questions

The fundamental questions that a research study seeks to answer are referred to as research questions. They direct the entire research process and aid in the study's focus. The following main research question is proposed:

1.3.1 Main Research Question

"How can data-driven methods, leveraging the camera data from the airport and the sensor data from the aircraft, be effectively used to enhance the accuracy of KLM's baggage loading and unloading duration predictions at Schiphol Airport?"

This research aims to determine how to accurately predict how long it takes to load and unload baggage during the KLM turnaround process by analyzing camera data from the airport. The research subject is crucial because it tackles a fundamental issue that KLM and other airlines confront when relying on flawed methods to estimate certain tasks, in this case, the baggage handling task, and the absence of sufficient data analysis and more complex prediction models.

1.3.2 Sub-research Questions

To achieve the research objective, a series of research questions have been developed and are divided into four distinct phases: literature review, current system analysis, empirical data analysis, modeling and validation, and performance evaluation and integration strategy for the proposed model. Each phase constitutes a chapter in the study, addressing specific research questions necessary to meet the main research question. In case a research question includes sub-questions, those sub-questions are required to answer the main research question.

Literature Review

The literature review aims to investigate the factors that affect the efficiency of the baggage loading and unloading process in the aircraft turnaround process. The review will also examine the significance of accurate estimation of baggage loading and unloading times. Additionally, it will explore various data-driven techniques used for extracting insights from data, such as descriptive and inferential statistical methods, data visualization techniques, feature selection techniques, and techniques for creating and evaluating prediction models for baggage loading and unloading duration.

1. What factors impact baggage unloading and loading efficiency in aircraft turnaround, and how does the accurate estimation of loading and unloading times contribute to improved efficiency?
2. What data-driven techniques for data analysis and feature selection can be used to understand and prepare the given data for predictive modeling?
3. What are the various types of prediction models currently utilized in the industry, and which among them are suitable for accurately predicting baggage loading and unloading task duration?
4. What evaluation metrics and design considerations are important for predictive modeling in baggage unloading and loading duration?

From System Analysis to Data Preparation

This chapter uses a comparative study design, employing descriptive research methods and methodological data preparation to analyze the baggage loading/unloading process. Additionally, primary data will be collected through observation and existing sources, such as the DeepTurnaround initiative and the Flight 720 platform. Through descriptive analysis, the accuracy and limitations of the current estimation tool will be assessed. To address system limitations, a mixed-methods approach, including surveys, interviews, and qualitative and quantitative analysis, will be implemented. Furthermore, data cleaning, preprocessing, and transformation will be conducted to establish a robust foundation for analysis. Finally, initial

data visualization will be utilized to aid in identifying patterns and trends, providing valuable insights for further investigation.

5. How does KLM currently estimate the baggage unloading and loading durations?
   (a) What is the current method used for estimation, and how accurate is the current estimation tool in predicting the duration of baggage loading and unloading in the aircraft turnaround process?
   (b) What metrics will be used to measure the accuracy of the current estimation tool?
6. How can data collected from cameras and sensors in the baggage loading and unloading process be effectively prepared for analysis in the prediction tool?

Feature Engineering

The goal of this chapter is to enhance the quality and relevance of available features, with the ultimate objective of improving the performance and interpretability of prediction models. The proposed approach involves carefully selecting, transforming, and creating new features to extract valuable information and patterns from the data. Descriptive statistical techniques will be employed to understand the central tendency and variability of features, while inferential analytics will be used to identify relationships and assess their impact on the target variables. Categorical variables will be encoded, and continuous variables will be scaled for optimal representation. Furthermore, feature selection techniques based on existing literature will be applied in the future to identify the most influential features for the prediction models.

7. How can statistical analysis and feature selection techniques aid in identifying influencing features for the target variable at the time of prediction?

Modelling and Validation

In this chapter, the primary goal is to construct and validate prediction models capable of accurately estimating the duration of baggage loading and unloading. The selection of the most appropriate prediction techniques will be based on the explanatory features and informed by the literature review. Additionally, the literature review will identify the appropriate methods for assessing the accuracy of the predictions.

8. How can the prediction models identified in the literature be adequately prepared and trained to ensure accurate estimation of baggage unloading and loading durations?

Results and Evaluation:

This chapter aims to achieve two objectives: evaluating the performance of the proposed prediction model and determining an integration strategy. To conduct a comprehensive evaluation, discrepancies between the new model and the current tool will be identified to understand their strengths and limitations. Accuracy metrics will be compared over the entire

dataset and specific subsets such as aircraft types or specific duration ranges. Additionally, a root cause analysis may be performed to identify contributing factors to lower predictability. Following this, an integration strategy will be developed with the help of the Business Platform Ground and the Terra Team to ensure seamless integration of the new prediction model with the Terra tool.

9. How to assess and compare prediction models, identify the best performer, and extract key insights for further improving the top-performing model?

10. What documentation and guidelines facilitate a smooth integration of the new prediction model with the Terra resource planning tool for KLM?

1.4 Methodology

To ensure a structured and reliable approach to data analysis and modeling in this project, we have adapted the well-known Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a well-known data mining and analytics methodology. Our adapted methodology, illustrated in Figure 1.1, consists of five phases tailored to meet the specific needs and requirements of our process.

Our methodology, while based on CRISP-DM, introduces certain modifications that enhance its effectiveness. It places a greater emphasis on comprehending the current system and ensuring appropriate data preparation, as well as understanding and validating the accuracy of the current tool. The introduction part has touched upon the CRISP-DM Business Understanding phase, where we have partially comprehended the business objectives and requirements of the baggage loading and unloading process, providing a clear direction for the subsequent steps. Furthermore, the Literature Review contains the theoretical concepts and procedures to be employed for this research. Subsequently, relationships between the adapted methodology and the CRSP-DM framework are as follows:

1. From System Analysis to Data Preparation: This phase combines the CRISP-DM Business Understanding, Data Understanding, and Data Preparation phases. It involves understanding the current estimation tool and baggage loading/unloading data, followed by data extraction, preprocessing, transformation, and initial analysis to identify patterns and trends.

2. Feature Engineering: Combining the CRISP-DM Data Preparation and Data Understanding phases, this phase continues the data understanding process using descriptive and inferential statistical analytics techniques, along with feature selection techniques.

3. Modelling and Validation: Similar to the CRISP-DM Modelling phase, this phase focuses on constructing and validating prediction models that accurately estimate the duration of baggage loading and unloading.

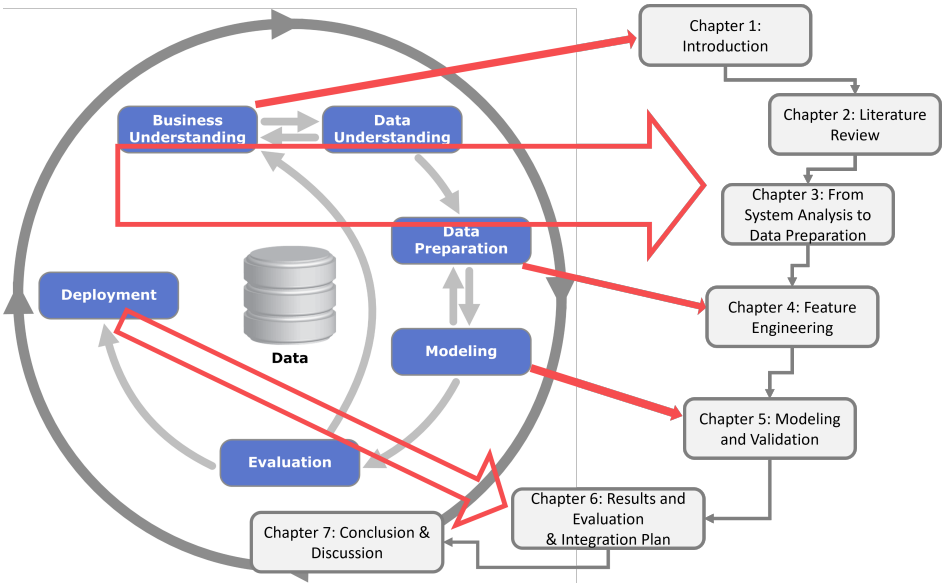


Figure 1.1: Adaptation of my research in the CRISP-DM cycle

4. Results and Evaluation: This phase combines the CRISP-DM Evaluation and Deployment phases. It involves evaluating the performance of the proposed prediction model against KLM's current estimation tool (Evaluation), and determining an integration strategy for the proposed model in the Terra tool (Deployment).

1.5 Document outline

This paper is structured in the following format: Following this chapter, Chapter 2 is a literature review that explores data-driven techniques for analyzing and predicting baggage duration. The chapter reviews relevant studies and research on feature selection and prediction models for baggage duration and identifies gaps in the existing literature. Chapter 3 focuses on data extraction and preprocessing, as well as understanding and validating the accuracy of the current estimation tool. This chapter will also identify any issues with the data and clean it to ensure its suitability for analysis. Chapter 4 conducts empirical data analysis using analytics techniques and selects the most influential features for predictive analysis. In Chapter 5, the study aims to construct and validate prediction models for baggage duration and selects the best-performing model. Finally, Chapter 6 evaluates the performance of the newly proposed model against KLM's current tool. The chapter discusses the criteria and metrics used to compare and determine if the proposed model is an improvement over the current model. Additionally, the chapter proposes an integration strategy with the company's Business Platform Ground and Terra Team. Finally, Chapter 7 will summarize the key findings of the study and discuss their implications, provide recommendations for future research, and conclude the thesis. Refer to Figure 1.2 for a visual outline of the project's structure and organization.

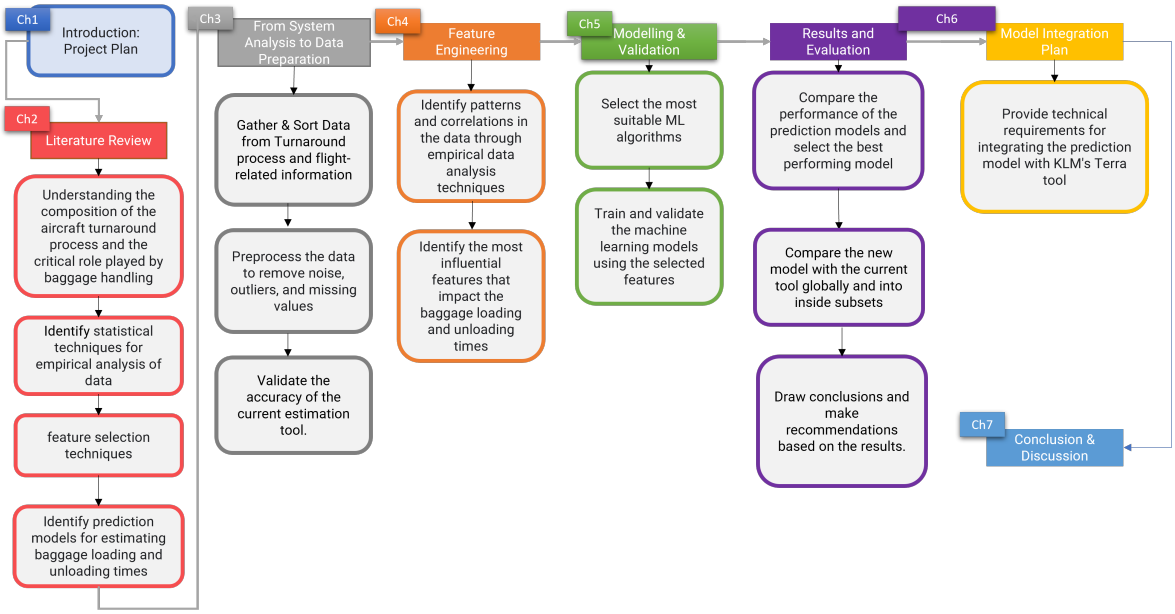


Figure 1.2: Project Structure

Chapter 2

Literature Review

This chapter provides an overview of the scientific background required to understand the context of the current study. First, in Sections 2.1 and 2.2 an overview of the aircraft turnaround process and the baggage loading and unloading process is respectively provided, as the focus of this study lies in improving the estimations of the baggage handling duration. Afterward, in Section 2.3 the empirical methods for data analysis, feature selection techniques, and prediction models are examined with the ultimate goal of identifying the best practices to be applied in this research study in order to improve the accuracy of KLM's baggage loading and unloading duration predictions at Schiphol Airport. The chapter concludes with Section 2.4, which addresses the existing literature gap and outlines the approach to be taken in this research study.

2.1 Overview of the Aircraft Turnaround Process

This section provides a comprehensive review of the aircraft turnaround process, discussing its stages in Sub-section 2.1.1, examining factors influencing efficiency in Sub-section 2.1.2, and emphasizing the importance of optimizing and accurately estimating its duration in Subsection 2.1.3. Understanding this process is vital to grasp the impact of baggage handling on overall turnaround efficiency.

2.1.1 Stages of the Turnaround

This subsection aims to present an overview of the stages of the aircraft turnaround process, which encompasses a sequence of activities that require efficient and effective completion to ensure the aircraft's readiness for the next flight.

The aircraft turnaround process is the period between an aircraft's arrival at the gate and its departure. This process involves numerous tasks that must be completed efficiently and safely to ensure the aircraft is ready for its next flight (Schmidt, 2017). The most important tasks in this process include refueling, loading and unloading of baggage, catering, cleaning the cabin, and servicing the aircraft. Fuel calculation is based on distance, aircraft weight, and other factors (Ashford et al., 2013). Factors that can affect baggage handling include

the number and size of bags and the number of ground crew Timajo et al. (2014). Catering, cleaning the cabin, and servicing the aircraft are important tasks that must be completed efficiently to ensure the safety and comfort of the passengers (Ashford et al., 2013; Schultz and Fricke, 2008).

Wu (2016) illustrates the primary activities involved in the turnaround process shown in Figure 2.1.1, including passengers' boarding and disembarking, luggage loading and unloading, refueling, routine maintenance, catering loading and unloading, cabin cleaning, security procedures, and pre-flight checklist.

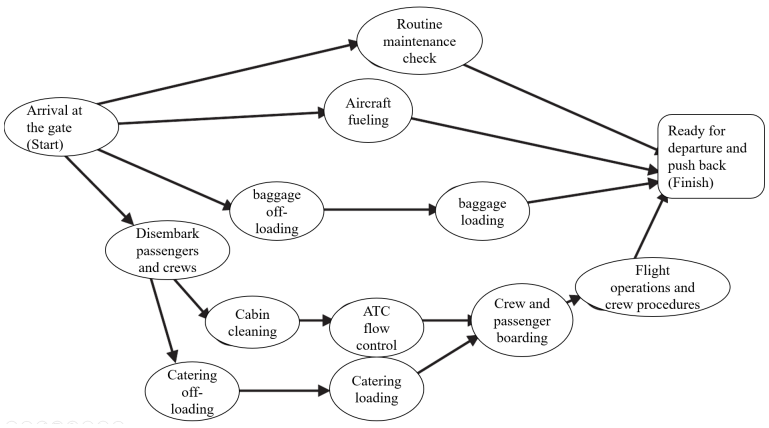


Figure 2.1: Activities in the Turnaround process, inspired in Wu (2016).

2.1.2 Factors that affect the Efficiency of the Turnaround

In this subsection, the factors that can influence the turnaround, as well as their potential impact on its various stages, are explored. Timajo et al. (2014) claim that several factors impact the turnaround time, such as ground staff, air traffic control, and weather. Ground staff performs tasks like loading and unloading luggage, embarking and disembarking passengers, servicing, and maintenance. Any delays caused by issues like a lodged cargo can lead to a succession of delays. Air traffic control is responsible for overseeing aircraft movement and issuing clearance for takeoff, landing, and taxiing. Some airports are prone to unexpected weather conditions such as high winds, fog, and snow.

Additionally, quality factors in transport processes and human factor reliability in air operations affect aircraft ground handling efficiency (Szabo et al., 2022). Flight-based and airport-based factors can affect an aircraft's turnaround time. Factors of flight-based factors include the number of passengers or bags, aircraft type, and others, while factors of airport-based factors include the capacity of an airport, delays, and weather conditions (Hassel, 2019). Postorino et al. (2020) outlines various effects of unanticipated interruptions on turnaround operations, including aircraft departure delays, missed connections, greater fuel consumption, additional crew and personnel expenditures, and lower passenger satisfaction. Ground staff disruptions are very important and have a substantial influence on turnaround operations and overall airport operations.

An analysis and simulation model can be used to evaluate the efficiency of aircraft turnaround operations and to identify critical activities. San Antonio et al. (2017) describes how simulation can be used to examine critical activities and paths in aircraft turnaround operations. A critical path is created by breaking down the process into smaller tasks and identifying their dependencies. Luo et al. (2021) used ABS to simulate the behavior of various agents participating in aircraft stand operations and to understand how their interactions affect total ground time. Moreover, Mota et al. (2017) used a discrete-event simulation, a method that models the aircraft turnaround process as a series of discrete events, to evaluate the turnaround performance of Lelystad Airport under different conditions weather disruptions, and technical problems. Their findings indicate that higher staffing levels result in faster turnaround times.

Finally, Gao et al. (2015) identified the key factors that influence the turnaround process through qualitative and quantitative analysis. These factors include the distance between the aircraft stand and terminal, aircraft type, domestic or international flight, airline agent, flight arrival time, flight arrival passenger number, and flight departure passenger number. Factors such as far or near aircraft stand, aircraft type, international flights, and airline agents significantly affect turnaround time. Flight arrival and departure times and the number of passengers also impact turnaround time.

2.1.3 Importance of Optimizing and Estimating the Turnaround Time

This subsection aims to briefly elaborate on the importance of optimizing and estimating the turnaround time to enhance operational efficiency, reduce costs, and improve customer satisfaction. Studies such as Evler et al. (2022) investigated how a resource-constrained turnaround scheduling model combined with an aircraft routing model may successfully forecast and minimize delay propagation in airline networks. This tactical decision support system can help airlines restore their schedules more quickly and effectively, lowering total costs and delays caused by schedule interruptions. Moreover, Wu and Caves (2004) used a Markovian simulation framework to recreate the various stages of the turnaround process and their interactions, accounting for stochasticity in flight punctuality and operational uncertainties by simulating the duration of each activity using probability distributions. This enabled them to evaluate the operational efficiency of turnaround operations.

The literature also emphasizes the importance of improving airport resource use throughout the turnaround phase to decrease expenses and delays caused by schedule setbacks (Evler et al., 2021), as well as, optimizing the turnaround duration to achieve on-time departure performance which is critical for airlines to stay competitive in the industry (More and Sharma, 2014). Mirza (2008) demonstrated how decreasing the turnaround time may enhance aircraft utilization, resulting in more flights per year. They show that decreasing this time's average by only 10 minutes - from 40 to 30 minutes - can increase utilization by 8.1 percent. Thus, by effectively managing airport resources and optimizing the turnaround time, airlines can improve their operational efficiency and reduce costs, which is crucial for remaining competitive in the industry.

2.2 Overview of the Baggage Unloading-loading Process

This section offers a comprehensive overview of the baggage unloading and loading process. Sub-section 2.2.1 provides a detailed description of the stages involved in this process, highlighting their sequence and impact on other processes. In Sub-section 2.2.2, the importance of the baggage unloading and loading process is underscored, along with the challenges associated with efforts to automate or optimize it. Additionally, Sub-section 2.2.3 focuses on examining the factors that influence the baggage loading and unloading process, which in turn, impact the overall aircraft turnaround time.

2.2.1 Stages of the Baggage Unloading and Loading

This process during the aircraft turnaround starts by unloading the baggage from the cargo hold onto the baggage carts. The baggage is then transported to the baggage handling area, where it is sorted and sent to the correct baggage carousel or transferred to connecting flights. Once the baggage has been unloaded, the next step is to load the new baggage onto the aircraft. The new baggage, which has been sorted and screened for safety already, is then transported to the aircraft on baggage carts and loaded into the cargo hold. Ground handling staff work closely with the flight crew to ensure a smooth operation, and following proper procedures is essential for safety and efficiency (AviationLearnings, 2020). Figure 2.2 illustrates this process.

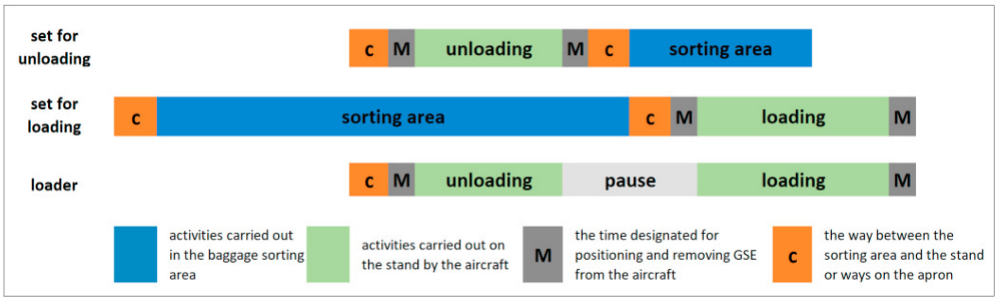


Figure 2.2: Aircraft unloading and loading schedule (according to Volt et al. (2022))

In the process of loading and unloading an aircraft, loaders are responsible for physically unloading and loading the baggage, while cart sets are used to transport the baggage between the sorting area and the aircraft stands. Typically, different sets of carts are used for loading and unloading, with a minimum of two sets needed for each flight. It is noteworthy that the set of carts used for loading begins its journey earlier than the set used for unloading to ensure that carts are ready and available when the loading process begins, even though the actual loading process occurs after unloading. This is due to the fact that loading is a more time-consuming and complex process that requires significant preparation compared to unloading. The passenger check-in process starts before the scheduled departure time, and the baggage loading and unloading process is closely coordinated with the flight schedule. The aim is to ensure that all baggage is transported to the correct endpoint before the

aircraft needs to be loaded, in order to achieve an on-time departure (Volt et al., 2022).

Ashford et al. (2011) outlines various types of baggage handling systems, including manual, semi-automated, and fully automated systems, and highlights the significance of appropriate equipment placement and layout to facilitate a seamless flow of baggage and minimize congestion. They also note that the efficiency of baggage handling is influenced by factors such as the amount, weight, and shape of the baggage, as well as the use of technology, such as automated systems and RFID technology, to track and organize the luggage. They, finally, emphasize the importance of regular maintenance and testing of baggage handling equipment to ensure its reliability and efficiency.

Schultz and Fricke (2008) investigated the stability and variability of processes involved in turnaround time, among these baggage (un)loading and passenger (de)boarding, and discovers technical shortcomings that contribute to longer times. The authors employed statistical methods to examine the influence of enhancing the dependability of important procedures on turnaround time and discovered that better boarding and de-boarding processes could dramatically lower this duration. Although the loading and unloading procedures were investigated, enhancing their dependability had little effect on turnaround time. They find that unloading is the most stable operation, followed by loading. Nevertheless, Frey (2014) observed that delays in the baggage handling process can, nowadays, result in lengthier turnaround times for flights due to increased air and passenger traffic, which results in larger luggage volume.

2.2.2 Importance and Challenges Associated with Baggage Handling

Efficient baggage handling is crucial to an airport's overall efficiency and has a substantial impact on airline competition. A baggage handling system's primary purpose is to guarantee that all checked-in or transfer bags arrive at their destinations before the aircraft is loaded (Tarău et al., 2009). Furthermore, improving all components of the aircraft turnaround process is critical to attaining on-time departure performance, which is a critical factor in the competitiveness of airlines (Rizal, 2016). Finally, Frey (2014) noted that delays in the baggage handling process can result in longer turnaround times for planes, affecting the airport's overall efficiency. As a result, airports and airlines must prioritize improving their baggage handling systems and turnaround operations in order to deliver a seamless and efficient travel experience for their passengers.

Volt et al. (2022) developed a mathematical model to reduce the amount of airport equipment required to load and unload airplanes, they explain that the loading or unloading time is influenced by a variety of factors such as the aircraft type, the load factor (proportion of occupied available seats), the quantity of baggage and cargo being loaded or unloaded, and the number of ground service equipment (GSE) available for the task such as loaders and carts. Other factors that can have an impact on loading or unloading time include the efficiency of the baggage sorting system, the number of passengers and their baggage, and any interruptions or delays in the boarding process. Tarău et al. (2010) discovered that the key control difficulties of a baggage handling system are processing unit coordination and

synchronization, route choice control of each bag, and velocity control of each destination-coded vehicle (DCV). They propose tackling these issues by improving control methods for determining the best route. Load patterns on the system vary greatly based on factors such as season, time of day, type of aircraft at each gate, and the number of passengers on each flight (De Neufville, 1994; Tarău et al., 2010).

2.2.3 Summary of Influencing Factors

The factors influencing both baggage handling and the overall aircraft turnaround process, discussed in Sub-sections 2.2.2 and 2.1.2, exhibit an overlap due to the inherent relationship between baggage handling and the turnaround process. These factors can be categorized as internal and external. The internal factors encompass flight-related aspects that are specific to each analyzed flight, including variables like the number of bags and the aircraft type. Conversely, external factors are predominantly associated with the airport environment and operations, impacting multiple flights and overall airport functionality. Examples of external factors include weather conditions and delay propagation. For a comprehensive overview of these influential factors, refer to Table 2.1.

Internal Factors	**External Factors**
Number of passengers	Unexpected delays duration
Number of general and special bags	Weather conditions
Weight of general and special bags	Aircraft departure delays
Number of ground service equipment (loaders & carts)	
Number of ground staff	
Aircraft ramp, stand	
Terminal Gate	
Aircraft type	
Time of the day + Day of the week	
Type of flight (domestic or international)	
Aircraft load factor	
Sorting system	
Aircraft load factor	
Late arriving passengers	
Loading/unloading strategy	

Table 2.1: Factors Influencing the baggage loading and unloading process

2.3 Data-driven Methods for Prediction Model Creation and Evaluation

Data-driven methods can be utilized to develop prediction models for the baggage unloading/loading process, which can lead to improved planning, scheduling, and resource allocation. In this literature review, the use of empirical methods for data analysis is explored in Sub-section 2.3.1, feature selection techniques are discussed in Sub-section 2.3.2, and predictive modeling techniques including modeling frameworks and machine learning models and principles are explored in Sub-sections 2.3.3 and 2.3.4. The review will provide an overview of the different methods, their strengths and limitations, and examples of their applications in the aviation industry. The goal is to identify the most effective and accurate methods for data analysis and prediction of baggage unloading/loading times in the aircraft turnaround process.

2.3.1 Empirical Methods for Data Analysis

Empirical models refer to models that are developed purely based on observations and data without any underlying theoretical principles. In aviation research, these models are widely used to understand various performance metrics such as aircraft turnaround times, baggage (un)loading times, delays, and other related factors. Han et al. (2012) covers a broad range of statistical analysis techniques that can be applied to identify patterns in data. These techniques can be useful for analyzing data related to the aircraft turnaround process.

Data Cleaning and Visualization

The initial step in the data analysis process involves cleaning the data to ensure its accuracy and completeness, including the identification and resolution of errors or missing values. Once the data has been cleaned, various visualization techniques are employed to gain insights and explore the dataset further. These techniques encompass scatter plots, histograms, box plots, bar charts, heat maps, and time series plots, each serving a specific purpose in visualizing different aspects of the data. The objective is to uncover trends, patterns, and outliers, as well as to investigate the relationships between variables. As a valuable resource for effective data visualization, Schwabish (2021) provides guidance and insights. Table 2.2 illustrates specific examples of visualizations that can be implemented for analyzing the baggage unloading and loading data.

Visualization technique	Purpose	Example
Scatter plot	Explore the relationship between two continuous variables	Actual duration of the baggage loading and unloading task and the number of ground personnel involved
Histogram	Explore the distribution of a continuous variable	Actual duration of the baggage loading and unloading task
Box plot	Explore the distribution of a continuous variable across different categories	Actual duration of the baggage loading and unloading task for different aircraft types or carriers
Bar chart	Explore the frequency or proportion of observations in different categories	Number of baggage loading and unloading events for different aircraft types or carriers
Heatmap	Explore the relationship between two categorical variables	Number of ground service equipment and the type of aircraft involved
Time series plot	Explore the change in a variable over time	Actual duration of the baggage loading and unloading task over different scheduled inbound and outbound dates

Table 2.2: Data Visualization Techniques with examples, inspired by (Schwabish, 2021)

Descriptive Statistics

To begin, computing descriptive statistics for the variables in the dataset such as mean, median, mode, standard deviation, range, and quartiles are necessary to understand the data's central tendency, variability, and distribution. Moreover, correlation analysis can be used to determine the relationship between the variables, and clustering analysis to find distinct groups of variables with similar features can help find factors that are highly correlated and may have a substantial influence on the luggage loading and unloading process.

Inferential Statistics

To gain a comprehensive understanding of statistical approaches for analyzing, interpreting, and inferring patterns from data, methods such as **ANOVA (Analysis of Variance)** can be used to determine whether there are any statistically significant differences between the means of two or more groups when both a continuous and a categorical variable are present if ANOVA indicates a significant difference, posthoc tests could be performed to identify which groups are significantly different from each other. Furthermore, the **chi-square test** could be employed to ascertain whether there is a statistically significant association between two categorical variables. Last but not least, when there are both a continuous dependent variable and an independent two-label categorical variable, the **T-test** could be used to compare the means of two groups (Nayak and Hazra, 2011). Nevertheless, one disadvantage of the t-test is that both samples must be normally distributed, or nearly so (Ghasemi and Zahediasl, 2012).

Multicollinearity

Multicollinearity refers to a high correlation among explanatory variables in a prediction model, leading to instability and difficulties in interpreting individual feature effects. Es-

sentially, it inflates regression coefficient variances, rendering them unreliable and less interpretable. To address this issue, the Variance Inflation Factor (VIF) quantifies correlations among continuous variables, where higher VIF values indicate stronger correlations. Furthermore, when dealing with categorical variables, Goodman-Kruskal's lambda coefficient measures the reduction in predictive error of one variable when another variable is known, revealing causality (Goodman et al., 1979). Alternatively, Cramer's V can also be used. However, unlike lambda, Cramer's V is a symmetrical measure that may not accurately capture the true one-way influence between variables (Towards AI Team, 2022).

Related work of empirical methods used for data analysis

In the literature, empirical methods have been employed to test data hypotheses and draw conclusions from the data by analyzing distributions, running statistical analyses and correlations, and conducting descriptive, inferential, and predictive analyses. Neumann (2019) used descriptive, inferential, and predictive analytics to study the relationship between the boarding procedure and flight turnaround time. The author gathered data on both processes and tested the hypothesis that the boarding process is on the crucial path of the turnaround process using OLS regression analysis and backward stepwise regression. The results validated the hypothesis, highlighting the importance of analyzing the relationship between different factors involved in the turnaround process to optimize operations and reduce delays. Hutter et al. (2019) did a similar statistical process to find what factors influence the boarding time and how they influence it they found that the number of passengers and the total capacity of the airplane in its selected configuration are the only variables that are required to obtain a good estimate of the boarding time.

Another paper that drew patterns through statistical analysis of a simulation study was (Wu, 2008). They present a real-time monitoring system for aircraft turnarounds and the author discovered that offloading goods during the turnaround process is normally trouble-free, unless there are equipment delays or failures. The loading procedure, on the other hand, might cause delays, especially for intricate connections between planes. According to the report, 22 percent of planes have late loading beginnings, which leads to 17 percent of late loading finishes and, ultimately, loading delays. They also demonstrate that early loading begins as a result of extended turnaround times. Loading-related difficulties account for 21 percent of flight delays, with half owing to load connections and the remainder due to late loading completion. Moreover, Horstmeier and de Haan (2001) looked at ways to shorten the turnaround time for an Airbus A380-200 with an 80 percent passenger capacity. To model passenger as well as cargo movements, several probability distributions are utilized. The article draws four insights for reducing turnaround time from the simulation: starting catering unloading upon arrival, employing belly catering (moving trolleys from the passenger deck to the lower belly), refueling while boarding and deboarding, and improving infrastructure. These technologies have the potential to reduce turnaround time by 12-17 minutes but also have downsides such as passenger discomfort, safety issues, and the requirement for infrastructure upgrades.

Overall, empirical methods play a crucial role in predicting aviation-related metrics when theoretical models or simulations are unavailable or not applicable. Not only do they allow researchers to draw important conclusions and identify influencing variables, but they also provide valuable guidance for developing more complex prediction models. Despite their advantages, these methods are limited by the quality and quantity of data available for analysis. Additionally, they may not be able to capture the intricate interactions between different factors that affect aviation performance. Nonetheless, empirical methods provide a useful starting point for predicting aviation-related metrics and can offer valuable insights.

