UNIVERSITY OF TWENTE.

PLATFORM GROUND

KLM Data & Technology

Hand luggage overflow prediction at KLM

Carl Vico Heinrich

Student number: 2643596

First Supervisor: Dennis Prak

Second Supervisor: Martijn Mes

KLM Supervisors:

Joey van Kempen

Ali Poursaeidesfahani

Industrial Engineering and Management

2023-08-27

MANAGEMENT SUMMARY

The high demand of hand luggage from airline passengers is a significant challenge for carriers who offer both a personal item as well as a trolley as a carry on luggage for every passenger. Due to the limited capacity of the overhead bins and a demand which on a lot of flights is exceeding the capacity, delays are a common occurrence. An accurate prediction of the expected demand of hand luggage which exceeds the available capacity is therefore vital in order to offer passengers the possibility to hand in their hand luggage to reduce delays which are associated with last minute hand luggage collection. Rule-based prediction have been the default way of predicting the overflow for carriers such as AirFrance as well as KLM. Such predictions are however inaccurate, and often more hand luggage has to be collected than which has been predicted. This typically happens at the gate, where time is already tight and small disruptions can have a big impact. This causes delays due to the collection of hand luggage at the gate, as well as customers which are unhappy since they potentially have to repack their luggage right before boarding so that the trolleys can be collected.

This work has the goal to increase the accuracy of the hand luggage overflow prediction as well as to decrease the variance within the prediction. Next to that, the goal is also that the model can guide the collection process improvements. For that, the following research question was developed:

> How can a model be designed that predicts the hand luggage overflow of a given flight and guides collection process improvements?

The work is proposing to switch from a rule-based model to regression models for the prediction, as well as to add to the feature selection step which is currently happening via data analysis the step of a correlation study. To increase the accuracy and stability of the machine learning models further, it proposes to combine the best performing machine learning models together in an ensemble model.

The study found that the combination of the Extreme Gradient Boosting regressor, the Random Forest regressor and the Multi-layer Perceptron regressor combined in a voting regressor results in the best prediction of the hand luggage overflow.

As process improvements, the work recommends changing the user interface at the self-service baggage drop-off machines to increase the collection rate at them. Next to that, it is proposed to implement an intuitive cabin crew feedback system to increase awareness about how full the overhead bins are on each flight so that the prediction can be made even more accurate.

CONTENTS

Management summary i

1 Introduction 1
- 1.1 Description of KLM . . . 1
- 1.2 Research environment . . . 1
- 1.3 Background . . . 2
- 1.4 Core problem . . . 3
- 1.5 Relevance . . . 4
- 1.6 Comparison of Norm and Reality . . . 4
- 1.7 Main research question . . . 4
- 1.8 Theoretical Framework . . . 5
- 1.9 Scope . . . 5
- 1.10 Sub-research questions . . . 6
- 1.11 Deliverables . . . 7

2 Context 8
- 2.1 Importance of the hand luggage prediction model based on problem statistics . . . 8
  - 2.1.1 Overflow . . . 8
  - 2.1.2 Delay . . . 9
  - 2.1.3 Accuracy of delay codes and duration . . . 11
- 2.2 Baseline performance . . . 11
- 2.3 Available data . . . 11
  - 2.3.1 Historical . . . 11
  - 2.3.2 Live . . . 12
  - 2.3.3 Cost prediction . . . 12
- 2.4 Customers . . . 12
  - 2.4.1 Databases (Flight720 & BlueLagoon) . . . 12
  - 2.4.2 Applications (Altea & Appy2Help) . . . 12
  - 2.4.3 Visualization (Spotfire & Plug) . . . 13
  - 2.4.4 Notification (CRMPush) . . . 13
  - 2.4.5 Scheduling (Variable task time) . . . 13
- 2.5 Prediction types . . . 13

3 Literature 14
- 3.1 Hand luggage . . . 14
- 3.2 Prediction models . . . 15
  - 3.2.1 Extreme Gradient Boosting Regressor . . . 15
  - 3.2.2 Random Forest Regressor . . . 16
  - 3.2.3 Multi-layer Perceptron Regressor . . . 16
  - 3.2.4 Hyperparameters . . . 16
- 3.3 Delay cost . . . 17

4 Methodology **18**
4.1 Research Design . . . 18
4.2 Feature Engineering . . . 18
4.2.1 Feature selection . . . 18
4.2.2 Feature inspection . . . 19
4.2.3 Feature encoding . . . 19
4.2.4 Feature reduction . . . 19
4.2.5 Eliminating values . . . 20
4.2.6 Scaling . . . 20
4.3 Overflow model selection . . . 20
4.3.1 Performance measurements . . . 21
4.3.2 Model selection . . . 21
4.3.3 Ensemble modeling . . . 22
4.3.4 Implementation plan . . . 22
4.4 Cost model . . . 22
4.4.1 Machine learning cost model . . . 23
4.4.2 Historic cost model . . . 23
4.4.3 Dynamic cost model . . . 23
4.5 Process improvement plan . . . 24

5 Results **25**
5.1 Overflow model . . . 25
5.1.1 Feature analysis . . . 25
5.1.2 Model selection . . . 28
5.1.3 Ensemble model . . . 29
5.1.4 Comparison with the baseline . . . 29
5.1.5 Implementation plan . . . 30
5.2 Cost model . . . 31
5.2.1 Machine learning cost model . . . 31
5.2.2 Historic cost model . . . 33
5.2.3 Dynamic cost model . . . 33
5.3 Process improvements . . . 33
5.3.1 Self-service baggage drop-off machine . . . 34
5.3.2 Cabin crew feedback system . . . 34

6 Conclusion & Recommendations **35**
6.1 Conclusion . . . 35
6.2 Discussion . . . 35
6.2.1 Overflow model . . . 35
6.2.2 Cost models . . . 36
6.2.3 Implications . . . 37
6.2.4 Process improvements . . . 38
6.2.5 Hand luggage declaration . . . 38
6.2.6 Expected cost savings of the model deployment . . . 39
6.3 Recommendation . . . 39
6.4 Limitations & further research . . . 40
6.4.1 Limitations . . . 40
6.4.2 Assumptions . . . 41
6.4.3 Future research . . . 41

References **44**

A Appendix **46**
A.1 User interface changes 46
A.2 Hyperparameters 47

1 INTRODUCTION

Until about one and a half decades ago, airlines commonly included a free checked bag with each fare. This however changed when in 2008 American Airlines started to charge a fee for a checked bag on their lowest fare [1]. In 2013, this trend came to Europe, where KLM was the first airline to charge for checked bags on their lowest fair for flights inside Europe [2]. As a result of this change, the amount of hand luggage which passengers are bringing to the plane has increased significantly [1]. At roughly the same time, airlines reduced staff as a cost-cutting measure, making it difficult for gate staff to know how full the overhead bins were, since with less staff reporting that information was no longer possible [3]. The combination of those two factors, together with the factor that passengers not only store their trolleys in the overhead bins but also a wide variety of personal items [1], led to the situation we have ever since, where the staff at the gate has to guess how much hand luggage will fit in the overhead bin compartments.

This thesis studies the before mentioned problem at KLM and aims to develop a solution for the lack of knowledge about the amount of hand luggage which has to be checked in. This chapter provides a general introduction to KLM and the problem it faces in regards to the hand luggage overflow.

1.1 Description of KLM

Koninklijke Luchtvaart Maatschappij, better known as KLM Royal Dutch Airlines, was founded on October 7, 1919, making it the world's oldest airline still operating under its original name. It is the flag carrier airline of the Netherlands and is based at the Amsterdam Schiphol Airport. In 2004 KLM merged with Air France, forming the Air France-KLM Group. Today the Air France-KLM Group is one of the big three European network carriers next to the German Lufthansa group as well as the International Airlines Group (IAG) of British Airways and Iberia. KLM is considered a network carrier airline, since it operates under the hub and spoke model. When an airline operates under the hub and spoke model, it has all flights centered on its hub. This means that each flight is part of a trip where the first flight of the trip is leaving the hub to a destination and the second flight of the trip is returning to the hub from the destination. This model offers passengers a great variety of destinations, since due to the bundling of passengers at the hub, connections become possible which would not be viable via a direct flight. Together with other network carriers such as Delta, KLM forms the airline alliance SkyTeam. KLM is active in three core fields: passengers, cargo, as well as engineering and maintenance. KLM has several subsidiaries, such as the airlines Transavia and Martinair. KLMs European flights are primarily operated by KLMs subsidiary KLM Cityhopper as well as KLMs fleet of 737 series planes.

1.2 Research environment

This thesis has been developed in collaboration with KLMs Platform Ground, which is the IT development department for Ground Services. Platform Ground forms together with Platform

Flight, Integral Planning & Control; Safety, Security & Crisis Management as well as People, Technology & Intelligence the Data & Technology section of KLM, which is responsible for developing digital solutions for KLM Passenger Operations & HR. Within Platform Ground, the thesis was primarily supported by the Terra team, which is a Data Science team focused on developing optimization models for the loading process of the planes, as well as the customer management team. This research environment has been chosen to bridge the gap of the practice of Customer Management with the data driven approach of the data science teams within platform ground.

1.3 Background

KLM is facing the problem that passengers are bringing more hand luggage to the flight than there is capacity for the hand luggage in the cabin of the planes. This is currently primarily an issue for European flights, in particular flights on the 737 planes. KLM is currently mainly coping with that issue by asking passengers to check-in their carry-on luggage at the gate in case there is not enough capacity for all bags in the cabin. This however is suboptimal since it is a big factor causing delays.

In order to manage the amount of hand luggage, KLM has added a rule to its cheapest fare. If there is too much hand luggage for the cabin, KLM is entitled to collect the trolleys of the passengers who booked the cheapest fare. In order to know if trolleys have to be collected on a given flight, as well as how many trolleys, KLM has developed three years ago a model to predict the amount of hand luggage overflow for a given flight. That model is rule-based and uses information KLM has observed and collected in the past about their flights and the effect certain parameters have on the hand baggage overflow. The goal of the model is that staff at the gate should know how much hand luggage they need to collect and check-in during the boarding process.

KLM wants to solve the issues with last-minute hand luggage check-in by taking several actions. First, they want to implement a new prediction model that is more accurate than their current one. Next to that, KLM is aiming for a process change, which involves pushing the hand luggage check-in from the gate to the check-in desks in order to reduce the risk of delays caused by hand luggage collection. Lastly, KLM wants to investigate if the hand luggage issue requires a redesign of their cheapest non-premium light fare in order to solve the issue, as well as if some other changes have to be made in order to resolve it. To support the business case and to enable KLMs business team to get insights into the incurred cost from hand luggage delay, another tool has been developed which estimates the incurred cost.

There are primarily two aspects that limit the performance of the current rule-based model and which could lead to improvements for the improved model. The first aspect is that the current model does not allow for automatic feedback, which could be used for automatic retraining of the parameters. Implementing such automatic feedback would allow the model to learn automatically the changes in hand luggage via the feedback of the cabin crew and adjust its prediction accordingly. This would be particularly helpful in spotting changes in customer behavior in terms of the hand luggage brought. Since the amount of hand luggage passengers bring changes a lot, the current model becomes quite inaccurate since it is inflexible due to the lack of automated updates.

The second aspect which could lead to an improvement is an improved feature selection. The feature selection of the current model was done using only business logic, rather than a hybrid approach of business logic and a correlation study. Excluding the step of a correlation study leads to the inclusion of both features which have only a very small correlation with the target, as well as the inclusion of features which are highly correlated with one another. Removing features which only have a very small correlation can decrease the computation time required for the model, since fewer features have to be considered, as well as avoid model overfitting where the model becomes not very generalizable. Removing highly correlated features can make the model more stable, which helps with accuracy.

1.4 Core problem

The problem KLM is facing has many aspects to it, which can be seen in Figure 1.1. The starting problem KLM is having is, that the capacity of their overhead bin compartment is limited and demand for hand luggage is often higher than the capacity they can offer. From that, the problem arises, that KLM currently has to check-in hand luggage at the gate in order to meet all hand luggage demand. This leads to the core problem, which is that:

> The inaccurate hand luggage overflow prediction as well as the current hand luggage collection procedures can lead to delay of KLMs flights.

The core problem seen in Figure 1.1 can be solved by tackling the two action problems of having inaccurate predictions as well as bad processes. This can reduce the delays caused by last minute hand luggage check-ins and therefore save KLM money. A nice side effect of the associated process improvement could also be, that the customer satisfaction gets increased by it since the trolley check-in process will be more seamless, which is also beneficial for KLM.

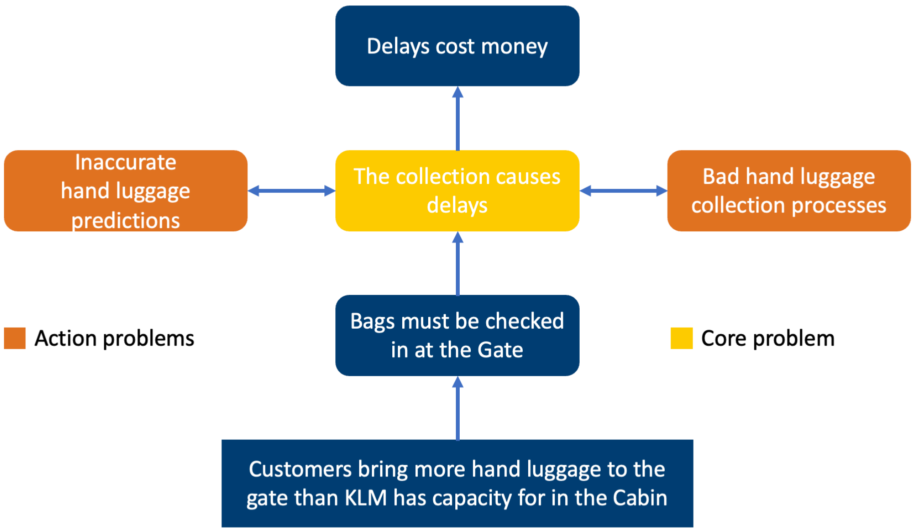

Figure 1.1: Problem Cluster

1.5 Relevance

The Core problem has both for the airline industry as well as academically a high relevance. For airlines, in particular for network carriers such as KLM, this research helps them to reduce delays, while still offering their passengers hand luggage rules which exceed the offerings of low-cost carriers. Delay caused by hand luggage collection is, with [at most 10% (exact number removed)] of the delays per season, the third-likeliest cause for delays for KLMs 737 fleet. The only causes of delays which happen more often are due to (cause of delay removed), as well as (cause of delay removed). Therefore, it is of high priority for KLM to reduce the delay factor of hand luggage. Thanks to the research, KLM has the opportunity to keep hand luggage changes of their non-premium light fare, which currently includes a personal luggage item as well as a trolley, to a minimum while minimizing costs and delays due to too much hand luggage.

