Probabilistic Prediction of Aircraft Turnaround Time and Target Off-Block Time

Case studies for Prague, Geneve, Arlanda and Fiumicino international airports using operational data

Paolino De Falco
EGSD/INO/ENG
EUROCONTROL Innovation Hub
Bretigny-Sur-Orge, France

Jan Kubat &
Vladimir Kuran
Prague Airport
Prague, Czech Republic

José Rodriguez Varela
Geneve Airport
Geneve, Switzerland

Salvatore Plutino &
Alessandro Leonardi
Aeroporti di Roma
Rome, Italy

***Abstract*—Collaborative decision making airports are extremely dependent on the precision of the target off-block time (TOBT) which is the target time set by an airline or ground handler agents for the off-block departure. This value reflects any delays that can be attributed to the aircraft operator or to the ground handling operations and must be updated by +/- 5 minutes when differing from the previous released value. Last-minute changes of the TOBT are undesired as they might alter the pre-departure sequencing resulting in very late air traffic flow management departure slots. Therefore, accurate predictions of turnaround times and last TOBT values are essential for better planning and tactical management of stands. This paper presents a set of probabilistic machine learning models to predict turnaround time and last TOBT values in nominal operational conditions at Prague, Geneve, Arlanda and Fiumicino airports. The turnaround models exhibit mean absolute errors ranging from 9 to 7 minutes during the strategic/pre-tactical planning phase, and from 6 to 4 minutes during the tactical planning phase. A validation exercise using ground handlers' data shows potential benefits for airport operations. Finally, a model trained on all the available data from the four airports demonstrates the potential to generalise the approach without compromising the quality of predictions. Prague and Fiumicino airports will deploy the model in the operational environment during the first quarter 2024.**

***Keywords*—Turnaround; Target Off-Block Time; Machine learning**

I. Introduction

The concept of airport collaborative decision making (A-CDM) was introduced at the end of the 90s to improve the efficiency and resilience of airport operations, and it is implemented at 33 European airports nowadays. A-CDM allows the airport partners, such as airport operators, aircraft operators, ground handlers, air traffic controllers (ATCs) and the Network Manager to work more transparently and collaboratively, whilst exchanging relevant, accurate and timely information focusing on predictability and pre-departure processes [1]. In A-CDM a target time relates to a milestone and serves as an agreement between partners who are thus committed to achieving a milestone at a specified time. These milestones are expected to trigger the decision-making process for downstream events and influence both the further progress of the flight and the accuracy with which the progress can be predicted [2].

One of the A-CDM milestones is the target off-block time (TOBT) which is defined as the time when an aircraft operator or ground handler agent estimates that an aircraft will be ready to depart (doors should be closed, boarding bridge removed and pushback vehicles available and ready to start-up). TOBTs are made available via a CDM Platform to the ATCs and other CDM partners. ATCs can confirm the TOBT of each flight by providing a corresponding target start-up approval time (TSAT) which is either equal to or later than the TOBT according to the possible air traffic flow management (ATFM) regulations or to the need of maximizing the runway throughput and ground movement interactions [1].

The TOBT must be updated by the turnaround coordinator based on the progress of the turnaround processes (such as catering, cleaning, fueling, and boarding of passengers) by +/- 5 minutes. The TOBT coordinator is at the very heart of the A-CDM process. Indeed, the achievement of substantial levels of local and network predictability requires that a culture of early and accurate TOBT updates is embedded at the CDM airport [3]. Although the common effort towards sharing more accurate information is well recognised, the TOBT coordinator might be reluctant to provide early TOBT updates to avoid that a flight become subject to ATFM delay upon its update or because additional TSAT delay maybe incurred by moving the flight into a period of higher departure demand [3]. While regular and accurate updates of early TOBTs are important, the final update of TOBT requires the highest accuracy and timeliness since it ties up the largest number of resources such as airspace, ground handling equipment, airport facilities, personnel, and passengers' time [2]. Current manual input of TOBT is prone to deficiencies because of human factor limitations, typos, and inefficient information flow, resulting in last-minute changes of TOBT. As the operational complexity on the ground increases, the effort of turnaround coordinators responsible for updating the TOBT also increases while trying to keep pace with the micro-adjustments for each aircraft. Consequently, the number of last-minute TOBT updates grows in such scenarios. It is during these critical moments that substantial delays in the network are triggered. Indeed, it has been experienced that very last-minute changes of TOBT tend to shift the TSAT value and trigger very late calculated take-

off time (CTOT) with small chances for improvement.

Possible reasons for delay during the turnaround can be allocated to: aircraft and ramp handling, cargo and mail, damaged aircraft, flight operations and crewing, passengers and baggage, and aircraft equipment. In 2019 11.1 million flights were recorded over Europe, including 8.3 million commercial flights. According to the Eurocontrol CODA reports [4], the average delay per flight was 13.1 minutes including 3.4 minutes (value provided by the CODA Eurocontrol team) of delays occurring during the turnaround (IATA delay codes 11-19, 21-29, 31-39). Although, the cost of delay is non-linear [5], a single average value can be used when making approximate calculations. Using the value reported in [6], the cost of delay per minute on the ground without network effect is €53.2 (values are computed from 2014 data), making the total cost of delay due to turnaround operations of approximately €1.5 billion. Several airports have been recently facing operational issues caused by a high rate of regulation and operations close to infrastructural capacity. Despite this, the predictive trend for next years is still rising thus indicating worsening of the situation in the future. Covid outbreak only provided some additional time to alleviate this saturation. Without any major infrastructural changes, there is a need of accurate predictions of turnaround time and TOBT values for better planning and tactical management of stands contributing to the cost reduction of delayed turnaround operations.

