Accepted refereed manuscript of: Wang X, Brownlee A, Woodward J, Weiszer M, Mahfouf M & Chen J (2021) Aircraft taxi time prediction: Feature importance and their implications. *Transportation Research Part C: Emerging Technologies*, 124, Art. No.: 102892. https://doi.org/10.1016/j.trc.2020.102892

Aircraft taxi time prediction: Feature importance and their implications

Xinwei Wang[a], Alexander E.I. Brownlee[b,*], John R. Woodward[a], Michal Weiszer[c], Mahdi Mahfouf[d], Jun Chen[c,*]

[a]*School of Electronic Engineering and Computer Science, Queen Mary University of London, UK*
[b]*Division of Computing Science and Mathematics, University of Stirling, UK*
[c]*School of Engineering and Materials Science, Queen Mary University of London, UK*
[d]*Department of Automatic Control and Systems Engineering, University of Sheffield, UK*

Abstract

Taxiing remains a major bottleneck at many airports. Recently, several approaches to allocating efficient routes for taxiing aircraft have been proposed. The routing algorithms underpinning these approaches rely on accurate prediction of the time taken to traverse each segment of the taxiways. Many features impact on taxi time, including the route taken, aircraft category, operational mode of the airport, traffic congestion information, and local weather conditions. Working with real-world data for several international airports, we compare multiple prediction models and investigate the impact of these features, drawing conclusions on the most important features for accurately modelling taxi times. We show that high accuracy can be achieved with a small subset of the features consisting of those generally important across all airports (departure/arrival, distance, total turns, average speed and numbers of recent aircraft), and a small number of features specific to particular target airports. Moving from all features to this small subset results in less than a 1 percentage-point drop in movements correctly predicted within 1, 3 and 5 minutes.

Keywords: air traffic management, feature importance, machine learning, prediction, taxi time

1. Introduction

The number of flights made globally by the airline industry has increased steadily since the early 2000s and has reached 39 million in 2019 [1]. Moreover, it is expected

*Corresponding authors.
Email addresses: xinwei.wang@qmul.ac.uk (Xinwei Wang), sbr@cs.stir.ac.uk (Alexander E.I. Brownlee), j.woodward@qmul.ac.uk (John R. Woodward), m.weiszer@qmul.ac.uk (Michal Weiszer), m.mahfouf@sheffield.ac.uk (Mahdi Mahfouf), jun.chen@qmul.ac.uk (Jun Chen)

that passenger numbers could double to 8.2 billion from 2017 to 2037, and importantly, there will be tighter environmental regulations [2, 3]. This trade-off between increasing demand of airport capacity and tighter environmental regulations has been recognised as one of the grand transport challenges [4]. To better address the challenge and limited capacity offered by the existing airport infrastructure, there is an urgent need to develop more intelligent and automated airport traffic management systems.

Developing airports' digital asset twins is the way forward, since it provides a high-fidelity simulation platform to identify bottlenecks and potential operational failures. Simply having digital data does not spontaneously move airports towards intelligence and automation. To realise this, airport traffic management systems based on trajectory-based taxiing operations deserve further investigation [5, 6]. Aligned with the European advanced surface movement, guidance and control systems [7] and the US next generation air transport system programme [8], airport operations based on trajectory-based taxiing will lead to more accurate and efficient aircraft ground movements. Furthermore, they are closely interconnected with other airport optimization problems, e.g. runway scheduling, stand allocation, and airport bus scheduling [9, 10].

In order to realise trajectory-based taxiing operations, accurate taxi time prediction has played an indispensable role. It is not only important to create more robust schedules and identify choke points between gate and runway for practitioners, but also helps the government analysts to estimate the optimal airport capacity and evaluate the regulation impacts [11]. Various modelling techniques have been utilised in the literature to estimate and predict the taxi time, e.g., queuing models [12], statistical regression approaches [11], fuzzy rule-based systems [13] and machine learning techniques [14]. Meanwhile, existing research has adopted different numbers of features (from 5 up to 42) that may affect taxi time from different data sources, including Airline Service Quality Performance (ASQP) and Preferential Runway Assignment System (PRAS) [12, 15], Aviation System Performance Metrics (ASPM) [14, 16, 17], Airport Surface Detection Equipment, Model X (ASDE-X) and Severe Weather Avoidance Programs (SWAP) [18, 19], Spot and Runway Departure Advisor (SARDA) [20, 19], FlightRadar24 (FR24) [21]. This leads to a potential risk that certain features related to the taxi time have not been included. For instance, only a few publications considered the impact of weather conditions on taxi time [22, 23, 20], and these relevant studies simply considered the weather conditions as *severe* and *fine* weather. Meanwhile, some defined features are only obtained when the taxi procedure is completed, which is not realistic for taxi time prediction in practise [11]. Moreover, more information can be obtained through digging into the raw data source, e.g., defining new metrics to directly reflect the surface traffic congestion [24, 25].

Another drawback we have observed falls in the feature importance identification process. Although as mentioned, various prediction methodologies have been applied in this domain, only a few of existing research addressed the identification of feature importance [26, 27]. Assuming there is no priority information for the features, Jordan

et al. [26] applied a sub-optimal subset selection method [28] to capture much of the variability in the data source, though with no guarantee of ensuring high-quality prediction accuracy. Herrema *et al*. [27] defined two metrics to rank the candidate features, and selected the top ten of them as the most important features. Note a sufficient analysis of the selected features number was not provided. Therefore, there is still a strong need to better understand which features and to what extent the features affect the taxi time. We aim to develop a feature importance identification procedure with quantitative analysis to address this issue, and better underpin subsequent applications of the taxi time prediction in uncertain airport runway and taxi scheduling problems.

Consequently, the overarching original contributions of this new study include the following: (1) To the best of our knowledge, a complete feature group for the taxi time prediction has not been provided; for the first time we introduce a set of weather features, aircraft speed and runway utilisation information that may affect the taxi time, and compare the prediction performance using different models and feature groups. (2) A backward feature importance identification process with quantitative analysis is applied to taxi time prediction. (3) The feature importance identification process integrating with prediction model is validated using real-world freely-available data for three international airports. (4) We show that the taxi time prediction models only including aircraft departure/arrival, distance, number of existing departure flight, average speed of recent aircraft and specific features for target airports can provide a high level of accuracy.

This paper is organized as follows: Section 2 provides a comprehensive literature review in the taxi time prediction; Section 3 introduces three international airports and a complete set of features that could impact taxi time; the prediction models, performance metrics as well as the developed feature importance identification process are described in Section 4, followed by the computational results and discussions presented in Sections 5 and 6; finally, conclusions are drawn in Section 7, highlighting the important contributions of our work and potential future directions. Appendices listing the abbreviations we use throughout the paper, and giving standard deviation of the results are also included.

2. Literature Review

To clearly present the state-of-the-art of taxi time prediction methodologies, we classify the reviewed literature in line with four categories: queuing models, statistical regression approaches, Fuzzy Rule-Based Systems (FRBSs) and other machine learning techniques. An overview of the studies in taxi time prediction is listed in Table 1.

The queuing model has been generally utilised in the early research of taxi time prediction. Pujet *et al*. [12] and Idris *et al*. [15] analysed real-world data at Boston Logan International Airport (BOS) and identified the runway configuration, the airline/terminal, the downstream restrictions and the takeoff queue size as the main features that affect the taxi time. Consequently, the queuing models were developed to further improve the accuracy of taxi time prediction. Simaiakis and Pyrgiotis [22] modelled the aircraft

Table 1: Overview of the studies in taxi time prediction.

Authors	Year	Methodology	Features considered in the model	Data source	Airport	Movements
Pujet *et al.* [12]	2000	Queuing model	Airport operational information, e.g. scheduled and actual pushback time, takeoff/landing time and gate arrival times, and departure/arrival runways and capacity	ASQP and PRAS	BOS	/
Idris *et al.* [15]	2002	Queuing model	Airport operational information, e.g. scheduled and actual pushback time, takeoff/landing time and gate arrival times, and departure/arrival runways and capacity, downstream restrictions	ASQP and PRAS	BOS	26302
Balakrishna *et al.* [14]	2008	Reinforcement learning	5 features including time spent in the runway queue, the number of departure/arrival aircraft during taxiing, average taxi time of the previous half hour, the time of day	ASPM	JFK	254 days
Balakrishna *et al.* [16]	2010	Reinforcement learning	8 features including number of flights features, average taxi time features, and time of day	ASPM	TPA	˜132200
Jordan *et al.* [26]	2010	Linear regression	17 features including taxi distance, number of flights features, airline, runway direction	Runway Status Lights System	DFW	4720
Simaiakis and Pyrgiotis [22]	2010	Analytical queuing model	Number of flights features, the departure capacity, and time of day	ASPM	BOS and EWR	99196
Chen *et al.* [13]	2011	Fuzzy rule-based systems	14 features including taxi distance, taxi turning angle, departure/arrival, number of flights features, operational modes	Airport	ZRH	679
Srivastava [18]	2011	Linear regression model	Taxi distance, number of flights features, average taxi time of the previous quarter, severe weather or not	ASDE-X and SWAP	JFK	43 days
Diana [23]	2013	Survival and frailty analytical models	Airport operational information, e.g.block delay, departure/arrival delay and percentage of capacity utilized, good weather or not	ASQP	JFK	1250
Ravizza *et al.* [11]	2013	Multiple linear regression	15 features including taxi distance, taxi turning angle, departure or arrival, number of flights features and some less important factors	Airports	ARN and ZRH	1340
Ravizza *et al.* [24]	2014	Multiple linear regression, least median squared linear regression, support vector regression, M5 model trees and fuzzy rule-base systems	16 features including taxi distance, taxi turning angle, departure or arrival, number of flights features and some less important factors	Airports	ARN and ZRH	7607
Lee *et al.* [20]	2015	Linear optimized sequencing, linear regression, support vector machines, k-nearest neighbors and random forest	Taxi distance, assigned gate, spot and runway, number of flights features	SARDA	CLT	332
Lee *et al.* [19]	2016	Linear regression, support vector machines, k-nearest neighbors, random forest and neural networks	12 features including terminal concourse, gate, spot, runway, departure fix, aircraft model, aircraft weight, taxi distance, time of day, number of flights features and unimpeded taxi time	SARDA and ASDE-X	CLT	246083
Lordan *et al.* [29]	2016	Linear regression	15 features including gate, runway, departure or arrival, number of flights features	Airport	BCN	35858
Chen *et al.* [21]	2017	Multi-objective fuzzy rule-based systems	same as Ravizza 2013	FR24	MAN	1413
Diana [17]	2018	Ensemble machine learning, ordinary least-squared and penalized algorithms	5 features including departure demand/throughput, percentage of airport capacity utilised, approach conditions and runway configurations	ASPM	SEA	2760
Herrema *et al.* [27]	2018	Neural networks, regression tree, reinforcement learning and multilayer perceptron	42 features including operational information, congestion/capacity level, unimpeded taxi time and number of departures	Airport	CDG	˜1 million
Lian *et al.* [30]	2018	Two improved support vector regression methods, generalized linear regression, softmax regression and artificial neural network	6 features including taxi distance, number of flights features, delay and takeoff/pushback times	Airport	PEK	17 days
Yin *et al.* [25]	2018	Machine learning with a macroscopic network topology	21 features including surface instantaneous flow indices, cumulative flow indices, aircraft queue indices and slot resource demand indices	Airport	PVG	/
Mirmohammadsadeghi *et al.* [31]	2019	Statistic regression, percentiles method and data surveillance	5 features including speed, waiting time, departure runway and wheels-on/off	ASDE-X	CLT *et al.*	/

departure process as a queuing system, aiming to predict taxi time through analytic taxiway and runway queues approximations. This developed runway queuing model is stochastic and easily transferred to different airports; it has been validated against real data at BOS and Newark Liberty International Airport (EWR).