Academically, this research has high relevance due to the way the prediction model is developed. Traditionally, one would try to predict the hand luggage overflow by modeling the amount of hand luggage passengers will bring. Based on that, one would then calculate the expected hand luggage overflow by comparing the hand luggage capacity of the plane with the expected amount of hand luggage brought. Such a traditional modeling approach could be developed in multiple ways, such as through a simulation study or a traditional forecasting model. In order to model the overflow with such a traditional model, both the dependent and the independent variable have to be observable. This is however not the case here, since one does not know how much hand luggage passengers are bringing on board. KLM does not know how much hand luggage passengers will bring on board, since all tickets are already including hand luggage. As a result of that, they never have to specify the amount of hand luggage they are bringing as well as the dimensions of it. This turns the amount of hand luggage brought by passengers into a latent variable. In order to still be able to make a prediction about the expected overflow. The unconventional approach of predicting the expected overflow directly is chosen.

1.6 Comparison of Norm and Reality

The current model used by KLM has a very low accuracy, and predictions of the model can be very far off. There are reports from cabin crew that on some flights more than [three-digit number (exact number removed)] pieces of hand luggage have to be collected right at the gate, and they were not prepared for that. At a recent flight it could be observed, that for that flight the Gate agents had to collect [more than double (exact number removed)] times more pieces of hand luggage than the current hand luggage tool told them to collect in order to fit all the hand luggage inside the cabin. In order to improve that, a new tool should be developed. The norm which should be achieved in order to classify the new tool as a success would be if a minimum of 80% of the predictions are within 5 trolleys of the actual value.

1.7 Main research question

Since the core of the research will be to find out how many pieces of hand luggage will have to be checked in, and the other aspects of the research are merely guiding a successful implementation of the prediction model in the current operations, the main research question will be:

> How can a model be designed that predicts the hand luggage overflow of a given flight and guides collection process improvements?

The aim of the research question is to guide the development of a prediction model which is able to recommend the number of trolleys which have to be check-in either at the Gate, or the check-in or one of the airports where a passenger for the flight is initially departing from.

1.8 Theoretical Framework

The theoretical framework of the research which can be seen in Figure 1.2 is based on three main pillars: The process improvement pillar, the overflow prediction pillar, as well as the cost model. The goal of the process improvement pillar is to optimize the collection processes, so that more hand luggage can be collected earlier than right before boarding. The second pillar is the overflow prediction model. It should replace the current inaccurate rule-based model, enabling KLM to better plan the hand luggage collection. The last pillar is the pillar of the cost model, which justifies the research solutions and supports the business case. The business case can be seen at the top of the theoretical framework. It is supported by all the pillars and is developed by KLMs business team to internally "sell" the solution. Between the process improvements as well as the overflow prediction model are future KLM changes that are enabled by the research. Those are changes which can be implemented by KLM due to the overall better performance of the hand luggage after the changes from the research have been implemented. Throughout the whole research, the focus is on the overflow prediction model as that is the primary driver for improvement within KLM, with both the cost model as well as the collection process improvements taking up supporting roles.

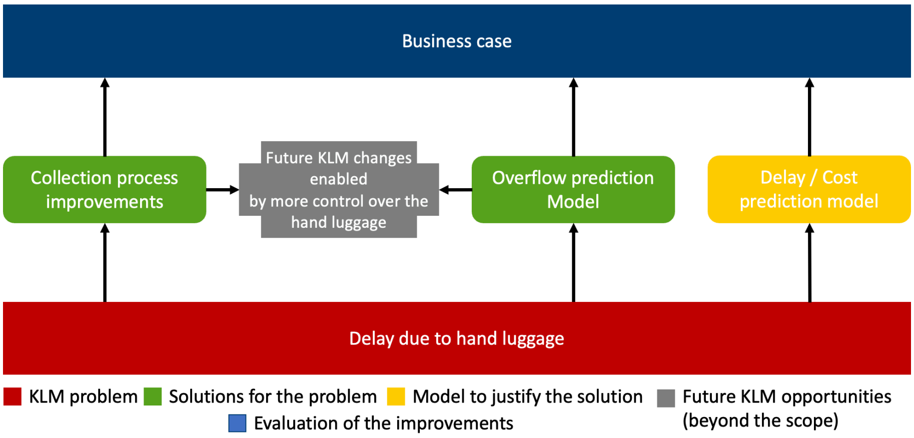

Figure 1.2: Theoretical Framework

1.9 Scope

For all research, the data from January 1, 2023, until June 1, 2023, is used. The reason for that is, that it turned out, that seasonality based features only had a very low correlation with the amount of hand luggage which was brought by passengers. The consequence of that is, that adding more historic data would increase the computation time for the model significantly, while the accuracy of the model would not improve significantly. The focus of the research is exclusively on flights which have been operated with the 737 fleet. The reason for that is, that other aircraft types most likely do require different features as well as different prediction models. The 737 was chosen, since the hand luggage problem is most severe with it.

1.10 Sub-research questions

The sub research questions are supporting the main research question. They are meant to give an overview on how KLMs current prediction tool functions, how the current collection process is structured, as well as what kind of new prediction tool would be fit for the project.

1. Analysis of the hand luggage collection process
   (a) How does the hand luggage collection currently take place? (Section 3.1)
   In order to identify how the hand luggage collection currently takes place, the current hand luggage collection at the check-in desks, the self-service drop off machines as well as at the gate have been analyzed. Additionally, conversations have been held with the stakeholders in order to learn more about the current collection procedures.
   (b) What is the effect on the process when the prediction is more accurate? (Section 4.4)
   To determine the effect of the implementation, a cost prediction model has been developed which can evaluate the effect on both the delay as well as the associated cost. Additionally, literature has been analyzed about it.
2. Analysis of the current hand luggage overflow prediction tool
   (a) What is the current method of predicting the hand luggage overflow? (Section 2.2)
   The current model has been analyzed to determine the as is state of the prediction. In order to set a baseline, the overflow prediction results from January 1, 2023 to June 1, 2023 have been used to represent the model performance.
   (b) Why does the current model not perform up to standard, what are its limitations? (Section 1.3)
   The weaknesses and limitations of the current model have been investigated, to find out the reason for its sub-par performance.
3. Analysis of the available data
   (a) What potential features are available? (Section 2.3)
   The available databases have been analyzed for potential relevant features. For those potential features, it has been analyzed if enough data is available for that feature to make sense in the prediction model. On top of that, it has been investigated, if the available data has a sufficient data quality.
   (b) What features have a correlation with the overflow? (Section 5.1.1)
   A correlation study has been done in order to determine features which have a correlation with the hand luggage overflow. For that the Pearson, Spearman, and Kendall correlation are used.
4. Model
   (a) When will the prediction be made? (Section 5.1.5)
   The prediction will be done in time intervals prior to departure. The duration of the intervals has been developed and is justified in the implementation strategy section of the report.
   (b) What model fits the need of the prediction? (Section 5.1.2)
   Based on the prediction characteristic, suitable models have been determined. Literature research has been done on the models which have been selected in order to put them into context.

(c) How will features be selected? (Section 4.2)

Features have been first selected based on their availability as well as the availability of data for that feature. In a second step, features have been selected using a Pearson, Spearman and Kendall correlation study, where the correlation of the features to the overflow is tested. It has then been investigated if features have a high correlation to other features which are in the selection. The feature with the highest correlation to the target have then been kept, while the other features were disregarded.

(d) What performance improvements can that model offer over the baseline? (Section 5.1.3)

The current model as well as the improved model have been tested on the same past data in order to determine the improvement. The performance will then be judged using the selected performance metrics.

1.11 Deliverables

As part of the research, three deliverables are created to support the main research question as well as the core problem KLM is facing. The deliverables are as follows:

1. Overflow prediction
   (a) Model
   A prediction model which is predicting the hand luggage overflow for a given flight.
   (b) Implementation plan
   A plan on how to implement the prediction model at KLM so that it performs to its fullest potential.
2. Cost model
   A model which determines the cost associated with hand luggage delays to allow for better business decisions associated with the hand luggage overflow.
3. Process improvement recommendations
   Recommendations about which processes should be improved and how they should be improved to allow for a better hand luggage collection.

2 CONTEXT

In this chapter, the context of the paper will be explained to give insights into the environment in which the research is taking place. The chapter will start by giving an overview about the importance of the hand luggage prediction model. After that, the performance of the current model will be covered and the available data is discussed. Lastly, the customers of the prediction model are introduced and the different types of prediction are explained.

2.1 Importance of the hand luggage prediction model based on problem statistics

There are two statistics which can be used to measure the model performance, which will be discussed in this section. Those are the resulting overflow caused by the prediction as well as the delay incurred due to the collection of hand luggage at the gate. For the latter one, the aspects of frequency as well as duration of the delay are covered.

2.1.1 Overflow

Even though KLM already has a prediction model in place, KLM is still often facing the problem that more hand luggage has to be collected at the gate than was predicted. The extent to which that is the case can be seen in the histogram of Figure 2.1. The x-axis of this histogram shows the difference between the amount of hand luggage which still has to be collected at the gate according to the prediction model, and the actual amount of hand luggage collected at the gate. The amount of hand luggage which still has to be collected at the gate is determined by subtracting the amount of hand luggage already collected before the gate from the total amount of hand luggage which should be collected according to the prediction. It can clearly be seen, that it is a common occurrence that up to [more than 10 (exact number removed)] extra pieces of hand luggage have to be collected in addition to the pieces of hand luggage which the prediction model already predicted which could not be collected earlier in the process.

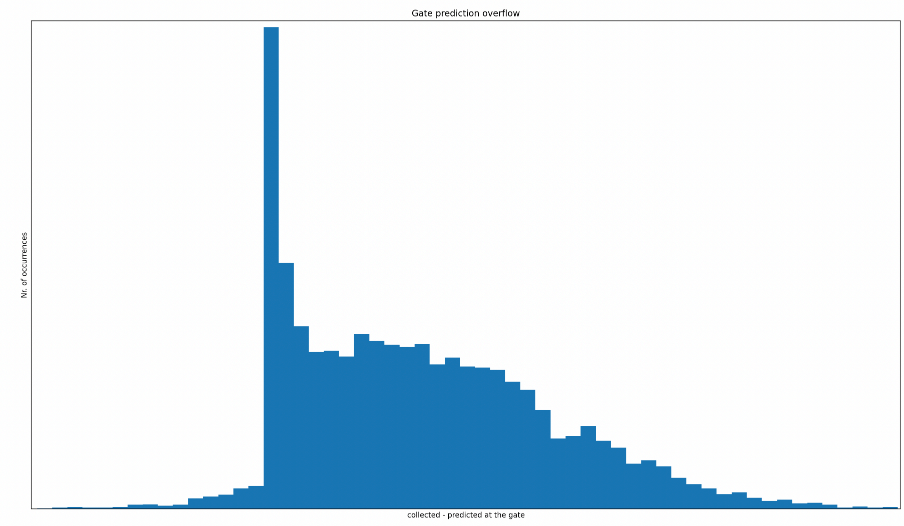

Figure 2.1: Hand luggage overflow at the Gate

2.1.2 Delay

The hand luggage which has to be collected because not enough hand luggage has been collected as well as hand luggage which was included in the prediction but could not be collected earlier in the passenger journey are a big cause for delays. From a cost perspective, two characteristics of delays are of importance. The frequency in which the delays occur (meaning what proportion of flights have delays) as well as how long the delays due to hand luggage overflow are.

Frequency

Delay caused by too much hand luggage are a common occurrence for KLM. For KLMs 737 fleet, the delays account for [at most 10% (exact number removed)] of the overall delays, which makes them the third-likeliest cause for delays for KLM on that type of aircraft, as can be seen in Figure 2.2. The frequency of hand luggage related delays does fluctuate depending on the route of the flight, with some routes having up to [more than 10% (exact number removed)] of the overall flights delayed, as can be seen in Figure 2.3. When comparing the routes with the highest proportion of delays related to hand luggage, it can be seen, that the flights departing from Amsterdam (orange chart in Figure 2.3) have a higher proportion of flights within the routes delayed compared to flights departing from outstations (blue chart in Figure 2.3). This is suggesting that while the problem is severe for both departures in Amsterdam as well as outstations, it is more significant for flights departing in Amsterdam. Aggregated onto the network, about [at most 10% (exact number removed)] of KLMs European flights are delayed due to hand luggage from January 1, 2023, to June 1, 2023.

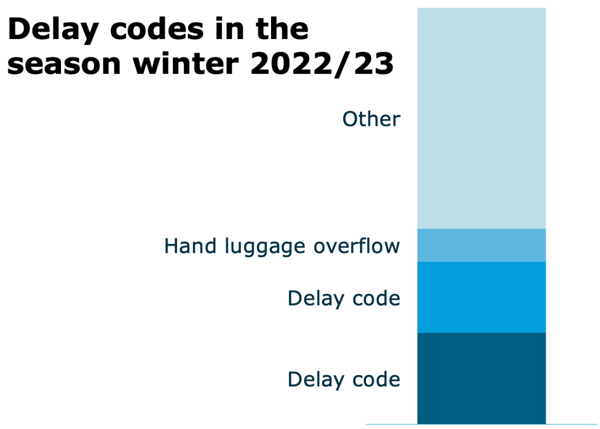

Figure 2.2: Delay code 10 proportion

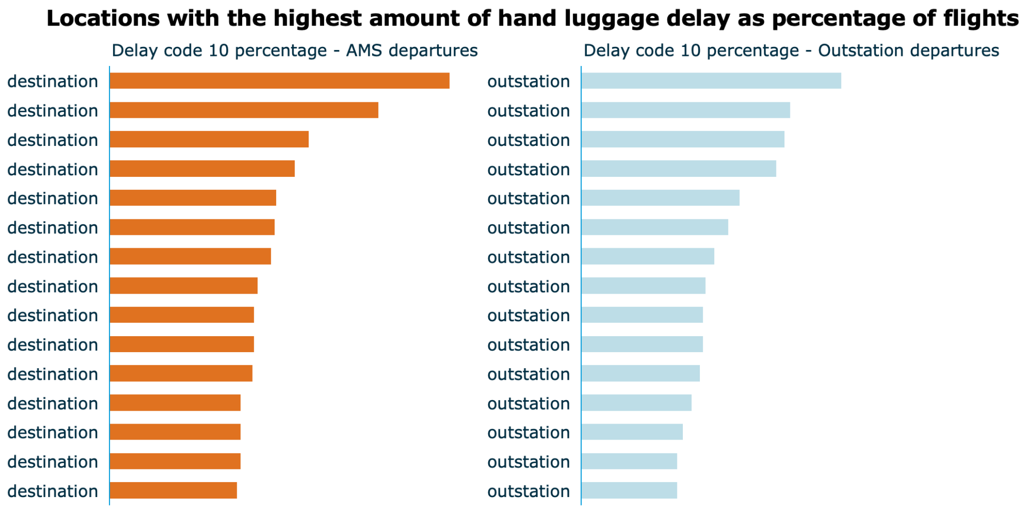

Figure 2.3: Proportion of flights on a route with delay due to hand luggage

Duration

When taking a look at the season 2022/2023 in figure 2.4, it can be seen, that delays caused by hand luggage have an average duration of [at most 10 (exact number removed)] minutes. This makes delays caused by hand luggage the delays with the fourth-longest duration for the 737 fleet. Since delays are very expensive, it is vital for KLM to reduce delays which have such a long duration.

