Department of Information and Computing Sciences
master Computing Science, Utrecht University

Predicting the amount of excess hand luggage on an aircraft using machine learning

Master's Thesis

Supervisors KLM:
J. van Kempen MSc
Dr. A. Poursaeidesfahani
Supervisors UU:
Dr. ir. J.M. van den Akker
Dr. A.J. Feelders
L.A. Heuseveldt MSc

Author:
S.A.J. Boot
1678973

September 2024
Public Version

Abstract

An excess amount of hand luggage is a growing issue for airlines, causing boarding delays and passenger dissatisfaction. This thesis investigates the factors that cause excess hand luggage and presents a model that predicts the amount of excess hand luggage using machine learning. Various models have been tested on flight data from KLM Royal Dutch Airlines including Linear Regression, Decision Tree Regression, Random Forest Regression and XGBoost. These models have been compared to the heuristic currently in use by KLM.

The results indicate that all Regression models gave a big improvement over the heuristic. The best-performing model is the Random Forest Regressor achieving an R^2 value of 0.83 on the intercontinental flight data. To improve the prediction for gate agents, a variation of the model estimates the amount of hand luggage that should be collected at each collection point. This addition decreased the Mean Absolute Error of the gate collection from 4.58 to 1.65. For KLM this means that, using the models, planning and assignment tasks become easier since it is known how much work needs to be distributed over the gate agents.

Despite these promising results, some challenges remain. The first challenge is the lack of a feedback loop from the cabin which can correct the model. Currently, the true target is the number of hand luggage pieces collected at the outstations or by the check-in agents or gate agents. Another challenge is the model drift. While still performing much better than the heuristic, results decrease in the next year. Future research could explore the use of quantile models to better manage prediction uncertainty.

Contents

List of Terms

Bax Baggage. Bax is a commonly used abbreviation for (multiple pieces of) luggage used in the travel industry.

Cabin The space in the aircraft where passengers sit.

EUR EURopean flights. These flights are done using narrow-body aircraft with one aisle. These flights do not necessarily always have a destination that lies in Europe, the Schengen zone, or the European Union.

Flight leg A Flight leg is the journey of an aircraft from one airport to another without any stops. A flight can consist of multiple flight legs ($A \rightarrow B, B \rightarrow C$).

ICA InterContinentAl flights. These flights are done using wide-body aircraft with two aisles. ICA flights do not necessarily have to go to another continent. As long as the flight is done using a wide-body aircraft, it is an ICA flight.

KLC KLM Cityhopper. A daughter company of KLM which flies with smaller Embraer aircraft (Embraer 175, 190, 195-E2).

KLM KLM Royal Dutch Airlines. The oldest airline in the world which has its hub at Amsterdam Airport Schiphol. KLM is part of the Air France-KLM group and a member of the SkyTeam airline alliance.

Outstation In the context of this thesis, an outstation is an airport of a passenger's previous leg. If this thesis talks about flight $B \rightarrow C$ and passenger travel from $A \rightarrow B \rightarrow C$, airport A is an outstation.

Passenger leg A flight leg of one passenger. If a passenger flies from A to C via B, it has two pax legs: $A \rightarrow B$ and $B \rightarrow C$.

Pax Passengers. Abbreviation of passengers used in the travel industry.

Transfer A transfer passenger is a passenger who traveled by plane to the departure airport and did not pass check-in.

Transit Transit passengers are passengers who stay on the same aircraft for multiple flight legs. For flight leg $B \rightarrow C$, a passenger who travels on the same aircraft from $A \rightarrow B \rightarrow C$, is a transit passenger.

Part I

Background & Methodology

Chapter 1

Introduction

Hand luggage is an increasing problem for airlines. Each year, more passengers carry a large amount of carry-on luggage. Even if all passengers adhere to the rules about the size and the quantity of hand luggage, there is a surplus. Each aircraft has a certain amount of space for hand luggage in the overhead bins, but sometimes there is not enough space. Even the Airbus A330 does not fit all the hand luggage of a flight that is fully booked with US passengers [3]. This means that airlines need to find ways to prevent such a surplus. One way many airlines are solving this issue is by collecting hand luggage, mostly trolleys, from passengers and putting them in the hold of the aircraft.

With all the overflows, airlines have created a procedure for collecting hand luggage. For example, Transavia let people board until they have counted 70 trolleys. After the limit is reached, Transavia starts collecting all the trolleys from the passengers still in the queue. [14].

KLM Royal Dutch Airlines has another procedure: before the boarding starts, they look through the queue and estimate whether they need to collect hand luggage, and how much. KLM Royal Dutch Airlines starts collecting from the passengers with the cheapest tickets. This procedure does not always work flawlessly, there may still be too much hand luggage on the aircraft. When there is a hand luggage surplus, and the gate agents do not collect enough hand luggage at the gate, gate agents need to go into the plane to label the excess trolleys and move them to the hold. While the gate agents are picking up the excess trolleys, communication is done over a walkie-talkie with a gate agent still at the gate to link the luggage to the right passenger. When the excess trolleys are picked, they need to be moved outside to the ground platform to be moved into the hold of the aircraft. These manual retrievals almost always cause the aircraft to have a delayed departure. That is why KLM Royal Dutch Airlines wants to collect the hand luggage as early as possible during the process. This research wants to create a model that can indicate if hand luggage needs to be collected to prevent an overflow, and how much hand luggage needs to be collected.

KLM Royal Dutch Airlines, henceforth called KLM, wants to improve its hand luggage estimation because looking through the queue does not always give an accurate estimation: People may arrive later, or a queue might be crowded in which case the gate agents can not scan through the queue. The desire is to have a prediction before the boarding starts. This prediction should tell the gate agents how many hand luggage pieces they need to collect to prevent a surplus inside the aircraft. Using this prediction, KLM wants to prevent an overflow in the aircraft itself.

Another use case for the prediction is to use the model in a flow network. If flight $B \to C$ is expected to be very busy, KLM wants to collect the hand luggage as early as possible. Passengers on flight $B \to C$ may come from another KLM flight ($A \to B$). KLM ultimately wants to already collect the hand luggage at station A to prevent delays at station B. The current situation sometimes causes chaos at station B because of the large number of passengers that have to board in a short amount of time. To be able to create this flow network, an accurate prediction model is needed for intercontinental and KLC flights.

Chapter 2

Literature Overview

2.1 Hand Luggage overflow prevention

The aviation industry has a lot of published research. Many papers have been written about the best boarding procedure to minimize boarding time for different amounts of hand luggage per passenger [2], [8], [11], [13]. Less research has been done into (the prevention of) hand luggage overflow.

At KLM, two studies have been done into (the amount of excess) hand luggage. In 2015, Xander van der Broek studied the hand luggage problem at KLM [3]. His research dived into the causes of the hand luggage surplus together with possible solutions.

Van der Broek surveyed 1150 people to analyze the trolley factor, the number of passengers that bring a trolley compared to the total number of passengers on a flight, based on different factors. In the survey, he asked people about the duration of their stay, the goal of the trip, and whether they had hand luggage or not. The survey showed that nationality, duration of stay, the goal of the trip, and whether the person checked baggage before were important factors. The exploratory survey in which the hand luggage of 150 passengers was weighed also showed that 17% of the respondents' hand luggage pieces exceeded the allowances and could have been checked.

In his study, Van der Broek focussed on European flights. He observed that the KLM Cityhopper aircraft (Embraer planes) had another boarding procedure which did not delay the aircraft by much. When passengers of such a cityhopper were asked to check their hand luggage, they could carry their trolley to the aircraft and leave it at the bottom of the stairs. After landing, they could retrieve their hand luggage immediately after disembarking the aircraft. For the passengers, this was a big advantage because this meant that they did not have to wait for their hand luggage at the baggage belt.

One of the prediction solutions Van der Broek introduced is a heuristic for calculating the hand luggage surplus. His report stops at the introduction of the equation, so no further information about the equation is available. Within KLM, some variant of the heuristic is being used for predicting the number of hand luggage pieces that need to be collected on intercontinental flights. The heuristic in the report shows the following structure:

$$HLsurplus = \#bookings * trolleyfactor + margin - airplanecapacity \tag{2.1}$$

Vico Heinrich continued the research of Van der Broek [6]. His research focussed on creating a machine learning model for estimating the amount of hand luggage on the EUR flights (Boeing 737 aircraft).

In the exploration of the data, Heinrich found that the following features had a high (> 0.3 absolute Pearson correlation coefficient) correlation with the target: TotalPax (passengers), TotalBax (bags), TotalLocalPax (not transfer passengers), scheduledFlightProcessDuration, and whether it is a departure from Southern Europe.

After the data research, Heinrich compared different types of models. After testing all models, he found that the random forest regressor, MLP, and XGBoost models gave the highest R^2 values. Other models with an R^2 value above 0.60 were: Linear Regression, SVR, Ridge regression, GBR, Bagging, and SGD. For all tested models, Heinrich used the default settings since there was not enough time to do hyperparameter tuning. In his final model, Heinrich combined the 3 best-performing models into an ensemble. After testing, the weighted-averaged voting regressor gave the highest R^2 value.

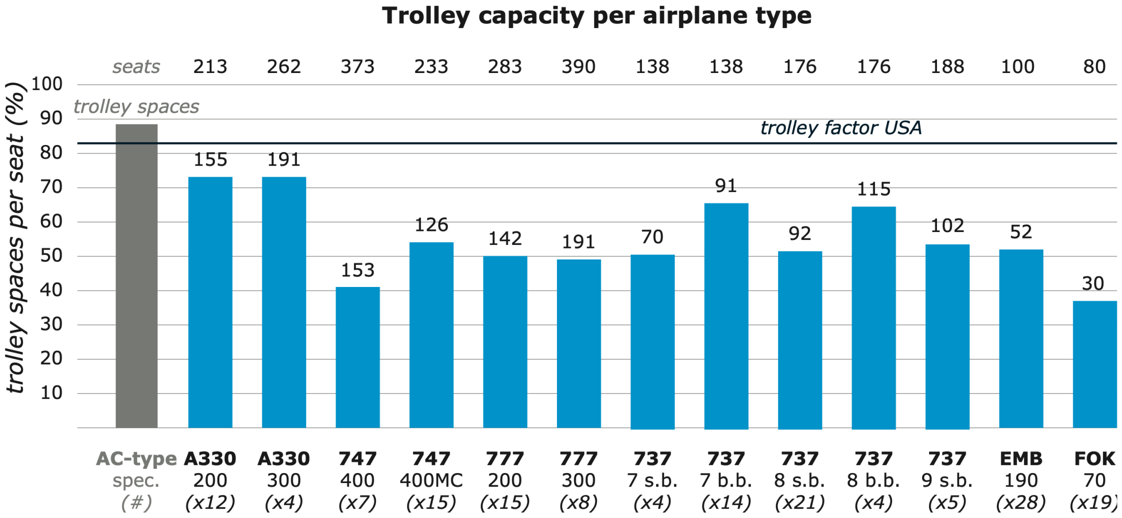
A public survey on the hand luggage problem asked respondents what they thought about different solutions [12]. The survey showed that 56% of the respondents traveled with only carry-on luggage. Research showed that stowing hand luggage in the overhead bins is the most time-consuming process when boarding [5] When the respondents were asked what their opinion was for an automated system onboard that would automatically store their carry-on luggage somewhere, 83% gave an affirmative response. On the other hand, only 51% accepted the consequence that, in the case of an automated system, there would be limited access to their stored items. This tells us that passengers are willing to hand in their hand luggage, as long as they are not picked individually and other passengers are not.

Milne and Kelly did research into new methods to board the aircraft. In their conclusions, they mentioned the move within the airline industry towards charging fees for carry-on luggage. Allegiant Air charges a lower fee for carry-on luggage when paid online prior to arrival at the airport [7]. This stimulates passengers to register their hand luggage before their arrival at the airport. Using this method, airlines can get more accurate information about the amount of carry-on luggage on a flight. They also state: "If an airline does not want to use fees for this purpose, the airline could estimate carry-on luggage from other known information. For example, a passenger departing Monday morning and returning the same day/night can be expected to have less carry-on luggage than a passenger departing that same Monday morning who is returning four days later without checking a bag.”

2.2 Aircraft types

One of the factors Heinrich analyzed is the type of aircraft. Van der Broek collected the trolley capacity for all aircraft types that KLM uses. This information is shown in Figure 2.1 where the percentage of trolley spaces per seat is shown. The black line indicates the average trolley factor for the USA [6]. The number above each bar shows the exact number of trolley spaces that that type of aircraft can store. The figure shows that an Airbus A330-300 has a higher relative trolley capacity than a Boeing 777-300 while both have space for 191 trolleys. (Figure 2.1). The figure shows that none of the aircraft types used by KLM have enough space to store all the trolleys brought by passengers, assuming that the trolley factor on a KLM flight is the same as the trolley factor in the USA.

Figure 2.1: Trolley capacity per aircraft type.



Chapter 3

Thesis outline

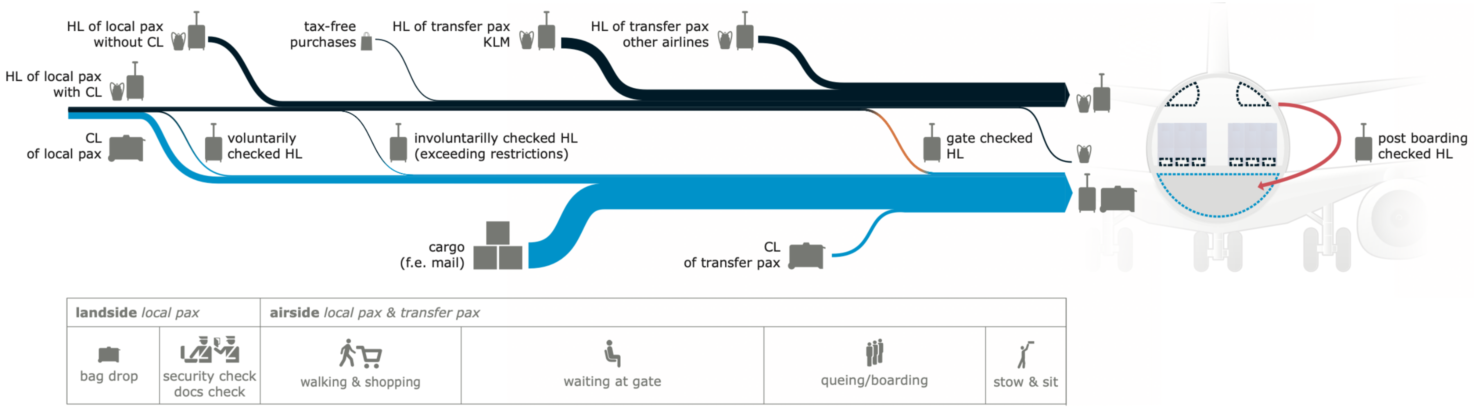
3.1 Problem description

The problem of estimating the amount of hand luggage that needs to be collected can be split up into several subproblems. One of the big questions is what we want to estimate. The overall problem is that a flight could have a hand luggage overflow.

At several points in the flight process, the hand luggage can be collected. However, 70% of the passengers on a flight ($B \rightarrow C$) are transfer passengers ($A \rightarrow C$) who do not pass the check-in at the departure airport of their current flightleg (airport B). Those transfer passengers could transfer from another airline such as Delta or AirFrance. In figure 3.1 from Van der Broek these different flows of hand luggage can be observed.

KLM wants to solve the hand luggage problem by collecting as early as possible. If the model only predicts when the passengers are already waiting at the gate, it is too late for KLM. The model should take into account how much hand luggage is collected at the check-in and the outstations. If in an early stage it is known that there is a great surplus of hand luggage, KLM can decide to collect more at the check-in stage by blocking boarding passes.

Figure 3.1: The flow of different types of Hand Luggage (HL), Checked Luggage (CL) and other types of luggage. Image from Van der Broek [3].



3.2 Research question

Combining all these problems, the following research question arises:

> How can a predictive model be developed and optimized, to accurately forecast the total number of hand luggage pieces that have to be collected for KLM intercontinental and KLM Cityhopper flights?

The prediction tool should be able to predict the number of hand luggage pieces that need to be collected given different criteria. In the first place, the model should be able to estimate the number of hand luggage pieces that have to be collected for days without big disruptions. When the model has a good performance for 'normal' days, the prediction for flights with disruptions will be studied. The research question can be split out into the following questions:

1. What is the target entity for this problem?
2. How can a prediction model solve this problem?
   - (a) What specific variables (both available and derivable) are most predictive of the number of hand luggage pieces collected for KLM flights departing from Schiphol Airport?
   - (b) What specific variables (both available and derivable) are most predictive of the number of hand luggage pieces collected for KLM flights arriving at Schiphol Airport?
   - (c) Among various machine learning models, which model minimizes prediction error, especially in terms of avoiding large prediction errors?
3. How does the collection rate (how much hand luggage is being collected) at outstations, affect the predictive accuracy for flights with more than 70% transfer passengers?
4. How can the model be made more robust to handle special events such as the Jewish Passover feast, or the Chinese New Year?
5. What is the effect of using the observed hand luggage collection at the check-in and the outstations as a feature in the model on the accuracy at the gate?

Part II

Data Analysis

Chapter 4

Current heuristic model

4.1 Functionality

Currently, KLM uses a simple model to estimate the total number of excess hand luggage pieces. This model is based on Van der Broek's proposed calculator [3]. The model is managed by someone from KLM who tunes the model by hand based on feedback from the gate agents. The model works like a linear regression model in the sense that features get multiplied by a certain factor to calculate the total amount of excess hand luggage [9]. The following features control the models output:

Aircraft type & Route category The first feature is a correction for each aircraft type for all flightlegs. Per city pair (pair of 2 airports), each aircraft type has an indication how much hand luggage should be collected as a basis.

Number of Skyprio pax Per x passengers, one additional excess hand luggage item is added to the prediction.

Point of Sale correction (nationality) Per x passengers, with point of sale in certain countries, add one additional item to the amount of hand luggage pieces that should be collected. For other selected countries, decrease the prediction by one for every x passengers that buy a ticket from one of the selected countries.

Season correction (winter coats) In the winter period, collect one hand luggage piece more for every x passengers.

Amount of checked in hold baggage Calculate the checked in bag factor: $\frac{n_{checkedbaggage}}{n_{acceptedpassengers}}$. A lower factor means that fewer bags are checked which means more hand luggage can be expected. A correction factor gets multiplied by the number of accepted passengers.

Correction for open seats For every x open seats, decrease the prediction by 1.

Restrict the maximum prediction [redacted]

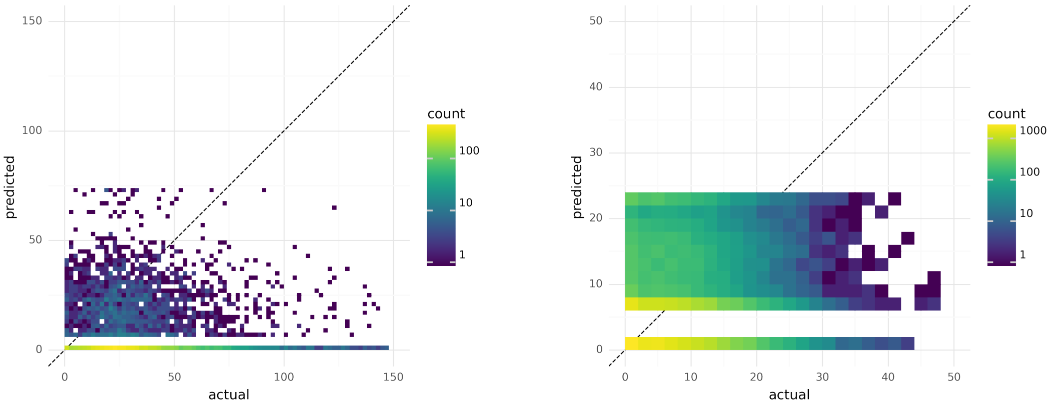
4.2 Performance

To measure the performance of the current model, the data from 2023 is used. The predicted total amount of excess hand luggage has been compared with the total number of collected hand luggage pieces. Table 4.1 and Figure 4.1 show the results of the current model. [redacted] The table shows a negative R^2 for both ICA (-1.63) and KLC (-1.17). Section 11.4 will further discuss the negative R^2 scores. The mean absolute error for the current model is ≈ 20 for ICA flights and ≈ 6 for KLC flights. This means that, on average, the model has a prediction error of 31 for ICA flights. When splitting the MAE into a positive and a negative error, the analysis shows that -when the error is positive- ≈ 23 hand luggage pieces get collected more than the predicted number for ICA flights. When the error is negative, on average ≈ 19 less hand luggage pieces get collected than predicted. For KLC, these numbers are ≈ 7 when underestimating and ≈ 5 when overestimating. The average error for each aircraft is around 10% of the trolley capacity (Table 4.2). For the Embraer 195-E2 (295), the model shows an underestimation of about 19%.

