Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions

Article *in* Journal of Air Transport Management · December 2019

DOI: 10.1016/j.jairtraman.2019.101737

Mihaela Mitici
Delft University of Technology
19 PUBLICATIONS **46** CITATIONS

Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions

Miguel Lambelho[a], Mihaela Mitici[a], Simon Pickup[b], Alan Marsden[c]

[a] *Faculty of Aerospace Engineering, Delft University of Technology, HS 2926 Delft, The Netherlands*
[b] *London Heathrow Airport, Nelson Road, TW6 2GW, United Kingdom*
[c] *Eurocontrol, Airport Research, Brussels, Belgium*

Abstract

The continuous growth of air traffic, together with limited existing airport infrastructures, have resulted in air traffic demand-capacity imbalances, airport congestion and arrival/departure flight delays. To address this, large European airports implement strategic flight schedules, i.e., up to 6 months prior to the day of the flight execution, flights are assigned arrival and departure time slots. The allocation of slots is performed by airport slot coordinators, which strive to accommodate the requests of airlines for arrival/departure slots, while taking into account airport's capacity and current IATA slot allocation guidelines. To support this process, several deterministic optimisation models for slot allocation have been proposed. However, the resulting flight schedules assume ideal conditions, .i.e., potential flight delays or cancellations are not considered. In this paper we propose for the first time a machine learning-based approach to assess the strategic flight schedules in terms of potential arrival/departure flight delays and cancellations. We first propose algorithms to classify the strategic, scheduled flights as delayed or cancelled. Further, we use these results as input for a generic flight schedule assessment methodology based on a relative on-time airport performance comparison of the considered strategic schedules. We demonstrate our methodology by assessing 10 strategic flight schedules in the period 2013-2018 at London Heathrow Airport, one of the busiest airports in Europe. Together with the development of dedicated strategic flight schedule optimization models, our proposed approach supports an integrated strategic flight schedule assessment, where strategic flight schedules are evaluated with respect to flight delays and cancellations.

Keywords: Strategic flight schedule, Delay prediction, Cancellation prediction, Machine learning, Schedule ranking

1. Introduction

The continuous growth of air traffic, together with limited airport expansion possibilities, have resulted in air traffic demand-capacity imbalances and arrival/departure flight delays at the largest airports in Europe. As an example, in 2017, the number of flights in Europe has increased by 4.3% relative to 2016 (EUROCONTROL, 2017). This corresponds to an additional 1191 flights per day on average. At the same time, 20.4% of the flights in 2017 experienced an arrival delay of 15min or more (EUROCONTROL, 2017).

To manage the demand-capacity imbalances, the busiest European airports make use of administrative demand management strategies to limit the number of flights scheduled to arrive/depart during busy hours. The main administrative demand management strategy currently in use is the airport slot allocation process, which follows the International Air Transport Association (IATA) Worldwide Slot Guidelines (International Air Transport Association, 2017). The IATA slot allocation process takes place two times per year (the Winter and the Summer season) and gives airlines the permission to use the full range of an airport's infrastructure to arrive or depart at a specific date and time at the airport. (Zografos et al., 2017; Pellegrini et al., 2017; Ribeiro et al., 2018) provide a detailed overview of the slot allocation process. The main inputs of the slot allocation process at an airport are 1) the requests of the airlines to access an airport's infrastructure for arrival or departure at a specific date and time, and 2) a deterministic, pre-defined airside and terminal capacity of the airport in terms of, for instance, maximum number of movements per day, per hour and per 15min. Following the IATA guidelines, the output of the slot allocation process is a strategic flight schedule containing the scheduled arrival and departure flight date and time up to 6 months prior to the day of the flight execution. The strategic schedule is in the form of a series of scheduled arrival and departure times. These series are commonly recurrent over a period of time, e.g., flight 123 is scheduled to arrive at 10AM every day from Monday to Thursday, in the months April and May.

In the past decade, several optimisation models to allocate slots to arriving/departing flights, have been proposed. (Zografos et al., 2012; Ribeiro et al., 2018) have developed optimization models for slot allocation at a single European airport. Network-wide slot allocation optimization models have been developed by (Castelli et al., 2012; Corolli et al., 2014; Pellegrini et al., 2017). The main objective of these models is to minimize the difference between the airlines' slot requests and the slots granted at the airport, following the IATA guidelines and taking into account the declared capacity limits of the airport. However, these models assume ideal conditions: the flights are assumed to be able to arrive and depart exactly within their scheduled slots, and the capacity of the airport is considered to be fixed, deterministic. In the day of the execution, however, flights often experience arrival/departure delays or cancellations. The strategic flight schedules, currently obtained following optimisation of the slots requests and the IATA guidelines up to 6 month prior to the day of the flight execution, do not give an indication on the potential flight delays and cancellations associated. In turn, the impact of the strategic schedules on the airport on-time performance is unknown at the moment of schedule generation. To address this, a methodology is needed to assess strategic flight schedules with respect to potential flight delays and cancellations and to provide airports with insights into potential performance bottlenecks. Such insights are particularly important to support the airport coordinators in developing strategic schedules that not only meet the IATA guidelines, but also enable a smooth and robust air traffic that benefits both airlines, airports and passengers.

In this paper we propose a machine learning-based approach to assesses the impact of strategic, IATA guidelines-compliant flight schedules on the on-time performance at an airport. In particular, we propose classification algorithms to predict whether flights scheduled in the strategic phase (6 months prior to the day of the execution) are subject to arrival/departure delays and cancellations during execution. Using the obtained flight delay and cancellation results, we propose a generic methodology to rank the strategic schedules by comparing and contrasting the associated flight delay and cancellation predictions. This analysis provides a means to assess strategic schedules based on their predisposition to have flight delays and cancellations. We demonstrate our assessment methodology using 10 strategic flight schedules from 2013-2018 at London Heathrow Airport (LHR), which is one of the busiest airports in Europe.

The contribution of this paper is three-fold. Firstly, we propose a machine learning-based algorithm to classify strategic, scheduled flights as being delayed or cancelled having a prediction horizon of 6 months. To the best of our knowledge, this is the largest prediction horizon assumed for flight delay and cancellation machine learning-based predictions. Most existing machine learning algorithms for flight delay and cancellation predictions assume a prediction horizon of only a few hours to a few days prior to the flight execution. Secondly, we characterize and compare the contribution of the classification features to a flight being classified as delayed or cancelled. Thirdly, we apply a generic methodology to assess strategic flight schedules and changes-to-schedule based on target Key Performance Indicators (KPIs) which are derived from the results of the flight delay and cancellation classification algorithms. The generality of our proposed assessment relies on the fact that we use a relative performance comparison between the assessed strategic schedules, rather than assigning user-defined weights to the target KPIs.

In summary, this paper provides an approach to assess the performance of strategic flight schedules with respect to associated flight delays and cancellations up to 6 months prior to the flight execution. To the best of our knowledge, we assess for the first time the robustness of strategic, IATA-compliant flight schedules with respect to delays and cancellations. Together with the development of dedicated optimization slot allocation models that minimize slot displacement in the presence of airport capacity constraints, our approach provides the airport coordinators with an integrated assessment of the performance of the IATA-compliant, airport slot allocation process.

The remainder of this paper is organized as follows. Section 2 discusses existing machine learning approaches for flight delay and cancellation predictions and their performance. Section 3 describes the flight schedules and the flight delay and cancellation data from LHR in the period 2013-2018. Section 4 presents our proposed machine learning approach for flight delay and cancellation classification. Section 5 describes a generic approach to assess strategic flight schedules based on KPIs that are derived in Section 4 using machine learning-based predictions for flight delay and cancellation. Section 6 discusses the implications of our results. as Section 7 provides conclusions and outlines future research directions.

2. Related work

The analysis of flight delays has been extensively addressed in the literature. (Mueller and Chatterji, 2002; Wu, 2014; Tu et al., 2008) propose data-driven models to estimate flight delay distributions at non-European airports. (Mueller and Chatterji, 2002) determines flight delay statistics for 10 major US airports by analyzing historical fight data. Based on these statistics, departure and arrival delays have been model as a Poisson process and a normal distribution, respectively. (Wu, 2014) estimates the probability density function of departure and arrival delays at Beijing Capital International Airport using historical flight delay data and an optimal generalized extreme value model. (Tu et al., 2008) proposes a statistical model to estimate flight departure delay distributions and seasonal trends at Denver International Airport. The authors consider in their model a seasonal trend, daily propagation patterns and random residuals. (Abdel-Aty et al., 2007) develops a frequency analysis to detect flight delay patterns at Orlando International Airport.