2.3.2 Feature Selection

Feature selection is a crucial step in building predictive models as it involves identifying a subset of relevant features that significantly impact the target variable. Feature selection methods can be broadly classified into three types: filter methods, wrapper methods, and regularization methods (Kuhn and Johnson, 2021). These methods can be used in regression models as they evaluate the relevance of the features to the target variable(Kuhn and Johnson, 2021; Wang et al., 2013).

Filter methods

Filter methods are utilized to rank features based on their relevance to the target variable, employing statistical measures like Pearson correlation, Spearman's rank correlation coefficient, ANOVA one-way test, and the two-sided T-test. Pearson correlation quantifies the linear relationship between two continuous variables, with values ranging from -1 to 1. A value of -1 indicates a perfect negative correlation, 0 implies no correlation, and 1 signifies a perfect positive correlation. Similarly, Spearman's rank correlation coefficient ranks normally distributed numerical data. In the case of categorical independent variables with more than two labels and a continuous target variable, the ANOVA one-way filter method determines if there exist significant differences in the means of the target variable across different categories of the independent variables, using the F-test statistic the variance of the target variable between the groups can be assessed, giving in this way, the total impact of the categorical variable. For two-labeled categorical variables, the two-sided t-test provides its own statistical measure to analyze their impact on the target variable (Kuhn and Johnson, 2021).

Wrapper methods

Wrapper methods are widely used to assess the performance of machine learning models by evaluating different subsets of features and selecting the subset that yields the best results. Two popular options for wrapper methods are Recursive Feature Elimination (RFE) and Sequential Forward/Backward Selection. RFE gradually eliminates less important features based on their importance as determined by the model. On the other hand, Sequential Forward/Backward Selection adds or removes features based on performance metrics.

These methods are preferred over alternatives like Genetic Algorithm (GA) and Exhaustive search due to their compatibility with various ML models, utilization of model-specific metrics, and computational efficiency (Kuhn and Johnson, 2021). However, Sequential Forward/Backward Selection has a limitation where once a feature is added or removed, it remains fixed throughout the selection process. To overcome this limitation, Sequential Floating Forward/Backward Selection periodically re-evaluates the decision, allowing for temporary removal and addition of features. By adding the floating part, flexibility and adaptability are incorporated in the search for the best subset of features (Raschka, 2018).

Crone and Kourentzes (2010) discusses the challenges faced in feature selection for time series data and argues that while wrappers are better than filters, they require significant computational power. Additionally, filters are limited in identifying non-linear interdependencies and their confidence intervals become narrower with larger sample sizes, leading to non-parsimonious models that rely on sample size rather than the data structure. To address these issues, the authors propose an iterative neural filter that uses a two-stage process. The first stage uses a filter to identify relevant features and reduce the search space for feature selection. The second stage employs a wrapper to evaluate the remaining features by computing forecasts for feature subsets, considering the inductive algorithm's biases and properties.

Regularization methods

Regularization methods are used to add a penalty term to the objective function of the model to encourage sparsity in feature selection and thus select relevant features in a dataset. Regularization methods are especially useful when the number of features is large, and there is a risk of overfitting the model. Lasso and Ridge's regression are two popular regularization methods that work by adding L1 and L2 penalties, respectively, to the objective function. Lasso regression helps to shrink the coefficients of less important features to zero, leading to feature selection, while Ridge regression helps to reduce the impact of multicollinearity between features (Hastie et al., 2009).

2.3.3 Modelling Framework for the Baggage Handling Process

Selecting the appropriate modeling framework is crucial for establishing a clear direction and determining the subsequent models to be employed. When it comes to predicting the duration of baggage unloading and loading, there are various frameworks to consider, such as regression analysis, time-series analysis, and their respective variations. In this section, we delve into these methods, offering a comprehensive description and assessing their suitability for the specific task at hand.

Simple Regression

This method considers every observation equally and ignores any temporal relationships or patterns. If the data does not show any temporal patterns and the purpose is to predict a

continuous outcome variable based on a collection of predictor factors, this method may be acceptable.

Enhanced Regression

This strategy employs a regression model to predict the target variable but incorporates extra time-related variables such as weekdays or hours of the day. This accounts for systematic fluctuations in the data caused by the time of day or day of the week, which can enhance the model's accuracy and interpretability.

Univariate Time-Series

This method entails constructing a distinct time series for each label of a categorical variable; in this case, it is conceivable to build separate time series for each gate or ground crew; and estimate them independently. This implies that there is no statistically significant relationship between the labels of the category considered and that this categorical variable is the only influential factor.

Multivariate Time-Series

This approach creates time series for each gate or ground crew, considering correlations and influences. It reveals relationships between different gates or ground crews and their impact on baggage loading/unloading. However, the unrealistic assumption that there are observations of different gates at exactly the same time should be made. Additionally, variables (e.g., bags, passengers, aircraft, weather) can be included as predictors using models like VAR or ARIMA, this can help describe the relationships. However, each variable should be stationary, and lag lengths and model specifications should align with the data (Andrews et al., 2013).

Hybrid Models

There are also hybrid models such as ARIMAX that can be used to create forecasts for time series data with explanatory variables. This approach combines the benefits of time-series modeling with regression modeling to create a more accurate and flexible model (Andrews et al., 2013). For example, Somyanonthanakul et al. (2022) demonstrated that the ARIMAX model has the potential to predict the number of COVID-19 cases by also incorporating the most associated variables such as weather conditions and population density.

In conclusion, considering that it is more likely to have more than one influential factor based on the factors found in section 2.2.3, the most appealing method for predicting KLM baggage unloading and loading is enhanced regression. This method can consider all flight-related, airport-related, weather, and time-related factors, providing significant results without unnecessary complexity.

2.3.4 Prediction Models

Machine learning (ML) has emerged as a powerful tool for developing sophisticated prediction models in recent years. ML, as a sub-field of Artificial Intelligence (AI), entails training algorithms on large datasets to make predictions or judgments without the need for explicit programming. As a result, this section will concentrate on machine learning and its various applications. The objective of this section is to elucidate the potential of machine learning in predicting airline operations, such as baggage loading/unloading duration, by providing a comprehensive overview of the existing literature.

Machine learning (ML) encompasses several types, such as supervised, unsupervised, semi-supervised, and reinforcement learning. Among these, supervised learning is widely used to train models on historical data, where inputs representing specific process features or conditions are used to predict the duration of the process. While the focus here is on aviation processes due to their similarities in uncertainty and sources, it is important to highlight that ML models have also shown innovation in other industries. Lastly, Figure 2.3.4 provides a summary of studies that have used ML approaches to predict specific turnaround stages or the entire process.

Ensemble methods

Ensemble methods, such as bagging, boosting, and stacking, are powerful techniques that improve predictive performance and robustness by combining multiple models. Bagging generates variations of samples to train base classifiers and reduce variance. Boosting trains a sequence of weak learners to reduce bias. Stacking combines heterogeneous models using a meta-learner to improve predictions (Diana, 2018; Sutton, 2005).

Random Forest is a bagging ensemble method that builds multiple decision trees and combines their predictions. It is a highly accurate and robust model that works well with numerical and categorical data. Random Forest can handle high-dimensional datasets with many features, which makes it a good fit for a dataset with multiple explanatory variables (Sutton, 2005).

Moreover, Gradient Boosted Trees (GBTs) are powerful ML models that combine decision trees with gradient boosting. This technique iteratively improves model performance by adjusting the weights of misclassified samples (Friedman, 2001). GBTs come in different types, each with unique features. The most basic form is the Gradient Boosting Machine (GBM), where each tree grows independently of previous trees' results (Friedman, 2001). Extreme Gradient Boosting (XGBoost) uses a regularized form of gradient boosting to prevent overfitting (Chen and Guestrin, 2016). LightGBM is a fast and efficient implementation that uses 'gradient-based one-side sampling' to reduce computation time (Ke et al., 2017). CatBoost includes a specialized preprocessing step to handle categorical features more effectively (Dorogush et al., 2018). Another well-known boosting strategy that was considered was Adaboost, which likewise strengthens the ensemble over time by combining weak learners into a strong learner by altering their weights (Sutton, 2005).

Neural Networks

Neural networks are a machine learning model that emulates the structure and function of the human brain, enabling them to recognize patterns and relationships in various types of data. The four main types of neural networks are feedforward, which processes information in a one-way flow; recurrent, designed for sequential data and equipped with feedback connections; Long Short-Term Memory (LSTM), a specific type of recurrent network that can selectively forget or remember information from previous time steps (Staudemeyer and Morris, 2019); and convolutional networks, which analyze input data through filters to extract features at different spatial scales.

Other ML Methods

Support Vector Regression (SVR) is a type of machine learning algorithm that can be used for predicting continuous variables. In SVR, the goal is to find a hyperplane in a high-dimensional space that maximally separates the training examples while minimizing the error on the test examples.

Bayesian Networks are probabilistic graphical models that can be used for modeling the relationships between variables and predicting the probability of specific events occurring. In Bayesian Networks, each node represents a variable, and the edges between nodes represent the relationships between variables.

Related work on the Machine Learning methods

Wang et al. (2022) utilized various base learners, such as linear regression, k-nearest neighbor, support vector regression, and light gradient boosting machine to predict flight delays, and found that the stacking method provided the best prediction performance for the test dataset. Moreover, Luo et al. (2021) utilized agent-based simulation and ML algorithms, such as decision trees, random forests, and XGBoost, to predict aircraft ground times at stands. The authors used classification models to predict the type of ground handling process required for a given flight and regression models to predict the duration of each process. They found that the XGBoost model performed the best.

Furthermore, Gao et al. (2015) used a feedforward Neural Network to predict the turnaround time at two major Chinese airports. The authors identified 7 key factors (mentioned also in Section 2.1.2) through qualitative and quantitative analysis and uses the TRAINLM algorithm from MATLAB as a learning algorithm. Other ML models, such as the Long Short-Term Memory (LSTM) neural network model, have also been used to predict boarding times based on passenger characteristics and the number of passengers seated, as proposed by (Schultz and Reitmann, 2019). Moreover, Schultz et al. (2021) used historical flight data and weather information to train several classification models, including decision trees, random forests, and support vector machines, to predict the impact of weather on airport performance. The authors reported a high overall accuracy of around 80% for their model. In addition to neural networks and support vector machines, other ML models have

also been used to predict aircraft turnaround tasks. For example, Hassel (2019) proposed a process-structure-aware approach to predict the turnaround time of an aircraft. The features used for prediction included the scheduled turnaround time, operating capacity, and passenger load factor. Two ML models, a Feed Forward Network and Random Forest, were utilized for prediction, and regression evaluation metrics such as MAE, RMSE, and MAPE were used to measure model performance. For classification, metrics such as accuracy, precision, recall, and f1-score were used to measure the quality of announcing an activity as ”critical.”

Jasra et al. (2018) conducted a literature review of ML techniques used to analyze flight data, including anomaly detection methods, such as distribution-based, depth-based, clustering-based, and distance-based methods, as well as unsupervised methods, such as multiple kernel anomaly detection and one-class support vector machines. In addition, Oreschko et al. (2012) used a Bayesian network and stochastic process model to estimate the delay patterns of each airport by utilizing a multivariate distribution. Finally, Yıldız et al. (2022) used a deep learning and computer vision-based approach to automatically detect and monitor the timestamps of ground service actions in airports to improve turnaround operations.

Finally, Carpinteiro et al. (2012) explores the performance of these three models on a time series of a Brazilian stock market fund. The authors found that the hierarchical model, which includes a self-organizing map and SVM on top, outperformed both the SVM and MLP models in terms of predictive accuracy. This study highlights the usefulness of hierarchical modeling approaches for time series analysis and prediction.

Table 2.3 shows an overview of the ML models used in the literature for each of the tasks in the Turnaround process and Table 2.4 depicts the different mentioned ML models with the type of features they require.

Predictions in the Turnaround process								
Machine Learning Models	**Aircraft Turnaround**	**Cleaning**	**Catering**	**Fueling**	**Baggage (un)loading**	**Aircraft (de)boarding**	**Aircraft Push-back**	**Flight Delays & Flight Data**
Neural Networks & variations	Gao et al. (2015) Hassel (2019)					Schultz and Reitmann (2019)		
Random Forest Regression	Luo et al. (2021) Hassel (2019)	Luo et al. (2021) Hassel (2019)	Luo et al. (2021) Hassel (2019)	Luo et al. (2021) Hassel (2019)	Luo et al. (2021) Hassel (2019)	Luo et al. (2021) Hassel (2019)	Luo et al. (2021) Hassel (2019)	
Gradient Boosted trees	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Wang et al. (2022)
Support Vector Regression	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Luo et al. (2021)	Wang et al. (2022) Jasra et al. (2018)
Bayesian Networks	Oreschko et al. (2012)							Oreschko et al. (2012)

Table 2.3: Machine learning models used in existing Literature for each task in the turnaround process

Machine Learning Model	Type of Feature they can handle
Decision Trees or Random Forests or Gradient Boosted Methods	Numerical and categorical features Non-linear relationships between the features and the target variable
Support Vector Machines (SVMs)	Numerical features Non-linear relationships between the features and the target variable, by using a kernel function to map the features to a higher-dimensional space
Neural Networks	Numerical and categorical features Complex non-linear relationships between the features and the target variable
Bayesian Networks	Numerical and categorical features Complex relationships between the features and the target variable, including non-linear relationships

Table 2.4: ML models with required explanatory features, inspired by (Géron, 2017; Sutton, 2005)

2.3.5 Machine Learning Fundamentals

This section covers key concepts in machine learning: overfitting, underfitting, and bias-variance trade-off. These concepts are crucial for understanding ML model performance and generalization.

Key concepts

The bias-variance trade-off is the balance between a model's ability to capture complexity and generalize well. Bias refers to simplified assumptions, while variance refers to sensitivity to training data fluctuations. High bias leads to underfitting, where a model is too simplistic to capture underlying patterns, while high variance results in overfitting, where a complex model performs well on training data but fails to generalize to unseen data.

To mitigate the issue of overfitting caused by high variance and control the model's complexity and generalization performance, it is recommended to employ the use of simpler models and apply regularization methods, such as L1 and L2 regularization, as suggested by Burkov (2019). Moreover, to address underfitting caused by high bias, Géron (2017) suggests to use of more complex models such as neural networks or ensembles of decision trees, and Kuhn and Johnson (2021) recommends enhancing data representation in the feature engineering and selection process.

2.3.6 Evaluating Regression Models

Evaluation metrics for regression models provide insights into the performance and accuracy of the predictions. In regression tasks, where accurately calculating the target variable

is challenging, accuracy, which is commonly used in classification tasks, is an inappropriate metric. Instead, regression evaluation metrics, also known as error metrics, focus on measuring the proximity between predicted and actual values.

Evaluation Metrics

Evaluation metrics in regression tasks focus on measuring the proximity between predicted and actual values. Based on Géron (2017), they include:

- **Mean Absolute Error (MAE):** MAE calculates the average absolute difference between observed and predicted values. It provides a straightforward assessment of the model's performance.
- **Mean Squared Error (MSE):** MSE computes the average of the squared differences between predicted and observed values. It emphasizes outliers and is more sensitive to large errors compared to MAE.
- **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, sharing the same unit as the target variable. It can be interpreted as the standard deviation of the errors.
- **R-squared (coefficient of determination):** R-squared measures the goodness-of-fit of the algorithm and represents the percentage of the dependent variable's variation that can be explained by the independent variables. It provides an indication of how well the model fits the data. Note that R-squared or Adjusted R-squared, an alternative that is unaffected by the number of features, may not perform well when non-linear relationships are present in the system, as they assume linear dependencies.

Cross-validation

Cross-validation is a technique used to estimate the performance of a model on unseen data. In regression problems, K-fold cross-validation is commonly employed. It involves dividing the data into K subsets or folds, training the model on K-1 folds, and evaluating it on the remaining fold. This process is repeated K times, and the results are averaged to estimate the model's overall performance. K-fold cross-validation helps detect overfitting and provides a reliable assessment of the model's ability to generalize to new data (Sammut and Webb, 2011).

Conclusion from evaluation metrics

We prioritize specific metrics for evaluating machine learning models to ensure a comprehensive assessment. The MSE takes precedence over the MAE due to its consideration of squared differences, which effectively captures outliers and robust performance. Given the potential presence of non-linear relationships in our regression problem, R-squared may

not provide significant insights. Moreover, RMSE, measured in the same unit as the target variable, is prioritized over all the measures since it offers accurate and interpretable results, accounting for outliers. Additionally, we employ k-fold cross-validation to compare and assess models, providing a clearer understanding of performance and mitigating overfitting risks. By considering these metrics and utilizing cross-validation, we ensure a robust evaluation of the models.

2.4 Literature Gap

The literature review conducted for this project revealed several gaps and discrepancies in the existing research related to enhancing the accuracy of baggage loading and unloading duration predictions at airports.

Firstly, there is uncertainty regarding the true impact of baggage handling operations on the overall turnaround time. While some authors suggest minimal influence (Schultz and Fricke, 2008), others emphasize its growing importance (Frey, 2014; Tarău et al., 2009), creating a gap in understanding the specific effect of baggage handling on turnaround efficiency.

Furthermore, there is a lack of clarity regarding the relationship between factors that impact the overall turnaround time and baggage handling duration, emphasizing the need for further investigation. Various authors propose different sets of factors, revealing a potential gap in the literature in terms of coherence and synchronization among factors impacting both processes. However, it is important to note that all of the proposed factors can still be utilized for analysis, and it is possible to identify the most influential factors from these varied perspectives.

Finally, a significant literature gap exists in the realm of data-driven methods for predicting baggage handling durations. While previous studies focus on optimizing the process (Tarău et al., 2009, 2010; Volt et al., 2022), none have specifically addressed prediction models for unloading and loading durations. Existing literature primarily concentrates on predicting the overall turnaround time or boarding-deboarding durations (Gao et al., 2015; Horstmeier and de Haan, 2001; Hutter et al., 2019; Luo et al., 2021; Schultz and Reitmann, 2019; Wang et al., 2022). However, considering that the baggage handling process constitutes a substantial portion of the turnaround time, accurately estimating its duration could greatly contribute to improved resource planning and allocation.

Chapter 3

From system analysis to data preparation

The literature review in Section 2.2 provided an overview of the baggage unloading and loading process globally. However, in the case of KLM, this process differs due to the presence of containerized and bulked-loaded flights. Containerized flights involve bags that are prearranged into a ULD (container or pallet) and directly unloaded/loaded from/into the aircraft using a high-loading device. On the other hand, bulked-loaded flights require bags to be brought in bulk and unloaded/loaded with the assistance of a belt-loading device.

KLM's fleet consists of wide-body, narrow-body, and regional jets. Wide-body aircraft, such as the Boeing 777 and Airbus A330, are designed for long-haul flights with high passenger capacity, requiring longer ground time and high-loading devices for baggage handling. Narrow-body aircraft, like the Boeing 737 and Airbus A320, are smaller planes for shorter flights, with shorter ground time and bulk baggage handling using belt loaders. Regional jets, including the Embraer E175, E190, and E195-2, have the shortest ground time and also use belt loaders for baggage handling.

The baggage handling process, as defined for this project, encompasses both unloading and loading activities. Unloading commences when the first baggage item is removed from the airplane and concludes when the last piece is placed onto the baggage cart. Loading, on the other hand, begins with the first piece being taken off the baggage cart and extends until the final piece is successfully loaded onto the aircraft.

The remainder of the chapter discusses the description of the current tool, its accuracy, and limitations in Section 3.1. Moreover, Section 3.2 describes the data extraction process. Later, Section 3.3 offers a brief overview of each dataset. Additionally, Section 3.4 presents the steps taken to prepare and transform the data for subsequent analysis and modeling. Afterward, Section 3.5 shows the preliminary data preparation and cleaning. In Section 3.6, initial insights from the dataset are provided. Finally, Section 3.7 marks the culmination of this chapter, providing a comprehensive conclusion to the discussed topics and findings.

3.1 Current Estimation Tool

KLM employs an integrated tool within its system to estimate baggage unloading and loading durations, with a focus on improving the variable aspects of the process. The tool takes into consideration factors such as the amount of dead load, the number of special handling items, and the number of handbags checked in at the gate. The estimation of baggage handling time consists of both variable and static components. The static components include durations for setup time, preparation time, and clean-up time, which remain constant. On the other hand, the variable components encompass the time required for unloading/loading the total dead load, hand luggage, and special cargo. The primary objective of the project is to enhance the estimation accuracy of the variable parts.

Moreover, it is crucial to acknowledge the limitations of the estimation tool. Firstly, the tool has been noted to consistently overestimate the actual time required for baggage handling, as observed by ground staff. This discrepancy could be attributed to the tool's reliance on assumptions and estimates, which may introduce inaccuracies in the predictions. Additionally, the tool assumes a steady flow of bags, disregarding real-world variations in baggage sizes, machine efficiency, and ground staff fatigue. Secondly, it struggles to handle unexpected events like flight cancellations or equipment failures that can significantly impact baggage handling. Additionally, the tool's limited feature set may overlook important factors, leading to less accurate predictions. It also struggles to capture nonlinear relationships, hindering its ability to model complex dependencies. Moreover, the tool lacks scalability and measurability due to the absence of data storage, limiting its capacity to learn, improve, and assess performance over time.

Furthermore, the lack of data storage in the current tool prevents the validation of its estimations, posing a significant challenge in understanding the effectiveness of the tool for addressing the main problem of this research. Without access to the necessary data, the validation process becomes uncertain. To address this limitation, an approximation of the current estimation tool is developed and can provide insights into its behavior and performance.

Approximation of the Current Tool

KLM's internal data analysis revealed valuable insights on the current tool's functionality and provided comprehensive information about unloading/loading speeds for different aircraft types. The estimation formula considers the unique characteristics of each aircraft type. For regional jets or narrow-body aircraft, the duration is estimated by dividing the total weight by the corresponding unloading speed in bags per minute. For wide-body aircraft, the duration is estimated by dividing the total number of bags by 40 (standard bags per ULD) and then by 0.20 (ULD unloading/loading speed per minute). An additional 4 minutes are allocated for each special item to accommodate potential delays.

$$\text{Estimated duration} = \begin{cases} \dfrac{\text{total weight}}{\text{(un)loading speed(aircraft type)}} + 4 \cdot \text{special items} & \text{If aircraft type is a regional jet or narrow-body} \\ \dfrac{\text{total bags}}{40} \cdot \dfrac{1}{0.2} + 4 \cdot \text{special items} & \text{If aircraft type is a wide-body} \end{cases}$$

It is essential to acknowledge that the accuracy of this approximation will be evaluated in subsequent stages, as the data collection and preparation process is currently ongoing. The assessment of accuracy will be documented in Chapter 6 (”Evaluation and Results”), where a comprehensive comparison between the proposed models and this approximation will be conducted.

3.2 Data Extraction

The data extraction process involved three data sources, namely DeepTurnaround, Platform Flight 720, and the Dutch Meteorological Institute (KNMI). These sources provide a comprehensive view of the flight turnaround process from both airport and airline perspectives, making it possible to analyze and optimize the process for better efficiency and passenger experience.

3.2.1 DeepTurnaround

DeepTurnaround is an initiative from Schiphol Airport that collects data about ground operations by locating cameras in the airport’s ramps. The cameras record events and timestamps of different operations such as baggage handling, catering, cleaning, and fueling. The data extracted from DeepTurnaround provides aggregated data for turnaround events of KLM aircraft at Schiphol airport, including timestamps, and some flight information. Two CSV files were extracted, one including data from August to November 2022, and the other with data from February and March 2023.

Moreover, the camera recordings capture the start and end points of the unloading and loading processes as follows: For unloading, the cameras initiate the recording when the first bags are detected leaving the plane and conclude when the cameras observe the last bags being loaded into the baggage carts. In contrast, for loading, the cameras begin recording when they detect the first bags being taken off the baggage carts, and they cease when the last bag is observed entering the plane.

3.2.2 Platform Flight 720

Flight 720 is a platform used by KLM to store and manage information related to their flights, including flight schedules, aircraft configuration, passenger information, and crew informa-

tion. This platform serves various purposes such as flight planning, revenue management, and customer service. It acts as a central repository for all flight-related information, enabling KLM to make real-time operational decisions based on the data collected. Flight 720 enables KLM to optimize its flight operations, improve customer experience, and enhance operational efficiency. Two CSV files were extracted, one including data from August to November 2022, and the other with data from February and March. Both files are concatenated to aid further examination.

The information in Flight 720 comes from various data sources, including ACARS (Aircraft Communications and Reporting System), which is a communication system linking aircraft to ground control. ACARS logs all system events generated by the aircraft before, during, and after a flight, including 'first cargo door open', 'last cargo door closed', 'norm baggage loading conform time', 'actual baggage loading started time', and others. This data is highly reliable as it is generated by the aircraft itself and can be used to check the reliability of the data from DeepTurnaround.

3.2.3 Weather data

The weather data was obtained from the Royal Dutch Meteorological Institute's (KNMI) official website (Koninklijk Nederlands Meteorologisch Instituut), which is a trustworthy source for weather information in the Netherlands. KNMI is responsible for collecting, processing, and archiving all-weather measurements in the Netherlands, and is considered to be one of the leading meteorological institutes in the world.

The official site of KNMI provides hourly measurements of various weather parameters, such as temperature, precipitation, wind speed, and cloud cover, among others. The data is validated and quality controlled, ensuring that it is accurate and reliable for use in research studies. For this project, hourly weather data was obtained for Schiphol Airport every day during the months of August to November and February to March.

The use of KNMI data for predicting the baggage unloading and loading duration at Schiphol Airport is a sound approach, as weather conditions can significantly impact ground handling operations at airports. Additionally, the KNMI data can help determine when cameras at the airport fail due to weather conditions, enabling us to ensure that the data is reliable. The KNMI data for Schiphol Airport covers the entire period for which the user has obtained data from turnarounds at the Schiphol cameras. This allows for the analysis of the impact of weather on baggage handling during different weather conditions and seasons.

3.3 Data Description

The goal of this stage is to identify any potential issues or limitations with the data and to determine whether it is suitable for analysis and modeling. Since weather data has already been cleaned up and processed, it will not be considered in this section.

3.3.1 DeepTurnaround

The concatenated DeepTurnaround dataset contains aggregated data for KLM turnarounds of inbound (arrival) flights and outbound (departure) flights recorded at Schiphol airport in the months of August to November 2022 and February and March 2023.

The DeepT dataset, consisting of 915,678 entries for a 6-month period, underwent thorough inspection to ensure data quality. Among the findings were duplicate values for the 'turnaround id' and a total of 10,479 unique turnaround keys. These keys represent a diverse range of baggage handling events, including aircraft arrival and departure, cargo door operations, and the start and end times of baggage unloading and loading activities.

To focus specifically on the analysis of baggage handling operations, the project emphasizes the actual unloading and loading processes. The recorded times of aircraft arrival and departure serve as reliable markers for the beginning and end of the turnaround process. Additionally, the accuracy of the DeepTurnaround data is compared with information obtained from the aircraft's door sensors through cargo door operations (opening and closing), as monitored by the ACARS system.

During the examination of the dataset, it became apparent that not all turnaround IDs had timestamps for every event, resulting in missing data. Filtering the dataset to isolate the desired turnaround events led to the identification of unique keys for specific activities. These included 10,450 keys for "AircraftAppears" and "AircraftDisappears," 9,771 keys for "BaxLoadingUnloadingStarts" and "BaxLoadingUnloadingStops," 9,423 keys for "FirstCargoDoorOpens," and 9,414 keys for "LastCargoDoorCloses."

The presence of duplicate values in the dataset arose from multiple timestamps associated with each event of the turnaround ID as shown in Figure 3.1. Notably, significant time gaps are observed between certain pairs of "BaxLoadingUnloadingStarts" and "BaxLoadingUnloadingStops" timestamps, indicating a shift from unloading to loading activities. To address these duplicates, a comprehensive data transformation process is implemented, as detailed in the data transformation section.

It is worth noting that the presence of duplicate and missing timestamps can be attributed to various factors, including camera flickering, network connectivity issues, software glitches, data entry errors, and technical challenges. Despite these obstacles, the cleaning and transformation of the data enable the utilization of the collected information from ramp cameras to enhance the estimation of baggage unloading and loading durations, as demonstrated in the data cleaning approach.

3.3.2 Flight 720

Flight 720 is a comprehensive dataset that provides information on all KLM flights. The data is collected from various sources, including the KLM reservation system, departure control system, baggage handling system, and the Aircraft Communications Addressing and Reporting System (ACARS). This data is then processed and contextualized in real-time to provide actionable insights for operational decision-making.

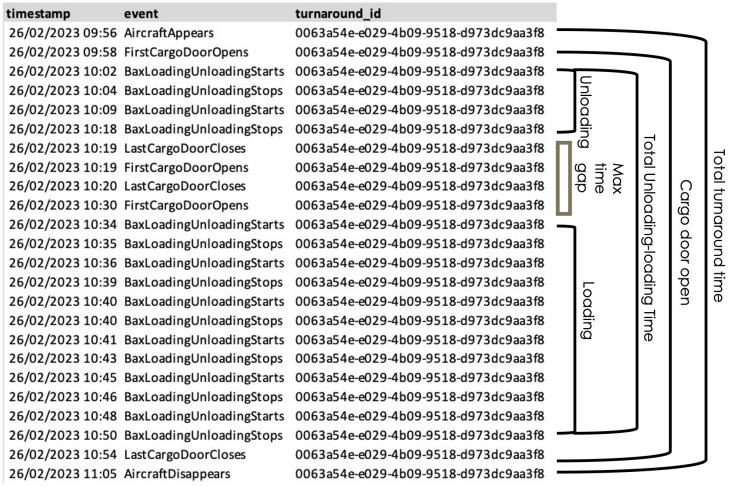


Figure 3.1: DeepTurnaround data example and approach

The dataset contains several categories of data, including ”Flight legs”, which provides information related to individual flights, such as flight numbers, dates, times, routes, and aircraft types. Other categories include ”Passenger data,” which contains information about individual passengers, and ”Baggage data,” which provides information about baggage, including the number of bags per passenger, weight, and tracking information.

For this project, the focus is on predicting the baggage unloading-loading duration, and the relevant information is contained in the ”Flight Legs” category. The concatenated CSV file contains only information in this category, with a total of 96,214 observations and 254 columns for the given six months.

Observations about the F720 Data

During the data collection process from the Flight 720 platform, the Flight Legs dataset provided several columns representing the number of bags. One column, labeled as the total number of check-in bags, was expected to represent the overall bag count for each flight. However, upon closer examination, summing the separate columns that captured the number of check-in bags for local, transfer, crew, and hand baggage did not yield a matching total. Additionally, the dataset included two columns named ”total bags loaded” and ”total bags not yet loaded,” which, when combined, were expected to provide the total number of bags. However, this approach did not consistently yield the expected total of check-in bags.

To address these inconsistencies, the separate variables representing the total number of check-in bags were aggregated and used as the definitive measure of the total number of bags. It is important to note that a complementary dataset called Flight Leg Loads was obtained from Flight 720, which provided total load information for KLM flights. To calculate the total number of bags, the ”deadload:number of items” column in the Flight Leg Loads dataset was summed only for the Baggage and CrewBaggage types, considering entries with an ACTIVE and Final status. While this data appeared similar to the previously discovered information, there were instances where the number of bags in the Flight Leg Loads

dataset exceeded the findings from the Flight Leg dataset. These discrepancies might be attributed to potential issues during the retrieval of the dead load data.

Due to time constraints, the variable derived from adding the separate types of check-in bags in the Flight Leg dataset is used, with the exception of the ”number of special items” column, which is exclusively provided by the Flight Leg Loads dataset. It is worth mentioning that the number of bags derived from the Flight Leg dataset demonstrates stronger explanatory power for the durations compared to the column obtained from the Flight Leg Loads dataset.

A similar situation arises concerning the weight information. The Flight Leg dataset only includes information about the total check-in weight, whereas separate check-in weight details for local and transfer bags are available, but no specific data is provided for crew and hand baggage. Consequently, the decision is made to utilize the total check-in weight for analysis purposes. Furthermore, the total weight obtained from the Flight Leg Loads dataset is not considered due to uncertainties regarding the accuracy of the data retrieval process for the number of bags.

In summary, after careful evaluation and consideration, the variables derived from the Flight Leg dataset are chosen as the primary source of information, with the inclusion of the ”number of special items” column from the Flight Leg Loads dataset. These variables demonstrate superior relevance and explanatory power for the duration analysis.