Table 4.1: This table shows the results of the current model for ICA and KLC flights. The results are from flight data from 2023.

	ICA	KLC
R^2	-1.63	-1.17
MAE	20.28	6.09
RMSE	27.72	8.22

Figure 4.1: Bin2d plots which show the prediction results of the current model for both ICA and KLC flights in 2023. The count tells how many flights have the score of that square.



(a) Predictions for ICA flights.

(b) Predictions for KLC flights.

Table 4.2: This table shows the average over- and underestimation of the current heuristic model. The percentages show the MAE in relation to the trolley capacity of the aircraft type.

	Aircraft type	Overestimated		Underestimated	
		MAE	%	MAE	%
ICA	Boeing 777-200ER	21.25	[redacted]%	25.61	[redacted]%
	Boeing 777-300ER	24.61	[redacted]%	34.11	[redacted]%
	Boeing 787-9	14.51	[redacted]%	16.78	[redacted]%
	Boeing 787-10	15.29	[redacted]%	19.28	[redacted]%
	Airbus A330-200	19.00	[redacted]%	18.54	[redacted]%
	Airbus A330-300	15.86	[redacted]%	18.12	[redacted]%
KLC	Embraer 175	6.04	[redacted]%	4.88	[redacted]%
	Embraer 190	4.86	[redacted]%	6.53	[redacted]%
	Embraer 195-E2	3.29	[redacted]%	10.33	[redacted]%

Chapter 5

The KLM data

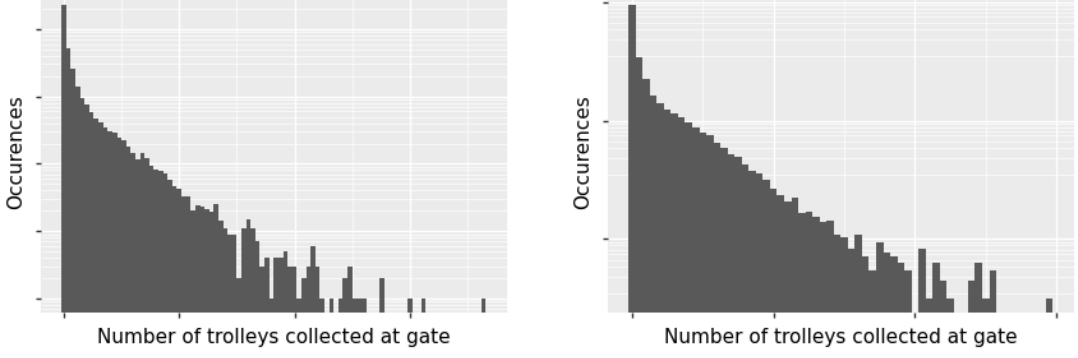
For this research, flight data from KLM will be used to train and test the models. For each flight, KLM stores data about the flightleg, the passengers, and the baggage. The level of observation is a flight leg. For each flightleg, a prediction will be made. The model will be trained and tested on 2023 data, since 2020-2022 is not representative due to the coronavirus, and KLM only started registering the hand luggage collection in the second half of 2019. Additionally, data up until May 2024 will be used to test how the models perform in future years. From the 2023 data, a random split will be made for the training (70%) and test (30%) set.

5.1 Censored data

The 2023 dataset stores all hand luggage that is collected at the outstations, check-in, and the gate. What is not recorded, is the number of empty trolley spaces in the aircraft bins. The stored data is count data, meaning that all variables are ≥ 0. This means that only data is available when hand luggage is collected. When no hand luggage is collected, it is unknown if it is a perfect fit or if there is still space left in the aircraft. Figure 5.1 shows the data from 2023 for the number of trolleys collected at the gate. This figure shows that in most cases 0, or only a few hand luggage pieces were collected.

When no collected hand luggage pieces are registered, gate agents don't know how much space there is left in the overhead bins of the aircraft. The gate agents check the passengers while boarding and will not know the situation inside the plane. The gate agents only will get feedback if they do not collect enough hand luggage. When they collect too few pieces, the boarding process is delayed because hand luggage pieces have to be moved from the plane to the aircraft's hold. On the other hand, if the gate agents collected too much, they would not know. It could be the case that there is space left for 10 more trolleys, but that information will not reach the gate agents. This means that even without a prediction for the gate agents, the data is censored.

Figure 5.1: Histogram showing the number of hand luggage pieces collected at the gate for all intercontinental and KLC flights in 2023.



(a) Plot for ICA flights.

(b) Plot for KLC flights.

To formalize this, let y^* denote the number of pieces that should be collected to get a perfect fit of the hand luggage in the aircraft where no space is left in the overhead bins. The value of y^* is not always observed and is therefore sometimes called a latent variable. Let y_c denote the number of pieces of hand luggage that have been collected at or before the gate. We observe y^* whenever $y^* > y_c$ because then there won't be enough room in the plane and additional hand luggage has to be collected from the cabin. On the other hand, if $y^* \leq y_c$, then we observe y_c, because there might still be room in the aircraft, but this is not noted. This has consequences for the model since the amount of room left is not known. If the censoring is not taken care of, the assumption has to be made that the amount of space left, is always 0.

All in all, we can summarize the situation as follows. Let y denote the observed number of pieces of hand luggage collected. Then we have

$$y = \begin{cases} y^* & \text{if } y^* > y_c \\ y_c & \text{if } y^* \leq y_c \end{cases} \tag{5.1}$$

5.2 Flightleg data

A flight leg is a flight from one airport to another airport, with no stops in between. Flight $A \rightarrow B \rightarrow C$ will result in 2 flight legs: $A \rightarrow B$ and $B \rightarrow C$.

5.2.1 Correlated features

The first analysis that is done is Pearson's and Spearman's correlation. The results (Figure A.1) show that for ICA flights arriving at Schiphol, the percentage of local (departing) passengers has a negative correlation of ≈ -0.5 with the number of hand luggage pieces collected at outstations. This means that on flight $B \rightarrow C$ more local passengers (departing from B) will result in less hand luggage collected at outstations (A). This is expected because more local passengers means that there are fewer transfer passengers from whom hand luggage can be collected.

A correlation of ≈ 0.75 can be observed between the percentage of transit passengers and the amount of hand luggage collected at outstations (Figure A.1). This is an observation that can be explained by the fact that transit passengers are passengers who continue their flight on the same flight number. This is the case for some destinations where KLM flies from $A \rightarrow B \rightarrow C$ with the same aircraft. Passengers who fly from A to C, are transit passengers on flight $B \rightarrow C$. This means that hand luggage that is collected at station A from passengers who fly to airport C, is registered as hand luggage collected at outstations for flight $B \rightarrow C$.

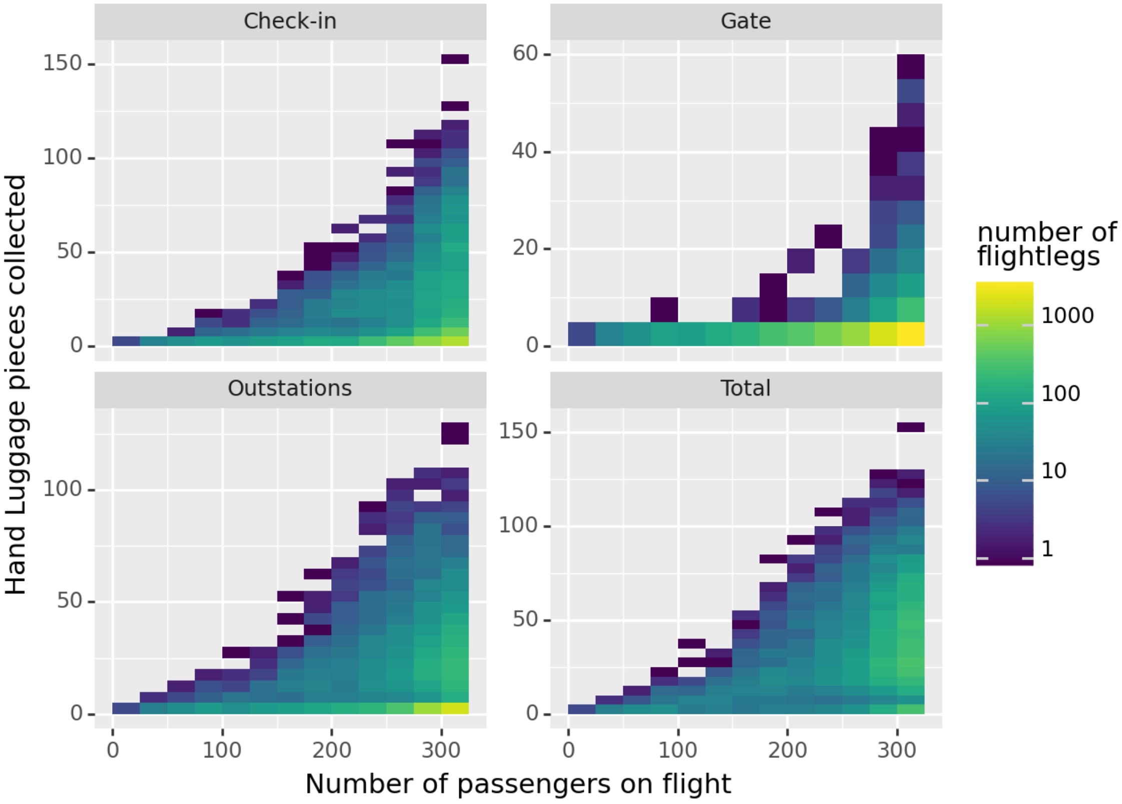
For intercontinental flights departing from Schiphol, two expected correlations can be observed: a correlation of ≈ 0.6 between the percentage of local passengers and the amount of hand luggage collected at check-in, and a negative correlation of ≈ -0.6 between the percentage of transfer passengers and the amount of hand luggage collected at check-in. This is explained by the fact that at check-in, hand luggage can only be collected from local passengers. The same correlation can be observed between the local/transfer percentage and the amount of hand luggage collected at outstations.

Correlation matrices A.4 and A.3 show that there are no features with significant correlated besides the negative correlation between the percentage of local and the percentage of transfer passengers. The distribution of passengers is spread between local, transfer, and transit passengers. This large negative correlation is mostly due to the lack of transit passengers which results in the passengers being local or transfer.

5.2.2 Number of passengers on board

It feels intuitive that there should be a relation between the number of passengers on board and the amount of hand luggage because each (intercontinental) passenger has a free trolley included. The maximum amount of luggage that passengers can bring depends on the destination. For some destinations, passengers can bring 2 pieces of check-in luggage and 1 piece of hand luggage. For other destinations, passengers can only bring 1 piece of check-in luggage together with 1 piece of hand luggage. When looking at the hand luggage collection, a flat line at $y = 0$ would be expected when plotted against the number of passengers on board for flightlegs that are less than ($< 80\%$) full. The data contradicts this. The lower right plot in Figure 5.2 corner shows that hand luggage collection is already observable from $x = 0$. This phenomenon can be explained by the check-in (upper left) and outstations (lower left) plots. These two plots show that there is a correlation between the number of passengers and the number of hand luggage pieces collected. The upper right plot shows minimal hand luggage collection when there are less than 275 passengers on board. This proves our hypothesis that there is little hand luggage collection at the gate when a flightleg is not full. This analysis increases the complexity of the research since KLM wants to reduce the gate collection by collecting at the check-in or the outstations, but still wants to know how much hand luggage to collect. If there were no gate collection -and there is voluntary collected hand luggage at the check-in and the outstations- the model would be trained to start collecting hand luggage. The model does not know if hand luggage is collected voluntary or involuntary. This means that the model can predict to collect hand luggage even when it may not be necessary.

Figure 5.2: Number of hand luggage pieces collected for Boeing 777-200ER in 2023.



5.3 Passenger data

5.3.1 Airport behaviour

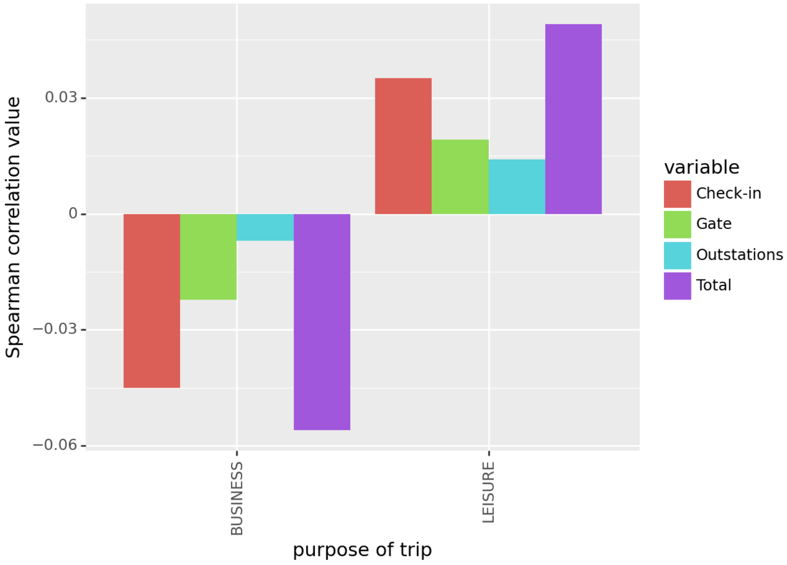
There is a big correlation visible between the amount of hand luggage collected and the arrival/departure airport. When training a simple decision tree, the arrival and departure airports were the most important features. For KLM this is unwanted behaviour: these features should not be the *most* important features since KLM could decide to fly to a new destination. The flights to these new airports need a prediction too.

Besides that, the influence was so big that other features such as the number of passengers on board were not used for the prediction. This is unwanted behaviour because automatically predicting an excess hand luggage amount of 200 when flying to destinations for which a lot of hand luggage gets collected is unwanted. Otherwise, if a flight is half empty, or when the passenger composition is different, the model would still base its estimation on which airport the aircraft is departing/arriving from. The resulting model would not be robust.

5.3.2 Journey purpose

The KLM dataset contains the journey purpose of passengers on their journey. For 2023 ICA flights, [redacted] are business flights, [redacted] are leisure flights, and for [redacted] the purpose of the journey is not registered. Pearson and Spearman correlation calculations have been run to measure the performance of these potential features. Figure 5.3 shows that there is no correlation between the purpose of the journey and the amount of hand luggage collected on ICA flights. This result contradicts the results from [3] which states that the goal of the trip is one of the four big factors. This might be due to the limited amount of journey purposes. The data only accounts for 2 types: leisure and business. Van der Broek held a survey which might have led to more possible answers to the question about the goal of the trip.

Figure 5.3: The figure shows the Spearman correlation results between the number of pax with journey purpose x and the number of hand luggage pieces collected. The data used is the flight data in 2023 for ICA flights.



5.3.3 Passenger nationality

Another feature that KLM stores about their passengers is the nationality of the passenger. This might be an interesting feature since gate agents state that passengers from some countries bring more (hand) luggage with them than passengers from other countries. This information is only useful with the assumption that all passengers from all nationalities have the same chance to be selected for handing in their hand luggage. The Spearman correlation results in Figures A.9 and A.10 show that there is a correlation between the nationality of the passenger and the amount of hand luggage collected for some countries.

The computed correlations show that there is a big difference between flights departing and arriving at Schiphol International Airport. For flights departing from AMS, all four different types of collected hand luggage appear in the graphs (Figures A.9a and A.10a). In Figure A.9a, the country with the most correlation with the check-in hand luggage collection is [redacted]. This can be explained by the fact that 70% of KLM passengers are transfer passengers, which means that they don't leave the airport. As a result, most passengers passing check-in are passengers who start their journey at that airport. These passengers will mostly be citizens or tourists.

In Figure A.10a, the highest bar shows a correlation of 0.20 between the number of [redacted] passengers and the number of hand luggage collected at the gate. This is a correlation gate agents agree with. The gate agents often see [redacted] passengers coming from a flight with the company [redacted] and continuing their journey with KLM. [redacted] has different hand luggage allowances than KLM. Passengers flying with [redacted] can bring bigger hand luggage trolleys that fit on their aircraft, but not on KLM aircraft. Coming from another flight means that they only pass the gate before embarking on their next flight leg. This results in the gate as the only option to collect their oversized hand luggage.

The flights arriving at Schiphol show more countries with a correlation with the amount of hand luggage collected at the gate (Figure A.9b)

5.4 What is the target?

A problem that needs to be solved is choosing the right target for our model. As analyzed in the previous sections, multiple variables are related to how much hand luggage is collected. However, for each of those variables, there are advantages and disadvantages to setting that variable as a target.

total This is the total amount of hand luggage pieces collected during the entire process for a flight. It includes pieces collected at the check-in and at the gate, but also pieces collected at outstations. Outstations are the airports where transfer passengers come from before boarding the current flight.

gate When the gate agents expect a lot of hand luggage pieces, they will collect hand luggage at the gate. The total number of pieces collected at the gate is recorded in this variable.

Intuitively the gate-variable would be picked: We want to predict whether hand luggage needs to be collected at the gate, or not. However, this does not take into account that hand luggage is already collected at the check-in and the outstations. If a lot of hand luggage is collected at outstations, the gate agents should have a lower prediction.

On the other hand, if we use the total number of HL pieces collected, other errors occur. The advantage of predicting the total amount of HL that needs to be collected in the entire process is that no other variables influence the predicted value. The disadvantage is that training on the total amount of collected pieces can create expectations from outstations. Let's use an example to explain this problem: All transfer passengers on our current flight ($B \rightarrow C$) come from many European flights ($A \rightarrow B$). For KLM, a European flight is a flight that is flown with a narrow-body aircraft (single aisle) with often limited space for hand luggage. This causes a lot of hand luggage collection on flight $A \rightarrow B$. The total amount of hand luggage collected *at outstations* for flight $B \rightarrow C$ is high because of flight $A \rightarrow B$. This causes the *total* amount collected to increase as well. The problem is that our target now is a big value while all those hand luggage pieces easily could have fitted in the aircraft on flight $B \rightarrow C$. Because of the count data, the model does not know that there is still plenty of space left in the aircraft on flight $B \rightarrow C$.

This is the reason that finding the right target is one of the subproblems.

For this research, the decision is made to predict the *total* number of hand luggage pieces that need to be collected for a given flight. The reason for this is that KLM eventually wants to stop to collect hand luggage at the gate. Collecting hand luggage at the gate often makes passengers uncomfortable because they do not want to be the person from whom hand luggage is taken. In the future, KLM wants to only collect hand luggage that exceeds restrictions at the gate. Currently, one of the strategies is by blocking the boarding passes as earlier explained. For these strategies, KLM still needs to know how much hand luggage to expect and how many boarding passes need to be blocked. That is why this research will focus on predicting the total amount of hand luggage that needs to be collected. The model should be able to estimate the right amount of hand luggage that should be collected for a flight without leaving too much empty space in the overhead bins.

The decision to predict the total amount of hand luggage still leaves the outstation's hand luggage problem. In this research, the effect of outstations will be studied when predicting how much hand luggage needs to be collected.

Chapter 6

Other datasets

This chapter discusses data from other sources than KLM that are used in this research.

6.1 Airports, Countries and Continents

When a classification tree was trained on the 2023 data, the first split was on the (one-hot) boolean feature that the flight departed from some airport. Not only was the top feature split on the airport, but all splits (except the bottom one or two) were on departure and arrival airports. This showed the importance of the origin and destination airports. This relation can be observed in figure B.1, where for each (IATA) airport code a box plot is plotted for the number of collected hand luggage pieces.

However, for KLM it is not desired to use the departure or arrival airport because it is -while being correlated- not the cause for the amount of collected hand luggage. A model in which (nearly) all splits are done on the departure or arrival airport is not robust to adding new destinations. A solution to this problem would be to use the country or the continent of the airport as a feature. Using the country as a feature was also not a good option because most countries only have one airport that is a KLM destination. This leaves the continent as the other option. The flight data does not store the country or continent of airports, so a dataset is used to perform this conversion. The dataset is from 2019 and was created by Andrian Zinovei [15]. In figure B.2, the number of hand luggage pieces collected for each continent is plotted. It can be observed that there are differences in hand luggage collection between continents. The continent feature will be used instead of the country or the airport features.