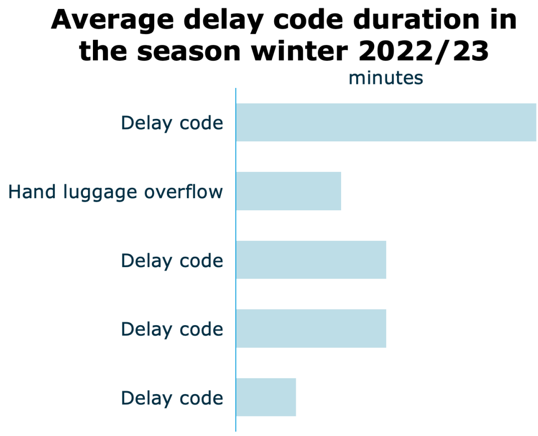

Figure 2.4: Average delay code duration

2.1.3 Accuracy of delay codes and duration

The delay minutes caused by the hand luggage collection are tracked manually. Each Flight can have up to two delay codes, with a corresponding delay duration per delay code. Tracking the delay codes as well as their duration manually does introduce some inaccuracies, since it is hard to grasp what exactly now caused a delay. There is however no mechanism which could track the delay more accurately, so it is the most reliable source of delay which is available. Additionally, airlines do have contracts with other partners which do include delay metrics as part of their contract. There is a possibility that in some cases, delays are assigned to a wrong delay code in order to meet the agreed upon performance targets, making delay codes even less reliable.

2.2 Baseline performance

Next to the practical results of the current model, it also makes sense to take a look at the performance of the current model from as statistics point of view to judge the model performance. The current rule-based model has a R^2 value of -2.2 which means that averaging the total hand luggage pieces collected (R^2 value of 0) leads to a more accurate prediction than the current rule-based model. The mean absolute error of the current model is 10.3, which means that the prediction is on average 10 hand luggage pieces of, and the root-mean-square error 24.6. In comparison, taking the average has a mean absolute error of 10.8 and a root-mean-square error of 13.67. This is suggesting, that while the absolute deviation is for the current tool compared to the average, it has more significant outliers than taking the average which is a problem. The metrics of the current tool show, that the current model of KLM is not very accurate in its predictions and that it should be replaced or improved.

2.3 Available data

There are three data sources depending on the data which is needed. Those are databases for both historical as well as live data, as well as an API for cost data.

2.3.1 Historical

The data used to analyze the current situation and model as well as the data which is used to develop the new prediction model is out of KLMs data lake “BlueLagoon”. Out of this data lake, the relational Microsoft SQL databases FlightLegs is used. FlightLegs is a database which is used to store historical flight information. Since most of the data in those tables is automatically collected, a high data quality is ensured. Only the before mentioned delay codes have to be considered with caution for the reasons mentioned above.

2.3.2 Live

Once the model is tried and tested, KLM has the plan to switch the model over to the Apache HBase NoSQL database called “Flight 720”. This is done, since with “Flight 720”, the model can receive data live and can therefore update the predictions in intervals. This is not possible on “BlueLagoon”, since it only receives daily replays. The content in “Flight 720” and “BlueLagoon” is about the same, with the only difference that each metric in “Flight 720” has a time stamp, while in “BlueLagoon” the data is aggregated (for example in number of hand luggage pieces collected before boarding).

2.3.3 Cost prediction

In order to predict the cost which is associated with delays, KLMs Forecost API is used. With it, KLM is predicting the cost they will be incurring for delays. With this API for each flight, the delay can be predicted for several delay durations. The prediction horizon of Forecost is two days into the future, and no information about cost prediction is stored after the flight has taken place.

2.4 Customers

The hand luggage prediction gets used by several customers within KLM. There are five different kinds of customers which use the data generated by the hand luggage prediction. Customers are Databases such as Flight720 & BlueLagoon, Applications such as Altea & Appy2Help, Visualization tools such as Plug & Spotfire, the SMS notification service CRMPush as well as the scheduling application Variable task time (VTT).

2.4.1 Databases (Flight720 & BlueLagoon)

Both the HBase database Flight720 as well as the SQL database BlueLagoon are customers of the hand luggage prediction. They are storing the prediction so that other applications can make use of the prediction. Flight720 additionally takes care of doing the calculation of how many pieces of hand luggage are left to collect for a given flight. This logic is subtracting the already collected hand luggage pieces from the pieces of hand luggage which the prediction tool recommends collecting. This calculation is done globally, so that all applications can make use of it.

2.4.2 Applications (Altea & Appy2Help)

The Applications Amadeus Altea & Appy2Help are use by the customer service agents when interacting with the passengers. Altea is the application which is running on the computers of the agents, while Appy2Help is the application which the service agents have on their iPads. Both applications show a ribbon at the top of the screen informing the agents about the remaining hand luggage pieces to collect for a given flight.

2.4.3 Visualization (Spotfire & Plug)

TIBCO Spotfire as well as Plug are Analytics applications which can be used to analyze performance metrics. For both of them, the hand luggage prediction is a metric which is being analyzed.

2.4.4 Notification (CRMPush)

On flights where the prediction tool is expecting a lot of hand luggage which has to be collected, KLM is sending out an SMS message to customers informing them about the expected large amount of hand luggage on board. This message is sent via CRMPush and informs the customers of the Non-Premium-Light-Fare about the option to check-in their hand luggage for free.

2.4.5 Scheduling (Variable task time)

Since loading hand luggage which has been collected at the gate has an effect on the time which is needed to load the luggage onto the airplane, the hand luggage prediction is taken into account in the planning of the time which is needed to load the luggage onto the airplane. This variable time is calculated with the tool “Variable Task time” (VTT) and then used in the scheduling optimizer Terra to develop a schedule for the baggage loading staff.

2.5 Prediction types

Prediction in the context of regression can generally end up being an overprediction as well as an underprediction. Underpredictions in our context are predictions where the actual amount of hand luggage for a flight exceeds the predicted number of pieces of hand luggage. This is highly undesirable, since this most likely means that KLM has to collect additional pieces of hand luggage at the gate in order to be able to depart. Since the collection of hand luggage costs time, this can easily cause delays which can turn out to be very expensive.

Overpredictions are predictions where the model has predicted, that KLM must collect more hand luggage prices than they actually had to collect. This is undesirable, since that means, that more passengers were asked to hand in their hand luggage than were actually required. When passengers are asked to hand in their hand luggage due to full overhead bin compartments and the compartments actually do not happen to be full, passengers are not happy since most of them prefer to keep their hand luggage to themselves rather than to check the luggage in.

For KLM, it is optimal when the model has as little underpredictions as possible, since underpredictions are directly effecting the delays and therefore has an effect on KLMs punctuality as well as cost. For overpredictions, it is important for KLM to keep the amount of hand luggage pieces which are over predicted quite small, since it leaves a bad impression on the customer, however a slight overprediction actually helps with the boarding since when the bins are not 100% filled passengers don’t have to search for a spot for their hand luggage.

3 LITERATURE

The literature chapter aims to cover three core aspects. First, literature related to hand luggage will be analyzed with a focus on how it was covered both from the industrial design perspective as well as from the prediction perspective. After that, relevant prediction models for the overflow prediction will be introduced and explained. As part of determining the optimal prediction model for the overflow prediction, a variety of prediction models have been tested. This literature section is only covering the model, which also ended up being used in the prediction model. Those are the “Extreme Gradient Boosting Regressor”, the “Random Forest Regressor”, and the “Multi-layer Perceptron Regressor”. To conclude the literature section, the hand luggage issue will be put into context in a final section, in which cost associated with delay for airline operations will be covered.

3.1 Hand luggage

During the boarding process, interferences are often a reason for delays as well as an increase in turnaround time. There are several reasons for such interferences, however the three main bottlenecks are too little storage space for hand luggage or not effectively used storage space for it, unprepared passengers as well as unclear and unfocused audio announcements [4].

A very important aspect about the hand luggage collection is to take a look at the baggage journey to identify where hand luggage collection is taking place. The baggage journey of KLM has been visualized in 2015 by Van der Broek [5] and can be seen in Figure 3.1. In the industrial design study by Van der Broek [5], he identified several ways on how to prevent the hand luggage overflow. In particular, he proposes to decrease the number of hand luggage pieces per passenger, to increase voluntary as well as involuntary hand luggage check-in. Next to that, he proposes to increase the hand luggage capacity of the overhead bins as well as to increase the storage efficiency inside the bins. However, he concludes, that the only aspect which actually has an impact on the reduction of the hand luggage overflow is the prevention of gate checks all together via a more accurate prediction.

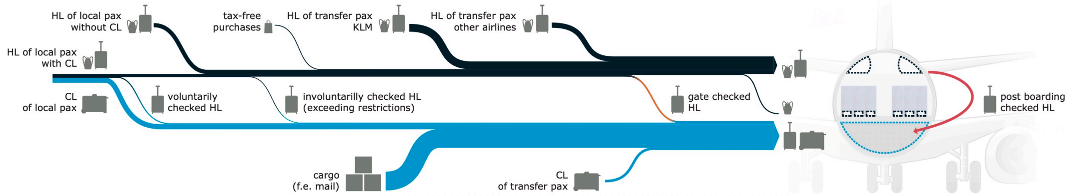

Figure 3.1: KLM baggage journey by Van der Broek[5]

In order to effectively prevent gate checks all together, it is important to raise passenger awareness about the possibility of checking in their hand luggage for free. Ham [6] has investigated in his research the effectiveness of showing passengers during the check-in flow a message outlining the benefits of checking in their hand luggage, however he was not able to show clear positive, nor negative effects.

Predictions in the context of hand luggage has been covered primarily as part of the boarding process, where prediction models were researched which predict the boarding times. For those models, hand luggage was primarily seen as a parameter [7].

Erland et al. [8] mention that it is best to have people board according to decreasing luggage handling time. Meaning that passenger who have a lot of hand luggage should therefore board before passengers without hand luggage. This is supported by Qiang et al. [9] who state that the passengers with a lot of hand luggage should board first in order to overall reduce boarding time. Bachmat [10] recommends to place passengers into different queues based on the expected boarding time as well as other characteristics. Also in this source, the amount of hand luggage is seen as an aspect by which passengers can be grouped.

There currently does exist a research gap between the field of industrial design as well as boarding time predictions, as there is no research which predicts the expected amount of hand luggage or the expected hand luggage overflow based on the flight parameters.

3.2 Prediction models

In this report, three different prediction models will be applied. Those are the Extreme Gradient Boosting Regressor, the Random Forest Regressor and the Multi-layer Perceptron Regressor. Below, an introduction into the types of models will be given.

3.2.1 Extreme Gradient Boosting Regressor

Extreme Gradient Boosting also known as XGBoost is a tree based regression model which was proposed by Chen and Guestrin [11] as part of the gradient boosting framework. Compared to a default gradient boosting model, XGBoost has an additional regularization term to avoid overfitting [11].

In order to set up XGBoost, first an initial guess of the independent variable is made. After the initial guess has been made, it is determined how good it has been in predicting the independent variable by calculating the residuals between the initial guess and the actual values. In order to now build the prediction model, the residuals from the initial guess are used to form an initial leaf, which is used as a starting point for the regression tree. To build the regression tree, the initial leaf is turned into a root for new leafs, which are determined based on the quantiles of the underlying data using either the exact or the approximate greedy algorithm [11]. Each quantile is used as a threshold to form a leaf which is below the threshold and one which is above the threshold, where each of the leafs is holding the corresponding residuals. Next, the regularized objective for the leafs as well as the root are calculated with the formula $loss\,value = \frac{(Sum\,of\,residuals)^2}{\#\,of\,residuals + regularisation\,term}$. The regularization term of this formula is used to regulate the sensitivity of the prediction to individual observations [11]. It can be defined as a hyperparameter for the model, where an increase in it makes it easier to prune the leafs of the model at a later stage to mitigate overfitting. When the residuals in a leaf are very different, they cancel each other out and the loss value is very low, while when they are very similar or there is only one residual in a leaf, then the loss value is very large. To find out how much better the new leafs are performing to the root, the gain is calculated by adding together the loss value of the leafs and subtracting the loss value of the root from it. In order to find out which threshold to use for the tree, the gains of the different thresholds are compared and the threshold with the highest gain is picked. Based on the resulting leafs, the process is now repeated to form new thresholds and determine the optimal one using the same logic [11]. This is repeated until either the minimum number of residuals in a leaf is reached or the maximum tree depth is reached. In order to avoid overfitting of the tree, pruning is used. In pruning, it is investigated if the gain value of a leave minus the hyperparameter gamma is a positive or a negative number. If the resulting number is positive, the leaf can stay and pruning will not continue to move down the tree from that branch, else the leaf will be removed, and pruning will be applied one branch further down the tree [12]. Once pruning has been finished, the output values of the leafs are calculated. In order to now make a prediction, the initial guessed value is taken and the output values of the XGBoost tree are added to it. The output of the tree is scaled down by the learning rate to limit the influence of the individual trees so that no overfitting occurs, and to leaves space for future trees to improve the model [11]. Based on this new prediction, the new residuals are determined and the process of building another tree is initiated using the residuals of the prior tree as starting point [13]. This is repeated until the maximum number of trees is reached or until the residuals are very small. At the end, all trees are scaled down by the learning rate and summed together with the initial guessed value to form the final prediction [11].

3.2.2 Random Forest Regressor

Random forest is an ensemble learning method based on decision trees. A significant drawback of using a decision tree model is, that it is prone to overfit. This problem is mitigated by using the random forest model by combining multiple decision trees together. The first step in making a random forest model is to extract multiple samples from the original sample using bootstrap resampling. For each of the samples, a decision tree is then modeled, where for each branch a random subset of features is tested and the best performing one is used as the branch [14]. This is repeated, until either the mean-squared-error of the tree is optimal or if the tree has run out of features [13]. In order to determine the prediction, the trees of the different samples are all asked to do a prediction, which then are averaged to determine the prediction of the random forest model [15].

3.2.3 Multi-layer Perceptron Regressor

The Multi-layer Perceptron regressor is a feed forward neural network, meaning that the output of each neuron is not affecting the neuron itself [16]. It is the most popular type of neural network and is defined by its unidirectional flow from input to output layer [17][18]. The Multi-layer Perceptron neural network has as many artificial input units as it has input variables and a fixed number of artificial units for both the hidden layers as well as the output layer [19][20]. The decisions of the Multi-layer perceptron regressor are made based on the activation function of the perceptrons, as well as on the weights connecting the perceptrons to the perceptrons of the prior layer. This is done by forming a weighted sum of the inputs and adding a constant/bias to it [16]. The weights and the bias initially get set when the model is set up and then optimized using back propagation [16]. Multi-layer Perceptron models are generally considered highly accurate, however they have a high possibility of finding a local minimum, which has as the consequence that no global optimum is found [17].