3.4 Data Transformation

The Data Transformation section pertains to the conversion of raw data into a more suitable form for analysis. This process was predominantly applied to the DeepTurnaround dataset, and the steps taken are depicted in figure 3.2. Additionally, the final subsection will discuss the integration of the newly transformed DeepTurnaround dataset with the flight-720 dataset.

DeepTurnaround

In order to prepare the DeepTurnaround dataset for analysis, a series of data transformations were applied. These transformations were designed to clean, pre-process and reformat the data in a way that facilitates analysis and modeling. Figure 3.2 provides a visual representation of all the transformations applied to the DeepTurnaround dataset, which will be explained in further detail in the following subsections.

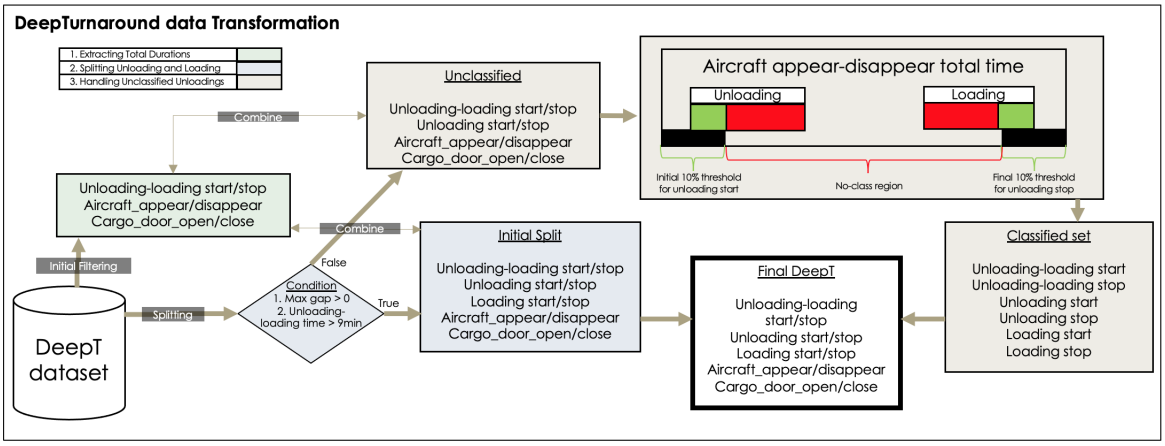


Figure 3.2: DeepTurnaround data transformation process, designed by the author

Extracting Total Durations

Previously in Section 3.3, it was observed that there were numerous start-stop, open-close, and appear-disappear timestamps for each turnaround id. To address this issue, the earliest start timestamp and the latest stop timestamp were selected for each turnaround id by grouping the start and stop timestamps, sorting them in ascending and descending order, and selecting the first timestamp from each list. As a result, the new further-processed dataset contains 9'771 data entries with unique turnaround ids, the previous columns, as well as the new timestamp formatted columns "unload-load start", "unload-load stop", "FirstCargoDoorOpens", "LastCargoDoorCloses", "AircraftAppears", and "AircraftDisappears".

Splitting Unloading and Loading Activities

Upon examining the raw data, a significant time gap between multiple "BaxLoadingUnloadingStarts" and "BaxLoadingUnloadingStops" timestamps was observed, indicating the transition from unloading to loading activities. To extract the necessary information and split the activities, these events were grouped based on their unique turnaround id, and within each group, the earliest start and latest stop times were selected. The maximum gap between consecutive start-stop pairs of timestamps was calculated, and if it exceeded X minutes and the total unloading-loading time was greater than Y minutes, the activity was divided into two.

To determine the minimum total unloading-loading time (Y), an interview was conducted with a ground operations expert, who stated that a regional jet's baggage process requires a minimum of around 10 minutes. From a practical perspective, the process captured by the airport's cameras involves activities from the initial unloading phase, when the first bags are observed leaving the plane, to the final loading phase, when the last bag is loaded into the aircraft. These activities include tasks such as closing doors, repositioning equipment, and opening doors during the transition period. This process can be assumed to take approximately 6 minutes, excluding bag handling and setup time, which adds can add additional 3

minutes in the case of the minimal number of bags. Taking these factors into account, the estimated minimum total unload-load time (Y) is 9 minutes. Furthermore, it is reasonable to assume that the greatest time gap between consecutive start-stop pairs (X) is greater than zero, accounting for fast turnarounds. In addition, the distribution of time gaps between unloading-loading activities in Figure 3.3 confirms this assumption, with significant values starting at around 1 minute.

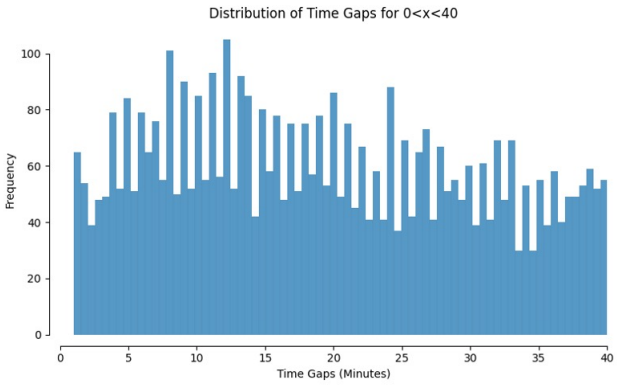


Figure 3.3: Distribution of the ”time gaps” column

The transformation process will generate five new columns: ’unload start’, ’unload stop’, ’load start’, ’load stop’, and ’time gaps’. These columns will contain the start and stop timestamps for unloading/loading, as well as the maximum gap between consecutive start-stop pairs. In cases where the maximum time gap or total duration didn’t meet the conditions, the first start and last stop timestamps were assigned to unloading, while loading had null values. If the number of ”BaxLoadingUnloadingStarts” events didn’t match the corresponding ”BaxLoadingUnloadingStops” events within a turnaround, both activities were assigned null values.

Out of 9,771 analyzed turnarounds, 8,116 met the conditions and were divided into separate unloading and loading events. Additionally, 9,761 turnarounds were classified as unloading events, even if they didn’t meet the splitting criteria. However, 10 turnarounds had discrepancies in the number of start and stop events, resulting in null values to indicate data inconsistency.

Handling Unclassified Unloading Activities

To improve the dataset, we merged the recently created dataset, which included separate unloading and loading activities, with the existing dataset that contained information on timestamps for the unloading-loading duration, door opening time, and aircraft appearance time. However, there were still unclassified unloading events that didn’t meet the previous criteria for splitting. These unclassified events might represent loading actions, emphasizing the need for additional analysis and categorization to assign them appropriately.

To address this, we used the variables ’AircraftAppears’ and ’AircraftDisappears’ as markers for the beginning and completion of the turnaround process. We compared the

start and end times of unclassified unloading events with these variables to determine their classification. A rule was established to classify undefined unloading events as actual unloading events if the start timestamp fell within the initial 10 percent of the total turnaround time and within five minutes of the observed airplane doors opening time. Similarly, undefined unloading events were classified as loading events if the stop timestamp fell within the last 10 percent of the total turnaround time and within five minutes of the observed airplane doors closing time. Any unclassified unloading events that didn't meet these criteria were assigned null values for both unloading and loading events.

Finally, we calculated the time differences between various start-stop, appear-disappear, open-close, and other relevant variables. These differences were transformed into decimal numbers representing minutes, ensuring a standardized and consistent representation of time throughout the dataset

3.4.1 DeepTurnaround, Flight-720, and Weather Data Integration

In order to integrate the DeepTurnaround dataset and the F-720 dataset, several steps were taken. Firstly, a filtering process was applied to the DeepT flight numbers to ensure that only inbound and outbound flights within the ranges of 400-899 (flight numbers for Intercontinental flights) and 900-1999 (flight numbers for European flights) were included. This reduced the dataset size from 9771 to 9601 unique rows.

Since the datasets had different unique identifiers, a solution was found by creating flight keys for both inbound and outbound flights in the DeepTurnaround data. A dictionary was created to match flight numbers with their corresponding departure and arrival airports, resulting in consistent flight keys across the datasets. Moreover, during the matching process, it was discovered that 167 inbound keys were missing from the F-720 dataset, while 236 instances involved different aircraft types, indicating incomplete turnarounds. These instances were excluded from further analysis to maintain accuracy.

Furthermore, the integration of the hourly weather data into the dataset followed a two-step approach: First, the weather variables corresponding to the hour of the scheduled arrival time of the inbound flight were retrieved. Then, the weather variables for the hour of the scheduled departure time of the outbound flight were obtained. By averaging these two sets of weather variables, the representative weather conditions for the entire duration of the turnaround process were obtained. This choice ensures that the relevant data will be accessible at the time of prediction.

The final integrated dataset can be observed in Table 3.1 and consists of 9,197 unique entries and serves as the foundation for predicting the duration of baggage unloading and loading operations, as well as their individual durations.

Attribute	*Description*	*Data type*
Timestamps		
scheduled_arrival_time_inbound (time,dd,mm,yy)	Scheduled arrival date of incoming flights at Schiphol airport	Date-time
Durations		
total_time_minutes_DeepT	Total time spent during the turnaround in minutes recorded by the airports cameras	Float (in minutes)
unload_total_minutes_DeepT	Total time spent unloading during the turnaround in minutes recorded by the airports cameras	Float (in minutes)
load_total_minutes_DeepT	Total time spent loading during the turnaround in minutes recorded by the airports cameras	Float (in minutes)
door_open_minutes_DeepT	Total time spent with the aircraft's cargo doors open in minutes recorded by the airports cameras	Float (in minutes)
door_open_minutes_ACARS	Total time spent with the aircraft's cargo doors open in minutes recorded by the aircraft sensors	Float (in minutes)
Flight Information		
flight_number_inbound/outbound	Flight number for inbound flight	integer
total_pax_inbound/outbound	Total number of accepted passengers	Integer
total_transfer_pax_inbound/outbound	Total number of accepted transfer passengers	Integer
load_factor_inbound/outbound	Total number of passengers boarded divided by total available seats	Float
total_bax_inbound/outbound	Total number of bags for the incoming or outgoing flight	Integer
total_bax_weight	Total weight of the bags	Float
total_handbags_inbound/outbound	Total number of hand bags	Integer
ramp	Ramp name where the aircraft was parked during the turnaround (e.g A1)	String
aircraft_type	Type of aircraft	String
aircraft_group	Group of aircraft type (e.g wide-body, narrow-body)	String
departure/arrival_airport_inbound/outbound	Departure airport for the incoming flight and arrival airport for the outgoing flight	String
departure/arrival_gate_inbound/outbound	Departure/arrival gate for incoming or outgoing flights	String
departure/arrival_parking_stand_inbound/outbound	The parking stand where incoming aircraft arrive to and where outgoing aircraft leave from.	String
departure/arrival_terminal_inbound/outbound	The terminal where incoming aircraft arrive to and where outgoing aircraft leave from.	String
flight_cancelled	Indicates if the flight was cancelled	Boolean
is_quick_turnaround	Indicator whether or not a quick turnaround should be applied	Boolean
scheduled_flight_duration_inbound/outbound	The difference between the scheduled arrival time and the scheduled departure time in minutes	Timedelta
aircraft_type	Type of aircraft	String
Weather Data		
WindSpeed	Hourly Wind Speed (in m/s)	Float
Temperature	Hourly Temperature (in Celsius)	Float
SunshineDuration	Duration of Sunshine per hour (in minutes)	Float
Radiation	Hourly Radiation (in J/cm2)	Float
HourlyPrecipitation	Hourly Precipitation (in mm)	Float
AirPressure	Air Pressure (in hPa)	Float
Humidity	Humidity (in percentage)	Float
Mist	Mist or Fog (0=not occurred, 1=occurred)	Boolean
Rain	Rain (0=not occurred, 1=occurred)	Boolean
Snow	Snow (0=not occurred, 1=occurred)	Boolean
Thunderstorm	Thunderstorm (0=not occurred, 1=occurred)	Boolean
IceFormation	Ice Formation (0=not occurred, 1=occurred)	Boolean

Table 3.1: Integrated dataset

3.5 Data Cleaning

This section will present the steps taken to clean the transformed and newly integrated datasets obtained from the DeepTurnaround and Flight-720.

3.5.1 Reliability Analysis and Data Cleaning

Reliable data is vital for making well-informed decisions. In this section, the reliability of the DeepTurnaround data is evaluated by comparing it with highly reliable ACARS aircraft sensor data. It's important to note that the ACARS data captures timestamps for cargo door opening and closing, but lacks specific information about baggage unloading or loading. To assess the reliability of the DeepTurnaround data, the timestamps for cargo door open-close events are used and it is assumed that their reliability extends to other events such as baggage unloading and loading. Additionally, the time differences between the open and close timestamps for each data source have been calculated and converted into decimal minutes. It's worth mentioning that these timestamps are solely used for evaluating data reliability and are not employed in subsequent stages.

To ensure reliability and accuracy, we filtered out outlier values in our analysis. We set minimum and maximum thresholds of 15 minutes and 600 minutes, respectively, for door opening durations based on observed distribution patterns. Out of the total 9,197 values, 6,997 fell within this range and were considered for analysis. It's worth noting that 960 ACARS values and 1,396 DeepTurnaround values fell outside this range, potentially representing different data entries. Although ACARS data can be influenced by factors like overnight turnarounds, long delays, and exceptional circumstances, the substantial number of outliers raises concerns about its reliability. However, given the study's context, ACARS data remains the most trustworthy source to identify reliable data from DeepTurnaround.

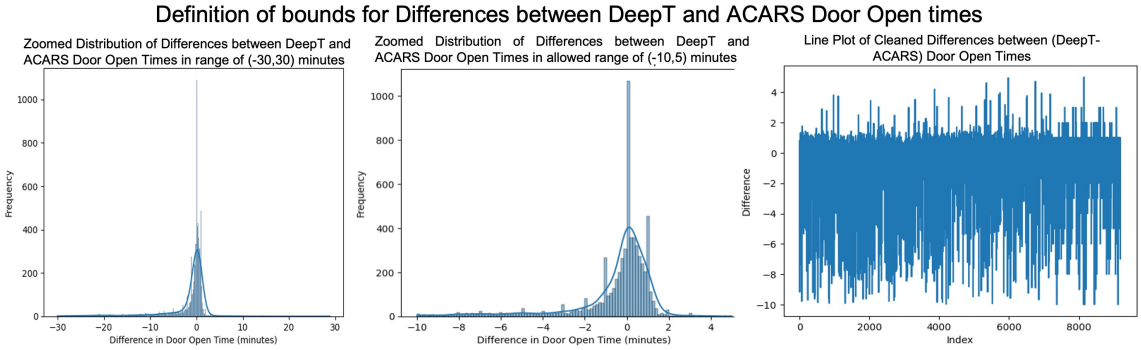
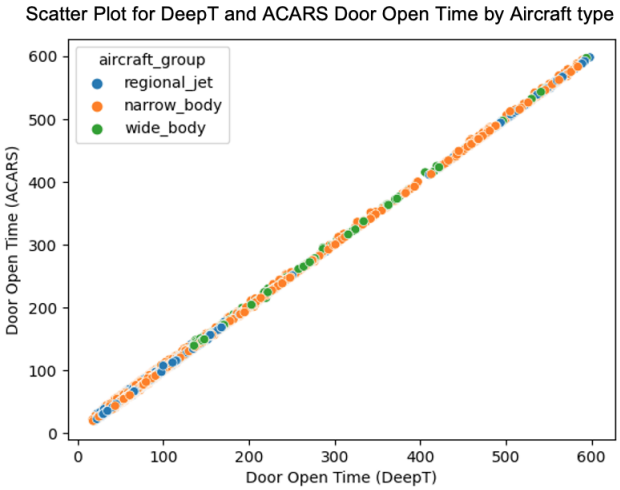


Figure 3.4: Definition of Bounds for DeepT-ACARS Differences

To determine the reliable range of data for DeepTurnaround, we calculated the residuals between the recorded door opening times in DeepTurnaround and ACARS. The majority of these residuals, as shown in the left-hand graph of Figure 3.4, fall within the range of -10 to 5 minutes. However, we identified 736 cases that lie outside this range and are therefore excluded from further analysis. By using this range, we ensure the retention of 6261 data

points, which represent a subset of reliable DeepTurnaround data. This range takes into account potential scenarios where the camera might experience minor delays in capturing the opening of the cargo doors or record the closing time with slight deviations compared to the immediate readings from the aircraft sensors. To address potential measurement errors, each scenario is assigned a buffer of 5 minutes. Consequently, the reliability of the DeepTurnaround data is supported by the similarities observed between the DeepTurnaround and ACARS door opening times, as depicted in Figure 3.5. This similarity further substantiates the reliability and accuracy of the DeepTurnaround data.



Statistical Discription of DeepT and ACARS Door Open Times

	Door Open Time DeepT	Door Open Time ACARS
count	6261.00	6261.00
mean	108.20	108.73
std	102.32	102.39
min	17.47	19.00
25%	57.52	58.00
50%	75.97	77.00
75%	110.00	110.00
max	598.00	598.00

Figure 3.5: Final comparison between DeepT-ACARS cargo door opening time

Moreover, an additional analysis was conducted to gain insights into the underlying causes of the outlier and extreme differences observed in the data. Specifically, the investigation focused on examining the potential influence of weather conditions on the reliability of airport cameras. Figure 3.6 provides a comprehensive overview of various weather conditions, including rain, mist, thunderstorm, snow, and ice formation. Upon thorough examination, a significant finding emerged. It became apparent that weather conditions do not have a clear impact on the functionality of the airport cameras, with the exception of the rain, which exhibited a discernible, albeit moderate, influence.

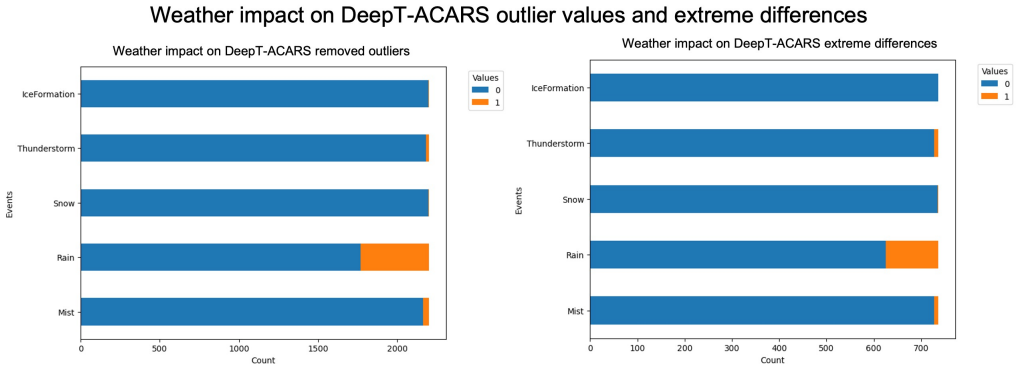


Figure 3.6: Weather Impact on DeepT-ACARS Outliers and Extreme Differences

3.5.2 Outlier Detection

Separate analyses are conducted for baggage unloading and loading durations in each group of aircraft types to establish a feasible range and remove outliers. The total unloading-loading duration, including the unpredictable time gap between activities, is not considered due to limited explanatory features. KLM's fleet comprises Wide-body, Narrow-body, and Regional jets, categorized in a provided dictionary (Figure 3.2). By filtering the dataset for each duration separately, utilizing histograms and box plots, two new filtered datasets are generated. The maximum bound for both unloading and loading durations is determined by diminishing significance in duration distribution, while the minimum bound is derived from observed bags after zero, accounting for weight and KLM's standard speed.

Aircraft Type Dictionary	
Group	Aircraft Type
Wide-body	Airbus A330-200 (332)
	Airbus A330-300 (333)
	Boeing 777-200 (772)
	Boeing 777-300ER (77W)
	Boeing 787-9 (789)
	Boeing 787-10 (781)
Narrow body	Boeing 737-700 (73W)
	Boeing 737-800 (73H)
	Boeing 737-900 (73J)
Regional jets	Embraer ERJ-175 (E75)
	Embraer ERJ-190 (E90)
	Embraer E195-2 (295)

Table 3.2: Description of the Aircraft type dictionary

Figure 3.7 depicts the initial distributions of durations categorized by aircraft type, utilizing both box plots and histograms.

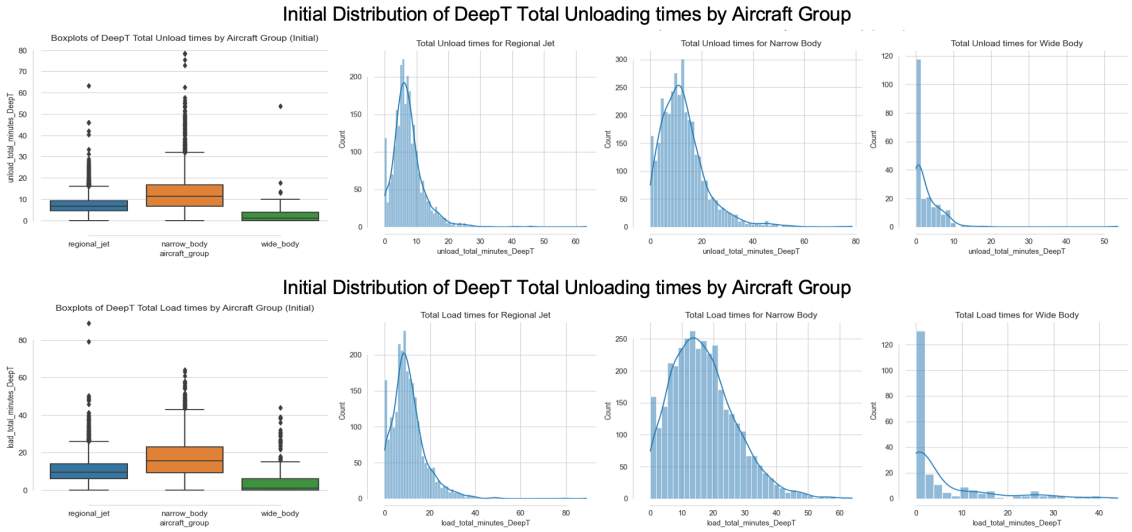


Figure 3.7: Initial Distributions of Durations Classified by Aircraft Group

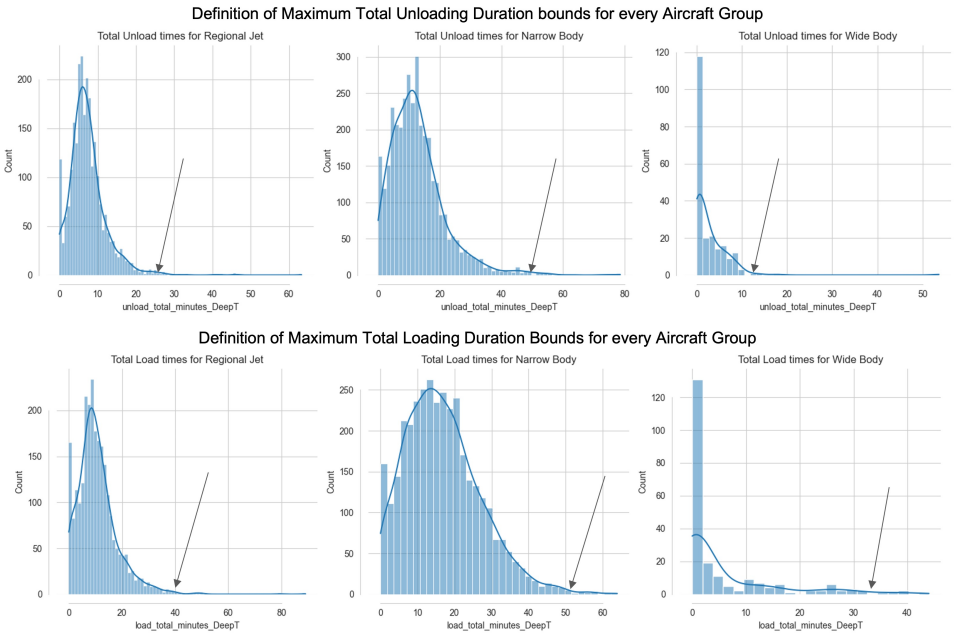


Table 3.3: Definition of Maximum Bounds for Unloading and Loading Separate Durations

Bounds Definition

For unloading durations, minimum bounds were established based on the number and weight of bags across different aircraft groups. Regional jets required a minimum of 5 bags weighing 69kg, taking approximately 0.75 minutes at KLM's standard speed of 89 kg per minute. Narrow-body planes had a minimum of 8 bags weighing 121kg, requiring around 1.35 minutes. Wide-body planes need a minimum of 76 bags, equivalent to one ULD, taking approximately 5 minutes.

Similarly, loading durations were determined to be 0.5 minutes for two bags weighing

46kg in Regional jets, 1.1 minutes for six bags totaling 96kg in Narrow-body planes, and 5 minutes for 66 bags or one ULD in Wide-body planes.

Furthermore, establishing the maximum bounds for the two separate durations in each aircraft group will involve a careful analysis of the histogram's tail end, as shown in Figure 3.3.

Resulting Datasets

The resulting filtering conditions and their corresponding values for each separate duration can be found in Figure 3.8.

Bounds for Total Unload Duration				
Aircraft type Group	**Min Bound**	**Max Bound**	**Min Bags**	**Remaining Data Entries**
Regional jet	0.75 min	25 min	5 bags (69kg)	2,346 entries
Narrow body	1.35 min	50 min	8 bags (121kg)	3,373 entries
Wide body	5 min	12 min	76 bags	44 entries

Bounds for Total Load Duration				
Aircraft type Group	**Min Bound**	**Max Bound**	**Min bags**	**Remaining Data Entries**
Regional jet	0.5 min	40 min	2 bags (46kg)	2,334 entries
Narrow body	1.1 min	51 min	6 bags (96kg)	3,410 entries
Wide body	5 min	34 min	66 bags	58 entries

Figure 3.8: Description of Conditions for Outlier Removal

In conclusion, the comprehensive examination of the distributions, supported by the corresponding histograms and the minimum number of bags, provides a comprehensive understanding of the unloading, loading, and total durations across different aircraft groups. The final distributions, along with their respective box plots, are visually presented in Figure 3.9, encapsulating the outcomes of the analysis and laying the groundwork for further insights into the baggage handling process.

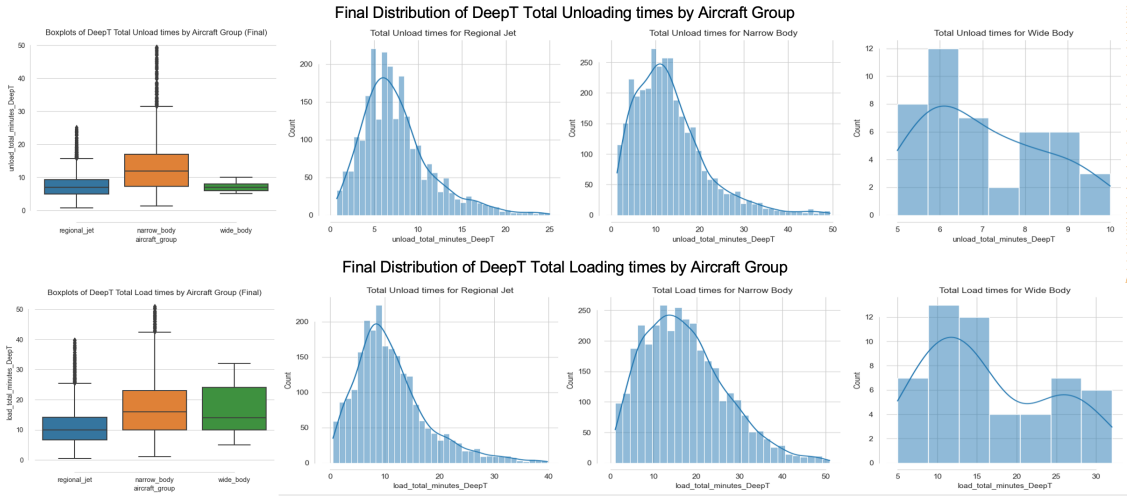


Figure 3.9: Distribution of target Durations After Outlier Removal

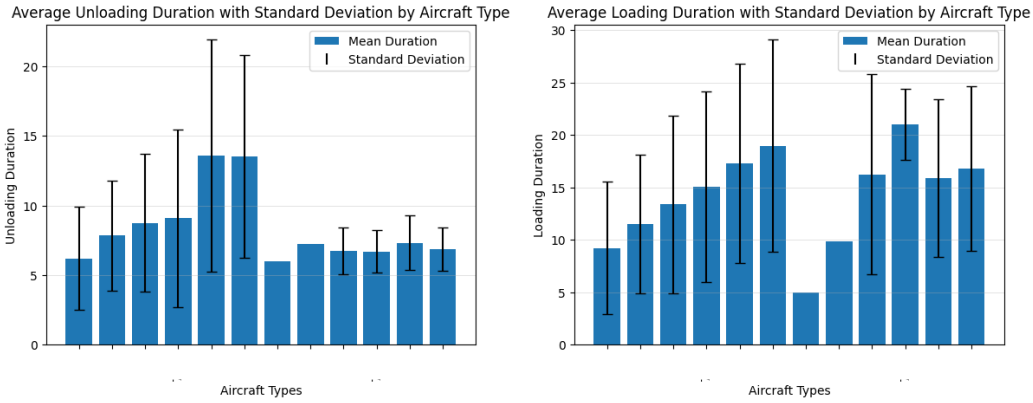


Figure 3.10: Average Durations Across Aircraft Types

3.6 Data Insights

This section provides insights into average durations for unloading and loading activities across different subsets of data. These insights inform decision-making and have the potential to improve estimation accuracy. We analyze the dataset to determine average durations for subsets such as aircraft types and combinations of aircraft types and groups of bags. Additionally, other combination of features we explored but did not really yield interesting insights. These averages serve as benchmarks for current and future estimation tools. Additionally, we derive average speeds from the data to enhance the current estimation tool. Subsequent subsections present key findings and insights.

3.6.1 Average Durations Across Aircraft Types

Figure 3.10 shows average unloading and loading durations for each aircraft type. Even though narrow-body planes have a larger quantity of data available, they have the highest variability, followed by regional jets and wide-body planes, please refer to Table 3.2 for a reference of the aircraft types. Additionally, despite high standard deviation, these average durations provide initial estimates for planners, enabling data-driven decision-making rather than reliance on assumptions.

3.6.2 Average Durations Across Aircraft Types and Number of Bags

In Figure 3.11, the data is segmented into bins based on the number of bags, and the average unloading/loading durations and variability are shown for each aircraft type and bin. This analysis serves as another benchmark model for planners. Generally, the duration increases with the number of bags. However, for the first two aircraft types, the 110-120 bags bin shows a significantly smaller average duration. This inconsistency may be due to a single data exception in this particular bin, which should not significantly impact the overall estimate. Therefore, it is recommended to use the average duration from the previous bin of the number of bags for these cases. Despite inconsistencies, these averages seem to

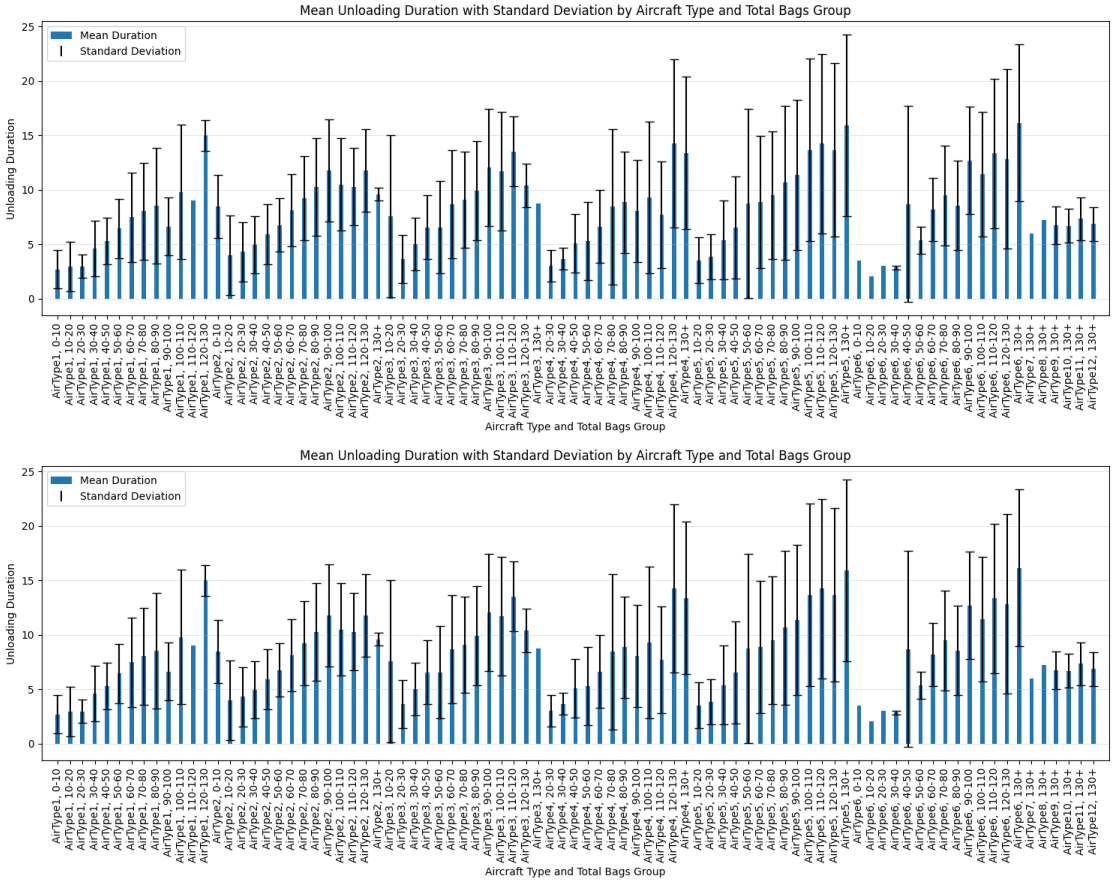


Figure 3.11: Average Durations Across Aircraft Types and Bag Groups

capture the behavior of the durations in a compelling manner. Therefore, it has been decided to adopt these averages as a simple model, assuming that planning users will utilize these average durations. This allows for straightforward comparison with more complex prediction models.