6.2 National Holidays

When talking to the gate agents, one of the common things heard was that national holidays have a large impact on the amount of excess hand luggage. The python package national holidays [1] was used to retrieve a dataset of the worldwide national holidays. The following features have been created:

- days until a national holiday at departure country
- days after a national holiday at departure country
- boolean value whether the current date is a national holiday at the departure country
- days until a national holiday at arrival country
- days after a national holiday at arrival country
- boolean value whether the current date is a national holiday at the arrival country

These six features have been tested for correlation with the target variable. Figure B.3 show a positive correlation of ≈ 0.2 between the number of days since a holiday at the departure country and the number of collected hand luggage pieces at outstations. This means that more days since the last holiday means that more hand luggage gets collected at the outstations. However, the same is true for the days until a holiday and the amount of hand luggage collected at outstations which shows a Spearman correlation score of ≈ 0.3 for ICA and ≈ 0.2 for KLC. This could mean that, at the outstations, the most hand luggage gets collected between two holidays.

6.3 Dutch School Holidays

When speaking to gate agents, they often see the school holidays as a busy period. Not only does KLM perform the most flights during the school holidays, but it is also the period that a lot of hand luggage gets collected. The company has the option to block boarding passes for some tickets if a hand luggage overflow is expected, to force these passengers to check in their hand luggage. This option is used during the busy summer times. To analyze the possible correlation between the Dutch school holidays and the hand luggage collection rate, an additional library is used to retrieve these school holidays. Felix Claessen contributed to the workalender Python package by adding the Dutch school holidays [4]. To test the correlation, the same features were tested as for the national holidays in the previous section (Section 6.2). The highest correlation is a Spearman correlation coefficient of ≈ 0.3 between the boolean variable whether the current date is a holiday at the departure country and the amount of hand luggage collected at outstations (Figure B.4a). A smaller Spearman correlation coefficient of ≈ 0.1 can be seen for the total amount of collected hand luggage. Figure B.4b shows a positive Spearman correlation value of ≈ 0.2 for the number of collected hand luggage pieces when the arrival country (The Netherlands) currently has a school holiday. Surprisingly, no correlation is discovered for the amount of hand luggage collected at the check-in when the departure country (The Netherlands) has school holidays.

Part III

Solution Approach & Implementation

Chapter 7

Proposed models

The hand luggage estimation problem is a supervised regression problem with censored data. The output of the model should be the number of hand luggage pieces that need to be collected in the entire process (check-in + gate + outstations). If the predicted number of excess hand luggage pieces is too low, the gate agents need to collect additional hand luggage pieces (otherwise, the aircraft can't depart). These corrections can be used to retrain the models.

7.1 Single output models

The first types of models that will be tested are the regression models that will estimate the total number of hand luggage pieces that need to be collected. For the experiment, all models will estimate the total amount of hand luggage that needs to be collected for a given flight leg. The following models were chosen to be used in the experiment:

Linear Regression A basic model that attempts to fit a linear relationship between the input variables and the target variable.

Decision Tree Regressor A single tree-like model that splits the set by making decisions using the features. The expectation is that this model will perform well since similar flights will have similar hand luggage collections. Flights with a large amount of excess hand luggage are easily separated from flights with no hand luggage overflow.

Random Forest Regressor A random forest model that will generate several trees.

XG Boost Regressor A different type of ensemble machine learning using Decision Tree Regressors. Trees are added one by one and are trained to correct the prediction errors made by prior models.

Multilayer Perceptron Regressor A feedforward neural network with a single output node.

In order to be able to calculate the performance at the gate, the gate predictions can be simulated by subtracting the hand luggage collected at the check-in and outstations from the predicted total collected hand luggage. This is calculated as shown in Equation 7.1. For calculating the $y_{check-in}$ and $y_{outstations}$, the following assumptions are made:

In order to be able to calculate the estimated number of excess hand luggage pieces, two assumptions have to be made. The first assumption is that the $\hat{y}_{total}$ perfectly estimates the number of excess hand luggage pieces, ensuring a perfect fit within the aircraft. A perfect fit means that all the trolley storage bins in the aircraft are filled with no excess hand luggage and no empty bins. The second assumption is that the total number of collected hand luggage pieces collected—whether at check-in, the gate, or outstations—equals the estimated number of excess hand luggage pieces. Based on these assumptions, the estimated number of excess hand luggage pieces at the check-in and outstations can be calculated as shown in Equation 7.1.

$$\begin{aligned} \hat{y}_{total} &= \hat{y} \\ \hat{y}_{check-in} &= \hat{y} - y_{gate} - y_{outstations} \\ \hat{y}_{gate} &= \hat{y} - y_{check-in} - y_{outstations} \\ \hat{y}_{outstations} &= \hat{y} - y_{check-in} - y_{gate} \end{aligned} \tag{7.1}$$

7.2 Multi-output models

The single output models only estimate the total amount of hand luggage that needs to be collected. These models do not take into consideration how much hand luggage is collected during the different collection moments. For the planning of the personnel, KLM needs to have an estimate of how much work the check-in agents and the gate agents have to do. Since the single-output models only predict the total, predictions of the check-in and gate collections need to be made. Predicting how much hand luggage needs to be collected at check-in and the gate is not possible since these numbers are not known before the flight process is started. During the flight process, calculating how much hand luggage pieces need to be collected at the gate can be calculated as shown in Equation 7.1. However, the $\hat{y}_{gate}$ can only be calculated after all passengers passed the check-in or departed from the outstations. For planning reasons, KLM wants these predictions earlier. Although KLM wants to bring the gate collection to zero in the future, the amount of hand luggage passengers bring to the gate can not (easily) be controlled for passengers who don't pass check-in. Up until three hours before departure, the planning of passenger services can change how many gate agents are assigned to a flight.

In order to be able to make more precise predictions for the check-in and gate, the multi-output models have been created. These models do not estimate the total amount of excess hand luggage, but they estimate how much hand luggage will be collected at the outstations, check-in, and the gate. The models will be trained individually on each collection process to learn the patterns for each collection point. The total amount of excess hand luggage can be calculated by computing the sum of these three predictions. Equation 7.2 shows how the estimations for all collection points are calculated.

$$\begin{aligned} \hat{y}_{total} &= \hat{y}_{check-in} + \hat{y}_{gate} + \hat{y}_{outstations} \\ \hat{y}_{check-in} &= \hat{y}_{check-in} \\ \hat{y}_{gate} &= \hat{y}_{gate} \\ \hat{y}_{outstations} &= \hat{y}_{outstations} \end{aligned} \tag{7.2}$$

These multi-output models are created using the `MultiOutputRegressor` function from the sk-learn package. This function creates a model (linear / decision tree / random forest / XG Boost / MLPR) for each output. This means that for each model, three separate models will be trained for each of the three outputs. However, for the multilayer perceptron, this function is not used since a multilayer perceptron has built-in multi-output functionalities. For the multilayer perceptron, a network with three output nodes is created.

Chapter 8

Experimental setup

To test the proposed models, flight data from KLM will be used. This data will be split into two categories: 2023 and 2024 flights. The data will be split on the flight group (ICA or KLC) to create 2 datasets. For both datasets, a 70/30% split will be made on the 2023 flight data to create the train and test set. Flight data from January 1st, 2024 until July 31st, 2024 will be used to evaluate the models on their predictive performance. The experiments will be performed in Python using the models from the sk-learn package [10].

For all models, hyperparameter tuning will be used to find the best values. To automate this tuning, the `GridSearchCV` function from sk-learn will be used with a cross-validation fold of 5.

After the data analysis and some preliminary experiments, the following features have been chosen to be included in the model:

1. Day of week (date at the departure airport, cyclic encoded)
2. Month number (date at the departure airport, cyclic encoded)
3. Departure hour (date at the departure airport, cyclic encoded)
4. Arrival hour (date at the arrival airport, cyclic encoded)
5. Number of booked passengers
6. Number of male passengers
7. Number of female passengers
8. Number of children
9. Number of babies
10. Number of booked economy seats
11. Number of booked premium comfort seats
12. Number of booked business seats
13. Percentage of local passengers
14. Percentage of transfer passengers
15. Percentage of transit passengers
16. Number of saleable seats (number of seats - number of blocked seats)
17. Departure continent (one-hot encoded)
18. Arrival continent (one-hot encoded)
19. Trolley capacity of aircraft
20. Days since last holiday (country of arrival airport)

21. Days until next holiday (country of arrival airport)
22. arrival day is holiday (country of arrival airport)
23. Days since last holiday (country of departure airport)
24. Days until next holiday (country of departure airport)
25. arrival day is holiday (country of departure airport)
26. Number of passengers with nationality x (for all nationalities)

Chapter 9

Metrics

This chapter describes which measure will be used in the experiment to compare the tested models. Not only performance measures will be used, but the models will be compared to a business rule too. This chapter describes the different metrics used and explains the reason for using them.

9.1 Performance metrics

To compare the models, a variation of metrics will be used to measure the performance. Those metrics include R^2, mean absolute error and root mean square error. The R^2 metric is used to measure how much of the data is explained by the model. The MAE is the average of all (absolute) errors. The RMSE is more sensitive to outliers due to the error being squared. The RMSE is an important metric because, for KLM, large outliers (on the low side) cause more risk for delay than small errors. A large error on the upside where too much hand luggage gets collected is bad for passenger satisfaction but does not directly impact the boarding process.

Besides the performance of the test data, KLM also wants to know the performance of the model in practice. More metrics will be introduced to evaluate the models. The first metric is the R^2_{2024} together with the MAE_{2024}. The model is trained on 2023 data and also tested on 2023 data. To know the predictive performance of the model, the models will be tested on data from 2024 which the model has never seen before in the training. The second metric is the R^2_{gate} which measures the accuracy of the model at the gate. Equation 7.1 shows how the y_{gate} is calculated. Since the observed collection at the gate is known, the R^2_{gate} can be calculated.

9.2 Business rule

In Heinrich's report, the following business norm was stated: [redacted]% of predictions have a deviation of less than or equal to [redacted] at the gate [6]. This means the following: the model predicts the total number of hand luggage pieces that should be collected in the entire process before the flight. The number of hand luggage pieces that still need to be collected gets reduced in the process when hand luggage is being collected at the outstations and check-in. The number of hand luggage pieces that still need to be collected at the gate, must have a difference of [redacted] or less with the actual amount of collected hand luggage pieces, for [redacted]% of the predictions. During the experiment, this number will be calculated using the test dataset.

Part IV

Results & Conclusion

Chapter 10

Results

This chapter presents and analyzes the results of all proposed models to predict the number of excess hand luggage for ICA and KLC flights. First, the results from the single output models are gathered and analyzed using basic statistics. After that, the multi-output model results will be shown and discussed. Following this, a comparison with the current heuristic model will be made after which the models will be discussed using the performance metrics and business rule.

10.1 Single output results

This section gives an overview of the hyperparameter tuning of the models and will give a showcase of the model output. The results of all tuned models are gathered and shown in Table 10.1. Table C.1 shows the results of the hyperparameter tuning.

Table 10.1 shows that all the models could explain the amount of excess hand luggage using the training data. The lowest R^2 score for the ICA models is from the Ridge regression model which has an R^2 value of 0.76, which is already above the results from Heinrich [6]. For KLC, the best R^2 is lower with a value of 0.47 for the random forest regressor. The random forest regressor has the lowest mean absolute error and the lowest root mean square error with values of 7.36 and 10.13. The results show that the random forest regressor gave the overall best results with an R^2 value of 0.83, an MAE of 7.36, and an RMSE value of 10.13.

Table 10.1: This table shows the results of the tests of the single-output models which were trained on 2023 flight data.

(a) Table for ICA single-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2	0.76	0.82	0.76	**0.83**	0.77
MAE	8.94	7.56	8.70	**7.36**	8.7
RMSE	11.91	10.25	12.07	**10.13**	11.62
$R^2_{check-in}$	0.70	0.77	0.69	**0.78**	0.71
R^2_{gate}	-1.39	-0.91	-1.51	**-0.82**	-1.24
$R^2_{outstations}$	0.62	**0.72**	0.61	**0.72**	0.63

(b) Table for KLC single-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2	0.36	0.45	0.33	**0.47**	0.42
MAE	4.09	3.73	4.13	**3.62**	3.83
RMSE	5.38	4.98	5.52	**4.90**	5.14
$R^2_{check-in}$	0.30	0.40	0.26	**0.42**	0.36
R^2_{gate}	-0.88	-0.80	-0.89	**-0.56**	-0.69
$R^2_{outstations}$	-0.98	-0.70	-1.08	**-0.64**	-0.81

Knowing how well these models would perform in practice is difficult to know without additional measures. As shown in Table 10.1, the models perform well with a mean absolute error of 7.36 (ICA) and 3.62 (KLC) for the random forest regressor. However, adding additional measures gives a different impression. The measures that give a different impression are the $R^2_{check-in}$, R^2_{gate} and $R^2_{outstations}$. These measures are calculated as stated in Equation 7.1 and are shown in Table 10.1. The results show that all models maintain their score for the check-in and outstations R^2 values, but have a negative R^2 score for the gate predictions. To solve this, the multi-output models have been created. The next section covers the results of the multi-output models which should give better results for the R^2_{gate} score.

10.2 Multi output results

In this section, the gathered results of the multi-output models are shown. These models predict 3 values: for each collection point in the flight process the number of collected hand luggage pieces is predicted. This means that the prediction for how many hand luggage pieces are collected at check-in is different from the prediction for the gate, or the outstations. The total number of collected hand luggage pieces is determined by the sum of these three predictions. The results can be found in Table 10.2. Table C.2 shows the results of the hyperparameter tuning.

The results of the total amount of excess hand luggage are similar to those of the single-output models. Table 10.2 shows that the R^2 values of the multi-output models do not have a difference bigger than 0.02 compared to their single-output counterpart. These multi-output models are created to solve the negative R^2_{gate} score from Table 10.1. Table 10.2 shows the R^2 results for each collection process. The results show that all multi-output models have a positive R^2_{gate} value. The random forest regressor gave the best results with R^2 values of 0.88 for the $R^2_{check-in}$, 0.62 for the R^2_{gate} and 0.90 for the $R^2_{outstations}$.

The KLC multi-output models do not show an improvement for any of the R^2, mean absolute error and root mean square error scores (except for the MAE of the MLPR, Table 10.2). For the $R^2_{check-in}$, R^2_{gate} and $R^2_{outstations}$, all models show improvements over their single-output counterpart.

Table 10.2: This table shows the results of the tests of the multi-output models which were trained on 2023 ICA flight data.

(a) Table for ICA multi-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2	0.76	0.81	0.74	**0.83**	0.78
MAE	8.96	7.72	8.96	**7.40**	8.64
RMSE	11.93	10.70	12.43	**10.21**	11.54
$R^2_{check-in}$	0.77	0.87	0.83	**0.88**	0.84
R^2_{gate}	0.39	0.55	0.54	**0.62**	0.57
$R^2_{outstations}$	0.80	0.88	0.85	**0.90**	0.87

(b) Table for KLC multi-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2	0.36	0.43	0.30	**0.46**	0.41
MAE	4.10	3.68	4.26	**3.66**	3.84
RMSE	5.40	5.10	5.65	**4.97**	5.16
$R^2_{check-in}$	0.45	0.56	0.47	**0.59**	0.54
R^2_{gate}	0.14	0.25	0.11	**0.27**	0.20
$R^2_{outstations}$	0.57	**0.66**	0.54	**0.66**	0.65

10.3 Including national holidays

The models above did not include the national holidays dataset. The results of the experiments *with* the holiday dataset can be found in Table 10.3. For ICA no significant changes were observed with the biggest change being an increase in the R^2_{gate} value for the Multilayer Perceptron from 0.55 to 0.57.

For KLC, the inclusion of the holiday dataset shows some minor improvements in the performances. The multilayer perceptron shows the biggest increase in R^2 score from 0.43 to 0.46. The Ridge Regression model shows an improvement of the R^2_{gate} from 0.14 to 0.20. The other metrics also show some minor improvements, but no big changes can be found.

Table 10.3: This table shows the results from the multi-output models that used the holiday data in addition to the flight data.

(a) Table for ICA multi-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2	0.76	0.80	0.74	**0.83**	0.78
MAE	8.95	7.83	8.95	**7.37**	8.63
RMSE	11.92	10.86	12.42	**10.16**	11.51
$R^2_{check-in}$	0.77	0.86	0.83	**0.89**	0.84
R^2_{gate}	0.39	0.57	0.54	**0.62**	0.57
$R^2_{outstations}$	0.80	0.88	0.85	**0.90**	0.87

(b) Table for KLC multi-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2	0.37	0.46	0.30	**0.47**	0.42
MAE	4.06	3.62	4.25	**3.61**	3.82
RMSE	5.35	4.97	5.64	**4.90**	5.13
$R^2_{check-in}$	0.46	0.59	0.48	**0.61**	0.55
R^2_{gate}	0.20	0.25	0.11	**0.27**	0.20
$R^2_{outstations}$	0.58	**0.67**	0.54	**0.67**	0.66

10.4 Using outstation collection rate as an input

Since the shortest KLM flight already takes 40 minutes (AMS → BRU), the number of collected hand luggage pieces at the outstations can be used as input for the gate prediction model. Table 10.4 shows the results of adding the number of collected hand luggage pieces at outstations as a feature to the model that predicts the amount of hand luggage that should be collected at the gate. The results show no significant increase in the scores for the intercontinental flights.

10.5 Comparison with current heuristic model

Comparing the proposed models to the current heuristic models is done using the metrics described in Chapter 9. These metrics are calculated by the predictions done by the current heuristic model. Table 10.5 shows the results from 2023 for the KLC and ICA flights. The current model shows a negative R^2 for both KLC and ICA flight groups. The bin2d plots (Figure 4.1) of the model show this complete misclassification of the models. The figures show that the model is mostly predicting too high for KLC flights and mostly predicting too low for ICA flights.

When comparing the proposed model with the current model, the first comparison is made using the metrics. For ICA flights, the best-performing proposed model is the Random Forest regressor which has an R^2 value of 0.83 (Table 10.2a), which is much higher than the current model which has an R^2 value of -1.63. The other metrics also show much better results compared to the current model. The Mean Absolute Error from the Random Forest Regressor is 7.40, against an MAE of 20.28 of the current model. The RMSE shows a decrease from 27.72 to 10.21.

For KLC flights, the differences are smaller since the amounts of excess hand luggage are lower. This is due fact that the KLC aircraft can carry fewer passengers compared to the wide-body aircraft. The best-performing KLC model is also the Random Forest Regressor. The model has an R^2 score of 0.46 (Table 10.2b) which is significantly higher than the -1.17 of the current heuristic

Table 10.4: This table shows the results from the gate prediction models for which the observed number of collected hand luggage pieces at the outstations is added as an extra feature to the models.

(a) Table for ICA multi-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2_{gate}	0.39	0.56	0.54	**0.63**	0.57
MAE_{gate}	2.24	1.8	2.04	**1.65**	1.84
$RMSE_{gate}$	4.32	3.68	3.76	**3.39**	3.64
KLM norm	0.91	0.92	0.92	**0.93**	0.92

(b) Table for KLC multi-output models.

	Ridge regression	Multilayer perceptron	Classification tree regressor	Random forest regressor	XG Boost
R^2_{gate}	0.15	0.25	0.11	**0.27**	0.20
MAE_{gate}	1.46	1.26	1.36	**1.23**	1.33
$RMSE_{gate}$	2.62	2.45	2.67	**2.42**	2.53
KLM norm	**0.96**	0.95	0.95	**0.96**	**0.96**

model. The Random Forest Regressor has an MAE and RMSE score of 3.66 and 4.97 against 6.09 and 8.22 of the current model. This shows that, although it is not perfect, the Random Forest Regressor improves the current heuristic model by a lot.

Table 10.5: This table shows the results of the current model for ICA and KLC flights. The results are from flight data from 2023.

	ICA	KLC
R^2	-1.63	-1.17
MAE	20.28	6.09
RMSE	27.72	8.22

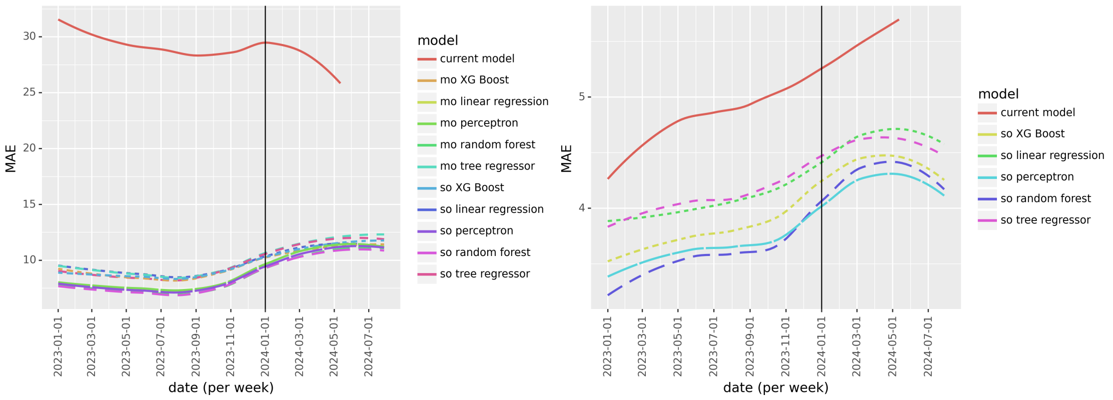
10.6 Performance of metrics and business rule

As discussed in Section 9, there are two custom measures to test the model's performance. The first measure tracks the model drift over time, and the second tests the accuracy of the model for the gate collection.