Meanwhile, statistical regression approaches have been applied in the taxi time prediction. Jordan *et al.* [26] presented a statistical linear regression (LR) approach to modelling aircraft taxi time at Dallas/Fort Worth Airport (DFW). Combining with a feature selection method, the developed model achieved 98.3% prediction accuracy within 3 min absolute error. Notice only data on good weather days were applied to train and test the model. Building on a historical traffic flow database, Srivastava [18] established an adaptive taxi time prediction model with LR analysis, where a set of explanatory variables including aircraft queue position, taxi distance are included. Using actual data from John F. Kennedy International Airport (JFK), the prediction model has been demonstrated with high accuracy.

Combining both airport layout and historic taxi time information, Ravizza *et al.* [11] presented a taxi time prediction model with a multiple linear regression (MLR) analysis. Data from Stockholm-Arlanda Airport (ARN) and Zurich Airport (ZRH) was utilized for the experiments, and the taxi distances, the sum of turning angle, aircraft departures or arrivals and the amount of traffic when the aircraft is taxiing were identified as the important features for taxi time prediction. Furthermore, various regression approaches including MLR and least-medium-squared LR were testified in [24]. Inspired by [11], Lordan *et al.* [29] considered route- and interaction-specific features and designed a log-LR model for taxi time prediction at Bacrelona-EI Airport (BCN). Experimental results have verified the strong predictive validity of the proposed model, while a sample size covering an extensive airport operational period is required. Recently, Mirmohammadsadeghi *et al.* [31] introduced a regression method, which is being applied by the Federal Aviation Administration for unimpeded taxi time estimation.

FRBSs [32], which can offer more explanations of the underlying behavior, were introduced for taxi time prediction by Chen et al [13]. The results at ZRH indicate that FRBSs are a valuable alternative to existing statistical methods. Detailed comparisons including various regression approaches and FRBSs were conducted in [24], demonstrating the FRBSs outperformed other approaches in terms of prediction accuracy at ZRH. To address the prediction accuracy as well as associated uncertainty, Chen *et al.* [21] further developed a multi-objective FRBS based approach for taxi time prediction at Manchester Airport (MAN), in which the structure of the FRBS was simplified and only one predominate rule accounted for one taxi scenario.

In addition to FRBSs, other machine learning techniques also contribute to the taxi time prediction literature. Based on one of the busiest US airport JFK, Balakrishna *et al.* [14] designed a probabilistic framework with reinforcement learning (RL) strategy to predict the aircraft taxi time. The results indicated the RL estimator is capable to

capture the dynamics at challenging airports such as JFK, while the prediction accuracy on individual flight needs to be improved. Furthermore, Balakrishna *et al.* [16] conducted a case study at Tampa International Airport (TPA), realising 81% prediction accuracy with a standard error of 2 min.

Lee *et al.* [20] applied the linear optimized sequencing approach to develop a discrete-event fast-time simulation tool for taxi time prediction. A data-driven analytical method using four machine learning techniques is introduced for comparison as well. These methods were evaluated with actual data at Charlotte Douglas International Airport (CLT), and experiments indicate that the developed simulation tool is competitive. Lee *et al.* [19] further considered weather conditions at CLT to improve the taxi time prediction. However, simulation results indicate that the prediction accuracy has not been improved. The reason could be that the weather conditions were simply divided into good weather and rainy days, and other weather properties such as wind, visibility and temperature have not been investigated.

Assuming the taxi time is a function of several features which may not be expressed in existing models, Diana [23] conducted a survival and frailty analysis, and revealed the block delay and capacity utilisation percentage would impact the taxi time as well. A comprehensive comparison of various prediction models was then presented at Seattle International Airport (SEA) [17]. The ensemble machine learning, ordinary least-squared and penalized approaches were testified and the results suggest that no algorithm outperforms others in all cases, and one should strike a balance between the prediction bias and variance. Herrema *et al.* [27] focused on Neural Networks (NN), Regression Tree (RT), RL and multilayer perceptron (MLP) methods for the taxi time prediction at Charles de Gaulle Airport (CDG). The top 10 out of 42 features, e.g., unimpeded taxi time, congestion level, and number of departures in the last 20 minutes were chosen in the feature selection process, and RT turned out to be the most efficient method.

Comparing with traditional taxi time prediction methods, e.g., LR, softmax regression and NN, Lian *et al.* [30] developed two improved support vector regression (SVR) approaches in a case study of Beijing International Airport (PEK). Several features including queue length, taxi distance and potential landing number were identified, and a high prediction accuracy up to 95% within 5 minutes was achieved. Based on a macroscopic network topology, Yin *et al.* [25] formulated the taxi time relevant features into four groups: surface instantaneous flow, cumulative flow, aircraft queue length and slot resource demand. Three machine learning methods including LR, SVR and Random Forest (RF) were then applied in the taxi time prediction. Using the historical data at Shanghai Pudong International Airport (PVG), computational results demonstrated the effectiveness of the proposed features and machine learning techniques.

Throughout the above literature review, we can identify the gaps in the current research for taxi time prediction as follows: due to the incomplete data source, several features that may influence the taxi time have not been comprehensively investigated.

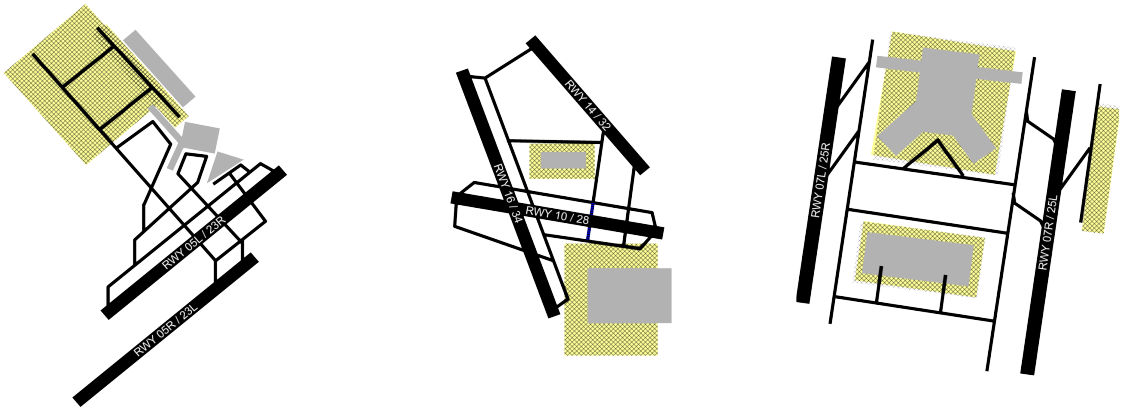
For instance, although existing literature noticed the presence of weather conditions, the weather features were simply classified as severe/ fine weather [22, 23, 20]. Meanwhile, to ensure taxi time prediction accuracy and underpin robust airport traffic management, e.g., airport ground movement routing/scheduling and runway scheduling, the feature importance identification with quantitative analysis requires to be further addressed. We aim to fill these gaps in this research.

3. Data

This study utilises data from three international airports in Europe and Asia: Manchester Airport (MAN), the third biggest airport in the UK; Zurich Airport (ZRH), the largest airport in Switzerland; Hong Kong International Airport (HKG), ranking among the global top ten busiest airports. The layout of the three airports are illustrated in Figure 1. The real-world aircraft movement information is taken from freely-available data on the website FlightRadar24, following the techniques described in [33, 34] (The tools are available at `https://github.com/gm-tools/gm-tools`). FlightRadar24, which has also been used to gather airborne flight tracks [35, 36, 37, 38], collects automatic dependent surveillance-broadcast (ADS/B) messages transmitted by many aircraft. These messages contain the latitude, longitude and altitude, usually every 5 to 10 seconds. The coordinates have a resolution of 10^{-4} degrees: approximately 10m for our target airports. While not all aircraft broadcast ADS/B data, and of the broadcast data has calibration errors or corruptions needing cleaned before use, enough flight movements are present to allow reliable taxi time estimation models to be derived. For this work, all tracks for aircraft with an altitude of zero within 5km of each airport's centre were collected. The raw tracks were snapped to the actual taxiways by searching for all taxiways within 10m of each coordinate and deriving the most likely route taken, taking the shortest path between the coordinate points from ADS/B except where those lead to sharp turns. Each movement contained the taxi route taken and the real time at the start and end. The corresponding weather information is extracted from the METAR Weather Service (`https://www.aviationweather.gov/metar`). We accessed 14 872 movements data for MAN from 21st Jan 2017 to 13rd Apr 2017, 19 808 movements for ZRH from 21st Jan 2017 to 18th Mar 2017, and 42 397 movements for HKG from 15th Jan 2017 to 21st Feb 2017. After snapping these raw movements to the known taxiways at each airport, the data contained 10 216, 11 271 and 33 095 tracks for MAN, ZRH and HKG respectively; the missing tracks did not have enough points aligned with the taxiways for the taxi routes to be determined with certainty. These were further reduced to 10 210, 11 248 and 33 060 tracks after removing any taxi routes longer than 45 minutes (substantial outliers and deemed to be erroneous).

For the purposes of this study, we consider taxi-time to be the actual push-back time to the actual line-up time for departures, and the actual time of leaving the runway to the actual on-block time for arrivals. The average taxiing times at MAN, ZRH, HKG are 9.6,

6.6 and 11.3 minutes respectively. Notice the movements without complete information (i.e. full path between runway and stand) have been removed in advance.



(a) MAN (b) ZRH (c) HKG

Figure 1: Sketches of the layout at three airports.

In order to ensure taxi time prediction accuracy, one should comprehensively consider relevant features that may affect taxi time. In this study, we use up to 33 features, aiming to provide a sufficient set of features for the taxi time prediction. These relevant features are divided into four categories, including aircraft and airport operational factors, airport congestion, aircraft average speed and weather information.

3.1. Aircraft and airport operational features

Eight aircraft and airport operational features which can be easily extracted from the raw data are listed in Table 2. The binary feature *depArr* indicates whether the movement is a departure or arrival flight. *Distance* is the entire taxi distance from the gate to the runway; *distance_long* is the sum of straight taxiway lengths exceeding 500 metres. Two features related to aircraft turning angles that may affect the taxi speed are included in the prediction model as well. Besides, aircraft weight has been considered as a potential candidate factor for taxi time prediction [11, 27]. In line with the aircraft wake vortex [39], we introduced feature *aircraft_weight* to categorise the aircraft as small, medium and large.

The *runway_info* feature reflects which runway as well as which direction is used for aircraft landing or take-off. For instance, as shown in Figure 1(a), MAN has two runways for aircraft departure and arrival. Typically the two runways are both used from 6 am to 10 pm of the day, while only one runway is utilised for the rest of the day. During the dual runway operation, the aircraft normally use runway 23L/R for arrival and

Table 2: Features of aircraft and airport operational information.

Features	Type	Description
depArr	Binary	Departure(0) or arrival(1)
distance	Numerical	Sum of taxi distance in metres
distance_long	Numerical	Sum of taxi distance in metres on straights of more than 500 metres
angle_sum	Numerical	Sum of turning angle in degrees
angle_error	Numerical	Count of 180 degree turns (usually 0, or 1 corresponding to a pushback)
aircraft_weight	Categorical	Aircraft weight category: small, medium and large
runway_info	Categorical	Aircraft runway information category
budget_airline	Binary	Budget (0) or non-budget (1) airline

departure, while it might use runway 05L/R when the wind direction changes. Therefore we define $2 \times 4 = 8$ runway information categories for MAN, where 2 indicates two runway operation conditions, and 4 the two directions of the two runways. This identifies the specific runway used by each flight. We use this in contrast to the more course-grained 'operating mode' (i.e. which runways were in use at the time of the flight) as it also services to indicate which area of the airport in which the aircraft either began or completed taxiing. Similarly, ZRH has 6 *runway_info* types since there are two directions of three runways. For HKG, 4 *runway_info* modes are categorised given two directions of two runways.

The last feature of aircraft information is *budget_airline*, aiming to investigate whether budget airline flight would impact taxi time for economic reasons. In this research, Ryanair and easyJet are categorised as budget airlines, while others as non-budget airlines.