This paper presents the main results of the project *OpTT (Optimisation of Turnaround Time)* which was initiated in response to a proposal from Prague airport within one of the Eurocontrol Air Transport Innovation Network (EATIN) initiatives (https://www.eurocontrol.int/project/eatin). As more airports have later joined the project, a set of machine learning (ML) models predicting the last release of TOBT and turnaround time has been developed using data provided by Prague, Geneva, Arlanda and Fiumicino airports from 2022. The models are expected to be used at any planning phase (Section III-A) by airport operators in nominal operational conditions (e.g., low in-bound flight delays and non-critical weather conditions). As a result of an approach combining regression and classification algorithms, the output of the models is a probability distribution of the turnaround time (TT), and the level of prediction uncertainty is quantified. The predictions in the test set, which includes observations never seen by the model during training, are compared to a baseline that assumes turnaround time being equal to the scheduled one. Furthermore, some of the models have been validated using ground handler data from Prague and Arlanda airports. As a generalization of the approach, a final model has been developed combining the four airports' datasets.

The paper is organised as follows: a literature review on turnaround time predictions is presented in Section II; Section III provides the description of the models developed for each airport and of the metrics that are used for their performances assessment. In Section IV a description of the main results is provided. Section V shows the results of a generalised model while Section VI provides an overview of the operational benefits and model implementation at Prague airport. Finally, Section VII includes the discussion and conclusions.

II. Literature Review

There are several works in literature on the modelling of turnaround sub-processes [7], [8], [9]. These studies generally use stochastic probability functions based on historic data, discrete event simulation and synthetic data generated by agent-based simulation approaches [10].

Models predicting single turnaround processes, that are not based on synthetic data, require historical data which are typically collected manually. Because of human behaviour during the acquisition phase, the usage of these data in predictive models might introduce aleatory uncertainty [11]. Recent initiatives focus on the monitoring and prediction of turnaround processes using camera systems overcoming this limitation (https://www.schiphol.nl/nl/aviation-solutions/pagina/deep-turnaround/; https://www.assaia.com/). However, for these technologies the quality of predictions is highly dependent on the capability to label accurately the duration of the single processes.

A possible approach when scheduling turnaround time is to add buffer time to the estimation of minimum turnaround time. This additional time is typically used to absorb possible delays from the inbound legs, although sometimes, it can be imposed on the airline due to airport slot availability or other scheduling constraints. In the POEM SESAR project, the minimum turnaround was computed as the 2nd percentile from turnaround distributions of historical data grouped by aircraft operator, airport size and wake turbulence category [12]. Studies on optimal buffer times to minimise delay propagation were conducted in [13] using Monte Carlo simulations. Here, it was highlighted dynamic buffering as a concept to overcome deficiencies of the typical buffer strategies for ground processes.

Turnaround time is affected by many random factors, such as passenger behavior, airport resource availability, and short-noticed maintenance activities. These factors make the realisation and prediction of turnaround time uncertain. The quantification of prediction uncertainty is an important task as it might affect the decisions of the model user [14]. In [15] the authors model the turnaround processes and compute the probability that a turnaround is completed within a specific TOBT by applying mathematical convolution and including the uncertainties related to the single processes in the modelling approach. Similarly, in [16] and [17] it was proposed an approach based on Monte Carlo simulation to identify critical paths of turnaround process and account for operational uncertainties. In [18] turnaround predictions during the tactical planning phase were approached with a regression model predicting the aircraft ground time and with a classification model predicting adherence levels of TOBT to the actual off-block time. Results from the XGBoost regression model show standard deviation values ranging from 4 to 7 minutes approximately.

III. Model development

This section presents the details of the models development. First, a description of the input data is provided. Secondly, a data-driven method to label and filter outliers is proposed. Finally, the algorithms, the models' output and the metrics for the assessment of their performances are described.

A. Data availability and input features

There are limitations when training a model that should perform predictions during the strategic or pre-tactical planning phase. Indeed, predictions can only be made using information that is available at the inference time [19]. Some information, such as weather or ATFM regulations is only available with a certain level of accuracy on the day of operations. This makes the set of features available for making predictions in the strategic or pre-tactical phase rather limited. The models proposed in this paper have been trained with the input features described in Table I, where the acronyms SOBT, SIBT and AIBT refer, respectively, to the schedule off-block time, schedule in-block time and actual in-block time.

TABLE I. Set of attributes used to train the models.

