Factors influencing charter flight departure delay

Abstract

This study aims to identify the main factors leading to charter flight departure delay through data mining. The data sample analysed consists of 5,484 flights operated by a European airline between 2014 and 2017. The tuned dataset of 33 features was used for modelling departure delay (e.g., if the flight delayed more than 15 minutes). The results proved the value of the proposed approach by an area under the receiver operating characteristic curve of 0.831 and supported knowledge extraction through the data-based sensitivity analysis. The features related to previous flight delay information were considered as being the most influential toward current flight being delayed or not, which is consistent with the propagating effect of flight delays. However, it is not the reason for the previous delay nor the delay duration that accounted for the most relevance. Instead, a computed feature indicating if there were two or more registered reasons accounted for 33% of relevance. The contributions include also using a broader data mining approach supported by an extensive data understanding and preparation stage using both proprietary and open access data sources to build a comprehensive dataset.

Keywords

Charter industry; flight delay; delay prediction; data mining; feature relevance.

1. Introduction

Change is swift at any given moment in the business world, especially in organisations needing to adapt to the continuous modifications of the market [1]. Nowadays, all

companies from all industries can create new products and services rooted in data analytics [2]. Thus, analysing historical business data may prove to be an enhancement opportunity to improve any company [3] and gain a competitive advantage over other competitors - a blue ocean-oriented strategy [4]. An airline business model is no exception; it includes several critical tasks to operate air flights successfully. These tasks include services associated with the aircraft turnover [5], either provided by the ground handling staff, such as aircraft movements [6] and loading [7], or by terminal passengers and luggage processing [8], while others by the technical and cabin crew members. An error in any of those tasks triggers a chain of events which may lead to departure delay, and subsequently, to unexpected financial expenses, and conducting to brand negative impact [9].

Furthermore, external factors may also affect flight operation. Some of those include the weather [10], air traffic [11], or even outliers such as personnel labour strike (both crew or airport staff) [12]. The impact of a flight delay extends well beyond the specific flight as the resources involved in operating flights are optimised toward operational efficiency [13]. Several types of resources and activities such as staff and taxiing aircraft need to be withheld for a more extended period to cope with the delay, resources which otherwise would be allocated to other flights or tasks [14]. Thus, the delay effect is propagated and may ultimately affect an entire network of flights, whether by delaying other tasks within the departure airport or by affecting arrival airport tasks if the aircraft arrives late at its destination [15]. Therefore, airlines can be profoundly affected by flight delays and incur in economic losses [16]. To address such issue, airlines facing successive delays tend to charge higher airfares to passengers [17]. The chain effect of the tasks involved in-flight operation, and the delay propagation and impact to both airline and airport operational management have motivated researchers

and practitioners toward predicting flight delay. The approaches typically consist of data-driven solutions based on statistical or machine learning techniques to model the departure delay [18]. However, most studies are focused on problems translated by a set of specific given features that directly characterise those problems, such as the departure-arrival connection [19] or the weather conditions [20]. Thus, the research effort is typically more dedicated to tuning the model in pursuit of the most accurate predictions, overlooking the critical step of data understanding and, particularly, feature enrichment in a data mining project [21]. Literature addresses several studies on this matter, although understanding the influencing factors of flight delays is an on-going research subject, given the myriad of factors and different perspectives that can be addressed [22, 17, 23]. Therefore, further studies are needed to address the understanding of flight delays.

This study's objective is to identify the main factors leading to charter flight departure delay (e.g., if the flight departure delayed more than 15 minutes or not) in a specific context. To address these research objectives, we adopted a data mining approach. This study takes a broader perspective of mining charter flights data to predict departure delay by emphasising the relevance of identifying the most meaningful features to model the delay using both proprietary and publicly available data to build an enriched dataset characterised by problem-relevant features [24]. The charter flight data is provided by a European-based airline operating both scheduled ahead and ad-hoc flights between 2014 and 2017.

This study brings three contributions. First, this study enables to highlight the features of influencing flight delay in the crucial European market. Most of the empirical research in flight delay prediction is conducted within the US context [19], given the US has a less restrictive data policy when compared to Europe and promotes data sharing

for research purposes. Second, this study was conducted in charter flights context. Third, this study also provides insights on the most relevant features to a flight delay by opening the trained models using the data-based sensitivity analysis [25]. This approach has already been successfully applied in other contexts such as tourism [26] and civil engineering [27]. Also, this study uses a broader number of distinct sources with *a-priori* known features when compared to existing literature [19, 20].

This paper is structured in six sections, beginning with the context and research problem definition (section 1), followed by the theoretical background (section 2). In section 3, the methodological approach and data sampling are presented. Section 4 presents flight prediction results and, in section 5, the features relevance results are discussed. Finally, in the last section, we draw the conclusions, and the theoretical and practical implications of this study.

2. Literature review

Flight operations translate into complex problems to solve, highlighted by a myriad of possible scenarios with many internal (e.g., traffic management) and external variables (e.g., weather). The studied problems include planning of flight routes [28], detecting flight trajectory anomalies [29], air traffic management [30], recognizing aircraft events such as landing [31], among others. Given the chained nature of air transportation business processes, most of those problems are interconnected, making it challenging to narrow the research focus in a specific problem. The challenges inherent of such complexity have led researchers to devote efforts to adopting state-of-the-art techniques such as modern optimisation and data mining to address them.

The departure delay prediction of a flight is a widely studied problem in air transportation literature [19]. The operational impact of flight delays is enormous,

resulting in economic losses estimated by Zou and Hansen [32] in $7.1-13.5 billion in 2007 for US airlines when comparing the ideal operational performance to the real one. Ball et al. [33] corroborate that range by estimating a value of around $8.3 billion to airline losses. Flight delays affect passengers, causing distress, and consequently undermining the airline brand image. Although airlines attempt to compensate for the delays, travellers are unable to perceive the different services offered by the competition, which does not favour brand loyalty, with travellers often choosing to shift to another airline [34]. Delays have a more emphasised impact on high-income managers who frequently need to air travel, with airlines being especially concerned to satisfy this type of travellers [35]. Besides passengers, delays can also cause labour disruption by affecting personnel work period, especially in the air transportation business, where there is strict legislation regarding flying times and off-times [36]. Also, as it was previously stressed, the propagation effect of delays to the daily airline and airport operation can cause other flights' delay which is difficult to overcome since resources are optimised to reduce slacks [13; 37; 38]. Another critical issue related to delay propagation is the air transportation network [39]. The complexity of such networks implies that a wide range of external factors (i.e., not related to the airline) can affect directly or indirectly flights, leading to delays [40]. As Sun et al. [41] pointed out, country networks are different between countries, with some of them being more influenced by the passenger traffic of their neighbourhood countries.

Table 1 summarises four articles focusing on predicting flight departure delays. The first noteworthy common characteristic is the fact that all four studies used the US publicly available data. This data was supplied by the U.S. Department of Transportation [18], the Federal Aviation Administration - FAA [42, 19], or the Bureau of Transportation Statistics - BTS [20]. Table 1 stresses out (1) the open data policy of

the US authorities [43], and (2) the need to study other countries' airspace and air transportation operations. Some studies aim to predict the delay time value, thus making of it a regression problem, although most of them are focused in predicting if the flight is delayed or not, therefore building a classifier [44]. The latter poses the challenge of defining the threshold above which the flight is considered delayed. Rebollo and Balakrishnan [19] considered a flight delayed if departures are 60 minutes after schedule, while Choi et al. [20] considered 15 minutes. However, the former authors stated that the US Department of Transportation "only counts a flight as delayed if it incurs a delay of more than 15 min" [19, pp. 240]. Moreover, literature acknowledges the 15 minutes delay published by the US Department of Transportation as a standard threshold definition [45, 32]. Furthermore, in Europe, the Eurocontrol [46] states that an "aircraft should take-off within 15 minutes of the time stated in its flight plan". Managers at the European-based airline used for the empirical research also provided support for the 15 minutes definition which is currently used in the aviation business.