3.2. Airport congestion features

The airport congestion features are first introduced to the taxi time prediction in [11, 24] inspired by the queuing model [15]. These features shown in Tables 3 and 4 have been demonstrated as effective factors to improve taxi time prediction performance. Eight features are designed to represent the airport congestion conditions via counting the number of arrivals and departures during the time the current aircraft is taxiing for departure or arrival. Note the features in Table 3 indicate the congestion information when the current aircraft starts taxiing, while the ones in Table 4 denote the congestion conditions until current aircraft has completed the movement.

The information of the features in Table 4 can only be obtained when the aircraft has completed its taxiing process. Therefore, these features actually cannot be utilised

Table 3: Features of airport congestion information.

Features	Type	Description
NDepDep	Numerical	Number of other aircraft on the way to runway when current aircraft pushes back
NDepArr	Numerical	Number of other aircraft on the way to stand when current aircraft pushes back
NArrDep	Numerical	Number of other aircraft on the way to runway when current aircraft lands and starts taxiing
NArrArr	Numerical	Number of other aircraft on the way to stand when current aircraft lands and starts taxiing

Table 4: Features of airport historical congestion information.

Features	Type	Description
QDepDep	Numerical	Number of other aircraft that reach runway and depart while current aircraft is on the way to runway
QDepArr	Numerical	Number of other aircraft that arrive at stand while current aircraft is on the way to runway
QArrDep	Numerical	Number of other aircraft that reach runway and depart while current aircraft is on the way to stand
QArrArr	Numerical	Number of other aircraft that arrive at stand while current aircraft is on the way to stand

for practical taxi time prediction. In light of this, we only choose the congestion features listed in Table 3, excluding the historical congestion information in Table 4.

Given the aircraft queuing length could have more impact on the taxi time of departure flight [12, 24], we expect features *QDepDep* and *QDepArr* could have stronger contributions to the models compared to *QArrDep* and *QArrArr*.

3.3. Aircraft average speed features

Table 5: Features of aircraft average speed information.

Features	Type	Description
AvgSpdLast5Dep	Numerical	Average speed of latest departing 5 aircraft
AvgSpdLast5Arr	Numerical	Average speed of latest 5 arriving aircraft
AvgSpdLast5	Numerical	Average speed of latest 5 aircraft
AvgSpdLast10Dep	Numerical	Average speed of latest departing 10 aircraft
AvgSpdLast10Arr	Numerical	Average speed of latest 10 arriving aircraft
AvgSpdLast10	Numerical	Average speed of latest 10 aircraft

The taxi speed of other moving aircraft can reasonably be expected to be related to the actual taxi time, acting as a proxy for many other factors influencing the taxi speed. The idea is that average speed of other aircraft is easy to calculate and might capture confounding factors that are difficult to explicitly measure. Thus, it is somewhat surprising that, until now, the features relevant to the speed of other aircraft have not been studied in the literature. We are the first to introduce the speed features into taxi time prediction, through defining six average speed features as shown in Table 5, where the units are metre per minute. Notice we collect the latest average speed for departure, arrival or both of them. This is because that the previous research indicated the airport congestion conditions of departure aircraft have larger impact on the taxi time [24]. In case the average speed features also have different impacts on departure and arrival aircraft, we define these speed features for departure and arrival aircraft respectively, better capturing the influence of speed features on the taxi time. Precisely how many aircraft should be counted for these features is somewhat arbitrary, so we explored two numbers (5 and 10) to determine whether the number made much difference. Similar to the airport congestion features, the average speed features related to departure flight, e.g., *AvgSpdLast5Dep* and *AvgSpdLast10Dep*, are expected to have more contributions to the taxi time predictions compared to arrival related features.

3.4. Weather information features

Although existing research did not indicate close correlation between the weather conditions and taxi time [22, 23, 20], it is expected that the airport local weather conditions should impact the taxi time to some extent [11, 27, 17]. To better address and

Table 6: Features of weather conditions.

Features	Type	Description
Pressure	Numerical	Air pressure in inHg
Temperature	Numerical	Temperature in Celsius
WindSpeed	Numerical	Wind speed in metres per second
Visibility	Numerical	Visibility in metres
isRain	Binary	Whether it is raining
isSnow	Binary	Whether it is snowing
isDrizzle	Binary	Whether it is drizzling
isFog	Binary	Whether it is fogging
isMist	Binary	Whether it is misting
isHaze	Binary	Whether it is hazing
isHail	Binary	Whether it is hailing

analyse possible influence of the weather conditions on aircraft taxi time, a list of eleven weather features rather than simply defining bad/fine weather are presented in Table 6. The abundant weather information is promising to reveal potential relations between the weather conditions and taxi time. However, as the flight could be postponed or cancelled under extreme weather conditions, e.g., blizzard or heavy rains, the collected aircraft taxiing data probably does not include terribly bad weather. This could limit the impact of the weather conditions on taxi time prediction.

4. Prediction models and feature importance

In this section, five prediction models are introduced for predicting the taxi time, and seven performance metrics are adopted to address and compare the model performance. We also develop a feature importance identification procedure, aiming to identify important features and provide high prediction accuracy with a narrow subset of features.

4.1. Models

The selection of prediction models could impact the prediction performance. In this study, we introduce five models for taxi time prediction, specifically MLP, LR, Polynomial Regression (PR), Gradient Boosted Regression Trees (GBRT) and RF.

4.1.1. Multilayer perceptron

NNs are brain inspired models which allow a machine to learn from available data [40]. As a class of feedforward artificial NN, MLP consists of at least three layers of nodes:

an input layer, a number of hidden layers and an output layer. The nodes are connected between each layer with different weight. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. For instance, the commonly used activation function is constructed as

$$y(v) = \max(0, v) \tag{1}$$

where v is the input and $y(v)$ denotes the output.

MLP trains the network by using back propagation [41], which is widely utilised for a general NN.

4.1.2. Linear regression

The concept of LR is that there is a relationship between an independent feature and a dependent one. If the two variables move in the same direction, then there is a positive relationship between the two variables. On the other hand, if the independent variable increases and the dependent variable reduces (and vice versa), there is a negative relationship between the two variables.

The general linear regression formula is the following:

$$\hat{y} = \alpha x + \beta \tag{2}$$

where $\hat{y}$ is the prediction output, x is the vector of input features, α and β are coefficients vectors.

When fitting this model, we aim to find α and β that minimize the defined cost function, e.g. the ordinary least squares as

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

where y_i and $\hat{y}_i$ are the ith actual and predictive output values respectively, and n is the number of data samples.

4.1.3. Polynomial regression

PR is a technique of regression analysis, where the relationship between the dependent feature and the independent feature is described by certain polynomial degrees in the dependent variable. LR is a special case of PR, in which the polynomial order equals one. This method is beneficial for describing curvilinear relationships, while it may easily be over-fitted. Therefore one should carefully select the polynomial orders for PR. The model with nth order polynomial regression is constructed as

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m \tag{4}$$

where $\hat{y}$ is the prediction value, $x^z = \{x_1^z, x_2^z, \ldots, x_f^z\}(z = 1, 2, \ldots, m)$ are the zth polynomial input features, f is the number of input features and $\beta_i(i = 0, 1, \ldots, m)$ are regression coefficients.

Similar to LR, the coefficients of the PR model are determined by minimising the cost function, e.g. the ordinary least squares.

4.1.4. Gradient boosting regression trees

GBRT is a flexible non-parametric statistical learning technique for regression. It builds the model in a stage-wise fashion, and allows optimization of an arbitrary differentiable cost function [42].

Like other boosting methods, gradient boosting combines weak learners into a single strong learner in an iterative manner. At each stage s of the gradient boosting, there exists a current prediction model F_s that can be further improved. Through adding an estimator h, a more accurate prediction model is obtained by constructing $F_{s+1}(x) = F_s(x) + h(x)$. This implies that the selection of estimator h can be expressed as

$$h(x) = y - F_s(x) \tag{5}$$

where y is the actual output value.

Specifically, the gradient boosted regression tree, i.e. GBRT, has been widely used due to its efficiency, accuracy and interpretability [43], making it a promising approach to accurately predict the aircraft taxi time. GBRT builds a series of regression trees, statistical models generated for supervised prediction problems. They make their predictions by a series of decisions represented in a tree structure, in which each node is a split in the possible values for one feature. The benefit of decision tree regression is that it is easy to interpret and visualise, and can possibly reveal patterns which may be difficult in being identified through traditional regression methods.

4.1.5. Random Forest

RF is a prediction model made up of many decision trees [44, 45]. RF can usually generate good results even without tuning the hyperparameters [46]. Each tree in a RF learns from a set of randomly sampled data during the training process, a technique known as bagging. Notice the samples would be repeatedly applied in a single tree, so that the entire forest will have lower variance without increasing the bias. The outputs of RF are obtained through averaging predictions of each decision tree in RF.

Whilst decision trees search for a split on each feature in each node, RF investigates for a split on only one feature in a node. First, a small subgroup of explanatory features is randomly selected. Next the node is split with the best feature among the small number of randomly selected features. After splitting, a new list of eligible features is chosen arbitrarily. This process continues until the tree is completely grown. Ideally, in every terminal node there will be only one observation. As the number of features increases,

the eligible feature set will be quite different from node to node. Nevertheless, significant features finally appear in the tree and their respective success in prediction will lead to more reliability.

4.2. Performance metrics

In this section, we provide seven prediction performance metrics to evaluate and compare the introduced prediction models.

Accuracy

The accuracy of prediction models measures the relative percentage difference between the predictive and actual values. It is defined as

$$\text{accuracy} = (100 - \frac{100}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}|)\% \tag{6}$$

where n is the number of data samples.

R^2

R^2, namely the coefficient of determination, denotes the variation in the dependent variable explained by the independent variables. The value of R^2 is no more than 1, which is the best possible value. Notice it could be negative when a model tries to fit nonlinear functions to sampled data [17]. The mathematical definition of R^2 is expressed as

$$\mathrm{R}^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \frac{1}{n}\sum_1^n y_i)^2} \tag{7}$$

Mean Absolute Error

Mean Absolute Error (MAE) is defined to measure the average absolute deviations between the predictive and actual values. It is formulated as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

Root Mean Squared Error

Similar to MAE, Root Mean Squared Error (RMSE) is also an important metric to evaluate the model performance. By squaring the errors it gives a greater weighting to data points with a larger error. It is characterised as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

Prediction accuracy within 1,3 and 5 minutes

In practical aircraft taxi time prediction, small prediction errors are tolerable; we are more interested in the prediction accuracy within the predefined threshold [24, 19, 27, 25]. In this study, we set the the threshold as 1, 3 and 5 minutes respectively, and provide the prediction accuracy within the corresponding thresholds.

4.3. Feature importance identification procedure

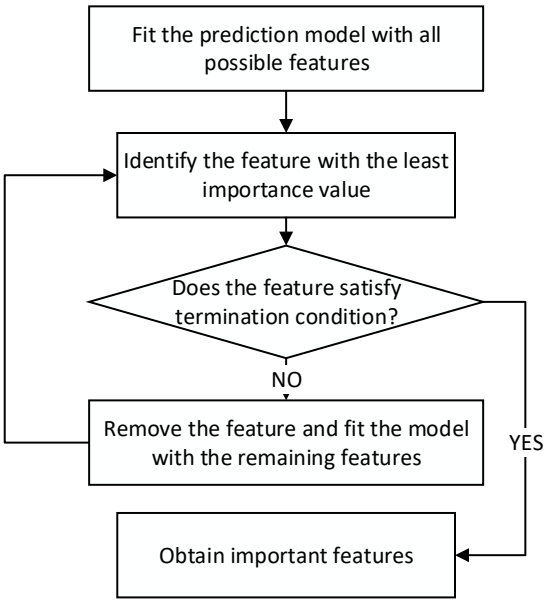


Figure 2: Feature importance identification with backward elimination.

As discussed in Section 1, previous research mainly focused on the performance of developed prediction models, while little work has been done on the feature importance identification with quantitative analysis. To better address this issue, we use an effective backward feature elimination procedure [47] to identify the important features that affect the taxi time. A flowchart of the proposed feature importance identification process is illustrated in Figure 2. The idea of backward feature elimination is to select one with the least importance value at one time (Note the importance value is calculated via the prediction model itself, and the value is dynamically changed when features are removed), and check whether the termination conditions have been met when removing the selected feature. If not, the selected feature would be removed and step into choosing another feature; otherwise the procedure terminates and the selected feature with remaining ones would be defined as important.