The first measures that will be discussed are the R^2_{2024} and the MAE_{2024} which measure the performance of the 2023-trained model in 2024 (data up until July). Figure 10.1 shows the MAE per week for all tested models. In Figure C.1 this score is split out per aircraft type. Table 10.6 shows a lower R^2_{2024} score for all models compared to their R^2 score on the test data in 2023. However, this decrease in the R^2 score does not start in January 2024. Figure 10.1a shows that an increase in the MAE started in September 2023. This increase will be further discussed in Section 11.1.

The second measure is the percentage of absolute prediction errors below [redacted] at the gate. This measure ensures the amount of trust that KLM can put in the predictions of the model. The company uses the predictions of this model to plan and schedule gate agents. A flight with a lot of predicted hand luggage collection at the gate means that more gate agents need to be scheduled for that flight to ensure minimal delays at the gate. Figure 10.2a shows that none of the single-output ICA models meet the criterion of 80%. Contrary to this, *all* multi-output ICA models meet the criteria with the multi-output random forest as the best-performing model. On the other hand, for KLC, all single-output models already meet the business rule as shown in Figure 10.2b. The multi-output models increase this score up to 96%.

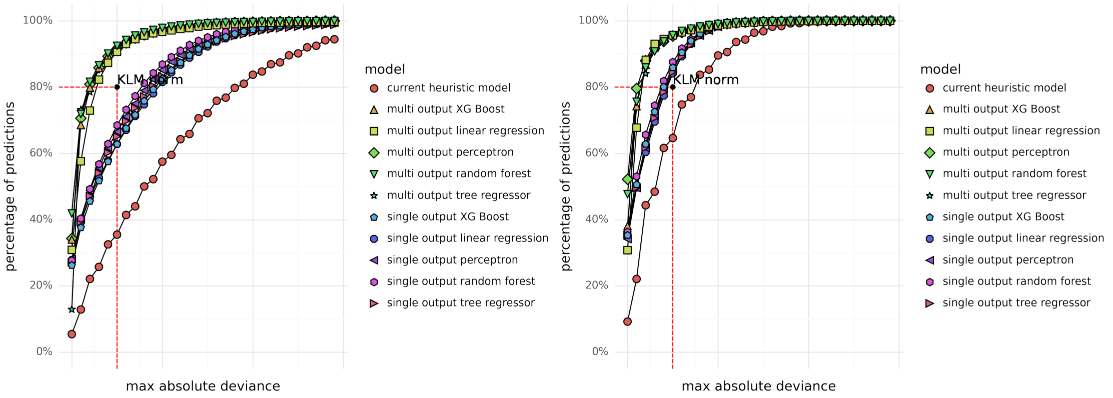
Figure 10.1: Mean Absolute Error over time for the trained single-output (SO) and multi-output (MO) models. The models are trained on data from 2023 (left of the black intercept). The MAE is calculated for the 2023 test set. For 2024, all flightlegs done for that flight group are evaluated.



(a) Plot for ICA flights.

(b) Plot for KLC flights.

Figure 10.2: Percentage of ICA predictions with a maximum absolute error of x.



(a) Plot for ICA flights.

(b) Plot for KLC flights.

Table 10.6: This table shows the results for two of the measures described in Chapter 9 for the Single-Output (SO) and Multi-Output (MO) variants of the models.

	Ridge regression		Multilayer perceptron		Classification tree regressor		Random forest regressor		XG Boost regressor	
	SO	MO	SO	MO	SO	MO	SO	MO	SO	MO
R^2_{2024}	0.61	0.61	0.64	0.63	0.58	0.59	0.65	**0.66**	0.64	0.64
MAE_{2024}	11.24	11.25	**10.71**	11.32	11.56	11.67	10.72	10.73	10.84	10.89
klm norm	0.62	0.91	0.68	0.92	0.64	0.92	0.69	**0.93**	0.63	0.92

10.7 Feature importance

After the models were trained, feature importance metrics were calculated. Figure C.5 shows the Gini importance metric for both the ICA and KLC multi-output models. The Gini importance shows the (normalized) sum of impurity reductions that a feature did achieve in all trees of the random forest. The figure shows that the percentage of local and transfer percentages were the features that achieved the highest impurity reduction for both ICA and KLC.

Because of the cardinality bias of the Gini importance metric, two other methods have been used to calculate the feature importance. Figure C.6 shows the impact on the MAE score if the values in the test set of a feature were randomized. Figure C.6b shows that randomizing the boolean feature that states whether a flight departs from Schiphol Airport gave the highest MAE increase. The total number of bookings and the total number of booked economy tickets showed an increase in the MAE when randomized.

The features that showed no role in any of the above metrics were most of the n_XX_passengers features. Besides some nationalities that came back in all the tests, other nationalities never came back in the splits. From the 267 nationalities in the data, 168 nationalities were never used to make a split in the random forest. Figure C.7 shows that the features that appeared the most in the splits differ from ICA and KLC. The three splits with the most splits on ICA were on features about the passengers. For KLC, the 4 features with the most splits were all on national holiday features.

10.8 Performance on flights with many transfer passengers

Section 5.4 describes the influence of transfer passengers on the total amount of collected hand luggage. This situation is summarized in research question 3. To test the prediction accuracy, the following statistics have been calculated using the multi-output ICA model for the ICA flight group. The MAE for the entire model is 7.37. Table 10.7 and Figure C.2 show the results of splitting the results based on the passenger composition. The MAE value for flights with a percentage of transfer passengers between 70% and 100% is 8.05. For flights with 0% and 70% of transfer passengers, this value is 7.12. This shows that the MAE is 13% higher for flights with $70\% - 100\%$ transfer passengers.

Table 10.7: The MAE of the multi-output Random Forest Regressor for ICA flights in 2023.

% of transfer passengers	MAE	samples
0-10	7.32	3840
10-20	7.74	1150
20-30	7.69	460
30-40	5.86	211
40-50	6.28	309
50-60	5.97	570
60-70	6.55	880
70-80	7.05	1234
80-90	8.57	1067
90-100	10.64	256

Chapter 11

Discussion

11.1 Decrease of score in September

In Figure 10.1a, an increase in the MAE score can be observed in September 2023 for the ICA flights. This increase is not expected since the training data contains flights from January 2023 until December 2023. The collection rates change from week to week, so the models should be able to capture this change in September. An increase from January 2024 onwards is expected since the training data does not include flights from 2024. After analyzing the data a change is observed that could potentially be the cause of this increase: the heuristic model is enabled for ICA flights in September 2023. The existence of a prediction model might have changed the collection behaviour of gate agents. This is something that has to be taken into account when the proposed model will be released in production.

11.2 Retraining on model output

The previous section discussed the increase in the mean absolute error when a prediction model exists. This means that the gate agents react in some way to the model. Before the boarding process starts, gate agents know how many hand luggage pieces they should collect during the boarding. After the boarding process, the number of collected hand luggage gets logged in the database. Due to the model drift, the model should be retrained periodically. However, since the gate agents first get the information on how much hand luggage to collect, the model could hypothetically be trained on its output. This happens if the gate agents always perfectly adhere to the output of the model and if there never is excess hand luggage in the cabin. The model will only be corrected if there is not enough hand luggage collected at the gate and gate agents need to retrieve excess hand luggage from the cabin. If too much hand luggage is collected, there is still space in the cabin for more hand luggage. This signal does currently not reach the gate agents. At this moment, KLM is developing an app for the cabin personnel to report how much space is left in the overhead bins. This app will create a feedback loop that is needed to prevent the model from drifting further from a perfect fit.

11.3 Censored data

As described in the previous section, the lack of a feedback loop has a risk of unnoticeable model drift. Another potential cause for this drift is the censored data as described in Section 5.1. Since collecting not enough hand luggage has big implications, gate agents want to stay from the edge (the perfect fit). This results in always having some free spaces for trolleys. However, since the gate agents collected more than predicted, the result is that the next model will predict higher collection numbers. Another potential solution besides the addition of a feedback loop is an algorithm that, instead of predicting the amount of hand luggage pieces that need to be collected, predicts the total amount of hand luggage pieces. Predicting the total number of hand luggage pieces will result in a number that can be easily interpreted by the data scientists and the gate agents.

11.4 Negative R squared score

In this thesis, multiple negative R^2 scores can be observed (Tables 4.1, 10.1, 10.5). These R^2 scores are calculated using the sk-learn package. A negative R^2 score means that the predictions were worse than if the prediction always were the $\bar{y}$ of the test set.

Table 10.1 shows a negative score for all gate estimations. This can be explained by the fact that the single-output has not been trained for the gate collection process. This means that the $\hat{y}_{gate}$ depends on how much hand luggage is collected at the check-in and the outstations. It still does not explain why the R^2_{gate} is negative since the check-in and outstation collection processes are highly correlated with the gate collection process. At the gate, the amount of collected hand luggage should be the number of hand luggage pieces that have not yet been collected at the check-in and outstations when assuming a perfect fit in the aircraft. In theory, the collection numbers should look like this:

$$perfectfit = HL^{capacity}_{aircraft} = TotalHL - collectedHL \tag{11.1}$$

The negative R^2 scores could indicate that there is *not* a perfect fit of hand luggage in the aircraft. The multi-output models give a decent R^2 score for the gate prediction, but this only means that the data can be explained using the models. It does not mean that following these predictions always gives a *perfect fit*.

11.5 Multi-output models

How should the multi-output models be treated? Should each collection point collect the predicted amount of hand luggage? Or does only the total number of collected hand luggage matter and do the gate agents only have to worry about the difference between $\hat{y}_{total}$ and the current collected number of hand luggage pieces?

The second seems more logical when comparing it to the research problem: predict the difference between the total number of hand luggage pieces and the aircraft capacity (for hand luggage pieces). This means that it does not matter *when* the hand luggage gets collected, but *how much* hand luggage gets collected.

The multi-output models main purpose is to give a rough estimation on how much hand luggage does have to be collected at the gate. The multi-output models are able to do this. For the operation, the advice is to use the $\hat{y}_{total}$ at the check-in and the gate and subtract all hand luggage pieces that get collected during the process (including the hand luggage pieces from transfer passengers).

Chapter 12

Conclusion

In this thesis, new machine learning models are proposed for predicting the amount of excess hand luggage. Two different types of flight groups are investigated: intercontinental flights done by wide-body aircraft and flights done with the KLM Cityhopper (Embraer 175, 190 and 195-E2).

12.1 The current situation

The problem that is solved with this research is that KLM Royal Dutch Airlines had a prediction tool that did not perform well. The prediction tool predicts the number of hand luggage pieces that have to be collected for a given flight leg. This amount should not be too low, that would cause an overflow in the cabin. The amount should also not be too high, which would cause passenger dissatisfaction. This dissatisfaction is increased if there is still space left in the overhead cabins (because of a too-high prediction). The prediction model currently in use is a heuristic that multiplies certain factors by a set value. This model has an R^2 score of -1.63 for ICA flights and -1.17 for KLC flights.

The data study showed no variables with high correlation values for flights originating from and flights flying to Schiphol International Airport. For ICA, the most significant feature for the total number of collected hand luggage pieces is the total number of economy bookings followed by the total number of passengers. Other features with a Pearson or Spearman correlation score above 0.3 were: the number of sale-able seats, the total number of female passengers and the percentage of transfer passengers. For KLC, the same features gave the highest Pearson or Spearman correlation result. In addition to this, the number of male passengers also gave a Spearman correlation value of ≈ 0.4.

12.2 The proposed models

For both flight groups, of the five tested models, the Random Forest Regressor gave the best results. This model gave an R^2 value for respectively ICA and KLC of 0.83 and 0.46. The Mean Absolute Error decreased significantly for both flight groups. The results show that adding the national holidays as a feature gave no (significant) changes for ICA and a minor decrease in the MAE and RMSE for the KLC Random Forest Regressor. Adding the number of collected hand luggage pieces at outstations as a feature for the model that predicts the amount of excess at the gate also showed no significant improvements for either flight group.

The KLC models showed worse results compared to their ICA models. This can be explained by the fact that the KLC encounters relatively more noise. The Embraer aircraft that KLM uses has a capacity of 88, 100 or 132 passengers. This is much lower than the wide-body aircraft used for ICA flights with the largest aircraft having a capacity of 408 pax. The result of this is that voluntarily checked hand luggage gives noise between the obliged checked hand luggage. This claim can be supported by Figure A.10a where a big spike can be observed for the number of passengers from [redacted] [redacted]. Gate agents observe that the passengers from this country often have hand luggage which does not comply with KLM's hand luggage size allowances. According to the gate agents, these passengers travel to Schiphol International Airport with an airline that allows for bigger hand luggage on their aircraft resulting in an involuntarily checking of their hand luggage at Schiphol.

12.3 The performance of the proposed models

Additional metrics have been defined to compare the models in specific situations. The first metric is the R^2 score in 2024. This metric was introduced to measure the predictive performance of the model in a year for which the model has not been trained on. The results in Table 10.6 show that all models performed worse in 2024 compared to 2023. However, Figure 10.1 shows that the MAE easily stays below the current models (which gets constantly tweaked to be improved in contrast to the proposed models which did not get tweaked for the 2024 data). This shows that the proposed models outperform the model currently in use.

The second introduced measure was the KLM norm that stated that 80% of the predictions (at the gate) had a maximum absolute error of [redacted]. The heuristic model currently in use does not reach this norm for both flight groups. The results show that for both ICA and KLC, all multi-output models satisfy the norm easily. For the single-output models, the flight groups show a different result with none of the ICA single-output models reaching the norm. The rule is reached by all single-output KLC models but by none of the single-output ICA models. This can be explained by the fact that the KLC flight group has to deal with a lot fewer passengers and thus less hand luggage. The result of this is that the business norm is more easily satisfied for KLC since the allowed maximum absolute error is the same for all flight groups.

12.4 Similar model performance

The results showed that all tuned (multi-output) models performed very similarly besides the Linear Regression model. This made it difficult to make a decision on which model should be adopted by KLM to be used in production. The best-performing model is the Random Forest Regressor, but the differences with the multilayer perceptron are small. This indicates a robust dataset for which multiple types of models can be used to predict the amount of excess hand luggage. To prevent too much drift, due to shifts in hand luggage trends, periodic retraining is advised. This will result in a robust model that can predict the amount of excess hand luggage for KLM's intercontinental and Cityhopper flights.

12.5 Summary

To summarize all these conclusions, the results show that the multi-output models are preferred above the single-output models since the overall performance does not drop while the individual performance (at the different collection points) does increase. The inclusion of the national holidays set and the number of collected hand luggage pieces at outstations did not improve the performance of the models. When the best model has to be chosen, the metrics show no significant differences. However, since most gate agents approach this problem as some kind of decision tree, the random forest regressor has been chosen as the best performing model on this data. In the end it is a regression problem, but before that a lot of questions are asked regarding the aircraft type, season and destination. Hence, the random forest regressor fits the spirit of how the decisions are made.

Chapter 13

Future work

13.1 Luggage allowances

How much (luggage) passengers can bring on their flight varies from airline to airline. Even within KLM, many factors impact the luggage allowance. The big factor is the ticket type, passengers flying economy can bring less luggage than business class passengers. For some flights, economy passengers can only bring 1 piece of hand luggage, whereas on other flights economy passengers can bring 2 luggage pieces: one hand luggage piece and a checked luggage piece. Other factors could not be discovered, but even for the same flights, the allowances could change from time to time. It is expected that these allowances impact the amount of excess hand luggage. If a passenger already needs to wait for their checked luggage at the arrival airport, waiting for two luggage pieces does not take much longer than waiting for one. Gate agents within KLM have the feeling that passengers who only travel with hand luggage do not want to hand in their hand luggage since they do not want to wait for their luggage at the arrival airport. These allowances were not included in the research because no dataset or algorithm contained this information.

13.2 Quantile models

The model that is proposed in this thesis does estimate the total number of excess hand luggage pieces for a given flight. What it does not predict is the amount of uncertainty. Currently, the gate agents see two numbers on their screen: a minimum and a maximum. The gate agents should aim to collect the number of hand luggage pieces that lie between these two bounds. However, the current heuristic model does not give two bounds, it only gives one prediction. KLM adds 5 to that number to calculate the upper bound. This does also not reflect any uncertainty. To solve this uncertainty, quantile models could be used to reflect how certain the model is about a prediction. The 95% (or other numbers) intervals of quantile models could be used as lower and upper bounds for the gate agents. This way gate agents not only know how much hand luggage they should collect but also know how big the uncertainty spread is. If the model is not certain about a prediction, the gate agents could use the higher spread to manually decide how much hand luggage they will collect. Further research is needed to investigate how gate agents would react to such quantile models.

13.3 Minimize baggage intake

The last recommendation is an extension towards predicting over multiple flight legs. This extension can minimize the total number of hand luggage pieces collected: Suppose KLM has a flight from $A \to B \to C$ and a surplus of hand luggage is expected for both flight legs ($A \to B$ and $B \to C$), Collecting from a passenger that travels from A to C decreases the total amount of hand luggage that needs to be collected because the hand luggage is collected for both legs. This is preferred compared to first collecting from a passenger that flies from $A \to B$ and then collecting from a passenger that flies from $B \to C$.

The final step in this extension is to not only minimize the baggage intake for flight legs, but for KLM's entire flight network. A second goal can be introduced to also minimize the aircraft delays due to an excess amount of hand luggage at the gate. If a passenger flies from $A \rightarrow C$ and flightleg $B \rightarrow C$ is expected to have a large amount of excess hand luggage at the gate, KLM can decide to collect the hand luggage of this passenger at airport A. This will reduce the work load for the gate agents on airport B and reduce the risk of having a delay due to an overflow of hand luggage.

Bibliography

[1] arkid15r, *Python holidays package*, Aug. 2024. [Online]. Available: https://pypi.org/project/holidays/.

[2] M. Ball, C. Barnhart, M. Dresner, M. Hansen, K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani, B. Zou, *et al.*, "Total delay impact study," in *NEXTOR Research Symposium, Washington DC*, 2010.

[3] X. van der Broek, "Solving klm's hand luggage problem," M.S. thesis, Delft University of Technology, Feb. 2015. [Online]. Available: http://resolver.tudelft.nl/uuid:f4730c6e-7c83-4d6b-a990-4a72b3493318.

[4] F. Claessen, *School holidays in python — unlocking an important feature for understanding consumption data*, Nov. 2020. [Online]. Available: https://seita.nl/2020/11/17/school-holidays-in-python-understanding-consumption-data/.

[5] J. Fuchte, N. Dzikus, and V. Gollnick, "Cabin design for minimum boarding time," Sep. 2011.

[6] C. V. Heinrich, "Hand luggage overflow prediction at klm," M.S. thesis, Aug. 2023. [Online]. Available: http://essay.utwente.nl/96988/.

[7] R. J. Milne and A. R. Kelly, "A new method for boarding passengers onto an airplane," *Journal of Air Transport Management*, vol. 34, pp. 93–100, 2014, ISSN: 0969-6997. DOI: https://doi.org/10.1016/j.jairtraman.2013.08.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969699713001166.

[8] R. J. Milne and M. Salari, "Optimization of assigning passengers to seats on airplanes based on their carry-on luggage," *Journal of Air Transport Management*, vol. 54, pp. 104–110, 2016, ISSN: 0969-6997. DOI: https://doi.org/10.1016/j.jairtraman.2016.03.022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969699715300235.

[9] S. Niemeijer, *Hand luggage prediction model eur/klc*, Internal powerpoint of KLM, Dec. 2020.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] X. Ren, X. Zhou, and X. Xu, "A new model of luggage storage time while boarding an airplane: An experimental test," *Journal of Air Transport Management*, vol. 84, p. 101 761, 2020, ISSN: 0969-6997. DOI: https://doi.org/10.1016/j.jairtraman.2019.101761. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969699718303405.

[12] F. I. Romli and N. H. Ariffin, "Malaysian public survey on the current carry-on luggage handling onboard commercial transport aircraft," *International Journal of Engineering & Technology*, vol. 7, no. 4.13, pp. 14–17, 2018.

[13] M. Schultz, "Implementation and application of a stochastic aircraft boarding model," *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 334–349, 2018, ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2018.03.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X18303735.

[14] Transavia, *Mag mijn bagage altijd mee in de cabine?* [Online]. Available: https://www.transavia.com/service/nl-nl/bagage/handbagage/plek-reserveren.