3.2.4 Hyperparameters

All the above-mentioned models have different hyperparameters which can be tuned in order to adjust the model to the project. As part of this study, the models were always applied with their default hyperparameters, since it makes more sense to do the hyperparameter optimization on KLMs servers rather than on a local machine. The most important hyperparameters which are effecting the prediction performance of the models can be found for the Extreme Gradient Boosting Regressor in Figure A.3, for the Random Forest Regressor in Figure A.4 and for the Multi-layer Perceptron Regressor in A.5 along with a short description of the hyperparameter, possible values for the hyperparameter as well as the default value of it, which is used in this study.

3.3 Delay cost

Due to the collection of the hand luggage at the gate, delays are not uncommon. According to Eurocontrol [21] who has done a combined study with the University of Westminster [22], it makes sense to distinguish delays into tactical delay cost as well as strategical delay costs. Strategic delay costs are described as the delay costs which are incurred in advance due to actions which are needed, such as adding buffers in the schedule or changing the route. Tactical delay costs on the other hand are incurred on the day of operations and include costs such as delay which gets generated by a prior delay. Within the tactical delay costs, the delay is differentiated, between delay incurred in different phases of the flight. The phases described can be seen in Figure 3.2 and are described as “at the gate, stable”, “at the gate, turnaround”, “Taxi”, “Takeoff/Landing”, “En route” and “arrival management”. Based on those delay classifications, the delay cost is then approximated, which can be seen in Figure 3.3.

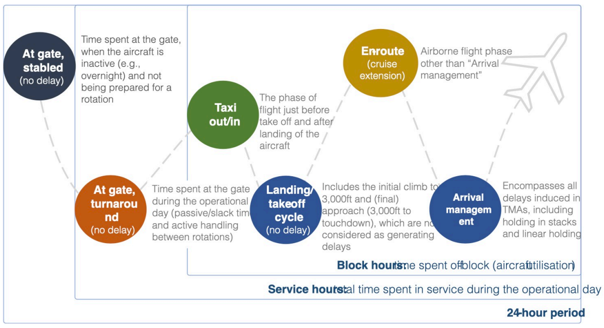

Figure 3.2: Stages of the flight according to Eurocontrol

Table 16.3:
Total tactical delay costs with network effect - base scenario

Aircraf type	At gate			Taxiing			En-route			Arrival management		
	5'	15'	30'	5'	15'	30'	5'	15'	30'	5'	15'	30'
A319	€83	€523	€1,901	€154	€724	€2,305	€286	€1,128	€3,112	€262	€1,057	€2,982
A320	€95	€594	€2,162	€178	€844	€2,672	€297	€1,199	€3,386	€297	€1,187	€3,350
A321	€119	€689	€2,566	€201	€927	€3,041	€357	€1,414	€4,015	€333	€1,343	€3,873
A332	€213	€1,176	€4,218	€404	€1,722	€5,310	€677	€2,542	€6,961	€558	€2,185	€6,237
AT43	€36	€213	€724	€71	€309	€915	€83	€345	€986	€83	€345	€986
AT72	€47	€286	€974	€83	€392	€1,199	€107	€463	€1,343	€107	€440	€1,295
B733	€83	€511	€1,842	€166	€749	€2,317	€297	€1,140	€3,112	€250	€1,010	€2,851
B734	€95	€570	€2,067	€178	€820	€2,577	€309	€1,199	€3,326	€297	€1,164	€3,243
B735	€83	€463	€1,663	€166	€712	€2,150	€274	€1,045	€2,816	€213	€879	€2,483
B738	€107	€641	€2,305	€178	€856	€2,745	€321	€1,283	€3,599	€297	€1,211	€3,457
B744	€286	€1,627	€5,939	€546	€2,400	€7,484	€1,104	€4,086	€10,85	€844	€3,279	€9,242
B752	€119	€736	€2,720	€237	€1,093	€3,445	€404	€1,592	€4,431	€345	€1,402	€4,062
B763	€201	€1,069	€3,802	€345	€1,497	€4,656	€606	€2,281	€6,225	€570	€2,173	€6,022
DH8D	€47	€297	€1,057	€83	€404	€1,272	€130	€534	€1,520	€130	€534	€1,520
E190	€71	€380	€1,366	€130	€558	€1,722	€213	€832	€2,269	€213	€820	€2,234

Source: *University of Westminster (2015), European airline delay cost reference values - version 4.1*

Figure 3.3: Delay cost according to the University of Westminster and Eurocontrol

4 METHODOLOGY

The methodology chapter consists out of four parts. First, the research design and the engineering of the features is covered. After that, the selection of both the overflow model as well as the cost models is explained. Finally, the approach for the process improvement plan is covered.

4.1 Research Design

The research will consist of several aspects which cover the different needs of KLM. At its core, the research will consist out of a prediction model which aims to better predict the hand luggage overflow for a given flight. This helps in better knowing what to expect in terms of hand luggage demand, so that the appropriate collection as well as staffing decisions can be made. In order to determine good features for the model, a correlation study will be conducted. In addition to the model, the current hand luggage collection process will be analyzed, and an improved collection process will be recommended.

4.2 Feature Engineering

Feature Engineering is the process of deciding on the features for the prediction model, as well as to prepare the features for the prediction model. As part of this section, six parts of the feature engineering will be covered. First, the steps of the feature selection and inspection will be covered. After that, feature encoding, feature reduction as well a feature elimination will be discussed. Finally, the scaling methodology is introduced.

4.2.1 Feature selection

The first step of feature engineering is the step of selecting the features which should be considered for the prediction model. Features which can be used could be on a flight level, such as how many passengers are on board, as well as on a passenger level, such as what is the age of the passengers on board. For this model, the decision was made to only focus on features on the flight level since that allows us to not be required to use complex models which take the distribution of the passenger population as the input, such as deep-neural-networks. That simplifies the development significantly and still archives reasonably good results.

It is of high importance that the features which are used for the prediction model are available at the time of the prediction. In order to ensure that, the documentation of the features was read on KLMs internal company wiki by Atlassian called “Confluence” and only features which are available at the time of prediction were considered. Features which meet that requirement can be of two types. The first types are features where the feature it self will be unchanged during the complete duration of the prediction and only the value of the feature might change. Such a feature is, for example, the number of available economy class seats. For that feature, the number of available seats can change when for example the aircraft is changed or the curtain between the business class and the economy class gets moved, however the feature itself stays the same during the complete duration of the prediction. The second type of feature which can be considered are features where the feature is changing over time of the prediction. Such a feature is for example the total number of passengers on the flight. The exact number of passengers on the plane are only known after boarding, which means that an alternative has to be used prior to boarding as a replacement. This can be done in two ways depending on the feature, either by using the expected value for that feature or by using other features which are known earlier in the process as a replacement until the feature becomes available. To determine the number of passengers during the prediction horizon with other features, the approach of using the number of booked as well as checked-in passengers as replacement can be used. To determine the expected value of the number of passengers on board, the information used for overbooking an aircraft can be used, since there it is already calculated how many passengers will probably be on board.

4.2.2 Feature inspection

In order to get a good understanding of the numerical features as well as their behavior, the first step was to plot all the numerical features. The first type of plot which was chosen for this was a scatter plot, in which the features were plotted against the dependent-variable of "total hand luggage pieces collected". The second plot which was made was a histogram in which the times of occurrence of the feature values were plotted.

4.2.3 Feature encoding

Next to the numerical features, some features were also time based or categorical. Before those features could be used in the prediction model, they had to be encoded. For the time based features, two types of encoding strategies were tried. The first encoding strategy which was used was label encoding, here the time based features were binned if necessary (for example the time based on hours) and then encoded to a number. The second encoding strategy which was used to encode time based features is cyclical encoding. This is an encoding strategy which helps to encode continuity into the features. With label encoding, there would be a small connection between 11pm and 1am, since 11pm would be encoded as 23 and 1am would be encoded as 1. With cyclical encoding, this continuity can be encoded by encoding the feature using sine and cosine functions. For categorical features, the encoding strategy of one hot encoding was used. Here each category a new column got created, and a one got placed inside that column for elements where the feature was present, while a zero was placed in the column for elements where the feature was not present.

4.2.4 Feature reduction

The feature reduction was done using a correlation study in two steps. In a first step, features were removed, which had a low correlation against the target variable of total hand luggage pieces collected. In a second step, features were removed which had a high cross correlation with other features.

Correlation against target

To test for the correlation against the target variable, the correlation of each independent variable against the target was tested with the Pearson, Spearman, and Kendall test to ensure that a variety of correlations could be detected. Features for which the test with the highest correlation had a correlation with the target variable which was significantly less than $|0.1|$ were disregarded, since they have too little explanatory significance.

Correlation between features

Since a lot of features had a high amount of cross correlation, the next step of the feature selection was to disregard those features. In order to do that, a correlation matrix was set up where the cross correlation was tested using the Spearman correlation. The Spearman correlation was chosen, since it was for most features in the prior test the correlation test with the highest correlation. For features which had a cross correlation greater than |0.9| with one or multiple other features, it was then inspected which of the features with high cross correlation has the highest correlation with the target variable of total hand luggage pieces collected. The feature which did have the highest correlation were then kept, while the other features which had a high cross correlation with that feature, but a lower overall correlation with the target were disregarded.

4.2.5 Eliminating values

In the data preparation, two elimination steps have taken place. Features which were constant were eliminated, since they do not have any predictive value and are therefore not contributing to a good prediction. Additionally, flights which had NULL entries were eliminated since a prediction with NULL values would not be possible.

4.2.6 Scaling

Since the features of the model are in completely different scales, they get scaled before they are inputted into the model. This is useful, since if there is a large discrepancy between the features the model might learn large weight values which make the model unstable, which is not desirable. As a model scalar, the SKLearn Min-Max-Scalar was chosen since it rescales all features between zero and one. Scaling values between zero and one is a good, since that is the range where floating point values are the most accurate. The model scaling was done using the Scikit-learn pipeline to prevent information leakage into the test dataset.

4.3 Overflow model selection

In order to find models which fit the purpose, several factors have to be considered. The first question which should be asked is if a Time series or cross-sectional model is a good fit. Then one has to decide if a Regression or Classification model fits the purpose better. Lastly, one has to decide what exact model within those specifications is a good fit and if it makes sense to use an Ensemble method.

Time series vs. cross-sectional

The first question which has to be asked is if the data is fluctuating over time. If the data changes a lot over time, such as with the stock market, then a time series model should be chosen. If the model is fairly static over time, a cross-sectional model is the better choice. For the hand luggage prediction, it can be seen when taking a look at the correlation of the different features with the target variable, that time based features only play a very small role in terms of correlation with the target variable. Therefore, the decision had been made to use a cross-sectional model.

Regression vs. Classification

The next decision which has to be made is if the model should be a regression or a classification model. The difference between the two types of models is, that for a regression model the output is a continuous quantity, while in a classification model the output are discrete class labels. Since the number of hand luggage pieces which have to be collected is a continuous variable, it makes sense to pick regression models for this purpose.

4.3.1 Performance measurements

In order to test the models, performance measurements are a vital part. In order to judge the performance of the model best, different performance measurements are important during different phases of the model development. During the model selection phase, the model performance gets tested using the Adjusted R^2 measurement as well as the root-mean-square error measurement. The Adjusted R^2 is a great measurement for model accuracy, which unlike the R^2 measurement only considers the independent variables which have an effect on the model performance. The root-mean-square error is a measurement which measures the difference between the predicted values of a model as well as the actual values. Unlike the mean-square error measurement which does measure that as well, the root-mean-square error does however penalize predictions which are further off more than predictions which are close to the target. This makes the root-mean-square error not only a great measurement for performance measurements but also makes it the ideal measurement to be used as a loss function for the models.

The drawback of both the adjusted R^2 and the root-mean-square error is, however, that they are both not using pieces of hand luggage which are overflowing as units of measurements, making the intuitive understanding of the measurements quite a challenge. Since it is crucial that the performance of the model can be understood by all stakeholders, additional performance measurements are used for that purpose. The first measurement which is used is the mean absolute error, which is measuring how large the average deviation from the actual value is. Next to that, a custom performance measurement is used, which is measuring what the maximum deviation of hand luggage pieces is for 80% of the predictions. The target set by KLM for this measurement is, that 80% of the predictions have a maximum deviation of 5 pieces of hand luggage. Next to doing all the performance measurements against the targets, the model gets also put the model into context, by being tested against the current model to see the improvements against the baseline.

4.3.2 Model selection

In order to find out which regression model does perform best, several common regression models are tested with a train test split. 70% of the historic data (Jan 1, 2023 – June 1, 2023) is used as training, while the other 30% is used for testing. The models which perform best were explained in more detail in section 3.2. For all models, the default hyperparameters were used, since extensive hyperparameter tuning with a model at this size is very resource intensive, and it is more practical to do that on the KLM servers rather than on a local machine. The considered models for the model selection step are:

1. Linear regression
2. Lasso
3. Elastic Net
4. k-nearest neighbors regression
5. Decision Tree regression
6. Epsilon-Support Vector regression
7. Random Forest Regressor

8. Ridge regression
9. Least-angle regression
10. Gradient Boosting regression
11. Bagging regression
12. Multi-layer Perceptron regression
13. Stochastic Gradient Descent regression
14. Extreme Gradient Boosting regression

4.3.3 Ensemble modeling

Once good performing models have been selected, it was tested if combining the best performing models together in an ensemble method would improve the model performance. As ensemble methods, a voting regressor model as well as a stacking regressor has been tested.

The voting regressor trains each of the selected models separately and then lets each of the models predict the result. It can be configured to either take the average or a weighted averaged of the predictions of the individual models to form the final prediction. In this case, the voting regressor has been tested with both a simple average as well as a weighted average based on the adjusted r^2 of the contributing models. The voting regressor is particularly helpful in case when the prediction of one of the models is far from the actual value, while the other models are closer to the actual. The reason for that is, that by using the voting regressor the impact of the model which is far off can be minimized by the other models making the overall predictions more accurate.

The stacking regressor also trains the selected models separately, but then combines the results of them together by using another model. The models which have been tested to combine the models were a linear regression model as well as a ridge regressor. Similarly to a voting regressor, the stacking regressor can be used to increase the accuracy, however compared to the voting regressor it can learn about the strength and weaknesses of the contributing models and use that in the combination of their input.

4.3.4 Implementation plan

In order to develop an implementation plan for the new model, the landscape in which the model has to be embedded has been analyzed. This included an analysis on what can be implemented so that the model is supported best, who has to be involved in the model deployment as well as how the model should be run when it is implemented.

4.4 Cost model

In order to predict the cost of the delay, KLMs Forecost API is used. The values received by the API are put into context using the cost calculations from literature of Eurocontrol [21] and the University of Westminster [22]. The delay associated with hand luggage is stored by KLM as delay code 10. Delay code 10 can occur as the first delay code or the second delay code. Both of those occurrences of delay code 10 are combined for all the cost models and a combined feature is developed which includes both the duration of delay code 10 as first delay code as well as, as second delay code. As part of the cost model development, three different cost models have been developed. One Machine learning cost model which works similarly to the overflow prediction model, one historic cost model as well as one dynamic cost model.