The core of the procedure is to set proper termination conditions with quantitative analysis, aiming to provide strong confidence in the feature importance identification.

Two termination conditions are provided as follows: 1) check whether current feature importance is greater than a certain value, where the feature importance value is calculated by the prediction model itself; or 2) check whether the prediction performance goes down by a certain value, in which the prediction accuracy within 5 minutes is selected as the performance metric.

5. Results: Model configuration and Tuning

In this section, we first present the experiment setup, and then report the performance of different models for taxi time prediction at three international airports. The goal of this section is to settle on a modelling approach that produces accurate results for the target application before we focus on our investigation of feature importance.

5.1. Experimental setup

The computational experiments are conducted on a laptop equipped with Intel Core i5 CPU at 2.5 GHz and 8 GB of RAM on a Windows 10 64-bit OS. The prediction models are all implemented in Python using the Scikit-Learn library. The polynomial order of PR is selected as 2 according to preliminary experiments. Detailed parameters for MLP, GBRT and RF are set as follows: the activation fuction, size of hidden layer, learning rate and maximum iteration number of MLP are Relu, 100, 0.001 and 10000 respectively. The loss function, learning rate, number of boosting stages and evaluation criterion of GBRT are least squares regression, 0.1, 1000 and mean squared error. For RF, the number of trees, evaluation criterion, maximum depth of the tree and minimum number of samples are 150, mean squared error, expanding until all leaves are pure, and 10 respectively. The information of the utilised data is referred to Section 3, in which 33 features are collected from three international airports, i.e., MAN, ZRH and HKG. 70% of the entire data is used for training, and the remaining 30% is used for testing. The training data was then split using 10-fold cross validation. To fairly compare different prediction models, the same seeds were used to randomly generate the subsets for cross-validation. The performance metrics marked *Training* presented in the tables below are the average values over the ten folds, and those marked *Testing* are on the 30% unseen test data.

For convenient performance comparisons on different features, we further divide the introduced features into five groups as follows. These feature groups will be applied and compared in the following experiments.

- Group A (8 features): Include general aircraft and airport operational information denoted in Table 2.
- Group B (12 features): Add airport congestion information (Table 3) to Group A.
- Group C (18 features): Add aircraft average speed information (Table 5) to Group B.

- Group D (29 features): Add airport local weather information (Table 6) to Group C.
- Group E (33 features): Add airport historical congestion information (Table 4) to Group D.

5.2. Base line

Table 7: Results of base line prediction methods.

	MAN		ZRH		HKG	
	EURO	SELF	EURO	SELF	EURO	SELF
Accuracy (%)	-24.93	41.28	-262.33	-32.30	-185.83	55.56
R^2	-1.19	0.32	-1.68	0.38	-3.16	0.63
MAE	6.59	3.16	7.54	2.74	14.57	3.40
RMSE	7.84	4.37	8.79	4.21	15.82	4.69
$<$ 1 min (%)	8.05	23.41	6.40	30.17	1.07	19.66
$<$ 3 min (%)	24.62	59.27	19.91	71.13	3.65	60.90
$<$ 5 min (%)	39.82	81.01	34.66	84.82	8.14	77.73

Before we proceed to our modelling results, as a base line to our experiments, Table 7 presents prediction results using the average taxi-in and taxi-out time from EUROCONTROL (EURO) and the collected data itself (SELF). The average taxi-in and taxi-out times at MAN, ZRH and HKG are 6.7/14.3, 5.4/13.5 and 6.6/23.7 minutes from EUROCONTROL report. The corresponding times from the data itself are 5.0/11.6, 2.8/9.6 and 4.8/17.2 minutes.

5.3. Performance of different prediction models

Since Group D contains all available features that are practical for predicting aircraft taxi time, it is utilised for model comparisons at three international airports. The results are provided in Tables 8 to 10, in which the best value for each performance metric is marked in bold.

Overall, RF outperforms other models across the three airports, closely followed by GBRT. RF also has a very similar performance on the training and testing sets. PR performs better than LR in all metrics for ZRH and HKG, while LR has close MAE and RMSE compared to PR. This is logical – LR can be viewed as a special PR model with only 1st linear order, and the addition of 2nd order in PR may sacrifice certain absolute errors to improve the prediction accuracy within time intervals. MLP always has poorest results, and the performance differences of the training and testing sets are relatively sinificant, specially for ZRH. We note that accuracy for ZRH is lower than for the other airports, whereas the other metrics vary much less. The results for the baseline suggest that simple averages yield poorer results at ZRH than the other two airports; in turn suggests a greater variation in times. This makes sense as the runway entry/exit points are much more spread out than at the other airports. With runway crossings too, it is a

Table 8: Prediction models performance comparisons for Manchester Airport. Bold numbers indicate the best performance in terms of the corresponding metrics.

Model		MLP	LR	PR	GBRT	RF
Training	Accuracy (%)	50.06	63.53	63.72	68.18	**69.89**
	R^2	0.30	0.55	0.49	0.58	**0.60**
	MAE	3.30	2.48	2.52	2.38	**2.29**
	RMSE	4.44	3.58	3.79	3.43	**3.35**
	$<$ 1 min (%)	21.10	31.13	30.90	31.97	**33.84**
	$<$ 3 min (%)	56.72	70.77	70.75	72.63	**74.38**
	$<$ 5 min (%)	78.63	87.82	87.89	89.20	**89.75**
Testing	Accuracy (%)	57.29	63.31	64.08	67.04	**70.34**
	R^2	0.43	0.56	0.56	0.59	**0.61**
	MAE	2.81	2.46	2.46	2.35	**2.25**
	RMSE	3.98	3.52	3.49	3.40	**3.31**
	$<$ 1 min (%)	26.04	30.90	30.34	31.75	**33.80**
	$<$ 3 min (%)	66.00	71.91	70.93	72.92	**75.40**
	$<$ 5 min (%)	84.31	87.93	88.03	**89.85**	**89.85**

Table 9: Prediction models performance comparisons for Zurich Airport. Bold numbers indicate the best performance in terms of the corresponding metrics.

Model		MLP	LR	PR	GBRT	RF
Training	Accuracy (%)	8.25	16.31	27.29	43.29	**47.24**
	R^2	0.57	0.60	0.12	**0.67**	**0.67**
	MAE	2.29	2.07	2.05	1.84	**1.76**
	RMSE	3.49	3.34	4.81	**3.06**	**3.06**
	$<$ 1 min (%)	34.35	41.73	43.99	48.17	**52.15**
	$<$ 3 min (%)	75.79	78.84	79.74	81.58	**83.03**
	$<$ 5 min (%)	90.00	90.86	91.34	92.64	**92.74**
Testing	Accuracy (%)	-22.15	21.88	29.13	44.65	**46.23**
	R^2	0.09	0.61	0.62	**0.67**	**0.67**
	MAE	4.16	2.12	2.06	1.87	**1.78**
	RMSE	5.24	3.44	3.38	3.15	**3.10**
	$<$ 1 min (%)	8.19	42.55	43.94	47.58	**51.74**
	$<$ 3 min (%)	36.93	78.33	79.30	81.73	**82.85**
	$<$ 5 min (%)	73.71	90.21	90.57	**92.58**	92.53

Table 10: Prediction models performance comparisons for Hong Kong Airport. Bold numbers indicate the best performance in terms of the corresponding metrics.

Model		MLP	LR	PR	GBRT	RF
Training	Accuracy (%)	47.65	75.40	79.02	79.63	**80.42**
	R^2	0.58	0.82	0.82	**0.85**	0.84
	MAE	3.79	2.19	2.02	**1.96**	**1.96**
	RMSE	4.99	3.30	3.27	**3.03**	3.10
	$<$ 1 min (%)	18.22	39.54	44.36	46.10	**47.87**
	$<$ 3 min (%)	50.60	75.24	77.61	**78.29**	77.90
	$<$ 5 min (%)	71.67	89.01	90.69	**90.73**	90.63
Testing	Accuracy (%)	45.11	76.43	79.55	80.31	**80.80**
	R^2	0.54	0.81	0.83	0.83	**0.84**
	MAE	4.16	2.28	2.08	2.03	**1.96**
	RMSE	5.32	3.45	3.27	3.23	**3.10**
	$<$ 1 min (%)	10.62	38.50	44.18	46.55	**48.09**
	$<$ 3 min (%)	40.52	73.74	76.68	77.36	**77.79**
	$<$ 5 min (%)	73.29	88.40	89.90	90.19	**90.53**

considerably more complex picture so unsurprising that it is harder to model. The taxi times at Zurich are on average a little shorter, so the metric accuracy can drop while the metrics related to the 1, 3 and 5 minute thresholds remain much the same.

We also observe that all prediction models have consistent performance across the three airports, e.g., the variation of prediction accuracy within 5 minutes for different airports is less than 5% for each model. In conclusion, RF has the best prediction performance among the introduced five models, and we apply RF as the prediction model for the following experiments using different feature groups.

6. Results: Feature importance

In this section we investigate the prediction performance with respect to the different features. The goal is to determine which features are most important for constructing accurate models of taxi time. We begin by using different feature groups. We then investigate and discuss the taxi time distributions and importance for individual features.

6.1. Performance of different feature groups

We introduced a set of new features in Section 3 that could improve the taxi time prediction performance. To investigate the impact of different features, we test the RF model on Groups A, B, C, D and E, respectively. Note Group E contains the historical congestion features which cannot be obtained in practice, and here we use the results on Group E to evaluate and compare to other feature groups.

The prediction performance of RF using different feature groups are reported in Tables 11 to 13, where the best performance value for each metric among Groups A, B,

C, D is in bold. The prediction results of the training and testing sets using different feature groups are consistent for each airport. Besides, as shown in Tables 18 to 20 in Appendix B, the standard deviations of the prediction results have very small variations across feature groups. There is no apparent trend that the standard deviations increase with adding more features, indicating the RF model with random parameters has stable performance when considering more features.

Table 11: Prediction performance comparisons on different feature groups for Manchester Airport. Bold numbers indicate the best performance in terms of the corresponding metrics excluding Group E.

Feature Group		A	B	C	D	E
Training	Accuracy (%)	70.07	**70.29**	70.13	70.09	73.32
	R^2	0.57	0.59	**0.61**	**0.61**	0.70
	MAE	2.34	2.29	**2.27**	2.28	2.04
	RMSE	3.48	3.40	**3.34**	**3.34**	2.92
	$<$ 1 min (%)	34.32	**34.39**	34.13	34.28	36.84
	$<$ 3 min (%)	73.47	74.30	74.54	**74.56**	77.88
	$<$ 5 min (%)	89.28	89.72	**89.79**	89.76	92.51
Testing	Accuracy (%)	70.67	**70.89**	70.66	70.47	73.69
	R^2	0.57	0.59	**0.60**	**0.60**	0.69
	MAE	2.35	2.30	2.30	**2.29**	2.05
	RMSE	3.50	3.43	3.39	**3.37**	2.96
	$<$ 1 min (%)	34.06	**34.30**	34.25	34.13	37.13
	$<$ 3 min (%)	73.75	74.04	73.93	**74.60**	77.66
	$<$ 5 min (%)	89.19	89.67	89.50	**89.80**	92.33

It is evident Groups C and D with additional aircraft speed and weather condition features have overall better performance comparing to Groups A and B. Moreover, the historical congestion features in Group E significantly improved the prediction results in terms of all performance metrics, indicating the congestion conditions indeed have huge impact on the taxi time. A possible explanation for this is simply that, even with constant traffic levels, the longer the taxi time, the more aircraft will stop or start moving.

When we closely look at the prediction results on Groups B and C, similar performances can be observed. Compared to Group B, the prediction accuracy on Group C increases slightly for HKG, while it has almost the same results (or even slightly worse in some metrics) for MAN and ZRH. Only when aircraft speed and weather condition features are considered together in Group D, does the model have better performance comparing to Groups A and B.