In the last years, an increasing number of studies have analyzed flight delays using machine learning approaches (Sternberg et al., 2017). Several studies (Kim et al., 2016; Choi et al., 2017; Alonso and Loureiro, 2015) consider a short prediction horizon of up to 1 day before the flight execution. (Kim et al., 2016) classifies delays at several US airports using recurrent neural networks and several weather-related features. The models achieve a classification accuracy of 0.874. (Choi et al., 2017) employs random forests and weather-related features to classify flight delay with an accuracy of 0.828.(Alonso and Loureiro, 2015) develops a multi-class classification algorithm to predict flight departure delay at Porto airport, achieving an accuracy of 0.57. One of the most important feature used for classification is the amount of delay experienced by the previous flight arrival.

(Choi et al., 2016; Belcastro et al., 2016; Horiguchi et al., 2017) propose machine learning approaches to classify flight delays with a prediction horizon of several days prior to the day of the flight execution. (Choi et al., 2016) achieves an accuracy of 0.268 using weather forecasts available 5 days prior to the day of the flight execution. The authors employ random forests classifier that are exclusively trained with weather-related features. (Belcastro et al., 2016) proposes a model to classify flights as being delayed exclusively as a result of unfavorable weather conditions. The authors use a balanced flight dataset, where a random under-sampling algorithm is used to decrease the number of delayed samples. The features considered are the scheduled departure/arrival time, the origin/destination airport and the weather conditions. The proposed random forests classifier obtains an accuracy of 0.858, with a recall of 0.869 with a 60 min delay threshold (a fight considered to be delayed if it has a delay of 60 min or more relative to the scheduled arrival time). (Horiguchi et al., 2017) considers a flight delay prediction horizon of 5 months before the day of the flight execution. A XGBoost classifier achieves an area under the ROC curve (AUC score) of 0.534 when predicting flight delay for 20 airports in Asia for a low-cost airline.

Several studies analyze flight delay cancellations at an airport. (Sridhar et al., 2009) proposes a neural network approach that aims at predicting the total aggregate number of flight cancellations. The accuracy of the predictions obtained is $r = 0.79$. Many studies also propose logit models to explain the influence of several variables on a flight being cancelled (Rupp and Holmes, 2006; Xiong and Hansen, 2009).

This paper expands this previous work on flight delay and cancellation prediction by developing machine learning classifiers to predict flight delays and cancellations with a 6-month prediction horizon at a large European airport. These flight delay and cancellation predictions are further used to assess strategic flight schedules on their impact on the airport's on-time performance.

3. Description of the case study data

The case studies presented in this paper are based on the strategic flight schedules, i.e., scheduled arrival and scheduled departure flight times, at London Heathrow Airport (LHR) in the period 30 March 2013 - 30 March 2018. These scheduled arrival and departure times correspond to 10 strategic slot allocation schedules, i.e., for each year in the period 2013-2018 there is a 6-month Summer Season schedule (end March to end September) and a 6-month Winter Season schedule. These schedules are the result of the IATA slot allocation process at LHR: airlines submit requests for arrival/departure slots; the airport coordinator grants slots while taking into account the available capacity at LHR and the IATA guidelines for slot allocation (International Air Transport Association, 2017).

Table 1 gives an example of scheduled flights at LHR, following the IATA slot allocation process. Each arrival/departure flight is assigned an arrival/departure time between 6:30-24:00, an interval of dates when the flight is scheduled for arrival/departure, the frequency of the arrival/departure over the indicated interval of dates, an LHR terminal (at LHR there are 4 passenger terminals and a cargo terminal), and the destination (origin) airport for an flight departing from (arriving at) LHR. The type of aircraft assigned to a scheduled flight is also known from the slot request submitted by the airline. For example, flight KL1031 in Table 1 is scheduled for arrival at 17:55 every day from Monday to Saturday in the period 01.04.2013-01.07.2013.

FlightID	Arrival/ Departure	Time	Start Date	End Date	Days of the week	Terminal	Origin/ Destination	Aircraft
KL1031	Arr.	1755	01.04.2013	01.07.2013	123456·	4	AMS	73W
BA830	Dep.	0930	01.04.2013	20.06.2013	12··567	1	DUB	320
DL100	Arr.	0800	01.04.2013	13.05.2013	1·· 45 ·7	3	JFK	764

Table 1: Example of strategic flight schedule with flights scheduled to arrive at/depart from LHR.

We consider 2.3 million flights scheduled to arrive and depart from LHR in the period March 2013 - September 2018. There are 177 types of aircraft assigned for these flights and a total of 542 distinct origin/departure airports for the flights arriving/departing from LHR. Moreover, 25% of these flights are short-haul, with a flown distance of at most 700 km, 25% of the flights have a flown distance of 700 – 1400 km, 30% of the flights 1400 – 6000 km, and 20% of the flights more than 6000 km.

Flight	Imbalance ratio	Delay STD (min)
Arrivals	3.15	44.68
Departures	3.70	32.87
Cancellations	58.88	-

Table 2: Actual delays and cancellations of flights that have been scheduled in the 2013-2018 strategic schedules.

We say that an arrival (departure) flight has an arrival/departure *delay* if, during execution, this flight arrives (departs) 16min or more after the scheduled time of arrival (departure) (EUROCONTROL, 2017).

We say that an arrival (departure) flight is *cancelled* if this flight is not executed in the day when it was scheduled to arrive (depart).

4. Machine learning algorithms for flight delay and cancellation prediction

In this Section, we present a machine learning approach to predict whether strategic, scheduled arrival/departure flights are delayed or cancelled. These predictions are based on strategic flight schedules from LHR and assume a 6-month prediction horizon, i.e., we predict whether flights are delayed or cancelled 6 months prior to the day of the flight execution. We make use of classification algorithms to predict: arrival flight delay, departure flight delay and flight cancellation. In Section 4.1 we discuss the selection of features used as input for the flight classification algorithms. Section 4.2 presents the performance of three flight classification algorithms: LightGBM, multilayer perceptron (MLP) and random forests (RF). In Section 4.3 we make use of model-agnostic interpretability methods to explain the predictions yielded by the classification algorithms.

4.1. Feature Selection

In this section we discuss the selection of features used to classify strategic, scheduled flights as being delayed or cancelled.

We select the features used for the flight delay and cancellation classification algorithms using the recursive feature elimination (RFE) method (Guyon et al., 2002), which recursively eliminates weak features using cross-validation and scoring of features subsets. Table 3 shows the RFE-based features selected to classify an arrival/departure flight with a 6-months prediction horizon. Table 4 provides a detailed description of these features.

The features Airline, Terminal, Aircraft, Airport, Year, Month, Hour, Day of year, Day of month, Day of week are obtained directly from the strategic flight schedules, which are compliant with the IATA slot allocation guidelines. The features Distance, Country and Seats are derived after processing the 6-months strategic schedules. We have also considered several other features such as the number of aircraft present at the airport at the moment of/1 hour before/1 hour after the flight arrival/departure time, the arrival stack, the standard instrumental departure (SID) route, the turnaround time determined as the difference between the scheduled departure time and the scheduled arrival time. However, these features have been eliminated by the RFE feature elimination algorithm, i.e., these features did not further improve the performance of the classification algorithms.

Classifier	**Features**
Departure Delay	Airline[a], Terminal[a], Aircraft[a], Distance, Airport[c], Country[a], Seats, Year, Month[b], Hour[b], Day of year[b], Day of month, Day of week
Arrival Delay	Airline[a], Terminal[a], Aircraft[a], Distance, Airport[c], Country[a], Seats, Year, Month[b], Hour[b], Day of year[b], Day of month, Day of week
Flight Cancellation	Airline[a], Terminal[a], Aircraft[a], Distance, Airport[a], Country[a], Year, Hour[b], Day of year[b], Day of month, Day of week

[a] Feature prepossessed with the target encoding method

[b] Feature transformed by trigonometric functions

[c] Categorical feature encoded using geographic coordinates

Table 3: Feature selection for flight delay and cancellation classifiers with a 6-months prediction horizon.