Attribute	Description
Day of week	Encoded as numerical
Hour of operations (at SOBT)	Encoded as numerical
Month of operations (at SOBT)	Encoded as numerical
Great circle distance (GCD)	From origin airport (Km)
Scheduled Turnaround Time (STT)	SOBT - SIBT
Available Turnaround Time (ATT)	SOBT - AIBT
Aircraft type	Encoded as numerical
Aircraft MTOW	Maximum Take-off Weight (Kg)
Aircraft operator	Encoded as numerical
Normalised congestion	Numerical

The user will be requested at inference time to input the most updated In-Block information for the feature *Available Turnaround Time*. It is expected that in the strategic/pre-tactical planning phases the user will adopt the SIBT while for the tactical phase the AIBT (which has been used for the model's training) or its estimation. It is expected that the level of congestion of ground operations for each airline is dependent on the number of flights of a certain airline performing the turnaround within a specific time frame. Therefore, the feature *normalised congestion* in Table I has been computed as the ratio between the hourly number of planned turnarounds on a specific day and their overall average values per airline.

Data starting from January 2022 were provided by the four airports covering the following amount of months: Prague: 9; Arlanda: 6; Geneve: 7; Fiumicino: 11. Prior to training the models, data was pre-processed to reflect the operational requests of the project partners. Specifically, only the occurrences with the following characteristics were kept:

- $15\text{min} \leq \text{TOBT}_{last}$ - AIBT $\leq 200\text{min}$
- $15\text{min} \leq$ SOBT - SIBT $\leq 200\text{min}$
- IATA Service categories: "Passenger only", "Normal service", "Technical stop", "Non-revenue", "Loose loaded cargo", "Cargo/Mail" and "Cargo"
- $-30\text{min} \leq$ AIBT - SIBT $\leq 60\text{min}$

Number of passengers, aircraft stand identifiers and weather information such as, wind speed, temperature, visibility distance, dewpoint, rain, snow and fog intensity have been included as input data in a first phase of model development. These data were later removed to simplify the data extraction at the airports during inference time since their contribution to the predictions was negligible (assessed by SHAP analysis [20] and metrics as in Section III-E). As a result of the data selection, no attribute is specific to the layout or other characteristics of a certain airport allowing a relatively easy generalisation of the model (i.e., trained with all the available data) as shown in a validation exercise later (Section V-B)

B. Probabilistic predictions

A method to compute probabilistic predictions has been implemented in this study. The details of this method can be found in [21] and [22], and are here described. First a regression model is trained to predict the target variable. The error distribution produced by this model is discretised into a number of bins and used as a target for a classifier after one-hot encoding. Finally, the outcome of the regression and classification models are combined producing an individual prediction as a discrete probability distribution. These steps are summarised in Figure 1.

It is important to highlight that in this approach the number of bins (i.e., number of classes for the classifier models) of the probability distribution is a parameter that might be selected to optimise the model performances. In the following sections the results obtained using 20 bin distributions will be shown. However, similar results were achieved when discretising the error domain with 10, 25 and 30 bins. The model performances will be derived from the probability distributions, as explained in Section III-E.

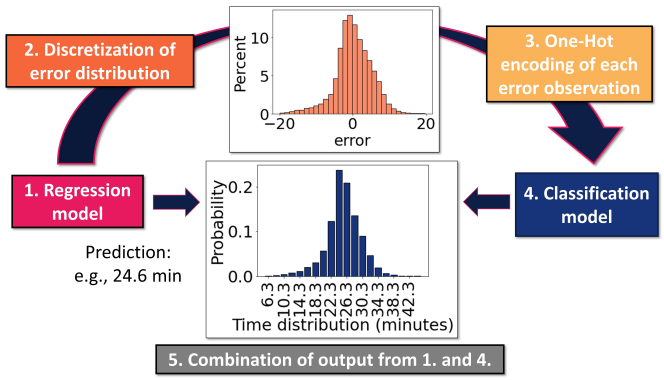

Figure 1. Method for probabilistic predictions in regression problems. The steps are: 1) Regression model predicting continuous target variable; 2) Discretisation and binning of the error distribution produced by the regression model; 3) One-Hot encoding of the error associated with each observation; 4) Classification model to predict the one-hot encoded errors; 5) Combination of the output of the two models

C. Machine Learning algorithm

Boosting is a machine learning ensemble technique that combines the predictions of multiple weak learners to create

a strong learner. XGBoost [23] is a boosting algorithm that builds a strong learner by training decision trees sequentially, with each tree correcting the errors made by the previous ones using gradient descent optimization [24]. It has recently been dominating applied machine learning and Kaggle competitions and has been selected as a ML algorithm for the models here presented.

In ML, a loss function is a crucial component that quantifies the difference between the predicted values of a model and the actual target values in the training dataset. The primary goal of a ML algorithm is to minimise this loss function and make the model's predictions as accurate as possible. In this study, the mean squared error (MSE) and the cross entropy loss functions have been used [25], respectively, to train the regression and classification models according to the methodology explained in Section III-B.

Several hyper-parameters can be optimised when implementing XGBoost allowing to control the learning process and the model performances. In this study, the maximum depth and the number of decision trees were optimised because they were found to have the most significant impact on the loss function. Cross-validation (CV) is a widely employed technique to evaluate how well a model performs with specific hyper-parameter settings. Among its variations, the fundamental k-fold CV involves partitioning the training dataset into k distinct subsets, referred to as folds. Subsequently, the ensuing process is repeated for each of these k folds: a duplicate of the model is trained using the remaining k-1 folds as the training set, while the current fold serves as the test set for calculating a performance score. The average of these k scores determines the CV score, which quantifies the model's efficacy for the given hyper-parameter configuration. There exist several methods to search the hyper-parameter space for the best CV score. In this study, the GridSearchCV [26], which evaluates all the possible combinations of hyper-parameters and return the one minimising the CV score, was implemented. The tuning process for the four models using five folds led to one hundred estimators and three-five edges (maximum depth) as optimal hyper-parameters.

