INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

No-Show Passenger Prediction for Flights

Wei-Song Chin [a,*], Choo-Yee Ting [a], Chin-Leei Cham [a]

[a] *Faculty of Computing and Informatics, Multimedia University, 63000 Cyberjaya, Selangor, Malaysia*

Corresponding author: [*]1191100961@student.mmu.edu.my

***Abstract*— In aviation, "no-show" refers to a customer who booked a reservation but failed to show up. No-shows can result in various resource wastes, such as vacant seats, leading to income loss and flight delays. As a result, no-show passengers can cause considerable problems for airlines, ultimately affecting their bottom line. Recent research has shown the use of machine learning algorithms to reduce the rate of no-shows. For example, a researcher in healthcare is using a predictive model to identify no-shows' patients to increase efficiency. Therefore, this study aimed to develop prediction models to predict passenger no-shows. In this work, we used a dataset supplied by a local airline company consisting of 1,046,486 rows and 8 columns. Additional datasets like weather data, public holiday data of different countries, aircraft details, and foot traffic data are used to carry out the dataset's feature enrichment task to complement the original dataset. As a result, feature selection has become an important stage in this research to identify and pick the most relevant and useful features from the enormous number of columns. The findings showed that the model built using Random Forest has the highest accuracy of 90.4%, while Decision Tree performed at 90.2%, Gradient Boosting at 86.5%, and Neural Networks at 67.6%. To enhance the accuracy of the models, further research efforts are essential to integrate supplementary passenger information.**

***Keywords*— No-show; aviation; prediction; machine learning; classification; feature enrichment; feature selection; neural networks.**

Manuscript received 25 Dec. 2022; revised 14 Jul. 2023; accepted 20 Aug. 2023. Date of publication 30 Nov. 2023.

I. INTRODUCTION

When a person is expected but does not show up or appear where they are supposed to, it is known as a no-show [1]–[11]. No-show behavior, a special kind of absenteeism, is frequently problematic, especially in the service sector. Although such behavior might harm the individual no-show in the long term, today, service providers, and sometimes uninvolved third parties, tend to bear the short-term consequences [3].

No-shows can have a big impact on the service sector, healthcare sector [4], [8], [12], and airline sector [13] through direct financial costs, operational costs, customer service costs, and many other things. Every time a patient or customer chooses to default, there is not only valuable time lost but also a risk that the appointment or booking may never be filled, leading to lower revenue, underutilized personnel, lost commissions, and demoralized employees. Regardless of the size and type of business, the cost associated with cancellations and no-shows is a genuine pain and inconvenience, and it can frequently occur across many specialties in many places. In a replicated facility, researchers found that no-shows caused a daily net loss of 16.4% ($725.42). However, intervention tactics could reduce this loss by half, ranging from 3.8% ($167.38) to 10.5% ($464.27). This means that the estimated gains from interventions could make up 36.0% ($261.15) to 76.9% ($558.04) of the losses caused by no-shows [14].

In the aviation industry, if an aircraft departs with vacant seats that might otherwise have been filled as a result of such a "no-show", the airline will lose the opportunity to resell the seat to another customer, resulting in a direct loss of revenue. Due to this reason, airlines often accept reservations that exceed the cabin capacity based on estimates of the number of no-shows. This is also a concern since sufficiently high levels of overbooking could result in problems including customer dissatisfaction, brand damage (especially during social media times when passenger complaints increase in scope and reach), and revenue effect [10].

Understanding the causes of the high occurrence of no-show incidents is critical. In our daily lives, numerous no-show incidents might not get much notice. For instance, people who reserved a table at a restaurant did not show up; participants were not present at the jury as scheduled; patients scheduled appointments but did not appear; flight passengers did not show up while boarding. Every no-show event appears

insignificant when seen in the context of the larger world. At a macro level, because thousands of identical events occur each year, these seemingly insignificant absences are not favorable to the long-term development of many industries [12]. Hence, it is also necessary to look into several strategies that may be used to lessen the consequences of no-shows and cut down on their frequency.

Subsequently, air travel is particularly susceptible to social, political, and economic changes, causing passenger purchasing patterns to alter significantly. Thus, it is challenging to determine the variables that influence no-show prediction. Thousands of variables are present in big data sets, making it challenging to handle and manage effectively using conventional methods. Consequently, many studies in other fields have focused on variable selection [15].

Lastly, developing precise prediction models is a significant challenge in various industries, such as business, finance, healthcare, etc. Recently, there has been a growing interest in creating predictive models that can forecast the future based on past data and relevant variables. Well-designed prediction models can help companies and organizations make wise choices, conserve resources, and enhance their financial performance. As a result, different industries must maintain their competitiveness by identifying their top-performing model and delivering insightful guidance for making decisions based on reliable forecasts [16]–[18].