Features	Feature type	Feature description
Airline	C	airline operating the flight
Terminal	C	arrival/departure airport terminal assigned to a flight
Aircraft	C	aircraft type
Distance	N	distance between origin and destination airport (km)
Airport	C	origin/destination airport of the flight
Country	C	country of origin/destination airport
Seats	N	number of seats of the aircraft assigned to a flight
Year	N	scheduled year of flight arrival/departure
Month	N	scheduled month of flight arrival/departure
Hour	N	scheduled hour of the day of flight arrival/departure
Day of year	N	scheduled day of the year of flight arrival/departure
Day of month	N	scheduled day of the month of flight arrival/departure
Day of week	N	scheduled day of the week of flight arrival/departure
Arrival ATFM delay	N	daily average ATFM arrival delay (min)

Table 4: Description of features used for flight delay and cancellation classification algorithms - 6 months prediction horizon. C=Categorical, N=Numerical.

The categorical features Airline, Terminal, Aircraft and Country are encoded using the target encoding method (Micci-Barreca, 2001). The categorical feature Airport has been encoded both using the geographic coordinates of the airport and target encoding. Binary encoding and one-hot encoding methods have not been employed due to the high cardinality of the categorical features. Ordinal encoding has not been used either since it misleadingly assumes an order within a feature. For example, an ordinal encoding of the airlines such as $1, 2, 3, ...$ would mean that an airline encoded as 1 is more similar to an airline encoded as 2 than an airline encoded as 8. Table 5 gives an example for each of the mentioned encoding methods, where one-hot encoding uses strings of bits with only one high bit (1) and the rest low bits (0) for each airline type, ordinal encoding uses ordered integers for each airline type, binary encoding uses binary strings of bits for each airline type. Lastly, the target encoding method (Micci-Barreca, 2001), taking the case of departure delay classifiers, encodes an airline type based on the probability that a flight from this airline is delayed (the target). For example, in Table 5, the airline BA is encoded as $\frac{2}{3} = 0.67$ since there are 2 BA flights delayed from a total of 3 BA flights.

Airline	Delayed	One-hot encoding	Ordinal encoding	Binary encoding	Target encoding
TAP	Yes	100	1	00	0.5
KLM	No	010	2	01	0
BA	Yes	001	3	10	0.67
TAP	No	100	4	00	0.5
BA	Yes	001	5	10	0.67
BA	No	001	6	10	0.67

Table 5: Example of encoding methods for feature Airline and classifier departure delay. Here we consider 3 airlines, i.e., TAP, KLM, BA, and a total of 6 flights.

The features Hour, Day of year and Month have been transformed by trigonometric functions to account for periodicity (Horiguchi et al., 2017). For example, for a specific hour of the day t, we use the trigonometric functions $sin\left(\frac{2\pi t}{24}\right)$ and $cos\left(\frac{2\pi t}{24}\right)$ to ensure a 24hrs periodicity. As a consequence, t=24:00 and t=1:00 become sequential hours. Similarly, we ensure a periodicity of 365 days for the feature Day of the year and a periodicity of 12 for the feature Month.

The feature Arrival ATFM delay is a daily average feature, i.e, assumes the same value for all flights in a specific day of operations, and corresponds to the duration between the last Estimated Take-Off Time (ETOT) and the Calculated Take-Off Time (CTOT) allocated by the European ATM Network Manager. A positive value of this parameter indicates traffic congestion due to, for instance, weather conditions.

The feature Seats has a range between 4 seats (for regional jets) to 800 seats (for A380-800). This feature has been derived from the type of aircraft assigned to a flight.

4.2. *Flight delay and cancellation classification algorithms*

In this Section we present three machine learning classification algorithms to classify flight delays and cancellations 6 months in advance of the day of the flight execution: LightGBM, multilayer perceptron (MLP) and random forests (RF). These three algorithms belong to different machine learning types of algorithms: gradient boosting decision tree, neural networks and random decision forests, respectively. We make use of three different classification algorithms to cross check our results.

LightGBM (Ke et al., 2017) is a tree-based machine learning algorithm where ensembles of decision trees are trained in sequence by fitting negative gradients of the loss. LightGBM uses Gradient-based One-Side Sampling, which excludes data instances with small gradients, and Exclusive Feature Building, which bundles mutually exclusive variables, thus, reducing the number of features. To estimate the hyperparameters that yield the best performance, we use the Python library `hyperot` (Bergstra et al., 2013) to optimize the $f1$-score metric Duda et al. (2012), i.e., the harmonic mean between precision and recall, by performing Bayesian optimization. Table 6 shows the hyperparameters of the LightGBM classifiers. The best performance is achieved with a high learning rate and with a relatively small number of decision trees.

Classifier	Number input features	Learning rate	Max depth of tree	Trees	Subsample	Weight positive class
Dep. Delay	18	0.1	15	300	0.578	2.265
Arr. Delay	18	0.1	39	200	0.573	2.100
Cancellation	14	0.1	30	100	0.781	7.300

Table 6: Hyperparameters of LightGBM flight delay and cancellation classifiers with a 6-month prediction horizon.

Multilayer perceptron (MLP) (Hinton, 1990) is a feed-forward neural network that has consecutive layers with adaptive weights. The vector of inputs of MLP was normalized $N(0,1)$. The initialization of the weights follows a normal distribution $N(0,0.01)$. To increase the stability of the neural network, all the hidden layers have batch normalization. Table 7 shows the hyperparameters of the MLP classifiers. All classifiers produced superior results when trained with two hidden layers and with the Adam optimizer (Kingma and Ba, 2014). Additionally, dropout layers were included to reduce overfitting. The small learning rate used increased the computational time of the MLP classifiers when compared with the LightGBM models, as shown in Table 10.

Classifier	Number input features	Number neurons for each layer	Batch size	Dropout rate	Learning rate
Dep. Delay	18	100 → 100	1000	0.15	1.0×10^{-4}
Arr. Delay	18	110 → 110	1000	0.05	1.0×10^{-3}
Cancellation	14	150 → 150	1000	0.05	1.0×10^{-6}

Table 7: Hyperparameters of MLP flight delay and cancellation classifiers with a 6-month prediction horizon, Adam optimizer.

Random Forests (RF) (Breiman, 2001) is an ensemble learning method that aggregates dissimilar decision trees. When building a forest tree, only a random part of the training set is used to build each tree. To increase the ensemble diversity, further randomness is introduced when building each tree by selecting a fraction of the the total number of features. Once all forest trees are created, the classification of each sample in the test set is executed by combining the predictions of each tree through majority voting. Table 8 shows the hyperparameters of the RF classifiers.

Classifier	Number input features	Number trees generated	Max depth of tree	Percentage features for each split
Dep. Delay	18	500	11	0.60
Arr. Delay	18	1000	12	0.55
Cancellation	14	500	10	0.60

Table 8: Hyperparameters of RF flight delay and cancellation classifiers with a 6-month prediction horizon.

Before presenting the results of the classification algorithms, we introduce the following metrics. We consider the True Negatives (TN), the False Positives (FP), the False Negatives (FN) and the True Positives (TP) (Marsland, 2011; Duda et al., 2012; Hossin and Sulaiman, 2015), where TN is the number of actual non-delayed flights that are predicted to be non-delayed, FP is the number of actual non-delayed flights that are predicted to be delayed, FN is the number of actual delayed flights that are predicted to be non-delayed and, TP is the number of actual flights that are delayed and are predicted to be delayed. Then, we determine Accuracy = $\frac{TP+TN}{TN+FP+FN+TP}$, Recall = $\frac{TP}{TP+FN}$, Precision = $\frac{TP}{TP+FP}$, $f1$-score is the harmonic mean between Precision and Recall, and AUC, i.e., the area under the curve determined by the rate of TP and FP Marsland (2011).