4.4.1 Machine learning cost model

The machine learning cost model has been developed with the purpose of being able to input flight characteristics and get as output the expected cost due to hand luggage for a flight with those characteristics. In order to output the cost, a mechanism was developed which allowed for the input of flight characteristics, which is then outputting the expected delay minutes incurred by the collection of hand luggage for such a flight. In order to be able to get the expected cost of that flight, the delay minutes can then be mapped to the cost associated with the delay minutes on that flight from Forecost to get the expected cost associated with the hand luggage policy on that flight. The delay minute calculation part of the model has been developed similarly to the hand luggage overflow prediction model. The difference between the two models is, that for the cost calculation model the target variable is now the delay minutes of delay caused by hand luggage rather than the total hand luggage pieces which have been collected. Additionally, the threshold of correlation for the features with the target has been set at |0.05| rather than |0.1|. The reason for that is, that in the case that it would have been set to |0.1| not enough features would have been of relevance. Next to those changes, the steps of predicting were identical to the overflow prediction. Features which had a high cross correlation were eliminated, as well as features which were constant. Flights which had NULL values were disregarded as well, and the features were scaled with the min-max-scalar. It was then tested which models performed best in predicting the delay, and it was tested if using an ensemble model would improve the prediction. The prediction of that model could then be used to find the associated cost for that flight on Forecost.

4.4.2 Historic cost model

Next to the machine learning prediction model, a model has been developed which allows KLM to see how high the cost of past delays were for them. This is of relevance for KLM, since KLM currently has no mechanism to track how expensive the occurred delays associated with hand luggage overflow have actually been. Since the cost calculations of the Forecost model are deleted as soon as a flight operates, there is currently no possibility for KLM to know how much delay they have had on a flight while at the same time knowing how expensive that delay has been. This model is saving the cost predictions of Forecost for all upcoming flights and averages them for the different delay minutes by aircraft type. This allows us to get a rough estimation of how much the average delay cost is per aircraft type. Since the delay minutes outputted by Forecost are [predicting partially less granular than needed for this model (exact prediction horizon and steps removed)], a linear approximation is implemented so that also delay minutes between the cost values outputted by Forecost can be measured. The cost prediction by Forecost then gets mapped to the delay minutes incurred in the past, which are saved inside “BlueLagoon”, to get a rough estimation of how much cost the delay in the past has generated for KLM.

4.4.3 Dynamic cost model

In order to allow the KLM team to make rough estimations about the cost which they can save by implementing a new hand luggage model, a small tool is implemented with which KLM can model improvements to their hand luggage prediction. With this model, which they can use to test “what-if” scenarios, the user has to input a date in thee past to get the proportion of different 737 types which operated on that day. Next to that, the proportion of flights which should incur delay code 10 as well as the average duration delay code 10 should have been inputted. Based on the amount of flights which are departing on the specified day as well as the selected metrics, an approximation of the delay costs is then made. Since the metrics can be changed in the model, it is useful to see how performance improvements as well decreases effect the cost KLM is incurring. In the model, the number of flights for each 737 subtype which operates on the day which was inputted gets multiplied by the proportion of flights which the user specifies should incur the delay code 10. The resulting values for the 737 subtypes are then the proportions of 737 planes which would be incurring a delay due to hand luggage. To get the cost those flights would then occur, again Forecost is used. Identically to the historic cost model calculations, the costs for the flights occurring in the upcoming two days are retrieved and averaged per plane type. The proportion of 737 planes which will incur a delay due to delay code 10 will then be multiplied by the respective average cost which was specified by the user to get the average cost in that scenario.

4.5 Process improvement plan

In order to determine weak points in the processes of KLM, the complete customer journey was analyzed with a focus on the hand luggage collection and the aim in mind to ensure a hand luggage collection as early as possible in the process. This included an analysis of a model of the current customer journey, as well as an analysis in the field to see where improvements can be made. This analysis was complemented by discussions with stakeholders to better understand the circumstances as well as the opportunities KLM has to improve the hand luggage collection.

5 RESULTS

The results section has three parts to it. It starts by covering the results of the development of the overflow prediction model. After that, the results of the different cost models which have been developed are covered. The chapter concludes with covering the process improvements which have been identified as part of the research.

5.1 Overflow model

There are several steps within the overflow model in order to get to the final result of the model. First, the results of the correlation study are presented, followed by the results of the model selection. Next, the ensemble model is covered and insights into the proposed implementation plan are given.

5.1.1 Feature analysis

The correlation study has two parts to it. The reduction of features with a low correlation to the target variable of total hand luggage pieces collected which holds for features which have a correlation of significantly less than |0.1|, as well as the reduction of features which have a cross correlation greater than or equal to |0.9| with another independent variable. As mentioned in the section 4 about Methodology, the features of the FlightLegs table were considered and were filtered by features which were available at the time of prediction and which made sense for the prediction. The features which proved to have both a sufficient correlation with the target variable as can be seen in Figure 5.1 as well as a low enough cross correlation with the other features which has been determined using the correlation matrix which can be seen in Figure 5.2 were kept.

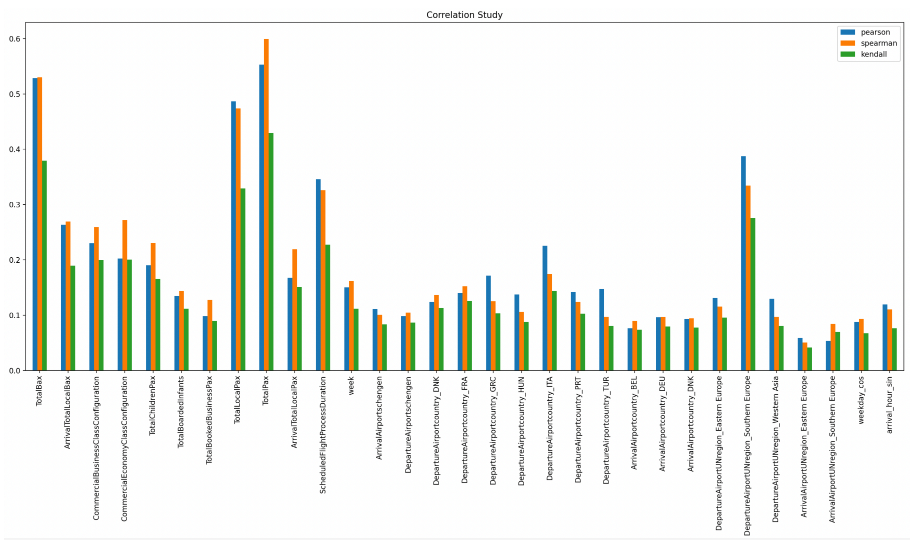

Figure 5.1: Correlation test against total hand luggage pieces collected

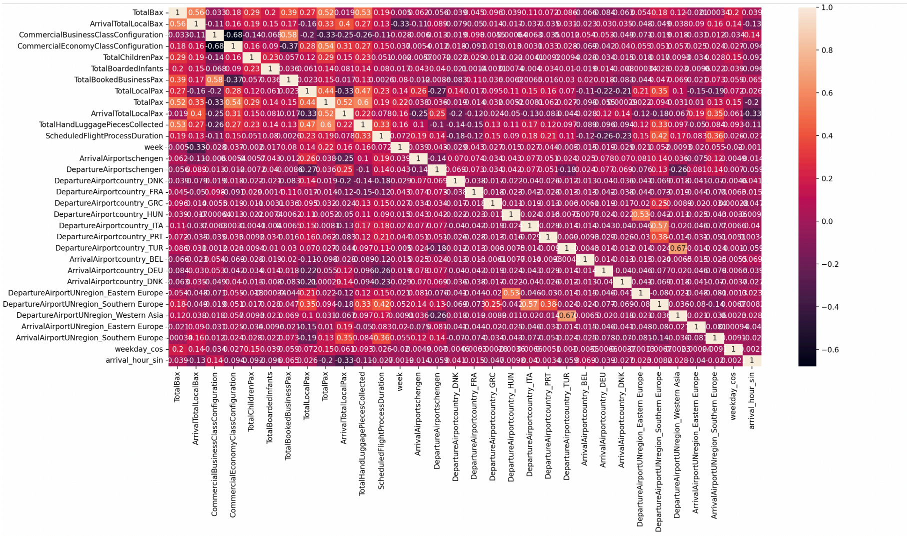

Figure 5.2: Correlation Matrix

There are three types of features remaining, “Flight based features”, “Location based features” as well as “time based features” as can be seen in Figure 5.3. The correlation with the target variable of total hand luggage pieces collected of the remaining features can be seen in the bar chart of Figure 5.1. The five features which in the end had the highest correlation were the features of “TotalBax” which is the feature describing the total amount of hand luggage checked-in, “TotalLocalPax” which is describing the total amount of local passengers, “TotalPax” which is describing the total amount of passengers, “Schedule flight process duration” which is describing how long the flight takes as well as “DepatureAirportUNregion_Southern Europe” which is a one-hot encoded feature which is one, when the departure airport is in Southern Europe.

Flight based	Location based (country)	Location based (region)	Time based
Total luggage pieces	Departure airport country DNK	Arrival airport Schengen	Week Number
Arrival total local luggage pieces	Departure airport country FRA	Departure airport Schengen	Weekday
Commercial business class configuration	Departure airport country GRC	Departure airport eastern Europe	Arrival hour
Commercial economy class configuration	Departure airport country HUN	Departure airport southern Europe	
Total children Pax	Departure airport country ITA	Departure airport western Asia	
Total infants	Departure airport country PRT	Arrival airport eastern Europe	
Total booked business Pax	Departure airport country TUR	Arrival airport southern Europe	
Total local Pax	Arrival airport country BEL		
Total Pax	Arrival airport country DEU		
Arrival total local Pax	Arrival airport country DNK		
Scheduled flight duration			

Figure 5.3: Overflow model features

When plotting the numerical flight based features against the target variable of total hand luggage pieces collected, it can be seen, that while the upper bound usually forms a function (either linear, cubic, or bi-model), the rest of the values are everywhere below the function up all the way down to the x-axis as can be seen in Figure 5.4. A notable observation which can be made beyond that is concerning the feature of “Commercial economy class configuration”. For it, it is noticeable that there is a gap around 140 economy seats in the plot. This is because there are three subtypes of the 737 family active withing KLM, namely the 73H, 73J and 73W. Those subtypes have different length and therefore also a different amount of economy class seats. Next to plotting the features against the target, the features have also been plotted as histograms. Here again the gap could be observed for the Commercial economy class configuration as can be seen in Figure 5.5.

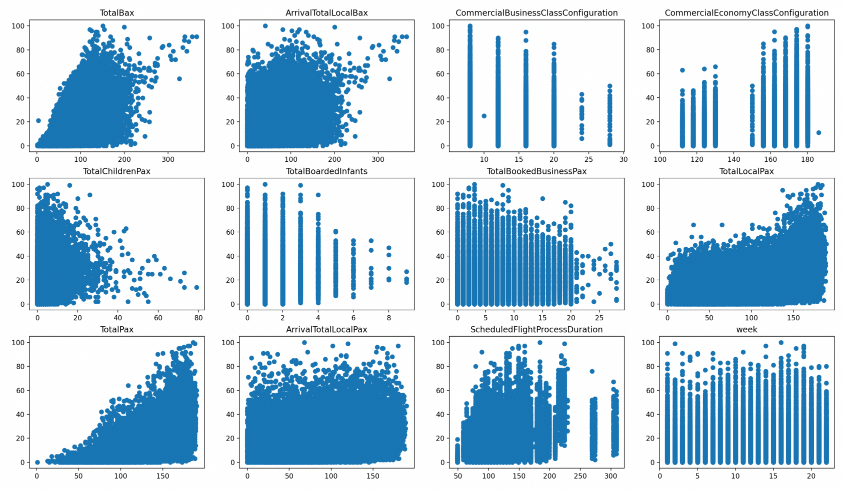

Figure 5.4: Feature against target plot

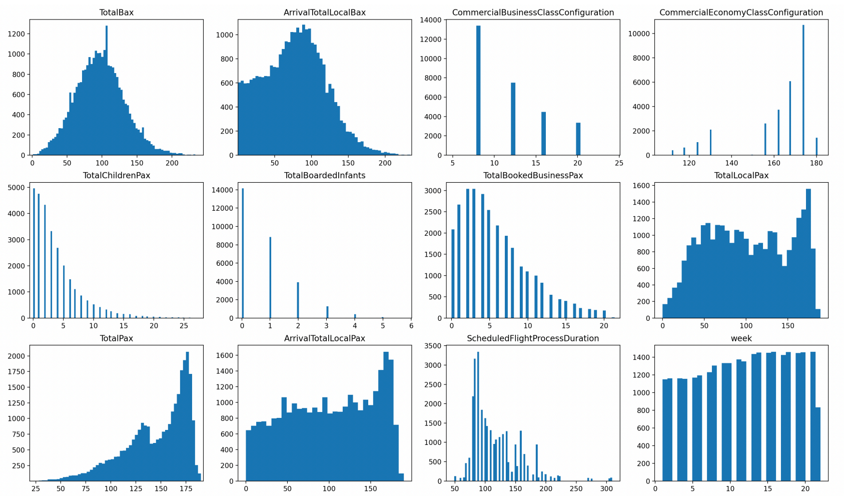

Figure 5.5: Feature histogram

In the process of the feature selection, the most surprising aspect which could be observed was the unimportance of seasonality based features. The only seasonality based features which exceeded the correlation threshold were the features of week number and week day. Features one would traditionally associate with seasonality such as Dutch school holiday, Dutch national holiday or the number of surfboard or ski on board had a correlation below the threshold with the total amount of hand luggage pieces collected. Next to that, it was also surprising to see, that there is also a low correlation between the planes who have so-called "Space Bins", which are extra large overhead bin compartments to accommodate more hand luggage, and the total number of hand luggage pieces collected. This could be either because they are not as effective as advertised by Boeing, or because KLM is already assigning those planes to routes where they expect a lot of hand luggage, making their effect not noticeable.

5.1.2 Model selection

For the model selection, the models mentioned in section 4.3.2 have been evaluated. The models which have been performing best in terms of the adjusted R^2 value as well as the root-mean-square error (RMSE) were the Extreme Gradient Boosting regressor with an adjusted R^2 of 0.7 and an RMSE of 7.45, the Random Forest regressor with an adjusted R^2 of 0.68 and an RMSE of 7.57 and the Multi-layer Perceptron regressor with an adjusted R^2 of 0.69 and an RMSE of 7.70. The full list of results is attached to Figure 5.6. When particularly focussing on the adjusted r^2 of the models, it could be seen, that the three models mentioned above are all close together, and after them is a significant gap to the next best performing models which is are the Gradient boosting regressor as well as the Bagging regressor who both have an adjusted r^2 of 0.66. It therefore makes sense to use those three models together in the ensemble models since they require models which are performing quite similar as their components.

