

Quantile-based Machine Learning Predictions of Luggage Arrival Time at Airport Reclaim Areas

A Data-Driven Approach Using Operational Data from Munich Airport

Paolino De Falco
EUROCONTROL
Bretigny-Sur-Orge, France

Pablo Gasco Casado
EUROCONTROL
Bretigny-Sur-Orge, France

Ivan Baruta
EUROCONTROL
Bretigny-Sur-Orge, France

Hanna Wulkow
Munich Airport
Munich, Germany

Abstract—Luggage waiting time in reclaim areas significantly influences the arrival passengers' experience at airports. This study leverages data from Munich Airport to develop quantile-based machine learning models that predict the time from the arrival of the aircraft to the time the first and last bags appear on the baggage reclaim belt, based on timestamps manually recorded by ground handlers. The ultimate aim is to enhance the passenger journey by sharing these predictions in real time, enabling passengers to make decisions on the use of their time while waiting. To account for the fact that the timestamp of the first bag does not always reflect the actual start of the luggage delivery process, often marking the deposit of only a few bags before the main set, quantile predictions above the 50th percentile were used to intentionally overestimate luggage delivery times. This approach would also ensure that passengers experience luggage delivery earlier than the predicted time communicated to them, potentially enhancing their overall perception of service quality. A set of developed models shows encouraging performance, indicating potential for providing passengers with timely and realistic predictions of luggage delivery time through airport displays or mobile apps. The next phase of the project will involve validation activities and post-deployment surveys to evaluate potential improvements in passenger satisfaction. The work has been conducted as part of the PRELUDE project, one of the recent EUROCONTROL initiatives.

Keywords—Luggage; airport; passenger experience; machine learning; quantile regression

I. INTRODUCTION

Air transportation has become a crucial pillar of modern global mobility, with airports acting as key hubs that link regions, drive economies, and connect people. Within the broader framework of Air Traffic Management (ATM), the passenger has increasingly become the focal point of operational strategies, driving a shift from purely capacity-oriented approaches toward service quality and experience-oriented performance metrics [1].

Airports today are complex, high-throughput ecosystems that integrate the operations of multiple stakeholders, such as airlines, ground handlers, security agencies, and air traffic control, into tightly coupled processes [2]. With growing passenger volumes and tighter turnaround schedules, even minor inefficiencies in baggage handling can propagate into broader operational disruptions. The smooth movement of passengers and baggage relies on the coordinated planning

of all airport stakeholders and could be improved by accurate data-driven forecasting. Consequently, the ability to reliably estimate luggage delivery times is not only a logistic necessity but also a strategic enabler for passenger-centric performance management in ATM.

Among key touchpoints in the passenger journey, timely delivery of checked luggage plays a crucial role, influencing both perceived service quality and overall airport efficiency [3]. Indeed, delays or uncertainty in baggage delivery can reduce overall airport efficiency and weaken the reliability of the air transport system perceived by passengers.

At Munich airport, particular emphasis is placed on enhancing the passenger experience through data-driven services. In this context, the baggage claim area represents a key touchpoint, where passengers are provided with dynamically updated predictions of luggage delivery times on existing monitors, generated from real-time ground handling data. This allows passengers to decide how to spend their time while waiting. However, in certain situations, such real-time data may be unavailable or incomplete. In these cases, a fallback prediction model is required to ensure that passengers still receive an estimated waiting time for their luggage.

Recent advances in data science have made machine learning (ML) an increasingly adopted tool in air transport operations [4]. ML models are becoming embedded in the decision support frameworks of airports and ATM stakeholders. In particular, probabilistic and distributional prediction methods offer significant advantages over traditional point estimates. By producing a range of possible outcomes along with their associated probabilities, these methods enable decision makers to manage uncertainty, allocate resources more effectively, and provide passengers with more accurate information. Additionally, quantile-based approaches allow predictions to be systematically biased upward or downward. For example, using quantiles above the 50th percentile results in deliberate overestimation of the target values.

By leveraging historical operational data, this study focuses on the development and evaluation of quantile-based machine learning models to predict two key time intervals in airport baggage handling: (1) the time difference between aircraft arrival and the first piece of luggage being placed on the



baggage carousel, and (2) the time difference between the delivery of the last and first pieces of luggage. The combination of these two predictions make it possible to infer the expected arrival time of the last bag. In baggage handling operations, the timestamp of the first bag deposited onto the carousel or belt does not always represent the true beginning of the luggage delivery process. In many cases, this timestamp reflects only the deposit of a small number of bags, before the bulk of the luggage arrives. To account for this practice, our approach models the 50th quantile as representative of typical operational performance, whereas higher quantiles are employed to deliberately overestimate delivery times and thus better approximate the typical passenger experience. This conservative prediction strategy serves a dual purpose: it reduces the likelihood of underestimating waiting times, and it can positively influence passenger perceptions. Indeed, once the solution is implemented and passengers can visualise the predictions, deliveries that occur earlier than the predicted could enhance passengers' perception of the ground handlers' efficiency.

This work is part of PRELUDE (PREdiction of LUGgage DELivery time) project, one of the latest EUROCONTROL Air Transport Innovation Network (EATIN) initiatives, which prioritise working closely with end-users, such as airports, air navigation service providers (ANSPs) and airspace users. The EATIN projects are in-kind contributions to the SESAR program and focus on the development of agile solutions that address operational challenges and deliver operational benefits. Munich Airport is actively involved in this innovative project aimed at enhancing its fallback model used for luggage delivery time predictions in case of missing real-time information.