3.6.3 Data-driven Current Tool

By leveraging the available data, average unloading and loading speeds can be integrated into the analysis. Although the specific average speeds cannot be disclosed due to confidentiality reasons, they demonstrate a notable difference compared to the assumed loading speeds by KLM. Furthermore, the dataset lacks information regarding the unloading or loading of special items, making it challenging to precisely estimate their impact. Consequently, this data-driven current tool approximation is to be used as another benchmark for comparing the future Machine Learning models

3.7 Conclusion

In conclusion, this chapter has provided a comprehensive overview of the current tool, its limitations, and an approximation formula to this is provided in Section 3.1. The data extraction process was described in Section 3.2, which involved obtaining data from sources such as DeepTurnaround, Flight 720, and the Royal Dutch Meteorological Institute (Koninklijk Nederlands Meteorologisch Instituut). Section 3.3 provided an overview of each dataset and the extracted columns. The steps taken to prepare and transform the data for subsequent analysis and modeling were presented in Section 3.4, with a focus on DeepTurnaround data and the classification of the separate unloading and loading durations, in addition to the transformation of the other timestamped data columns. Section 3.5 highlighted the preliminary data preparation and cleaning, including the assessment of DeepTurnaround data reliability by comparing with data from ACARS data from the aircraft sensors, and the determination of minimum and maximum bounds for outlier cleaning. Finally Section 3.6 provides preliminary insights from the dataset and suggests two data-driven models that can be used as a baseline for future comparisons with more complex prediction models. Overall, the found final datasets provide a solid framework for the upcoming chapters, which will include additional analysis and exploration.

Chapter 4

Feature Engineering

”Garbage in, garbage out” is a widely recognized adage within the machine learning community, underscoring the significance of high-quality input data for generating accurate and reliable output. Driven by this principle, the central objective of this chapter is to meticulously prioritize the selection, preparation, and comprehensive exploration of the most pertinent features.

In order to achieve this objective, Section 4.1 explores the crucial task of feature engineering, focusing on selecting available features at the prediction stage and deriving new features to enhance their quality. Section 4.2 addresses multicollinearity, reducing redundancy and improving model stability. Visual analysis of feature-target relationships is presented in Section 4.3 to assess their impact. Section 4.4 applies preprocessing techniques to ensure optimal format and range for prediction models. Lastly, Section 4.5 employs feature selection models to further reduce complexity, enhance interpretability, and improve prediction accuracy. By curating and selecting pertinent features, this chapter sets the stage for optimal performance and insightful analysis in the subsequent modeling phase.

	Unloading-Duration	Loading-Duration
Variables	**Naming for features**	
target duration	unloading_minutes	loading_minutes
total_bax total_hand_bax total_weight total_pax total_transfer_pax load_factor scheduled_flight_duration number_special_items	_inbound	_outbound
aircraft_group aircraft_type departure_airport_inbound arrival_airport_outbound month (scheduled arrival time) hour (scheduled arrival time) day (scheduled arrival time) WindSpeed Temperature SunshineDuration Radiation HourlyPrecipitation AirPressure Humidity Rain IceFormation Snow Thunderstorm Mist		

Table 4.1: Starting set of variables for each duration

4.1 Feature Selection and Derivation

A data cleaning process was conducted prior to variable selection to address variables with a significant number of missing values. Variables with missing values exceeding 20 percent were excluded from the analysis. The selected variables needed to be available within a specific time frame, at least one hour prior to departure and possibly as early as three hours before departure. The resulting list of starting variables is shown in Table 4.1.

4.1.1 Deriving New Features

When it comes to the categorical derived features, one of the key derived features that have greatly contributed to the analysis is the aircraft group. It involves categorizing different aircraft types into wide-body, narrow-body, and regional-jet planes. This feature has been instrumental in filtering, preprocessing, and cleaning the information.

In addition to the aircraft group, another valuable derived categorical feature is the inbound or outbound departure continent. This feature provides a more concise representation of the data compared to individual departure or arrival airports. Please refer to Table 4.2 for its description.

Moreover, the derivation of continuous features is intended to potentially enhance explanatory power or mitigate multicollinearity issues. It is necessary to test the hypothesis that these features will improve the understanding of the baggage handling process. Only the derived features that validate the hypotheses should be retained and considered for

Derived Data		
Derived attributes	***Description***	***Data type***
continent_inbound/outbound	Type of flight depending on the flight number (EU[900,1999], MENA[400,499], AF[500,599], NAM[600,699] LATAM[700,799], AS&OC[800,899]) for inbound/outbound flights. It provides a more concise representation of the data compared to individual departure or arrival	String
max_possible_bax	the maximum number of bags identified in the dataset for each aircraft type. It helps in understanding the baggage capacity of different aircraft types.	Integer
avg_bags_per_passenger _inbound/outbound	total number of passengers divided by total number of bags. It provides insights into the baggage volume per passenger during the unloading-loading process.	Float
avg_bag_weight_per_passeng er_inbound/outbound	total number of passengers divided by total weight of bags. It helps in evaluating the average weight of baggage carried by passengers.	Float
baggage_load_factor _inbound/outbound	Total number of bags divided by max_possible_bax. It provides insights into the utilization of the aircraft's baggage capacity during the unloading-loading process.	Float

Table 4.2: Derived features

further analysis. For the description of the continuous features, please refer to Table 4.2.

4.1.2 Scaling Numerical Data

The min-max scaling approach will be employed to scale the numerical features, transforming them to a range of 0 to 1. This approach prevents any individual variable from dominating the analysis due to its large scale and preserves the interpretability and original values of the data. It also facilitates data comparison and interpretation by ensuring consistent scales across all features. While the standard scaler is another recommended method that assumes a normal distribution, the MinMax Scaler is chosen to address non-normal distributions and ensure fair comparisons without bias from varying feature scales.

4.2 Multicollinearity Cleaning

4.2.1 Dealing with Multicollinearity for Continuous Explanatory Features

To address multicollinearity, variables with direct relationships were evaluated, retaining the one with the highest correlation coefficient with the target durations and eliminating others. The Variance Inflation Factor (VIF) was used to assess variance inflation due to multicollinearity and remove variables contributing to it.

In the unloading and loading datasets, the highly influential variables ”total number of bags” and ”total weight of bags” had a significant correlation of 0.99. To address multicollinearity, only the ”total number of bags” was included in the prediction model. Other variables strongly correlated with the number of bags were considered redundant and excluded, except for the ”Scheduled flight duration” due to its high explanatory power. Similarly, radiation was chosen over correlated weather variables (such as sunshine duration and humidity) for its stronger explanatory power in relation to the target durations. Additionally, among highly correlated derived features, the number of bags provided sufficient explanatory power, rendering other bag-related variables unnecessary. Only the average weight per passenger, encompassing both bag and passenger weight, was included due to its potential influence on prediction.

Unloading-duration			Loading-duration		
Continuous features	Corr. Coeff.	VIF Value	Continuous features	Corr. Coeff.	VIF Value
total_bax_inbound	0.47	8.87	total_bax_outbound	0.48	7.43
weight_per_passenger_inbound	0.28	7.54	weight_per_passenger_outbound	0.29	9.10
load_factor_inbound	0.17	7.78	load_factor_outbound	0.32	9.00
scheduled_flight_duration_inbound_min	0.06	3.43	scheduled_flight_duration_outbound_min	0.06	2.80
Radiation	0.05	1.88	Radiation	0.09	1.94
			special_cargo:number_of_items_outbound	0.06	1.04

Table 4.3: Final set of Continuous Variables for each Dataset

Variables with correlation coefficients below 0.05 with the target durations were removed as they indicated weak or non-significant relationships. In the unloading-duration dataset, variables such as hourly precipitation, visibility, air pressure, wind speed, average bag weight, and temperature fell below this threshold. Similarly, in the load-duration dataset, all variables mentioned, except the number of special items, did not meet the correlation coefficient threshold. Finally, the remaining variables were assessed for multicollinearity using VIF, eliminating those with high VIF values and low explanatory power through an iterative process.

Note that while some level of multicollinearity is retained for models capable of handling it, feature selection techniques discussed in the next section can aid in identifying the appropriate variable set for models unable to handle multicollinearity. As a result, the final list of continuous variables with their correlation coefficients and VIF values can be observed in Table 4.3.

4.2.2 Dealing with Multicollinearity for Categorical Explanatory Features

In this subsection, the analysis focuses on addressing multicollinearity in categorical variables and their subsequent removal. Two measures are employed to assess explanatory power: the two-sided t-test for categorical variables with two sub-categories and the F-statistic from the one-way ANOVA test for variables with more than two sub-categories. These measures quantify the impact on the target variable. Multicollinearity among categorical variables is evaluated using Goodman-Kruskal's lambda coefficient, which provides insights into causality. Visualizing relationships between categorical variables is facilitated through the use of heat maps, as demonstrated in Appendix A.

To eliminate correlated variables, the lambda association factor and explanatory power are considered. "Aircraft type" is excluded in favor of the stronger explanatory power of the derived variable "aircraft group". Despite a relatively high lambda measure between "aircraft group" and "continent inbound", both variables are retained due to their predictive capability. Variables with a statistic value below 1, indicating a lack of significant influence, are removed. In the unloading-duration dataset, "Mist" and "Thunderstorm" are removed, while in the loading-duration dataset, all binary weather variables, except "Rain", are excluded. Table 4.4 presents the resulting categorical features and their statistical measures of relationship with the target durations.

Unloading-duration		Loading-duration	
Categorical Features	**Respective Statistic**	**Categorical Features**	**Respective Statistic**
aircraft_group	483.43 (f-stat)	aircraft_group	354.3 (f-stat)
hour	18.89 (f-stat)	month	18.9 (f-stat)
month	9.8 (f-stat)	hour	13.8 (f-stat)
day	7.27 (f-stat)	day	4.1 (f-stat)
continent_inbound	4.41 (f-stat)	Rain	2.71
Rain	1.96	continent_inbound	1.21 (f-stat)
IceFormation	1.7		
Snow	1.35		
Mist	1		

Table 4.4: Final set of Categorical Variables for each Dataset

4.2.3 Insights from Multicollinearity Cleaning

The multicollinearity analysis provides key insights into variable influences on the durations. Among continuous variables, the number of bags has the greatest influence, followed by the weight per passenger and load factor. Interestingly, the outbound load factor has a greater impact on loading duration compared to the inbound load factor's influence on unloading duration. Scheduled flight duration and weather-related variables have minimal influence. Surprisingly, the number of special items has minimal to no significant influence on the target durations, particularly for loading duration and very low influence on unloading duration.

In terms of categorical variables, aircraft group is the most influential feature, while continents inbound and outbound have relatively small impacts. Time-related variables, such as hour and month, have stronger influence than the day variable. Weather variables demonstrate minor influence on the durations.

4.3 Exploratory Data Analysis

This section delves into the Exploratory Data Analysis (EDA) for each dataset, aiming to gain insights and uncover patterns that can inform the predictive modeling process. This analysis focuses on examining the relationships between the explanatory variables previously found and the target variables of interest.

4.3.1 Exploring the Impact of Continuous Numerical Features

This subsection visualizes the selected continuous variables in relation to the target durations. Scatter plots with regression lines are utilized to observe trends and correlations between these variables and the target durations. It is important to note that the numerical columns used for this analysis remain unscaled to preserve the real values and ensure accurate interpretation of the results. Scaling the data had minimal or no impact on the visual representation of the variables in the scatter plots.

Number of Bags

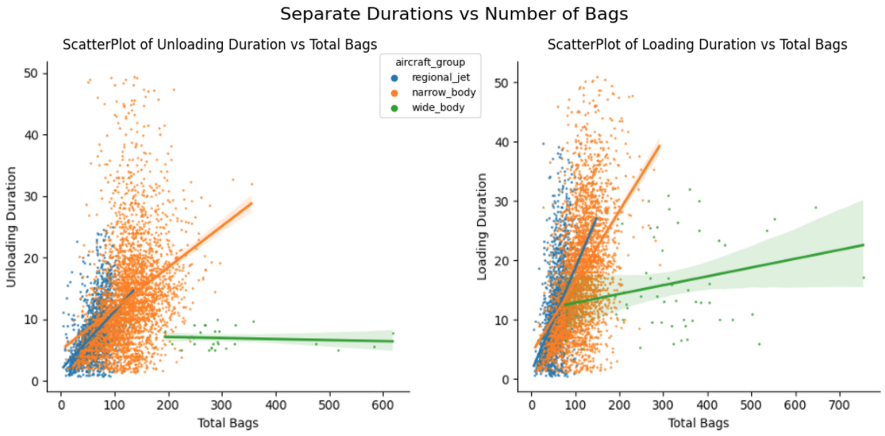


Figure 4.1: Target Durations vs Number of Bags

When examining the relationship between the total number of bags and the target durations in Figure 4.1, consistent positive trends can be observed, with some variations among aircraft groups. Wide-body planes exhibit deviations from the expected trend, indicating a unique relationship likely influenced by their handling processes, such as containerization and different loading devices. Outliers in the scatter plot for narrow-body planes suggest unidentified factors affecting the duration for this group. Overall, the scatter plots demonstrate dense clusters of data points, primarily concentrated around or below 200 bags. These clusters can be categorized into regional jets and narrow-body planes. Each group shows a similar increasing trend, however, values for the wide-body aircraft show no clear increasing trend and have significant spread and variability. The high density within a moderately increasing trend indicates that while there is a relationship between the number of bags and the duration, it is not particularly very strong.

Flight Duration

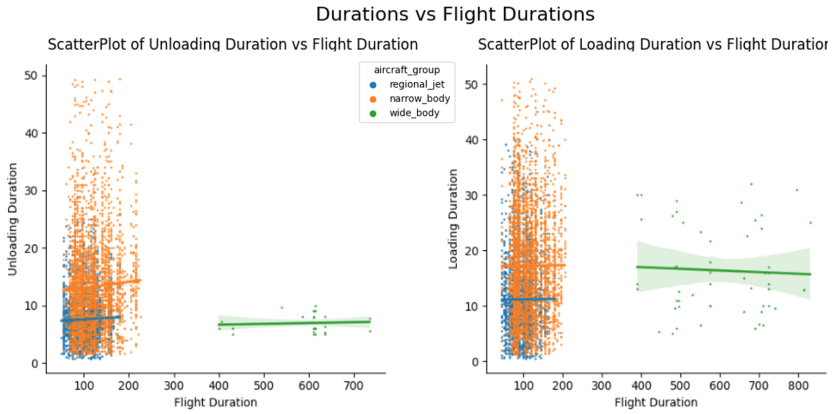


Figure 4.2: Target Durations vs Scheduled Flight Duration

Figure 4.2 illustrates the relationship between flight duration and each target duration classified by aircraft group.

The scatter plots indicate a moderately increasing trend between flight duration and target durations for regional jet and narrow-body aircraft groups, with data points clustering within a 200-minute by 200-minute square. However, the correlation is not distinctly linear. The durations exhibit higher density and wider variability, suggesting a more subtle correlation.

Moreover, a notable gap exists between the data points of regional jet and narrow-body aircraft groups and wide-body planes, challenging the observed positive correlation trend. Wide-body planes demonstrate sparser data points and a less prominent positive trend compared to other groups.

Overall, the correlation between flight duration and target durations, particularly the unloading and loading separate durations, is not strongly established, indicating the influence of additional factors.

Weight per Passenger

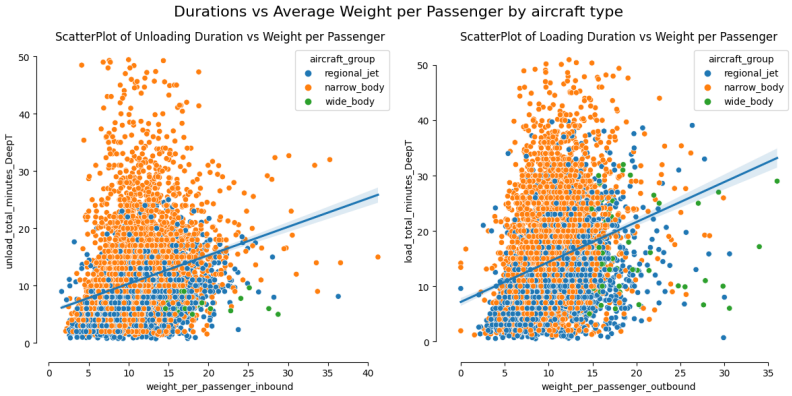


Figure 4.3: Target Durations vs Weight per Passenger

Figure 4.4 shows scatter plots illustrating the relationship between the ”weight per passenger” feature and each target duration. The plots display an ascending linear trend but with significant spread which means no clear influence, with values primarily clustered in the (5-20) range. Wide-body aircraft data points exhibit greater spread, especially in the loading duration plot, likely due to fewer data points and containerized handling. As a result, this feature introduces significant variability but still holds the potential for contributing explanatory power to the prediction model.

Load factor

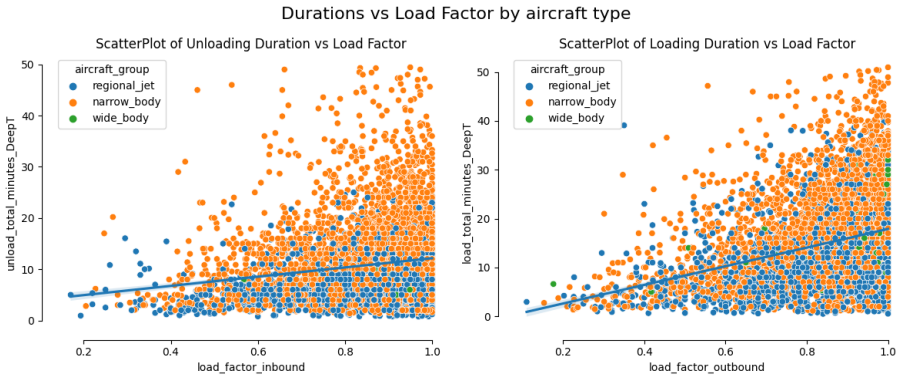


Figure 4.4: Target Durations vs Load Factor

Figure 4.4 presents the scatter plots illustrating the relationship between the ”load factor” feature and unloading and loading separate durations.

The scatter plots display an increasing trend, although scattered values introduce variability and outliers to this relationship. Specifically, the graph for the loading duration demonstrates a sharper upward trend. While it is challenging to identify distinct data groups, a clear observation is that a smaller load factor corresponds to a shorter duration. Surprisingly, it is also evident that a high load factor can result in a shorter time, as indicated by numerous data points. This observation aligns with the reality that the number of passengers on the plane does not necessarily correlate with the quantity of bags to be loaded. However, this variable does offer some aid in the prediction of the durations.

Solar Radiation

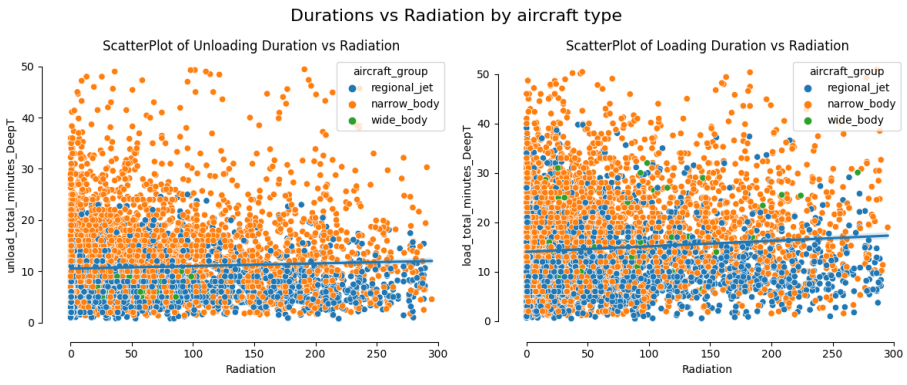


Figure 4.5: Target Durations vs Radiation

Figure 4.5 displays scatter plots illustrating the relationship between the ”Radiation” feature and each target duration. Despite examining the potential impact of high solar radiation levels on the baggage unloading-loading process, the correlation analysis indicates a weak relationship, as indicated by the small correlation coefficient. Consequently, it is challenging

to discern a clear trend in the scatter plots, with only a slight upward trend. These findings suggest that while solar radiation may have some influence on the duration, its impact is not significant or consistent in this case.

Special Items

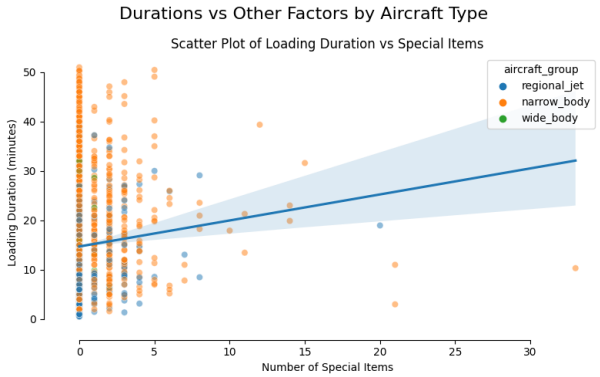


Figure 4.6: Loading Duration vs Special Items

Figure 4.6 shows the relationship between the loading duration and the presence of special items. The scatter plot reveals a slightly increasing trend, with a dense cluster of data points around zero, as most flights do not involve special items. However, when special items are present, longer durations are expected. It is worth noting that for values starting from 20 on the special items axis, the duration remains relatively low, indicating some inconsistency. Overall, the presence of special items has the potential to modestly enhance predictions due to the discernible increasing trend.

4.3.2 Exploring the Impact of Categorical Features

This subsection visually examines the influence of remaining categorical variables on the target durations. Box plots are used to analyze the distributions of durations across different categories. Box plots provide a summary of the data distribution, highlighting the range, outliers, and central tendency using whiskers, the interquartile range (IQR), and the median line.

Aircraft Group

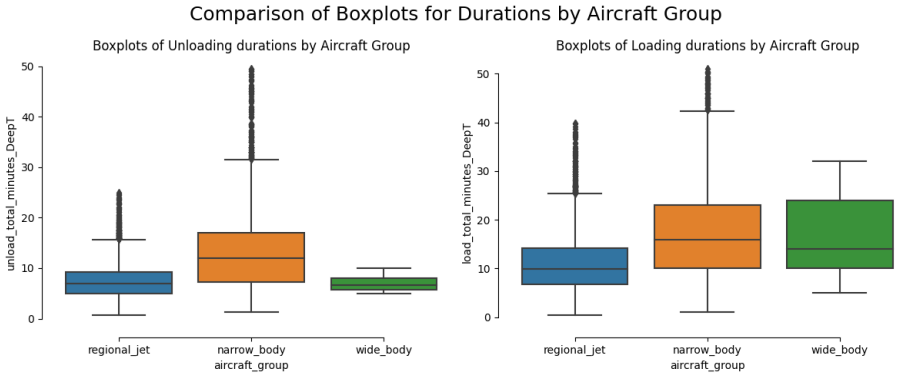


Figure 4.7: Variation of Durations Across Aircraft Groups

The Box-plot analysis in Figure 4.7 highlights the impact of aircraft groups on durations. It is important to say that the limits for the distribution of this specific variable have been defined beforehand since they aid in the process of outlier cleaning for the durations in Section 3.5.

Unloading duration analysis reveals wider distribution for narrow-body planes, while the other two groups show smaller variations. Wide-body planes exhibit a strong correlation with duration, indicated by a small interquartile range (IQR). Regional jets have outliers that may impact predictions. In the loading duration, each group exhibits variability, with regional jets showing slightly less variability but more outliers, potentially affecting accuracy. Notably, wide-body planes demonstrate shorter durations due to containerized handling, while narrow-body planes have longer durations compared to regional jets due to greater bag storage capacity.

Continent Inbound/Outbound

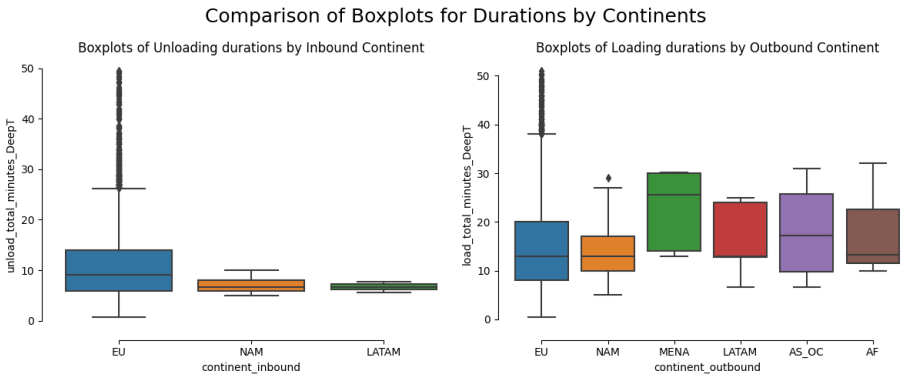


Figure 4.8: Variation of Durations Across Inbound/outbound Continents

The Box-plot analysis in Figure 4.8 highlights the impact of the inbound and outbound continents on durations.

In to the unloading duration, North and Latin America offer excellent predictive power, supported by a small interquartile range (IQR) and no outliers. Europe, on the other hand, has a larger IQR, indicating greater variability in unloading duration and a higher presence of outliers. This can be attributed to the limited data available for wide-body planes which are the only aircraft group that fly intercontinental routes.

Regarding the loading duration, the IQR across outbound continents reveals variations ranging from 6 to approximately 18 minutes. Latin America shows a skewed distribution without a clear median line, but its IQR and whisker range does not warrant its removal from the analysis.

Time-Related Variables

To analyze time-related variables such as the hour of the day, day of the week, and month of the year, histograms are utilized to visualize the mean duration across different time periods. It is worth noting that these variables are derived from the scheduled inbound arrival timestamp, which will also be used for prediction.

The visualizations that can be found in Appendix B reveal the variations of the target durations across time-related variables, highlighting their impact. In addition, the analysis provided a useful realization of categorizing each time-related variable into peak, medium peak, and low peak categories, as depicted in Figure 4.9. These newly categorized variables will be used for prediction, providing a more concise approach compared to using the variables in their original form.

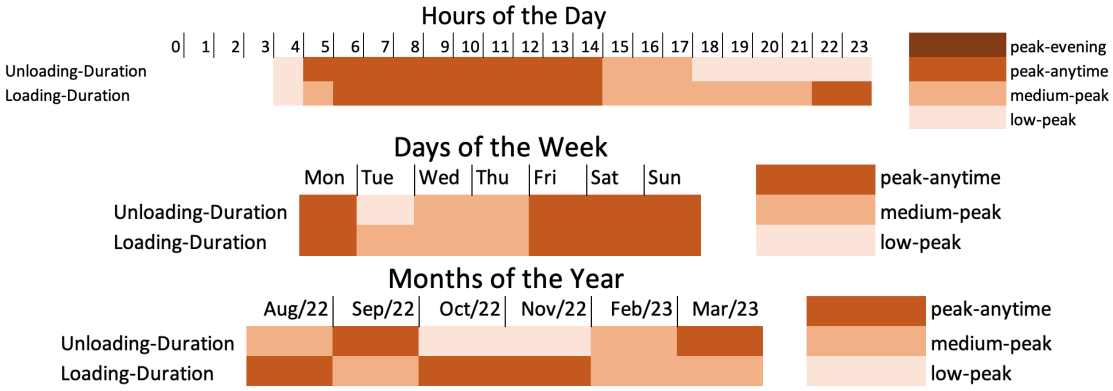


Figure 4.9: Classification of Time-related Variables

Weather-Related Variables

While weather variables were not included in the analysis due to their limited insights, they can be found in Appendix B. However, some observations can still be derived from them. Firstly, it was noted that there was no presence of snow for wide-body aircraft, and when snow was present, it led to increased baggage handling durations. Furthermore, the absence of mist was observed for wide-body planes, and although the presence of mist made durations more variable, its overall impact was not substantial. In terms of rain, it appeared

to have a significant effect primarily on wide-body planes. Additionally, the presence of thunderstorms relatively increased the durations. Lastly, while ice formation had a minor impact on duration variability, it was not observed for wide-body planes. These insights provide some context about the weather variables, albeit their limited influence on the analysis.

4.4 Data Preprocessing: Encoding, and Imputation

This section focuses on further preprocessing the selected list of variables from the two datasets to predict the baggage unloading and loading durations. The following subsections describe the specific preprocessing steps applied to the data.

4.4.1 Encoding of Categorical Features

Based on the exploratory data analysis in Section 4.3, the time-related variables are encoded into peak, medium peak, and low peak categories. A value of 1 is assigned if the time-related variable falls under that category, and 0 if it does not. The weather-related variables are already binary encoded, where a value of 1 indicates the presence of a certain weather pattern, such as rain, and 0 indicates its absence. Finally, the remaining nominal categorical variables for aircraft groups and inbound/outbound continents are encoded using one-hot or dummy encoding. This technique transforms each category into a binary vector representation, creating new binary variables. Each variable represents a single category and indicates its presence (1) or absence (0) in the original data. This encoding ensures that the machine learning algorithm treats each category equally and avoids assuming any ordinal relationship among them.

4.4.2 Imputation Methods

For continuous variables, missing values can be replaced with either the mean of the entire column or zero. For categorical columns, the most frequent values were used for imputation. However, the only column with missing values was the one for inbound special items, which has 29 out of 5445 missing values. Since it is not known exactly if there were special items in those certain flights, and considering that most of the values for this column are zero, it made sense to fill these missing values with zero.

4.5 Feature Selection

Taking the advice of Crone and Kourentzes (2010), who recommends a first reduction in space using filter methods, followed by a second reduction assisted by wrapper methods. Previously, in Section 4.2, filter methods like correlation coefficients and statistical tests help identify and rank influential features. This section focuses on advanced techniques to further reduce variables and find the optimal feature set for each ML model in the subsequent phase.

Three widely recognized wrapper methods are used for the purpose of feature selection. These methods, namely Recursive Feature Elimination and Sequential Floating Forward/Backward Selection, have been extensively discussed and identified as popular techniques in the literature review conducted in Chapter 2. Additionally, we designed a method that combines the filter correlation ranking method with forward feature selection.

The primary criterion for selecting the best feature selection algorithm is to strike a balance between a reduced number of variables and sufficient explanatory power while minimizing error evaluated using the mean squared error (MSE) metric. This metric is chosen because it gives outliers an increased weighting. Moreover, the run time of these feature selection models is not a major concern in this case, as they are only run once to select the optimal subset of features for each Machine learning model. However, if additional variables are introduced to improve model performance and the ML model needs to be run alongside the feature selection algorithms, considering the run time becomes important. In such cases, it is recommended to choose the algorithm with the lowest runtime for faster decision-making and model development.

4.5.1 Feature selection setup

The feature selection methods employ a 5-fold cross-validation to evaluate their performance, using RMSE as the evaluation metric. This approach involves dividing the data into subsets (folds), cycling through each fold as both training and validation data. As a result, it provides a reliable estimate of the feature selection methods' performance. Finally, selecting 5 folds is a common choice for medium-sized datasets as it strikes a good balance between bias and variance.