Model	RMSE	MAE	Adj. R^2
Linear Regression	8.385	6.41	0.61
Lasso	11.98	9.47	0.21
Elastic Net	12.90	10.2	0.09
KNN	8.67	6.56	0.59
Decision tree regressor	10.95	10.96	0.34
SVR	8.11	6.11	0.64
Random forest regressor	7.57	5.74	0.68
Ridge regression	8.39	6.42	0.61
Lars	11.3	8.79	0.29
GBR	7.90	5.98	0.66
Bagging	7.92	6	0.66
MLP	7.70	5.86	0.69
SGD	8.43	6.44	0.61
XGB	7.45	5.61	0.70

Figure 5.6: Prediction results of the regression models for hand luggage overflow

5.1.3 Ensemble model

When comparing the model performance of the four analyzed ensemble model variations in Figure 5.7, it can be seen that the weighted average voting regressor model is performing best in all metrics. When taking KLMs custom metric of the 80% maximum absolute deviation into account, the weighted-averaged voting regressor is reaching in 80% of the cases a deviation of less than 8.5 hand luggage pieces.

Model	RMSE	MAE	Adj. R^2
Voting regressor (weighted average)	7.29	5.50	0.72
Voting regressor (average)	7.32	5.59	0.71
Stacking regressor (Linear regression)	7.31	5.53	0.71
Stacking regressor (Ridge regression)	7.27	5.51	0.71

Figure 5.7: Performance of the ensemble models

It is unlikely that the model performance can be increased by further fine-tuning the voting regressor, since all tested ensemble models are performing similarly well. However, the overall performance of the combined model can be increased by doing hyperparameter tuning on the underlying models. The voting regressor only makes sense when the model performance of the included models is close together. If the model tuning has, as a result, that the contributing models are increasing unevenly in performance, then it should be considered to do changes to the voting regressor. This should namely be to reduce the voting regressor to only two models which are used in case one of the three is not improving as well as the others or to discontinue the voting regressor all together if one of the models is outperforming all the others significantly.

5.1.4 Comparison with the baseline

When taking a look at KLMs custom metric of measuring the percentage of predictions with a maximum deviation of n pieces, as can be seen in Figure 5.8. It can be seen, that despite the fact that the performance of the current model can not be measured for predicting too much hand luggage, the new model is outperforming it. The reason it can not be tracked how much the current model is predicting too much is, that the collection of hand luggage is guided by the current model, and it therefore does not very often happen, that less hand luggage is collected than predicted. This, combined with the problem of not knowing how much hand luggage was collected too much since no measurements from within the cabin are possible, results in us only knowing for the current model if the prediction was too low. This artificially increases the model performance for this metric of the current model, making it appear significantly better than it is actually performing.

When investigating, the more meaningful metric of the percentage of predictions within a maximum underestimation in Figure 5.8 which can be accurately measured for both models and has most likely the highest connection to the delays incurred, it can be seen that the gap between the two models are apart even more. This metric can be accurately measured for both models

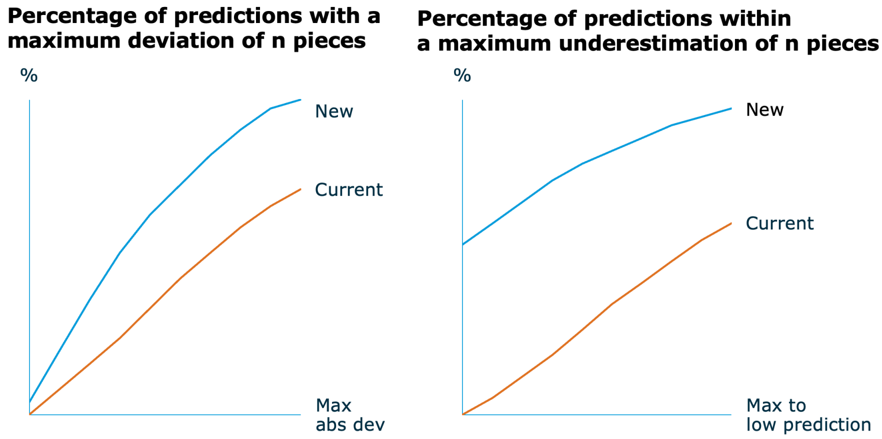

Figure 5.8: Model performance with custom metrics

5.1.5 Implementation plan

In order to ensure a good implementation of the model, three steps are essential. The first step which is recommended for a successful implementation is that all customers of the current model are informed about the model change. This is relevant, since the machine learning model has only been developed for planes of the 737 variant, and therefore the accuracy the customers get displayed differs depending on the plane type. Due to the inaccuracy of the current model, it could be the case, that customers of the current prediction have implemented business rules, which take the inaccuracy of the current model into account. One such business rule which is currently deployed is, that there are always five pieces of hand luggage added to the prediction in the apps of the ground staff in order to ensure that enough pieces are collected. Since the new model will be more accurate, decision-rules have to be revised for the 737 fleet. It is expected that due to the accuracy increase, it is no longer needed to add to the actual predicted number of five extra pieces as collection recommendation, so that in practice there will almost always be more collected than predicted. Instead, it can be preceded with recommending the actual prediction as number of pieces which shall be collected. In terms of model performance it could be seen, that the new model in its current state is predicting in 80% of the cases up to five hand luggage pieces too little. It would therefore be advised as a business rule to still have all staff in the collection process ready to collect up to 5 pieces of hand luggage at the gate.

Another aspect of the implementation plan is to decide when to run the prediction model. The new prediction model is using not that many resources, so KLM is advised to run the model frequently, since there are only very few drawbacks to that. Since the model prediction can change due to a variety of reasons, it is crucial, that the prediction has been run with relatively recent data of the flight which should be predicted. The first customer which needs the data is CRMPush in order to send out the SMS to passengers. This message is sent out 12 hours prior to the first flight. In order to ensure that all passengers who have as a first flight a long haul flight can be reached via SMS, the recommendation would be to run the model 72 hours prior to departure of the flight the first time. Since the model is very light to run, it would make sense to run the model from that point onwards each time when there is either a change with the passengers or when changes are made to the flight (for example, the plane type gets changed and therefore the amount of seats changes).

Lastly, it is important to decide when to retrain the model based on new data. Since the model itself is not that heavy to retrain, the recommendation is to retrain the model on a daily basis. Retraining the model that frequently is especially important for the feedback control system in which the model is embedded in. By getting up-to-date information about the actual amount of collected hand luggage pieces of other flights as well as about the overhead bin fill level of those flights from the cabin crew, the model is able to learn from that information at the moment when it is happening so that it can consider it in the recommendation for future flights. Daily retraining is especially important prior to events where either the location features change such as during a schedule change or when the time based features are likely going to influence the model the most such as when there is a sudden change of customer behavior expected. Thanks to the rolling horizon retraining of the model, this also means that the last day of the current training dataset is then dropped so that the time horizon always stays at half a year. The recommended implementation into the existing infrastructure for training and predicting can be seen in Figure 5.9.

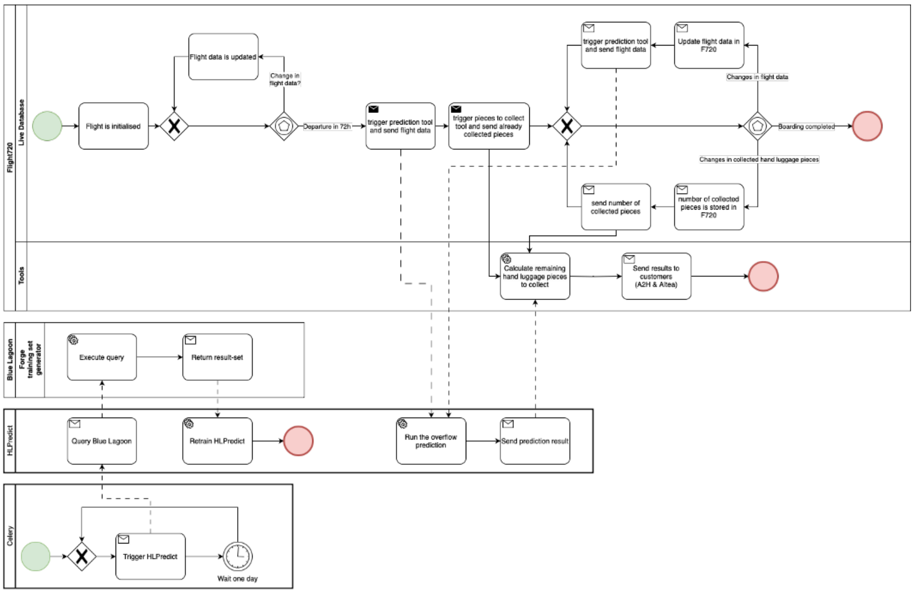

Figure 5.9: Infrastructure BPMN model

When a new country or region is added to KLMs flight schedule, an investigation has to take place to check if the new country is valuable for the model. By default, new countries are automatically added to the model, since it is better to include a country which is not contributing rather than not include one which is valuable for the model. This investigation has to take place manually using the correlation study proposed in this research.

5.2 Cost model

As mentioned in the methodology section, three cost models have been developed. A machine learning cost model, a historic cost model as well as a dynamic cost model where each of the models has a separate purpose.

5.2.1 Machine learning cost model

The machine learning model has the similar steps as the machine learning model for the overflow prediction model. While the three first steps are also that the results of the correlation study are presented, the results of the model selection are shown and the results of the implementation of a voting regressor are covered, no implementation plan shown thereafter.

Correlation Study

Similarly to the correlation study for the hand luggage overflow model, the features have been selected from the FlightLegs table. The features which meet the threshold of |0.05| for this model and at the same time do not have a high cross correlation with other features can be seen in the bar chart of Figure 5.10.

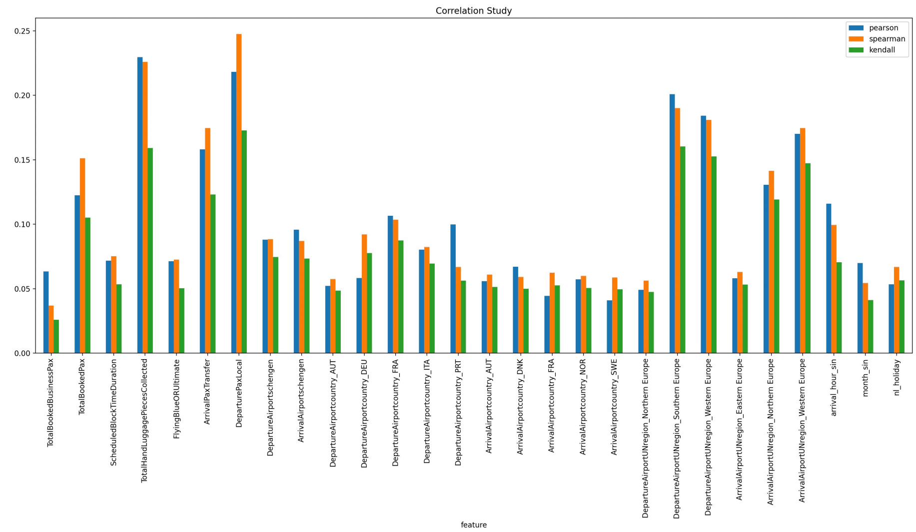

Figure 5.10: Machine learning cost correlation study

Model selection

For all the models which have been used to predict delays due to hand luggage overflow, the adjusted R^2 values have been negative, as can be seen in appendix 5.11. A negative adjusted R^2 implies that a constant line taken at the average is outperforming the model and does indicate a very poor model performance. The models are therefore not suitable to predict the delay caused by hand luggage.

Model	RMSE	MAE	Adj. R^2
Linear Regression	5.21	3.87	-0.04
Lasso	5.55	4.09	-0.08
Elastic Net	5.55	4.09	-0.08
KNN	5.85	4.27	-0.20
Decision tree regressor	7.67	5.56	-1.06
SVR	5.50	3.81	-0.06
Random forest regressor	5.46	4.09	-0.04
Ridge regression	5.20	3.86	-0.05
Lars	6.09	4.34	-0.39
GBR	5.44	3.96	-0.04
Bagging	5.69	4.21	-0.13
MLP	5.26	3.87	-0.02
SGD	5.24	3.88	-0.03
XGB	5.94	4.39	-0.24

Figure 5.11: Machine learning cost prediction model results

5.2.2 Historic cost model

The historic cost model has been run for the duration from January 1, 2023, to July 2, 2023, which is the halfway point of a year. When the cost which is predicted for the upcoming two days is used for the delays incurred during that time period, a cost for that half year of €[7-single digit number (exact number removed)] is incurred. Assuming that the second half of the year incurred the same amount of cost, this results in an annual cost for KLM of €[7-single digit number (exact number removed)] due to delays associated with hand luggage.

5.2.3 Dynamic cost model

The dynamic cost model can be run for any arbitrary day to predict the cost. When taking July 9, 2023, as an example date, we can see, that when the delay is occurring on [arbitrary single digit percent number (exact number removed)] of the flights on that day and the average duration is [arbitrary single digit number (exact number removed)] minutes the expected incurred cost is €[5-digit number (exact number removed)] for that day. When the same day is however taken and the delay would only occur on average for [arbitrary lower single digit number (exact number removed)] minutes, the cost incurred on that day would go down to €[lower 5-digit number (exact number removed)] for that day. The horizon of a day was chosen, in order to both include the proportion of the flights which are delayed as well as the average delay minutes as factors.

5.3 Process improvements

As part of the investigation of the hand luggage overflow problem, two process improvements could be identified which could help with the hand luggage problem. The first improvement which could be identified was the user interface of the self-service baggage drop-off machine, and the second process improvement was the lack of a feasible feedback system for the cabin crews.

5.3.1 Self-service baggage drop-off machine

When investigating the passenger as well as the baggage journey, the observation was made, that KLM not only has a problem with their hand luggage prediction, but that KLM can improve on requesting their passengers to collect the hand luggage prior to boarding. A weak point which has been identified are the Self-service baggage drop-off machines at Amsterdam airport, who have a user interface which makes it for the passengers very unclear what KLM wants from them and that they have the possibility to hand in their hand luggage. In order to improve upon that, and to make the user interface more intuitive, a concept has been developed together with a designer at KLM. The concept which can be seen in appendix A.2 introduces a “home screen”. Home screens are a concept most users are already used to from their laptops and smartphones, and are therefore easy to understand. The home screen has large touchpoints for both the check-in of hand luggage as well as normal luggage. Each of the touchpoints displays how much luggage pieces the passenger has booked, since most passenger do not remember what they have booked. It also points out, that the hand luggage can be checked in at the self-service baggage drop-off machine, which currently does not get communicated. By reminding passenger at the step of the luggage drop off about the possibility of checking in the hand luggage, we are hoping to increase the amount of voluntarily dropped off luggage.

5.3.2 Cabin crew feedback system

When analyzing the available data for the overflow prediction model, it became obvious, that KLM currently has no way to track how full the overhead bins are. While the voluntary crew feedback form, has a spot in which the cabin crew can give feedback about the hand luggage situation on board, the option seldom gets used since it is buried under several sub menus. Also, since the feedback form often gets filled out with a long time delay to the flight, the results are often inaccurate. It therefore gets recommended to include a mandatory feedback field which automatically appears shortly after the overhead bins are closed on the iPad of the cabin attendants. The feedback option is recommended to cover the full screen and consist out of three buttons where different fill levels can be indicated, as can be seen in appendix A.1.