D. Output target

The models output the probabilistic distribution and the expected value of turnaround time defined by Equation 1, where $TOBT_{last}$ is the last release of TOBT. The term $TOBT_{last}$ can be also computed as an output by adding the predictions of TT to the most updated in-block time information at a specific time horizon (Section IV-A).

$$\textit{Turnaround time (TT)} = TOBT_{last} - AIBT \tag{1}$$

E. Metrics

For the assessment of the model performances, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) have been used (Equations 2 and 3). In the equations y_i are the target values, $\hat{y}_i$ are the expected values of the probability distributions resulting from the method described in Section III-B and n is the number of data points.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{2}$$

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

As a new metric, we will refer to the term *uncertainty* (Figure 2) as the time domain which is centred around the peak of the distribution underlying the 95% of the probability distribution [22]. Since this metric expresses the confidence interval of single predictions, it is expected that the narrower the confidence interval, the lower the level of uncertainty. The results that will be shown in Section IV-A were computed on a randomly selected testing dataset that the models have not used for training.

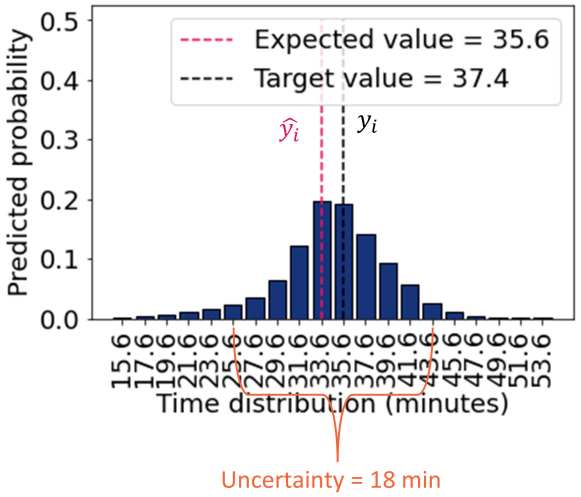

Figure 2. Visual description of the metrics. Uncertainty is computed as range of time covering the top 95% of the probability values. The MAE and RMSE are computed from the difference between the target and expected value of the distribution (respectively, y_i and $\hat{y}_i$)

F. Outliers filtering

A data-driven methodology was implemented to identify and filter outliers in the datasets, i.e., flights with unusual turnaround time. The models are expected to be used in nominal operational conditions. In these situations the presence of outliers (data points that are significantly different from the rest of the dataset) is undesired. In a decision tree the values in the leaves (terminal nodes) are typically computed as the average of the target values of the training samples that reach that leaf node during the tree construction. In this work, a regression decision tree model [27] was trained to predict the turnaround time using the available input features listed in Section III-A and the target described in Section III-D. As a loss function the mean squared error was used and as a constraint at least 5% of samples were forced to

fall in each leaf. The 2.5% and 97.5% percentiles of the target distribution (i.e., turnaround time) in each leaf were computed. In each leaf the data distributed between the maximum and the 97.5% percentile or between the minimum value and the 2.5% percentile were labelled as outliers, if:

- the difference between the maximum value and the 97.5th percentile was greater than $0.3 \cdot 97.5$th percentile, or
- the difference between the minimum value and the 2.5th percentile was greater than $0.3 \cdot 2.5$th percentile.

The approach for the outlier labeling is described in Figure 3. When computing the resulting target distribution in each leaf, the observations falling in the tails introduce the highest variance. Specifically, the observations falling within *heavy tails* introduce higher variance than the ones falling in *light tails* since their difference with the mean value is expected to be higher. Indeed, heavy tailed distributions tend to have many outliers with very high values. The heavier the tail, the larger the probability to find more disproportionate values in a sample [28].

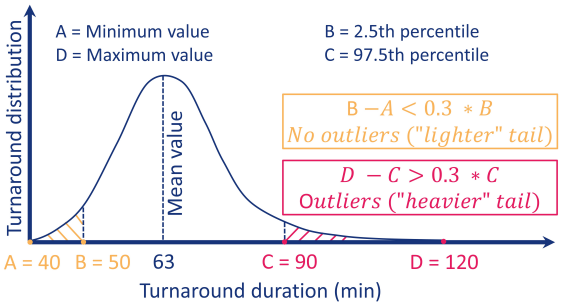

Figure 3. Description of the methodology for outlier labeling. An example of the target distribution in a generic leaf of the decision tree is represented. All the observations falling within the points C and D are labeled as outliers.

This methodology allowed to label as outliers 3 - 5% of data and to improve the model predictions by approximately 15% when excluding these outliers from each of the airports' dataset. In this approach, a set of parameters has been introduced, such as the coefficient *0.3* that was used for the identification of outliers. It is authors' intention to perform a full parametric analysis in a later work. However, an analysis on the minimum number of sample in each leaf was conducted using the value 4, 5 and 6%. Results show no significant impact of this parameter on the amount of data labelled as outliers.