[15] A. Zinovei, *Iata airport codes*, Nov. 2019. [Online]. Available: https://www.kaggle.com/datasets/zinovadr/iata-airport-code.

Part V

Appendices

Appendix A

Data analysis results from KLM data

Figure A.1: Correlation calculation results for ICA flights in 2023 arriving at Schiphol.

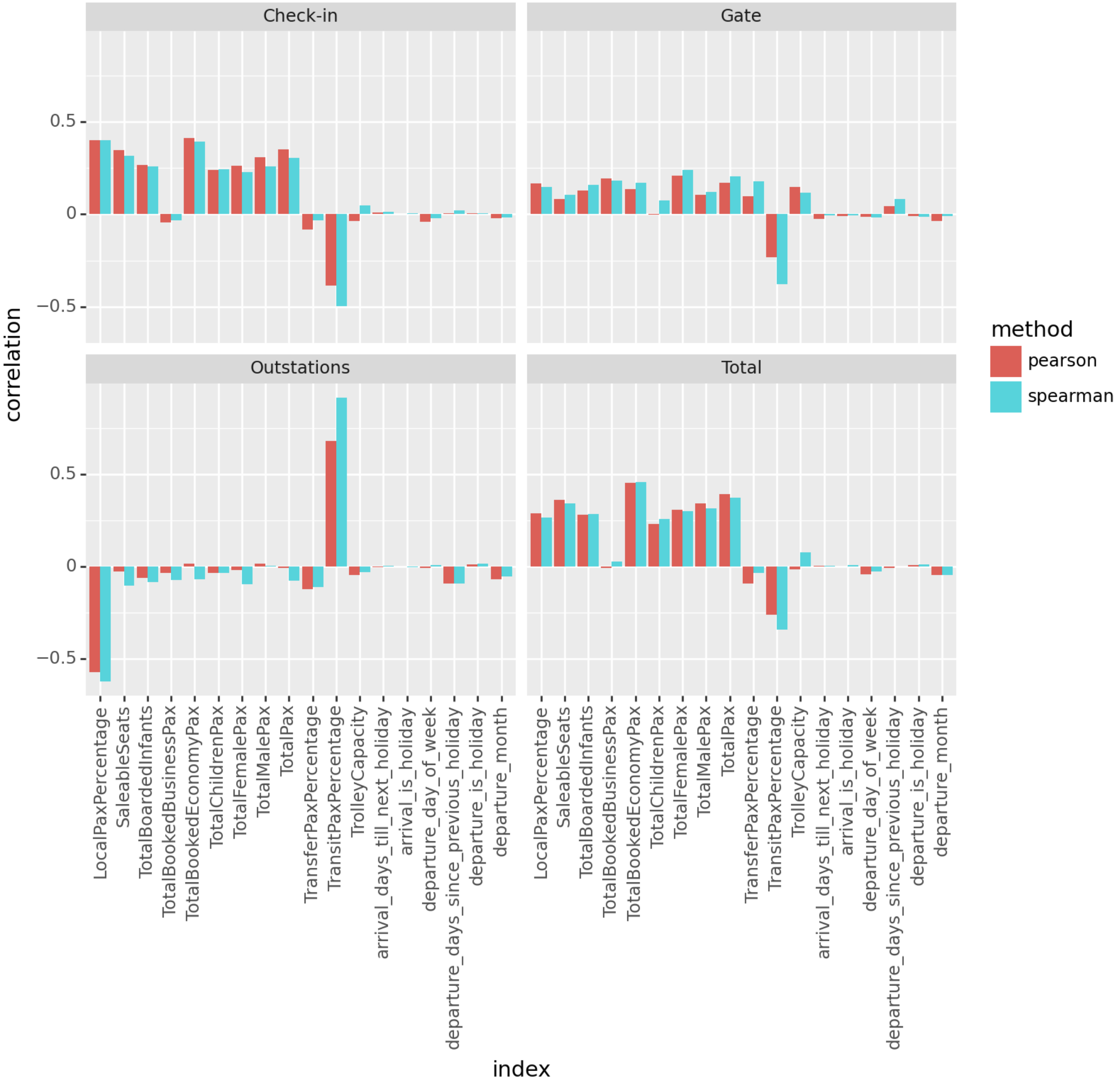


Figure A.2: Correlation calculation results for ICA flights in 2023 departing from Schiphol.

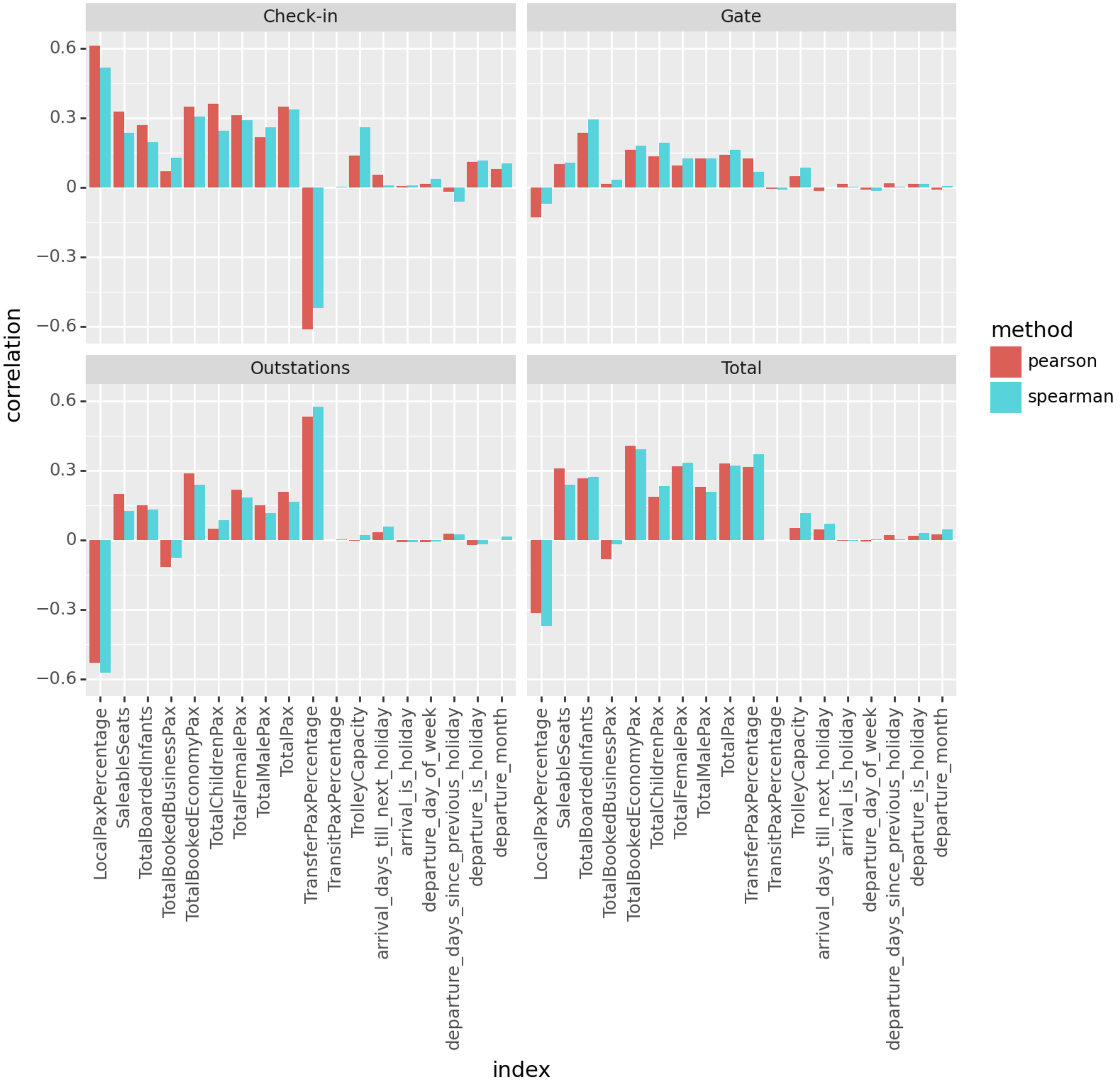


Figure A.3: Correlation matrix for Spearman correlation for ICA flights in 2023 departing from Schiphol.

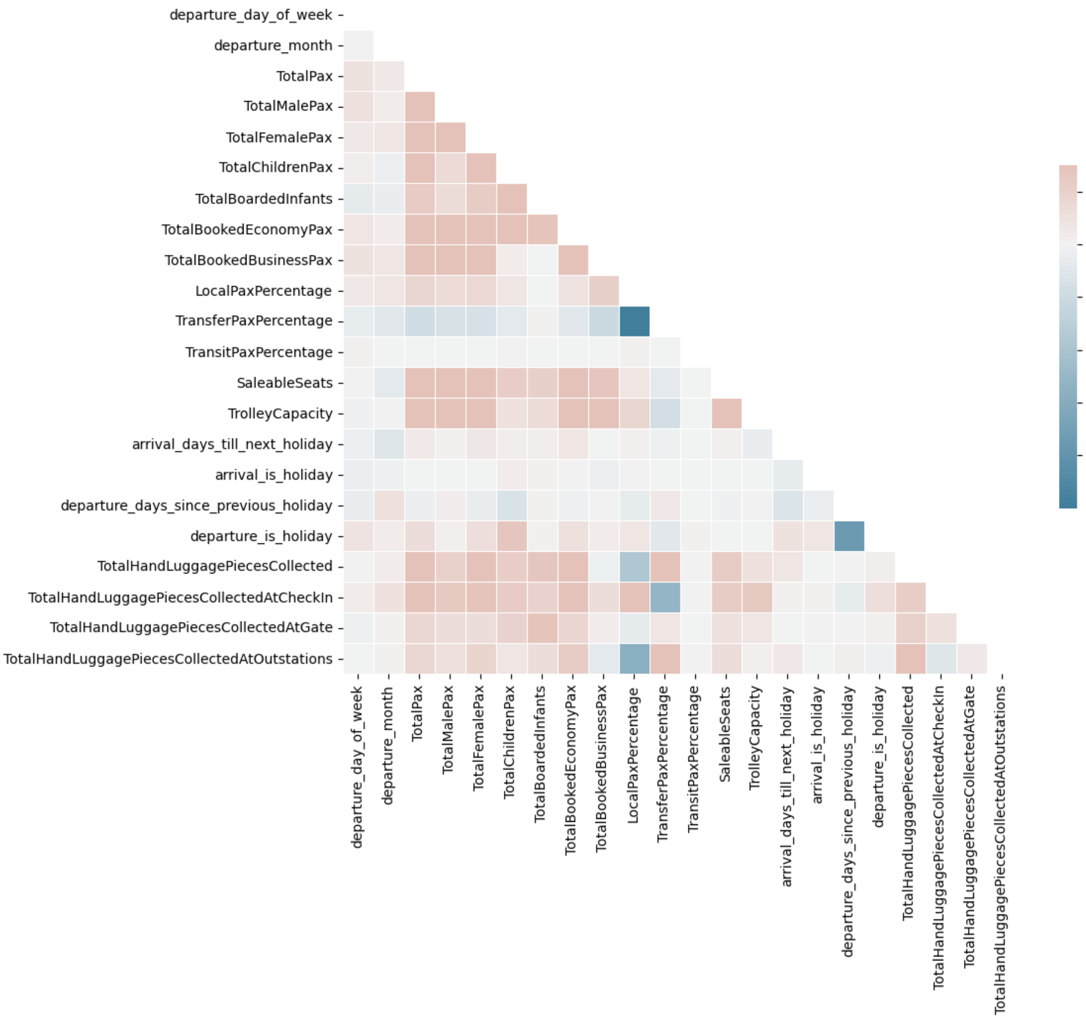


Figure A.4: Correlation matrix for Spearman correlation for ICA flights in 2023 arriving at Schiphol.

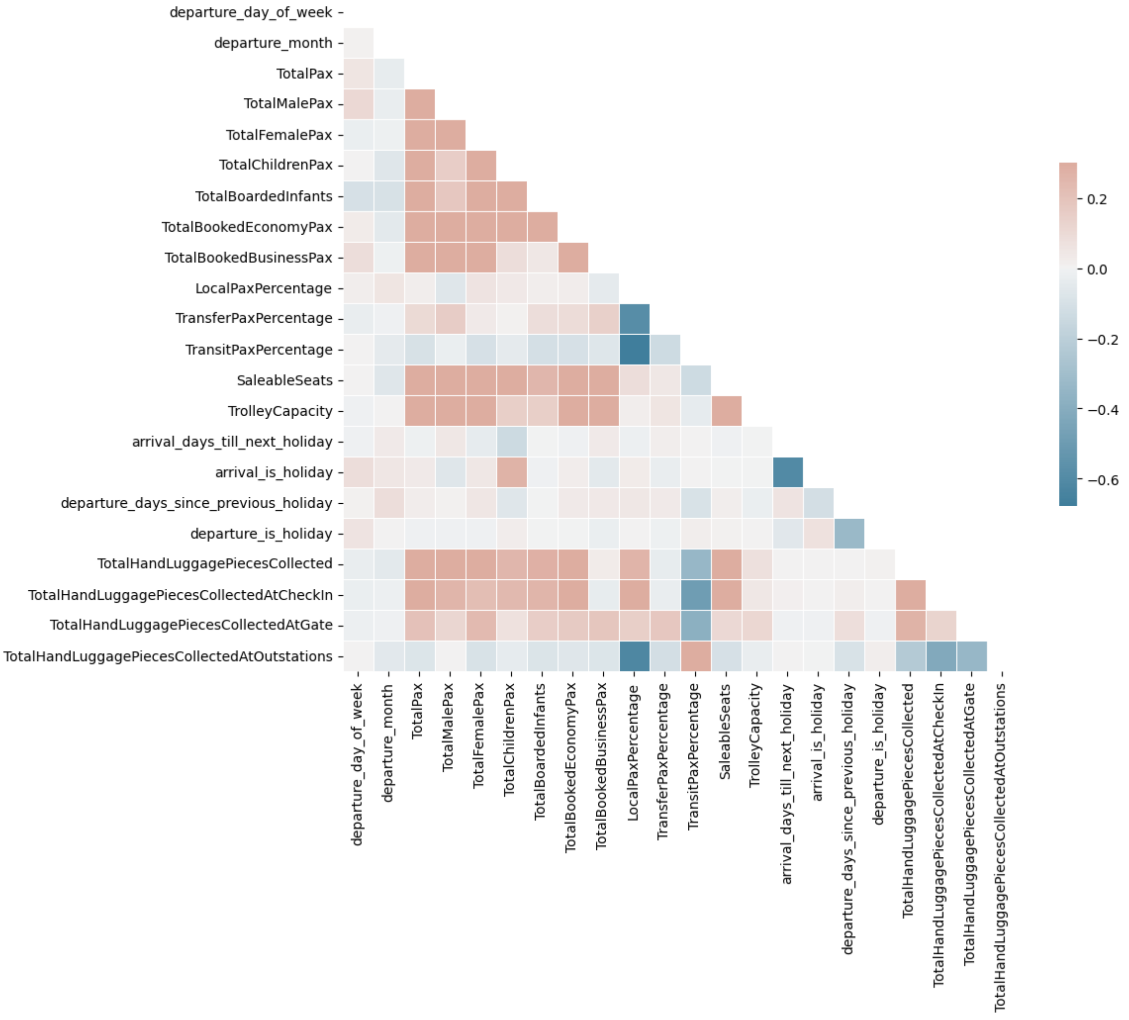


Figure A.5: Correlation calculation results for KLC flights in 2023 arriving at Schiphol.

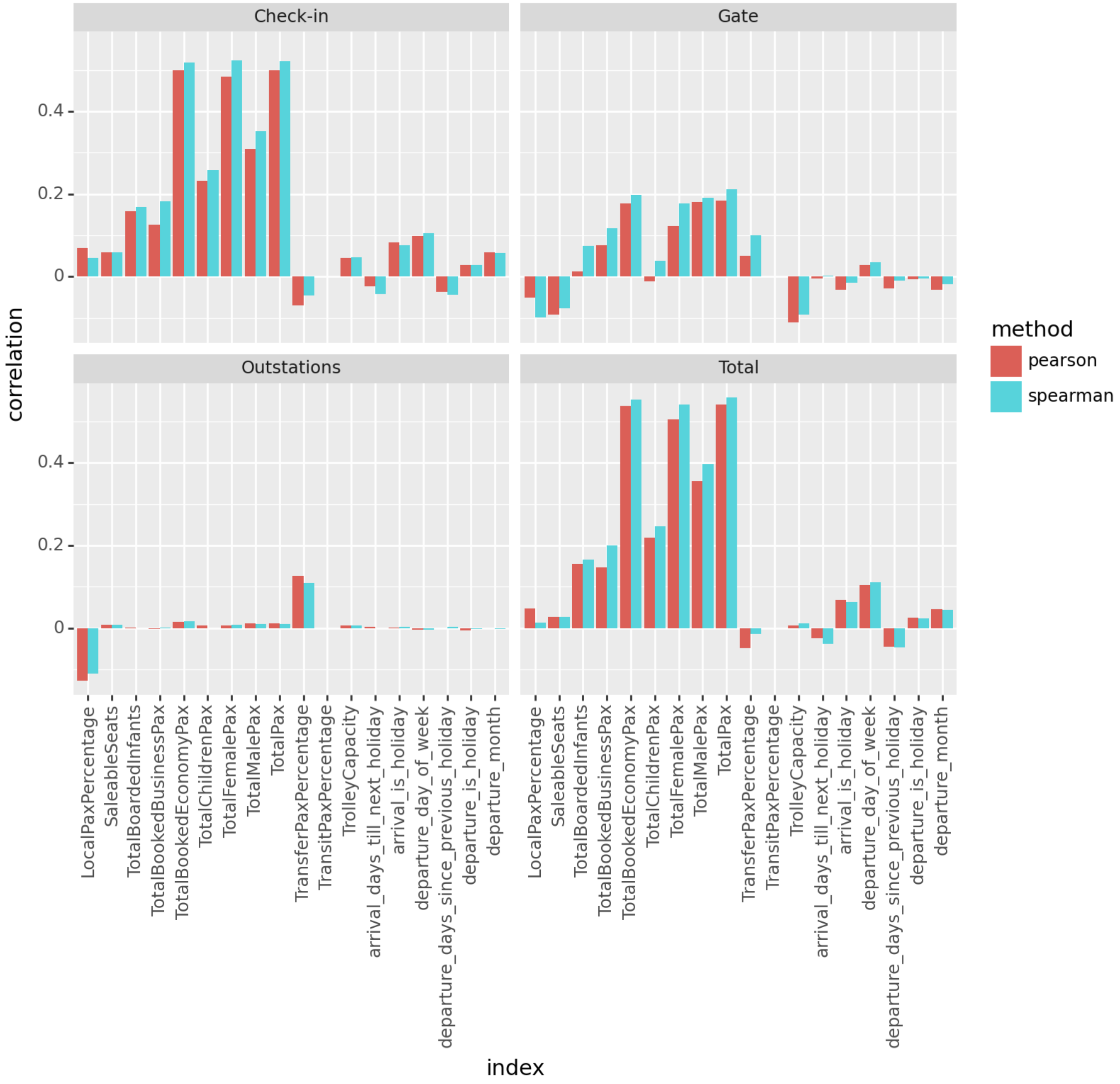


Figure A.6: Correlation calculation results for KLC flights in 2023 departing from Schiphol.