The results achieved by the studies mentioned in Table 1 are not directly comparable since the cases are different (i.e., different timeframe, airports, and features considered). Also, the goals do not precisely match, with Balakrishna et al. [42] focusing specifically on taxi-out (a likely reason for better accuracy than the remaining), while the others aimed to predict departure delay in general, although using different metrics for evaluating the results. The two most recent studies [19, 20] achieved a similar accuracy of 81%, although considering different features. In fact, apart from Tu et al. [18], the remaining are too narrow in the features used, focusing on understanding a specific phenomenon (e.g., weather influence; or the connection - origin-destiny airport - influence). Also, most studies are using the features directly obtained from a single or two public datasets, without attempting to unveil other interesting features that may

influence flight delays. Thus, the focus is on improving the model's accuracy by testing different models [20]. Finally, while predicting departure delay is a challenging task, it would be interesting to get insights from the most accurate models on the features' contribution to the delay. Such knowledge may help air transportation managers to understand this complex problem. However, from the studies analysed, only Rebollo and Balakrishnan [19] provided and discussed their findings regarding understanding features' contribution.

Table 1 - Departure delay prediction studies.

Reference	Goal	Case	Method	Features	Results	Most relevant features
[18]	Estimating flight departure delay distributions	US Denver International Airport, flights in 2000/2001 (92,865 records)	Genetic algorithm based on expectation-maximisation (smoothing spline model to estimate the relationship between delay, seasonal trend, and daily propagation pattern). Tested with 2001 data.	Three components: seasonal trend (including weather), daily propagation pattern, and random residual	Predicted upper 3.00% tail holds 2.65% of the delays (see Tu et al. [18] for additional metrics)	NA
[42]	Taxi-out delay prediction (mean taxi-out time for the same quarter of hour)	US Tampa Bay International Airport, June-August 2007	Nonparametric reinforcement learning (testing in 26th to 31st August flights)	Gate OUT, wheels OFF, wheels ON, and gate IN	ACC=93.7%	NA

[19]	(1) Predict if departure delay is above 60 min (2) Predict the delay time	US flights between 2007 and 2008 - 10 sets (4,000 records, 3,000 for training and 1,000 for testing)	Random forest; Kruskal–Wallis parametric ANOVA test to evaluate feature relevance	origin - destination (OD) pair	(1) ACC=81%; TPR=76.4% (2) 21 min median error	DCA (Washington) and JFK (NYC) departures
[20]	Predict if departure delay is above 15 min	US domestic airline traffic data and weather data from 2005 to 2015 (8,833 records)	Random forest; 10-folds cross-validation	5 schedule features (e.g., month) + 12 weather features (e.g., wind speed)	ACC=81.37%	NA

3. Data sampling and methods

The empirical research presented in this study is supported on a dataset of an European-based airline operating flights between airports around the world. The goal is to predict if a flight departure was delayed by more than 15 minutes and simultaneously understand why delays are happening. Thus, this research is aligned with the airline's requirement for efficiency improvement derived from reducing and mitigating departure delays. As such, this study takes a single airline's perspective instead of the broader flight analyses conducted by the studies cited in Table 1. In the light of previously supported 15 minutes delay definition, the focus is to obtain an accurate classifier of departure delay (i.e., classify if a flight departure will be delayed 15 or more minutes).

The initial dataset compiled consists of 5,484 flights occurred from the 1st of January 2014 to the 23rd of March 2017. By using an airline's anonymised data, it is possible to access specific features such as the type of flight (scheduled ahead or ad-hoc) or the

type of aircraft which would otherwise be concealed. As Wong and Tsai [47] study suggested, weather plays a significant role in delayed flights. Also, more extensive and busier airports can have a higher impact on propagating delays [48]. Thus, the airline dataset was merged with other publicly available data sources (Table 2), which were chosen by assuring that an identity field existed that enabled to merge sources. Some of the features retrieved from those sources hold high numbers of missing values (e.g., "fl.arr.DST" holds 1,807 of unknown values, Table 2, "OpenFlights" source for further details). Those features were discarded. Nevertheless, the new data gathered through the features that remained added important information that is known to influence flight departure (e.g., weather), justifying, including those data sources.

Table 2 - List of data sources used.

Source	**Description**
Airline	Data provided by the European-based airline
OpenFlights	2017 OpenFlights Airports Database (https://openflights.org/data.html)
OurAirports	OurAirports dataset (http://ourairports.com/data/)
Weather	IOWA State University METAR data (https://mesonet.agron.iastate.edu/request/download.phtml)
CountryRank	Country ranking by flight volume (http://databank.worldbank.org)
Top10Prize	World Airport Awards Top 10 (http://www.worldairportawards.com/Awards/worlds_best_airport.html)
Top10Surface	Top 10 airports by surface (https://www.worldatlas.com/articles/the-world-s-10-largest-airports-by-size.html)

Table 3 shows the list of 68 initially considered features for analysis. The column "source" highlights the diversity of data sources used and described in Table 2. A "source" attribute with the value "computed" indicates that the corresponding feature was computed using other features. Two examples are: the month, computed based on

the date, and the continent, computed based on the country (both for departure and for arrival). Features are labelled using prefixes according to the types: “ac” for aircraft; “dt” for date; “fl” for flight (with “dep” for departure and “arr” for arrival data); and “wt” for the weather.

The rationale for choosing the features, such as season, aircraft size, charter/regular flight, and longitude, is grounded in the literature [49, 50, 35]. However, there is a lack of a holistic model encompassing all those features to predict flight delays. Also, some factors such as the analysis of how previous flight delay information affects current delay have not been analysed to the extension here presented, i.e., by including an array of relevant features (those with prefix “fl.prev.fl” described in Table 3). Following a careful analysis of all features, some of them were discarded for being categories concealing too many possible values (e.g., “fl.dep.country” holds 108 different countries; “wt.weatherID” holds 5,431 different combinations of date, city code, and hour).

Table 3 - List of features analysed.

Feature	Description	Included?	Source
ac.age	Aircraft age (in years)	Y	Airline
ac.code	Aircraft internal code	N	Airline
ac.max.fuel	Aircraft maximum fuel capacity	N	Airline
ac.max.pax	Aircraft maximum PAX capacity	Y	Airline
ac.model	Aircraft model (e.g., Airbus A340)	N	Airline
ac.msn	Aircraft manufacturer serial number	N	Airline
ac.nr.engines	Aircraft number of engines (2 or 4)	Y	Airline
ac.registered	Aircraft registered country (Portugal or Malta)	Y	Airline
dt.date	Flight date	N	Airline

Feature	Description	Included?	Source
dt.day.of.month	Day of month (1 to 31)	Y	Computed
dt.month	Month (January to December)	N	Computed
dt.season	Season (Spring, Summer, Autumn, Winter)	Y	Computed
dt.weekday	Weekday (Sunday to Saturday)	N	Computed
dt.weekend	If the flight was on the weekend	Y	Computed
dt.year	Year of flight (2014 to 2017)	Y	Airline
fl.arr.altitude	Arrival airport altitude	Y	OpenFlights
fl.arr.continent	Arrival airport continent	N	Computed
fl.arr.country	Arrival airport country	N	OpenFlights
fl.arr.DST	Arrival airport Daylight Saving Time (Table 4)	N	OpenFlights
fl.arr.hour	Flight expected arrival hour	Y	Airline
fl.arr.IATA	Arrival airport IATA code	N	Airline
fl.arr.latitude	Arrival airport latitude	Y	OpenFlights
fl.arr.longitude	Arrival airport longitude	Y	OpenFlights
fl.arr.night.office	If arrival occurred in night office time	Y	Airline
fl.arr.UTC	Arrival airport Coordinated Universal Time	N	OpenFlights
fl.dep.airport.type	Departure airport size (medium; large)	Y	OurAirports
fl.dep.altitude	Departure airport altitude	Y	OpenFlights
fl.dep.continent	Departure airport continent	N	Computed
fl.dep.country	Departure airport country	N	OpenFlights
fl.dep.country.flights.rank	Ranking position of the country in nr. flights	Y	CountryRank
fl.dep.DST	Departure airport Daylight Saving Time (Table 4)	N	OpenFlights
fl.dep.hour	Flight expected departure hour	Y	Airline
fl.dep.IATA	Departure airport IATA code	N	Airline
fl.dep.ICAO	Departure airport ICAO code	N	OpenFlights
fl.dep.is.capital	If the departure airport is in a country's capital	Y	Computed
fl.dep.latitude	Departure airport latitude	Y	OpenFlights
fl.dep.longitude	Departure airport longitude	Y	OpenFlights
fl.dep.night.office	If departure occurred in night office time	N	Airline
fl.dep.top10.airp.prize	If it is one of the top 10 airports for PrizeAwards	Y	Top10Prize