It demonstrates that the introduced new features, at least some of them, contribute to more accurate taxi time prediction. However, which and to what extent the features are important to the taxi time remains unclear. Therefore, we further analyse and identify the feature importance in the next section.

Table 12: Prediction performance comparisons on different feature groups for Zurich Airport. Bold numbers indicate the best performance in terms of the corresponding metrics excluding Group E.

Feature Group		A	B	C	D	E
Training	Accuracy (%)	48.92	**48.95**	47.37	47.24	64.16
	R^2	0.63	0.66	0.66	**0.67**	0.84
	MAE	1.93	1.78	1.79	**1.76**	1.33
	RMSE	3.24	3.11	3.13	**3.06**	2.10
	$<$ 1 min (%)	49.14	**52.28**	52.00	52.15	57.13
	$<$ 3 min (%)	79.79	82.58	82.70	**83.03**	89.30
	$<$ 5 min (%)	91.54	92.51	92.44	**92.74**	96.63
Testing	Accuracy (%)	**48.67**	48.01	46.14	46.23	64.37
	R^2	0.64	0.66	0.65	**0.67**	0.85
	MAE	1.94	1.81	1.82	**1.78**	1.33
	RMSE	3.27	3.16	3.18	**3.10**	2.15
	$<$ 1 min (%)	48.74	**52.26**	51.66	51.74	58.01
	$<$ 3 min (%)	79.42	82.17	82.17	**82.85**	89.50
	$<$ 5 min (%)	91.41	92.29	92.27	**92.53**	96.69

Table 13: Prediction performance comparisons on different feature groups for Hong Kong Airport. Bold numbers indicate the best performance in terms of the corresponding metrics excluding Group E.

Feature Group		A	B	C	D	E
Training	Accuracy (%)	76.36	79.53	**80.44**	80.42	85.86
	R^2	0.72	0.82	**0.84**	**0.84**	0.93
	MAE	2.61	2.07	**1.96**	**1.96**	1.34
	RMSE	4.06	3.27	**3.10**	**3.10**	1.99
	$<$ 1 min (%)	43.95	47.15	**47.94**	47.87	55.72
	$<$ 3 min (%)	68.13	76.30	77.80	**77.90**	88.90
	$<$ 5 min (%)	81.83	89.26	90.57	**90.63**	97.31
Testing	Accuracy (%)	76.80	79.93	80.79	**80.80**	86.52
	R^2	0.73	0.82	**0.84**	**0.84**	0.93
	MAE	2.60	2.07	1.97	**1.96**	1.36
	RMSE	4.06	3.28	3.11	**3.10**	2.07
	$<$ 1 min (%)	44.21	47.41	**48.17**	48.09	55.49
	$<$ 3 min (%)	68.49	76.03	77.70	**77.79**	88.32
	$<$ 5 min (%)	82.10	89.01	90.42	**90.53**	97.08

6.2. Feature importance identification

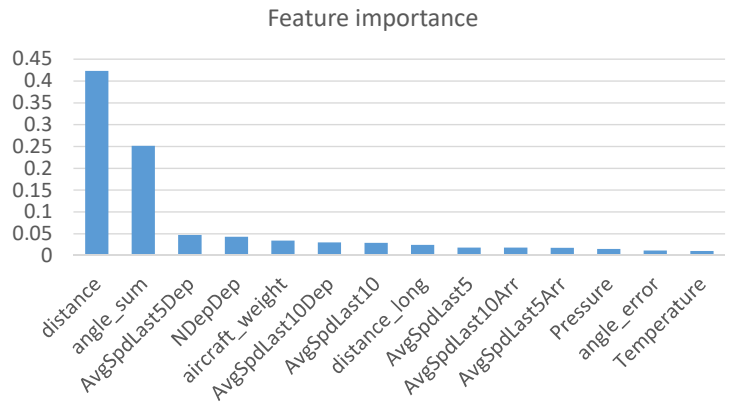
As discussed in Section 4.3, the feature importance value extracted from the prediction model is necessary for the identification procedure. The RF model is a combination of tree predictors which can provide useful internal estimates, e.g. error, correlation and feature importance [45]. Specifically, the importance of a feature in RF is computed as the normalized total reduction of the criterion brought by that feature. It is also known as the Gini importance. Detailed explanations of Gini importance and feature importance calculation are referred to [48]. We therefore can directly use the feature importance metric from RF, enabling the proposed feature importance identification procedure.

The feature importance rankings from RF on Group D (29 features) are illustrated in Figure 3, where the features with importance value less than 0.01 are omitted. Note the sum of importance value for all features is one. To better understand and evaluate the feature importance values, some representatives of taxi time distributions with different features are illustrated in Figures 4 to 6.

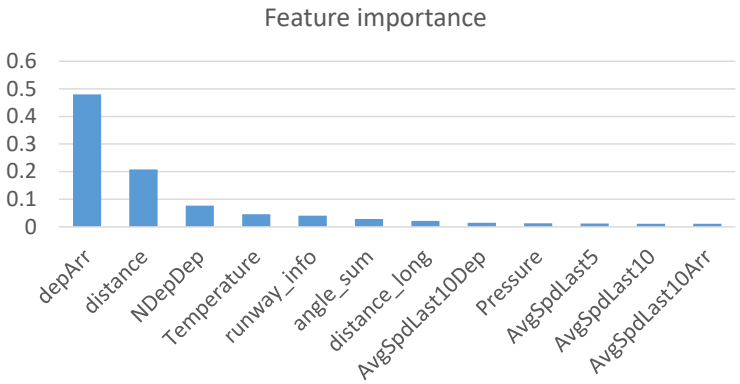
Clearly *depArr*, denoting whether the aircraft is departure or arrival, is always the most important. In particular, the importance value of *depArr* is up to 0.7 for HKG, indicating whether the flight is departure or arrival has dominant influence on the taxi time. The same conclusions can be made when we look at the taxi time distributions with *depArr*, in which the arrival aircraft indeed have an average shorter taxi time than that of departure flight.

Following *depArr*, *distance*, *NDepDep* (the number of other aircraft on the way to runway when the current aircraft starts departing), *angle_sum*, *distance_long*, and average speed features also have large importance values across the three airports. In line with the distribution plots with these features, the taxi time indeed varies against corresponding feature values. For example, as the values of *distance* and *NDepDep* increase, the aircraft mean taxi time also has an increasing trend. Meanwhile, we observe that the importance values of other congestion related features, e.g., *NDepArr*, *NArrArr* and *NArrDep*, are less than 0.01 for three airports. These findings demonstrate that only the congestion information, in particular the departure queuing length, heavily impacts the departure aircraft. The above findings are in line with the queuing model [15] applied to taxi time prediction. Similarly, the average speed features, especially the ones for departure flights *AvgSpdLast5Dep* and *AvgSpdLast10Dep*, have evident impacts on the taxi time. This observation also reveals that the departure aircraft are easily influenced by the airport operational conditions.

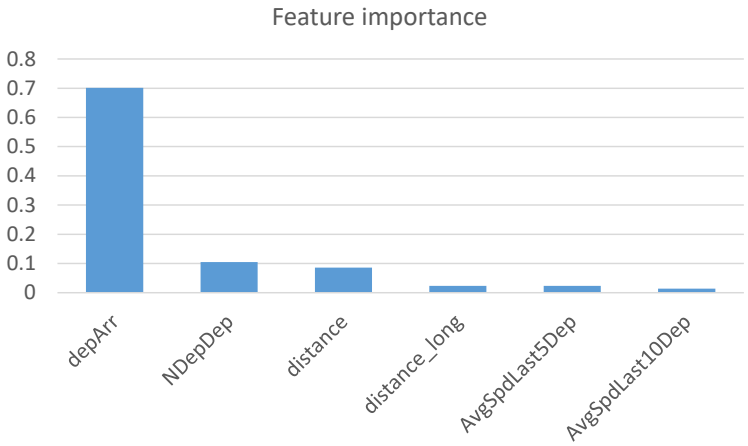
The *budget_airline* and *runway_info* features are considered in the prediction model following discussions with airport practitioners. However, *budget_airline* only has a small importance value. The distributions of taxi time with *budget_airline* in Figure 4(h) confirm this finding; there is no evident differences of the taxi time distributions between the budget and non-budget airlines. It is interesting to see that *runway_info* only has a large importance value for ZRH, which is consistent with its taxi time distributions in



(a) MAN

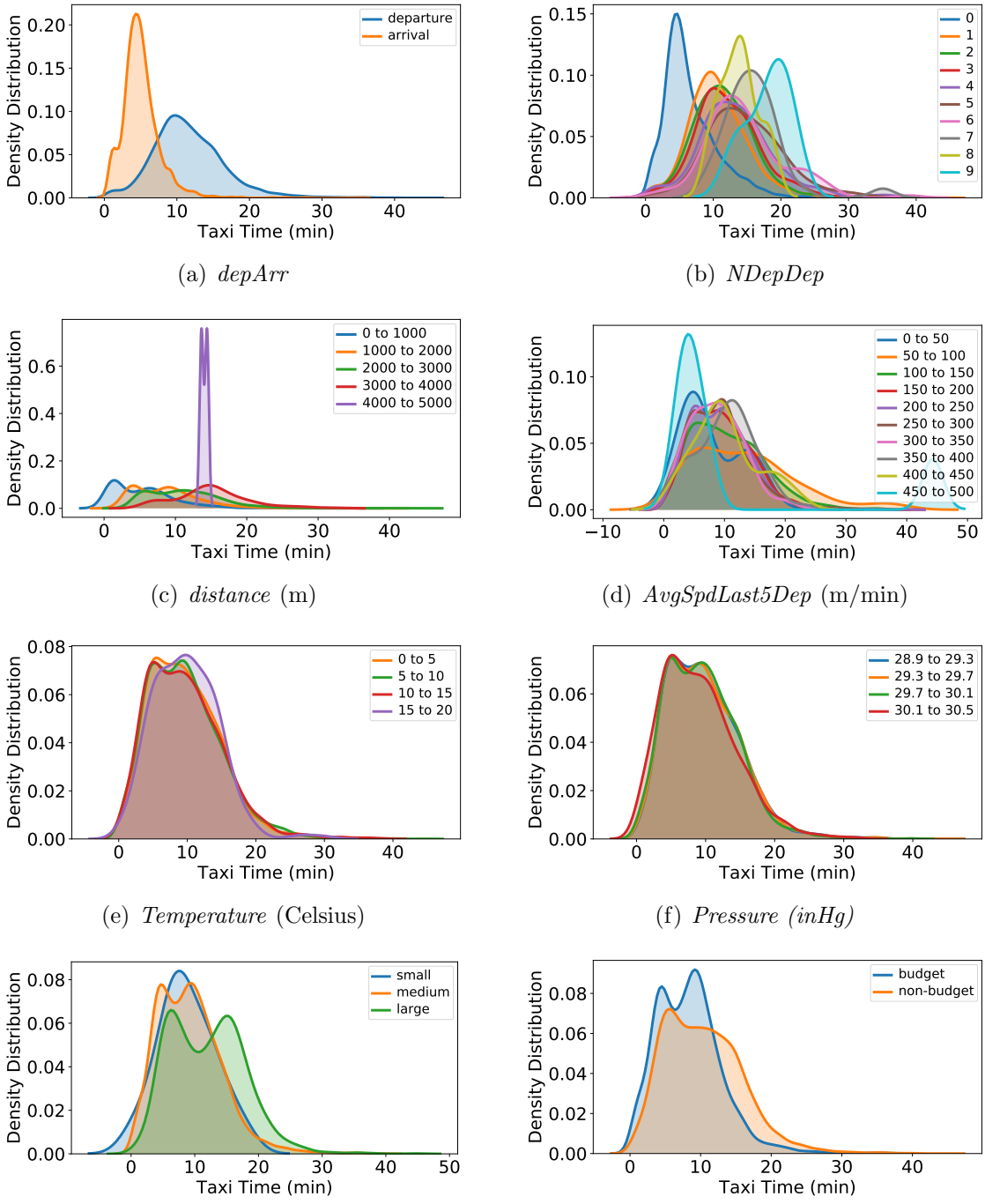


(b) ZRH



(c) HKG

Figure 3: Feature importance ranking for the three airports.