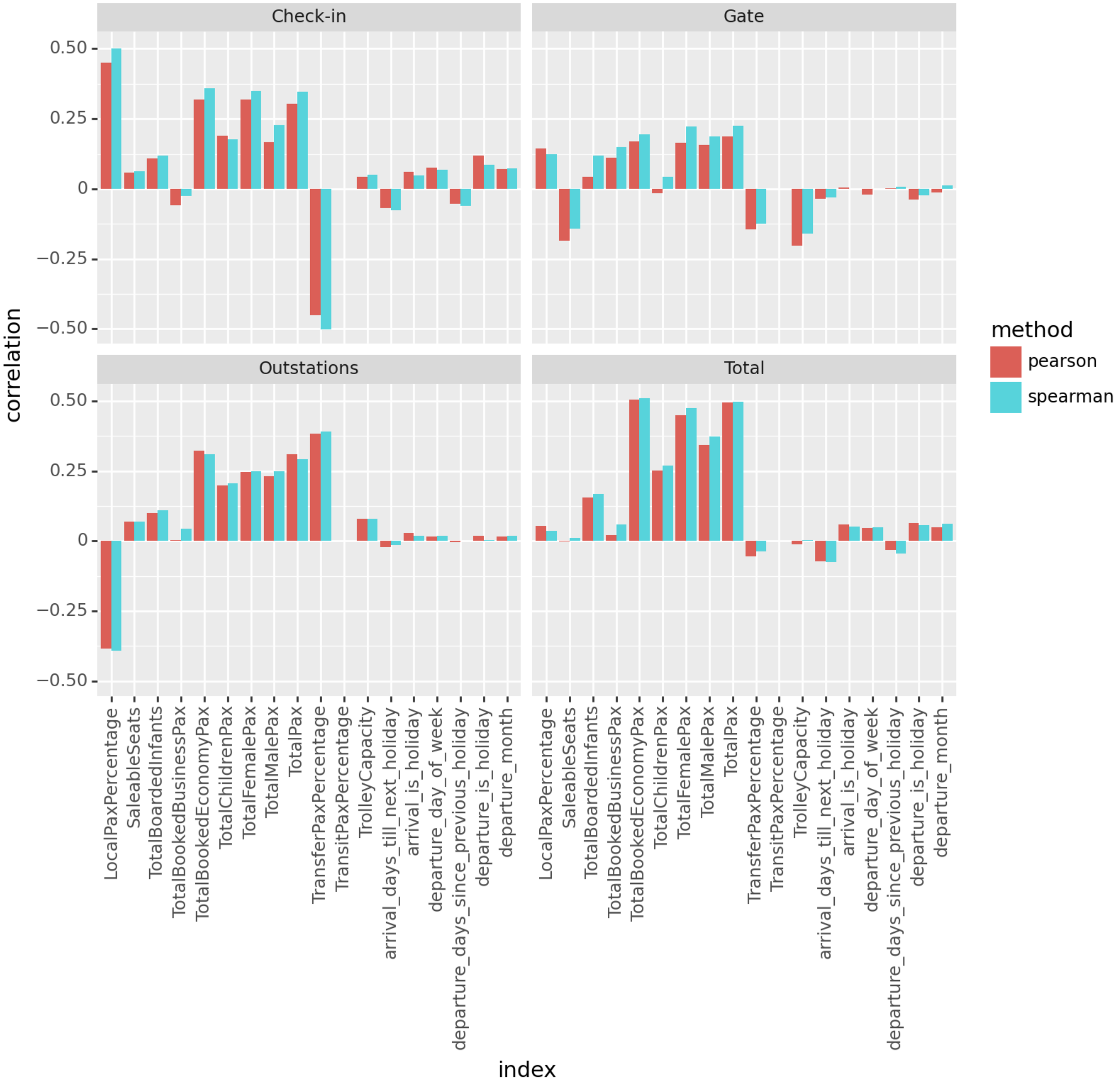


Figure A.7: Correlation matrix for Spearman correlation for KLC flights in 2023 departing from Schiphol.

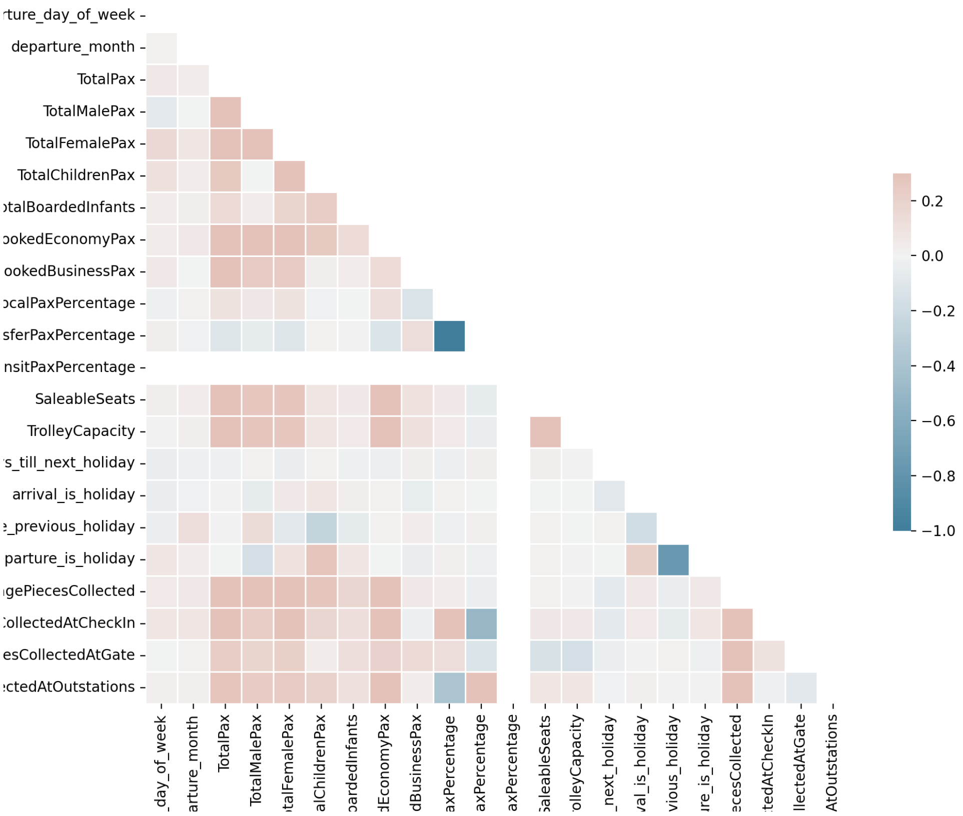


Figure A.8: Correlation matrix for Spearman correlation for KLC flights in 2023 arriving at Schiphol.

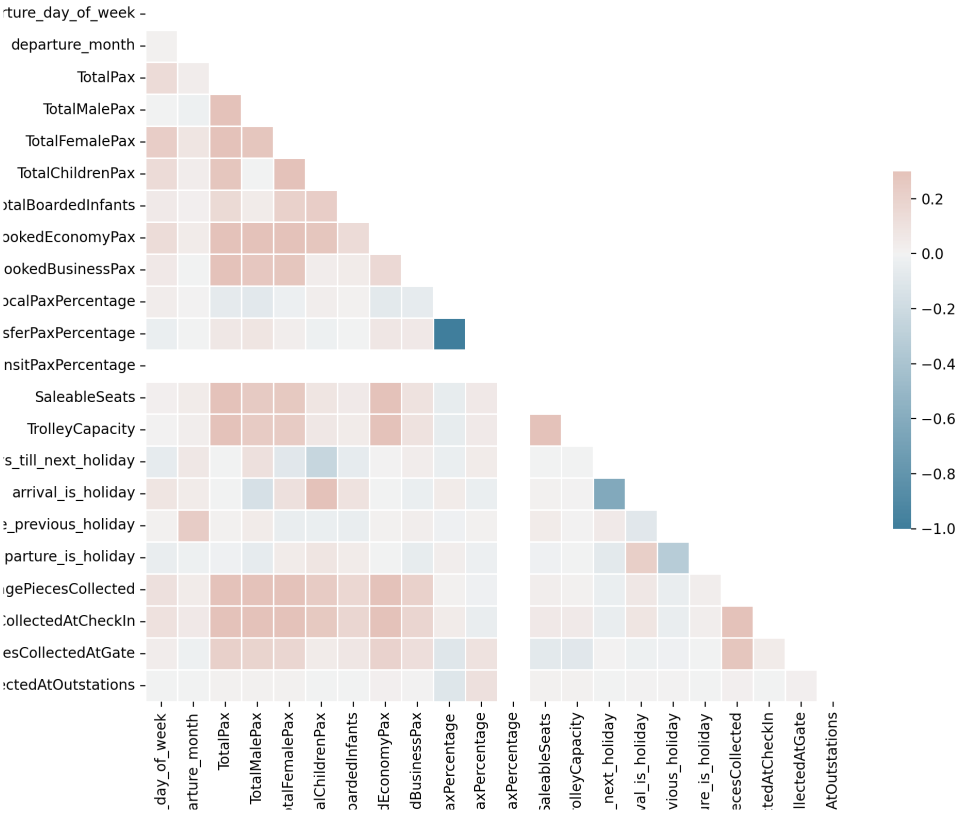
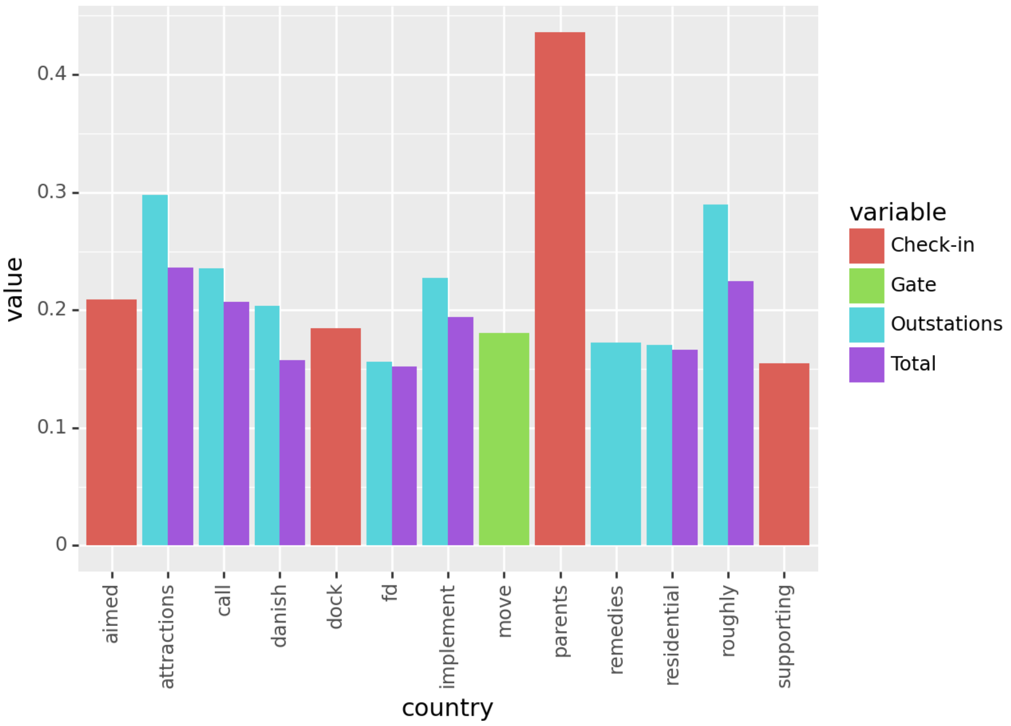


Figure A.9: Spearman correlation results for ICA flights in 2023 departing from or arriving at AMS. Country names have been replaced by random words due to confidentiality. Only countries with a correlation ≥ 0.15 are shown.

(a) Departing from AMS.



(b) Arriving at AMS.

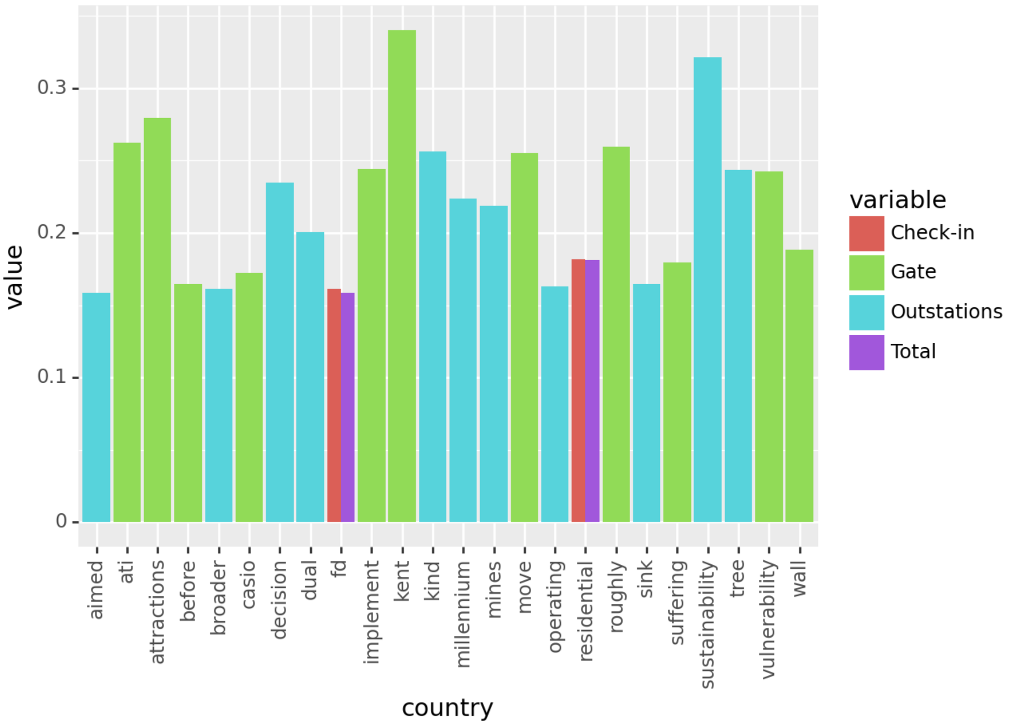
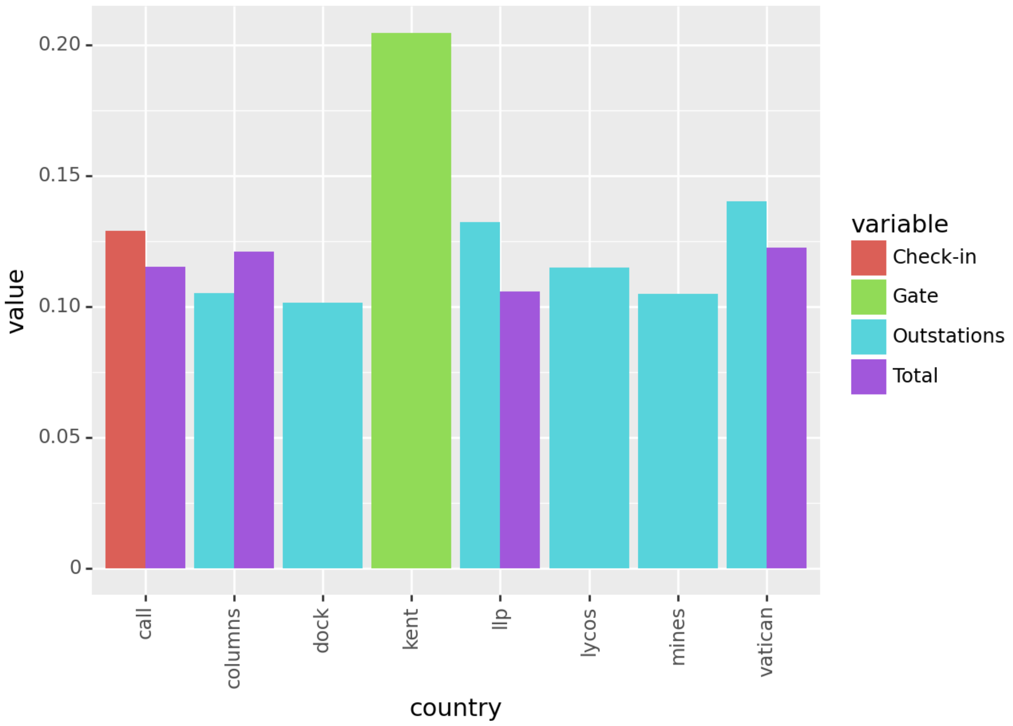
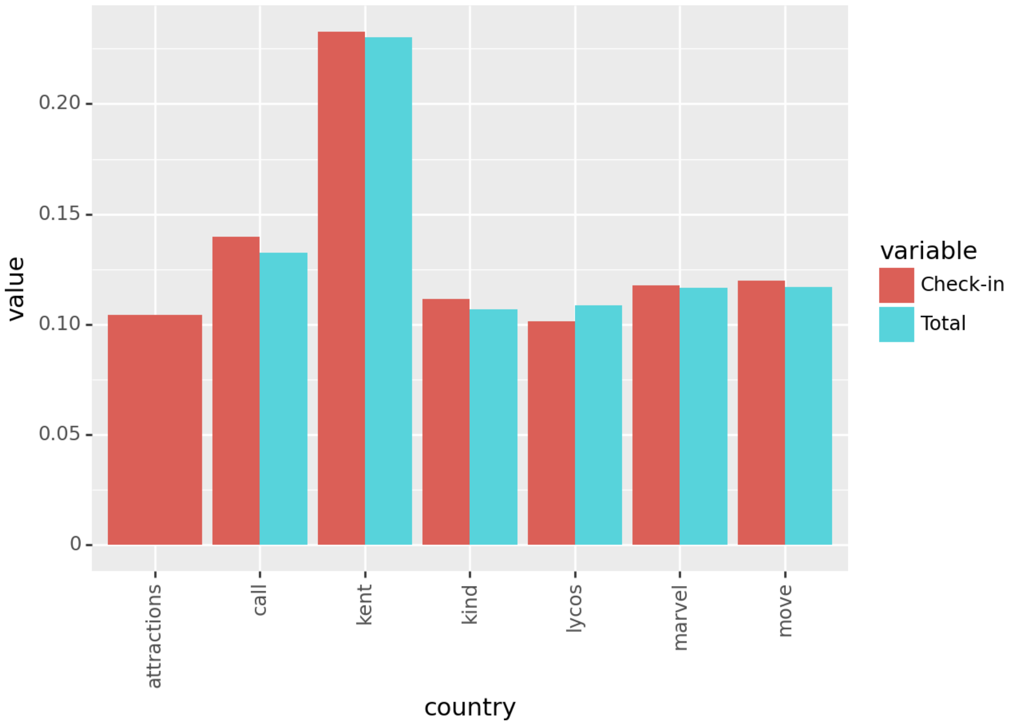


Figure A.10: Spearman correlation results for KLC flights in 2023 departing from or arriving at AMS. Country names have been replaced by random words due to confidentiality. Only countries with a correlation ≥ 0.10 are shown.

(a) Departing from AMS.



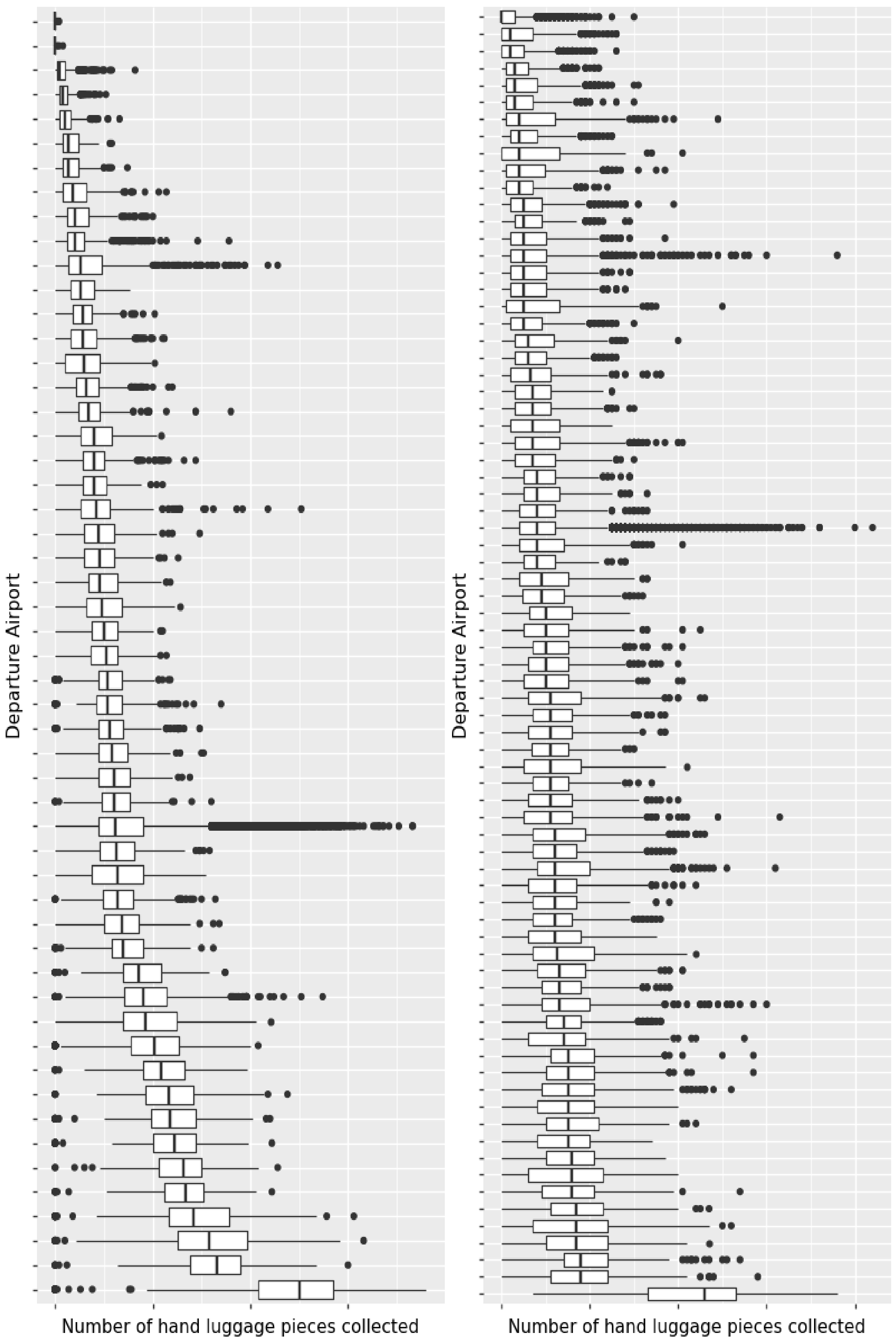
(b) Arriving at AMS.