Table 9 shows the performance of LightGBM, MLP and RF to classify arrival/departure flights as being delayed and cancelled. We note that the prediction horizon is 6 months prior to the day of the flight execution. A 5-fold cross validation is performed using the data on the flights arriving and departing in the period 2013-2018 from LHR. Among the 3 classification algorithms, LightGBM performs the best with respect to accuracy, precision, recall and area under the ROC curve (AUC) (Duda et al., 2012).

Classifier	Metric	LightGBM Mean	LightGBM STD	MLP Mean	MLP STD	RF Mean	RF STD
Departure Delay	Accuracy	0.794	5.8×10^{-3}	0.772	3.8×10^{-3}	0.771	9.3×10^{-4}
	Precision	0.516	2.1×10^{-3}	0.467	7.6×10^{-3}	0.460	2.4×10^{-3}
	Recall	0.516	2.2×10^{-3}	0.488	1.0×10^{-2}	0.455	2.7×10^{-3}
	$f1$-score	0.516	1.4×10^{-3}	0.478	2.8×10^{-3}	0.458	1.9×10^{-3}
	AUC	0.786	1.0×10^{-3}	0.754	2.4×10^{-3}	0.744	8.3×10^{-4}
Arrival Delay	Accuracy	0.791	5.0×10^{-4}	0.771	3.9×10^{-3}	0.759	8.0×10^{-4}
	Precision	0.567	2.2×10^{-3}	0.525	9.3×10^{-3}	0.500	3.0×10^{-3}
	Recall	0.553	2.2×10^{-3}	0.527	1.1×10^{-2}	0.515	1.3×10^{-3}
	$f1$-score	0.560	1.6×10^{-3}	0.526	3.6×10^{-3}	0.507	1.8×10^{-3}
	AUC	0.803	1.2×10^{-3}	0.774	2.1×10^{-3}	0.758	1.4×10^{-3}
Cancellation	Accuracy	0.987	1.8×10^{-4}	0.984	2.0×10^{-4}	0.984	8.5×10^{-5}
	Precision	0.608	3.8×10^{-3}	0.532	1.5×10^{-2}	0.529	7.0×10^{-3}
	Recall	0.592	2.0×10^{-3}	0.530	1.0×10^{-2}	0.529	3.0×10^{-3}
	$f1$-score	0.600	1.8×10^{-3}	0.531	8.3×10^{-3}	0.529	4.1×10^{-3}
	AUC	0.929	4.3×10^{-3}	0.840	7.2×10^{-3}	0.862	3.3×10^{-3}

Table 9: 5-fold cross validation results for machine learning models with 6-month prediction horizon.

All classifiers achieve an accuracy of 0.75 or higher for flight delay classification and 0.98 or higher for cancellation classification. We note that the prediction horizon is 6 months prior to the day of the flight execution. Given that the training and test data consists of a larger number of negatives, i.e., not-delayed/not-cancelled flights, than positives, i.e., delayed/cancelled flights (see also the imbalance ratios in Table 2), it is necessary to analyze closely both the recall, i.e., how many of the true positives are predicted as positive, and the precision, i.e., how many of the predicted positives are correctly predicted. As such, when estimating the hyperparameters of the classification algorithms, the aim has been to obtain the highest value of the $f1$-score and similar values for precision and recall, i.e., an equal number of false negatives and false positives.

Figures 1, 2 and 3 show the ROC curves achieved using LightGBM, MLP and RF for the three classifiers: arrival flight delay, departure flight delay and flight cancellation classifiers. Again, the results obtained using LightGBM show the largest AUC, i.e., the ability to identify actual delayed (not delayed) flights as delayed (not delayed) Duda et al. (2012).

The computational times required for the LightGBM, MLP and RF classifiers are given in Table 10. All the three classifiers have been trained and tested on the same dataset (see Section 3). The RF classifiers require the most computational time, despite the small depth of the trees generated and the low percentage of features considered for each split (see Table 8).

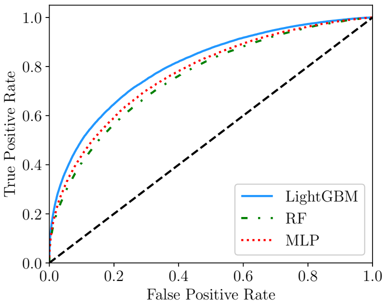

Figure 1: ROC curves of flight arrival delay classifiers.

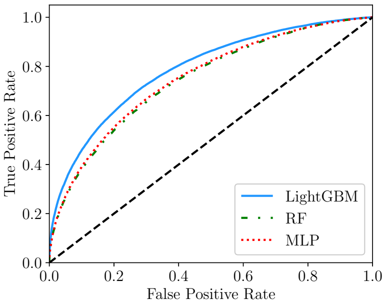

Figure 2: ROC curves of flight departure delay classifiers.

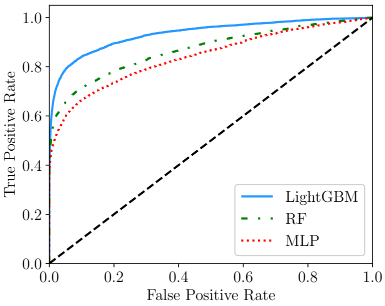

Figure 3: ROC curves of flight cancellation classifiers.

4.3. *Model-agnostic interpretability - LightGBM flight classification algorithms*

In this Section we interpret the results yielded by the LightGBM flight classifiers, which has the best performance with respect to accuracy, precision and recall (see Table 9).

Classifier	LightGBM (sec)	MLP (sec)	RF (sec)
Departure Delay	55.91	715.43	943.97
Arrival Delay	33.66	1554.75	1874.01
Cancellations	27.31	2131.48	1849.15

Table 10: Computational time for LightGBM, MLP and RF classifiers.

To determine the impact of a feature on the output of the LightGBM classifiers, we determine the Shapley additive explanations (SHAP) value (Lundberg and Lee, 2017) of a feature i, which we denote as ϕ_i, for every flight classified, as follows (Lundberg and Lee, 2017):

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f(S \cup \{i\} - f(S))],$$

where F is the set of all features considered for the classification algorithm, $S \subseteq F$ is a subset of features obtained from the set F except feature i, and $f(S)$ is the expected classification output given by the set S of features.

The SHAP values show which features have a significant positive or negative impact on the delay/cancellation flight classification and what is the magnitude of the impact, i.e., how much a specific feature value drives the classification of a flight as delayed/cancelled. For a specific flight, a large positive (large negative) SHAP value of a feature indicates that this feature has a large contribution for the flight to be classified as delayed/cancelled (not delayed/cancelled). A SHAP value of a feature close to zero indicates that this feature does not contribute/does not help deciding in classifying a flight as being delayed/cancelled or not. In this paper, the SHAP values are expressed in *log odds*, where the log odd of a variable A is defined as:

$$log\left(\frac{P(A)}{1 - P(A)}\right), \text{ with } P(A) < 1.$$

Figures 4, 5 and 6 are summary plots that show the SHAP values for all features for *all* flights considered for classification, i.e., these figures show an aggregation of dots, where each dot corresponds to a flight. For a given feature, each dot corresponds to a flight and an associated SHAP value. For a given feature, the color blue of a dot (flight) indicates that, for this flight, the value of the feature is small, while the color red indicates that the value of the feature is large. For example, in Figure 4, for the feature Arrival ATFM Delay, the dots (flights) colored red have large Arrival ATFM delays, while the dots (flights) colored blue have small Arrival ATFM delays. For a given feature, an accumulation of dots indicates that there is a large number of flights that have similar SHAP values. As an example, in Figure 4, for the feature Arrival ATFM Delay, there is a significant number of flights where this feature has a SHAP value between -1 and 0, i.e., an accumulation of blue dots corresponding to a SHAP value between -1 and 0. Again, the color blue indicates that all these flights have a low Arrival ATFM delay. The blue dots (flights) that have a negative SHAP value are those flights with a low Arrival ATFM delay (blue color) and that are classified as not delayed (negative SHAP). In particular, for the blue dots (flights) where the SHAP value is close to zero, the Arrival ATFM delay is low (blue color), but it does not significantly impact the classification of these flights (SHAP value close to zero).