Figure 4 shows the evolution over most of 2022 of the amount of data samples that are detected as nominal and non-nominal by using the data-driven approach described in Section III-F. As an example, data from Fiumicino airport were used. For this analysis the data were grouped by month and the percentage of non-nominal turnaround operations was computed from the ratio between all the turnaround operations and the non-nominal ones. Interestingly, the peaks of non-nominal turnaround operations (values are shown also in percentages) were found in January and over the summer months which are typically expected to be very challenging for airport management.

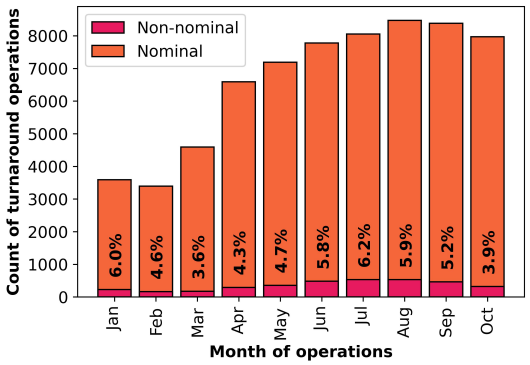

Figure 4. Analysis of nominal and non-nominal turnaround operations that were detected with the data-driven approach described in Section III-F for Fiumicino airport during most of 2022. The percentage of non-nominal turnaround operations is also shown.

IV. Results

A. Predictions of turnaround time and TOBT

In this section a description of the model performances is provided according to the metrics defined in Section III-E.

Results are shown in Figure 5 and Figure 6. As a baseline to compare the model performances the term in Equation 4 was used. Indeed, in a scenario where only schedule information is available, the user might rely on the STT to assess the duration of the turnaround. The model performances in the Strategic/Pre-Tactical planning phase are obtained by substituting the term SIBT to the term AIBT when providing the *available turnaround time* as an input feature (Table I) since the SIBT is the only in-block information at that time horizon.

$$\text{Baseline error} = (SOBT - SIBT) - TT \tag{4}$$

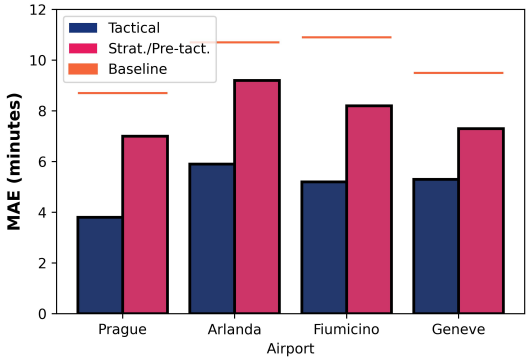

Figure 5. MAE values measuring the performances of the models for the four airports.

Figures 5 and 6 show that the models perform better in tactical rather than in strategic/pre-tactical planning phase, as expected. Prague model is the most performing in the tactical

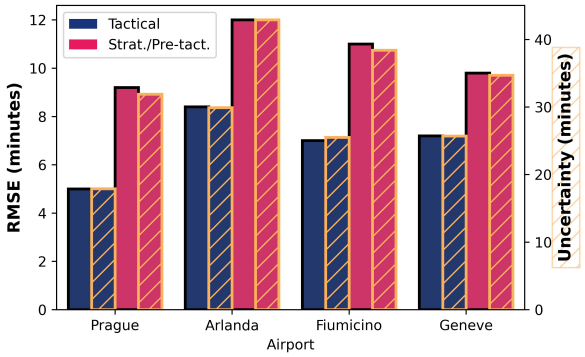

Figure 6. RMSE and uncertainty values measuring the performances of the models for the four airports. The uncertainty values are computed using the approach in Section III-E. The values of the filled bars should be read on the left y axis while the value of the hatched bars should be read on the right y axis.

phase according to all the available metrics (MAE is 3.8 minutes, RMSE is 5.0 minutes and the uncertainty is 17.9 minutes). The performances of Fiumicino and Geneve models in this planning phase are quite similar while Arlanda shows the highest values in terms of MAE, RMSE and uncertainty (respectively 5.9, 8.4 and 29.9 minutes). In the strategic/pre-tactical phase, a similar analysis shows that Prague model is still the most performing (MAE is 7.0 minutes, RMSE is 9.2 minutes, and the uncertainty is 31.9 minutes) although its performances are very similar to the ones of Geneve model.

The percentage improvements of the model in the tactical and strategic/pre-tactical phases when compared to the baseline (Equation 4) in terms of MAE and RMSE are, respectively, 56.0% and 20.0% for Prague, 44.0% and 14.0% for Arlanda, 52.3% and 24.8% for Fiumicino, and 44.2% and 23.2% for Geneve model. Interestingly, Figure 6 shows that the average *uncertainty* of the models is proportional to their RMSE.

The outcome of the models allows to compute also the $TOBT_{last}$ as an output. In the strategic/pre-tactical planning phases the predicted $TOBT_{last}$ was computed by adding the predicted values of TT to the SIBT. This value has been later compared to the corresponding actual $TOBT_{last}$ leading to the results reported in Table II. During the tactical planning phase, instead, once the AIBT is available the TOBT can be predicted with the same accuracy of the models (blue bars in Figures 5 and 6).