Feature	Description	Included?	Source
fl.dep.top10.airp.surface	If it is one of the top 10 airports by surface	Y	Top10Surface
fl.dep.UTC	Departure airport Coordinated Universal Time	N	OpenFlights
fl.diff.altitude	Altitude difference (departure and arrival)	N	Computed
fl.diff.latitude	Latitude difference (departure and arrival)	N	Computed
fl.diff.longitude	Longitude difference (departure and arrival)	N	Computed
fl.diff.UTC	UTC difference (departure and arrival)	N	Computed
fl.duration	Scheduled flight duration	Y	Airline
fl.during.night.office	If the flight occurred during night office	N	Airline
fl.hired	If it is a flight hired for another airline	Y	Airline
fl.intercontinental	If it is an intercontinental flight	Y	Computed
fl.is.adhoc	If it is a requested ad-hoc (as opposed to a scheduled ahead) flight	Y	Airline
fl.more.10.hours	If the flight is expected to last more than 10 hours	N	Airline
fl.domestic	If it is a domestic flight (i.e., departure and arrival city are located in the same country)	Y	Computed
fl.prev.fl.delay.1st.reason	Standard IATA Delay Codes (two main delay reasons for previous flight)	N	Airline
fl.prev.fl.delay.1st.reason.gr		N	Airline
fl.prev.fl.delay.2nd.reason		N	Airline
fl.prev.fl.delay.2nd.reason.gr		N	Airline
fl.prev.fl.delayed.duration	Previous flight delay duration (negative if the flight arrived earlier)	N	Airline
fl.prev.fl.delayed.reason	Based on 1st reason: Airline (0-9); Pre-flight (31-39,61-69); Processing (11:19,21:29); Aircraft (41-48,51:58); External (71-99)	Y	Computed
fl.prev.fl.delayed.more1r	If previous flight delayed for more than 1 reason	Y	Computed
fl.was.prev.fl.delayed	If the previous flight was delayed	N	Airline
wt.dwpc	Dewpoint temperature at departure city in Celsius	Y	Weather
wt.relh	Relative humidity in % at the departure city	Y	Weather
wt.sknt	Wind speed in knots at the departure city	Y	Weather
wt.station	Station site identifier at the departure city	N	Weather
wt.tmpc	Air temperature at departure city in Celsius	N	Weather

Feature	Description	Included?	Source
wt.valid	Timestamp of the observation at the departure city	N	Weather
wt.vsby	Visibility at departure city in miles	Y	Weather
wt.weatherID	Present weather code at the departure city, composed by date, city code, and hour	N	Weather
fl.delayed.above15	Target: if the flight departure was delayed	Y	Airline

Also, some features were discarded as these were used to compute other more aggregated features (e.g., "fl.prev.fl.delay.1st.reason" and "fl.prev.fl.delay.2nd.reason" were both used to compute "fl.prev.fl.delayed.more1r"). Finally, a correlation matrix was computed using the Pearson correlation coefficient, enabling to assess which features were overlapping in terms of the information held [51]. Of each pair of moderately to highly correlated features (i.e., $\rho \geq 0.5$), the feature holding less information was removed (e.g., "fl.dep.night.office" was removed, leaving "fl.dep.hour", because the former has only two possible values, {"Y", "N"}, while the latter has 24 possible values). Thus, modelling took as input a tuned dataset with 33 input features and the output. Table 6 in the Appendix displays the statistics for each of these selected features.

The classifier for departure delay modelled the "fl.delayed.above15" feature. Three modelling techniques were chosen: the neural network (NN), support vector machine (SVM), and random forest (RF). The first two were chosen as these have offered sustainably better performances when compared to other techniques (e.g., logistic regression) in previous studies [52, 53], while the latter was chosen as it achieved the best performance in the two most recent studies mentioned in Table 1 [19, 20]. For the NN model, the multilayer perceptron, the most popular NN architecture [54], was adopted. It consists of one hidden layer constituted by several hidden nodes (or neurons)

and one output node. The activation function of each node is computed by weighting previous nodes' outputs [55]. The SVM uses a nonlinear mapping dependent on a kernel (the popular Gaussian kernel, which presents fewer parameters when compared to others, was adopted [56]) to transform the complex R^M feature space in a high *m*-dimensional feature space. This new space enables to find the best separating hyperplanes by selecting support vectors [57]. Finally, the RF is an ensemble of decision trees which make individual contributions to the overall model, improving the overall performance by benefiting from several decision trees' heterogeneity [58]. Pal [59] found that RFs perform equally well when compared to SVMs and are more straightforward to define and require fewer parameters. The three chosen techniques can apprehend non-linear relations between several features, making them suitable for the proposed classification problem. Furthermore, the three have shown their superiority in training classifiers in a wide variety of problems [e.g., 53; 60].

Model tuning is an important issue in machine learning algorithms. The best hyperparameters were set using grid searchers for the three cases. The SVM adopted the popular Gaussian kernel (i.e., $K(x,x')=\exp(-\gamma||x-x'||2)$, and the search was performed using the following parameters: $\gamma \in 2^k$: $k \in \{-15,-11.4,-7.8,-4.2,-0.6,3\}$, as recommended by Moro et al. [53]. The second SVM parameter was fixed using the heuristic C=3 proposed by Cortez [61] for standardized input data. In the case of the NN, the search for the optimal number of hidden nodes was set by choosing from the set of values: {0,2,6,8,10,12} [53]. For the RF, the number of estimators (trees) was set by choosing from the set of values:{10,20,30,40,50,60} [62].

Classifiers can be evaluated through several metrics, which essentially compare the predicted to the real category. Binary problems, such as the one addressed can be easily evaluated using a confusion matrix [63], composed of four cells: True Positives (i.e., the

flight is correctly predicted as delayed); False Positives (i.e., the flight is incorrectly predicted as delayed); True Negatives (i.e., the flight is correctly predicted as on-time); and False Negatives (i.e., the flight is incorrectly predicted as on-time). Model's accuracy (ACC) can be computed as ACC=(TP+TN)/N, with N being the full number of records. Another interesting metric is the True Positive Rate, or sensitivity (TPR=TP/(TP+FN)). TPR emphasises the relevance of accurately predicting delayed flights, neglecting flights predicted as delayed but had an on-time departure. TPR is considered by Rebollo and Balakrishnan [19] a relevant metric for flight delay prediction. As opposed to TPR, the False Positive Rate (FPR) is computed by the following formula: FPR=FP/(FP+TN). The FPR represents the predictive performance of the classifier in predicting the "negative" class. Thus, the higher the FP (i.e., the cases when the model wrongly predicted as a "positive" class), and subsequently, the higher the FPR, the worse is the model. The abovementioned metrics assume a classifier's output consists of two or more categories. However, it is more interesting to build a classifier that computes a probability of a flight being delayed, instead of just "yes"/"no". Thus, by varying the possible thresholds for the considered probability above which the flight is considered delayed, one would obtain distinct confusion matrices. By plotting the TPR versus the FPR in a single graphic, it is possible to assess a classifier's performance through the range of thresholds for considering the flight as delayed. Such graphic is named the receiver operation characteristic (ROC) curve. The higher the TPR, the better are the classifier's predictions [64]. Therefore, the area under the ROC curve (AUC) represents a more generic metric for evaluating a classifier's global performance [65]. Given the premises above (i.e., a higher TPR representing a better classification), an AUC of 0.5 represents a random classifier (the baseline), whereas an AUC of 1.0 is the perfect classifier. Thus, the higher the AUC, the better are

the predictions. Since the dataset built is static (i.e., it is not fed with new data), the *k*-fold cross-validation was adopted. It splits the dataset into *k* partitions with the same number of records and trains the model using the first *k-1* partitions and tests it by using the last *k* partition. Then, the procedure rotates the train/test partitions used until all *k* partitions have been used at least once for testing and *k-1* times for training [65]. As recommended by Refaeilzadeh et al. [66], *k* was set to 10.