This paper aims to discuss the construction of an analytical dataset that identifies the factors contributing to passengers not showing up for their flights. This will involve identifying the variables that caused passengers to not show up for their flights. Following the dataset's creation, predictive models will be created to anticipate the likelihood of passengers failing to show up for their flights to select the best-performing model.

A. Reasons For No-Show

1) Weather Conditions: Bad weather conditions significantly increase the likelihood of appointment failure or no-show, which is considered an uncontrollable factor [5], [8], [19]. Studies have shown that snowfall and extremely low and high temperatures are particularly associated with higher probabilities of no-shows [20]. In response to severe weather predictions, imaging personnel have options such as adjusting exam scheduling, prioritizing patients on the waitlist for extended appointment times, or implementing dynamic scheduling strategies. Research in the United States studied the association between Extreme weather events and HIV-clinic attendance rates [21]. The study discovered that the risk of no-shows increased along with the heat index at 90°F (32.2°C), with a 14% increase on days above 100°F (37.8°C). Similarly, days with over an inch of precipitation had a higher probability of no-shows compared to dry days. The relative risk increased by 16% for 1-2 inches of precipitation and by 13% for more than 2 inches. Days with reported severe natural occurrences had a 10% higher chance of no-show visits compared to non-disaster days.

2) Foot Traffic: The flow of people through airport terminals, encompassing check-in, security checks, and boarding gates, is commonly known as foot traffic. This foot traffic is crucial to flight schedules and customer experience, particularly for no-shows. During peak travel periods, such as holidays or weekends, the high foot traffic can contribute to an increased number of no-shows due to long queues and overcrowding at customs and immigration counters, as well as security checkpoints. This issue is further compounded when airports have limited infrastructure capacity, resulting in insufficient resources like a shortage of open common check-in counters and inadequate personnel availability [22]. These factors significantly impact passengers' efficiency and overall experience and can lead to no-shows.

When airports experience high passenger volumes, factors such as age, gender, and seasonal clothing can slightly impact the security process [23], [24]. Different age groups may require modified screening procedures, potentially leading to longer screening times. Additional screening procedures may be necessary based on gender-specific security concerns, conducted discreetly to ensure passenger privacy and safety. During colder seasons, bulkier clothing worn by passengers may require further inspection, resulting in slightly longer screening times. Despite these considerations, airports and security personnel strive to maintain efficient screening procedures while accommodating high passenger volumes and upholding safety standards for all individuals.

According to a study in manufacturing and service operations management [25], stores that experience higher interday traffic variability face increased uncertainty, leading to significant errors in labor requirement forecasting. This results in mismatches between required and actual store labor, decreasing customer service, and fewer purchases. Similarly, in airports, high foot traffic combined with fewer immigration officers on duty can lead to increased processing time for passengers at customs and immigration. This can cause passengers to struggle to reach their boarding gates on time, mirroring the challenges faced by stores with high traffic variability and inadequate staffing, resulting in diminished customer service. Consequently, passengers may not be able to make it to their boarding gates promptly.

3) Commute Distance: In the healthcare sector, Researchers have found a correlation between the distance between a patient's home and the appointment location and the likelihood of a no-show [1], [6], [11], [19], [26]. Longer distances are associated with a higher risk of no-shows. For example, a study showed that patients have a 1.4% greater chance of missing appointments for every 10-mile increase in their journey [20]. The effect is more significant for patients in the low- and medium-income brackets, with a roughly 2% increase in the risk of no-shows for every 10-mile increase in commute distance, compared to the minimal change observed for high-income patients.

4) Transportation Problems: Transportation problems significantly contribute to no-shows, being an important uncontrollable factor [5], [19]. Insufficient transportation options can result in missed appointments [8]. In Saudi Arabia, female patients' reliance on others for transportation directly affects their rate of no-shows [1]. The recent permission for women to drive in the country has the potential to impact the no-show rates among these patients.

Expanding public transportation systems has reduced the no-show rate, especially for patients living near newly established rail lines [3]. Transportation has been identified as a key factor in maintaining patients' appointment attendance in the healthcare industry [20]. Access to transportation and choosing suitable options are important social determinants of health, as patients cannot receive adequate care if they cannot attend appointments. Lower- and middle-income groups are particularly affected by longer commuting times, relying more on public transportation than wealthier patients who often have private vehicles. Addressing transportation issues can involve scheduling patients at local clinics or outpatient imaging facilities within the same healthcare system to mitigate transportation challenges. Table 1 below presents the authors on the reasons for no-shows.