TABLE II. MAE AND RMSE OF PREDICTED $TOBT_{last}$ VALUES WHEN USING THE MODEL DURING THE STRATEGIC/PRE-TACTICAL PLANNING PHASE.

Airport	MAE (min)	RMSE (min)
Prague	10.8	23.9
Geneve	10.4	12.1
Fiumicino	14.4	21.2
Arlanda	11.1	38.9

B. Shapley analysis

Principles from game theory can be used to interpret the prediction of a ML model for a given set of observations where each input feature is a player and the model output is the payout. Assuming that all the input features participate in the game (i.e., are included for the model development) and join the game in a random order, the contribution of a feature could be calculated as the average change in the payout received by the coalition which already joined the model when the feature joins them. This contribution measure is commonly known in the literature as the SHAP or Shapley value [20].

Computing Shapley values for an arbitrary model is an NP-hard problem. In this paper, a new implementation (called *TreeExplainer*) allowing for tractable computation of Shapley values in polynomial time has been used [29].

For sake of simplicity, only the outcome of the SHAP analysis that was performed on the testing dataset from Prague airport is presented in Figure 7. However, similar qualitative results have been observed for the other developed models. According to Figure 7 the most relevant feature in terms of mean absolute Shapley value is the *Available Turnaround Time (ATT)*.

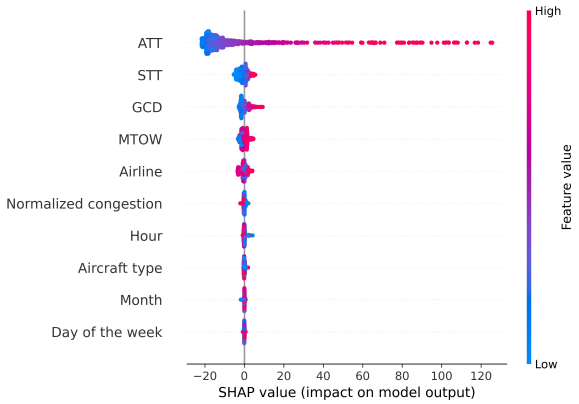

Figure 7. SHAP analysis showing the importance of input features for Prague model. On the vertical axis the name of the features is indicated, in order of relevance from the top to the bottom. Each dot in the horizontal axis represents the Shapley value of the associated feature for a single observation and the colour indicates the magnitude of that feature ranging from blue (low values) to red (high values).

V. VALIDATION EXERCISE

A. Model validation using operational metrics

This section provides a comparison of the model performances with the operational TOBT values provided by Prague and Arlanda ground handlers. Specifically, the first TOBT release ($TOBT_{1st}$) by ground handlers and the model predictions are compared to the actual value of $TOBT_{last}$. In the A-CDM implementation it is recommended (whenever there is a need) to update the TOBT values by +/- 5 minutes. According to this rule, the average number of TOBT updates per flight and the percentage of flights requiring at least one update of TOBT were estimated. For the first metric the

average number of TOBT updates per flight was computed as the ratio between the term $TOBT_{last} - TOBT_{1st}$ and the 5 minutes limit value which would trigger a new TOBT release (rounded to the integer digit). For the second metric, the focus shifted to consider only turnarounds where the time difference between the initial and final TOBT values exceeded 5 minutes, indicating the need for at least one TOBT update. Similar metrics were computed by comparing the model predictions of $TOBT_{last}$ values and their actual values (Figures 8 and 9).

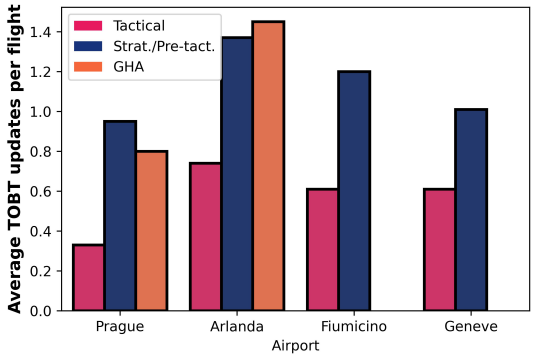

Figure 8. Average number of TOBT updates per flight. A comparison using operational metrics and models predictions.

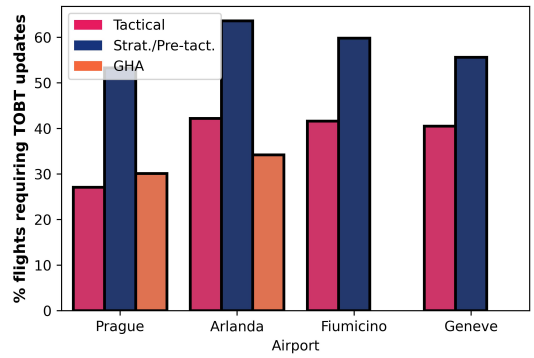

Figure 9. Percentage of flights requiring at least one TOBT update. A comparison using operational metrics and models predictions.

Figure 8 shows that for Prague airport in a tactical phase the model outperforms (by 59%) the performances of ground handlers that were computed based on the first release of TOBT while in the strategic/pre-tactical phase the model would require slightly a higher number of TOBT updates per flight (19% performance reduction). For Arlanda in both scenarios the model outperform the values computed with the ground handler's data by respectively 49% and 6%.