When it comes to methods to use, Recursive Feature Elimination (RFE) eliminates less important features from a dataset by relying on feature importances provided by the underlying algorithm. It works for algorithms like Random Forest and XGBoost that offer feature importances. However, it is not suitable for Multi-layer Perceptron Artificial Neural Networks (MLP ANN) or Support Vector Regression (SVR), as these algorithms do not provide feature importance due to their complex nature. SVR seeks an optimal hyperplane, while ANN MLP learns weights and connections through complex optimization, making feature importance extraction challenging. On the other hand, the Sequential Floating Forward and Backward Selection (SFFS/SFBS) algorithms can be applied to all machine learning algorithms as it iteratively adds or removes features based on their impact on model performance, without relying on specific feature importances.

Recursive Feature Elimination with Cross Validation (RFECV):

In Python's Scikit-learn library, the Recursive Feature Elimination (RFE) algorithm is available for feature selection. The library also provides RFECV, an extension of RFE that incorporates cross-validation and automatically selects the optimal number of features. The main parameters for these algorithms include the estimator model, the scoring measure (MSE in

this case), the step (which determines the number of features to remove at each iteration, set to 1), and the number of folds for cross-validation (set to 5) which is a standard number used in practice (Pedregosa et al., 2011).

Sequential Floating Forward/Backward Selection (SFFS/SFBS):

The Sequential Floating Forward and Backward Selection algorithms can be applied using the Mlxtend library in Python. The feature selection parameter is set to 'best', indicating that the algorithm aims to find the best subset of features. The Forward parameter can be set to either True or False, depending on whether forward selection or backward elimination is desired. Additionally, the Floating parameter is set to true to enable the additional opposite step if it improves the objective function. For cross-validation, a 5-fold validation strategy is applied (Raschka, 2018).

Sequential Forward Correlation Ranking (SFCR):

This method combines correlation ranking techniques, including Spearman's correlation coefficient for normally distributed features and Pearson's correlation coefficient for others. The goal is to rank all categorical encoded and continuous scaled features in order to obtain a final ranking. After the ranking is ready, forward sequential selection is applied, adding one variable at a time, and the mean squared error (MSE) is evaluated for each subset by using 5-fold cross-validation. The subset of variables with the highest score is chosen as the optimal feature set.

This method is employed after addressing multicollinearity in the previous section and with the hypothesis that only the most influential variables should be selected.

4.5.2 Results of Feature Selection

Table 4.5 showcases the feature selection results, with green highlighting the selected feature selection techniques. Additionally, the machine learning models were also run with all variables as a baseline comparison, allowing for performance evaluation of the effectiveness of applying feature selection models.

Duration	ML model	Feature selection technique	Evaluated features	# Selected features	MSE	RMSE	Running time (in seconds)
Unloading duration	Random Forest	All Features	24	24	41.55	6.4459	0
		Recursive Feature Elimination		11	41.58	6.4483	189
		Sequential Feature Correlation Ranking		17	41.47	6.4397	132
		Sequential Floating Forward Selection		6	40.8	6.3875	1066
		Sequential Floating Backward Selection		2	40.8	6.3875	1277
	XGBoost	All Features		24	46.19	6.7963	0
		Recursive Feature Elimination		2	41.04	6.4062	35
		Sequential Feature Correlation Ranking		3	41.03	6.4055	22
		Sequential Floating Forward Selection		2	40.71	6.3804	286
		Sequential Floating Backward Selection		2	40.71	6.3804	279
	Support Vector Machine (SVM)	All Features		24	40.77	6.3851	0
		Sequential Feature Correlation Ranking		7	39.31	6.2698	47
		Sequential Floating Forward Selection		14	39.09	6.2522	704
		Sequential Floating Backward Selection		9	39.07	6.2506	494
	Multi-Layer Perceptron ANN	All Features		24	38.56	6.2097	0
		Sequential Feature Correlation Ranking		5	38.49	6.204	228
		Sequential Floating Forward Selection		10	38.38	6.1952	3359
		Sequential Floating Backward Selection		7	38.43	6.1992	3146
Loading duration	Random Forest	All Features	23	23	58.82	7.6694	0
		Recursive Feature Elimination		8	58.84	7.6707	167
		Sequential Feature Correlation Ranking		19	58.58	7.6538	241
		Sequential Floating Forward Selection		12	58.41	7.6426	1700
		Sequential Floating Backward Selection		3	58.27	7.6335	1737
	XGBoost	All Features		23	66.22	8.1376	0
		Recursive Feature Elimination		3	59.22	7.6955	34
		Sequential Feature Correlation Ranking		1	59.22	7.6955	51
		Sequential Floating Forward Selection		10	58.42	7.6433	417
		Sequential Floating Backward Selection		3	58.47	7.6466	442
	Support Vector Machine (SVM)	All Features		23	58.09	7.6217	0
		Sequential Feature Correlation Ranking		6	54.9	7.4095	76
		Sequential Floating Forward Selection		13	54.79	7.402	797
		Sequential Floating Backward Selection		12	54.78	7.4014	993
	Multi-Layer Perceptron ANN	All Features		23	55.18	7.4283	0
		Sequential Feature Correlation Ranking		6	54.72	7.3973	254
		Sequential Floating Forward Selection		10	54.51	7.3831	1973
		Sequential Floating Backward Selection		7	54.5	7.3824	3658

Table 4.5: Feature selection results

The Sequential Feature Correlation Ranking (SFCR) method, which combines filter and wrapper techniques, demonstrates higher efficiency compared to other methods and performs well with ML models SVR and MLP ANN. However, it may not perform as effectively with tree-based ML models. This can be attributed to the fact that tree-based models have the ability to extract explanatory power not only from highly influential variables but also from the complex relationships within the data. As a result, they may exhibit better performance when features combine appealing relationships rather than solely relying on the highest-ranked influential variables.

4.5.3 Final Subset of Features

	Random Forest (RF)	XGBoost	Support Vector Reg. (SVR)	Multi-Layer Perceptron ANN
Unloading Duration	Number of incoming bags Hour low-peak Month peak Month medium-peak aircraft_group_wide_body continent_inbound_LatinoAmerica	total_bax_inbound aircraft_group_wide_body	total_bax_inbound load_factor_inbound weight_per_passenger_inbound Hour peak Hour medium-peak Hour low-peak Day low-peak aircraft_group_regional_jet aircraft_group_wide_body	total_bax_inbound weight_per_passenger_inbound Hour peak Hour medium-peak Day low-peak aircraft_group_narrow_body aircraft_group_regional_jet
Loading Duration	Number of outgoing bags aircraft_group_wide_body continent_outbound_NorthAmerica	Number of outgoing bags Hour low-peak Month peak Month medium-peak Day peak aircraft_group_wide_body continent_outbound_Africa continent_outbound_EU continent_outbound_LatinoAmerica continent_outbound_Middle-East	Number of outgoing bags load_factor_outbound weight_per_passenger_outbound special_items_outbound Rain Hour medium-peak Hour low-peak Month peak aircraft_group_regional_jet aircraft_group_wide_body continent_outbound_Asia-Oceania continent_outbound_NorthAmerica	Number of outgoing bags load_factor_outbound weight_per_passenger_outbound Rain Hour medium-peak Month medium-peak continent_outbound_EU

Table 4.6: Final selected features

The results in Table 4.6 confirm the consistency between the previously identified important features and the selections made by the feature selection models. These features include the number of bags, aircraft groups (particularly wide-body planes), hours of the day, months, and the influence of outbound continents on loading duration. Additionally, the weight per passenger consistently affects the unloading duration.

Among the features with the highest importance identified by the tree-based methods shown in Appendix B, the number of bags and the wide-body aircraft group as the most important features. XGBoost assigns greater significance to the wide-body aircraft group compared to the Random Forest model. Additionally, XGBoost identifies the European outbound continent as highly important for predicting the loading duration.

4.6 Conclusion

In conclusion, this chapter focused on enhancing prediction models by improving the quality and relevance of available features. The process involved feature selection and derivation in Section 4.1, correlation analysis and multicollinearity cleaning in Section 4.2, exploratory data analysis in Section 4.3, data preprocessing, in Section 4.4, included one-hot encoding of categorical variables and imputation of missing values, and finally, 4 feature selection techniques were explored in Section 4.5 resulting in the selected feature subsets for each subsequent ML model, which can be found in Table 4.6.

Chapter 5

Modelling and validation

After selecting the best features for each machine learning technique, the focus shifts to the chosen subset of features mentioned in Subsection 4.5.3. This chapter aims to construct reliable prediction models by providing a comprehensive understanding of the process. It is divided into several sections that cover various aspects. In Section 5.1, the most suitable Machine Learning models for predicting the target durations are selected. Section 5.2 focuses on dataset partitioning, and cross-validation to ensure the validity and generalizability of the prediction models during evaluation. Section 5.3 provides detailed information about the chosen models' architectures and their hyperparameters. In Section 5.4, the hyperparameter search approach is explained, with selection based on the RMSE. Finally, in Section 6, key findings are summarized and the process is reflected upon.

5.1 Selection of ML models

Based on the literature review in Sub-section 2.3.4, several studies (Hassel, 2019; Luo et al., 2021; Schultz et al., 2021) have identified Random Forest and XGBoost as the best ensemble methods for predicting airline-related problems. These methods have demonstrated promising results in terms of predictive performance and are widely used in the field. Moreover, Support Vector Regression (SVR) has also been chosen due to its significance as a regression predictor and the favorable outcomes reported in various studies (Carpinteiro et al., 2012; Jasra et al., 2018; Schultz et al., 2021). SVR is known for its ability to handle complex relationships in data and has been successfully applied in the prediction of airline-related variables. Finally, the Multi-layer Perceptron Artificial Neural Network (MLP ANN) has been selected based on its utilization in prior studies (Carpinteiro et al., 2012; Gao et al., 2015; Hassel, 2019; Schultz and Reitmann, 2019). MLP ANN is a type of neural network that can effectively capture non-linear relationships and has been applied in various domains, including airline-related prediction tasks.

The chosen ML models, namely Random Forest, XGBoost, SVR, and MLP ANN, have a solid research foundation and practical applicability. They excel in handling complex non-linear relationships, crucial for accurately understanding factors and improving predictions for baggage unloading and loading durations. Additionally, these models exhibit robustness

to noise and outliers. Even after data cleaning, the impact of remaining outliers can be mitigated through their effective mechanisms. Moreover, these models are scalable and flexible, adept at handling large datasets and varying complexities. Random Forest and XGBoost handle high-dimensional data, while SVR and MLP ANN can adapt to different configurations. Furthermore, these models provide valuable insights into feature importance, with Random Forest and XGBoost estimating importance and SVR and MLP ANN revealing feature contributions in magnitude and direction.

5.2 Model Development Process

This section addresses the critical task of partitioning the dataset and employing cross-validation techniques to guarantee reliable and generalizable results during the evaluation of prediction models.

5.2.1 Data Split

The dataset was split using a common practice of a 75/25 ratio for the training and testing sets. The 75% training data is used to fit the model, while the 25% test data is used to evaluate its predictive performance. This partitioning step is crucial to prevent overfitting or underfitting, ensuring that our model can generalize well to unseen data.

5.2.2 Cross-validation

As identified in the Literature Review in Chapter 2, 5-fold cross-validation is employed for this prediction task. It involves training the model on 4 folds of data and using the remaining fold as the test set. This choice of 5 folds is widely used as a standard in cross-validation. With 5 folds, a significant portion of the data (80% for training, 20% for testing) is utilized in each fold. Compared to the 75/25 ratio of the main data split, this 5-fold cross-validation with an 80/20 ratio ensures a robust training process. It offers advantages like a slightly larger training set for improved generalization and enhanced model learning capacity. Additionally, it allows for a comprehensive assessment of the model's performance across various data distributions.

The 5-fold cross-validation serves multiple purposes in this study. Initially, it was employed to select the best subset of features for each model predicting each duration. Subsequently, it will be used to identify the optimal hyperparameter configuration. Finally, it will be utilized for the overall evaluation of the selected ML models. The model with the best-averaged evaluation metric across all folds will be chosen as the best-performing model.

5.3 Model Architecture and Hyperparameters

To construct an accurate prediction model, it is essential to carefully consider the architecture and parameters of the selected machine learning models. In this section, we provide

comprehensive insights into the architecture and hyperparameters of the following models: Random Forest, XGBoost, SVR, and MLP ANN. The following explanation was inspired by (Géron, 2017; Pedregosa et al., 2011)

5.3.1 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest independently learns from a random subset of the training data and features, reducing overfitting and improving generalization. The number of trees in the forest is set by the n_estimators hyperparameter, allowing for more robust predictions with a larger ensemble. The max_depth hyperparameter controls the maximum depth of each decision tree, preventing overfitting by limiting the complexity of the trees. Additionally, the min_samples_split hyperparameter determines the minimum number of samples required to split an internal node, influencing the tree's ability to capture finer patterns in the data. Lastly, the max_features hyperparameter governs the number of features considered at each split, promoting diversity among the trees and enhancing their collective predictive power.

5.3.2 XGBoost

XGBoost is an optimized gradient-boosting framework that leverages an ensemble of weak prediction models to build a stronger predictive model. It sequentially adds new models to the ensemble, with each new model learning from the errors made by the previous models. The learning_rate hyperparameter controls the step size shrinkage during the boosting process, influencing the contribution of each weak learner. A lower learning rate makes the model more conservative, while a higher learning rate allows for faster learning but increases the risk of overfitting. The max_depth hyperparameter limits the complexity of the individual weak learners, preventing overfitting and improving generalization. The reg_alpha and reg_lambda hyperparameters add L1 and L2 regularization terms to the model, respectively, helping to control overfitting by penalizing large weights in the ensemble.

5.3.3 SVR (Support Vector Regression)

SVR is a variant of Support Vector Machines (SVM) adapted for regression tasks. It aims to find an optimal hyperplane that approximates the relationship between the input features and the target variable. The choice of kernel function, controlled by the kernel hyperparameter, determines the type of decision boundary used by the SVR model. The linear kernel is appropriate for linear relationships, while the polynomial, radial basis function (rbf), and sigmoid kernels can capture more complex relationships. The C hyperparameter adjusts the trade-off between fitting the training data and allowing deviations within the margin, influencing the model's tolerance to errors. The gamma hyperparameter, relevant for the rbf, poly, and sigmoid kernels, defines the kernel coefficient and affects the shape of the decision

boundary. Lastly, the degree hyperparameter specifies the degree of the polynomial kernel function, allowing for flexibility in capturing non-linear relationships.

5.3.4 MLP ANN (Multi-Layer Perceptron Artificial Neural Network)

MLP ANN is a type of artificial neural network with multiple layers of interconnected nodes (neurons). It excels at learning complex patterns and relationships in the data. The hidden_layer_sizes hyperparameter determines the number of neurons in each hidden layer of the network. A larger number of neurons allows the network to capture more intricate patterns but may increase the risk of overfitting. The activation hyperparameter specifies the activation function used in the hidden layers, influencing the network's non-linear behavior. The alpha hyperparameter controls the L2 regularization term (weight decay), helping to prevent overfitting by penalizing large weights in the network. Finally, the learning_rate_init hyperparameter sets the initial learning rate for weight updates during training, affecting the convergence speed and the network's ability to find an optimal solution.

5.4 Hyperparameters Tuning

This section involves fine-tuning the values of the hyperparameters to achieve the best performance of the ML models. To achieve this, search techniques are used to explore different combinations of hyperparameters within predefined ranges.

For hyperparameter tuning, the randomized search method was employed. This method involves randomly sampling a specified number of combinations from a defined parameter space (Géron, 2017). In this case, we conducted 100 to 150 iterations of random combinations to identify the optimal hyperparameters for our models. The randomized search was chosen over grid search due to its ability to efficiently explore a wide range of hyperparameter combinations (100-150 iterations) without exhaustively evaluating every single combination. This, in the case of large search spaces, can save computational resources and reduce the time required for optimization, making it suitable for this purpose. Finally, the mean squared error (MSE) was selected as the evaluation metric instead of the RMSE. Because of its squared nature, the MSE allows for a more precise comparison and facilitates the identification of the hyperparameter configuration that minimizes prediction errors, thereby improving the accuracy of our models.

When determining the range for hyperparameters, we rely on the values specified in the Scikit-learn library (Pedregosa et al., 2011), as well as commonly used ranges for hyperparameters. For the Multi-Layer Perceptron Artificial Neural Network (MLP ANN), the number of hidden layers and their neurons is a crucial parameter. To determine this parameter, we refer to advice provided by Gao et al. (2015); Hassel (2019). We use the formula:

$$m = \sqrt{p + q} + a$$

where m, p, q, and a represent the number of hidden layer nodes, input layer nodes, output layer nodes, and a constant, respectively. The authors recommend setting the constant

Model	Hyperparameters	Range	Unloading duration: Optimal values	Loading duration: Optimal values
Random Forest	n_estimators	[60, 100, 150, 250, 350, 500, 700, 850, 1000, 1200]	700	700
	max_features	[1.0, 'sqrt', 'log2']	"sqrt"	"sqrt"
	max_depth	[5, 10, 15, 20, 25, 30, None]	5	5
	min_samples_split	[2, 5, 10, 15, 100]	100	100
	min_samples_leaf	[1, 2, 4, 6, 8, 10]	1	1
	bootstrap	[True]	TRUE	TRUE
	# features		6	3
XGBoost	n_estimators	[100, 200, 300, 400, 500]	500	500
	min_child_weight	[4, 6, 8, 10, 12]	6	6
	max_depth	[3, 4, 5, 6, 7, 8, 9]	4	4
	learning_rate	[0.5, 0.25, 0.1, 0.01, 0.001]	0.01	0.01
	subsample	[0.6, 0.8, 0.9]	0.9	0.9
	colsample_bytree	[0.6, 0.8, 0.9]	0.9	0.9
	gamma	[0.0, 0.1, 0.2, 0.3, 0.4]	0	0
	reg_alpha	[0, 0.001, 0.005, 0.01, 0.05]	0.005	0.005
	# features		2	10
SVR	kernel	['rbf', 'poly']	"rbf"	"rbf"
	C	[0.1, 0.25, 0.5, 0.75, 1, 5, 10, 100]	10	100
	epsilon	[0.01, 0.1, 0.25, 0.5, 0.75, 1]	1	1
	gamma	['scale', 'auto', 0.1, 1, 10]	10	"scale"
	# features		9	12
MLP ANN	hidden_layer_sizes	(ranges specified in the code)	(64, 64, 64)	(44, 24, 44)
	activation	['relu']	Relu	Relu
	solver	['adam', 'sgd']	"adam"	"adam"
	max_iter	[300, 400, 500]	400	500
	alpha	[0.0001, 0.001, 0.01, 0.1]	0.0001	0.01
	learning_rate	['constant', 'adaptive']	"constant"	"constant"
	# features		10	7

Table 5.1: Hyper parameters of each model (ADD NUMBER OF VARIABLES)

between 1 and 10. However, due to the small number of features in our dataset, applying such a constant would result in excessively small hidden layers. Therefore, we have chosen to use a constant range of 20 to 120, with a step size of 20. This ensures that our hidden layers have a sufficient number of neurons for accurate predictions.

Through systematic variation of hyperparameters during the Randomized search, the analysis reveals the impact of different hyperparameter settings on the models' MSE performance. By recording and comparing the MSE for each configuration, valuable insights are gained regarding the relationships between hyperparameters and model outcomes as shown in Appendix D. It is worth noting that due to the randomized nature of the search process with 100-150 iterations, specific results cannot be directly attributed to individual hyperparameter configurations. Finally, the results of the hyperparameter tuning can be found in Table 5.1, which provides an overview of the selected hyperparameters for each of the selected ML models.

5.5 Conclusion

This chapter aimed to construct robust prediction models by implementing reliable tuning and evaluation methods. In Section 5.1, Machine Learning models, namely Random Forest, XGBoost, SVR, and MLP ANN were carefully selected based on their extensive use and proven effectiveness in the literature. In Section 5.2, the dataset was partitioned into

training and testing sets using a 75-25% ratio to mitigate the risk of overfitting or underfitting. Additionally, a 5-fold cross-validation was applied to assess the models' validity and generalizability during evaluation. Section 5.3 provided comprehensive insights into the architectures and functionality of the selected models, including their crucial hyperparameters. Section 5.4 introduced a systematic approach to finding the optimal hyperparameter configuration using a Randomized search method, cross-validation, and MSE as the evaluation metric. The resulting hyperparameter configurations are summarized in Table 5.1 and their sensitivity can be found in Appendix D. Overall, this chapter offered a detailed and organized framework to train the models effectively, addressing potential issues such as overfitting, underfitting, multicollinearity, and ensuring generalizability.

Chapter 6

Results and evaluation

This chapter focuses on the evaluation and comparison of the available estimation models for the unloading-loading duration. To begin, the focus is to analyze the performance of the trained ML models and identify the models that excel in each task. Section 6.1 shows the main outcomes of training and testing the models for each duration. Furthermore, Section 6.2 focuses on the selection of the best-performing machine learning model for each duration based on the averaged RMSE and its standard deviation from the 5-fold cross-validation. The performance of these models is then evaluated by analyzing the differences between predicted and actual values. Moving on to Section 6.3, the comparison is made between the previously identified approximation of the current tool (Section 3.1), the two data-driven methods (Section 3.6), and the best-performing ML models. This comparison is conducted for each duration across the entire dataset and specific subsets for comprehensive evaluation. In Section 6.4, a high-level integration plan for implementing the best-performing models is outlined. To conclude, a summary of the chapter's findings and contributions is provided.

6.1 Model Training and Testing Outcomes

In this section, the models will be trained on 75 percent of the data, which constitutes the training set. The specific features and hyperparameters determined earlier and shown in Table 5.1 will be used for training. After the training phase, the predictions generated by each model will be evaluated using the test set. The evaluation metrics are presented in Table 6.1, providing insights into the performance of each model for different subsets of data. The subsequent sub-sections will analyze the visual results of each ML model separately, focusing on the specific target durations. Furthermore, scatter plots depicting the predicted versus actual values from the test set will be provided, a good performance is given if data points tend to gather across the regression line which is the line of perfect fit for the actual values.

6.1.1 Unloading Duration

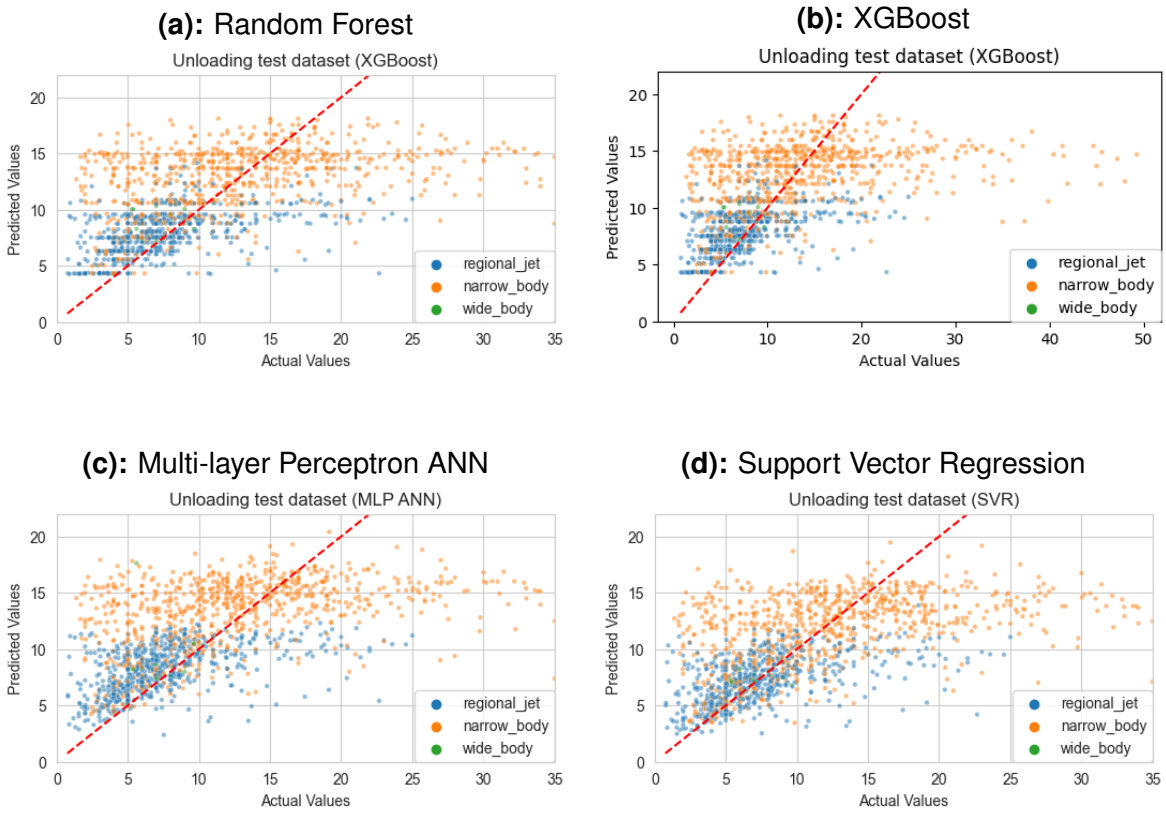


Figure 6.1: Results of ML Models for the unloading duration

Figure 6.1 presents a visual representation of the ML model's unloading duration predictions in comparison to the actual values from the test dataset. The RF model demonstrates strong predictive performance for durations up to 20 minutes, effectively clustering a significant number of data points around the regression line. However, it tends to overpredict, likely due to a cluster of data points located before the regression line. The model's predictions exhibit a distinct pattern of levels, indicated by the horizontal line of data points, rather than a smooth and continuous range. Moreover, the XGBoost model follows a comparable pattern to the RF model but with slightly fewer levels of prediction. Consequently, both models exhibit comparable performance in terms of RMSE.

The MLP ANN model exhibits a continuous linear trend that closely aligns with the regression line. However, this trend appears slightly shifted to the left, suggesting a tendency toward overprediction. Additionally, the MLP ANN model performs well for both low and high durations, particularly in predicting durations up to 20 minutes. This prominent performance could be attributed to its ability to capture a dense cluster of data points from the narrow-body group situated near the regression line. In addition, the SVR model, ranking as the second-best performer in terms of the RMSE metric, initially exhibits a strong linear increasing trend across the regression line. However, it demonstrates greater variation

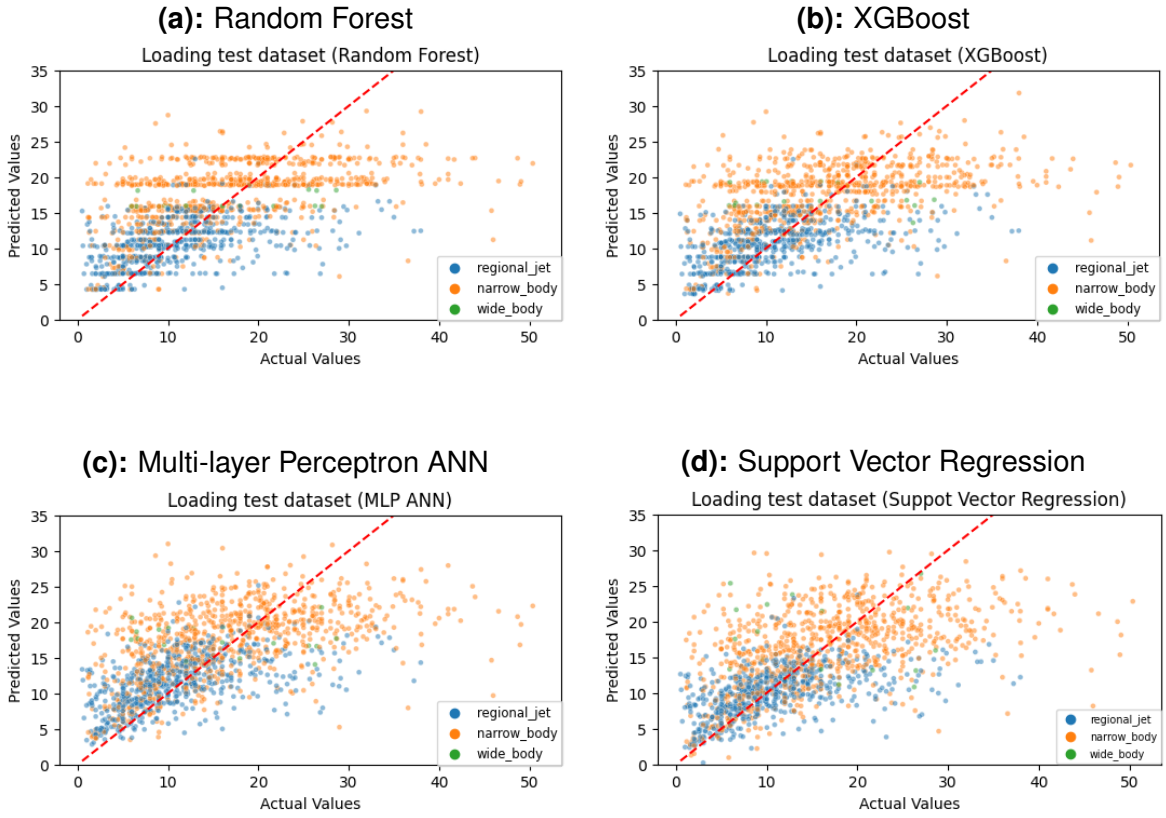


Figure 6.2: Results of ML Models for the loading duration

when predicting values for the narrow-body group, leading to the MLP ANN model slightly outperforming it in this regard.

6.1.2 Loading Duration

Figure 6.2 depicts the outcomes of the ML model loading duration predictions when compared to the actual values of the test dataset. The RF model demonstrates strong performance in terms of variability, with data clustered in distinct levels. These clusters exhibit promising predictive capabilities until around the 15-minute mark, after which they become wider and encompass more outliers. The RF model's good performance in terms of RMSE can be attributed to its ability to minimize the presence of scattered outliers that deviate from the regression line. Similar to the RF model, the XGB model follows a comparable pattern, but the clusters of data points appear slightly shifted to the left of the regression line. This suggests a tendency towards overprediction.

The MLP ANN model initially exhibits a notable upward trend, although this trend is slightly offset to the left of the regression line, indicating a propensity for overprediction. Moreover, the data points for narrow-body planes do not align closely with the regression line, indicating a lack of strong clustering. Finally, the SVR model shares a similar pattern with the MLP ANN model, but the values for narrow-body planes display fewer clusters

around the regression line. There is no distinct visible predictive trend, indicating a relatively weaker performance compared to the other models.

6.1.3 Overall Discussion

For the unloading duration, all models perform similarly and achieve good results with an RMSE of approximately 6 minutes, indicating low variability between predicted and actual values as observed in Figure 6.1. Accuracy is high for regional jets and wide-body planes, but difficulties arise when predicting unloading durations for narrow-body planes. This category has the majority of outliers, and for durations exceeding 20 minutes, the models tend to converge around the value of 15, limiting their predictive capability. Tree-based models also struggle with accurately predicting low durations compared to the other two models.

Based on Figure 6.2, the overall analysis of the ML models' performance for the loading duration, shows that predictions for regional jets and wide-body planes are satisfactory, similar to the unloading duration. However, narrow-body planes exhibit deviations from the actual values, particularly for durations exceeding 30 minutes. The predicted values cluster within a range of 10 to 25 minutes (as reflected by the limited y-axis range), which includes outliers. It's important to note that these outliers should not significantly affect the overall accuracy of the predictions.

6.2 Best Performing Model Selection

In this section, the best-performing model for each duration will be determined by using the methodology from Sub-section 5.2.2, where it was found advantageous to use 5-fold cross-validation in order to provide a more robust assessment of model performance. In addition, the results of the cross-validation are presented in Table 6.1. The table includes the average and standard deviation of the evaluation metrics across the five folds. The metric used to select the best-performing model is the Root Mean Squared Error (RMSE), as it measures the overall accuracy in the same unit as the target durations and it gives extra weight to outlier predictions. The model with the lowest average RMSE and the lowest standard deviation is considered the best-performing model for each duration.

6.2.1 Overall Results

After conducting the 5-fold cross-validation, the results reveal the MLP ANN model emerged as the best performer for the unloading duration with an averaged RMSE of 6.19 minutes and a standard deviation of 0.65. Although the other models were close contenders, the MLP ANN model exhibited slightly superior performance. Finally, in the case of the loading duration, the RF model closely stood out as the best-performing model, demonstrating the lowest averaged RMSE of 7.43 and the lowest standard deviation of 0.38 among the models.