The paper is structured as follows: Section II presents a literature review of previous studies addressing passenger experience and airport-related topics. Section III details the algorithms used to develop the models, while Section IV presents the key results. Finally, Section V discusses the findings and draws conclusions.

II. LITERATURE REVIEW

Airports represent critical nodes in the air transport system, where operational efficiency intersects with passenger experience. Research on airport passenger experience has grown steadily, recognising that satisfaction is shaped not only by functional service delivery but also by the psychological and emotional states of travelers. Studies have identified key determinants of satisfaction such as cleanliness, seating availability, signage, shopping options, luggage claim and security procedures [5], with queue times emerging as a particularly influential factor in shaping perceptions [6].

Based on content analysis of approximately 1000 traveler comments posted between 2010 and 2013 [5], baggage handling emerged as the second most frequently referenced airport service attribute. Sentiment analysis revealed a mixed perception, with 40% of mentions in a satisfactory context and 60% in a dissatisfying one, indicating that baggage services are highly performance sensitive. In a recent study [7],

1200 TripAdvisor reviews were analysed to explore travelers' experiences with air transport services. The analysis revealed that airport operations, including check-in, baggage handling, and boarding gates, are key factors that influence overall passenger satisfaction. Surveys are another important method for gathering passenger feedback and are widely used to evaluate airport satisfaction and service quality. Numerous studies have performed customer surveys to identify key areas for improvement. For example, it was found [8], through passenger surveys at Lamezia Terme Airport (Italy), that cleanliness should be a priority for improving service quality. Another study [9], performed through face-to-face interviews, highlighted that satisfaction levels vary with passengers' background. Similarly, from an analysis of 937 surveys from Riga International Airport (Latvia) [10], it was found that effective signage, clean restrooms, and staff courtesy are key determinants of passenger satisfaction. Taken together, these findings suggest that while surveys often emphasise hygiene and service delivery, online reviews tend to highlight baggage handling and queuing, pointing to a potential methodological bias depending on the type of data source.

Given the recurrent challenges identified in both surveys and online reviews, technological solutions have been proposed to mitigate passenger dissatisfaction, particularly in baggage handling. There is an increasing interest in providing real-time baggage insights to passengers, showing the location and expected arrival time of their luggage. RFID-based technologies, for example, enable real-time tracking and detailed information on baggage handling progress. While they do not directly provide delivery time predictions, the data they generate can be used to feed predictive models, resulting in more reliable estimates for passengers [11]. These systems offer high accuracy and reduced handling effort compared to barcodes, but they also face challenges such as high upfront costs, signal interference (particularly from metal or liquids), and integration difficulties with existing airport systems [12]. Recent camera-based systems, instead, detect and interpret various aircraft turnaround events, which, combined with operational data, enable predictive models to estimate baggage delivery times [13]. These emerging technologies show a clear shift from descriptive measurement of service quality toward predictive passenger information systems.

With passengers increasingly recognising the capacity of artificial intelligence (AI) to improve efficiency, minimise disruptions, and support sustainable operations [14], applying AI models to enrich the information provided to travelers appears to be a logical approach to enhance not only operational decision-making but also improve airport passenger experience and satisfaction. The integration of ML into ATM systems is rapidly advancing, offering improved predictive capabilities, optimised traffic flows, and enhanced operations [15], [16]. Probabilistic ML approaches are essential in ATM to handle the inherent unpredictability of flight operations, enabling risk-aware planning and decision making under uncertain conditions [17]. Various approaches have been proposed in the literature to provide individual ML predictions along with an



associated measure of uncertainty [18], [19]. As an example, quantile regression can be used to estimate a range of potential values for a given prediction as a measure of uncertainty [20]. A recent study [21] tackled the challenge of predicting flight delays a few days before departure by employing a quantile machine learning model with a pinball loss function to generate probabilistic prediction intervals. The work emphasised that for low quantiles (e.g., 10th percentile), the pinball loss penalises overestimation more than underestimation, so the model tends to produce predictions below or near the target. Conversely, for high quantiles, the model tends to overestimate rather than underestimate the target.

III. METHODOLOGY

In this section, the input data used for model development are introduced, together with the description of the quantile machine learning algorithm implemented for two sets of models: one predicting the luggage delivery time, defined as the interval between the arrival of the first bag on the baggage belt and the actual in-block time (AIBT) of the aircraft, and another predicting the loading time, defined as the difference between the arrival of the last and the first bags. For simplicity, the first model will be referred to as *1st bag* and the second one as *Loading* or *Load*.

A. Input Data

A dataset provided by Munich Airport, covering the full year 2024 and the period from January through July 2025, was used for model development. Due to confidentiality constraints, the dataset used in this study cannot be publicly shared. However, detailed descriptions of the data and pre-processing steps are provided to support reproducibility. The selection of input variables available at AIBT, together with the definition of the target for the *1st bag* model, which incorporates the AIBT, ensures that the models can generate predictions immediately upon the aircraft's arrival at the stand.

Prior to model development, the dataset was preprocessed and cleaned to ensure consistency and completeness. Specifically, two filtering processes were implemented. The first involved removing errors in the data. To this end, samples where the time from AIBT to placing the first bag on the belt was equal to zero were removed. The same filter was also applied to the loading time between the first and last bag. A second filter was applied to address extreme values that could potentially bias the performance of the model. For instance, extreme cases where the time to place the first bag was either excessively low (less than one minute) or high (more than two hours) were removed. Finally, only observations between the 1st and 99th percentiles of the distributions of the two model targets were retained.

The set of input features that was selected for this study is described in Table I. It includes temporal, operational, and flight-specific variables. Temporal features comprise the hour, weekday, and month of arrival. Operational and passenger-related features include the number of normal and special luggage items, as well as the number of arriving passengers.