In Figures 4, 5 and 6, the features are sorted by the sum of the SHAP values magnitudes over all samples such that the feature at the top of the graph has the highest impact on the flight classification, whereas the feature at the bottom of the graph has the lowest impact. For example, in Figure 4, the feature Arrival ATFM delay is at the top of the graph since it has the highest impact on the flight delay classification. The features Hour and Airline have the second and third largest impact on the flight delay classification. Similarly, Figure 5 shows that the features Arrival ATFM Delay, Airline and Hour have the highest impact on the arrival flight delay classification. Both Figures 4 and 5 show that the feature Seats also has a high importance for both arrival and departure delay classifiers. The feature Terminal has the lowest feature importance for departure flight delay classification. For the arrival flight delay classification, however, the feature Terminal has a larger importance, whereas the features Month and Day of the month have the lowest feature importance. Figure 6 shows that the features origin/destination Airport and Airline have the highest feature importance for the flight cancellation classification algorithm. The feature Aircraft also have a high importance in the cancellation classifier when compared with the flight delay classifiers. The feature Day of the month shows the lowest feature importance from for the cancellation classifier.

Figures 4, 5 and 6 also allow for a detailed analysis of the impact of each feature. For a given feature, the color red used of a dot, i.e., flight, and a corresponding large SHAP value shows that large values (red) of this feature are very significant (large SHAP value) for the classification. For example, Figure 4 shows that for the feature Arrival ATFM Delay, there are several dots, i.e., flights, that are red and that have a positive and large SHAP value. This means that a large (red) value for the Arrival ATFM Delay is very significant (large SHAP value) and drives the classification of a departure flight as being delayed (positive SHAP value). In Figure 4 there is a larger accumulation of blue dots (flights), with SHAP values between -1 and zero. The color blue indicates that these departure flights have low Arrival ATFM Delays. Here, the blue dots (flights) with negative SHAP values away from zero indicate that, for these flights, small (blue) Arrival ATFM Delays drive the classification of these departure flights as being not delayed (negative SHAP value). Also, the blue dots (flights) with negative SHAP values close to zero indicate that, for these departure flights, the small (blue) Arrival ATFM Delays do not significantly impact the classification of these flights. In Figure 4, for feature Arrival ATFM Delay, we also note that for the dots (flights) with SHAP values around zero, i.e., the feature does not significantly drive the classification of a flight as being delayed or not delayed, the values of the Arrival ATFM Delay are low (blue color). Thus, low Arrival ATFM Delays have a low classification importance.

A similar analysis can be made for all feature in Figures 4, 5 and 6. We note that for the categorical features Airline, Country, Aircraft, Terminal which are encoded using the target encoding method, high feature values means high probabilities of delay. Here, it can be seen that, for these features, high values of these features, i.e., high probabilities of delay, correspond to high SHAP values. For the features encoded with trigonometric functions (see also Section 4.1), we note that a detailed analysis of the summary plots is not straightforward as we apply sin and cos transformations. As such, for these features, we make use of the summary plots to determine the feature importance, as discussed above.

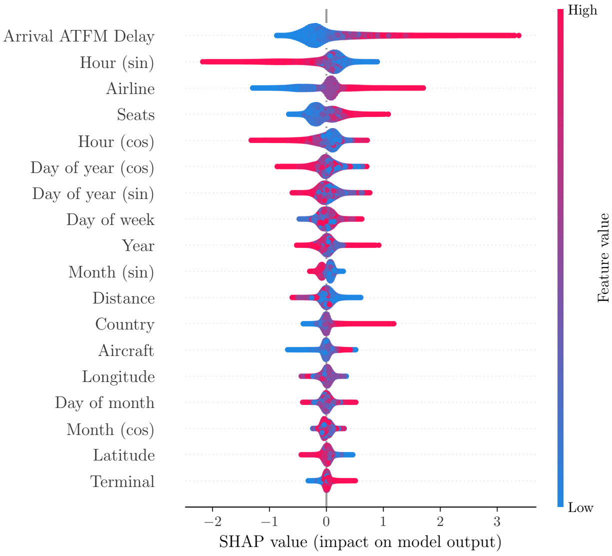

Figure 4: SHAP values (log odds) of the features used for delayed departure flight classification using LightGBM - 6 months prediction horizon.

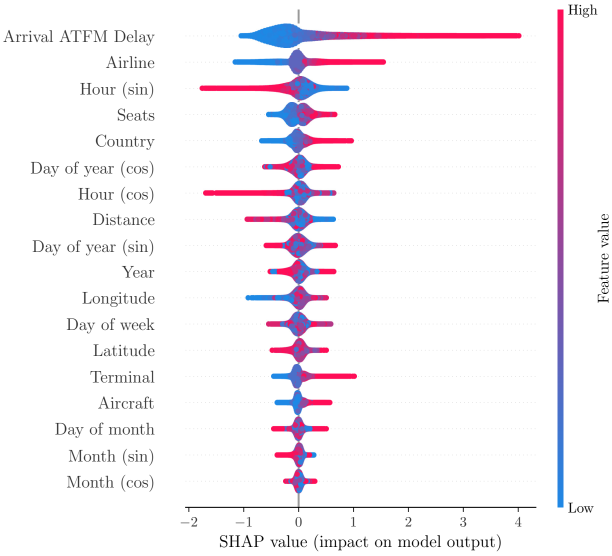

Figure 5: SHAP summary plot (log odds) of the features used for delayed arrival flight classification using LightGBM - 6 months prediction horizon.

5. Assessing strategic flight schedules using machine-learning flight delay and cancellation predictions

In this Section we present a generic approach to assess strategic flight schedules in terms of associated flight delays and cancellations. In Section 5.1 we present a generic method to assess flight schedules based on a set of general KPIs. Rather than assigning weights to the considered KPIs, which are user-specific and may be difficult to define, we propose a generic ranking of the schedules based on the relative improvement in KPIs of some schedules against others. However, in the case of strategic flight schedules, i.e., 6 months prior to the flight execution, the value of flight delay and cancellation-related KPIs are not known. In fact, a strategic flight schedule does not give any indication on the associated number of delayed/cancelled flights. To address this, in our pro-

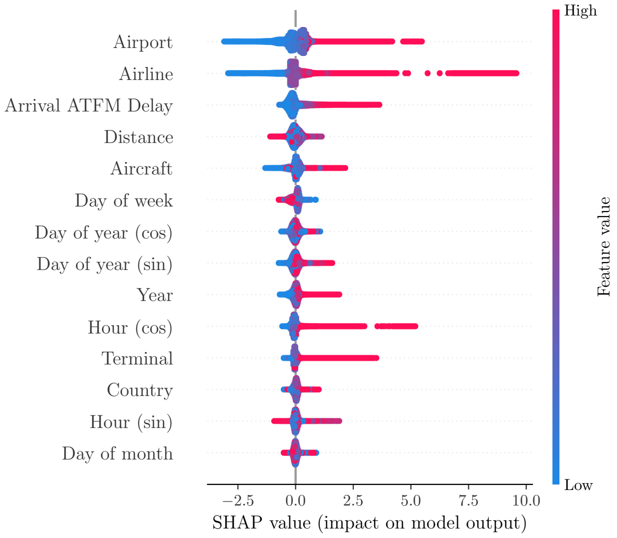

Figure 6: SHAP values (log odds) of the features used for flight cancellation classification - LightGBM, 6 months prediction horizon.

posed schedule assessment we make use of KPIs derived from the flight delay and cancellation predictions (see Section 4). In Section 5.2 we apply this assessment method to rank 10 strategic schedules using KPIs derived from machine learning-based flight delay and cancellation predictions. The numerical examples in this section are based on strategic flight schedules from LHR in the period 2013-2018.

5.1. Generic schedule assessment methodology

In this Section we apply an assessment methodology for strategic flight schedules using a set of general KPIs. This assessment can be done both before and after the execution of the flights, as long as the values of the KPIs are known. When assessing the flight schedules, we make use of the notion of schedule domination, which we define below, rather than assuming user-defined weights for the KPIs considered. Thus, we propose a generic assessment methodology that does not depend on the weights of the KPIs, which are user-specific.