(a) *depArr*

(b) *NDepDep*

(c) *distance* (m)

(d) *AvgSpdLast5Dep* (m/min)

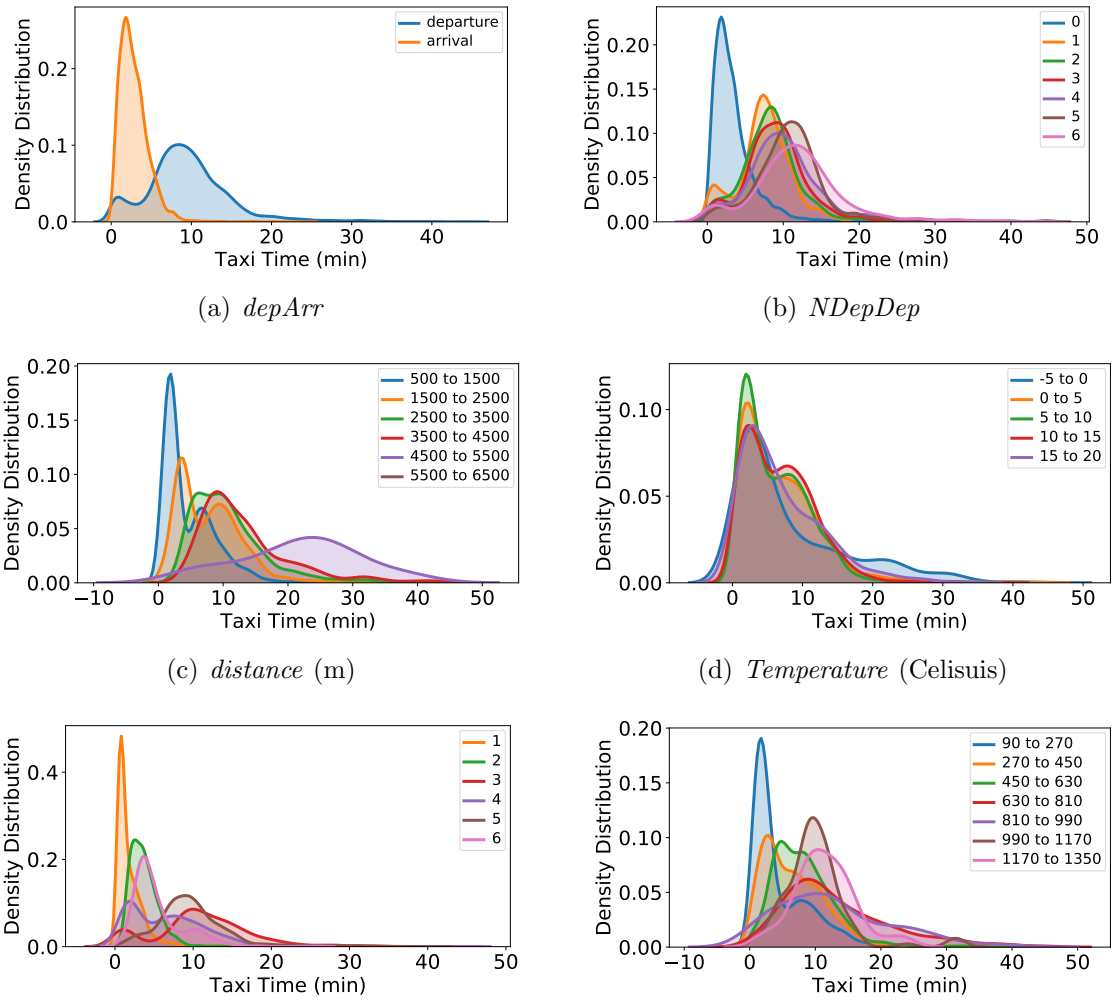
(e) *Temperature* (Celsius)

(f) *Pressure (inHg)*

(g) *aircraft_weight*

(h) *budget_airline*

Figure 4: Taxi time distributions with different features for Manchester Airport.

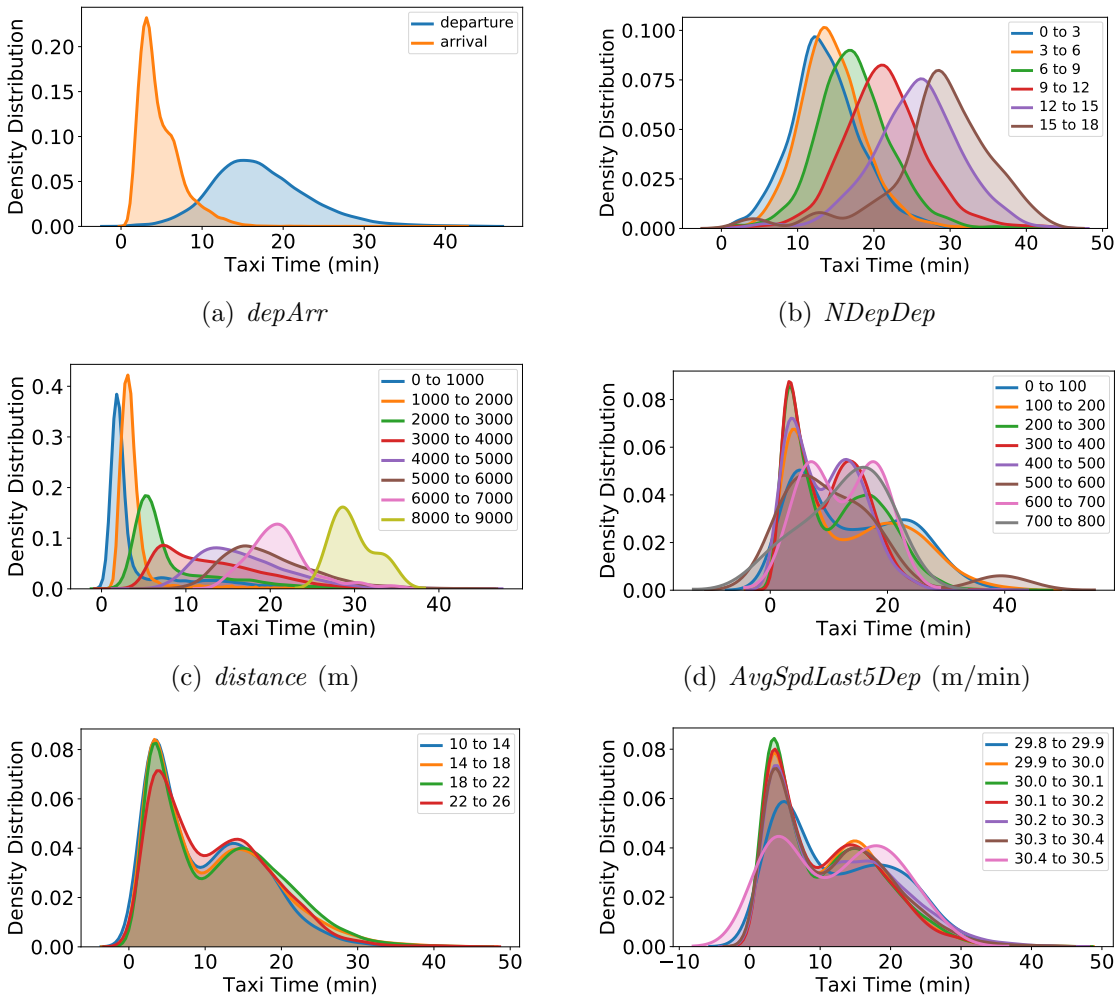


(a) *depArr* (b) *NDepDep*

(c) *distance* (m) (d) *Temperature* (Celisuis)

(e) *runway_info* (f) *angle_sum* (degree)

Figure 5: Taxi time distributions with different features for Zurich Airport.



(a) *depArr*

(b) *NDepDep*

(c) *distance* (m)

(d) *AvgSpdLast5Dep* (m/min)

(e) *Temperature* (Celisuis)

(f) *Pressure* (inHg)

Figure 6: Taxi time distributions with different features for Hong Kong Airport.

Figure 5(e). This could be due to the crossing runways at ZRH. As seen in Figure 1(b), ZRH has three runways and two of them cross each other. Figure 5 shows that the taxi times are very different for the modes 32/34 and 14/16. The crossing has as well a small influence for cases where runway 28 is in use as arrival runway but most flights try to leave the runway before the intersection. Therefore, the runway operational conditions, namely *runway_info*, probably have a huge impact on the taxi time for ZRH. The importance value of *aircraft_weight* ranks 5th for MAN. Its importance can be verified with the distributions in Figure 5(g), where the heavy aircraft typically have longer taxi time. This can be further attributed to the fact that MAN has the largest gradient variation of its taxiways (MAN has a glide slope of 3 degrees equal to varying 52.3 metres per kilometer, while ZRH and HKG typically have flat taxiways). Heavy aircraft in general require more time to accelerate in such a case.

As for the weather condition features, although 11 features are introduced in the prediction model, only *Temperature* has relatively large importance value for MAN and ZRH, and *Pressure* has importance value greater than 0.01 for ZRH. These results are also consistent with taxi time distribution plots. For instance, as indicated in Figure 5(d) for ZRH, the temperature range above 0 degree is prone to a higher density for shorter taxi time. In contrast, when the temperature is below 0 degree, which normally indicates not good weather, longer taxi time may be required. Notice weather conditions seem not impacting taxi time prediction for HKG. This is probably due to its dominant *depArr* feature with a 0.7 importance value; the importance values of other features are more difficult to be greater than 0.01 for HKG comparing to that for MAN and ZRH.

Next, the feature importance identification procedure with backward elimination manner is applied. The feature importance value less than 0.01 and prediction accuracy within 5 minutes are used as termination conditions, respectively. Note the importance value for the same feature dynamically changes during the backward elimination process.

The identification results with > 0.01 feature importance and $< 1\%$ prediction accuracy are shown in Tables 14 and 15 respectively, where features identified as important across the three airports are marked in bold, and the ones identified at two airports are in underline. As highlighted in Table 14, *depArr*, *distance*, *NDepDep*, *angle_sum*, *distance_long* and departure related speed features have high importance values across the three airports. These features can be identified as generally important features. Besides, several features are important only for specific airports, e.g., *aircraft_weight*, *Temperature* and *Pressure* for MAN; *Temperature*, *runway_info*, *Pressure* for ZRH. We regard these features as airport specific ones.

As the termination condition switches to the prediction performance, the number of selected features is reduced as shown in Table 15. Notice that all features in Table 15 are presented in Table 14. Moreover, *depArr*, *distance* and *NDepDep* are identified as the most three important features across the three airports.

As for the accuracy provided by the selected features, the prediction results are pre-

Table 14: Feature importance identification with importance value (Group F). Bold features are common to all 3 airports, underlined common to any 2.

No.	MAN	ZRH	HGK
1	***depArr***	***depArr***	***depArr***
2	**distance**	***distance***	***NDepDep***
3	***angle_sum***	***NDepDep***	***distance***
4	***AvgSpdLast5Dep***	*Temperature*	***AvgSpdLast5Dep***
5	***NDepDep***	*runway_info*	***distance_long***
6	*aircraft_weight*	***angle_sum***	***AvgSpdLast10Dep***
7	***AvgSpdLast10Dep***	***distance_long***	***angle_sum***
8	*AvgSpdLast10*	***AvgSpdLast10Dep***	
9	***distance_long***	*Pressure*	
10	*AvgSpdLast5*	*AvgSpdLast10Arr*	
11	*AvgSpdLast10Arr*	*AvgSpdLast5*	
12	*AvgSpdLast5Arr*	*AvgSpdLast10*	
13	*Pressure*	*AvgSpdLast5Arr*	
14	*Temperature*	***AvgSpdLast5Dep***	
15	*angle_error*		

Table 15: Feature importance identification with prediction performance (Group G). Bold features are common to all 3 airports, underlined features common to any 2 airports.