Flight-specific characteristics are captured by the aircraft type and operator, the departure airport, and whether the flight arrives from a country in or outside the Schengen area. Since the model is intended to generate predictions at the AIBT, the delay, computed as the difference between AIBT and the Scheduled In-Block Time (SIBT), was included as an input feature. Additional operational variables include the stand position (i.e., remote vs contact stand), the stand identifier and the assigned ground handling company. Features are represented as either categorical or numerical, as appropriate.

TABLE I. SET OF INPUT FEATURES USED FOR MODEL DEVELOPMENT.

Input Feature	Description	Type
hour, weekDay, month	Computed at AIBT	Cat
numBags	No. normal bags	Num
numSpecialBags	No. special bags	Num
numPax	Number of passengers	Num
aircraftType	IATA Code	Cat
aircraftOperator	IATA Code	Cat
groundHandler	Ground handler	Cat
adep	Dep. airport (IATA code)	Cat
standPosition	Stand position (or type)	Cat
schengen	Referred to dep. airport	Cat (Y/N)
stand	Stand identifier	Cat
delay	AIBT - SIBT	Num
Model target	Description	Type
firstLuggageTime	AIBT - first bag on belt	Num
loadLuggageTime	Last bag - first bag on belt	Num

B. Model evaluation metrics

To assess the performance of the predictive models described in Section III-C, three error metrics, widely used for regression problems, have been evaluated: Mean Absolute Error (MAE), Mean Error (ME), and Root Mean Squared Error (RMSE) [22].

The Mean Absolute Error (MAE) measures the average magnitude of the errors between predicted and observed values, without considering their direction. It is defined as shown in Equation 1, where $\hat{y}_i$ is the predicted value, y_i is the observed value, and n is the total number of observations.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (1)$$

The Mean Error (ME) captures the average signed difference between predictions and actual values (Equation 2). A positive ME indicates a tendency to overestimate the target values, whereas a negative ME indicates underestimation.

$$\text{ME} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i) \quad (2)$$

The Root Mean Squared Error (RMSE) quantifies the square root of the average squared differences between predicted and actual values (Equation 3). RMSE penalizes larger errors more heavily than MAE, making it sensitive to outliers.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3)$$

C. Quantile Machine Learning Regression

The algorithm for this study was selected by initially training and comparing three models based on LightGBM [23], XGBoost [24], and CatBoost [25]. The dataset and the model targets used for this comparison are summarised in Table I. The evaluation used the Mean Absolute Error (MAE) as the loss function, along with two performance metrics, MAE and Root Mean Squared Error (RMSE) [22], described in detail in Section III-B. CatBoost algorithm was selected due to its superior predictive performance, as shown in Table II. As a gradient boosting algorithm based on decision trees, it is particularly effective in handling categorical features and capturing complex nonlinear relationships through ordered boosting and regularisation [25].

TABLE II. COMPARISON BETWEEN DIFFERENT ALGORITHMS FOR A STANDARD REGRESSION APPROACH.

Model	MAE		RMSE	
	First	Load	First	Load
CatBoost	6.1	4.0	8.0	7.2
LightGBM	6.4	3.9	8.3	7.3
XGBoost	8.0	5.5	9.8	8.0

Upon selecting Catboost as the algorithm, two sets of models (namely the *1st bag* and *Loading* models) were trained to predict multiple quantiles of the two targets described in Table I. Quantile regression enables the estimation of conditional quantiles of the response variable, providing a detailed understanding of the distribution of arrival times beyond the mean. The models were trained using the pinball loss function, as shown in Equation 4, where y_i and $\hat{y}_i$ are the actual and predicted target values for a single observation, and $\alpha \in [0, 1]$ is the quantile to learn.

$$\text{Pinball}(y, \hat{y}, \alpha) = \alpha \max(y - \hat{y}, 0) + (1 - \alpha) \max(\hat{y} - y, 0) \quad (4)$$

The quantile level α is a model parameter allows the model to systematically overestimate or underestimate a given fraction of predictions within the training and testing datasets. For instance, assigning a value of 0.7 to α guarantees that approximately 70% of the observations are overestimated with respect to the target value. Figure 1 shows that predictions below the target value correspond to underprediction, while predictions above the target value correspond to overprediction. The pinball loss increases asymmetrically for under- and overpredicted values depending on the quantile, reflecting the quantile-specific weighting of residuals. Higher quantiles (e.g., 90th percentile) incur larger penalties for underprediction, whereas lower quantiles (e.g., 10th percentile) incur larger penalties for overprediction.

In this study, each of the two models were trained to predict the 50th, 70th, and 90th quantile values, ensuring that a specified proportion of observations in the dataset were consistently overestimated, for the operational considerations outlined in Section I.

It is also important to note that the selection of the training and test sets was not carried out according to a conventional

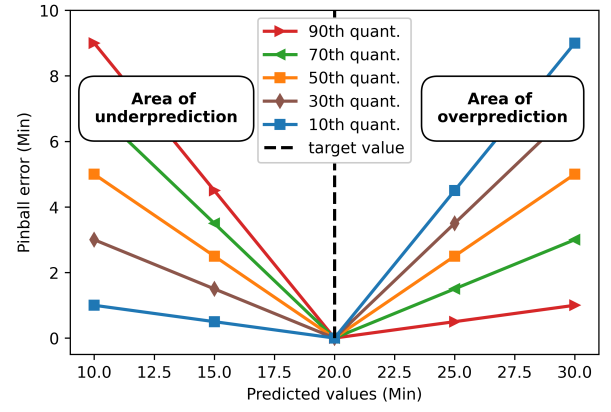


Figure 1. Pinball error plots illustrate the relationship between predicted values and quantile-specific errors for an example of a target value of 20 minutes.