We characterize a strategic flight schedule i by a set of n KPIs, i.e. $S_i : (KPI_1^i, \dots, KPI_n^i)$. We are interested in those schedules where the values of all n KPIs are *minimal*. To this end, we define the concept of schedule domination as follows. We say that schedule i, $S_i : (\mathrm{KPI}_1^i, \mathrm{KPI}_2^i, \dots, \mathrm{KPI}_n^i)$, dominates schedule j, $S_j : (\mathrm{KPI}_1^j, \mathrm{KPI}_2^j, \dots, \mathrm{KPI}_n^j)$, if: $\forall u \in \{1, 2, \dots, n\}$, $\mathrm{KPI}_u^i \leqslant \mathrm{KPI}_u^j$ and there exists at least one KPI $K_m, m \in \{1, 2, \dots, n\}$ such that $\mathrm{KPI}_m^i < \mathrm{KPI}_m^j$ (Boyd and Vandenberghe, 2004).

We consider the set $\mathscr{S} = \{S_1, \dots, S_N\}$ of N schedules. The Pareto front of the schedules $i \in \mathscr{S}, 1 \le i \le N$, with respect to the KPIs $KPI_1, \dots, KPI_n$, is the subset $\mathscr{S}_1$ of schedules that are not dominated by any other schedule (Boyd and Vandenberghe, 2004), where $\mathscr{S}_1 \subset \mathscr{S}$. We say that layer 1, which we denote by L^1, consists of all the schedules in $\mathscr{S}_1$. We next partition the set of remaining schedules $\mathscr{S} \setminus \mathscr{S}_1$ into additional layers. We define layer 2, i.e., L^2, of the schedules $i \in \mathscr{S}, 1 \le i \le N$ as the set of schedules that are in the Pareto front of the schedules $\mathscr{S} \setminus \mathscr{S}_1, \mathscr{S} \setminus \mathscr{S}_1 \neq \emptyset$. In general, we define layer m, denoted by L^m, of the schedules $i \in \mathscr{S}, 1 \le i \le N$ as the set of schedules in the Pareto front of the schedules $\mathscr{S} \setminus (\mathscr{S}_1 \cup \cup \mathscr{S}_2 \cup \ldots \cup \mathscr{S}_{m-1})$, with $\mathscr{S} \setminus (\mathscr{S}_1 \cup \mathscr{S}_2 \cup \ldots \cup \mathscr{S}_{m-1}) \neq \emptyset$.

Figure 7 shows an example of dominance relationships between 6 schedules $S_1, S_2 \dots, S_6$. Schedules S_1, S_2, S_3 form layer 1. Schedules S_4, S_5 form layer 2. Schedule S_6 is layer 3. Figure 7 also shows the dominance boundaries for each schedule, i.e., the bounds of the set of points that a schedule dominates. All the schedules $i, 1 \le i \le 6$, located within the dominance boundaries of a given schedule $j \neq i, 1 \le j \le 6$, are dominated by schedule j. Here, schedules S_1, S_6 and S_3 do not dominate any other schedule, schedules S_4 and S_5 dominate schedule S_6, schedule S_2 dominates schedules S_4, S_5 and S_6.

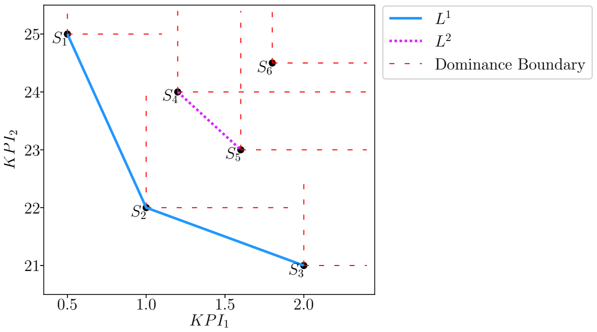

Figure 7: Example of schedule dominance for 2 KPIs.

We next define the dominance power of a schedule $i \in \mathscr{S}, 1 \le i \le N$, as introduced in Valkanas et al. (2014). We say that the dominance power of schedule $i, i \in \mathscr{S}$, which we denote by $D(S_i)$, is as follows:

$$D(S_i) = \sum_{\substack{k=1 \\ k \neq i}}^{N} \frac{1}{L^j} \mathbf{1}_{(S_i \text{ dominates } S_k) \cap (S_k \in L^j)}, 1 \le j \le m, \quad (1)$$

where $\mathbf{1}_A$ is an indicator function that takes value 1 if

A is true and zero otherwise.

As an example, in Figure 7, $D(S_2) = \frac{1}{2} + \frac{1}{2} + \frac{1}{3}$ since S_2 dominates S_4 from layer 2, S_5 from layer 2 and S_6 from layer 3; $D(S4) = D(S_5) = \frac{1}{3}$ because both S_4 and S_5 dominate S_6 from layer 3; and $D(S_1) = D(S_3) = D(S_6) = 0$ since S_1, S_3, S_6 do not dominate other schedules.

Lastly, we rank the strategic flight schedules $i \in \mathscr{S}, 1 \leq i \leq N$ based on their dominance power.

We are interested in those schedules with the highest dominance power. We assign a ranking position of 1 for the schedule(s) with the highest dominance power, a ranking position of 2 for the schedule(s) with the second highest dominance power and so on.

5.2. Assessing strategic flight schedules - results

In this Section we assess 10 strategic flight schedules using the methodology introduced in Section 5.1. In doing so, we consider 5 KPIs, which are commonly used in practice to analyse flight schedules: 1) percentage of flights cancelled, 2) percentage of departure flights delayed, 3) percentage of arrival flights delayed, 4) percentage of days in which the number of flights delayed is equal or higher than 25% (percentage departures yellow days) and, 5) percentage of days in which the number of flights delayed is equal to or higher than 25% (percentage of arrivals yellow days).

Since the strategic flight schedules do not give an indication on the potential flight delays and cancellations, the KPIs above are unknown prior to the execution of the flights, when the schedules are defined. As such, the schedule coordinators do not have an insight into potential flight delays and cancellations when setting the strategic schedules. To address this, we derive the flight delay and cancellation-related KPIs above using machine learning flight delay and cancellation classification algorithms (see Section 4).

Figure 8 shows the percentage of predicted cancelled flights and the percentage of predicted delayed arrival flights for each of the 10 strategic schedules. Layer 1 consists of schedules $\{S_1, S_3, S_4, S_9\}$. Layer 2 consists of schedules $\{S_2, S_5, S_6, S_7, S_8\}$. Layer 3 consists of schedule $\{S_{10}\}$. Using Equation 1, we determine the dominance power of these 10 strategic schedules when taking into account the percentage of cancelled flights and the percentage of delayed arrival flights. Table 11 shows the dominance power of the schedules. Table 11 shows that schedules S_3, S_1, S_4 have the best performance with respect to flight cancellations and delayed arrival flights. We note that all three schedules are part of the Pareto front (layer 1).

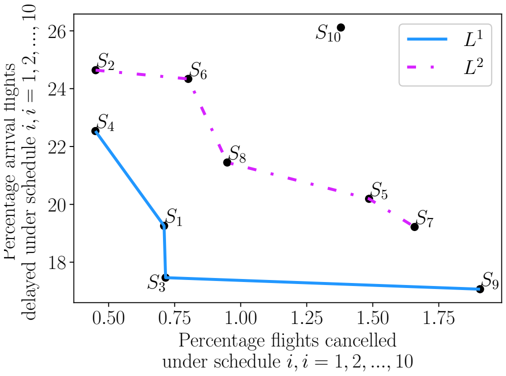

Figure 8: Layers of the strategic flight schedules when considering the percentage of predicted flights cancelled and the percentage of predicted delayed arrival flights.

	S_3	S_1	S_4	S_2	S_6	S_8	S_{10}	S_5	S_7	S_9
$D(S_i)$	2.33	1.83	1.33	0.33	0.33	0.33	0	0	0	0

Table 11: Dominance power of the 10 strategic schedules $S_1, ..., S_{10}$, when considering the percentage of cancelled flight and percentage of delayed arrival flights.