Figure 9 shows that in the tactical phase for Prague airport the model outperforms (by 10%) the performances of ground handlers that were computed based on the number of flights requiring at least one TOBT update. Instead, in the strategic/pre-tactical phase the model features 77% of performance reduction. For Arlanda, in both phases, a reduction of the performances by respectively 23% and 86% is recorded.

For Fiumicino and Geneve airports these two new metrics were computed only using the model predictions since the values of $TOBT_{1st}$ were not provided. Their average number of TOBT updates and percentage of flights requiring at least one TOBT update were 0.6 and approximately 40% in the tactical phase and, respectively, 1-1.2 and 60-55% in the strategic/pre-tactical planning phase (Figures 8 and 9).

B. Towards a generalised model

Training and maintaining several ML models could be complex and inefficient. Therefore, it was decided to train a unique model using all the available data from the four airports. The datasets were combined to compare the performances of a single generalised model with the ones of the models presented in the previous sections that will be now referred to as *ad-hoc*. As a common practice, the dataset was randomly split into a training and testing set using the ratio 80:20, as for all the models presented in this manuscript.

TABLE III. COMPARISON BETWEEN THE GENERALISED AND *ad-hoc* MODELS (PRESENTED IN THE PREVIOUS SECTIONS) WHEN USED IN THE TACTICAL PHASE. THE METRICS FOR THE GENERALISED MODEL WERE COMPUTED ON DATA SAMPLES OF SPECIFIC AIRPORTS FROM THE TESTING DATASETS.

Airport	Model	MAE (min)	RMSE (min)	Uncertainty (min)
Prague	Ad-hoc	3.8	5.0	17.9
	Gener.	3.9	5.3	23.4
Geneve	Ad-hoc	5.3	7.2	25.7
	Gener.	5.1	6.9	25.8
Fiumicino	Ad-hoc	5.2	7.0	25.5
	Gener.	5.2	7.0	26.5
Arlanda	Ad-hoc	5.9	8.4	29.9
	Gener.	5.6	7.9	26.1

For sake of simplicity, here only the performances of the models in the tactical phase are compared. Table III shows that the generalised model maintains overall similar levels of performances of the *ad-hoc* models when tested on specific airport datasets. It is interesting to notice that the performances of the generalised models improve (by 5% in terms of MAE) for Arlanda (vice versa for Prague airport which shows a reduction of 3% in terms of MAE). The outcome of this analysis triggered a follow-up EATIN project called *OpTT 2.0*, which is currently running and aims to develop a unique predictive model of turnaround time and TOBT for all the European CDM airports.

VI. MODEL IMPLEMENTATION AT PRAGUE AND FIUMICINO AIRPORTS

Prague and Fiumicino airports has decided to implement the model in their operations centre. The model will be a crucial part of the automatic TOBT calculation algorithm that will be incorporated in the live operations by Q1 2024. While the technological solution for the deployment at Fiumicino airport is under study, Prague airport will use its infrastructure in the

means of the Integration Service Hub, where the predictive model will be one of its micro-services. Such implemented model would lower the ground handler's workload and in return should decrease the number of late TOBT updates which cause late CTOTs. In result, this should support optimal stand allocation planning and better turnaround predictability.

VII. Discussion & Conclusions

Machine learning models require a set of input information during inference. A key aspect is the availability, at a specific time horizon of interest, of this information that feeds the model. The air traffic management (ATM) system is characterised by three phases: strategic, pre-tactical and tactical. The strategic phase encompasses measures taken earlier a few days prior to the day of operation up to two months or more in advance. The pre-tactical phase encompasses measures taken a few days or one day prior to the operation (in certain cases up to few hours before operations) and the tactical measures are adopted during the day of operations. In this paper a set of models has been presented to predict the turnaround time and the last release of TOBT during any of the three planning phases. The input feature differentiating the usability of the model in the tactical rather than in the strategic/pre-tactical phases is the $ATT = SOBT–AIBT$.

These ML models provide predictions in nominal conditions (i.e., in absence of high delayed inbound flights, etc.) according to the points described in Section III-F. The inclusion of outliers in the final dataset could be problematic especially when the model is intended to provide predictions in nominal conditions. A new data-driven approach for outlier identification has been developed during this project allowing to filter a small percentage of data samples (3-5% of the total) and improve significantly the model performances (Section III-F) Since the models are intended for deployment in operational environments, it becomes crucial to inform the user on the potential occurrence of certain exceptional events that were labelled as outliers. Therefore, the authors believe that, first, the user should have access to a detailed analysis of the operations that were detected as non-nominal by the approach described in Section III-F. This will allow to prepare in advance and to know the characteristics of the turnarounds that are more likely to present challenges as being non-nominal. Secondly, as an additional information to the user, a binary classifier with the newly labelled data could be later developed to inform the user on the probability of a specific observation to be an outlier or not. Overall, it is expected that the model will provide a support to the operations without excluding the intervention of the user in the process of TOBT updates.

A two-step approach has been developed to provide the potential user with additional information on each prediction (Section III-B). This approach allowed to introduce a new metric (*uncertainty*) which is defined as the time domain (centred around the peak) containing 95% of the probability distribution. While conventional metrics allow to assess the quality of the overall performances of a model, the *uncertainty* metric quantifies the level of confidence of each prediction. Interestingly it was found that the overall uncertainty of the testing data sample is proportional to the RMSE of the predictions (Figure 6).