80–20 split (respectively, indicating the percentage of observations in the training and testing datasets), but rather on a temporal basis. Specifically, the training set comprised all available data from 2024 in addition to the first three months of 2025 (up to 31 March). This approach ensured that the model was exposed to potential operational changes or procedural adjustments that may have been introduced at the beginning of 2025. The test set, in turn, consisted of data collected between 1 April and 31 July 2025, thereby enabling an evaluation of model performance on more recent and unseen observations.

With regard to hyperparameter tuning, a grid search procedure was employed in order to identify the most suitable configuration across multiple parameters, including the number of iterations, learning rate, and tree depth. Following this process, it was found that the optimal parameter set includes 300 iterations, a learning rate of 0.05, and a depth of 10. These hyperparameters were subsequently applied to all the models developed in this study.

IV. RESULTS

The results of the analyses are presented in this section, including the performance metrics of the machine learning model, findings from SHAP-based evaluations, and additional analyses in which observations were grouped by selected features to identify relevant patterns.

A. Main results

To assess the tendency of the model to systematically overestimate or underestimate predictions for specific quantile values, the cumulative distribution functions (CDF) of the prediction errors for the *1st bag* model were calculated (Figure 2). For this analysis, the errors, computed using the testing dataset, are first sorted in ascending order, after which the cumulative proportion of observations up to each error value is computed. This representation enables an intuitive interpretation of the accuracy of the prediction. As an example, for any of the CDFs, the proportion of errors that are under- or overestimated can be determined by projecting onto the y-axis the point corresponding to the curve at the zero error value.

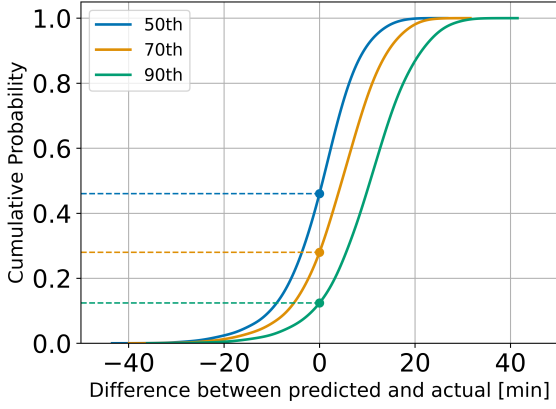


Figure 2. Cumulative distribution functions (CDFs) of prediction errors for the model at three quantiles: 50th, 70th and 90th.

Here, the curve for the 50th quantile is centered closest to zero, indicating that the median predictions exhibit the lowest overall bias. The curve for the 70th quantile crosses the y-axis at a cumulative probability of approximately 0.3, indicating that the model overestimates the target for roughly 70% of the samples. The curve representing the 90th quantile is further shifted to the right, with zero error occurring at an even lower cumulative probability (i.e. 0.1), reflecting that overestimation dominates for most samples at this higher quantile.

Table III presents the evaluation metrics across different quantiles (50th, 70th, and 90th) for the *1st bag* and *Loading* models. As expected, the MAE increases with higher quantiles, ranging from 6.1 to 11.7 minutes for the *1st bag* model and from 4 to 8 minutes for the *Loading* model, indicating slightly larger errors in predicting extreme values. The ME shows a trend from small negative values at the median (-0.1 and -1.7 minutes, respectively, for the *1st bag* and *Loading* models at the 50th quantile) to positive errors at higher quantiles (up to 10.1 minutes for the *1st bag* model and 6.1 minutes for the *Loading* at the 90th percentile), suggesting the expected tendency of the model to underestimate the target for lower quantile values and vice versa. RMSE increases similarly with quantile, from 8 to 13.6 minutes for the *1st bag* model and from 7.2 to 10.3 minutes for the *Loading* one, reflecting greater variability in the prediction errors in the upper quantiles.

TABLE III. KEY RESULTS FOR THE MODEL PREDICTING THE ARRIVAL TIME OF THE FIRST BAG AND THE INTERVAL BETWEEN THE FIRST AND LAST BAG ARRIVALS (LOADING TIME).

Quantile	MAE		Mean Error		RMSE	
	1 st bag	Load	1 st bag	Load	1 st bag	Load
50	6.1	4.0	-0.1	-1.7	8.0	7.2
70	7.6	4.6	4.1	0.6	9.4	7.2
90	11.7	8.0	10.1	6.1	13.6	10.3

B. Shapley analysis

SHAP (Shapley Additive Explanations) is a game-theoretic approach providing a unified framework to quantify the contri-

bution of each feature to a model's prediction [26]. It improves model interpretability by assigning each feature a contribution to the deviation of a prediction from the model's average output, taking into account all possible feature interactions to provide a transparent explanation for individual predictions. Figure 3 shows the results of the SHAP analysis for the two 50th quantile models, highlighting the stand position, number of bags and passengers as the most impactful input features.

C. Further Analyses of the Results

Two analyses have been conducted to identify how seasonal variations and operational differences affect the performance of the *1st bag* model.