Table 12 shows the ranking of the 10 strategic schedules based on the dominance power of the schedules (see Equation 1). When the dominance power of two or more schedules is the same, we further discriminate between these schedules by ranking them based on the percentage of predicted flight cancelled, where the schedule with the lowest percentage of flight cancellations is preferred. Table 12 shows the schedule ranking obtained using i) the predicted values KPIs, as a result of the machine learning algorithms, and ii) the values of the KPIs from the actual flight data. The results show that the best-ranked 4 schedules when considering the actual flight data are also captured by the schedule ranking when using the predicted values of the KPIs.

Ranking position	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
KPIs prediction models	S_3	S_1	S_4	S_2	S_6	S_8	S_{10}	S_5	S_7	S_9
KPIs real-data	S_1	S_3	S_2	S_4	S_7	S_5	S_6	S_8	S_{10}	S_9

Table 12: Schedule ranking with respect to percentage of flights cancelled and percentage of arrival delays under schedule S_i.

Figure 9 shows the percentage of predicted cancelled flights and the percentage of predicted delayed departure flights for each of the 10 strategic schedules. Layer 1 consists of schedules $\{S_1, S_3, S_4\}$. Layer 2 consists of schedules $\{S_2, S_5, S_6, S_7, S_8\}$. Layer 3 consists of schedules $\{S_9, S_{10}\}$.

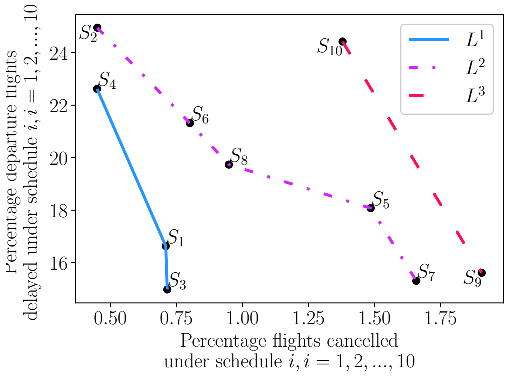

Figure 9: Layers of the strategic flight schedules when considering the percentage of predicted flights cancelled and the percentage of predicted delayed departure flights

Table 13 shows the dominance power of the schedules, where schedules S_3, S_1, S_4 have the best performance with respect to flight cancellations and delayed departure flights.

	S_3	S_1	S_4	S_6	S_8	S_7	S_2	S_{10}	S_5	S_9
$D(S_i)$	2.67	1.83	0.83	0.33	0.33	0.33	0	0	0	0

Table 13: Dominance power of the 10 strategic schedules $S_1, ..., S_{10}$, when considering the percentage of cancelled flight and percentage of delayed departure flights.

Table 14 shows the ranking of the 10 strategic schedules based on the dominance power of the schedules. When the dominance power of two or more schedules is the same, we further discriminate between these schedules by ranking them based on the percentage of predicted flight cancelled, where the schedule with the lowest percentage of flight cancellations is preferred. Table 12 shows the schedule ranking obtained using i) the predicted values KPIs, as a result of the machine learning algorithms, and ii) the values of the KPIs from the actual flight data. The results show that the 2 best-ranked schedules when considering the actual flight data are also captured when considering predicted KPIs.

Ranking Position	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
KPIs prediction models	S_3	S_1	S_4	S_6	S_8	S_7	S_2	S_{10}	S_5	S_9
KPIs real-data	S_3	S_1	S_2	S_4	S_{10}	S_7	S_5	S_6	S_8	S_9

Table 14: Schedule ranking with respect to percentage of flights cancelled and percentage of departure delays under schedule S_i.

Figure 10 shows the percentage of predicted percentage of delayed departure and arrival flights for each of the 10 strategic schedules. Layer 1 consists of schedules $\{S_3, S_9\}$. Layer 2 consists of schedule $\{S_7\}$, Layer 3 consists of schedule $\{S_1\}$. Layer 4 consists of schedule $\{S_5\}$. Layer 5 consists of schedules $\{S_8\}$. Layer 6 consists of schedule $\{S_4, S_6\}$. Layer 7 consists of schedules $\{S_2, S_{10}\}$.

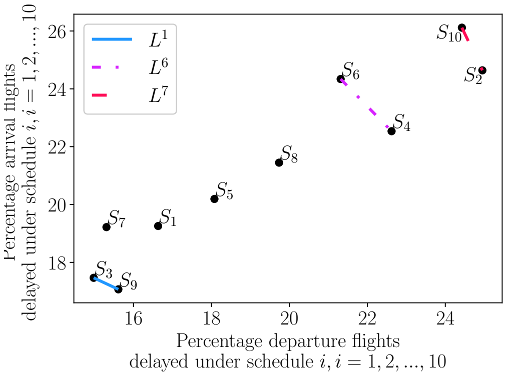

Figure 10: Layers of the strategic flight schedules when considering the percentage of predicted delayed departure and arrival flights.

Table 15 shows the dominance power of the schedules, where schedules S_3, S_7, S_9 have the best performance with respect to delayed departure and arrival flights.

	S_3	S_7	S_9	S_1	S_5	S_8	S_6	S_4	S_{10}	S_2
$D(S_i)$	1.90	1.40	1.40	1.07	0.82	0.62	0.29	0.29	0	0

Table 15: Dominance power of the 10 strategic schedules $S_1, ..., S_{10}$ when considering the percentage of delayed departure and arrival flight.

Table 16 shows the ranking of the 10 strategic schedules based on the dominance power of the schedules. When the dominance power of two or more schedules is the same, we further discriminate between these schedules by ranking them based on the percentage of predicted delayed departure flights, where the schedule with the lowest percentage of delayed departure flight is preferred. Table 16 shows the schedule ranking obtained using i) the predicted values KPIs, as a result of the machine learning algorithms, and ii) the values of the KPIs from the actual flight data. The results show that 3 from the 4 best-ranked schedules when considering the actual flight data are also captured when considering predicted KPIs.

Figure 11 shows the predicted performance of the 10 strategic schedules when considering 3 KPIs: the percentage of cancelled flights, the percentage of yellow departure days and the percentage of yellow arrival days. Layer 1 consists of schedules $\{S_1, S_3, S_4\}$.

Ranking position	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
KPIs prediction models	S_3	S_7	S_9	S_1	S_5	S_8	S_6	S_4	S_{10}	S_2
KPIs real data	S_5	S_7	S_1	S_3	S_{10}	S_9	S_2	S_8	S_4	S_6

Table 16: Schedule ranking with respect to percentage of delayed departure and arrival flights under schedule S_i.

Layer 2 consists of schedules $\{S_2, S_6, S_8\}$. Layer 3 consists of schedules $\{S_5, S_7, S_9, S_{10}\}$.

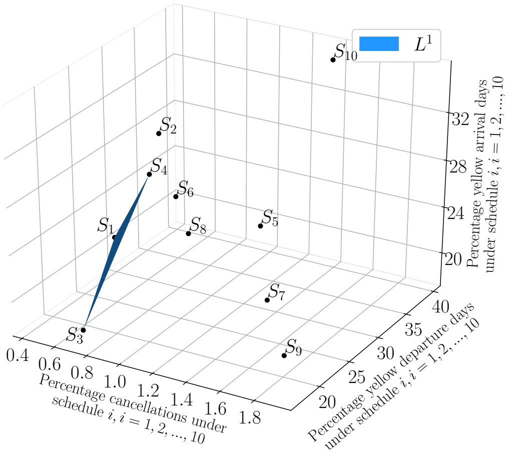

Figure 11: Pareto Front (L_1) obtained using 3 predicted KPIs: the percentage of cancelled flights, the percentage of departure yellow days and the percentage of arrival yellow days.

Table 17 shows the dominance power of the schedules, where schedules S_3, S_1, S_4 have the best performance with respect to delayed departure and arrival flights.

	S_3	S_1	S_4	S_2	S_6	S_8	S_{10}	S_5	S_7	S_9
$D(S_i)$	2.83	1.33	0.83	0.33	0.33	0.33	0	0	0	0

Table 17: Dominance power of the 10 strategic schedules $S_1, ..., S_{10}$ when considering the percentage of cancelled flights, the percentage of yellow departure days and the percentage of yellow arrival days.