After the initial feature analysis previously described, an automated feature selection procedure was executed [24]. Such procedure consists of training a model with all the previously selected features and assess its performance using the *k*-fold cross-validation scheme previously described and the AUC metric. Then, in a loop scheme, each individual feature was removed, and a new model was trained without that feature and evaluated using the same method (i.e., *k*-fold cross-validation and AUC). If the obtained AUC was the same or above the one without removing the feature, then the feature was removed, and the loop proceeded to the next iteration. Since the learning process of modelling techniques needs to be contained under a reasonable processing time, having a more significant number of features requires the algorithm to perform more tests to assess each feature's usefulness to the model [67]. Thus, having more features can result in slightly more unsatisfactory performances if there are features non-related to the model's target. The loop ended when removing any of the remaining features resulted in a decrease in performance. However, any removal of the remaining features resulted in more reduced performance, thus supporting the chosen features (those marked with "Y" in the "included" column in Table 3).

Finally, a sensitivity analysis (SA) was adopted to unveil the most interesting features. The SA aims to find a model's feature relevance by assessing the model's results to input feature variation. The one-dimension sensitivity analysis procedure varies one

input at a time while keeping the remain constant. Therefore, it is not able to measure the influence of each feature on the remaining. On the opposite, the global sensitivity analysis consists in varying the set features simultaneously through their range of possible values [62]. This is a computationally very demanding alternative, which is only suited for small numbers of features, given the procedure needs to go through the combination of the values possible within each feature. Another approach that was proposed and is described by Cortez and Embrechts [25] is the data-based sensitivity analysis (DSA). The DSA is a procedure that addresses the limitation of the one-dimension SA but without the computational effort required for a full variation of the whole set of features. The same authors also present two alternatives: the cluster-based sensitivity analysis (CSA), and the Monte-Carlo sensitivity analysis (MSA). However, as they recognize, CSA failed to detect the most relevant input in one of their tests, while MSA is more suited for cases where the data used for training the model is not available. Thus, we chose DSA, which uses a randomly selected sample of instances from the dataset and then varies the possible combinations of features only through the values stored in the selected instances to assess output variation (i.e., sensitivity). Cortez and Embrechts [25] tested DSA with four real-world datasets and concluded that their approach provided robust results when compared to the alternatives. All experiments were implemented using the open-source R statistical tool, which offers a myriad of packages suited for data analysis [68]. Notably, the "rminer" package was chosen as it implements simple-to-use data mining functions and the DSA [69]. It should be noted that the functions implemented by this package enable users to focus more on the data mining problem and less on the technical details. For example, the model training process can be fed using both numeric and categorical features.

4. Flight prediction results and discussion

Table 4 shows the achieved metrics using the three abovementioned techniques. The trained classifier's output is one of two classes: the flight departure was delayed (our target, considering the flight was delayed if its departure was more than 15 minutes of the scheduled hour), or the flight departure was on time (less than or equal to 15 minutes of the scheduled hour). As stated by Hand and Till [64], the AUC measures a classifier's performance through the range of values for the threshold probability above which a flight is deemed delayed (the target). Considering an AUC of 0.5 represents a random uninformed model, and a perfect model holds an AUC of 1.0, it is possible to observe the RF achieved the highest performance when compared to both NN and SVM. This result is consistent with both Rebollo and Balakrishnan [19], and Choi et al. [20]. Figure 1 helps to support the differences in AUC by displaying the ROC curves for the three models under the same graphic - the RF curve stands above the remaining for all cases.

Table 4 - Modelling performance evaluation.

	Evaluation metric		
Modelling technique	**AUC**	**ACC**	**TPR**
Neural Network (NN)	0.789	72.85%	79.56%
Support Vector Machine (SVM)	0.789	72.62%	82.19%
Random Forest (RF)	0.831	76.12%	83.27%

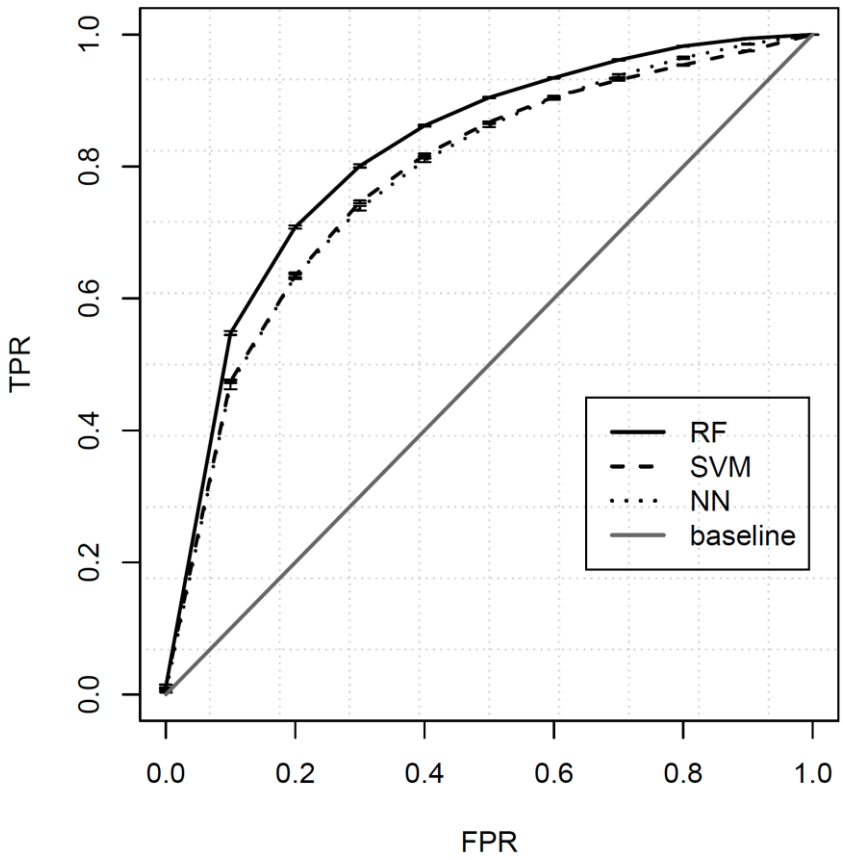

Figure 1 - ROC curves (FPR=FP/(FP+TN)).

The three computed metrics (AUC, ACC, TPR) confirm the classifier models departure delays with reduced errors.

5. Feature relevance extraction results and discussion

Considering the RF model clearly represent the best classifier for this problem, it was chosen for knowledge extraction. The knowledge of flight departure delay was extracted using the DSA, in the form of features' relevance to the model. Thus, the DSA computes each feature's contribution to the RF model. The result is a list with percentages distributed per feature, shown in Table 5. To avoid a lengthier table, only features with relevance above 1.0% are individually displayed, in a total of 21 from the 33 used for modelling. The remaining 12 features' aggregated relevance is shown in the last row, labelled "all the remaining," summing a total of 7.33% of relevance. However,

it should be stressed that all features have some contribution, since removing any of them resulted in lower performance.