TABLE I
TABLE OF AUTHOR-REASONS FOR NO-SHOW

Author	Transportation Problems	Commute Distance	Weather Condition	Foot Traffic
[1]	✓	✓		
[3]	✓	✓		
[4]		✓	✓	
[5]	✓	✓	✓	
[6]		✓		
[8]	✓		✓	
[9]	✓			
[11]		✓		
[19]	✓	✓	✓	
[20]	✓	✓	✓	
[21]			✓	
[22]				✓
[23]				✓
[24]				✓
[25]				✓
[26]		✓		
[27]		✓		

B. *Overcoming No-Shows in Airline*

1) Overbooking: The airline management strongly emphasizes an all-out promotion of continuous improvement and uses successful operations strategies to maximize operational efficiency and boost profit. Since they cannot determine whether a passenger will occupy a seat until the flight takes off, airline business management frequently and effectively employs overbooking as a revenue management strategy [28], [29]. Overbooking is a regular phenomenon in the airline business in which carriers sell more tickets than seats available on the plane. With an overbooking system in place, when demand is low during the low season, the cost can be covered by demand from the peak season. The service sector frequently uses overbooking to guard against undesirable events like cancellations and no-show customers that would result in missed opportunities for increased revenue [30], [31].

A "no-show" in the airline industry is when a passenger who has made a reservation for a flight does not show up. As a result, airlines overbook flights to generate revenue because they expect a certain number of passengers will not show up for the flight [10]. Without overbooking, every cancellation and no-show would leave empty seats in the aircraft, denying the airline the chance to earn extra money. For example, if an airline never overbooked a flight with 100 seats and a 10% no-show rate, there would be 10 empty seats. If a one-way flight from Kuala Lumpur International Airport to Bali, Indonesia, costs roughly RM 1100 on an ordinary day, we can determine that the airline company has already passed up the opportunity to earn RM 11,000 on a single flight.

Overbooking is a policy of selling tickets above the seating capacity. This policy has a risk and might potentially cost the airline if the number of customers who show up for departure exceeds seat capacity because the airline must give specified compensation for overbooking penalties [32]. In these situations, the airline usually looks for volunteers to take a later trip in exchange for payment, like a voucher for a future flight. If the airline cannot recruit enough volunteers, they may "bump" some passengers by denying them boarding inadvertently. Although involuntary bumping entitles passengers to compensation from the airline, it can still be inconvenient and frustrating for passengers. Because of the delay, some travelers can miss connections or significant occasions, and altering their travel arrangements could incur additional costs.

Consequently, this makes airline overbooking a problem that could make passengers feel disappointed if the airline cannot conduct flights efficiently. In addition, if the airline's handling of passengers who were inconvenienced by overbooking also did not satisfy passengers with the decision made by the airline, and the compensation received by passengers did not alleviate their unhappiness, customers will be less likely to utilize the airline's services again and are more likely to fly with other airlines as a result, which could have an impact on the airline's ticket sales [30]. Other than that, if the airline fails to handle the matter properly, the worst that can happen is that it can cause serious damage to the company's brand image. In April 2017, a video circulated online showing a customer being forcefully removed from United Airlines Express Flight 3411 due to overbooking. The passenger resisted, and a security guard dragged him out of his seat and down the aisle. The incident caused global outrage, and the security officer involved was suspended. The U.S. Federal Department of Transportation investigated the airline's compliance with overbooking regulations [33].

Determining an optimal overbooking limit has become increasingly important to prevent customer dissatisfaction and protect the company's reputation. One study uses an overbooking model to find a closed-form solution simultaneously for both the optimal booking limit and the optimal overbooking limit [34]. In this work, passenger airline data from Thailand is applied to a mathematical model that combines two of the most crucial airline revenue management tactics: overbooking and seat inventory control. The performance of the two-class overbooking model was then evaluated, and three assumptions were tested using actual data in numerical research. Another study examines the voluntary overbooking model in the context of rational expectation equilibrium to encourage consumer cooperation with airlines, preserve customer goodwill, and optimize expected total returns to airlines [29]. The researchers created A decision tree analysis for both customers and airlines. Simulated and real no-show random variables are subjected to sensitivity

analysis for validation. The results indicate that a "voluntary overbooking" policy that promotes cooperation between passengers and commercial airlines provides significant mutual benefits.