Appendix B

Data analysis results from other datasets

Figure B.1: Number of hand luggage pieces collected per airport for ICA and KLC flights in 2023. Only airports with more than 200 flights in 2023 shown.



(a) Plot for the ICA flights.

(b) Plot for the KLC flights.

Figure B.2: Number of hand luggage pieces collected per continent for KLM intercontinental flights in 2023.

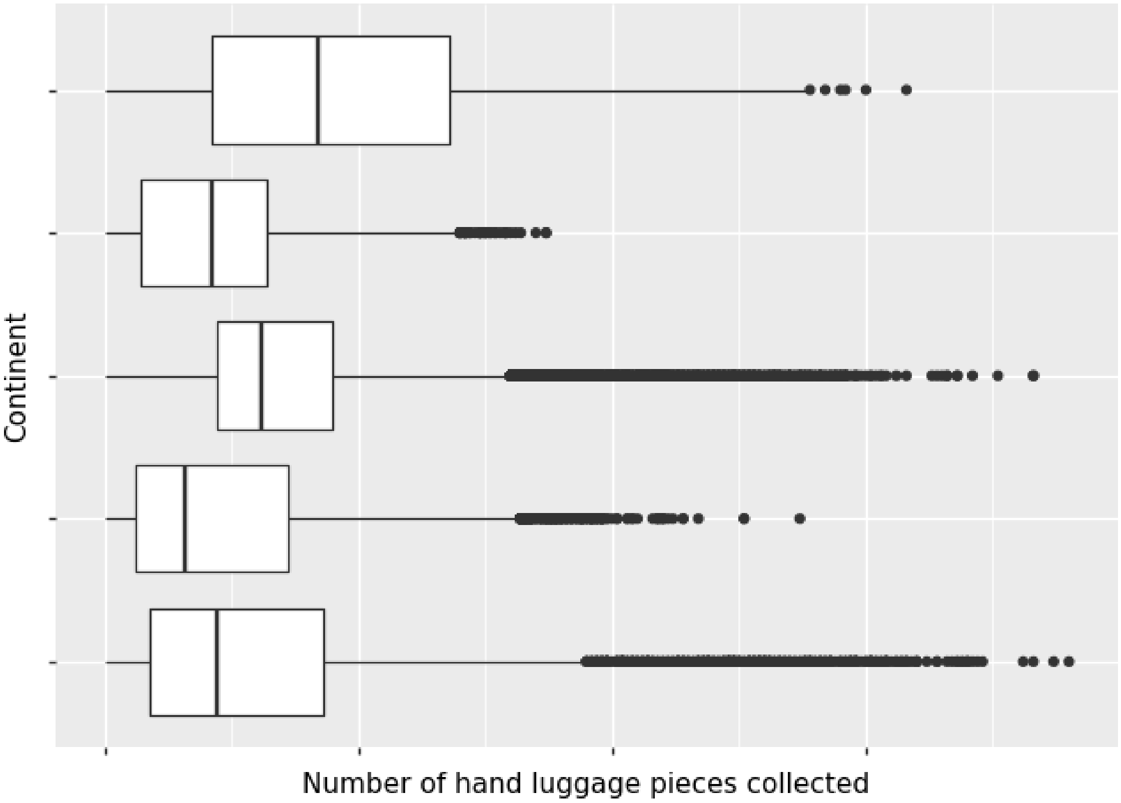
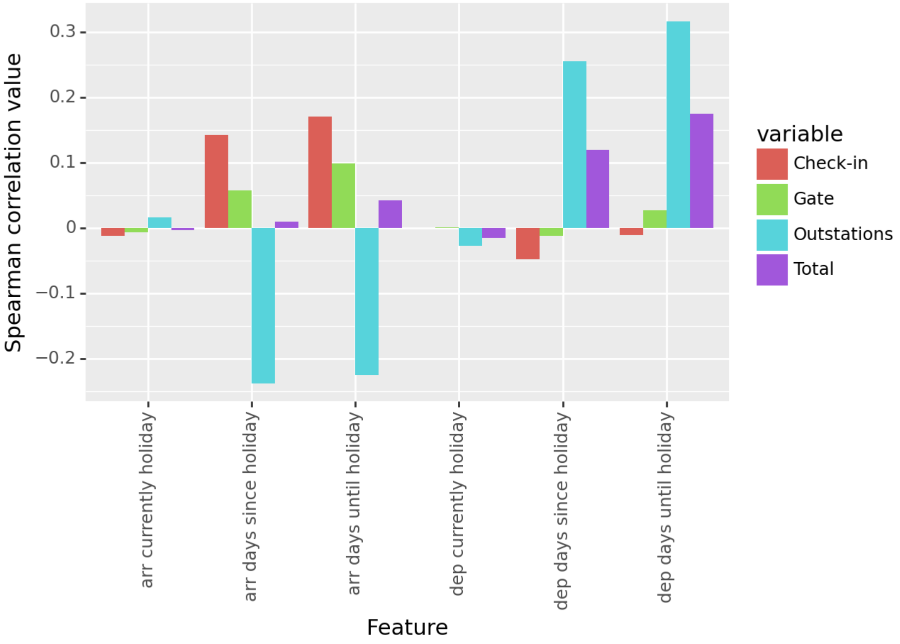


Figure B.3: Correlation of the national holiday dataset against the collected hand luggage features for ICA and KLC flights flown in 2023.

(a) Plot for ICA flights.



(b) Plot for KLC flights.

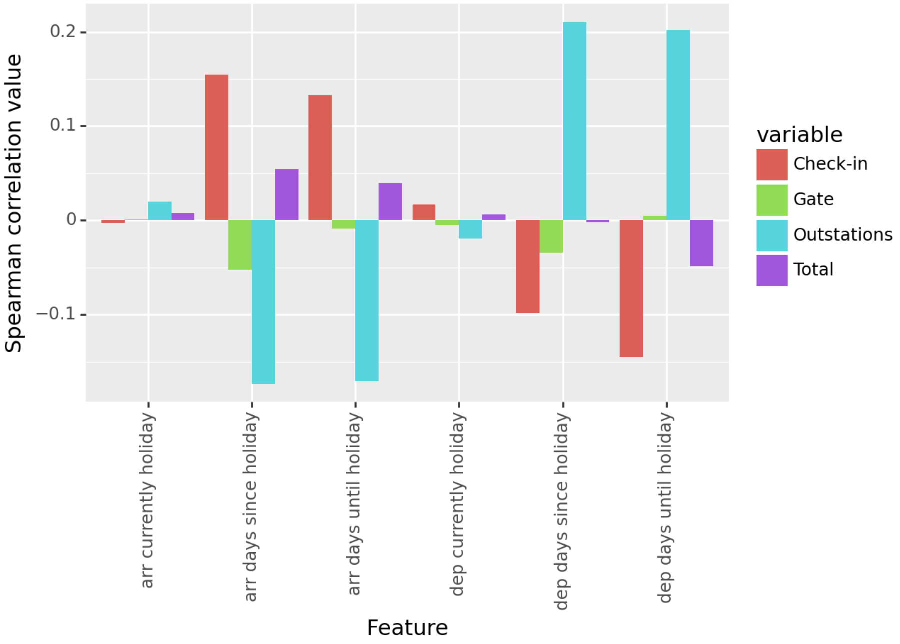
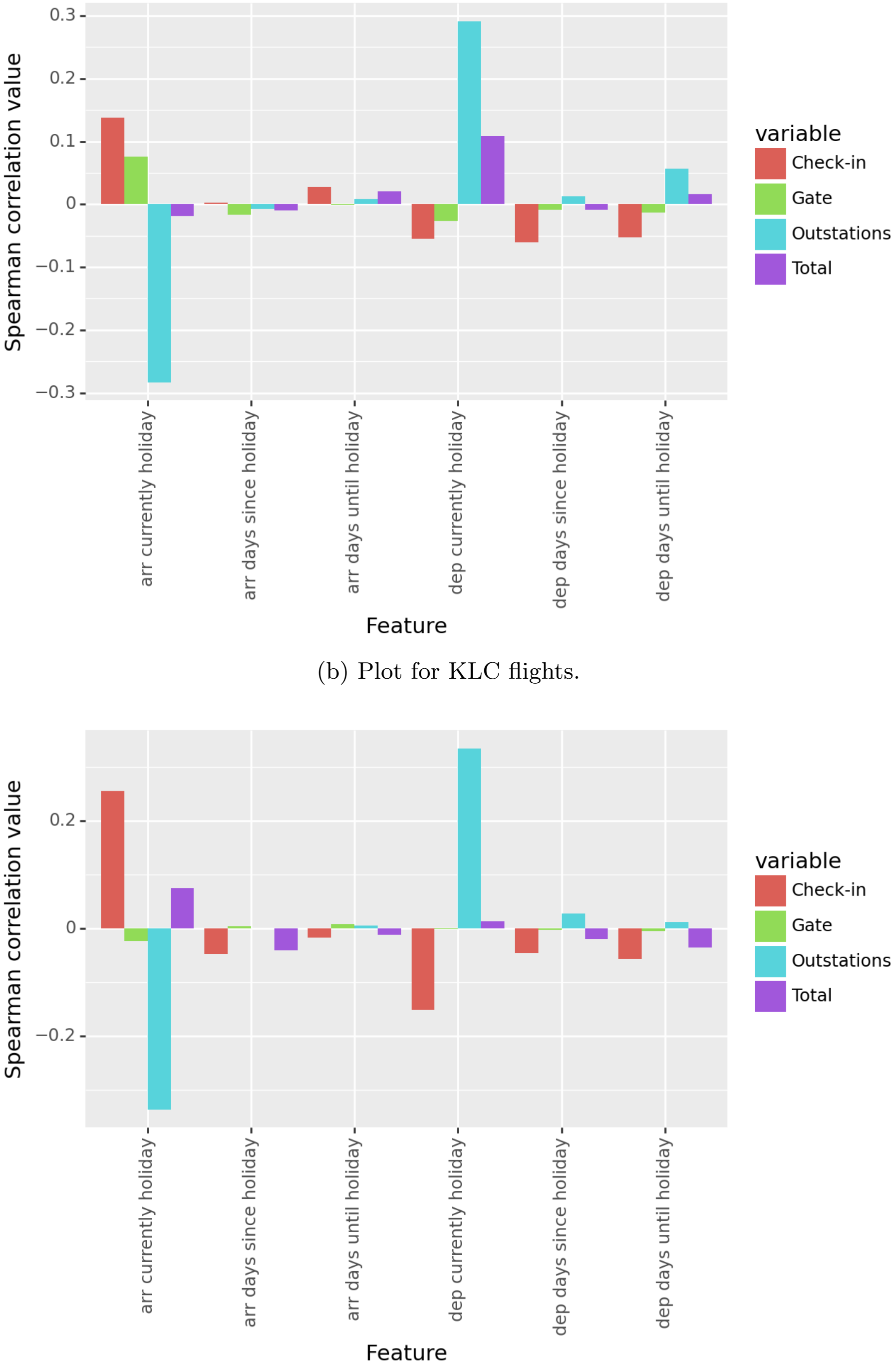


Figure B.4: Correlation of the Dutch school holiday dataset against the collected hand luggage features for ICA and KLC flights flown in 2023.

(a) Plot for ICA flights.



(b) Plot for KLC flights.

Appendix C

Experiment results

Table C.1: Hyperparameter Tuning results for the single-output models.

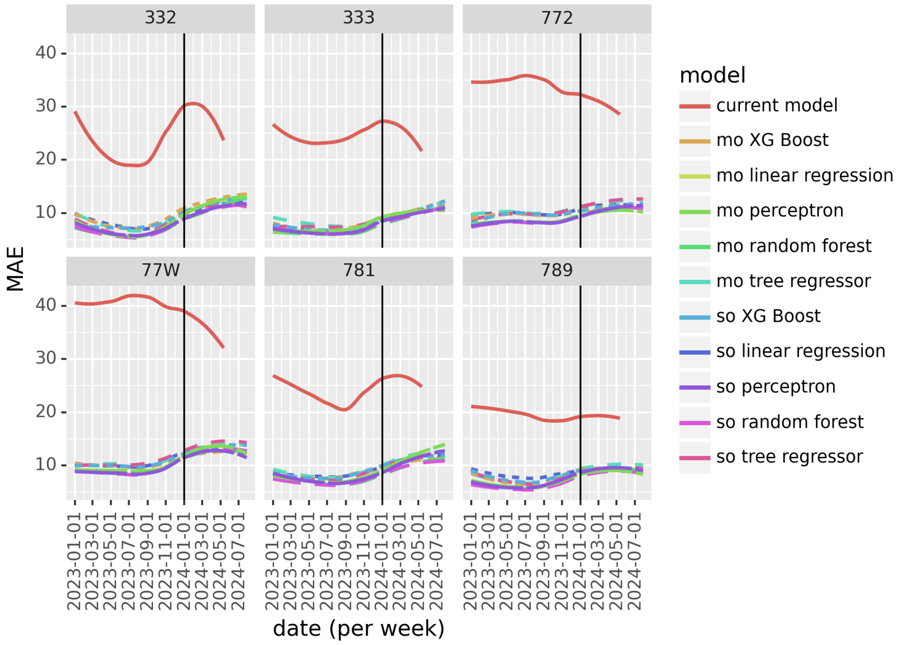
Model	Feature	ICA	KLC
Ridge Regression	alpha	1	1
Decision Tree Regressor	min samples split	100	30
	min samples leaf	10	10
	max depth	None	20
	ccp alpha	0.2	0.1
Random Forest Regressor	n estimators	140	160
	min samples split	3	2
	min samples leaf	3	200
	max depth	20	10
	ccp alpha	0	0
XG Boost	learning rate	0.1	0.1
	n estimators	90	130
	max depth	10	5
	min samples leaf	20	10
	min samples split	20	100
	ccp alpha	0	0
Multilayer Perceptron Regressor	solver	adam	adam
	activation	relu	relu
	layer sizes	(16,)	(32,)
	alpha	0.001	0.01
	learning rate init	0.001	0.001
	batch size	32	32

Table C.2: Hyperparameter Tuning results for the multi-output models.

Model	Feature	ICA	KLC
Ridge Regression	alpha	1	1
Decision Tree Regressor	min samples split	20	6
	min samples leaf	25	10
	max depth	15	10
	ccp alpha	0	0
Random Forest Regressor	n estimators	140	170
	min samples split	10	3
	min samples leaf	15	11
	max depth	25	40
	ccp alpha	0	0
XG Boost	learning rate	0.4	0.2
	n estimators	80	150
	max depth	30	5
	min samples leaf	20	20
	min samples split	20	10
	ccp alpha	0.1	0
Multilayer Perceptron Regressor	solver	adam	adam
	activation	relu	relu
	layer sizes	(32,)	(48,)
	alpha	0.01	0.01
	learning rate init	0.001	0.001
	batch size	16	32

Figure C.1: Mean Absolute Error over time for the trained models. Data from the ICA flights from 1st of January 2023 up until 31st of July 2024. Each subplot covers one aircraft type. The line shows the end-of-training-data date which means that the predictions on the left side are of the 30% test data and the predictions on the right side are from 100% of the ICA flights on which the model has not been trained.

(a) Plot for ICA flights.



(b) Plot for KLC flights.

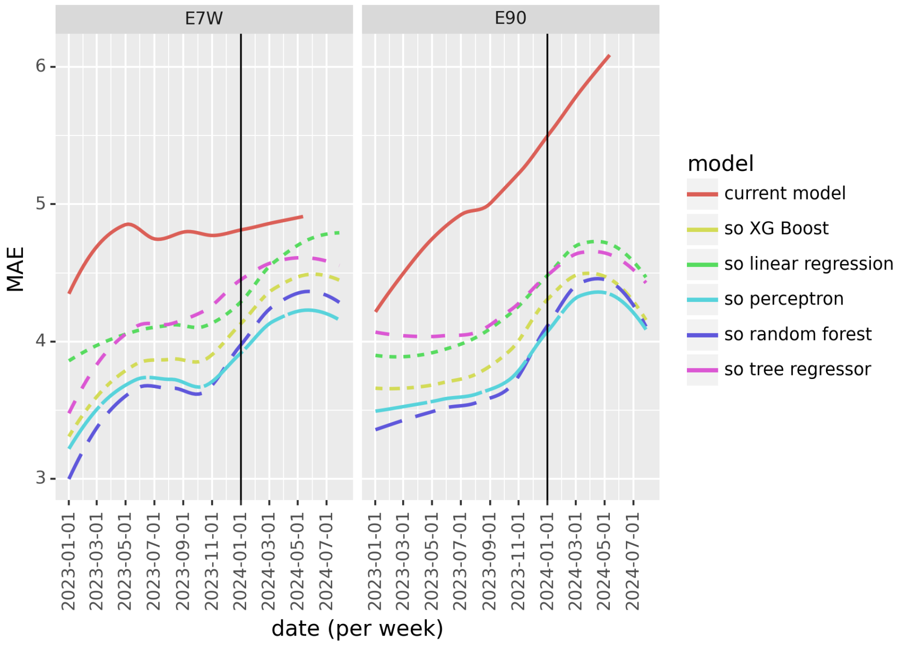
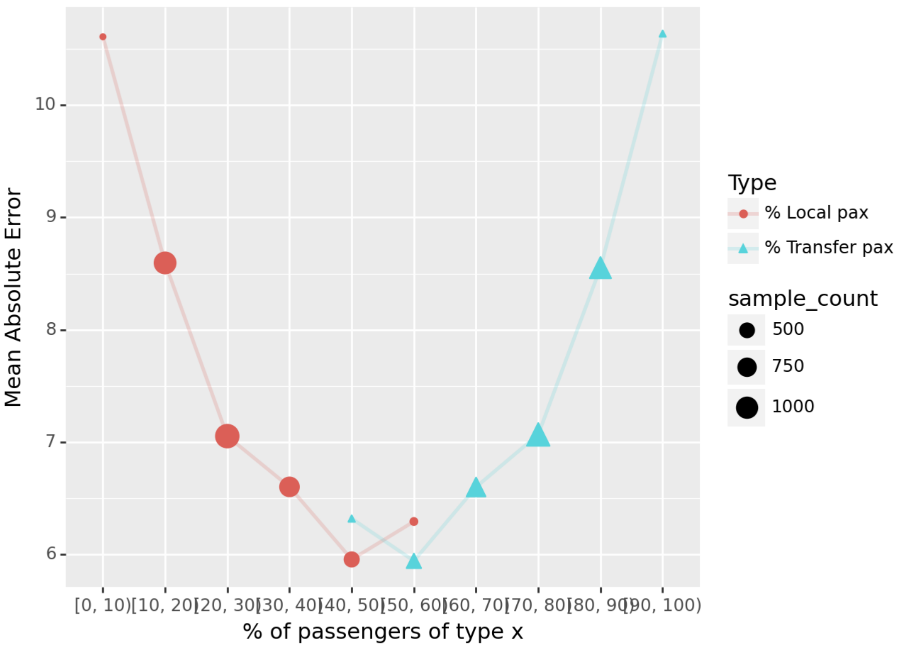


Figure C.2: The MAE of the multi-output Random Forest Regressor for ICA and KLC flights in 2023, departing from Schiphol, split by passenger composition. Each decile shows the mean absolute error of all flights within that decile. Only deciles that contain more than 2% of the data are shown. The transit passenger group is omitted since all data was grouped in the same decile causing the other points to become very small.

(a) Plot for ICA flights.



(b) Plot for KLC flights.