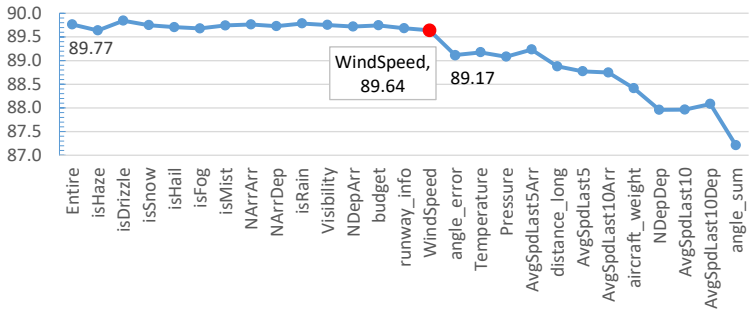
No.	MAN	ZRH	HGK
1	***depArr***	***depArr***	***depArr***
2	***distance***	***distance***	***NDepDep***
3	*angle_sum*	***NDepDep***	***distance***
4	*AvgSpdLast5Dep*	*angle_sum*	*distance_long*
5	*AvgSpdLast10Dep*	*Temperature*	*AvgSpdLast5Dep*
6	*AvgSpdLast10*		*AvgSpdLast10Dep*
7	***NDepDep***		
8	*aircraft_weight*		

Table 16: Prediction performance of RF using selected features.

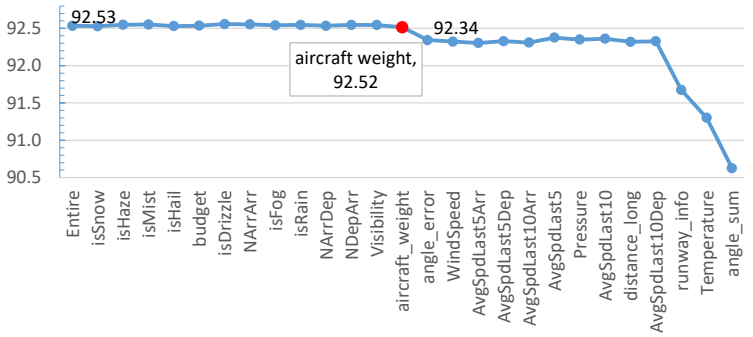
Airport		MAN			ZRH			HKG		
Feature Group		D	F	G	D	F	G	D	F	G
Training	Accuracy	70.09	69.90	69.06	47.24	46.57	40.34	80.42	80.14	79.92
	R^2	0.61	0.60	0.59	0.67	0.67	0.63	0.84	0.84	0.83
	MAE	2.28	2.28	2.33	1.76	1.77	1.87	1.96	1.99	2.01
	RMSE	3.34	3.34	3.41	3.06	3.08	3.24	3.10	3.14	3.16
	$<$ 1 min	34.28	34.18	33.82	52.15	52.21	51.04	47.87	47.60	47.29
	$<$ 3 min	74.56	74.54	73.57	83.03	82.83	81.22	77.90	77.49	77.28
	$<$ 5 min	89.76	89.72	88.96	92.74	92.59	91.55	90.63	90.11	89.92
Testing	Accuracy	70.47	70.31	69.47	46.23	45.54	39.01	80.80	80.48	80.26
	R^2	0.60	0.60	0.58	0.67	0.67	0.63	0.84	0.84	0.83
	MAE	2.29	2.30	2.35	1.78	1.79	1.90	1.96	2.00	2.01
	RMSE	3.37	3.37	3.44	3.10	3.12	3.29	3.10	3.15	3.17
	$<$ 1 min	34.13	34.22	33.93	51.74	51.81	50.61	48.09	47.65	47.40
	$<$ 3 min	74.60	74.18	73.58	82.85	82.65	80.86	77.79	77.35	77.18
	$<$ 5 min	89.80	89.74	88.85	92.53	92.34	91.62	90.53	89.95	89.76

sented in Table 16. Clearly the performance with more selected features whose importance values are greater than 0.01 (denoted as Feature Group F) is better than that with fewer features identified with $< 1\%$ accuracy threshold (denoted as Feature Group G). Moreover, it should be noted that the performance of the RF model using Group F are very close to that using Group D, and even slightly better for certain performance metrics, e.g., prediction accuracy within 1 minute for MAN and ZRH.

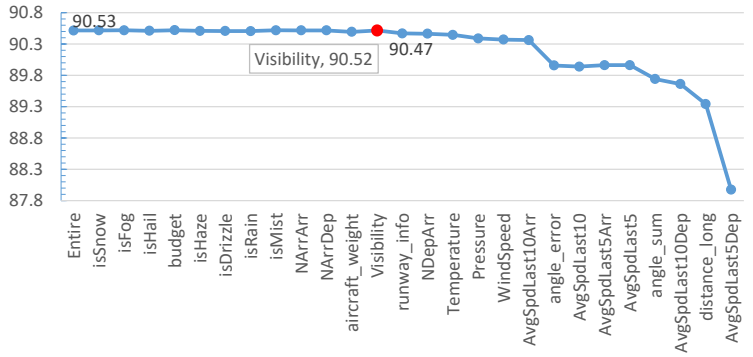
To further investigate which subsets of the available features can provide high prediction accuracy, we remove the termination conditions in the feature importance identification procedure, iteratively reducing features until only containing the most important three features, i.e., *depArr*, *distance* and *NDepDep*. The results are presented in Figure 7, where the red dot denotes the corresponding prediction accuracy within 5 minutes with the finally selected features subset. The prediction accuracy remains quite stable during the first-stage feature elimination process; therefore we aim to select fewer features which can maintain the same accuracy level compared to Group D. We observe that for MAN, the first dramatic decrease is from 89.64% to 89.17% when removing feature *angle_error*. The feature group identified to provide high prediction accuracy happens to be Group F (15 features) for MAN. When we iteratively reduce the features for ZRH, the prediction accuracy has very limited variations, until its value decreases to 92.34% when feature *angle_error* is removed. Therefore, we need to select the subset including Group F with extra features *angle_error* and *WindSpeed* (16 features in total) for ZRH, so that the high prediction performance can be realised. Similarly, the critical point is found when removing feature Visibility for HKG, and its corresponding subset consists of Group F and *runway_info*, *NDepArr*, *Temperature*, *Pressure*, *WindSpeed*, *AvgSpdLast10Arr*, *an-*



(a) MAN



(b) ZRH



(c) HKG

Figure 7: Prediction performance with iteratively reducing features until only containing the most important three ones. The Red dot denotes corresponding prediction accuracy within 5 minutes with the finally selected features subset.

gle_error, *AvgSpdLast10*, *AvgSpdLast5Arr* and *AvgSpdLast5* (17 features).

In conclusion, we have demonstrated that high prediction accuracy of taxi time can be achieved with a narrow subset of the features consisting of generally important ones across all airports and specific ones for target airports. These findings have provided

quantitative insights for aircraft taxi time prediction.

7. Conclusions

In this article, we have investigated aircraft taxi time prediction models and introduced a set of features that may affect the taxi time, among which the runway operational mode, budget/non-budget airlines and aircraft speed features are considered for the first time. We aim to select a prediction model that works best for our target airports, i.e. Manchester Airport, Zurich Airport and Hong Kong Airport. Five regression models for taxi time prediction are compared and extensive experiments demonstrate that the Random Forest model outperforms other prediction models for the three airports.

The major focus of our study was a feature importance identification procedure with backward elimination, where the feature importance value is extracted from the Random Forest model. Quantitative analysis shows that the *depArr* (aircraft departure or arrival), *distance* (sum of taxi distance) and *NDepDep* (number of other aircraft on the way to runway when current aircraft pushes back) features are identified as the most important features across the three airports, followed by *angle_sum* (sum of turning angle), *distance_long* (sum of taxi distance on straights of more than 500 metres) and departure related speed features. Moreover, the proposed feature identification method can define specifically important features for target airports, e.g., the *aircraft_weight* (aircraft weight category), *Temperature* and *Pressure* for MAN, and *runway_info* (airport operation mode in line with runway utilisation), *Temperature* and *Pressure* for ZRH. An important result is that high accuracy can be achieved with a narrow subset of the features consisting of generally important ones across all airports and specific ones for target airports.

Future research for aircraft taxi time prediction can be oriented in several directions. 1) The slope of the airport could have a huge impact on taxi time for specific airports, e.g. Manchester Airport. How to model the slope factor and define corresponding feature value remains unexplored. Currently, our hypothesis is that *aircraft_weight* somehow acts as a good proxy for such information. However, a more dedicated feature that can capture the slope information may further improve the prediction accuracy for certain airports. 2) The taxi time prediction underpins practical aircraft routing and scheduling system. To address online decision making requirements, current taxi time prediction model should be revised to support adaptive prediction with respect to dynamically varying features information. 3) Improving the feature importance identification procedure is also an avenue for future research. Alternatively we could apply model compression techniques [49], in which the generated single decision tree with high interpretability is promising to provide better feature importance metrics, while maintaining high prediction accuracy. 4) The range of machine learning approaches is constantly growing, and it would also be interesting to test alternative models, such as Lasso Regression, Elastic Net Regression, Regression Trees, Support Vector Regression, k-Nearest Neighbors and

Extra Trees Regression. 5) It would, of course, also be interesting to further extend the features covered. We could explicitly include whether or not an aircraft crosses a runway during its manoeuvre, or uses single or bidirectional taxiways. Wind direction and humidity could be useful additional indicators. The present study only covered Winter and Spring months at each airport: expanding the study to an entire year would also provide further insight into the impact of changing weather. There is certainly room for improving the accuracy of the models over our results and these directions may offer a means to achieving that.

Acknowledgements

This work was funded by the UK Engineering and Physical Sciences Research Council [grants EP/N029577/2, EP/N029496/2, and EP/N029356/1]. Grateful thanks are also extended to MSc students Nikolaos Prousalis and Anoosid Kiatkamolvong for their assistance in the experimental studies, and to the anonymous reviewers for their helpful comments and feedback.

Appendix A Abbreviations

Appendix B Standard deviation of RF for three airports

References

[1] International Air Transport Association, "International air transport association annual review 2019," https://www.iata.org/en/publications/annual-review/, accessed February, 2020.

[2] ——, "20 year passenger forecast," https://www.iata.org/en/publications/store/20-year-passenger-forecast/, accessed February, 2020.

[3] European Commission, "Airport capacity and quality," https://ec.europa.eu/transport/modes/air/airports/airport_capacity_and_quality_en, accessed February, 2020.

[4] ——, "Roadmap to a single european transport area," https://www.eea.europa.eu/policy-documents/roadmap-to-a-single-european, accessed February, 2020.

[5] J. Chen, M. Weiszer, P. Stewart, and M. Shabani, "Toward a more realistic, cost-effective, and greener ground movement through active routing—part I: Optimal speed profile generation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1196–1209, 2015.

Table 17: Abbreviations list.

ARN	Stockholm-Arlanda Airport
ASDE-X	Airport Surface Detection Equipment, Model X
ASPM	Aviation System Perfor-mance Metrics
ASQP	Airline Service Quality Performance
BCN	Bacrelona-EI Airport
BOS	Boston Logan International Airport
CDG	Charles de Gaulle Airport
CLT	Charlotte Douglas International Airpor
DFW	Dallas/Fort Worth Airport
EWR	Newark Liberty International Airport
FR24	FlightRadar24
FRBSs	Fuzzy Rule-Based Systems
GBRT	Gradient Boosted Regression Trees
HKG	Hong Kong International Airport
JFK	John F. Kennedy International Airport
LR	Linear Regression
MAE	Mean Absolute Error
MAN	Manchester Airport
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
NN	Neural Networks
PEK	Beijing International Airport
PR	Polynomial Regression
PRAS	Preferential Runway Assignment System
PVG	Shanghai Pudong International Airport
RF	Random Forest
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RT	Regression Tree
SARDA	Spot and Run-way Departure Advisor
SEA	Seattle International Airport
SVR	Support Vector Regression
SWAP	Severe Weather Avoidance Programs
TPA	Tampa International Airport
ZRH	Zurich Airport

Table 18: Standard deviation of RF for Manchester Airport.