In the first analysis, the observations according to stand type were grouped to measure the model performance. The dataset included four stand types: *Stations*, where passengers disembark in a separate building and are transported to the main terminal by bus; *Remote*, where passengers are transferred from the plane to the terminal by bus; *Terminal*, where passengers disembark directly to the terminal by using a passenger bridge; and *Walkboarding*, where passengers walk through the apron from the airplane to the terminal. When comparing the four types of stands, Figures 4b and 4d show that when an aircraft is assigned to a *Remote* stand, the model's performance slightly reduces (relative to the *Terminal* case), with the MAE averaging around 7 and 8 minutes, respectively, for the 50th and 70th quantile models. It is worth noting that *Walkboarding*, although rarely used at Munich Airport, exhibits higher predictability than the other stand types, with a MAE of 4 minutes for the 50th quantile, compared to approximately 6 minutes for the others.

The effect of the month on model performance is shown in Figures 4a and 4c. The model's performance improves toward the end of the year. In particular, November and December show the best prediction performance (below 5 minutes in terms of MAE for the 50th quantile). In contrast, the largest errors occur during the summer months, specifically in July and August where the error is closer to 7 minutes. When examining the Mean Error, the 70th quantile model exhibits a stronger tendency to overestimate the target values during the summer period. For these analyses, because only the most recent data from 2025 was used as the testing dataset, the training dataset was also included to represent a full year. This justifies potential discrepancies between the results of the analysis and the values shown in Table III which have been computed using only the testing dataset.

D. Comparison with Munich model

Passengers at Munich Airport typically visualise predictions of the first bag delivery time either based on real-time ground handling data or based on a fallback model. Here, a comparison between *1st bag* PRELUDE models and the Munich fallback model is provided.

Table IV reports the percentage of observations in the testing dataset whose prediction errors (defined as the difference between predicted and actual target values) fall within specified

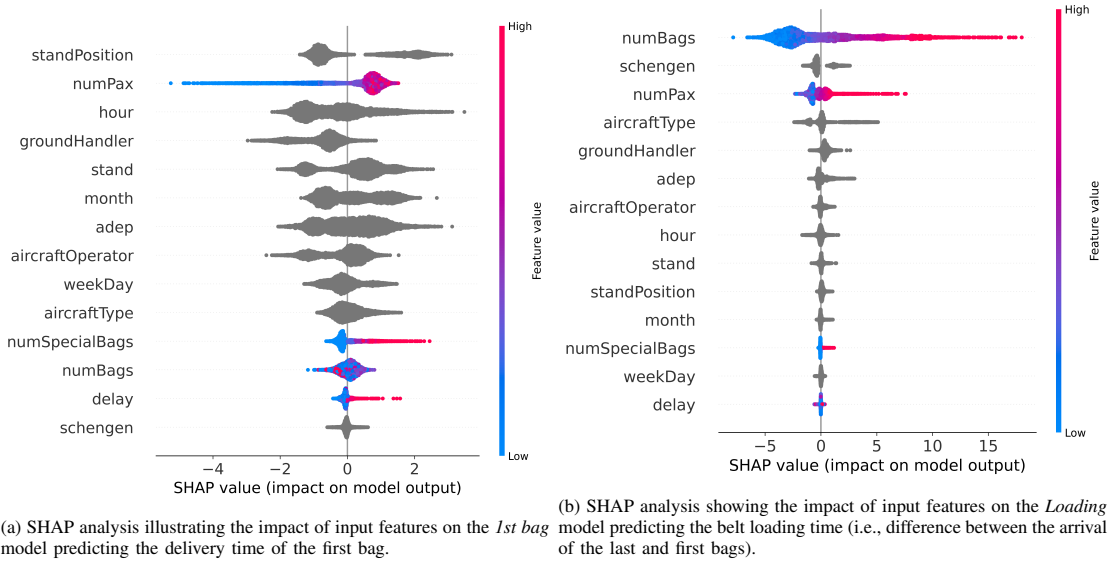


Figure 3. SHAP summary plots. The vertical axis displays the feature names, ordered from most to least influential. Each point along the horizontal axis represents the Shapley value of that feature for an individual observation, with color reflecting the feature's value, blue for low and red for high. Categorical features are shown in gray, as their values do not have a numeric magnitude.