Several interesting findings emerge from Table 5. First, the combined relevance of all features related to the previous flight the aircraft has made (fl.prev.fl.delayed.more1r, fl.prev.fl.delayed.duration, and fl.prev.fl.delayed.reason) sums around 48%. This result is not surprising since literature has exhaustively reflected on the propagating effect of flight delays [15]. However, the reason for the previous delay accounted for only 2.03% of relevance. Nevertheless, the most remarkable result is the fact that the feature indicating if there was more than one reason registered to justify the previous delay contributes to explaining more than 32% of the current flight being or not delayed. This feature is computed based on "fl.prev.fl.delay.1st.reason" and "fl.prev.fl.delay.2nd.reason". Thus, the flight team can insert up to two delay reasons. If they did insert two reasons, then "fl.prev.fl.delayed.more1r" was computed as "Y"; otherwise, "N." While managers in the business aviation industry may intuitively guess this was an important feature, current state-of-the-art literature has not yet raised this issue, nor it has quantified its relevance. Our approach is a valuable insight that scholars exploring the propagating effect of delays may use to weight the nodes of a directed graph representing a network of related flights.

Another interesting finding raised from Table 5 is the relevance of the type of flight, i.e., being an ad-hoc requested or a planned flight. Literature shows little evidence of the flight type influence in departure delay. Aviation managers treat both types differently, as ad-hoc requests often occur under stressing conditions where there is a shortage in supply that needs to be quickly suppressed. Hence, this study is the first presenting a measure of such impact, corroborating with airlines and airport managers' behaviour. The results also emphasise the relevance of the size of an airport and both departure and

arrival longitudes. Next, the five most relevant features (i.e., with relevance above 4%) are scrutinised. Using the results obtained through the DSA, it is possible to draw variable effect characteristic (VEC) curves that visually show how each feature influences the output (i.e., the flight departure delay) [25]. Therefore, the VEC plots the probability of a flight being delayed in the y-axis and the range of possible values for the analysed input feature in the x-axis. Explicitly, the probability in the y-axis is computed directly by the DSA through assessing how the outcome changes by varying each input feature.

Table 5 - Individual features' relevance.

Feature	Relevance
fl.prev.fl.delayed.more1r	32.57%
fl.prev.fl.delayed.duration	13.22%
fl.is.adhoc	5.62%
fl.dep.top10.airp.surface	4.41%
fl.dep.longitude	4.28%
fl.arr.longitude	3.57%
ac.max.pax	3.22%
fl.dep.latitude	2.90%
fl.dep.hour	2.48%
ac.nr.engines	2.46%
fl.dep.is.capital	2.05%
fl.prev.fl.delayed.reason	2.03%
ac.registered	1.96%
wt.relh	1.96%
fl.duration	1.63%

Feature	Relevance
dt.day.of.month	1.57%
fl.arr.hour	1.50%
fl.domestic	1.45%
dt.year	1.44%
ac.age	1.27%
wt.vsby	1.10%
all the remaining	*7.33%*

Figure 2 shows that a previous flight of an aircraft delayed on departure for two registered reasons has more than 30% of probability of being delayed in current flight. It should be noted that although the number of flights previously delayed for two reasons is an unbalanced feature, 524 from the total of 5,484 (almost 10%) were in this condition. Such numbers emphasise the representativeness of this feature and are directly translated into more than 30% of relevance to the model. These results show that, no matter what the reasons for the previous delay are, if there are at least two of them, i.e., if the crew bothered to register two or more reasons, it is likely that the next flight will also be delayed. However, the results should be carefully interpreted, as it may not be a typical standard for airlines to have 10% of flights delayed for more than one reason per flight. Further studies in other airlines could help to shed additional light on this subject. Nevertheless, this discovery is relevant to this specific airline, which may lead to preventive actions to avoid multiple delays.

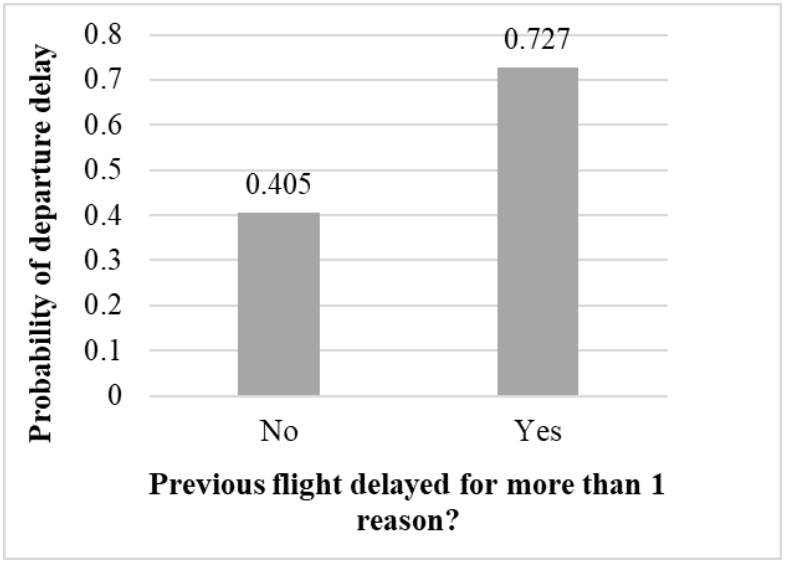

Figure 2 - Influence of previous flight being delayed for more than one reason.

The effect of the delay duration of previous flights can be observed in Figure 3. The previous flight delay seems to mainly affect current departure delay, increasing between 100 minutes earlier and 100 minutes late. After reaching the plateau (above 100 minutes delayed), the effect slightly decreases and remains steady afterwards. The observed effect derives from the fact that the airline only operates charter flights. When flights leave outside an appropriate interval of 100 minutes, the probability observed in Figure 3 can be explained by commercial factors. If a flight leaves above 1 hour and forty minutes (100 minutes) behind schedule, there is an increasing probability that the customer hiring this airline's flight may need to cancel or postpone future flights, thus updating future scheduled hours, which will become on-time. Therefore, the [-100, +100] minutes period is when there is a focus on solving problems, whereas outside that interval it is likely that the next flight for the same aircraft needs to be rescheduled.

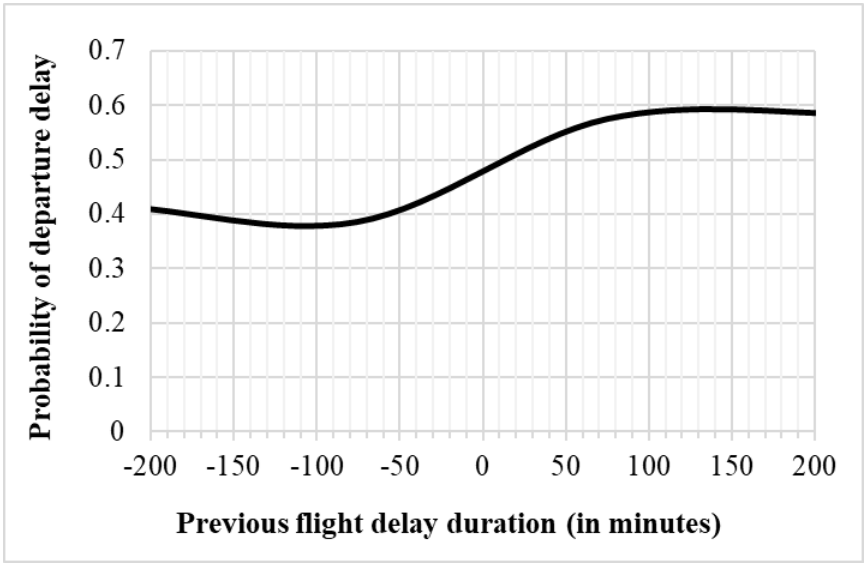

Figure 3 - Influence of previous flight delay duration.