Feature Group		A	B	C	D
Training	Accuracy	0.315	0.332	0.258	0.300
	R^2	0.009	0.009	0.010	0.010
	MAE	0.023	0.026	0.025	0.025
	RMSE	0.039	0.037	0.038	0.035
	< 1 min	0.662	0.579	0.503	0.506
	< 3 min	0.243	0.396	0.413	0.428
	< 5 min	0.441	0.244	0.239	0.218
Testing	Accuracy	0.413	0.401	0.319	0.259
	R^2	0.015	0.016	0.019	0.019
	MAE	0.039	0.046	0.043	0.044
	RMSE	0.071	0.072	0.073	0.076
	< 1 min	0.746	0.846	0.758	0.674
	< 3 min	0.596	0.752	0.859	0.939
	< 5 min	0.401	0.446	0.436	0.416

Table 19: Standard deviation of RF for Zurich Airport.

Feature Group		A	B	C	D
Training	Accuracy	0.780	0.698	0.768	0.681
	R^2	0.005	0.008	0.005	0.006
	MAE	0.014	0.017	0.015	0.014
	RMSE	0.037	0.049	0.046	0.042
	< 1 min	0.373	0.434	0.341	0.289
	< 3 min	0.373	0.236	0.263	0.225
	< 5 min	0.192	0.181	0.147	0.181
Testing	Accuracy	1.832	1.856	2.070	1.835
	R^2	0.009	0.014	0.014	0.017
	MAE	0.028	0.021	0.023	0.026
	RMSE	0.092	0.095	0.103	0.110
	< 1 min	0.915	0.606	0.684	0.687
	< 3 min	0.477	0.609	0.350	0.416
	< 5 min	0.398	0.427	0.392	0.422

Table 20: Standard deviation of RF for Hong Kong Airport.

Feature Group		A	B	C	D
Training	Accuracy	0.148	0.241	0.175	0.174
	R^2	0.005	0.000	0.000	0.000
	MAE	0.015	0.008	0.011	0.009
	RMSE	0.021	0.014	0.018	0.017
	< 1 min	0.180	0.156	0.227	0.167
	< 3 min	0.249	0.245	0.155	0.153
	< 5 min	0.186	0.130	0.130	0.092
Testing	Accuracy	0.330	0.579	0.425	0.431
	R^2	0.005	0.003	0.003	0.003
	MAE	0.021	0.017	0.020	0.019
	RMSE	0.038	0.037	0.041	0.039
	< 1 min	0.524	0.345	0.399	0.326
	< 3 min	0.270	0.360	0.378	0.304
	< 5 min	0.330	0.189	0.193	0.201

[6] J. Chen, M. Weiszer, G. Locatelli, S. Ravizza, J. A. Atkin, P. Stewart, and E. K. Burke, "Toward a More Realistic, Cost-Effective, and Greener Ground Movement Through Active Routing: A Multiobjective Shortest Path Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3524–3540, Dec. 2016.

[7] International Civil Aviation Organization, "Advanced surface movement guidance and control systems manual," https://www.icao.int/Meetings/anconf12/Document%20Archive/9830_cons_en%5B1%5D.pdf, accessed February, 2020.

[8] Joint Planning and Development Office, "Concept of operations for the next generation air transport system," https://www.hsdl.org/?abstract&did=747519, accessed February, 2020.

[9] J. A. Atkin, E. K. Burke, J. S. Greenwood, and D. Reeson, "Hybrid metaheuristics to aid runway scheduling at london heathrow airport," *Transportation Science*, vol. 41, no. 1, pp. 90–106, 2007.

[10] M. Weiszer, J. Chen, and G. Locatelli, "An integrated optimisation approach to airport ground operations to foster sustainability in the aviation sector," *Applied Energy*, vol. 157, pp. 567–582, 2015.

[11] S. Ravizza, J. A. D. Atkin, M. H. Maathuis, and E. K. Burke, "A combined statistical approach and ground movement model for improving taxi time estimations at airports," *Journal of the Operational Research*

Society, vol. 64, no. 9, pp. 1347–1360, Sep. 2013. [Online]. Available: https://www.tandfonline.com/doi/full/10.1057/jors.2012.123

[12] N. Pujet, B. Delcaire, and E. Feron, "Input-Output Modeling and Control of the Departure Process of Busy Airports," *Air Traffic Control Quarterly*, vol. 8, no. 1, pp. 1–32, Jan. 2000. [Online]. Available: https://arc.aiaa.org/doi/10.2514/atcq.8.1.1

[13] J. Chen, S. Ravizza, J. A. D. Atkin, and P. Stewart, "On the Utilisation of Fuzzy Rule-Based Systems for Taxi Time Estimations at Airports," in *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, 2011, pp. 134–145. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2011/3273/

[14] P. Balakrishna, R. Ganesan, L. Sherry, and B. S. Levy, "Estimating taxi-out times with a reinforcement learning algorithm," in *2008 IEEE/AIAA 27th Digital Avionics Systems Conference*, Oct. 2008, pp. 3.D.3–1–3.D.3–12.

[15] H. Idris, J.-P. Clarke, R. Bhuva, and L. Kang, "Queuing Model for Taxi-Out Time Estimation," *Air Traffic Control Quarterly*, vol. 10, no. 1, pp. 1–22, Jan. 2002. [Online]. Available: https://arc.aiaa.org/doi/10.2514/atcq.10.1.1

[16] P. Balakrishna, R. Ganesan, and L. Sherry, "Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 6, pp. 950–962, Dec. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X1000029X

[17] T. Diana, "Can machines learn how to forecast taxi-out time? A comparison of predictive models applied to the case of Seattle/Tacoma International Airport," *Transportation Research Part E: Logistics and Transportation Review*, vol. 119, pp. 149–164, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S136655451830543X

[18] A. Srivastava, "Improving departure taxi time predictions using ASDE-X surveillance data," in *2011 IEEE/AIAA 30th Digital Avionics Systems Conference*, Oct. 2011, pp. 2B5–1–2B5–14.

[19] H. Lee, W. Malik, and Y. C. Jung, "Taxi-Out Time Prediction for Departures at Charlotte Airport Using Machine Learning Techniques," in *16th Aviation Technology, Integration, and Operations Conference*. Washington, D.C.: AIAA, Jun. 2016, pp. 1–11. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2016-3910

[20] H. Lee, W. Malik, B. Zhang, B. Nagarajan, and Y. C. Jung, "Taxi Time Prediction at Charlotte Airport Using Fast-Time Simulation and Machine Learning

Techniques," in *15th AIAA Aviation Technology, Integration, and Operations Conference*. Dallas, TX: American Institute of Aeronautics and Astronautics, Jun. 2015. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2015-2272

[21] J. Chen, M. Weiszer, E. Zareian, M. Mahfouf, and O. Obajemu, "Multi-objective fuzzy rule-based prediction and uncertainty quantification of aircraft taxi time," in *20th International Conference on Intelligent Transportation Systems (ITSC)*. Yokohama: IEEE, Oct. 2017, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/8317826/

[22] I. Simaiakis and N. Pyrgiotis, "An Analytical Queuing Model of Airport Departure Processes for Taxi Out Time Prediction," in *10th Aviation Technology, Integration, and Operations (ATIO) Conference*. Fort Worth, Texas: AIAA, Sep. 2010. [Online]. Available: http://arc.aiaa.org/doi/10.2514/6.2010-9148

[23] T. Diana, "An application of survival and frailty analysis to the study of taxi-out time: A case of New York Kennedy Airport," *Journal of Air Transport Management*, vol. 26, pp. 40–43, Jan. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969699712001287

[24] S. Ravizza, J. Chen, J. A. Atkin, P. Stewart, and E. K. Burke, "Aircraft taxi time prediction: Comparisons and insights," *Applied Soft Computing*, vol. 14, pp. 397–406, Jan. 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494613003384

[25] J. Yin, Y. Hu, Y. Ma, Y. Xu, K. Han, and D. Chen, "Machine Learning Techniques for Taxi-out Time Prediction with a Macroscopic Network Topology," in *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, London, Sep. 2018, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/8569664/

[26] R. Jordan, M. A. Ishutkina, and T. G. Reynolds, "A statistical learning approach to the modeling of aircraft taxi time," in *29th Digital Avionics Systems Conference*, Oct. 2010, pp. 1.B.1–1–1.B.1–10.

[27] F. Herrema, R. Curran, H. Visser, D. Huet, and R. Lacote, "Taxi-Out Time Prediction Model at Charles de Gaulle Airport," *Journal of Aerospace Information Systems*, vol. 15, no. 3, pp. 120–130, Mar. 2018. [Online]. Available: https://arc.aiaa.org/doi/10.2514/1.I010502

[28] P. Pudil, K. Fuka, K. Beranek, and P. Dvorak, "Potential of artificial intelligence based feature selection methods in regression models," in *3rd International Conference on Computational Intelligence and Multimedia Application*. IEEE, 1999, pp. 159–163.

[29] O. Lordan, J. M. Sallan, and M. Valenzuela-Arroyo, "Forecasting of taxi times: The case of Barcelona-El Prat airport," *Journal of Air Transport Management*, vol. 56, pp. 118–122, Sep. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0969699716301570

[30] G. Lian, Y. Zhang, J. Desai, Z. Xing, and X. Luo, "Predicting Taxi-Out Time at Congested Airports with Optimization-Based Support Vector Regression Methods," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–11, 2018. [Online]. Available: https://www.hindawi.com/journals/mpe/2018/7509508/

[31] N. Mirmohammadsadeghi, S. Hotle, A. Trani, and J. Gulding, "Taxi event extraction from surveillance for surface performance evaluation," *Journal of Air Transportation*, pp. 1–8, 2019.

[32] J. M. Mendel, "Uncertain rule-based fuzzy systems," in *Introduction and New Directions*. Springer, 2017, p. 684.

[33] A. Brownlee, J. Atkin, J. Woodward, and E. Burke, "Methods and sources for underpinning airport ground movement decision support systems," University of Stirling, Tech. Rep., 19 March 2020. [Online]. Available: http://hdl.handle.net/1893/30962

[34] A. E. Brownlee, M. Weiszer, J. Chen, S. Ravizza, J. R. Woodward, and E. K. Burke, "A fuzzy approach to addressing uncertainty in airport ground movement optimisation," *Transportation Research Part C: Emerging Technologies*, vol. 92, pp. 150–175, 2018.

[35] C. Petersen, M. Mühleisen, and A. Timm-Giel, "Evaluation of the aircraft distribution in satellite spotbeams," in *Adv. in Comm. Networking*, ser. LNCS, T. Bauschert, Ed. Springer, 2013, vol. 8115, pp. 46–53.

[36] R. Turner, S. Bottone, and C. Stanek, "Online variational approximations to non-exponential family change point models: With application to radar tracking," in *Proc. of NIPS 26*, 2013, pp. 306–314.

[37] P. Ptak, J. Hartikka, M. Ritola, and T. Kauranne, "Long-distance multistatic aircraft tracking with VHF frequency doppler effect," *IEEE T Aero Elec Sys*, vol. 50, no. 3, pp. 2242–2252, 2014.

[38] A. J. Eele and J. M. Maciejowski, "Sequential Monte Carlo Optimisation for Air Traffic Management," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.693, 2015.

[39] J. N. Hallock and F. Holzäpfel, "A review of recent wake vortex research for increasing airport capacity," *Progress in Aerospace Sciences*, vol. 98, pp. 27–36, 2018.

[40] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[41] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[42] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[43] 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*, 2017, pp. 3146–3154.

[44] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Computing Surveys*, vol. 27, no. 3, pp. 326–327, 1995.

[45] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] S. Bernard, L. Heutte, and S. Adam, "Influence of hyperparameters on random forest accuracy," in *International Workshop on Multiple Classifier Systems*. Springer, 2009, pp. 171–180.

[47] Z. Bursac, C. H. Gauss, D. K. Williams, and D. W. Hosmer, "Purposeful selection of variables in logistic regression," *Source Code for Biology and Medicine*, vol. 3, no. 1, p. 17, 2008.

[48] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[49] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint arXiv:1711.09784*, 2017.