Unloading duration					Loading duration				
	MSE:					MSE:			
CrossV fold	RF	XGB	SVR	ANN	CrossV fold	RF	XGB	SVR	ANN
1	33.59	33.52	33.04	32.78	1	49.15	49.31	49.55	49.38
2	48.54	49.54	49.08	47.99	2	62.66	62.93	62.25	62.36
3	36.91	37.09	37.63	36.99	3	54.34	54.55	54.59	53.17
4	28.45	28.22	27.83	28.04	4	49.69	49.81	47.89	49.56
5	47.31	47.73	48.80	47.92	5	61.09	61.90	63.86	66.01
Average	38.96	39.22	39.28	38.75	Average	55.39	55.70	55.63	56.10
StaDev	7.81	8.21	8.48	8.04	StaDev	5.62	5.79	6.47	6.84
	RMSE:					RMSE:			
CrossV fold	RF	XGB	SVR	ANN	CrossV fold	RF	XGB	SVR	ANN
1	5.80	5.79	5.75	5.73	1	7.01	7.02	7.04	7.03
2	6.97	7.04	7.01	6.93	2	7.92	7.93	7.89	7.90
3	6.08	6.09	6.13	6.08	3	7.37	7.39	7.39	7.29
4	5.33	5.31	5.28	5.30	4	7.05	7.06	6.92	7.04
5	6.88	6.91	6.99	6.92	5	7.82	7.87	7.99	8.12
Average	6.21	6.23	6.23	**6.19**	Average	**7.43**	7.45	7.45	7.48
StaDev	0.63	0.66	0.68	**0.65**	StaDev	**0.38**	0.39	0.43	0.45
	R2:					R2:			
CrossV fold	RF	XGB	SVR	ANN	CrossV fold	RF	XGB	SVR	ANN
1	0.27	0.28	0.29	0.29	1	0.33	0.33	0.33	0.33
2	0.27	0.25	0.26	0.28	2	0.31	0.30	0.31	0.31
3	0.28	0.28	0.26	0.28	3	0.31	0.31	0.31	0.32
4	0.30	0.30	0.31	0.31	4	0.32	0.32	0.35	0.33
5	0.24	0.23	0.21	0.23	5	0.29	0.28	0.25	0.23
Average	0.27	0.27	0.27	0.28	Average	0.31	0.31	0.31	0.30
StaDev	0.02	0.03	0.03	0.03	StaDev	0.02	0.02	0.03	0.04
	MAE:					MAE:			
CrossV fold	RF	XGB	SVR	ANN	CrossV fold	RF	XGB	SVR	ANN
1	3.99	3.98	3.81	3.89	1	5.09	5.07	4.88	4.95
2	4.61	4.66	4.54	4.67	2	5.88	5.88	5.71	5.91
3	4.07	4.09	3.99	4.05	3	5.39	5.40	5.21	5.29
4	3.78	3.79	3.59	3.72	4	5.23	5.22	4.98	5.13
5	4.48	4.49	4.35	4.52	5	5.84	5.87	5.85	5.98
Average	4.19	4.20	4.05	4.17	Average	5.48	5.49	5.33	5.45
StaDev	0.31	0.33	0.35	0.37	StaDev	0.32	0.33	0.39	0.42

Table 6.1: Results of the ML models with cross-validation

6.2.2 Assessing Performance of the Selected Models

The unreliability of predictions for the full duration, with a significant RMSE of over 30 minutes indicates substantial variability. The uncertain and variable time gap between activities within the full duration makes accurate estimation and analysis challenging. Therefore, the predictions for the full duration are excluded from further analysis and the focus shifts more toward the more reliable unloading and loading durations, which offer valuable insights for operational planning and resource allocation.

The analysis focuses on residuals, which are the differences between predicted and actual values for unloading and loading durations. Analyzing the residuals for unloading duration predictions reveals an increasing trend in prediction errors with the number of bags, particularly within the range of 100 to 150 bags. The narrow-body aircraft group shows the highest disparities, especially for longer durations due to outlier entries during the cleaning process. Among the narrow-body aircraft, the ”73H” type has the most mistaken predictions and numerous outlier residuals. Weather variables and special items provide limited insights into the residual analysis. The analysis of inbound continents provides limited insights, with most flights originating in Europe and no significant outlier residuals observed.

Regarding the residuals of loading duration predictions, significant differences lie outside the range of (-15, 15) minutes. Over-predictions with the highest residuals, around 20 minutes, tend to increase with the number of bags. Under-predictions do not exhibit a clear trend, but there are instances where over-prediction residuals exceed 30 minutes. The narrow-body group, across all aircraft types, shows notable outlier residuals, while the wide-body planes also display over- or under-prediction with relatively high residuals. Weather-related variables, special items, and outbound continents do not provide further insights. More detailed information on specific airports with the largest residuals can be found in Appendix E

6.3 Comparison between Proposed Model and Available Estimation Models

This section compares the best-performing ML models with existing and data-driven and approximated estimation models for unloading and loading durations. It includes the approximation of the current tool, the improved current tool based on data insights, and the model utilizing average durations for different ranges of number of bags and aircraft types.

The comparison involves evaluating performance metrics and graphs on the overall dataset as well as its subsets based on aircraft groups, types, and duration intervals. Scatter plots are used to assess the fit of prediction values in relation to actual values, examining how closely the data points cluster around the regression line. Boxplot graphs are used to visually represent the deviation of predicted values from actual values (residuals). The boxplot displays the distribution of the residuals, a desirable fit is indicated by a boxplot with a median close to zero, suggesting that the predicted values align closely with the actual

values. Additionally, line plots illustrate the performance of the models across different data subsets, with a selection of fifteen random points from each subset to prevent overcrowding.

6.3.1 Overall comparison

The overall comparison between different duration estimations is conducted using the above-mentioned visualizations for each duration.

The scatter plots in Figure 6.3 provide insights into the characteristics of each model for both unloading and loading durations.

For unloading duration, the approximation of the current tool consistently overpredicts, exhibits significant variability and lacks a clear fit with the actual values. The ANN model performs relatively well for short durations (below 20 minutes), but struggles to capture longer durations. The data-driven current tool's approximation shows improvement, particularly for regional jets, but exhibits variability for other aircraft groups. The scatter plot for the data-driven model with average durations aligns closely with the previous models.

For loading duration, the current tool's approximation shows an unclear spread, overestimating wide-body aircraft and demonstrating overprediction and variability for narrow-body aircraft. The RF model fits better but misses predictions for longer durations. The data-driven current tool's approximation improves upon the initial tool but still exhibits variability and a pattern of over- and underprediction. The scatter plot for the data-driven model with average durations aligns closely with the RF model but struggles with longer durations.

The Box plots in Figure 6.4 confirm that the last three models exhibit similar estimation behavior for both unloading and loading durations. Determining the best performing model based solely on the median is challenging. However, considering the variability and presence of outliers, both the the ML models and the estimation model based on average duration outperform the data-driven current tool's approximation, which displays more outliers. Further analysis is required to determine the optimal estimation model.

6.3.2 Comparison over subsets of data

This analysis is conducted for each duration using various subsets of data such as aircraft groups, aircraft types, and duration ranges. This involves filtering the data and evaluating the performance of the models within these subsets. Above-mentioned visualizations are used visual aid.

Observations from the line plots in Figure 6.5 reveal common trends across the models. Durations below 10 minutes (unloading) or 15 minutes (loading) are consistently overpredicted, with the current tool's approximations showing the highest level of overprediction, followed by the data-driven current tool model. As durations exceed 10 or 15 minutes for unloading and loading respectively, the estimations become more stable and closer to the actual durations, although some differences persist. The ML models and the model utilizing average durations exhibit similar estimation behavior, as seen in the green and purple lines. In the subset for regional jet aircraft, the estimations closely align with the actual un-

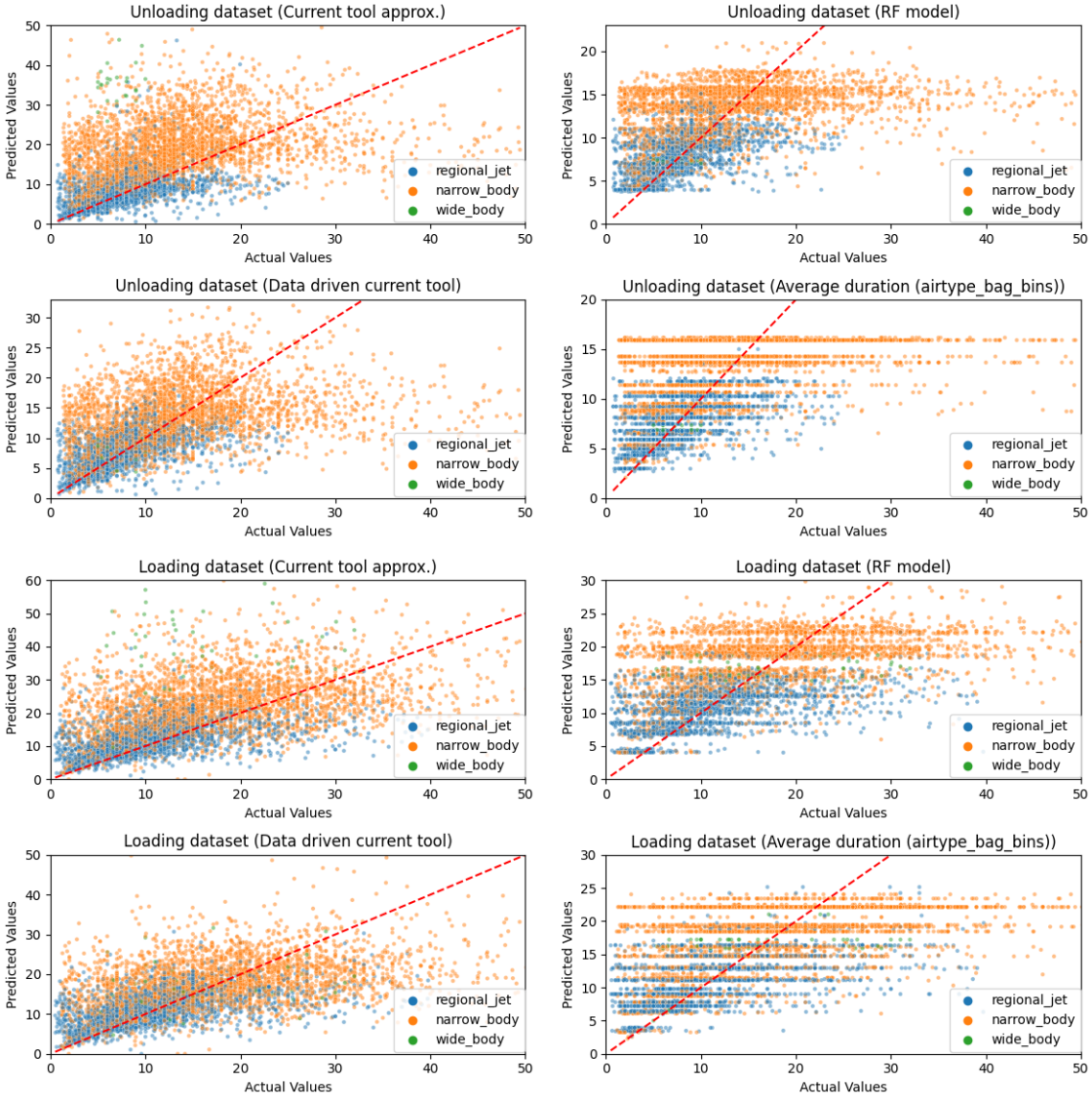


Figure 6.3: Scatter plots comparing the estimations vs the actual durations

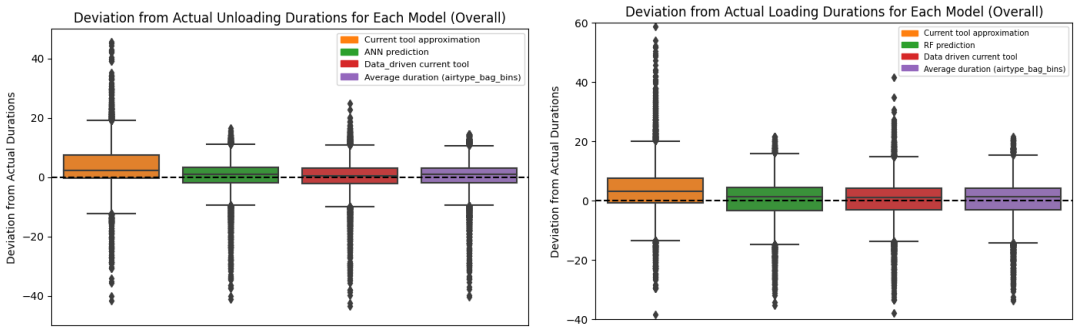


Figure 6.4: Boxplots for prediction residuals

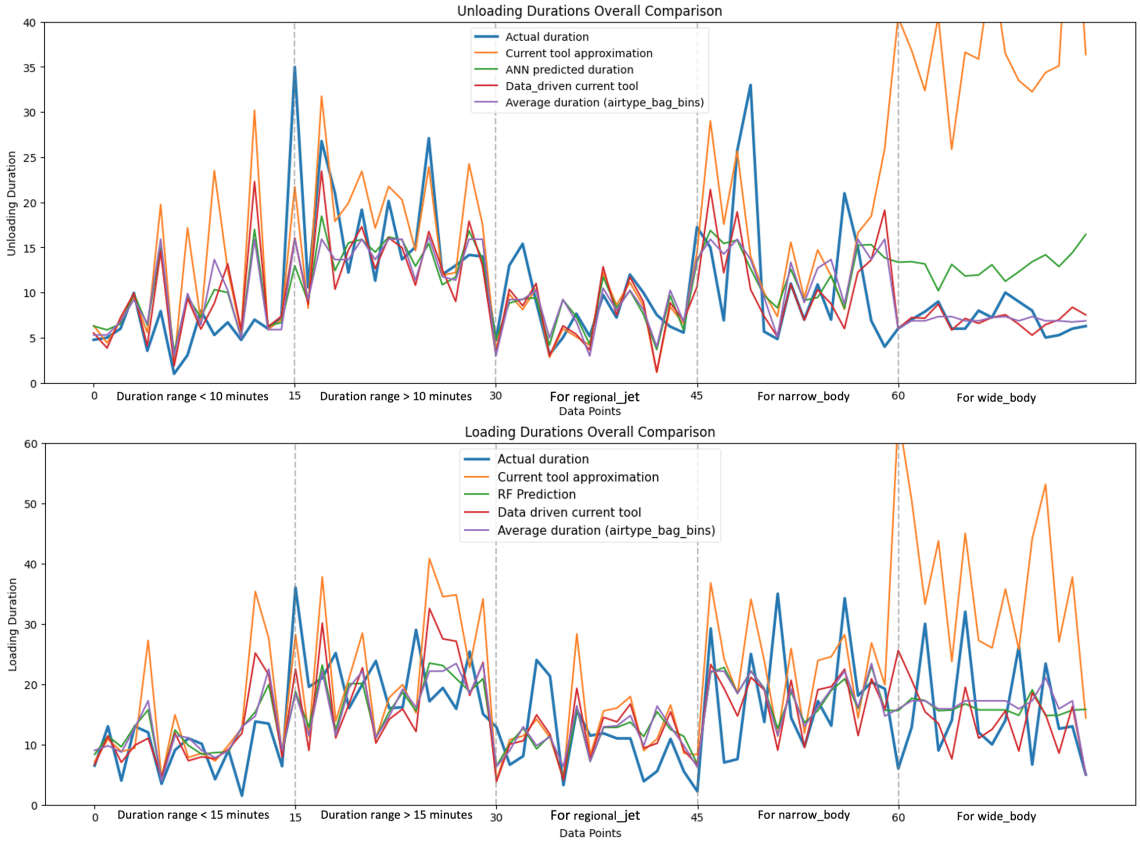


Figure 6.5: Line plots comparing model's estimations and actual durations over subsets of data

loading durations, but this is less clear for loading durations. For narrow-body aircraft, the actual durations display considerable variability, and at certain points, the estimations from the data-driven current tool overlap with these actual durations. In the case of wide-body aircraft, the data-driven models closely align with the actual duration line, indicating better estimations. However, the ML models tends to overpredict, and the current tool's approximation shows significant overprediction.

Given the limited display of only 15 random data points per subset, it is important to interpret the conclusions from these line plots with caution. While they offer a general understanding of the models' performance, they do not provide definitive conclusions. Further analysis using the entire dataset is necessary to obtain more reliable insights. Therefore, a more comprehensive and reliable analysis is conducted using the Box plots in Figure 6.6, visually illustrating the deviation between the models' estimations and the actual duration. The Box plots on the left side depict the deviations from unloading durations, while the Box plots on the right side depict the deviations from the loading duration.

For durations under 10 minutes (unloading) or 15 minutes (loading), all models tend to overpredict, with the current tool's approximation exhibiting the highest variability and the largest number of outliers. The ML models and the model based on subset averages perform similarly and outperform the other models in terms of low presence of outliers outside

the whiskers, although the median of the data-driven current tool is closest to zero in both durations.

For durations exceeding 10 minutes (unloading) or 15 minutes (loading), models tend to underpredict, with the exception of the current tool's approximation, which exhibits significant overprediction and underprediction. The underprediction is more evident in the loading duration, where models also display higher variability. There is no definitive best-performing model based on the median being closest to zero. However, the averages-based model shows fewer outliers compared to the others.

In the case of regional jets, the current tool's approximation shows a median very close to zero for unloading duration and close to zero for loading duration. However, it still has a significant number of outliers for both underprediction and overprediction. For regional jets and narrow-body aircraft, the data-driven current tool demonstrates a median aligned with zero in unloading duration and close to zero in loading duration, it also exhibits similar variability to the ML and averages-based models but has a larger presence of outliers.

Finally, the boxplots for wide-body aircraft confirm that the averages-based model and the data-driven current tool provide close-to-accurate estimations, followed by the ML models, with the ANN model tending to overpredict for the unloading duration and the RF model having a bit more deviation variability for the loading duration, and the current tool's approximation, which consistently overpredicts.

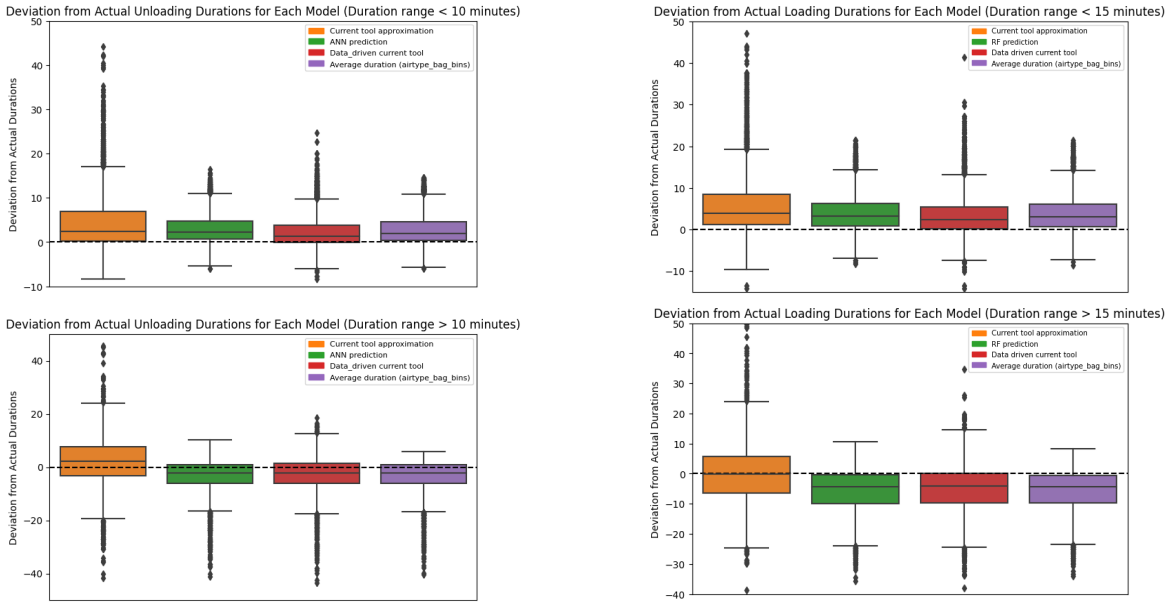


Figure 6.6: Box plots comparing the estimations vs. the actual durations for each subset of data

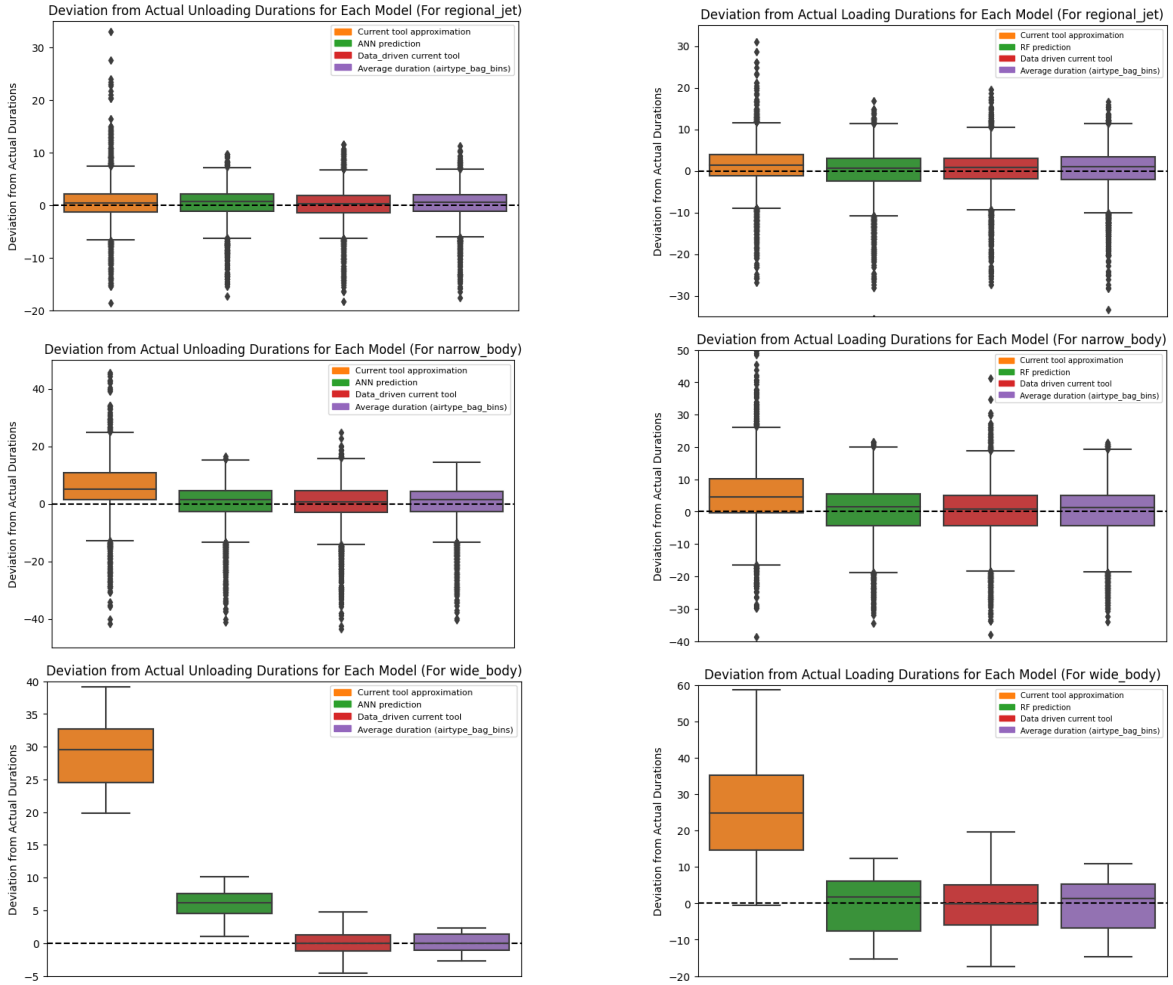


Figure 6.6: Continued from the previous page.

6.3.3 Results from Comparison

The averages-based model and the ML models perform similarly in terms of RMSE, with the averages-based model achieving the best estimation results, followed closely by the data-driven current tool. However, the approximation of the current tool performs significantly worse than the other models. This could be attributed to its reliance on assumptions and user instincts. Interestingly, the ML models do not provide a significant performance improvement compared to the data-driven methods, despite incorporating additional factors, capturing complex relationships, and handling large datasets effectively. As highlighted in Chapter 4, the number of bags, combined with aircraft types, proves to be the most influential variables and yields outstanding results in the averages-based model. Moreover, the data-driven tool shows a significant improvement over the approximation of the current tool, indicating the need to rely on evidence-based estimation rather than unproven assumptions about unloading/loading speeds. The assumption about the strong impact of the number of special items, as assumed by the current tool's approximation, does not hold, leading to considerable over-prediction, particularly in the wide-body aircraft subset.

	Data amount	Estimation model	RMSE	MAE	R-squared
Overall result	**5724**	Current tool approx.	10.43	6.35	-1.03
		ANN predictions	6.22	4.23	0.28
		Data-driven current tool	6.45	4.25	0.22
		Average duration (airtype_bag bins)	**6.18**	4.15	0.29
Duration range (0-10 minutes)	**3229**	Current tool approx.	9.73	5.37	-16.89
		ANN predictions	4.63	3.43	-3.05
		Data-driven current tool	**4.6**	3.08	-2.99
		Average duration (airtype_bag bins)	**4.6**	3.33	-2.99
Duration range (10+ minutes)	**1936**	Current tool approx.	11.28	7.62	-1.66
		ANN predictions	7.82	5.26	-0.28
		Data-driven current tool	8.25	5.76	-0.43
		Average duration (airtype_bag bins)	**7.77**	5.21	-0.26
regional_jet	**2346**	Current tool approx.	4.28	2.78	-0.13
		ANN predictions	3.46	2.5	0.26
		Data-driven current tool	3.49	2.44	0.25
		Average duration (airtype_bag bins)	**3.4**	2.42	0.28
narrow_body	**3348**	Current tool approx.	12.73	8.62	-1.42
		ANN predictions	7.58	5.42	0.14
		Data-driven current tool	7.91	5.55	0.06
		Average duration (airtype_bag bins)	**7.56**	5.38	0.14
wide_body	**30**	Current tool approx.	35.3	32.82	-540.3
		ANN predictions	6.35	5.92	-16.53
		Data-driven current tool	2	1.55	-0.73
		Average duration (airtype_bag bins)	**1.49**	1.28	0.03

Table 6.2: Final results from estimation comparison in unloading duration

	Data amount	Estimation model	RMSE	MAE	R-squared
Overall result	**5801**	Current tool approx.	10.29	7.05	-0.3
		RF predictions	7.44	5.48	0.32
		Data driven current tool	7.65	5.48	0.28
		Average duration (airtype_bag bins)	**7.38**	5.41	0.33
Duration range (0-15 minutes)	**3433**	Current tool approx.	10.1	6.54	-6.56
		RF predictions	6.03	4.55	-1.7
		Data driven current tool	6.14	4.31	-1.79
		Average duration (airtype_bag bins)	**5.98**	4.47	-1.65
Duration range (15+ minutes)	**2368**	Current tool approx.	10.54	7.79	-1.23
		RF predictions	9.1	6.84	-0.66
		Data driven current tool	9.43	7.18	-0.78
		Average duration (airtype_bag bins)	**9.03**	6.79	-0.64
regional_jet	**2335**	Current tool approx.	6.34	4.36	0.11
		RF predictions	5.86	4.11	0.24
		Data driven current tool	**5.68**	3.97	0.29
		Average duration (airtype_bag bins)	5.77	4.08	0.26
narrow_body	**3413**	Current tool approx.	11.74	8.59	-0.51
		RF predictions	8.34	6.4	0.24
		Data driven current tool	8.74	6.49	0.16
		Average duration (airtype_bag bins)	**8.3**	6.31	0.25
wide_body	**53**	Current tool approx.	30.47	26.32	-14.02
		RF predictions	7.9	6.84	-0.01
		Data driven current tool	8.51	6.76	-0.17
		Average duration (airtype_bag bins)	**7.43**	6.12	0.11

Table 6.3: Final results from estimation comparison in unloading duration

ML models and data-driven methods perform well for unloading durations under 10 minutes and for regional-jet aircraft, achieving a low RMSE of 4.6. Among the four models, all show relatively good performance for regional jet aircraft, with the worst RMSE of 4.28 from the current tool approximation and 3.4 from the averages-based model. Notably, data-driven models excel for wide-body aircraft, with the averages-based model achieving an RMSE of 1.49 and the data-driven current tool having an RMSE of 2. On the other hand, the current tool approximation significantly overestimates durations in this subset, resulting in a large RMSE of 35. All models struggle to accurately estimate durations exceeding 10 minutes and for narrow-body planes, where the best RMSE is around 7.6. This presents a challenge as a considerable amount of data belongs to narrow-body aircraft, indicating the presence of notable outlier predicted-actual differences within this aircraft group.

Similarly, the loading duration is best predicted for durations under 15 minutes and for regional jets. However, it has a higher overall best RMSE compared to unloading duration, indicating increased difficulty in prediction. Other than that, the behavior of the results for this duration follow a similar pattern to unloading durations.

6.4 Integration Plan for the Proposed Model

Even though, no significant improvement when using the ML models was observed in comparison to simpler data-driven methods, the following integration plan focuses on seamlessly incorporating the proposed prediction model or a data-driven model into the existing Terra tool for enhanced unloading and loading duration estimations. This section provides a high-level overview of the integration plan, highlighting the key steps for successful integration and utilization of the model within the Terra tool.

- **Model Deployment and Infrastructure**: Since the ML and data-driven prediction models give estimated durations similar as the current tool, these can use the existing infrastructure and connections employed by the current model, ensuring a smooth transition of information to the Terra tool.

- **Code Transfer**: The model, along with its associated code and instructions, will be shared with the Terra team. A detailed README file will be provided, outlining the necessary steps for data preparation, feature selection, hyperparameter tuning, model training, prediction generation, and model evaluation. Additionally, excel files will be provided containing the average values that data-driven methods use.

- **Model Training and Utilization**: The proposed model has undergone rigorous feature selection and hyperparameter tuning processes. As long as there are no significant changes to the data features, the model can be directly utilized with the current set of features and hyperparameters. However, in the event of new feature acquisition, the data will undergo preprocessing and preparation, and the model will require feature selection and hyperparameter tuning to ensure optimal performance.

6.5 Conclusion

Four ML models were evaluated and compared in this chapter using a 5-fold cross-validation in Section 6.1. Later in Section 6.2, the MLP ANN emerged as the best model for predicting the unloading duration, and RF for the loading duration, in addition, residual analysis for unloading and loading durations revealed trends and outliers in prediction errors, particularly related to the number of bags and aircraft types. Insights from weather variables, special items, and outbound continents were limited. More detailed airport-specific information can be found in Appendix E. Finally, Section 6.3 shows the performance of the best-performing ML models in comparison with that of data-driven and approximated estimation methods, where the approximation of the current tool was largely outperformed by data-driven and ML modes, and ML models showed similar or slightly lower performance than data-driven models. However, there is still room for improvement for data-driven and ML models, especially in predicting large durations and outlier values. Lastly, a high-level integration plan was proposed for seamless integration of the best performing models into the Terra tool's resource planning and allocation process.

Chapter 7

Conclusion and Discussion

This thesis work was undertaken in collaboration with the Business Platform Ground of KLM's Data & Technology department. This chapter presents the major findings from the research by addressing the research questions raised in Section 1.3. It is followed by a discussion on the impact of this study on science and practice. Furthermore, the limitations of the study, potential future work, and recommendations are discussed.

7.1 Conclusion

The research focuses on accurately predicting baggage loading and unloading durations for the KLM turnaround process using data from airport cameras, aircraft sensors, flights, and weather. A customized version of the CRISP-DM methodology is applied, emphasizing system comprehension, data preparation, and empirical evaluation. The literature review offers essential background information. Additionally, the current estimation tool is described and approximated, and insights and data-driven estimation methods are retrieved from data extraction, cleaning, preprocessing, and transformation. Afterward, techniques such as correlation analysis, multicollinearity removal, visual analysis, preprocessing, and feature selection are employed to identify influential features. Subsequently, careful selection and preparation of machine learning models contribute to improving estimation accuracy. These models are evaluated and compared with the current tool approximation and data-driven methods using appropriate metrics. Finally, an integration strategy seamlessly incorporates predictions from the best model into the existing Terra tool, ensuring a smooth integration of the improved estimation process.