6 CONCLUSION & RECOMMENDATIONS

The conclusion and recommendations chapter has three parts to it. First, the results of the research are summarized in the conclusion section. After which they are discussed and further recommendations are given. Lastly, this chapter is concluded with the limitations and further research section.

6.1 Conclusion

The research has shown, that the features seen in Figure 5.3 should be considered for the hand luggage overflow prediction of KLMs 737 fleet based on the results of the correlation study. It could be seen, that the Multi-layer Perceptron regressor, the extreme gradient boost regressor as well as the random forest regressor are the best performing models with the default hyperparameters and that combining them into a voting regressor which is taking the weighted average based on the adjusted r^2 of the models is further increasing the prediction performance.

The overall performance which could be achieved with this setup is already a significant improvement over KLMs current model, however still below the initially set goal. The goal of the study was to find out how a model can be designed that predicts the hand luggage overflow of a given flight and guides collection process improvements. It could be determined, that a machine learning model which uses features which are determined by using a combination of business logic and a correlation study, and which combines a multilayer perceptron regressor model, an extreme gradient boosting regressor, and a random forest regressor inside a weighted-average voting regressor is a feasible way to develop such a model which does perform well enough to guide the collection process of KLM even without hyperparameter tuning of the models. The study is confident that with the future step of tuning the hyperparameters, the proposed model will also be able to meet KLMs set performance targets.

6.2 Discussion

The discussion section of this thesis has six parts to it. First, the overflow model is discussed, where the most surprising findings all related to the feature selection step. Next to that, the cost models are discussed, followed by the implications of the research. After that, process improvements are discussed, as well as why a simpler solution in which the customer is declaring their hand luggage during booking or check-in is suboptimal. Lastly, a rough estimation is made about potential cost savings due to the implementation of the overflow model.

6.2.1 Overflow model

The correlation study of this research shows, that the amount of checked-in luggage on board is with a correlation coefficient of about 0.5 for both the Pearson and Spearman correlation surprisingly highly correlated with the total amount of hand luggage collected. This is making it the feature with the second-highest correlation overall with the target, behind the total amount of passengers on board. Despite the relatively high correlation, the feature of total amount of luggage on board is however also just correlated with a maximum of 0.5 to the other features. This is especially surprising in regard to the feature of total amount of passengers on board, since a higher correlation would be expected there. This combination makes it a surprisingly useful feature for the prediction model, since it allows the feature to give valuable insights which are not redundant.

Besides the checked-in luggage feature, the location based features were also notable. The two aspects which specifically stood out there were that most location features had a rather small correlation with the target, suggesting that despite the fact that some locations appear as hot-spots in the analysis of the delay minutes, their influence on the amount of hand luggage which is collected seems to be limited. This is suggesting that at those hot-spot locations, a lot of the delay results from bad handling of the hand luggage collection rather than the overflow itself.

In terms of features, it was also surprising, that date based features do not perform well in terms of correlation. It was particularly, surprising, that the feature which was checking if a flight was departing during a Dutch national holiday or one of the Dutch school holidays had a very low correlation. This is suggesting that despite the fact that the demographic is changing during that time period [more local passenger (exact proportion removed)], the behavior in terms of hand luggage is not.

6.2.2 Cost models

For the cost model it was surprising that the cost predictions of the Forecost API are generally significantly higher, as can be seen in the table in Figure 6.1 as well as the associated chart in Figure 6.2, than the cost predictions made within the literature by the University of Westminster and Eurocontrol in Figure 3.3. This has the effect, that all the cost calculations which are made in that section are significantly higher than if they had been done with the cost calculations by the University of Westminster and Eurocontrol. The reason for this difference is most likely not inflation, since the cost calculations on the website of Eurocontrol are inflation adjusted, but rather a result of the fact that airlines were only partially willing to provide cost information as stated by the authors themselves in the study. Next to that, also only a limited amount of airlines currently have the capabilities to determine their cost as detailed as KLM, making limiting the input they can provide to the study.

This picture has been removed
due to confidentiality

Thank you very much for your understanding

Figure 6.1: Forecost delay cost table

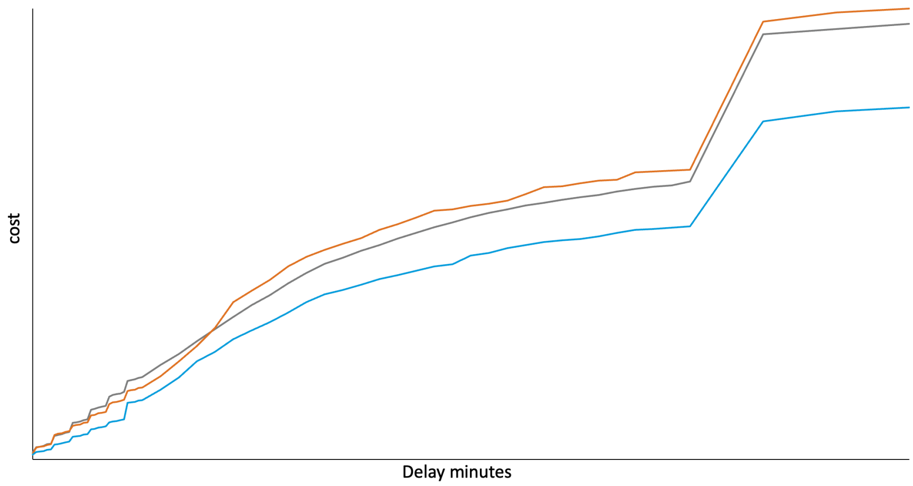

Figure 6.2: Forecost delay cost graph

The machine learning cost model did not perform as well as envisioned, as it was even less accurate than taking the average of the delays associated with the collection of hand luggage. A likely cause for such bad performance of the model is, that there are many factors playing a role in the delay of an aircraft which have not been considered in the prediction model, i.e., the amount of traffic at the airport at that time, the way the gate crew is handling the collection or the cooperation of the passengers. An additional aspect which is making the prediction of the collection time a challenge is, that it is a manually tracked value, leading to low accuracy. A potential approach which could be done in the future to predict the delay as well as the cost KLM is therefore incurring due to hand luggage would be to develop a digital twin of the airports in which different scenarios are simulated in the hope that this would at least partially accommodate the vast variety of factors effecting the delay.

The historic cost model is giving reasonable predictions for the cost KLM has likely incurred in the past, and is therefore a valuable tool for the business team to know how urgent the hand luggage problem is for KLM. It enables KLM for the first time to not only know how much delay there has been, but also how much it costs. Ideally, KLM would store the past cost prediction of Forecost so that the predictions could take the actual predicted costs into account rather than the averaged, however for the time being it is a significant upgrade for KLM.

The dynamic cost model is a very simple tool to predict the cost since it does only take one cost value rather than a distribution. It is therefore not very accurate compared to more advanced models such as the historic cost model. The dynamic cost model does however enable the business product analysts to do rough estimations in a “what if” fashion, which helps a lot in better understanding several potential decisions which can be made.

6.2.3 Implications

The academic implications of the research are, that it enables the research body to bridge the gap between the hand luggage brought by passengers and the expected amount of hand luggage for prediction models. This can be valuable for past research in the field of airline operations, such as in the prediction of boarding times, but also for future research which is using hand luggage as an input parameter. Practically the implications are that thanks to a more accurate hand luggage prediction and the process changes which can be implemented thanks to the more accurate prediction a lot of cost-cutting opportunities appear which helps airlines to stay competitive.

Comparisons of the model proposed in this research with the currently deployed model, always have to happen with caution, since the active model is always influencing the amount of collected hand luggage pieces. This results in the active model always looking better in comparison compared to the model which is trained on the total amount of hand luggage pieces collected number.

The prediction model proposed in this research should be generalizable with only minor changes to other flights and planes where the customer behavior is very similar. For KLM, this means that most likely an adaption to KLM Cityhopper is not very challenging. The generalizability to intercontinental flights is however most likely not easily possible, since the overhead bin capacity is quite different there and the customer behavior is most likely also significantly different to European flights. Nevertheless, KLM should however still be able to use both the feature exploration pipeline as well as the model selection pipeline to determine suitable features and models for their intercontinental flights, even though the resulting features might be quite different and different models might perform best.

6.2.4 Process improvements

Both the process improvement of the improved user interface of the self-service baggage drop-off machine as well as the improvement of the cabin crew feedback system are on their way of being implemented. The user interface proposal has the weak point, that the self-service drop off machines currently do not have the possibility of targeting passengers based on their travel class or status. This means that the machine would only be able to ask all passengers to hand in their hand luggage, rather than being able to exclude premium passengers. It is currently discussed how we could implement targeting of passengers with the help of other departments within KLM so that we can serve the message more granular. The cabin crew feedback system currently does not seem to have any weak points and should be implemented hopefully soon.

6.2.5 Hand luggage declaration

When analyzing the hand luggage overflow problem, one would intuitively think, that the problem could be solved way simpler than with the model proposed in this research by just having the customers declare their hand luggage in advance either during booking or during check-in. While this sounds good in theory, it has two primary problems in practice. The first problem is the variation of the hand luggage, not all hand luggage pieces are of the same size, dimension, and compress the same amount (hard shell vs. soft shell luggage). In order to know accurately how much of the overhead bin would therefore be occupied, more specific information would be needed from the passengers than just the amount of trolleys they are bringing. This would either have to be done via a multistep survey or a scan of the hand luggage with a camera, either by the customer themselves or automatically at the airport. Requiring the passenger to either fill out a survey or to scan their luggage is adding another step to the customer journey which would degrade the passenger experience, making the option undesirable. Scanning the hand luggage with cameras at the airport would be challenging as well, since they are owned and operated by the airports and are therefore inaccessible for KLM. Next to that, another problem is, that some passengers are not only placing their trolley in the overhead bin compartment, but also their personal item as well as for example their coat. This is adding variability which either has to be taken into account in the calculation of the overflow, or it must strictly be enforced that only trolleys are allowed in the overhead bin compartment. Taking the variability into account in a model which is predicting the overflow is making such a model less accurate and is adding complexity to it. Additionally, it would require regular updating of the variability when customer behavior changes. Enforcing that only trolleys can be placed in the overhead bin would not be a pleasant experience for both the cabin crew as well as the passengers, making that option also less viable. As a result of those limitations, the model proposed in this research seems to be the most viable option and has therefore been chosen as a solution for the problem.

6.2.6 Expected cost savings of the model deployment

It is difficult to estimate how much cost KLM will be able to save thanks to the deployment of the new overflow model. The reason for that is, since the study was not able to predict the delay incurred due to the collection of hand luggage using the machine learning cost calculation model. As a result of that, the delay could also not be mapped to the delay cost, making it difficult to quantify how much money KLM could save by having to collect less hand luggage at the gate. Nevertheless, some assumptions can be made to approximate potential savings given the cost KLM is incurring due to delays and the hand luggage prediction performance statistics. Since it can be seen in Figure 2.1 that usually the amount of hand luggage which is predicted is also collected as a minimum. It is therefore assumed, that unless more hand luggage has to be collected, the exact amount of hand luggage which is predicted is collected. In addition, it is assumed, that the amount of hand luggage which has been predicted has already been collected prior to the passengers arriving at the gate. At the gate, it is assumed, that staff can handle up to 5 unexpected hand luggage pieces before delay is incurred. When comparing the prediction performance of the current model with the newly proposed model in terms of in what percent of flights they are a maximum of 5 pieces too low as can be seen in Figure 5.8, it can be seen that the current model is in [more than 20% (exact percentage removed)] of the flights a maximum of 5 pieces off, while the new one will be in [more than 50% (exact percentage removed)] a maximum of 5 pieces off. It is now assumed, that the rest of the flights are considered "critical" which would mean, that for the current model [more than 40% (exact percentage removed)] of the flights are critical and for the new model [more than 10% (exact percentage removed)] would be critical. Next, it is assumed, that the delayed flights are from within the subset. Currently, [less than 10% (exact percentage removed)] of KLMs European flights are delayed, which is [less than 10% (exact percentage removed)] of the critical flights. Under the assumption that this proportion will not perfectly transfer to a more accurate model, since a lot of the delay is also caused by the collection and not only the prediction and therefore [more than the current proportion of critical flights (exact proportion removed)] of the critical flights will have a delay, then the delay proportion of the fleet would go down to [less than original (exact number removed)]. Under the assumption that the proportion of the delay distribution would stay the same, this would result in annual savings of about [7-digit number (exact number removed)] euros.

Next to the direct cost savings, KLM is expected to have, which are discussed above it is however also likely that KLM will have indirect cost savings namely in the field of load control thanks to better knowing how heavy the luggage is because it can get weight when it is collected earlier than at the gate. As well as via being better able to predict the loading time needed for the aircraft since more luggage is known in advance.

6.3 Recommendation

In order to reduce the delays caused by hand luggage and to make use of the improved overflow prediction, it is critical that the collection of hand luggage also gets improved and that most of the hand luggage collection is happening before the passengers are at the gate. In order to decide what communication measures are feasible from a cost perspective, the cost historic cost model from this study can be used to determine the money KLM is loosing due to hand luggage and project how much different measures could cost KLM relative to that. In order to ensure, that the communication with the customers is effective, it is recommended for KLM to set up a study which determines how effective different communication strategies are. A good baseline for such research is the study made by Ham in collaboration with Transavia [6].

In order to properly implement the findings of the research, KLM is also advised to only implement the model in tandem with the feedback control system, as can be seen in 6.3. This system is currently developed and is going to use the crew input discussed in section 5.3.2 in order to ensure that the model is kept in check during operation and does not start to fall into an unbalanced state in which the model is learning from itself that the best cause of action is to collect as much hand luggage as possible.

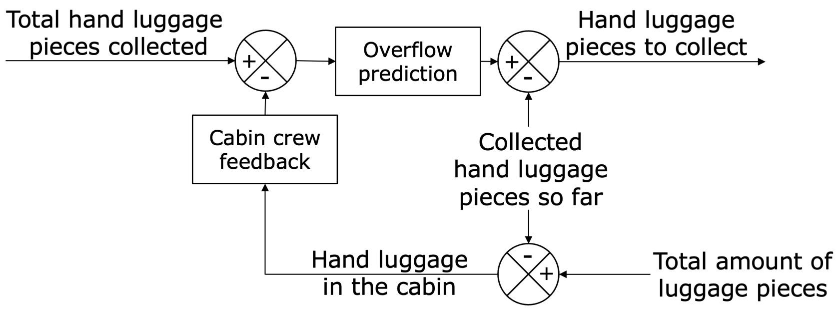

Figure 6.3: Feedback control system

Next to that, KLM is advised to tune the hyperparameters of the contributing models for optimal performance of the prediction model. Lastly, once the prediction model is deployed, KLM is advised to retest the location features when there is a schedule change in order to ensure that non-contributing features are removed and contributing features are included.

6.4 Limitations & further research

The study has several limitations which are covered in the limitation section, such as the limitation of the plane type as well as the limitation due to hyperparameters. Next to that, this section includes the assumption as well as recommendations for further research.