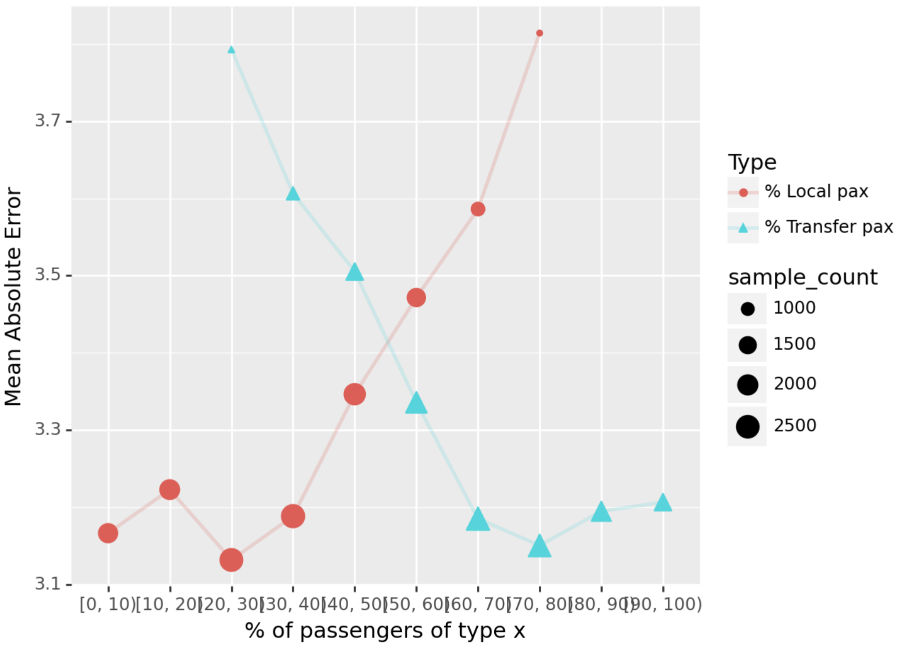
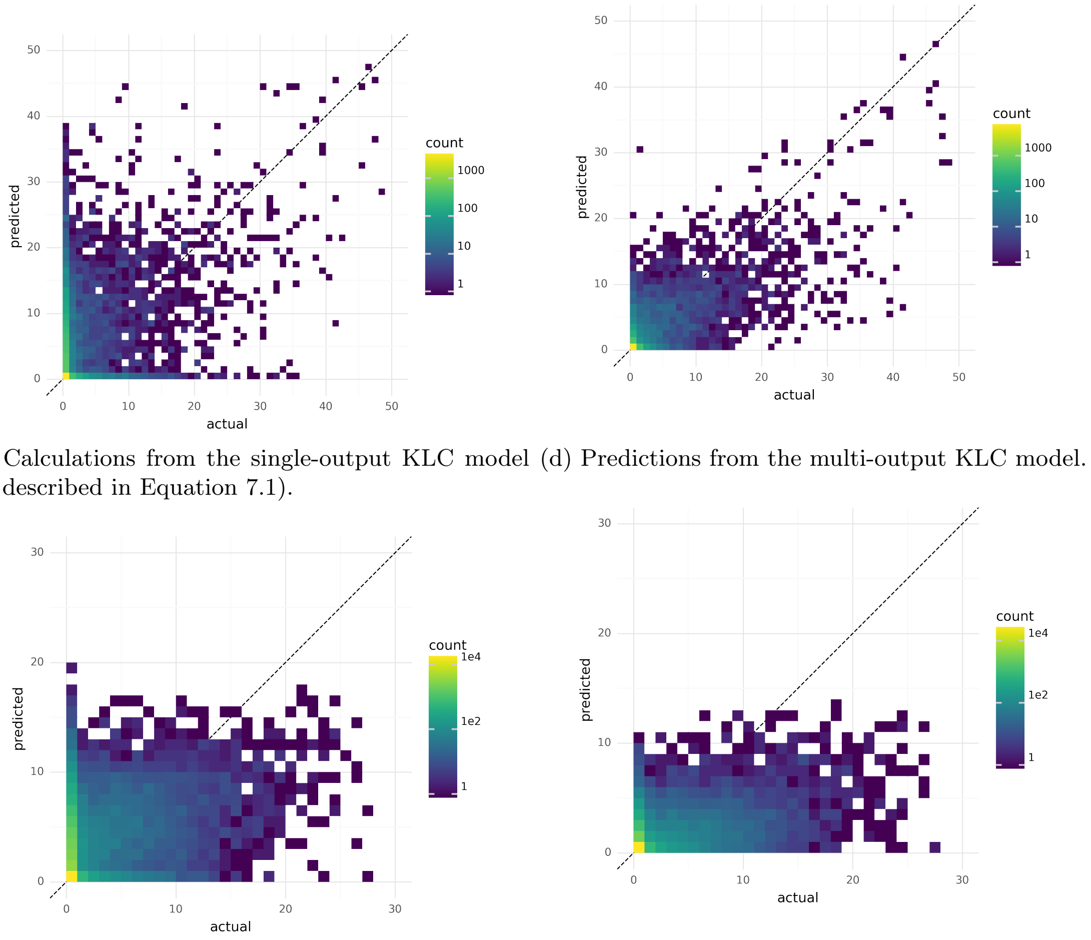


Figure C.3: Bin2d plots which show the gate prediction results of the random forest for both single-output and multi-output models for 2023 ICA flights.



(a) Calculations from the single-output ICA model (as described in Equation 7.1).

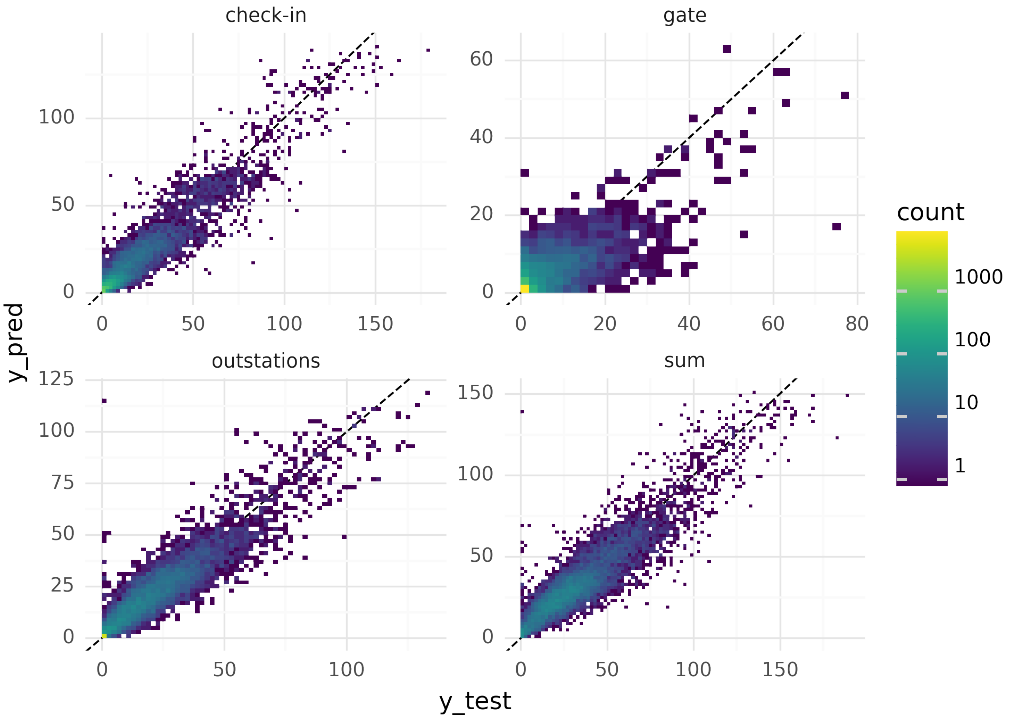
(b) Predictions from the multi-output ICA model.

(c) Calculations from the single-output KLC model (as described in Equation 7.1).

(d) Predictions from the multi-output KLC model.

Figure C.4: Bin2d plots for all collection moments from the best performing model: the multi-output random forest.

(a) Results from the ICA model.



(b) Results from the KLC model.

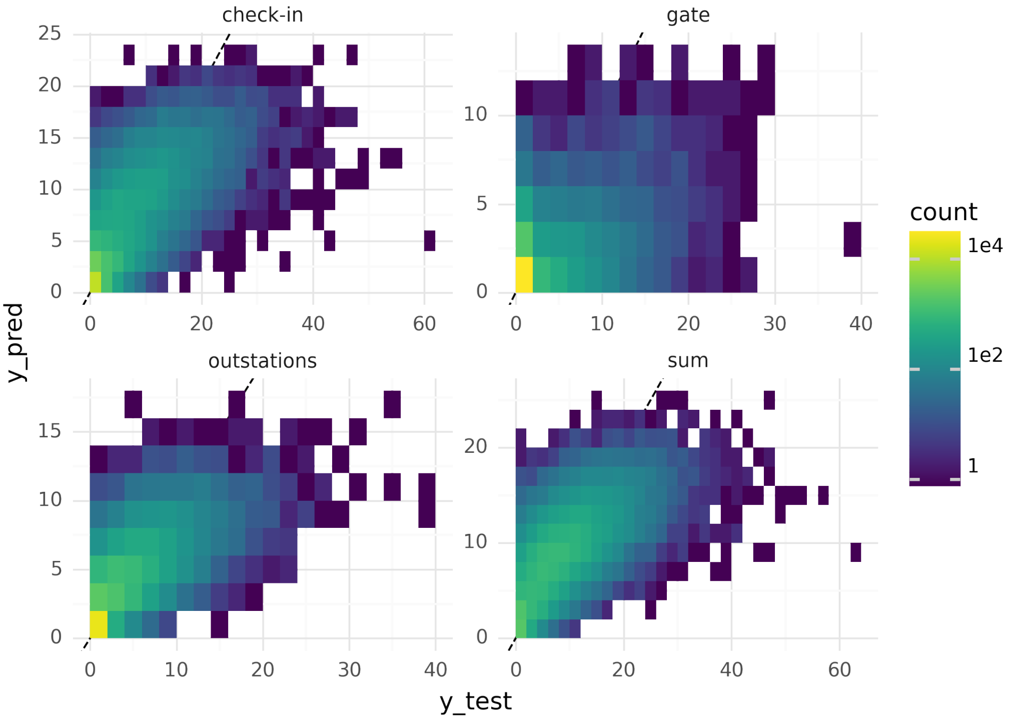
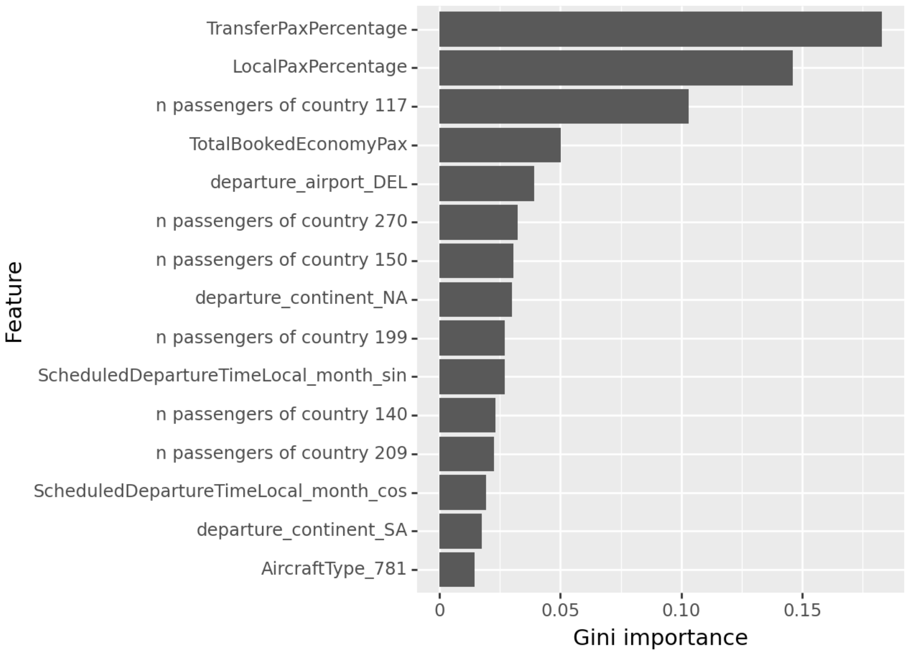


Figure C.5: Feature importance plot calculated using the Gini importance method. This method shows the (normalized) total reduction of impurity of a feature. The calculations are done on the random forest model trained on 2023 flight data.

(a) Results for the ICA model.



(b) Results for the KLC model.

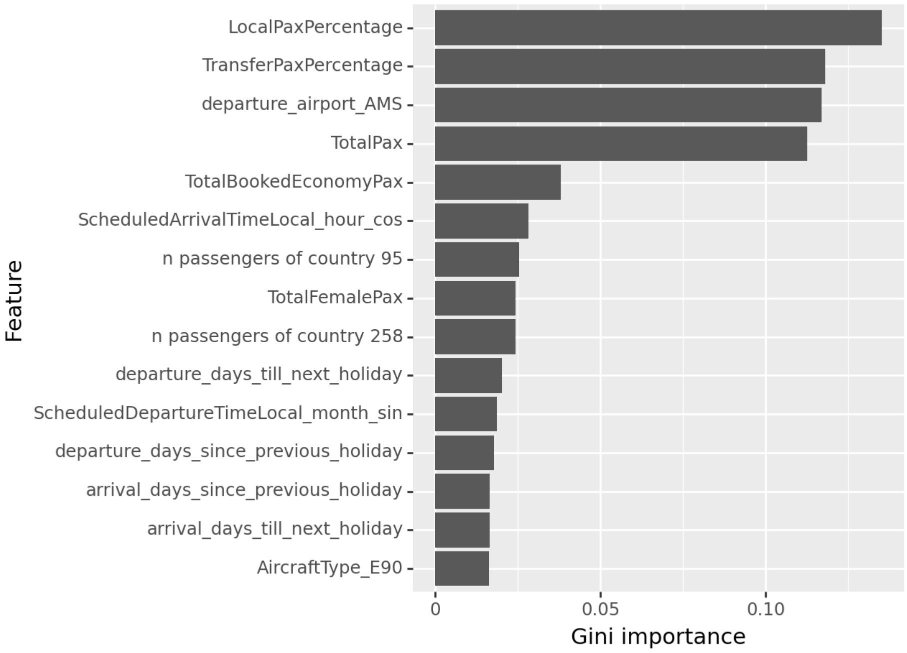
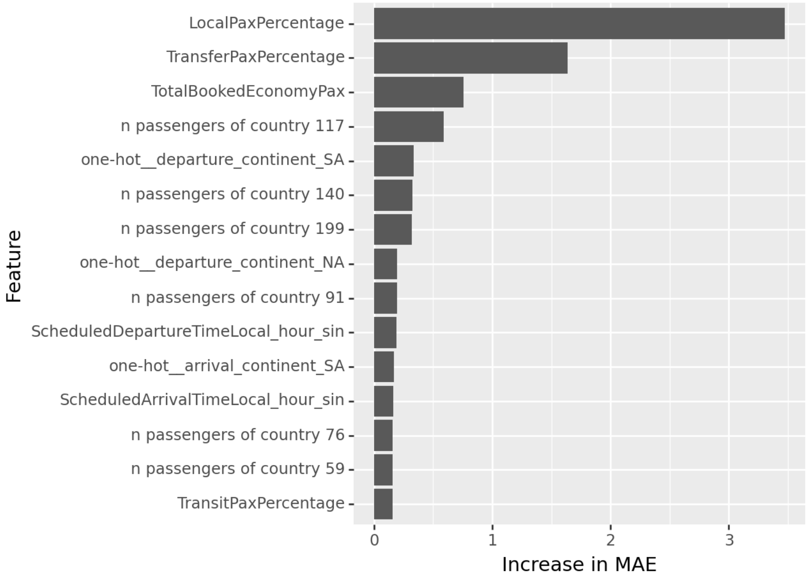


Figure C.6: These figures show the feature importance of the multi-output random forest models. The length of the column is the increase in the MAE score if the values (in the test set) of that feature got randomized.

(a) Results for the ICA model.



(b) Results for the KLC model.

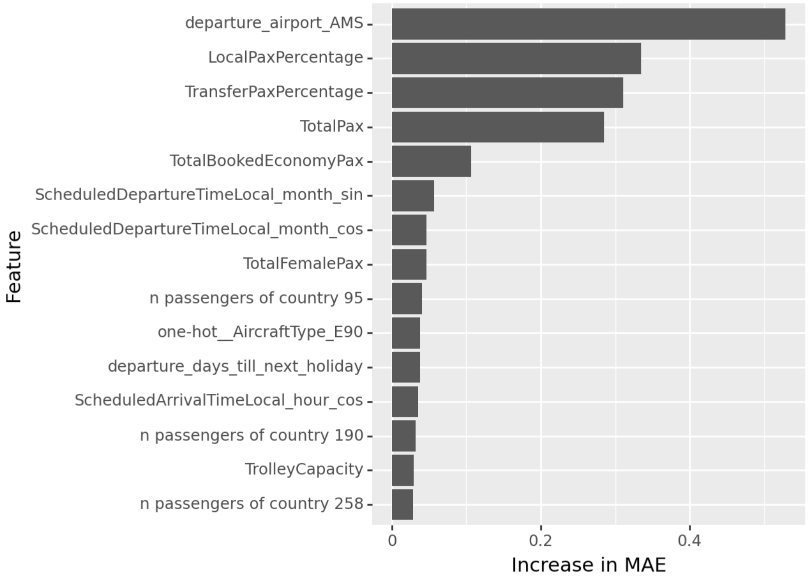
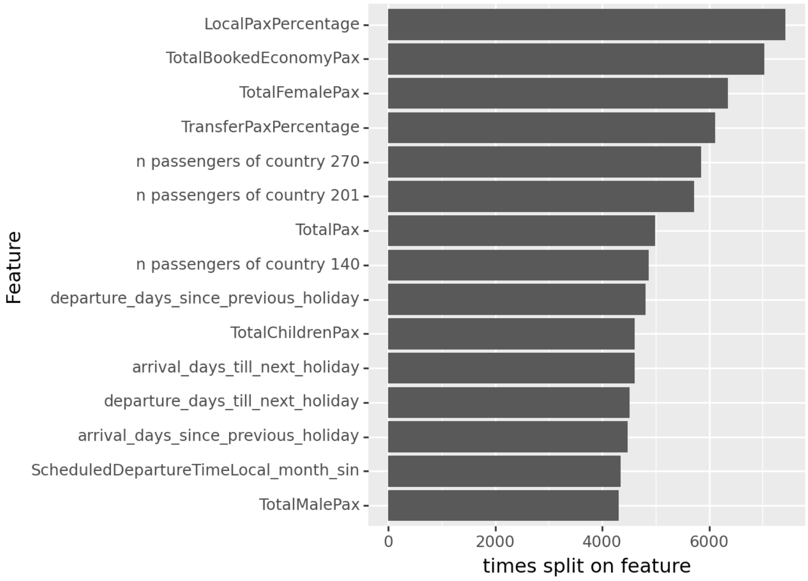
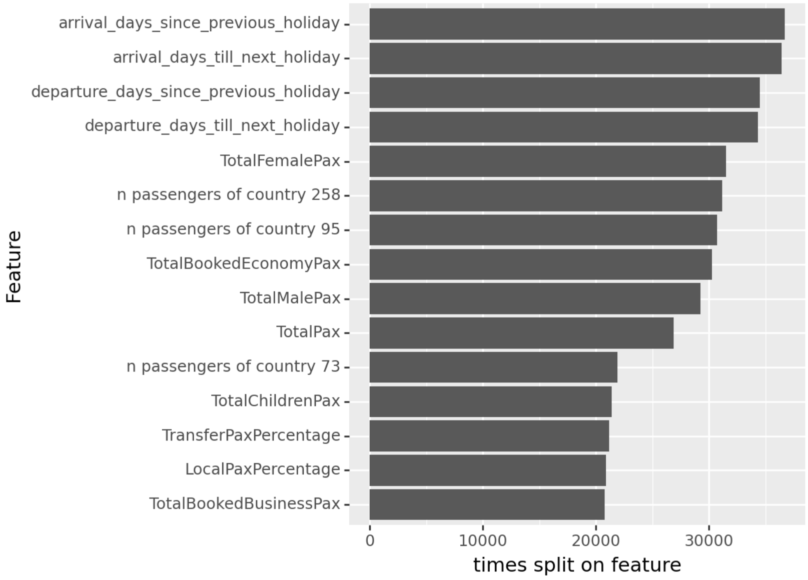


Figure C.7: These figures show the feature importance using the number of times a feature has been used to split in one of the trees. These results are from the multi-output random forest models.

(a) Results for the ICA model.



(b) Results for the KLC model.


arrival_days_since_previous_holiday
arrival_days_till_next_holiday
departure_days_since_previous_holiday
departure_days_till_next_holiday
TotalFemalePax
n passengers of country 258
n passengers of country 95
TotalBookedEconomyPax
TotalMalePax
TotalPax
n passengers of country 73
TotalChildrenPax
TransferPaxPercentage
LocalPaxPercentage
TotalBookedBusinessPax
Feature
times split on feature