The models have been validated on the testing datasets according to several indicators. In a hypothetical scenario where neither the model predictions nor complementary information is available, the user might rely on scheduled turnaround to assess the actual duration of turnarounds. As shown in Section IV-A, the predictions of the developed models lead to an improvement of 44.2-56.0% for the four models in the tactical phase and of 14.0-24.8% in the strategic/pre-tactical phases in terms of MAE when compared to the MAE derived from the use of scheduled turnarounds (baselines). However, it could be also interesting to compare the overall accuracy of the models with the error that a user would make by relying on the first (or following) TOBT release of the ground handlers as the final TOBT. A generic recommendation is to update the TOBT when varying by +/- 5 minutes. Using this criterion, we built two metrics for Prague and Arlanda airports demonstrating the validity of the predictions for operational purposes (Section V-A).

From an airport perspective, TT predictions could be used to improve airport capacity for level 3 coordinated airports that still have a linear declared capacity throughout the day. The model predictions could also be used to identify optimal turnaround buffers as well as to evaluate the investment for the creation of new aircraft positions. Based on the presented results, the authors believe that the development of a complete generalised model, that will be trained using the data from all the CDM airports, could become a useful tool not only for the operation planning of single airports but also for the operations of the Network Manager. In this implementation new input features could be potentially included. As an example, the duration of handling operations such as refuelling, catering, cleaning, etc. could be used as an input for the model. However, gathering this information might be problematic due to poor data sharing between the companies managing the ground handling activities and the other ATM stakeholders. Improving the predictability of turnaround duration might lead to reduction of delay due to ground operations and, therefore, to potential cost saving for all the stakeholders in ATM. Whether such a predictive model will lead to these results could be tested in a later work using simulators such as the Eurocontrol R-NEST (https://www.eurocontrol.int/solution/rnest).

References

[1] E. A. C. Team, "Airport cdm implementation - the manual," *EUROCONTROL, March*, 2017.

[2] M. Groppe, "Influences on aircraft target off-block time prediction accuracy," 2011.

[3] D. Huet, D. Booth, and S. Pickup, "A-cdm impact assessment-final report," *EUROCONTROL, March*, 2016.

[4] E. C. team, "Coda digest all-causes delay and cancellations to air transport in europe," *EUROCONTROL, Annual report for 2019*, 2019.

[5] A. Cook, "The management and costs of delay," in *European air traffic management*. Routledge, 2016, pp. 116–141.

[6] E. Cost, "Standard inputs for eurocontrol cost benefit analyses," 2013.

[7] A. Vidosavljevic and V. Tosic, "Modeling of turnaround process using petri nets," in *Air Transport Research Society (ATRS) World Conference*, 2010.

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

[9] H. Fricke and M. Schultz, "Improving aircraft turn around reliability," in *Third International Conference on Research in Air Transportation*, 2008, pp. 335–343.

[10] M. Schmidt, "A review of aircraft turnaround operations and simulations," *Progress in Aerospace Sciences*, vol. 92, pp. 25–38, 2017.

[11] Y. Li, J. Chen, and L. Feng, "Dealing with uncertainty: A survey of theories and practices," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2463–2482, 2012.

[12] A. Cook, G. Tanner, S. Cristobal, and M. Zanin, "Poem final technical report," SESAR, Tech. Rep., 2013.

[13] H. Fricke and M. Schultz, "Delay impacts onto turnaround performance," in *ATM Seminar*, 2009.

[14] J. Evler, E. Asadi, H. Preis, and H. Fricke, "Airline ground operations: Optimal schedule recovery with uncertain arrival times," *Journal of Air Transport Management*, vol. 92, p. 102021, 2021.

[15] E. Asadi, J. Evler, H. Preis, and H. Fricke, "Coping with uncertainties in predicting the aircraft turnaround time at airports," in *Operations Research Proceedings 2019: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR), Dresden, Germany, September 4-6, 2019*. Springer, 2020, pp. 773–780.

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

[17] C.-L. Wu and R. E. Caves, "Modelling and simulation of aircraft turnaround operations at airports," *Transportation Planning and Technology*, vol. 27, no. 1, pp. 25–46, 2004.

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

[19] R. Dalmau, P. De Falco, M. Spak, and J. Rodriguez, "Probabilistic Pretactical Arrival and Departure Flight Delay Prediction with Quantile Regression," in *15th USA Europe Air Traffic Management Research and Development Seminar*, Savannah, Georgia, 2023.

[20] L. S. Shapley *et al.*, "A value for n-person games," 1953.

[21] P. De Falco and L. Delgado, "Prediction of reactionary delay and cost using machine learning," in *Airline group of the International Federation of Operational Research Society (AGIFORS)*, 2021.

[22] S. Mas-Pujol, P. De Falco, E. Salamí, and L. Delgado, "Pre-Tactical Prediction of ATFM Delay for Individual Flights," in *41th Proceedings of the AIAA/IEEE Digital Avitonics Systems Conference (DASC)*, 2022.

[23] 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.

[24] J. Brownlee, *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery, 2016.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[27] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[28] M. C. Bryson, "Heavy-tailed distributions: properties and tests," *Technometrics*, vol. 16, no. 1, pp. 61–68, 1974.

[29] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.