6.4.1 Limitations

The biggest limitation of the research is the lack of hyperparameter tuning of the model. Since hyperparameter tuning is quite resource intensive especially for larger models such as this model, it was decided, that it is more practical to do the hyperparameter tuning on KLMs servers which have more resources after the main research has ended. As a result of that decision, all used models are only run with the default parameters, which is limiting their performance.

Another limitation of the research is, that the proposed model only focus on features on the flight level and does not include features which are based on the passenger population, such as the distribution of the age of the passengers. This decision has been made since it allows us to not be required to use complex models which take the distribution of the passenger population as the input, such as deep-neural-networks. That simplifies the development significantly and still archives reasonably good results. Since the gate staff has made the observation that the features of the passenger distribution do have some effect on the amount of hand luggage passengers are bringing, there is the potential that including such features could increase the model performance. It would be recommended that if such model will be developed, that it will be concatenated with the current model to include the benefits of a model which includes the passenger population and the model which does focus on flight level features.

Next to those limitations, the model also has the limitation that it has only been developed for KLMs 737 fleet. The reason for that is, that other aircraft types most likely do require different features as well as different prediction models. The 737 was chosen, since the hand luggage problem is most severe with it.

The final limitation of the model is, that the model is predicting the hand luggage overflow rather than the amount of hand luggage the passengers are bringing. This has to be done, since at this point in time, KLM does not want to add a step within their passenger journey in which passengers have to declare the amount and dimension of the hand luggage they are intending on placing in the overhead compartment, which would allow for a significantly simpler prediction. Predicting the variable indirectly instead of directly can however have an effect on the model performance. In some cases it is the case that in such models the accuracy of the model will be higher than in a model with a direct prediction, while the precision of the prediction is lower [23]. It therefore has to be carefully considered, if a model with an indirect prediction is desirable. Since it is the only viable option in the scenario, this trade off has to be accepted.

In order to ensure that the model is also working in the long term, it is crucial to embed it in a feedback control system which is adjusting the variable of total hand luggage pieces collected with feedback from inside the cabin. The feedback from the crew is currently not done for every flight and is highly inaccurate. In order to ensure that the data is always available, we are working together with the team who is developing the software for the crew iPads to make a simple screen with which feedback can be given for every flight.

6.4.2 Assumptions

An assumption which had to be made for the model development is, that the final value of each feature is representative of the expected value of that feature prior. More concert, that for example the final number of luggage pieces on board is about the amount which KLM is expecting. This assumption was made since it makes most sense to train the model on the final value, since that is the value which is reflecting reality most closely. For testing of the model, the same final values of the features were also used as feature values, since they are the values which are available in BlueLagoon. In practice, the prediction will however never use the final values for prediction, since during the prediction horizon only the expected value of the feature or an alternative feature with the same meaning (for example: total booked passengers/total checked in passengers) instead of the actual feature (total passengers) is known. Using the final values of the features is therefore not an ideal assumption, however a necessary assumption given the circumstances.

6.4.3 Future research

In the process of doing the research, two new research questions have become clear which are worth a further investigation, which are discussed in the future research questions subsection. Besides that, an alternate way in predicting delay is proposed besides the simulation approach mentioned above in the delay prediction section. After that, it is proposed to improve porter scheduling and customer communication to reduce the hand luggage delay. Lastly, hyperparameter tuning as well as a custom loss function are proposed as potential model improvemnts.

Future research questions

In the process of doing the research, two new research questions have become clear which are worth a further investigation. The first one has come to light during the correlation study, where space bins did not play a significant role despite Boeing advertising them as a solution for the hand luggage overflow problem. The reason for the low correlation could either be that KLM is already assigning space bin aircraft to routs where they are expecting a high demand for hand luggage, and the effect of the Space bins is therefore not that visible. Or it could be that space bins are not as effective as advertised. A potential new research question could therefore be how effective space bins are in the real world in increasing hand luggage storage space. This would be off value for airlines such as KLM, since it could help them to decide if space bins are worth the extra investment. The second potential new research question has come to light while it was tried to predict the delay minutes caused by hand luggage using machine learning in order to predict how costly certain hand luggage decisions are. Predicting the future delay minutes based on the amount of collected hand luggage pieces as well as the flight parameters using regression models was not successful. It could however be possible that the delay minutes can be determined using a simulation model. The resulting research question is therefore if it is possible to predict the delay for an aircraft using agent based modeling.

Delay prediction

Based on this study as well as past studies which predict boarding times, future researches could try to develop a model which predicts the delay caused by the hand luggage collection and therefore the cost associated with certain hand luggage related decisions. This could work by using the overflow model study of this research to predict how much hand luggage has to be collected at the gate. The researcher could then use the prediction model by Schulz [7] to predict the boarding time of the plane, taking into account the amount of hand luggage which is predicted. A potential delay could then be calculated by subtracting the predicted boarding time plus the predicted time which should be after the boarding from the planned time. This approach could be an alternative approach for predicting delay caused by hand luggage to the agent based modeling approach mentioned above.

Porter scheduling

Porters are the airport staff which transport the hand luggage from the gate to the loading crew of the aircraft. Currently, the schedule of the porters is planned far in advance and not adjusted on the fly. Additionally, both the gate crew as well as the aircraft loading crew does not know if porters are assigned to a flight. Since porters can have a significant impact on the loading time of hand luggage which is collected at the gate onto the flight, it is recommended to include them into KLMs “Chip” application, which is a custom version of Inform GroundStar. This would allow for a better planning of the loading times of the aircraft (Variable task time), which could reduce delays due to the loading of hand luggage which is collected at the gate. In a second step, KLM should also consider, if the scheduling of the porters could be supported by an optimizer which schedules the porters based on the amount of hand luggage which is predicted to be collected at the gate.

Customer communication

Currently, the only customer communication in terms of the collection of hand luggage does take place via SMS 12 hours prior to departure. KLM should investigate possibilities of including that messaging also in more of their digital products such as their app, the digital boarding pass or the online check-in procedure. This could increase customer awareness of the possibility of being able to check-in hand luggage for free, since many customers are not aware of the possibility.

Passenger population model

Next to the model proposed in this study, there is also the possibility to implement a model which is using the distribution of the passenger population to predict the hand luggage overflow. This could be done using a deep neural network which takes as input the distribution of features of the passengers on the plane such as origin airport, destination airport, age as well as nationality and encode them into a four dimensional tensor. This distribution could then be mapped to the Flight using the Flight ID. Such a deep neural network does not necessarily have to stand by itself, it could also be concatenated with the linear network proposed in this study to make the prediction even more accurate by including passenger data which could otherwise only be included using one hot encoding for all passenger combinations which would not be practical.

Hyperparameter tuning

In order to further increase the accuracy of the overflow prediction model, it is crucial to tune the hyperparameter of the included models. Since the overflow prediction model consists out of 4 sub models which each are trained with a large dataset, it is not practical to train the hyperparameters on a local machine. It is therefore recommended to train the hyperparameters of the overflow prediction model on one of KLMs serves during one or two weekends. Hyperparameter tuning is especially effective in large models, so the expected gains of this step are significant. There are several wrapper methods for hyperparameter tuning, such as grid search with cross validation or random search with cross validation. Since the model is relatively large, it would be recommended to use a wrapper method which is not searching through all parameters in a specified range such as grid search, but rather tests the hyperparameters in an informed manner such as the Bayesian optimization.

Custom loss function

A potential future improvement of the model could be the inclusion of a custom loss function for model training and hyperparameter tuning instead of the current loss function of root-mean-square error. Such a custom loss function could take into account that it is worse for KLM to collect not enough hand luggage, since that is the course for hand luggage related delays, compared to collecting too much hand luggage. It would also be advisable if that custom loss function would, similarly to the root-mean-square error loss function, be able to penalize a higher deviation from the actual value more than a small deviation. This would be important for both the part of the loss function which is applied to too much hand luggage predicted, as well as for the part of the loss function which is applied to not enough hand luggage predicted. For only a small amount of hand luggage is left at the gate to be collected due to the prediction, that is way easier to handle for the staff than if there is still a lot of luggage which has to be collected. Similarly, if it was predicted that a little too much hand luggage has to be collected and all the hand luggage which is predicted has been collected, then it makes a better impression to the customer if only a bit of space is left in the overhead bin compartments compared to when still a lot of space is left.

REFERENCES

[1] Terry Nguyen. The reason carry-on luggage became such a hassle, Dec 2019.

[2] Alex McWhirter. Klm to charge for checked baggage, Feb 2013.

[3] Aditi Shrikant. Why airlines make you check your carry-on bags, Dec 2018.

[4] Suzanne Hiemstra-Van Mastrigt, Richard Ottens, and Peter Vink. Identifying bottlenecks and designing ideas and solutions for improving aircraft passengers' experience during boarding and disembarking. *Applied Ergonomics*, 77:16–21, 2019.

[5] Xander van der Broek. Solving klm's hand luggage problem, 2015.

[6] M.J. Ham. Changing the passenger's 'hand luggage handling' experience and behavior, 2017.

[7] Michael Schultz and Stefan Reitmann. Machine learning approach to predict aircraft boarding. *Transportation Research Part C: Emerging Technologies*, 98:391–408, 2019.

[8] Sveinung Erland, Eitan Bachmat, and Albert Steiner. Let the fast passengers wait: Boarding an airplane takes shorter time when passengers with the most bin luggage enter first. *European Journal of Operational Research*, 2022.

[9] Sheng-Jie Qiang, Bin Jia, Dong-Fan Xie, and Zi-You Gao. Reducing airplane boarding time by accounting for passengers' individual properties: A simulation based on cellular automaton. *Journal of Air Transport Management*, 40(C):42–47, 2014.

[10] Eitan Bachmat. Airplane boarding meets express line queues. *European Journal of Operational Research*, 275(3):1165–1177, 2019.

[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[12] Sukhpreet Singh Dhaliwal, Abdullah-Al Nahid, and Robert Abbas. Effective intrusion detection system using xgboost. *Information*, 9(7), 2018.

[13] Wengang Zhang, Chongzhi Wu, Haiyi Zhong, Yongqin Li, and Lin Wang. Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization. *Geoscience Frontiers*, 12(1):469–477, 2021.

[14] Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29, 2020.

[15] Massimo Aria, Corrado Cuccurullo, and Agostino Gnasso. A comparison among interpretative proposals for random forests. *Machine Learning with Applications*, 6:100094, 2021.

[16] Marius-Constantin Popescu, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8, 07 2009.

[17] Paraskevas Tsangaratos, Andreas Benardos, and Dimitrios Rozos. Use of artificial neural network for spatial rainfall analysis. *Journal of Earth System Science*, 123:457–465, 04 2014.

[18] Deying Li, Faming Huang, Liangxuan Yan, Zhongshan Cao, Jiawu Chen, and Zhou Ye. Landslide susceptibility prediction using particle-swarm-optimized multilayer perceptron: Comparisons with multilayer-perceptron-only, bp neural network, and information value models. *Applied Sciences*, 9(18), 2019.

[19] Wei Chen, Xiaoshen Xie, Jiale Wang, Biswajeet Pradhan, Haoyuan Hong, Dieu Tien Bui, Zhao Duan, and Jianquan Ma. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, 151:147–160, 2017.

[20] Zhuo Chen, Shouyun Liang, Yutian Ke, Zhikun Yang, and Hongliang Zhao. Landslide susceptibility assessment using evidential belief function, certainty factor and frequency ratio model at baxie river basin, nw china. *Geocarto International*, 34:1–20, 11 2017.

[21] EUROCONTROL. Eurocontrol recommended values, 2022.

[22] Andrew Cook and Graham Tanner. European airline delay cost reference values. 2015.

[23] Alison Ledgerwood and Patrick Shrout. The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of personality and social psychology*, 101:1174–88, 08 2011.

[24] xgboost developers. Xgboost parameters. *dmlc XGBoost - XGBoost Parameters - xgboost 1.7.6 documentation*, 2023.

[25] scikit-learn developers. sklearn.ensemble.randomforestregressor, June 2023.

[26] scikit-learn developers. Sklearn.neural _network.mlpregressor, June 2023.

A APPENDIX

A.1 User interface changes

This picture has been removed due to confidentiality

Thank you very much for your understanding

Figure A.1: My-Flight app proposal

This picture has been removed due to confidentiality

Thank you very much for your understanding

Figure A.2: Self-service baggage drop-off machine user interface proposal

A.2 Hyperparameters

Hyperparameter	Description	Possible values	Default
booster	Boosters which can be selected	• gbtree • gblinear • dart	gbtree
eta	Learning rate	Between 0 and 1	0.3
gamma	Minimum loss reduction to partition a leaf node	Between 0 and ∞	0
max_depth	Maximum depth of a tree	Between 0 and ∞	6
min_child_weight	Minimum sum of weight needed in a child instance	Between 0 and ∞	1
subsample	fraction of observations to sampled randomly for each tree	Between 0 and 1	1
colsample_bytree	subsample ratio of columns for each tree	Between 0 and 1	1
colsample_bylevel	subsample ratio of columns for each level	Between 0 and 1	1
colsample_bynode	subsample ratio of columns for each node	Between 0 and 1	1
lambda	L2 regularization term	Between 0 and ∞	1
alpha	L1 regularization term	Between 0 and ∞	0
tree_method	tree construction algorithm	• auto • exact • approx • hist • gpu_hist	auto
objective	Loss function	• squarederror • squaredlogerror • logistic	squarederror (regression with squared loss)
eval_metric	Metric used for validation data	Any performance metric such as MAE, RMSE, etc.	rmse (Root-mean squared-error)

Figure A.3: Performance based hyperparameters of the extreme gradient boost regressor model [24]

Hyperparameter	Description	Possible values	Default
n_estimators	Number of trees in the Random Forest	Between 0 and ∞	100
criterion	quality of split measurement	• squared_error • absolute_error • friedman_mse • poisson	squared_error
max_depth	Maximum depth of each tree	Between 0 and ∞	None
min_samples_split	Minimum number of samples to split a node	Between 0 and ∞	2
min_samples_leaf	Minimum number of data point requirement in a node	Between 0 and ∞	1
max_features	Number of features considered for the split	• Sqrt • Log2 • None • Between 0 and ∞	1
bootstrap	If bootstraped samples are used	• True • False	True
oob_score		• True • False	False
ccp_alpha	Complexity parameter for pruning	Between 0 and ∞	0.0
max_samples	Max data that can be used in each tree	Between 0 and ∞	None

Figure A.4: Performance based hyperparameters of the random forest regressor model [25]

Hyperparameter	Description	Possible values	Default
hidden_layer_sizes	Neurons in the hidden layer	Array with values for the different hidden layers	100
activation	Activation function	• identity • logistic • tanh • relu	relu
solver	Solver for weight optimisation	• lbfgs • sgd • adam	adam
alpha	L2 regularization term	Between 0 and ∞	0.0001
batch_size	Size of minibatches	• auto • Between 0 and ∞	auto
learning_rate	Learning rate	• constant • invscaling • adaptive	constant
learning_rate_init	Initial learning rate	Between 0 and ∞	0.001
tol	Optimization tolerance	Between 0 and ∞	1e-4

Figure A.5: Performance based hyperparameters of the multi-layer perceptron regressor model [26]