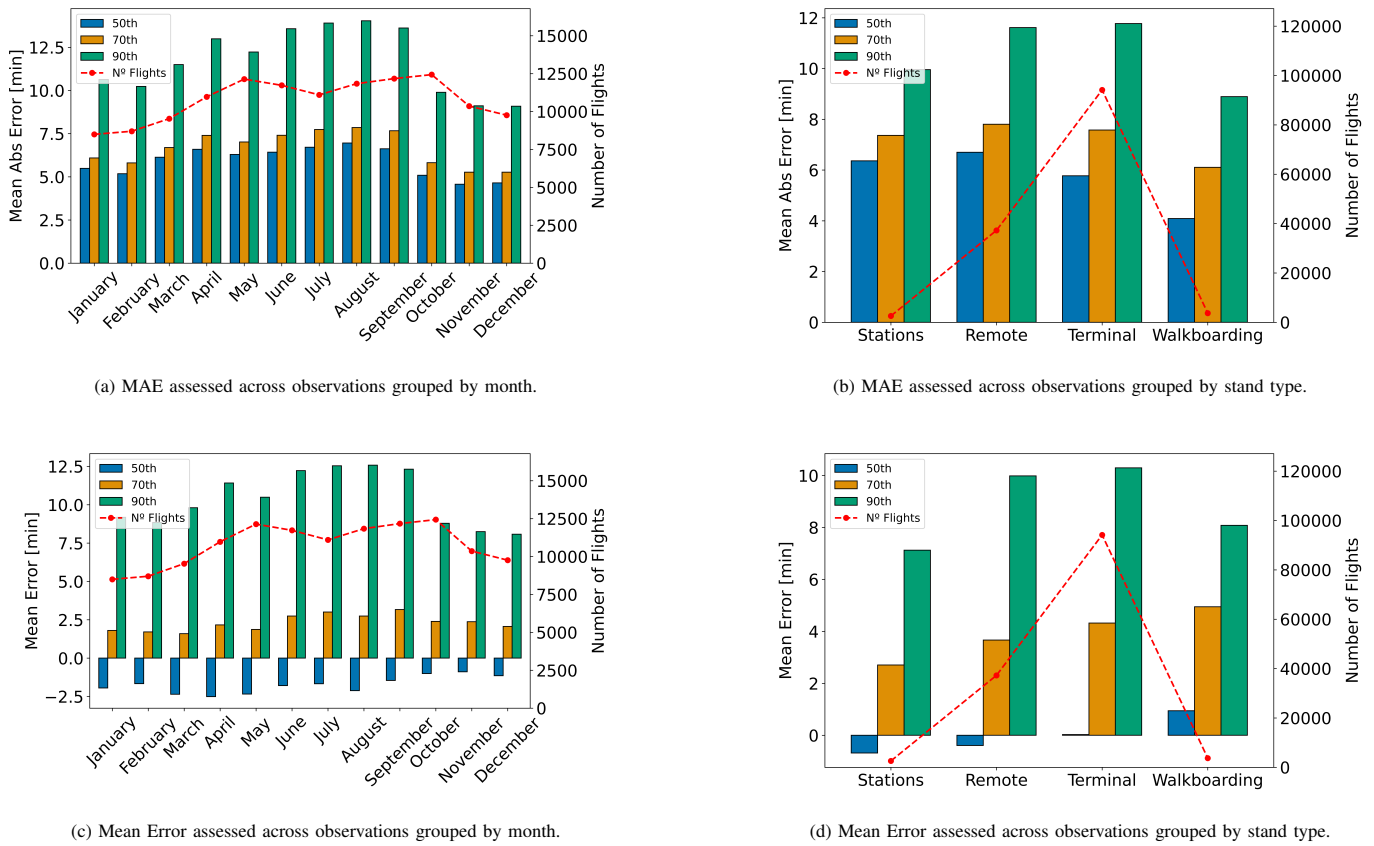


Figure 4. Performance of the *1st bag* model, computed for observations grouped by stand type and month.

ranges. For the 50th quantile model, the largest share of observations lies in the interval $(0, 5]$ (29.1%), followed by $(-5, 0]$ (22.8%) and $(5, 10]$ (17.1%). For the 70th quantile model, the distribution is similar, with $(5, 10]$ slightly more

dominant (24.9%) and $(-10, -5]$ reduced to 6.7%. At the 90th percentile, the proportions in $(0, 5]$ and $(5, 10]$ become more balanced (13.5% and 21.4%, respectively), whereas observations falling in the ranges $(-10, -5]$ and $(-5, 0]$ reduce

further, as expected. The Munich model shows the lowest percentages across all ranges, with the majority of residual errors concentrated in $(5, 10]$ (17.5%), followed by $(0, 5]$ (10.6%). Total percentages decrease from 81.5% (50th percentile) to 36.2% (Munich model), indicating that the Munich model frequently predicts values outside the error range of $[-10, 10]$ minutes.

TABLE IV. COMPARISON BETWEEN PRELUDE *1st bag* MODELS AND THE MODEL IMPLEMENTED AT MUNICH AIRPORT. THE ERROR IS COMPUTED AS THE DIFFERENCE BETWEEN PREDICTED AND ACTUAL TARGET VALUES.

Error (min)	50th	70th	90th	Munich
$(-10, -5]$	12.5%	6.7%	3.2%	2.9%
$(-5, 0]$	12.5%	6.7%	3.2%	2.9%
$(0, 5]$	29.1%	23.8%	13.5%	10.6%
$(5, 10]$	17.1%	24.9%	21.4%	17.5%
Total	81.5%	70.1%	44.4%	36.2%

A set of machine learning models was developed to predict both the arrival time of the first bag on the belt and the duration of belt loading. Specifically, the 50th, 70th, and 90th quantile models were developed. The 50th quantile model provides the most likely estimate and achieves the highest predictive accuracy with the MAE of 6.1 minutes for the *1st bag* model and 4.0 minutes for the *Loading* model. However, in operational contexts, slight overestimation may be preferable because recorded timestamps for the first bags do not always reflect the delivery of the main bulk of luggage. In such cases, predictions that slightly overestimate arrival times can lead to scenarios where bags arrive earlier than predicted, enhancing the perceived efficiency of ground handling.

For these reasons, the 70th quantile model is considered the most appropriate candidate for future trials, as it tends to slightly overestimate the target values with a Mean Error of 4.1 minutes for the *1st bag* model and 0.6 minutes for the *Loading* model. Meanwhile, the 50th quantile model remains useful for supporting airport operator decision-making by providing the most likely arrival time.

The model predicting the arrival time of the first bag identifies the stand position as the most influential feature. This is expected since the stand position reflects its distance from the baggage reclaim area, which is naturally correlated with the time required to transport bags from the stand to the reclaim belt (Section IV-B). Instead, the model predicting the duration of baggage loading operations on the belt is primarily influenced by the number of bags (Section IV-B). This result is also expected, as a larger volume of luggage naturally requires more time to load. Overall, these results suggest that the models capture the main dynamics effectively. It is therefore reasonable to hypothesise that a similar modeling approach could perform well for other airports.

While previous research [18] has shown that, in cases of delay (i.e., when AIBT occurs after the SIBT), aircraft turnarounds tend to be performed more quickly to prevent delay propagation, further analyses conducted within this work

revealed a different pattern for the delivery of the first bag: when delays are greater than 5–10 minutes, the delivery of the first bag tends to take slightly longer by one or two minutes compared to cases with minimal or no delay. This behavior is well captured by the model, as shown in Figure 3a, where higher delay values correspond to higher SHAP values, indicating a stronger contribution to predicting longer delivery times.