Figure 4 exhibits the different probability of a flight departure delay dependent on being a requested ad-hoc or a scheduled ahead flight. Although the difference is just of around 0.05, ad-hoc flights do not leave late as many as planned flights. In the context of this specific European airline, there is frequently a higher pressure in ad-hoc flights to leave on-time, in the result of a supply shortage. Also, ad-hoc flights are usually requested in known routes to the airline (both to managers and cabin crew members), by contractors that know the airline and have previously request its services to the same routes. Hence, those contractors usually have a long-term relationship with the airline and take advantage of that for requesting non-planned flights. Such flights create the abovementioned additional pressure due to lack of prior planning. On the opposite, planned flights are contracted in advance, with the airline flying to lesser-known destinations, for more extended periods, where there may occur planning constraints more frequently, such as maintenance tasks. Also, charter companies have usually a low market share, which leads to being lesser-known and often more neglected in unknown destinations when compared to larger airlines.

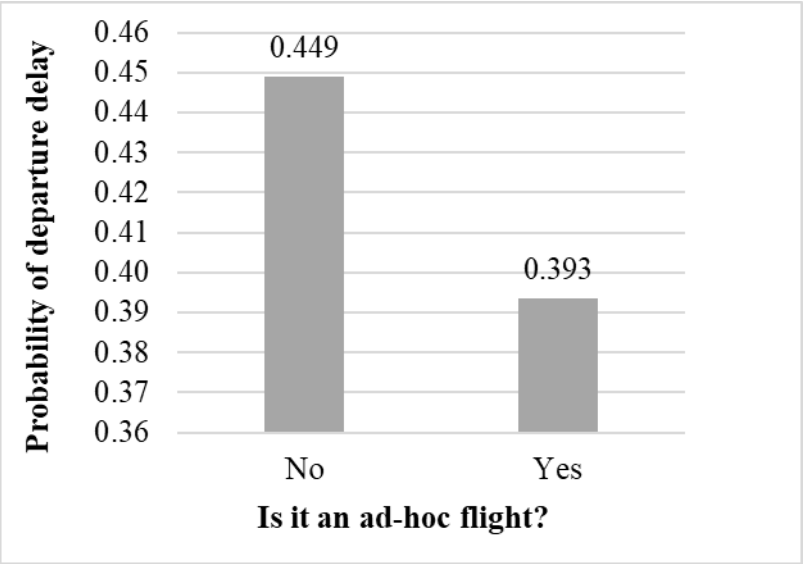

Figure 4 - Influence of being an ad-hoc flight.

It is interesting to note that, from the many airport-related features included in the model, the size of the departure airport outstands from the remaining, being the fourth most relevant when modelling departure delay. This result is supported by Baluch et al. [70] study on US flights. The same authors also identified a strong influence of the airline on delays by analysing several of them. In our study, this feature is constant by narrowing the analysis to a single airline. The fact that a feature extracted from an open data source of a known ranking was deemed influential to delays corroborates the importance of enriching corporate data with open data sources, emphasizing the contribution of the proposed approach.

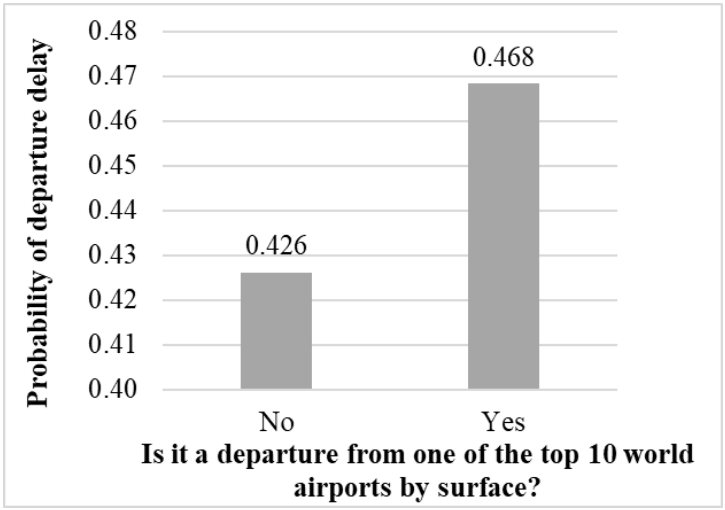

Figure 5 - Influence of being one of the top 10 airports.

The influence of the fifth most relevant feature in departure delay is reflected in Figure 6. As a result, that directly reflects the geographies from where the flights departed. The lower longitudes, where the probability of delay is higher, represent the cities in the American continent, with most of them located in South America and the Caribbean. Next, the delays are slightly higher in Central and South-Eastern geographies (50º - 120º) when compared to Europe and Far-East Asia and Oceania. Thus, more developed economies, with more efficient operational setups and personnel at their airports, can be more efficient, resulting in less delayed departures.

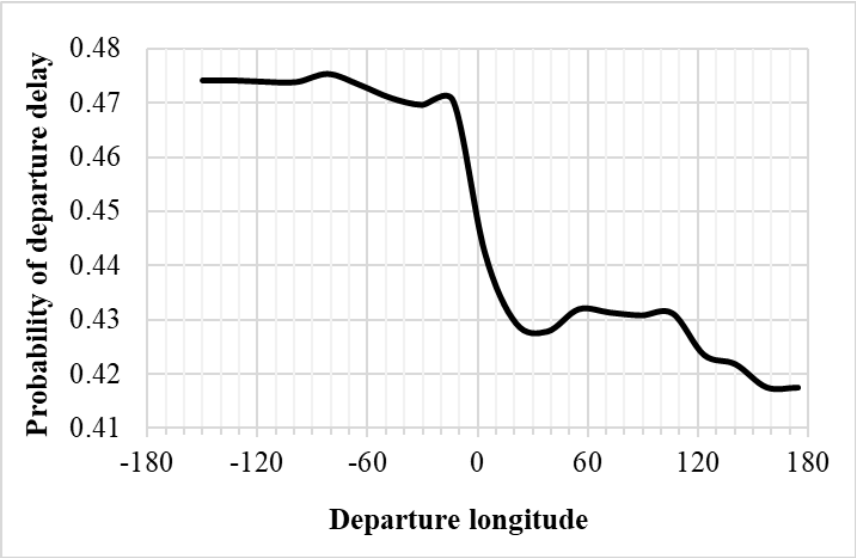

Figure 6 - Influence of departure longitude.

6. Conclusions, implications and future work

Flight departure delays affect travellers, airlines, and airport management. Understanding what drives delays is imperative for improving air transportation management. In this article, we present a theoretical review of air flight studies, and we conducted an empirical study. The empirical part of the study adopts a charter airline perspective to unveil the features that influence departure delays. The used dataset consists of 5,484 flights worldwide between 2014 and 2017 operated by an European airline, including both ad-hoc requested and scheduled ahead flights. The airline dataset provided was enriched by merging it with online open data sources on the airports, and the weather. Thus, a total of 33 features related to the aircraft, the airport (both departure and arrival), the weather, and the flight were used for modelling. A large number of features justified the choice of a data mining approach and, mainly, the data-based sensitivity analysis as suitable to extract knowledge. The model achieved a good performance, proving to be a valuable tool to predict unforeseen flights, as it was validated through a 10-fold cross-validation procedure.

As theoretical implications, this study is among the first studies on charter flights delay prediction, using both proprietary and open access data sources. This study unveiled insights on the most relevant features of a flight delay, via a support vector machine model using the data-based sensitivity analysis. When compared to commercial flights operated on regular schedules, charter airlines are directly pressured by unbalanced demand. The need to fulfil the higher possible number of contracts to assure profitability and make up for demand shortfall periods implies that in some cases, there is no slack available for compensating previous delays. As a result, our study provided empirical evidence of the relevance of previous delays for charter flights, in almost half of the total feature relevance.

Apart from providing an accurate model for predicting departure delay, which is a valuable tool for the airline, the practical contributions of this study are grounded on the knowledge extracted from the model. By computing the features' individual relevance, interesting insights could be brought to light. Although most of the results are consistent with the literature, combining in a single model a myriad of factors to explain departure delay of flights in an airline provides a deeper understanding of each factor's contribution to delays. The previous flight delay features proved to be the most relevant information by accounting for 48% of relevance in explaining departure delay. However, surprisingly, the reason for the previous flight delay in itself held relevance of 2%; on the opposite, the feature indicating if the previous flight had two or more registered reasons accounted for around 33% of relevance. Also, interesting to note is the fact that differences between origin and destination (e.g., if it was an intercontinental or a domestic flight) had less relevance than other features such as the type of flight (scheduled ahead or ad-hoc requested) or the size of the airport.