The main research question was: **"How can data-driven methods using camera and sensor data enhance the accuracy of KLM's baggage duration predictions at Schiphol Airport?"**. Chapters 3 to 6 provide key insights and directly address the main research question and the sub-research questions (RQs) are presented as follows:

(**RQ1**) **What factors impact baggage unloading and loading efficiency in aircraft turnaround, and how does the accurate estimation of loading and unloading times contribute to improved efficiency?**

The study identified internal or flight-specific and external or associated with the airport environment and operations which have an effect on both baggage handling and the aircraft turnaround process and can be found in 2.1. Moreover, it was found that accurate estimation of baggage handling times directly impacts on-time departure performance, enhances the overall travel experience for passengers, and reduces unnecessary costs or inefficiencies. Moreover, this estimation process becomes crucial in optimizing resource allocation by ensuring the timely availability of suitable personnel and equipment at designated locations. Moreover, it serves as a valuable tool for identifying bottlenecks within the baggage handling process for targeted improvements in operations.

(**RQ2**) **What data-driven techniques for data analysis and feature selection can be used to understand and prepare the given data for predictive modeling?**
This study revealed that descriptive statistics and visualizations aid in understanding data characteristics. Additionally, inferential statistics, such as correlation analysis and T-test/ANOVA, effectively identify patterns and generate insights. Moreover, techniques like Variance Inflation Factor and Lambda's association factor can be employed to address highly correlated features. Finally, feature selection methods, including filter techniques and wrapper methods, were found to be useful for identifying relevant features.

(**RQ3**) **What are the various types of prediction models currently utilized in the industry, and which among them are suitable for accurately predicting baggage loading and unloading task duration?**
Regression was found to be the most suitable framework for baggage handling prediction, considering flight-related, airport-related, weather-related, and time-related features. Moreover, the selected machine learning models (Random Forest, XGBoost, SVR, MLP ANN) capture complex relationships and are robust, scalable, and flexible for handling large datasets.

(**RQ4**) **What evaluation metrics and design considerations are important for predictive modeling in baggage unloading and loading duration?**
Design considerations such as the bias-variance trade-off in model development involves balancing a model's simplicity and ability to generalize. High bias means underfitting, where the model is too simple, while high variance leads to overfitting, where the model is too complex and fails to generalize well. Regularization methods like L1 and L2 can help control overfitting, while more complex models and improved data representation can address underfitting. Additionally, the evaluation metrics prioritized RMSE due to its alignment with the target variable, interpretability, and outlier handling. Finally, cross-validation was emphasized to enhance result reliability and mitigate overfitting.

Additionally, the examination of evaluation metrics prioritized the RMSE metric due to its alignment with the target variable's unit, interpretability, and ability to handle outliers. Finally, the importance of cross-validation was emphasized to enhance result reliability and mitigate the risk of overfitting.

(RQ5) How does KLM currently estimate the baggage unloading and loading durations?
KLM's current estimation tool considers factors like dead load, total number of bags, special

handling items, and gate-checked handbags to estimate baggage handling time. Moreover, the current tool has limitations in handling unexpected events, missing data, limited features, capturing nonlinear relationships, scalability, measurability, and data storage capabilities. These limitations undermine its ability to provide accurate and reliable predictions.

Furthermore, an approximation of the tool was developed in response to its lack of storage capabilities for understanding and comparison with proposed prediction models. Finally, using the RMSE metric, the accuracy evaluation in Chapter 6 showed significant errors of 10.43 for unloading and 10.29 for loading durations, indicating challenges faced by KLM.

(RQ6) How can data collected from cameras and sensors in the baggage loading and unloading process be effectively prepared for analysis in the prediction tool?
Data was collected from three sources: DeepTurnaround (airport camera data), Flight-720 (flight-related and sensors data), and the Royal Dutch Meteorological Institute (weather-related data). The data transformation process addressed challenges like missing and duplicate timestamps, as well as time gaps. Cleaning involved comparing DeepTurnaround data with ACARS data, setting outlier bounds, and removing columns with inadequate data.

(RQ7) How can statistical analysis and feature selection techniques aid in identifying influencing features for the target variable at the time of prediction?
The process of identifying influential features started with deriving new features to enhance explanatory power. Subsequently, correlation analysis and multicollinearity cleaning were performed. Then, exploratory data analysis provided insights and simplified categorical features. Moreover, preprocessing involved one-hot encoding and scaling. Finally, feature selection models determined the optimal subset of features for each prediction model. See Table 4.6 for the final list of variables.

(RQ8) How can the prediction models be adequately prepared and trained to ensure accurate estimation of baggage unloading and loading durations?
Initially, the dataset was split into training and testing sets using a 75-25% ratio to prevent overfitting or underfitting. then, a 5-fold cross-validation was performed to assess the models' validity and generalizability. Finally, the optimal hyperparameter configuration for each model was determined through a Randomized search method with 100-150 iterations, in each iteration using 5-fold cross-validation and with the MSE metric for evaluation The final hyperparameter tuning results and the sensitivity of this process can be found in 5.1 and in the in Appendix D.

(RQ9) How to assess and compare prediction models, identify the best performer, and extract key insights for further improving the top-performing model?
ML models were evaluated and compared using RMSE results from 5-fold cross-validation. XGBoost performed best for the full duration, MLP ANN excelled in the unloading duration, and RF stood out for the loading duration. Furthermore, the ML models consistently outperformed the current tool's approximation, with significant reductions in RMSE for unloading and loading durations. The accuracy of the full duration was uncertain due to a time gap that the current tool cannot predict. Finally, improvements are needed for accurately predicting large durations, as errors proportionally increased with the number of bags and were observed in both narrow-body and wide-body planes.

(RQ10) What documentation and guidelines facilitate a smooth integration of the new prediction model with the Terra resource planning tool for KLM?
The integration plan aims to seamlessly integrate the proposed prediction model into the existing Terra tool for improved unloading and loading duration estimations. The plan includes steps such as leveraging the current infrastructure, transferring the model code and instructions to the Terra team, and providing guidance on data preparation, feature selection, hyperparameter tuning, model training, prediction generation, and model evaluation. The proposed model, with its optimized features and hyperparameters, can be directly utilized with the current data, but adjustments may be needed if new data is acquired.

7.2 Discussions

7.2.1 Influential Features

The initial current tool approximation considered the weight of bags, number of special items, and assumed durations for each aircraft type. However, the data-driven tool, incorporating average speeds based on historical data, showed significant improvement by reducing the impact of special items and assumed durations.

Further analysis using ML models and the inclusion of time-related, weather-related, and flight-based variables resulted in filtered features and addressed multicollinearity in Section 4.2. The number of bags was preferred over bag weight for practicality, the aircraft group variable demonstrated higher explanatory power compared to specific aircraft types, and the classification of the airport variables into continents proved useful, considering the large number of airports. The feature selection models in Section 4 consistently suggested distinguishing the wide-body aircraft group, and it was found that tree-based ML models favored fewer variables, particularly the number of bags and aircraft group, and outbound continents in the case of the RF model for the loading duration. Additionally, time-related features, categorized into peak, medium, and low peak hours of the day, were highly preferred by the ML models. Weather-related features, with rain being the most influential, showed limited explanatory power.

The best-performing model relied on average duration for each combination of aircraft type and bag range, surpassing the performance of the ML models. This indicates that the number of bags and aircraft type are the most influential features for predicting unloading and loading durations, with similar results expected using bag weight or aircraft groups.

7.2.2 Machine Learning Models

Initially, it was expected that the ML models would be the most accurate for predicting baggage unloading and loading durations due to their complexity. However, during the development phase, it became apparent that there were no highly influential features that could accurately predict these durations. The variability in the number of bags and the absence of key variables related to workers and loading machinery in the data may have contributed to

the similar accuracy between ML models and data-driven models. Weather-related variables also did not have a significant impact. Limited data, especially for wide-body aircraft, could have hindered the ML models' performance. Additionally, the simplicity of the relationships between features and target durations may have limited the improvements of ML models over simpler average duration models. It should be noted that ML models require data without hints about the target durations to avoid overfitting, while data-driven models, based on average durations, clearly takes the target variables into account. This should explain why data-driven methods can be better but it can also recognize that their long-term reliability may not match that of ML models.

7.2.3 Data-driven models and Current Tool Approximation

The current tool approximation performs poorly compared to even simple data-driven methods, such as averaging durations for each aircraft type. This project provides valuable insights into unloading and loading speeds through data analysis, which can enhance the current tool. However, data-driven methods do not adequately consider the impact of special items since there is no specific data on their unloading or loading durations. The project explores including all items, check-in bags, hand bags, and special items, as a feature, but separate inclusion of bags and handbags, and special items yields similar or slightly better results.

7.2.4 Performance and Impact of Estimation Models

Table 7.1 offers an assessment of various estimation models and their impact on planning and resource allocation, considering an example where KLM aims to ensure sufficient capacity in 95 percent of cases. It presents the minimum and maximum error bounds for each model, along with the overpredicted and underpredicted hours, all at a 95 percent confidence interval. These metrics provide valuable insights into the accuracy and reliability of each model, supporting effective resource planning for KLM. As a result, models with less overpredicted hours improve resource allocation accuracy, preventing wasteful utilization and unnecessary costs. Also, they optimize resources, boosting efficiency, reducing idle time, and increasing profitability. On the other hand, models with less underpredicted hours prevent resource shortages and capacity constraints.

Table 7.1 highlights the performance differences among the estimation models. The current tool approximation exhibits the lowest underprediction, while the ML models and data-driven methods significantly improve overprediction. This improvement is evidenced by over 200 additional hours allocated to the current tool approximation in comparison with the other models. However, caution is advised with the data-driven and ML models as they may result in 70-90 more hours of underprediction compared to the current tool approximation. Notably, the averages-based model achieves the best balance between underprediction and overprediction, followed by the data-driven current tool, ML models, and current tool approximation.

Performance of Models in Unloading with 95%CI				
Estimation Model	**Lower Bound**	**Upper Bound**	**Overprediction Hours**	**Underprediction Hours**
Current tool approx.	-12.49	21.9	397.71	-77.31
ANN predictions	-15.94	10.47	179.21	-139.33
Data-driven current tool	-16.32	11.88	168.91	-146.01
Average duration (airtype_bag bins)	-16.09	10.86	168.22	-142.5
Performance of Models in Loading with 95%CI				
Estimation Model	**Lower Bound**	**Upper Bound**	**Overprediction Hours**	**Underprediction Hours**
Current tool approx.	-14.34	24.06	427.41	-122.43
ANN predictions	-18.04	13.41	226.11	-210.05
Data-driven current tool	-17.73	14.7	219.69	-209.98
Average duration (airtype_bag bins)	-17.74	13.18	222.85	-207.06

Table 7.1: Performance Analysis of Estimation Models

7.3 Contribution

This project has provided valuable insights into the baggage handling process, shedding light on its significance, which has often been underestimated and overlooked. Through the application of data-driven and machine learning methods, this study has demonstrated the potential of accurately estimating baggage handling durations and identifying influential features. The subsequent subsections delve into the scientific and practical contributions of this project

7.3.1 Scientific Contribution

This research project makes significant scientific contributions to the field of data-driven decision-making and Artificial Intelligence (AI) in airline baggage handling operations. It addresses several key challenges and provides valuable insights and methodologies. Firstly, the project tackles the issue of handling duplicate timestamps in datasets encountered in surveillance or recording cameras and IoT systems, which can occur due to flickering or system malfunctions. To overcome this issue, it introduces an approach that ensures event sequencing, which can be used in domains where precise temporal information is essential for analysis and prediction. Furthermore, the project conducts a comprehensive empirical analysis of the baggage handling process that addresses multicollinearity and selects the most influential variables. Lastly, the project contributes to the field of predictive modeling in baggage handling operations by developing three advanced feature selection models and evaluating four state-of-the-art machine learning algorithms (Random Forest, XGBoost, SVR, and Multilayer Perceptron Artificial Neural Network). Through this evaluation, the project provides valuable insights into the strengths and weaknesses of these models in this particular operation, aiding researchers and practitioners in selecting the most suitable approach for baggage duration prediction.

7.3.2 Practical Contribution

From a practical standpoint, this research project has implications for the aviation industry, particularly for KLM and other airlines. Firstly, the project addresses the challenge of preprocessing and handling duplicate timestamps, ensuring accurate event sequencing from camera data at Schiphol Airport. Moreover, it delves deep into the KLM baggage handling process, identifying the most influential features and gaining valuable insights. By considering various factors such as flight-related data, airport-related data, time-related data, and weather-related data, the project not only captures the complexities and interdependencies of airline operations in a prediction model but also allows us to get insights into the process. By improving predictions of the baggage unloading and loading durations and identifying areas of improvement, the project directly tackles inefficiencies in the planning process with data-driven support. This can enable KLM to streamline ground operations planning, resulting in increased operational efficiency, cost savings, and improved customer satisfaction. Finally, the project's incorporation of data-driven and predictive modeling techniques paves the way for automated decision-making processes, reducing reliance on manual interventions and enhancing prediction accuracy.

7.4 Limitations

The overall available data has limitations due to its scarcity, covering only specific months, which may not capture the complete year-round patterns. Additionally, a significant amount of data had to be excluded due to unreliability, raising concerns about the dataset's representativeness. Additionally, the data for wide-body aircraft is extremely limited, compromising statistical significance and generalizability for this aircraft category. Subsequently, data limitations discovered in each step of the process are discussed.

DeepTurnaround data preprocessing

The accuracy of capturing the start and stop times of the baggage handling process using airport cameras is uncertain. The criteria for camera recording are unclear, introducing discrepancies and uncertainty in identifying precise timings. This lack of clarity impacts the reliability of the analysis. Additionally, data quality issues, such as camera flickering, network connectivity problems, software glitches, and data entry errors, compromise the accuracy of the dataset and introduce noise or missing data, further affecting the reliability of the results. The accuracy of the transformation approach relies heavily on the quality and consistency of the timestamps in the raw data. Any errors or inconsistencies in the data could lead to incorrect splits and inaccurate activity durations. Moreover, the determination of the minimum total unloading-loading time (Y) and the maximum time gap (X) relies on expert input and assumptions. These thresholds may not capture all scenarios accurately and could vary across different types of aircraft or operational conditions. Moreover, the handling of unclassified event timestamps has a rule that assumes that the start and stop

timestamps fall within the initial/final 10 percent of the total turnaround time and within five minutes of the observed airplane doors opening/closing time. However, these assumptions may not hold true in all turnaround scenarios, potentially leading to the misclassification of unloading and loading events.

Data from Flight-720

During data collection from the Flight 720 platform, discrepancies emerged when summing separate columns for check-in bags and comparing them with the total check-in bags. Similarly, the columns for ”total bags loaded” and ”total bags not yet loaded” did not consistently align with the previous counts. Consequently, the addition of separate columns for check-in bags was chosen due to its reliability and consistency. Moreover, the weight information lacked specific details for different bag types, casting doubt on its accuracy. These limitations can affect the accuracy of the bag count and weight data used for analysis.

Reliability Analysis

ACARS data from the aircraft sensors, while reliable, has limitations due to factors like overnight turnarounds, long delays, and exceptional circumstances. Outlier observations raise concerns about its accuracy, impacting the assessment of DeepTurnaround data reliability. Moreover, camera-recorded cargo door opening times are shorter than those obtained from aircraft sensors, creating uncertainties in DeepTurnaround data accuracy. The allowed error range of -10 to 5 minutes accounts for potential camera-related variations but lacks a precise understanding of the actual differences. Other sources of measurement errors, like sensor accuracy and data transmission delays, are not considered. Further investigation is needed to comprehensively assess data reliability.

7.4.1 Model Limitations

The analysis’s empirical nature relies heavily on practical experience and observation, introducing subjectivity and potential biases in the initial search for influential variables and multicollinearity cleaning. Moreover, the limited focus on three feature selection methods may overlook better alternatives like exhaustive search or Genetic Algorithm. However, the potential improvement from these methods is deemed insignificant, balancing computational complexity and performance gains.

In addition, potential but well accounted for limitations for this project can include the analysis overlooks the possibility of a separate prediction model for specific and distinct subsets of data such as aircraft groups, potentially missing valuable insights. Nevertheless, this is a calculated risk since the ML models used can handle underlying patterns and non-linear relationships. Furthermore, the range of hyperparameter evaluation for the ML models is limited, with a randomized search of 100-150 iterations, potentially overlooking optimal hyperparameter settings. Finally, another limitation is the lack of reliable uncertainty

quantification in the point estimate prediction approach using ML models. This prevents analyzing a range of possible values or confidence levels for predictions and overlooks potential variability in the data.

Lastly, relying on an approximation for the current tool's durations instead of actual values can introduce errors and biases, limiting the accuracy and reliability of insights into the tool's performance. Moreover, comparisons between the proposed model and the approximation may not accurately reflect the true performance improvement achieved.

7.5 Future Work and Recommendations

In order to address the limitations of this study and further refine the predictions of baggage handling durations, several areas of future work and recommendations have been identified.

1. Data Collection, Preprocessing and Validation:
    (a) Develop outlier detection algorithms or statistical techniques to effectively identify and handle outlier observations in ACARS data.
    (b) Investigate discrepancies between camera-recorded cargo door opening times and those obtained from aircraft sensors, considering factors like sensor accuracy, data transmission delays, and potential measurement errors to refine the accuracy of the DeepTurnaround data.
    (c) Gather more data specifically for wide-body aircraft to enhance statistical significance and generalizability.
    (d) Validate and reconcile important columns such as the number of bags or weight of bags for data reliability.
    (e) Enhance DeepTurnaround data quality through regular maintenance, calibration, improved network connectivity, and bug fixes. In addition, collect a year-round extended and diverse dataset to cover all seasons and months, and capture accurate year-round patterns.
    (f) When splitting unloading and loading activities from duplicate event's timestamps, refine thresholds for minimum unloading-loading time and maximum time gap, considering aircraft type, operational conditions, and expert input to accurately capture a wider range of turnaround scenarios.
2. Refinement of Prediction Models:
    (a) Incorporate the number of workers involved in the baggage handling process in the model, once reliable data becomes available. This would allow Terra to input a specific number of workers and get a task time estimation accordingly.
    (b) Expand hyperparameter evaluation through a more extensive search, larger iterations, or advanced optimization algorithms to identify optimal settings and enhance model performance.

(c) Implement prediction intervals rather than point estimate predictions with methods like bootstrapping, Bayesian inference, or ensemble modeling to provide a range of values or confidence levels for predictions and capture more effectively the associated uncertainty.

Bibliography

Andrews, B.H., Dean, M.D., Swain, R., Cole, C., 2013. Building arima and arimax models for predicting long-term disability benefit application rates in the public/private sectors. Society of Actuaries , 1–54.

Ashford, N.J., Martin, S.H.P., Moore, C.A., 2013. Airport operations. McGraw-Hill Professional.

Ashford, N.J., Mumayiz, S., Wright, P.H., 2011. Airport engineering: planning, design, and development of 21st century airports. John Wiley & Sons.

AviationLearnings, 2020. How cargo baggage is loaded unloaded from an aircraft – the role of aircraft cargo loaders, belt loaders, cargo dollies ulds. URL: https://aviationlearnings.com/how-cargo-is-loaded-and-unloaded-from-an-airplane/.

Burkov, A., 2019. The hundred-page machine learning book. volume 1. Andriy Burkov Quebec City, QC, Canada.

Carpinteiro, O.A., Leite, J.P., Pinheiro, C.A., Lima, I., 2012. Forecasting models for prediction in time series. Artificial Intelligence Review 38, 163–171.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Crone, S.F., Kourentzes, N., 2010. Feature selection for time series prediction–a combined filter and wrapper approach for neural networks. Neurocomputing 73, 1923–1936.

De Neufville, R., 1994. The baggage system at denver: prospects and lessons. Journal of Air Transport Management 1, 229–236.

Diana, T., 2018. Can machines learn how to forecast taxi-out time? a comparison of predictive models applied to the case of seattle/tacoma international airport. Transportation Research Part E: Logistics and Transportation Review 119, 149–164.

Dorogush, A.V., Ershov, V., Gulin, A., 2018. Catboost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 .

Evler, J., Asadi, E., Preis, H., Fricke, H., 2021. Airline ground operations: Schedule recovery optimization approach with constrained resources. Transportation Research Part C: Emerging Technologies 128, 103129.

Evler, J., Lindner, M., Fricke, H., Schultz, M., 2022. Integration of turnaround and aircraft recovery to mitigate delay propagation in airline networks. Computers & Operations Research 138, 105602.

Frey, M., 2014. Models and methods for optimizing baggage handling at airports. Ph.D. thesis. Technische Universität München.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

Gao, Y., Huyan, Z., Ju, F., 2015. A prediction method based on neural network for flight turnaround time at airport, in: 2015 8th International Symposium on Computational Intelligence and Design (ISCID), IEEE. pp. 219–222.

Géron, A., 2017. Hands-on machine learning with scikit-learn and tensorflow: Concepts. Tools, and Techniques to build intelligent systems .

Ghasemi, A., Zahediasl, S., 2012. Normality tests for statistical analysis: a guide for non-statisticians. International journal of endocrinology and metabolism 10, 486.

Goodman, L.A., Kruskal, W.H., Goodman, L.A., Kruskal, W.H., 1979. Measures of association for cross classifications. Springer.

Han, J., Kamber, M., Pei, J., 2012. Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University .

Hassel, O.v., 2019. Predicting the turnaround time of an aircraft : a process structure aware approach l. URL: `https://research.tue.nl/en/studentTheses/predicting-the-turnaround-time-of-an-aircraft`.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction. volume 2. Springer.

Horstmeier, T., de Haan, F., 2001. Influence of ground handling on turn round time of new large aircraft. Aircraft Engineering and Aerospace Technology 73, 266–271.

Hutter, L., Jaehn, F., Neumann, S., 2019. Influencing factors on airplane boarding times. Omega 87, 177–190.

Jasra, S., Gauci, J., Muscat, A., Valentino, G., Zammit-Mangion, D., Camilleri, R., 2018. Literature review of machine learning techniques to analyse flight data .

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30.

Koninklijk Nederlands Meteorologisch Instituut, . Klimaat van nederland. [Online]. URL: https://www.knmi.nl/klimaat.

Kuhn, M., Johnson, K., 2021. Feature engineering and selection: A practical approach for predictive models. Chapman amp; Hall/CRC.

Luo, M., Schultz, M., Fricke, H., Desart, B., Herrema, F., Montes, R.B., 2021. Agent-based simulation for aircraft stand operations to predict ground time using machine learning, in: 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), IEEE. pp. 1–8.

Mirza, M., 2008. Economic impact of airplane turntimes. Aero Quarterly 4, 2008.

More, D., Sharma, R., 2014. The turnaround time of an aircraft: a competitive weapon for an airline company. Decision 41, 489–497.

Mota, M.M., Boosten, G., De Bock, N., Jimenez, E., de Sousa, J.P., 2017. Simulation-based turnaround evaluation for lelystad airport. Journal of Air Transport Management 64, 21–32.

Nayak, B.K., Hazra, A., 2011. How to choose the right statistical test? Indian journal of ophthalmology 59, 85.

Neumann, S., 2019. Is the boarding process on the critical path of the airplane turn-around? European Journal of Operational Research 277, 128–137. doi:10.1016/j.ejor.2019.02.001.

Oreschko, B., Kunze, T., Schultz, M., Fricke, H., Kumar, V., Sherry, L., 2012. Turnaround prediction with stochastic process times and airport specific delay pattern, in: International Conference on Research in Airport Transportation (ICRAT), Berkeley.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Postorino, M.N., Mantecchini, L., Malandri, C., Paganelli, F., 2020. A methodological framework to evaluate the impact of disruptions on airport turnaround operations: A case study. Case Studies on Transport Policy 8, 429–439. doi:10.1016/j.cstp.2020.03.007.

Raschka, S., 2018. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. The Journal of Open Source Software 3. URL: https://joss.theoj.org/papers/10.21105/joss.00638, doi:10.21105/joss.00638.

Rizal, R., 2016. Aircraft turnaround: A descriptive study and analysis optimisation using mathematical model. Available at SSRN 2851286 .

Sammut, C., Webb, G.I., 2011. Encyclopedia of machine learning. Springer Science & Business Media.

San Antonio, A., Juan, A.A., Calvet, L., i Casas, P.F., Guimarans, D., 2017. Using simulation to estimate critical paths and survival functions in aircraft turnaround processes, in: 2017 Winter Simulation Conference (WSC), IEEE. pp. 3394–3403.

Schmidt, M., 2017. A review of aircraft turnaround operations and simulations. Progress in Aerospace Sciences 92, 25–38. doi:10.1016/j.paerosci.2017.05.002.

Schultz, M., Fricke, H., 2008. Improving aircraft turnaround reliability. URL: https://www.researchgate.net/publication/263038959_Improving_Aircraft_Turnaround_Reliability.

Schultz, M., Reitmann, S., 2019. Machine learning approach to predict aircraft boarding. Transportation Research Part C: Emerging Technologies 98, 391–408.

Schultz, M., Reitmann, S., Alam, S., 2021. Predictive classification and understanding of weather impact on airport performance through machine learning. Transportation Research Part C: Emerging Technologies 131, 103119. doi:10.1016/j.trc.2021.103119.

Schwabish, J., 2021. Better data visualizations: A guide for scholars, researchers, and wonks. Columbia University Press.

Somyanonthanakul, R., Warin, K., Amasiri, W., Mairiang, K., Mingmalairak, C., Panichkitkosolkul, W., Silanun, K., Theeramunkong, T., Nitikraipot, S., Suebnukarn, S., 2022. Forecasting covid-19 cases using time series modeling and association rule mining. BMC Medical Research Methodology 22, 281.

Staudemeyer, R.C., Morris, E.R., 2019. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586 .

Sutton, C.D., 2005. Classification and regression trees, bagging, and boosting. Handbook of statistics 24, 303–329.

Szabo, S., Pilát, M., Makó, S., Korba, P., Čičváková, M., Kmec, L., 2022. Increasing the efficiency of aircraft ground handling—a case study. Aerospace 9, 2.

Tarău, A., De Schutter, B., Hellendoorn, J., 2009. Route choice control of automated baggage handling systems. Transportation research record 2106, 76–82.

Tarău, A., De Schutter, B., Hellendoorn, J., 2010. Predictive control for baggage handling systems using mixed integer linear programming. IFAC Proceedings Volumes 43, 16–21.

Timajo, L., Chakraborty, S., Chakraborty, B., 2014. Analysis of aircraft turnaround time. Europian Academic Research 2, 9982–9988.

Towards AI Team, 2022. Data Science Essentials - Multicollinearity. https://towardsai.net/p/l/data-science-essentials-multicollinearity. Accessed: June 6, 2023.

Volt, J., Stojić, S., Had, P., 2022. Optimization of the baggage loading and unloading equipment. Transportation Research Procedia 65, 246–255. doi:10.1016/j.trpro.2022.11.029.

Wang, Q.G., Li, X., Qin, Q., 2013. Feature selection for time series modeling. Journal of Intelligent Learning Systems and Applications 5, 152–164.

Wang, X., Wang, Z., Wan, L., Tian, Y., 2022. Prediction of flight delays at beijing capital international airport based on ensemble methods. Applied Sciences 12, 10621. doi:10.3390/app122010621.

Wu, C.L., 2008. Monitoring aircraft turnaround operations–framework development, application and implications for airline operations. Transportation Planning and Technology 31, 215–228.

Wu, C.L., 2016. Airline operations and delay management: Insights from airline economics, networks, and Strategic Schedule Planning. Routledge.

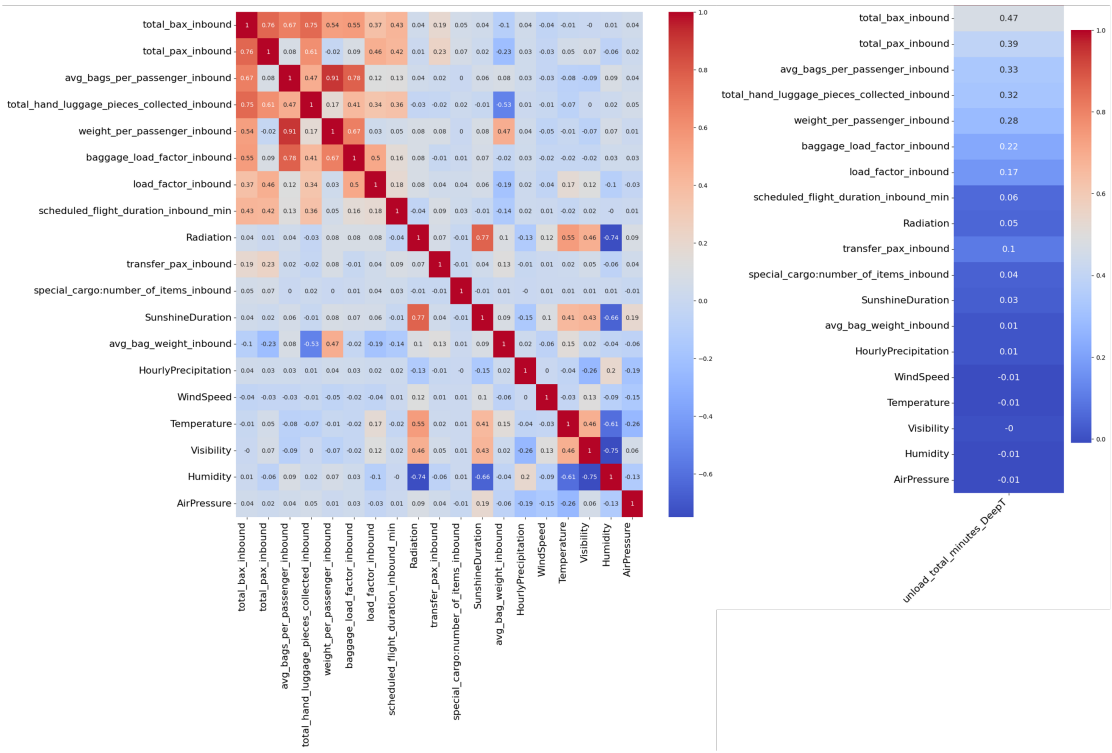
Wu, C.L., Caves, R.E., 2004. Modelling and simulation of aircraft turnaround operations at airports. Transportation Planning and Technology 27, 25–46.

Yıldız, S., Aydemir, O., Memiş, A., Varlı, S., 2022. A turnaround control system to automatically detect and monitor the time stamps of ground service actions in airports: A deep learning and computer vision based approach. Engineering Applications of Artificial Intelligence 114, 105032. doi:10.1016/j.engappai.2022.105032.