Uncertainty in predictive modeling can generally be divided into aleatoric uncertainty, which arises from inherent variability in the data, and epistemic uncertainty, which reflects the limited knowledge of the model due to its limited exposure to input data [27]. In this study, the difference between predicted quantile values could be interpreted as a measure of aleatoric uncertainty, since it captures the conditional variability of the target variable given the input features. Considering how the models' outcome could be presented to passengers, one possible approach is to display a single prediction derived from one quantile model, while a model supervisor at the airport retains additional quantile-based predictions that quantify uncertainty. This allows informing airport operators that the actual event duration may deviate, being shorter or longer than predicted, with an associated probability, thereby enabling risk-mitigating actions. Alternatively, a decision-making logic could be implemented when deploying the model. Specifically, the choice of which quantile prediction to communicate to passengers could explicitly depend on the quantified uncertainty, allowing an automatic adaptive dissemination of model outputs according to operational conditions.

The outcome of the two presented models could be communicated to passengers using various visualisation strategies. At Munich Airport, passengers can currently see the expected arrival time of the first bag on the monitors positioned in the luggage claim areas. Once the first bag reaches the belt, this message is replaced by an indication that baggage loading operations are in progress. Using the two PRELUDE models, the information presented to passengers could be enriched: alongside the expected time of the first bag, a visual bar could represent the ongoing loading duration, with color coding to indicate whether operations have just started or are nearing completion. Another possible strategy is to include the expected completion time of the loading operations from the start, once the notification on the screen for the given flight appears. This approach could increase transparency for passengers and potentially reduce anxiety by providing a clearer understanding of the baggage handling process. The precise textual and graphical presentation of these insights is currently under discussion and will be evaluated within the definition of the following validation activities.

As part of the project, face-to-face interviews were conducted with 1071 passengers during May and June of 2025 to assess their preferences regarding the timing of notifications about the expected arrival time of their bags on the baggage belt. Passengers were asked at which stage of their journey they would like to receive this information. The results, summarised in Figure 5, indicate that the majority of

respondents (29%) would prefer announcements made directly on the airline app. Notifications via screens at luggage claim area would be also highly preferred (28%), followed by announcement on the plane (26%) and airport apps (12%). A small proportion of respondents (5%) indicated that they did not have any preference. To address the most frequent response in the survey, in-flight announcements could be implemented, although this would require close coordination with airlines. Overall, as passengers generally prefer to receive notifications as early as possible, information could be provided not only at the luggage claim area but also using dedicated monitors immediately upon disembarkation. Furthermore, airport and airline applications could be enhanced by incorporating the model's predictions. All of these aspects are currently under discussion and will be considered in the next steps.

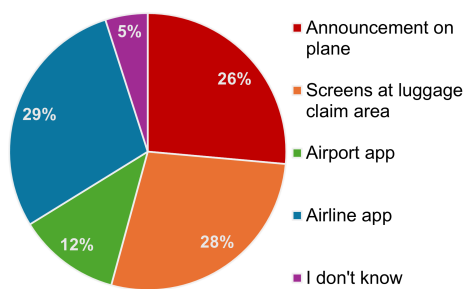


Figure 5. Results from the survey conducted at Munich Airport showing the percentage of interviewed passengers interested in visualising predictions for luggage arrival time at different phases of their journey.

At Munich Airport there is strong interest in further testing the PRELUDE model. The comparison of prediction errors shows that PRELUDE consistently achieves a much higher proportion of accurate estimates than the current system. For example, with the 50th quantile model, more than 50% of predictions fall within ± 5 minutes of the actual time, compared to roughly 16% for the Munich model Section IV-D. Even when considering a wider error margin of ± 10 minutes, PRELUDE 50th and 70th quantile models reach over 80% and 70% accuracy, while the current performance of the Munich model is approximately 36%. These results indicate that adopting the PRELUDE model could substantially improve the reliability of luggage time displays, thereby enhancing transparency and increasing passenger satisfaction. Next steps involve integrating the model with the airport's data systems and conducting a quality assessment during live operations using a shadow display, in which predicted values are not yet shown on luggage claim monitors but are used for initial evaluation. Following this assessment, predictions could be presented to passengers, while accompanying surveys could evaluate potential improvements in passenger satisfaction. Once validated, the model could be deployed for regular operations.

The current set of model inputs enables predictions once the arriving aircraft is in-block. A potential improvement would be to replace, among the set of input data, the AIBT with an estimated time of aircraft arrival at the gate, based on data available earlier, for example, when the aircraft is airborne. This adjustment would shift the prediction horizon forward (e.g., a few hours before flight arrivals), making the model suitable for applications such as belt allocation planning or on-shift resource management.

Beyond assisting passengers in planning subsequent stages of their journey or informing others of their arrival, these predictions may also be leveraged by transport operators to optimize connections across multimodal and intermodal travel chains. [28].