Table 18 shows the ranking of the 10 strategic schedules based on the dominance power of the schedules. When the dominance power of two or more schedules is the same, we further discriminate between these schedules by ranking them based on the percentage of predicted cancelled flights, where the schedule with the lowest percentage of cancelled flights is preferred. Table 18 shows the schedule ranking obtained using i) the predicted values KPIs, as a result of the machine learning algorithms, and ii) the values of the KPIs from the actual flight data. The results show that the 4 best-ranked schedules when considering the actual flight data are also captured when considering predicted KPIs.

Ranking position	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
KPIs prediction models	S_3	S_1	S_4	S_2	S_6	S_8	S_{10}	S_5	S_7	S_9
KPIs real data	S_3	S_1	S_2	S_4	S_{10}	S_7	S_5	S_6	S_8	S_9

Table 18: Schedule ranking with respect to the percentage of cancelled flights, the percentage of yellow departure days and the percentage of yellow arrival days under schedule S_i.

6. Discussion

To address the lack of insight into potential flight delays and cancellation at the moment when flight arrival and departure slots are allocated at an airport, we propose to rank the strategic slot schedules based on KPIs derived from flight delay and cancellation predictions. We determine the values of these KPIs using machine learning-based predictions for flight delay and cancellation with a prediction horizon of 6 months prior to the flight execution day.

In practice, the implications of being able to assess strategic airport slot schedules with respect to potential flight delays and cancellations are multilateral. One implication is that airport slot coordinators are able to identify at an early stage potential airport on-time performance bottlenecks associated with the strategic schedules. Such bottlenecks can be in the form of congested days, i.e., days where the scheduled flights are expected to experience many delays and cancellations, as well as in the form of more detailed indicators such as the type of airline or the type of terminal associated with large delays and cancellations. Most important, in the case when airport performance bottleneck are expected, airport coordinators are provided with support to propose, in the limits of the IATA slot allocation guidelines, changes to schedule such as alternative arrival/departure slots or aircraft size. Thus, the results of this assessment provide an early, quantified motivation for potential schedule alternatives in the slot allocation negotiation between airport coordinators and airlines.

From a methodological point of view, when ranking the strategic slot schedules, in this paper we consider 5 flight delay and cancellation-based KPIs, which are often considered in practice. However, our proposed methodology supports the analysis of larger sets of schedule-related KPIs. In fact, the larger the size of the

KPI set, the more detailed the dominance relationship between schedules is defined. Also, for our schedule ranking methodology, we do not assume weights for the considered KPIs, since these weights are user-specific and. However, in the case when the weights are known, our ranking methodology can still be applied for weighted KPIs.

7. Conclusion

We have developed a machine learning approach to classify scheduled flights as being delayed and cancelled with a horizon of 6 months prior to the day of the flight execution. We have implemented our proposed model on a representative set of flights scheduled to arrive and depart to and from London Heathrow Airport in the period 2013-2018. Our proposed prediction models have achieved an accuracy of 0.79 or higher, with the LightGBM decision-trees achieving better prediction results than feed-forward neural networks and random forests. We have analyzed the impact of the model features on the outcome of the prediction algorithms. In particular, we have determined the Shapley additive explanation values for all the features used in the prediction models. We have shown that for the delay classification of both arriving and departing flight, the features with the highest feature importance are the Arrival ATFM delay, hour of the day, the airline type and the number of seats of the aircraft executing the flight. For flight cancellations predictions, the features with the highest feature importance are the origin/destination airport and the airline executing the flight.

Further, we have proposed a generic approach to rank strategic flight schedules based on pre-defined Key Performance Indicators (KPIs). We have used this approach to rank 10 strategic flight schedules considering as KPIs the predicted flight cancellations and delays associated with the strategic schedules. Together with flight schedule optimization models, this approach supports an integrated strategic flight schedule assessment, where strategic flight schedules are evaluated with respect to on-time airport performance.

As future work, we consider extending the set of features for the prediction algorithms to improve the accuracy of the predictions. In addition, we will evaluate the impact of considering flight delay and cancellation predictions in the flights scheduling optimization models, at the strategical phase.

References

M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak. Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, 13(6):355–361, 2007.

H. Alonso and A. Loureiro. Predicting flight departure delay at porto airport: A preliminary study. In *Computational Intelligence (IJCCI), 2015 7th International Joint Conference on*, volume 3, pages 93–98. IEEE, 2015.

L. Belcastro, F. Marozzo, D. Talia, and P. Trunfio. Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):5, 2016.

J. Bergstra, D. Yamins, and D. D. Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*, pages 13–20. Citeseer, 2013.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Castelli, P. Pellegrini, and R. Pesenti. Airport slot allocation in europe: economic efficiency and fairness. *International Journal of Revenue Management*, 6(1-2):28–44, 2012.

S. Choi, Y. J. Kim, S. Briceno, and D. Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, pages 1–6. IEEE, 2016.

S. Choi, Y. J. Kim, S. Briceno, and D. Mavris. Cost-sensitive prediction of airline delays using machine learning. In *Digital Avionics Systems Conference (DASC), 2017 IEEE/AIAA 36th*, pages 1–8. IEEE, 2017.

L. Corolli, G. Lulli, and L. Ntaimo. The time slot allocation problem under uncertain capacity. *Transportation Research Part C: Emerging Technologies*, 46:16–29, 2014.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

EUROCONTROL. Performance review report - an assessment of air traffic management in europe during the calendar year 2017, performance review commission, 2017.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

G. E. Hinton. Connectionist learning procedures. In *Machine Learning, Volume III*, pages 555–610. Elsevier, 1990.

Y. Horiguchi, Y. Baba, H. Kashima, M. Suzuki, H. Kayahara, and J. Maeno. Predicting fuel consumption and flight delays for low-cost airlines. In *AAAI*, pages 4686–4693, 2017.

M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

C. International Air Transport Association, Montreal. Worldwide slot duidelines 8th edition, 2017.

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. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

Y. J. Kim, S. Choi, S. Briceno, and D. Mavris. A deep learning approach to flight delay prediction. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, pages 1–6. IEEE, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

S. Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.

D. Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM*

SIGKDD Explorations Newsletter, 3(1):27–32, 2001.

E. Mueller and G. Chatterji. Analysis of aircraft arrival and departure delay characteristics. In *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, page 5866, 2002.

P. Pellegrini, T. Bolić, L. Castelli, and R. Pesenti. Sosta: An effective model for the simultaneous optimisation of airport slot allocation. *Transportation Research Part E: Logistics and Transportation Review*, 99:34–53, 2017.

N. A. Ribeiro, A. Jacquillat, A. P. Antunes, A. R. Odoni, and J. P. Pita. An optimization approach for airport slot allocation under iata guidelines. *Transportation Research Part B: Methodological*, 112: 132–156, 2018.

N. G. Rupp and G. M. Holmes. An investigation into the determinants of flight cancellations. *Economica*, 73(292):749–783, 2006.

B. Sridhar, Y. Wang, A. Klein, and R. Jehlen. Modeling flight delays and cancellations at the national, regional and airport levels in the united states. In *8th USA/Europe ATM R&D Seminar, Napa, California (USA)*, 2009.

A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara. A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*, 2017.

Y. Tu, M. O. Ball, and W. S. Jank. Estimating flight departure delay distributions a statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, 103(481):112–125, 2008.

G. Valkanas, A. N. Papadopoulos, and D. Gunopulos. Skyline ranking à la ir. In *EDBT/ICDT Workshops*, pages 182–187, 2014.

Q. Wu. A stochastic characterization based data mining implementation for airport arrival and departure delay data. In *Applied Mechanics and Materials*, volume 668, pages 1037–1040. Trans Tech Publ, 2014.

J. Xiong and M. Hansen. Value of flight cancellation and cancellation decision modeling: ground delay program postoperation study. *Transportation Research Record: Journal of the Transportation Research Board*, (2106):83–89, 2009.

K. G. Zografos, Y. Salouras, and M. A. Madas. Dealing with the efficient allocation of scarce resources at congested airports. *Transportation Research Part C: Emerging Technologies*, 21(1):244–256, 2012.

K. G. Zografos, M. A. Madas, and K. N. Androutsopoulos. Increasing airport capacity utilisation through optimum slot scheduling: review of current developments and identification of future needs. *Journal of Scheduling*, 20(1):3–24, 2017.