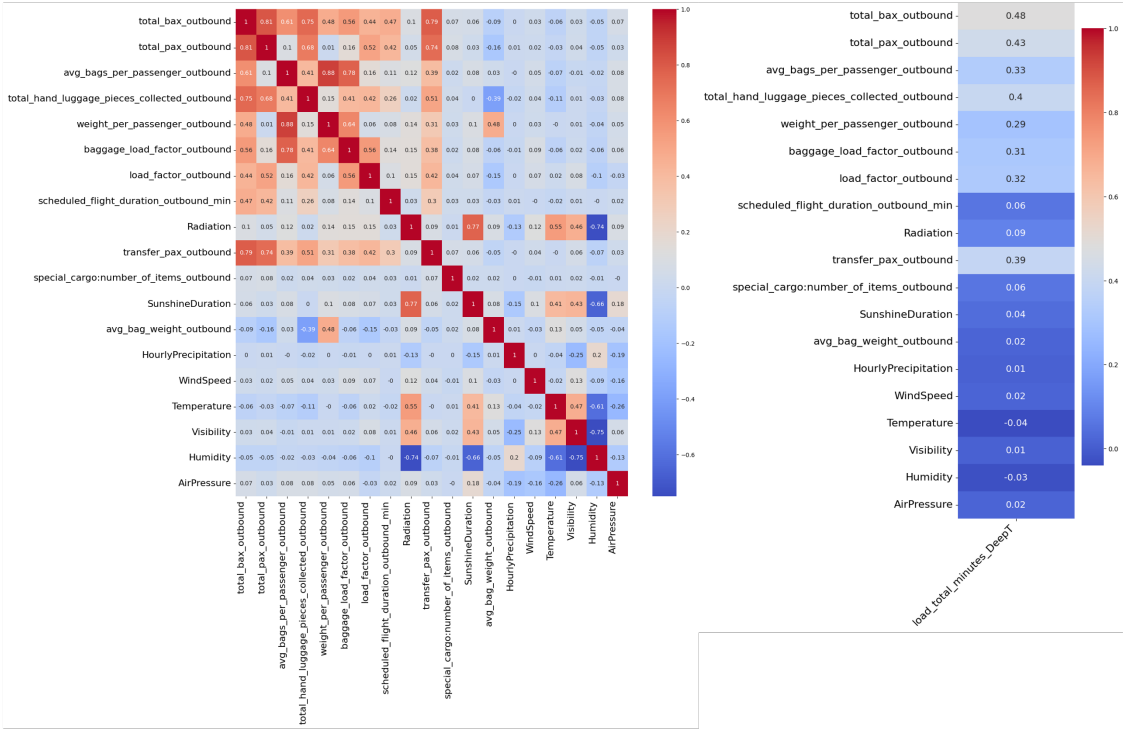
Appendix A

A.1 Multicollinearity Handling for Continuous Variables

A.1.1 Unloading-Dataset



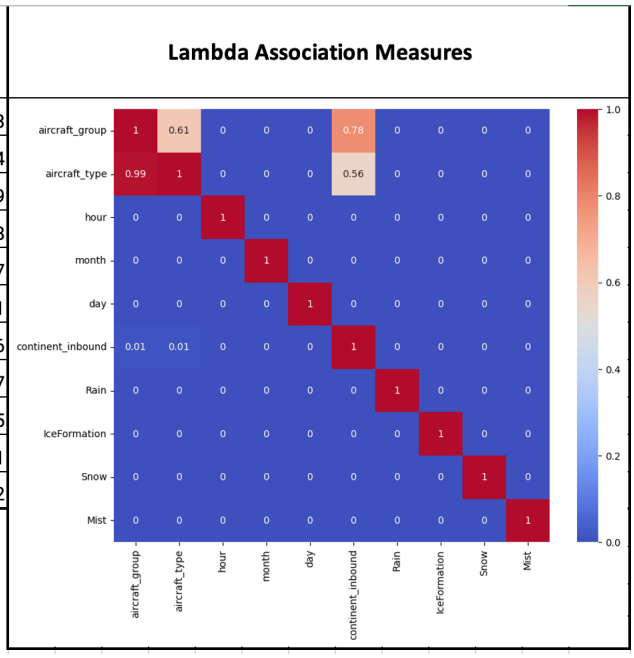
A.1.2 Loading-Dataset



A.2 Multicollinearity Handling for Categorical Variables

A.2.1 Unloading-Dataset

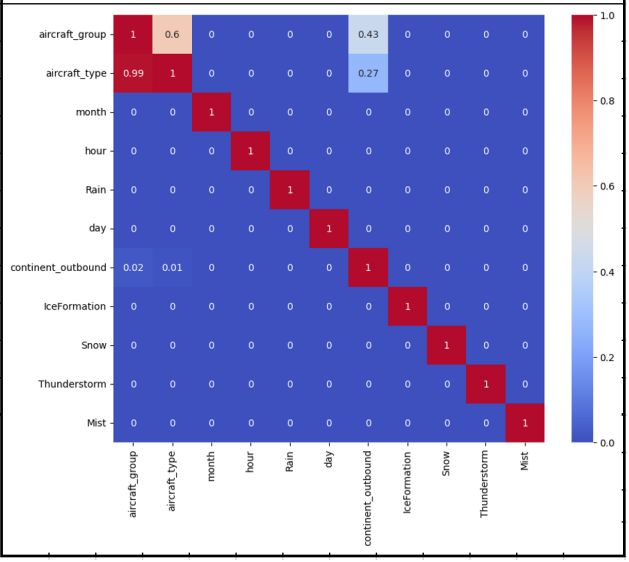
Anova and T-test statistics for association between variables and the unload_duration	
aircraft_group	483.43
aircraft_type	101.74
hour	18.89
month	9.8
day	7.27
continent_inbound	4.41
Rain	1.96
IceFormation	1.7
Snow	1.35
Mist	1
Thunderstorm	0.02



A.2.2 Loading-Dataset

Anova and T-test statistics for association between variables and the load_duration	
aircraft_group	354.3
aircraft_type	71.29
month	18.9
hour	13.8
day	4.1
Rain	2.71
continent_inbound	1.21
IceFormation	0.95
Snow	0.93
Thunderstorm	0.3
Mist	0.12

Lambda Association Measures



Appendix B

B.1 Time-related variables

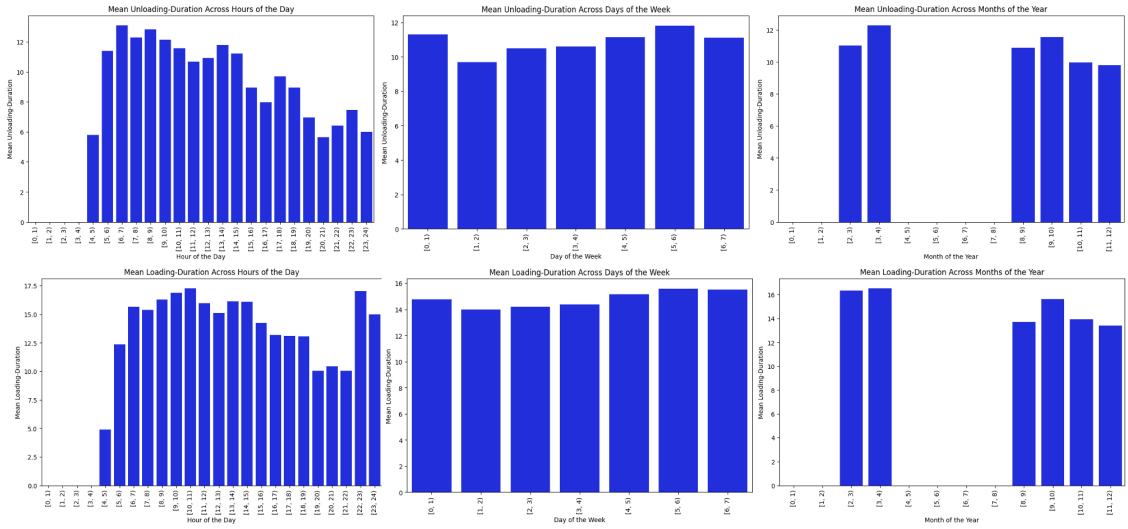


Figure B.1: Mean Durations Across Time-related Features

B.2 Weather-related variables

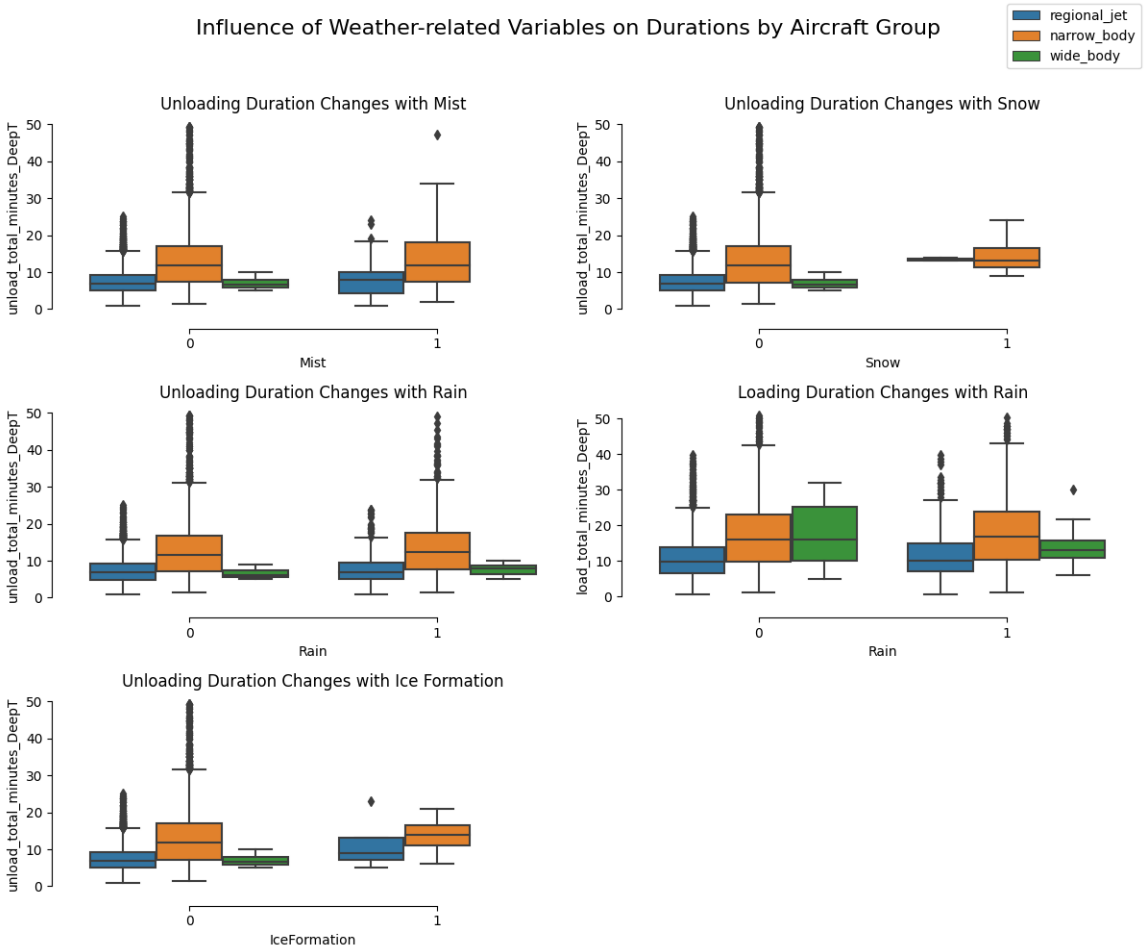


Figure B.2: Variation Durations in the Presence of Weather Features

Appendix C

C.1 Feature Importance

C.1.1 Unloading duration

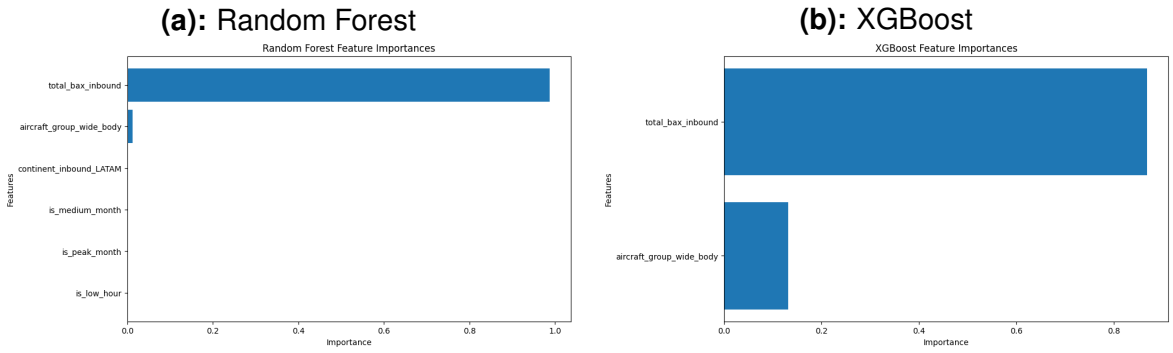


Figure C.1: Feature importance for tree-based ML models predicting the unloading duration

C.1.2 Loading duration

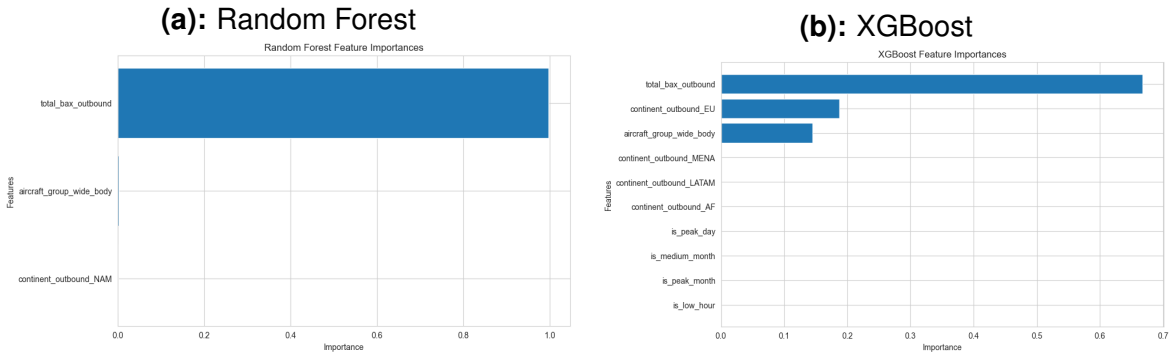


Figure C.2: Feature importance for tree-based ML models predicting the loading duration

Appendix D

D.1 Predictions and hyperparameters

D.1.1 Unloading duration

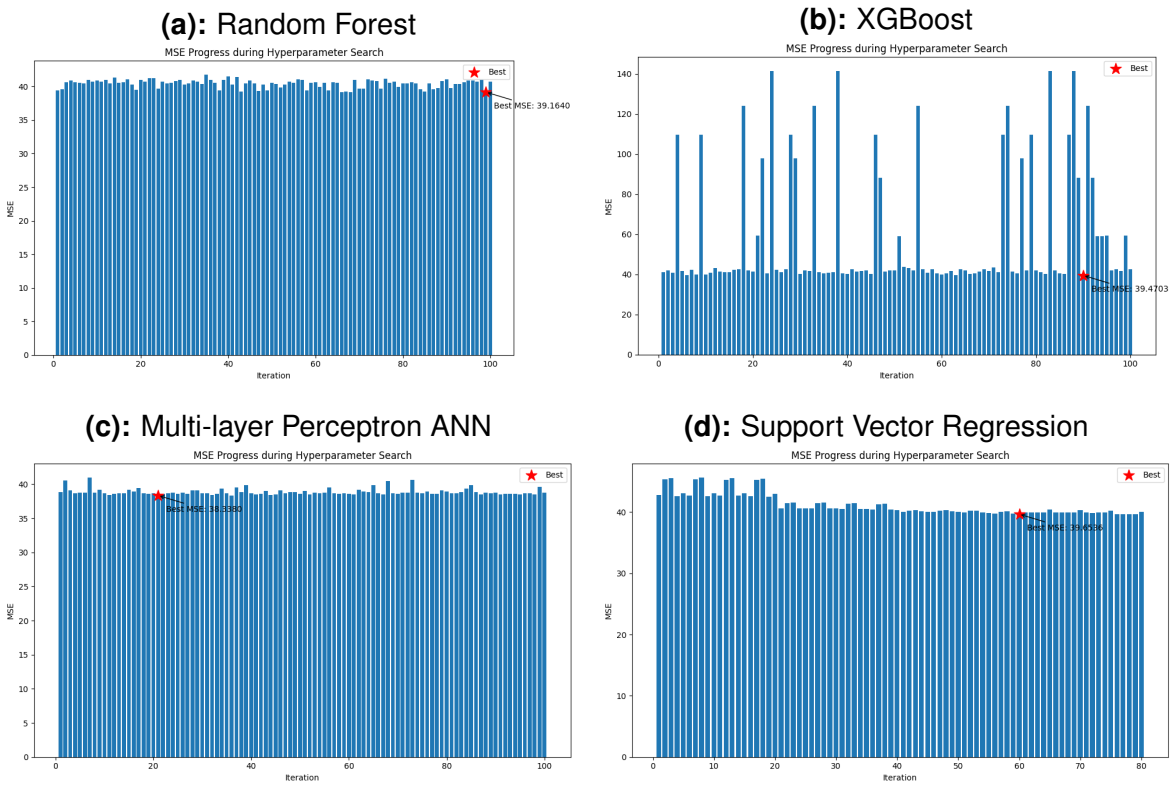


Figure D.1: Sensitivity of Hyperparameter tuning for ML models predicting the unloading duration

D.1.2 Loading duration

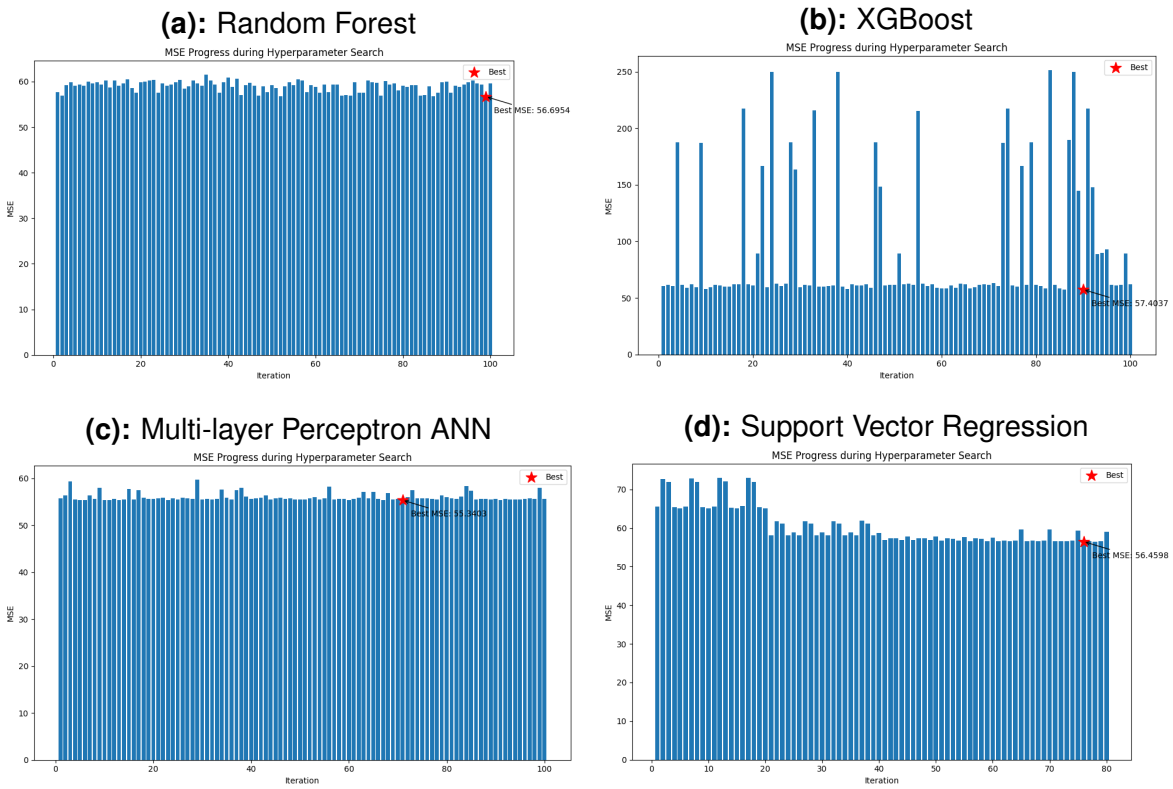


Figure D.2: Sensitivity of Hyperparameter tuning for ML models predicting the loading duration

Appendix E

E.1 Residuals for Unloading duration Predictions

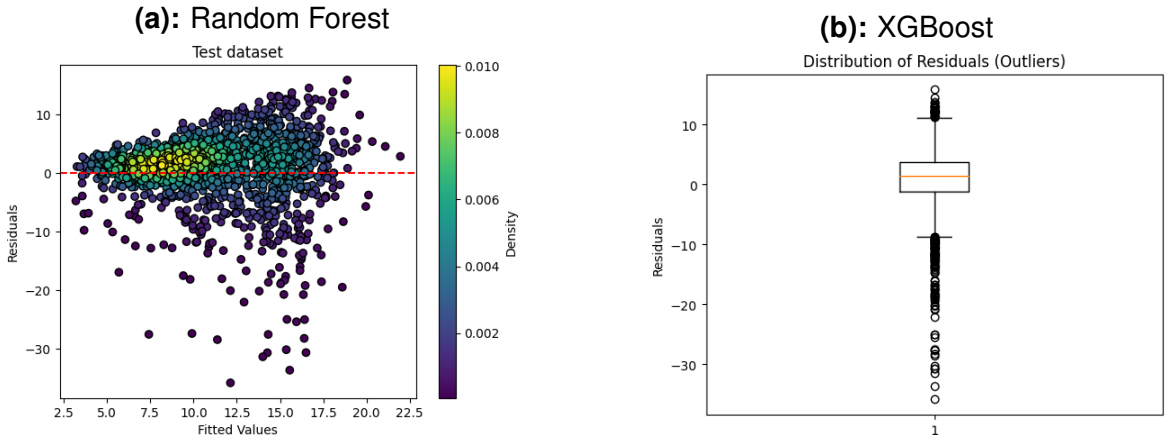


Figure E.1: Unloading residuals vs predicted values density graph & Residuals distribution box plot

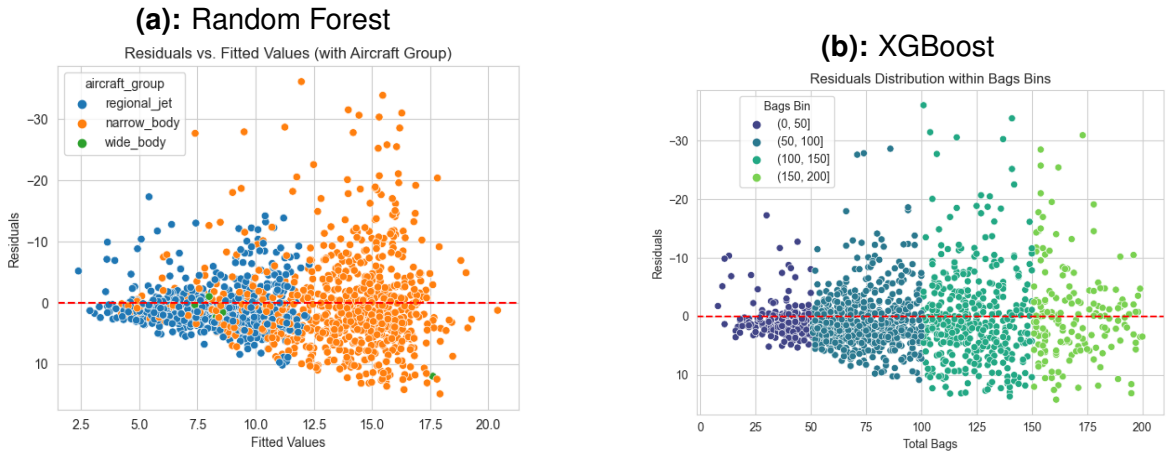


Figure E.2: Unloading residuals vs predicted values with aircraft group and number of bags classification

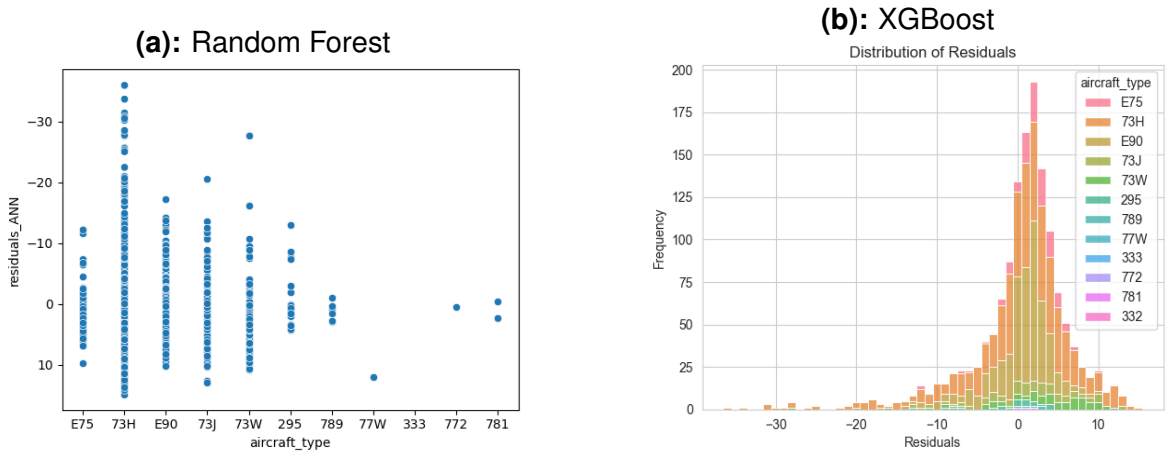


Figure E.3: Unloading residuals vs predicted values classiffied by aircraft type & residual distribution classified by aircraft types

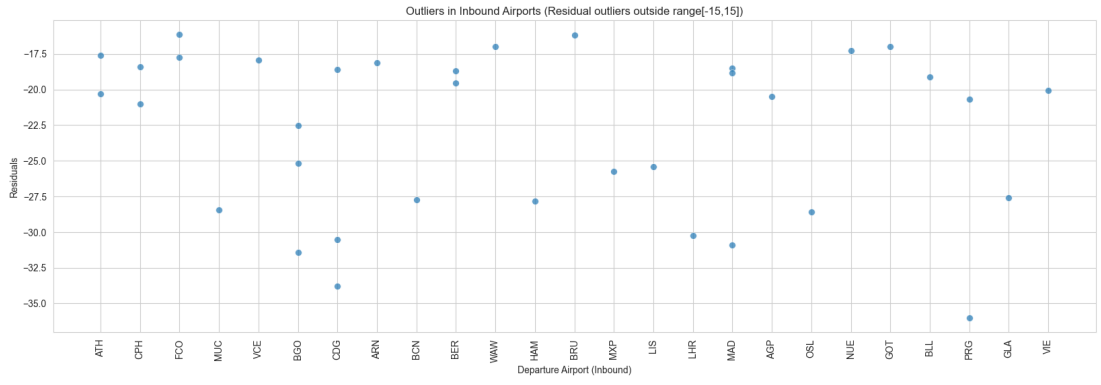


Figure E.4: Airports with the largest unloading outlier residuals

E.2 Residuals for Loading duration Predictions

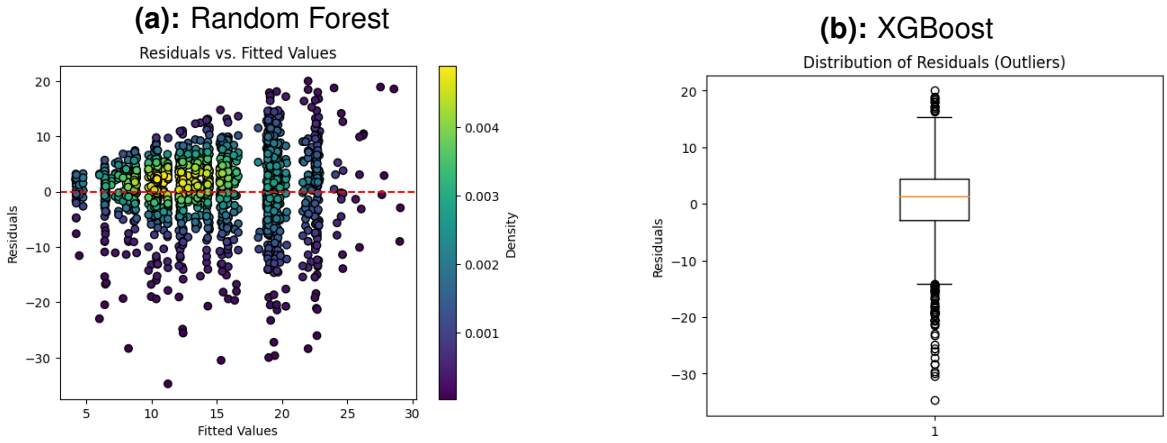


Figure E.5: Loading residuals vs predicted values density graph & Residuals distribution box plot

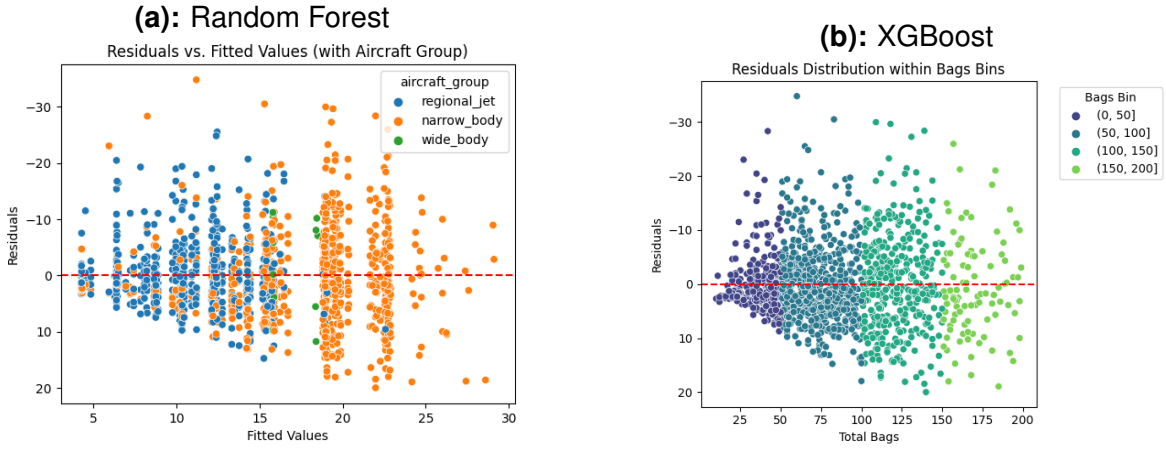


Figure E.6: Loading residuals vs predicted values with aircraft group and number of bags classification

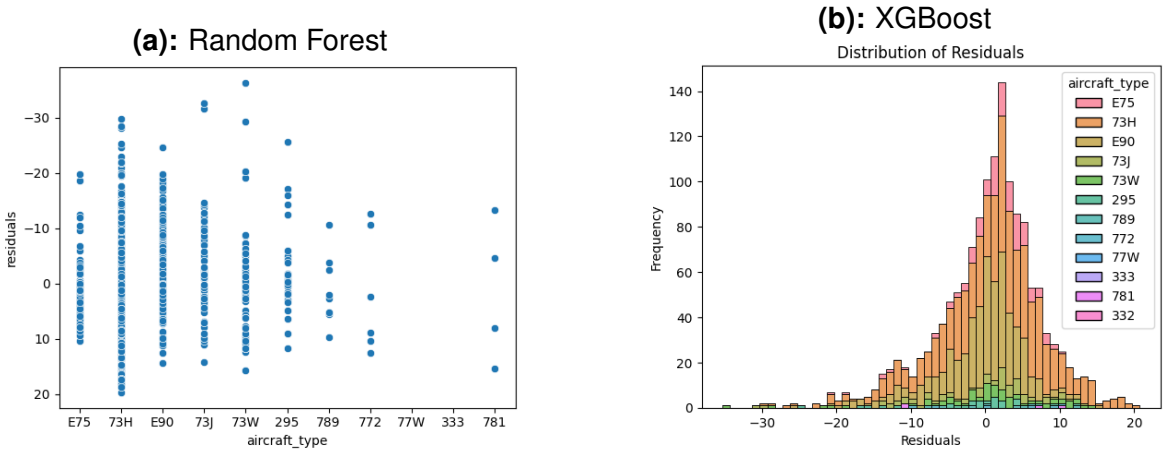


Figure E.7: Loading residuals vs predicted values classiffied by aircraft type & residual distribution classified by aircraft types

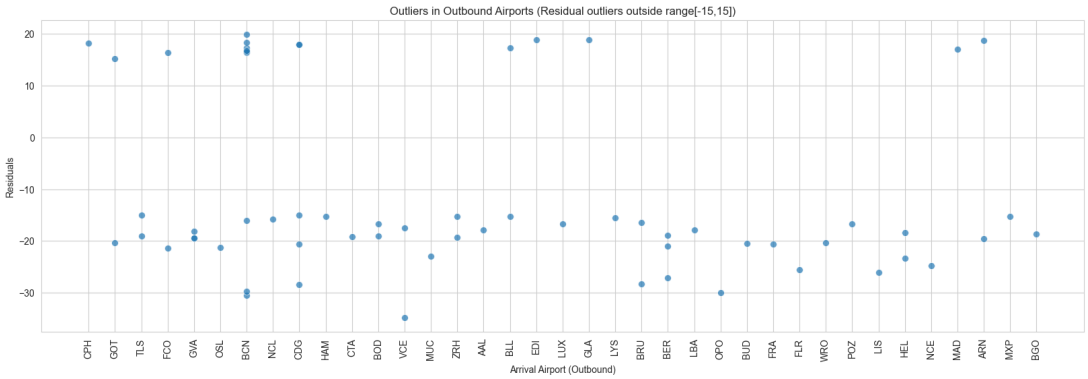


Figure E.8: Airports with the largest loading outlier residuals