The proposed approach and consistent results stemming from a real and up-to-date case of an European airline (instead of the most widely used cases based on flight data freely provided by the US authorities) has proven useful for predicting delays on the whole departure process. However, this study has important limitations that should be stated. First, the trained classifier only predicts if a flight departure was delayed by more than 15 minutes, whereas predicting the delay duration itself would render additional relevant knowledge, since a flight delayed by 16 minutes is different from one delayed by 80 minutes (i.e., although a client could file a complaint in both cases, it is more likely to do so for the latter case). Thus, we propose as an extension of the current study to develop a regression model to assess which factors are contributing the most to different delay durations. Also, since no air traffic data was available at the hour of each flight for each of the analysed airports, the complex interactions within air traffic networks were not considered [71]. For example, congestion features are, for sure important for understanding delays, it is clearly an issue that can be addressed in future research. One possibility would be to conduct a more focused study on a small set of airports for which there is data available about the traffic on at least an hour-level granularity. Also, as an avenue for further research, this approach can be enhanced by modelling specific flight departure sub-problems, such as predicting the boarding process of passengers [5]. We conclude this approach can serve as a baseline for implementing an adaptive business intelligence system benefitting from several weighted models and including optimisation algorithms for a more accurate delay prediction. Additionally, we intend to collect more data from the company and train a new model using deep learning approaches through networks combining multiple layers, such as convolutional neural networks [72]. This type of networks is known to

outperform the traditional multilayer perceptron network, especially when handling large amounts of data [73].

Conflict of interests' statement

The authors have no conflict of interests to declare.

References

1. Todnem By R (2005) Organisational change management: A critical review. J Change Management 5(4):369-380.
2. Liang TP, Liu YH (2018) Research Landscape of Business Intelligence and Big Data Analytics: A Bibliometrics Study. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2018.05.018.
3. Seng JL, Chen TC (2010) An analytic approach to select data mining for business decision. Expert Syst Appl 37(12):8042-8057.
4. Kim WC, Mauborgne R (2004) Blue ocean strategy. Harv Bus Rev 82(10):69-80.
5. Schultz M (2018) A metric for the real-time evaluation of the aircraft boarding progress. Transp Res Part C: Emerg Technol 86:467-487.
6. Norin A (2008) Airport logistics: modeling and optimizing the turn-around process. Dissertation, Linköping University Electronic Press.
7. Chan FT, Bhagwat R, Kumar N, Tiwari MK, Lam P (2006) Development of a decision support system for air-cargo pallets loading problem: A case study. Expert Syst Appl 31(3):472-485.
8. Guizzi G, Murino T, Romano E (2009) A discrete event simulation to model passenger flow in the airport terminal. Math Methods Appl Comput 2:427-434.
9. Yimga J (2017) Airline on-time performance and its effects on consumer choice behavior. Res Transp Econ 66:12-25.
10. Erdil A, Arcaklioglu E (2013) The prediction of meteorological variables using artificial neural network. Neural Comput Appl 22(7-8):1677-1683.

11. Sadiq A, Ahmad F, Khan SA, Valverde JC, Naz T, Anwar MW (2014) Modeling and analysis of departure routine in air traffic control based on Petri nets. Neural Comput Appl 25(5):1099-1109.
12. Forbes SJ (2008) The effect of service quality and expectations on customer complaints. J Ind Econ 56(1):190-213.
13. Kohl N, Larsen A, Larsen J, Ross A, Tiourine S (2007) Airline disruption management - perspectives, experiences and outlook. J Air Transp Management 13(3):149-162.
14. Sölveling G, Solak S, Clarke JPB, Johnson EL (2011) Scheduling of runway operations for reduced environmental impact. Transp Res Part D: Transp and Environ 16(2):110-120.
15. Pyrgiotis N, Malone KM, Odoni A (2013) Modelling delay propagation within an airport network. Transp Res Part C: Emerg Technol 27:60-75.
16. Ferguson J, Kara AQ, Hoffman K, Sherry L (2013) Estimating domestic US airline cost of delay based on European model. Transp Res Part C: Emerg Technol 33:311-323.
17. Zou B, Hansen M (2014) Flight delay impact on airfare and flight frequency: A comprehensive assessment. Transp Res Part E: Logist Transp Rev 69:54-74.
18. Tu Y, Ball MO, Jank WS (2008) Estimating flight departure delay distributions - a statistical approach with long-term trend and short-term pattern. J Am Statistical Assoc, 103(481):112-125.
19. Rebollo JJ, Balakrishnan H (2014) Characterization and prediction of air traffic delays. Transp Res Part C: Emerg Technol 44:231-241.

20. Choi S, Kim YJ, Briceno S, Mavris D (2016) Prediction of weather-induced airline delays based on machine learning algorithms. In: Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th (pp.1-6). IEEE.
21. Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10):78-87.
22. Cook AJ, Tanner G (2015) European airline delay cost reference values. Technical Report. EUROCONTROL Performance Review Unit, Brussels.
23. Sternberg A, Carvalho D, Murta L, Soares J, Ogasawara E (2016) An analysis of Brazilian flight delays based on frequent patterns. Transp Res Part E: Logist Transp Rev 95:282-298.
24. Moro S, Cortez P, Rita P (2017) A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features. Neural Comput Appl 28(6):1515-1523.
25. Cortez P, Embrechts MJ (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci 225:1-17.
26. Moro S, Rita P, Oliveira C (2017) Factors influencing hotels' online prices. J Hosp Mark Management 27(4):443-464.
27. Tinoco J, Gomes Correia A, Cortez P (2018) Jet grouting column diameter prediction based on a data-driven approach. Eur J Environ Civ Eng 22(3):338-358.
28. Amin-Naseri MR, Yazdekhasti A, Salmasnia A (2018) Robust bi-objective optimization of uncapacitated single allocation p-hub median problem using a hybrid heuristic algorithm. Neural Comput Appl 29(9):511-532.

29. Di Ciccio C, Van der Aa H, Cabanillas C, Mendling J, Prescher J (2016) Detecting flight trajectory anomalies and predicting diversions in freight transportation. Decis Support Syst 88:1-17.
30. Wandelt S, Sun X (2014) Efficient compression of 4D-trajectory data in air traffic management. IEEE T Intell Transp, 16(2):844-853.
31. Zhang N, Chandrasekar P (2017) Sparse learning of maximum likelihood model for optimization of complex loss function. Neural Comput Appl 28(5):1057-1067.
32. Zou B, Hansen M (2012) Impact of operational performance on air carrier cost structure: evidence from US airlines. Transp Res Part E: Logist Transp Rev 48(5):1032-1048.
33. Ball M, Barnhart C, Dresner M, Hansen M, Neels K, Odoni AR, ... & Zou B (2010) Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States. Federal Aviation Administration report.
34. Torlak G, Sevkli M, Sanal M, Zaim S (2011) Analyzing business competition by using fuzzy TOPSIS method: An example of Turkish domestic airline industry. Expert Syst Appl 38(4):3396-3406.
35. Pai V (2010) On the factors that affect airline flight frequency and aircraft size. J Air Transp Management 16(4):169-177.
36. Clausen J, Larsen A, Larsen J, Rezanova NJ (2010) Disruption management in the airline industry - Concepts, models and methods. Computer Oper Res 37(5):809-821.