REFERENCES

- [1] Airports Council International (ACI) and the International Civil Aviation Organization (ICAO), "Joint ACI World-ICAO Passenger Traffic Report, Trends, and Outlook," <https://aci.aero/2025/01/28/joint-aci-world-icao-passenger-traffic-report-trends-and-outlook/>, 2025.
- [2] A. Medany, I. Blumbergs, and K. Elsakty, "Analysis of international air hubs: a competitiveness review," in *International Conference TRANS-BALTICA: Transportation Science and Technology*. Springer, 2022, pp. 645–654.
- [3] SITA, "Passenger satisfaction at baggage collection jumps to a new high with mobile notifications," <https://www.sita.aero/pressroom/news-releases/passenger-satisfaction-at-baggage-collection-jumps-to-a-new-high-with-mobile-notifications/>, 2019.
- [4] E. C. Pinto Neto, D. M. Baum, J. R. d. Almeida Jr, J. B. Camargo Jr, and P. S. Cugnasca, "Deep learning in air traffic management (atm): a survey on applications, opportunities, and open challenges," *Aerospace*, vol. 10, no. 4, p. 358, 2023.
- [5] V. Bogicevic, W. Yang, A. Bilgihan, and M. Bujisic, "Airport service quality drivers of passenger satisfaction," *Tourism Review*, vol. 68, no. 4, pp. 3–18, 2013.
- [6] A. G. de Barros, A. Somasundaraswaran, and S. Wirasinghe, "Evaluation of level of service for transfer passengers at airports," *Journal of Air Transport Management*, vol. 13, no. 5, pp. 293–298, 2007.
- [7] A. Brochado, P. Rita, C. Oliveira, and F. Oliveira, "Airline passengers' perceptions of service quality: Themes in online reviews," *International Journal of Contemporary Hospitality Management*, vol. 31, no. 2, pp. 855–873, 2019.
- [8] J. Allen, M. G. Bellizzi, L. Eboli, C. Forciniti, and G. Mazzulla, "Identifying strategies for improving airport services: Introduction of the gap-ipa to an italian airport case study," *Transportation Letters*, vol. 13, no. 3, pp. 243–253, 2021.
- [9] M. G. Bellizzi, L. Eboli, C. Forciniti, and G. Mazzulla, "Air transport passengers' satisfaction: an ordered logit model," *Transportation Research Procedia*, vol. 33, pp. 147–154, 2018.
- [10] S. Paramonovs and K. Ijevleva, "Factor analysis of passengers' satisfaction at "riga international airport"," *Economics and Business*, vol. 27, pp. 46–52, 2015.
- [11] G. E. Vastianos, D. M. Kyriazanos, V. I. Kountouriotis, and S. C. Thomopoulos, "An rfid-based luggage and passenger tracking system for airport security control applications," in *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, vol. 9091. SPIE, 2014, pp. 412–424.
- [12] RFID Card, "Rfid baggage tracking in airlines: A complete guide," https://www.rfidcard.com/rfid-baggage-tracking-in-airlines-a-complete-guide/?srsltid=AfmBOoo1eAw-74bcTwaT71kR9jIWdirLw2G4mqAMEbmTwIRbwVfA_qpq, 2024, accessed: 2025-09-22.
- [13] S. A. Solutions, "Real-time baggage insights with deep turnaround," <https://www.schiphol.nl/nl/aviation-solutions/blog/real-time-baggage-insights-with-deep-turnaround/>, 2024, accessed: 2025-09-22.

- [14] J.-F. Fondevila-Gascón, Ó. Gutiérrez-Aragón, D. Lopez-Lopez, G. Curiel-Barrios, and J. Alabart-Algueró, "Passenger perceptions of artificial intelligence in airline operations: Implications for air transport management," *Journal of Air Transport Management*, vol. 129, p. 102874, 2025.
- [15] Y. Xie, N. Pongsakornsathien, A. Gardi, and R. Sabatini, "Explanation of machine-learning solutions in air-traffic management," *Aerospace*, vol. 8, no. 8, p. 224, 2021.
- [16] A. Degas, M. R. Islam, C. Hurter, S. Barua, H. Rahman, M. Poudel, D. Ruscio, M. U. Ahmed, S. Begum, M. A. Rahman *et al.*, "A survey on artificial intelligence (ai) and explainable ai in air traffic management: Current trends and development with future research trajectory," *Applied Sciences*, vol. 12, no. 3, p. 1295, 2022.
- [17] P. De Falco and M. Karaarslan, "Probabilistic constraints prediction with uncertainty quantification in trajectory based operations (tbo)," *14th SESAR innovation days*, 2024.
- [18] P. De Falco, J. Kubat, V. Kuran, J. Rodriguez Varela, S. Plutino, and A. Leonardi, "Probabilistic prediction of aircraft turnaround time and target off-block time," *13th SESAR innovation days*, 2023.
- [19] S. Birolini and A. Jacquillat, "Day-ahead aircraft routing with data-driven primary delay predictions," *European Journal of Operational Research*, vol. 310, no. 1, pp. 379–396, 2023.
- [20] S. J. Moon, J.-J. Jeon, J. S. H. Lee, and Y. Kim, "Learning multiple quantiles with neural networks," *Journal of Computational and Graphical Statistics*, vol. 30, no. 4, pp. 1238–1248, 2021.
- [21] R. Dalmau, P. De Falco, M. Spak, and J. D. R. Varela, "Probabilistic pretactical arrival and departure flight delay prediction with quantile regression," *Journal of Air Transportation*, vol. 32, no. 2, pp. 84–96, 2024.
- [22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [23] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [25] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *preprint arXiv:1810.11363*, 2018.
- [26] L. S. Shapley *et al.*, "A value for n-person games," 1953.
- [27] A. Malinin, "Uncertainty estimation in deep learning with application to spoken language assessment," Ph.D. dissertation, 2019.
- [28] D. Babić, M. Kalić, M. Janić, S. Dožić, and K. Kukić, "Integrated door-to-door transport services for air passengers: From intermodality to multimodality," *Sustainability*, vol. 14, no. 11, p. 6503, 2022.