37. Wu CL, Law K (2019) Modelling the delay propagation effects of multiple resource connections in an airline network using a Bayesian network model. Transp Res Part E: Logist Transp Rev, 122:62-77.
38. Kafle N, Zou B (2016) Modeling flight delay propagation: A new analytical-econometric approach. Transp Res Part B: Method, 93:520-542.
39. Wandelt S, Sun X, Zhang J (2019) Evolution of domestic airport networks: a review and comparative analysis. Transportmetrica B, 7(1):1-17.
40. Hansen M, Zou B (2013) Airport Operational Performance and Its Impact on Airline Cost. In: Zografos, K., Andreatta, G., & Odoni, A. (Eds.), Modelling and managing airport performance (pp. 119-143), John Wiley & Sons.
41. Sun X, Wandelt S, Cao, X (2017) On node criticality in air transportation networks. Netw Spat Econ, 17(3):737-761.
42. Balakrishna P, Ganesan R, Sherry L (2010) Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. Transp Res Part C: Emerg Technol 18(6):950-962.
43. Huijboom N, Van den Broek T (2011) Open data: an international comparison of strategies. Eur J ePractice 12(1):4-16.
44. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
45. Lan S, Clarke JP, Barnhart C (2006) Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions. Transp Sci 40(1):15-28.
46. Eurocontrol (2018) What is a slot? https://www.eurocontrol.int/news/what-slot Accessed January 11, 2018.

47. Wong JT, Tsai SC (2012) A survival model for flight delay propagation. J Air Transp Management 23:5-11.
48. D'Ariano A, Pistelli M, Pacciarelli D (2012) Aircraft retiming and rerouting in vicinity of airports. IET Intell Transp Syst 6(4):433-443.
49. Abdel-Aty M, Lee C, Bai Y, Li X, Michalak M (2007) Detecting periodic patterns of arrival delay. J Air Transp Management 13(6):355-361.
50. Wei W, Hansen M (2005) Impact of aircraft size and seat availability on airlines' demand and market share in duopoly markets. Transp Res Part E: Logist Transp Rev 41(4):315-327.
51. Haining R (1991) Bivariate correlation with spatial data. Geogr Anal, 23(3):210-227.
52. Chao CM, Yu YW, Cheng BW, Kuo YL (2014) Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst 38(10):1-7.
53. Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 62:22-31.
54. Haykin SS (2009) Neural networks and learning machines (Vol. 3). Upper Saddle River, NJ, USA. Pearson.
55. Moro S, Cortez P, Rita P (2015) Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Comput Appl 26(1):131-139.
56. Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition), NY, USA. Springer-Verlag.

57. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273-297.
58. Gashler M, Giraud-Carrier C, Martinez T (2008) Decision tree ensemble: Small heterogeneous is better than large homogeneous. In:Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on (pp.900-905). IEEE.
59. Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217-222.
60. Liu M, Wang M, Wang J, Li D (2013) Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensor Actuat B-Chem, 177:970-980.
61. Cortez P (2010) Data mining with neural networks and support vector machines using the r/rminer tool, Advances in Data Mining. Applications and Theoretical Aspects, 6171, Springer, 2010, pp.572–583.
62. Muzammal M, Talat R, Sodhro AH, Pirbhulal S (2020) A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks. Inform Fusion, 53:155-164.
63. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861-874.
64. Hand DJ, Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 45(2):171-186.
65. Bengio Y, Grandvalet Y (2004) No unbiased estimator of the variance of k-fold cross-validation. J Mach Learn Res 5:1089-1105.

66. Refaeilzadeh P, Tang L, Liu H (2009) Cross-validation. In: L Liu & MT Özsu (Eds.), Encyclopedia of database systems (pp.532–538). USA: Springer.
67. Williams N, Zander S, Armitage G (2006) A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Comp Com, 36(5):5-16.
68. Cortez P (2014) Modern optimization with R. Springer.
69. Cortez P, Embrechts MJ (2011) Opening black box data mining models using sensitivity analysis. In 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (pp. 341-348). IEEE.
70. Baluch M, Bergstra T, El-Hajj M (2017) Complex analysis of united states flight data using a data mining approach. In: Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual (pp.1-6). IEEE.
71. Sun X, Wandelt S, Zanin M (2017) Worldwide air transportation networks: a matter of scale and fractality?. Transportmetrica A, 13(7):607-630.
72. Gonçalves S, Cortez P, Moro S (2019) A deep learning classifier for sentence classification in biomedical and computer science abstracts. Neural Comput Appl, DOI:10.1007/s00521-019-04334-2.
73. Luo C, Wu D, Wu D (2017) A deep learning approach for credit scoring using credit default swaps. Eng Appl Artif Intel, 65:465-470.

Appendix

Table 6 - Statistics on the included features (features described in Table 3).

Feature	Description *
ac.age	{Min; Q1; Med; Q3; Max}={6; 15; 17; 19; 26}; Avg=16.3; SD=4.7
ac.max.pax	{Min; Q1; Med; Q3; Max}={10; 247; 267; 324; 330}; Avg=263.2; SD=57.7
ac.nr.engines	2 engines = 3692; 4 engines = 1792
ac.registered	Portugal = 5285; Malta = 199
dt.day.of.month	Distribution of flights between 161 for day 17th and 218 for day 26th
dt.season	Spring=1481; Summer=1424; Autumn=1143; Winter=1436
dt.weekend	Weekend=1591; Working day=3893
dt.year	2014=3713; 2015=1243; 2016=427; 2017=101
fl.arr.altitude	{Min; Q1; Med; Q3; Max}={-12.2; 15.0; 53.9; 119.5; 2580.4}; Avg=179.0; SD=333.2
fl.arr.hour	Distribution of flights arriving between 125 at 4 am, and 382 at 12 am
fl.arr.latitude	{Min; Q1; Med; Q3; Max}={-51.8; 22.0; 31.5; 43.7; 69.1}; Avg=28.9; SD=21.4
fl.arr.longitude	{Min; Q1; Med; Q3; Max}={-149.6; 2.0; 11.8; 39.7; 174.8}; Avg=21.1; SD=47.6
fl.arr.night.office	No=2622; Yes=2862
fl.dep.airport.type	Medium=1165; Large=4319
fl.dep.altitude	{Min; Q1; Med; Q3; Max}={-12.2; 15.0; 41.1; 116.4; 2580.4}; Avg=152.9; SD=297.3
fl.dep.country.flights.rank	{Min; Q1; Med; Q3; Max}={1; 13; 39; 53; 171}; Avg=41.5; SD=41.1
fl.dep.hour	Distribution of flights departing between 101 at 1 am, and 406 at 8 am
fl.dep.is.capital	No=3629; Yes=1855
fl.dep.latitude	{Min; Q1; Med; Q3; Max}={-51.8; 22.0; 30.1; 48.7; 69.1}; Avg=28.6; SD=21.7
fl.dep.longitude	{Min; Q1; Med; Q3; Max}={-149.6; 2.0; 17.9; 39.2; 174.8}; Avg=26.0; SD=44.7
fl.dep.top10.airp.prize	No=5314; Yes=170
fl.dep.top10.airp.surface	No=5093; Yes=391
fl.duration	{Min; Q1; Med; Q3; Max}={30; 130; 210; 390; 810}; Avg=259.3; SD=174.7
fl.hired	No=4983; Yes=501
fl.intercontinental	No=2939; Yes=2545
fl.is.adhoc	No=4135; Yes=1349
fl.domestic	No=4670; Yes=814

Feature	Description *
fl.prev.fl.delayed.reason	Aircraft=81; Airline=3151; External=1265; Pre-flight=390; Processing=597
fl.prev.fl.delayed.more1r	No=4960; Yes=524
wt.dwpc	{Min; Q1; Med; Q3; Max}={-24; 7; 14; 20; 28}; Avg=12.7; SD=8.6
wt.relh	{Min; Q1; Med; Q3; Max}={2.7; 51.0; 69.5; 83.8; 100}; Avg=65.7; SD=23.1
wt.sknt	{Min; Q1; Med; Q3; Max}={0; 4; 6; 10; 110}; Avg=7.0; SD=4.9
wt.vsby	{Min; Q1; Med; Q3; Max}={0.03; 6; 6.2; 6.2; 30}; Avg=5.7; SD=1.5
fl.delayed.above15	No=3113; Yes=2371

* For numeric features: {Min; Q1; Med; Q3; Max}={Minimum; Quartile 1; Median; Quartile 3; Maximum}; Avg=Average; SD=Standard Deviation