2) Revenue Management: Revenue management is an important part of the airline sector that entails adopting demand management strategies to forecast price and capacity control for various demands effectively. Airlines operate on a business model similar to that of perishable goods, which means that if seats or cargo space remain unsold before a flight, the opportunity to generate revenue is lost [32]. With high fixed costs, revenue management is utilized to predict demand uncertainty issues since excess inventory cannot be stored or carried over to the next period due to the limited capacity of seats and cargo space.

The ultimate goal of revenue management in the airline business is to improve revenue or yield. Revenue management determines how to assign undifferentiated capacity units to satisfy available demand to accomplish the goal. This is done by controlling inventory and price in line with micro-market-level forecasts of consumer behavior [13], [35]. Although airlines would prefer to have higher-fare passengers, they confront market demand uncertainties and frequently offer lower-fare tickets to avoid empty seats and associated opportunity costs. However, airline companies must strike an appropriate equilibrium between the number of tickets supplied at lower rates and those at higher prices to maximize income while providing sufficient capacity for higher-paying customers. Table 2 below indicates the details of ways the airline industry overcame no-shows.

TABLE II
TABLE OF AUTHOR-WAYS AIRLINE INDUSTRY OVERCOME NO-SHOW

Author	Overbooking	Revenue Management
[10]	✓	
[13]		✓
[28]	✓	
[29]	✓	
[30]	✓	
[31]	✓	
[32]	✓	✓
[33]	✓	
[35]		✓

B. Machine Learning Techniques in Resolving Other Challenges

In addition to employing overbooking and revenue management to address the "no-show" problem, researchers use machine learning techniques to make predictions to address additional issues that arise in the aviation and healthcare sectors. For instance, a Chinese researcher employed multivariate linear regression to predict aircraft delay problems [16]. The study presents a method to model the arriving flights and a multiple linear regression technique to estimate delay, comparing with Naive-Bayes and C4.5 approach. After carefully analyzing the data, the researcher discovered a strong correlation between arrival and departure delays. As a result, they use departure delays to forecast arrival delays. According to trials with a realistic dataset of domestic airports, the accuracy of the proposed model is roughly 80%, which is an improvement over the Naive-Bayes and C4.5 approach techniques. Another researcher in China is forecasting passenger distribution in airport terminal main areas based on flight arrangements [36]. In the research, a mathematical optimization (Gamma Distribution) predictive model is created and using the passengers' dwell time in each area of an airport terminal as the input of the predictive model to predict how people would be distributed throughout an airport terminal's main regions based on the configuration of the flights. The findings show that these areas saw peak passenger dislocation due to the airport's departure procedure.

When deciding how many seats to allow for overbooking, airlines depend on predictions regarding the expected number of passengers who may not show up for a specific flight. To address this challenge, a paper proposed a decision support system that integrates the Case-Based Reasoning (CBR) method with Interpolative Boolean Algebra (IBA), takes recommendations from both experts and algorithms, and predicts the number of no-show passengers [37]. Aside from that, a study uses the Box-Jenkins model and an artificial neural network model to anticipate the number of passengers flying in Malaysia based on lag variables as input variables [17].

Following that, another Malaysian researcher used geometric Brownian motion (GBM) to forecast the number of passengers over time [38]. Geometric Brownian motion (GBM) is a relatively simple mathematical model usually used to anticipate the future share price for a brief period. The researcher wishes to compare the distributional behavior data from two local airline firms' passengers. Besides that, a Swedish researcher forecasted the no-show rate of travelers using information from passenger booking data [31]. The researcher used several approaches to perform the prediction to determine whether decision trees, gradient boosting, or neural networks could outperform the simple baseline model. Consequently, gradient boosting produced outcomes comparable to but marginally lower than decision trees in the given KPIs (Key Performance Indicators). Out of all the models, the neural network did the worst.

Patients who miss outpatient visits for diagnostic or clinic tests are known in the healthcare sector as "patient no-shows". Physicians and healthcare facilities must identify these patients to use resources and effectively increase healthcare efficiency. For instance, one researcher created a predictive model to forecast patients' failure to attend scheduled visits using decision trees and AdaBoost [2]. The models' analysis indicates that those patients who missed their appointments tend to be younger, male, with morning appointments, and did not receive a text message on the phone or a reminder. Chinese researchers also developed a prediction model for patient no-shows in online outpatient appointments to help hospitals make informed decisions and decrease the likelihood of patient no-show behavior [6].

As we know, the product in the airline industry is the seat, which is a non-stackable product. The seat demand is unpredictable, the capacity is limited and hard to expand, and the variable costs are extremely expensive. As a result, the airline industry places an extremely high priority on predicted demand prediction. One researcher designed and created the best-fit model using a multilayered feed-forward neural network to estimate passenger demand at the flight origin and

destination levels based on historical data [39]. The researchers employed data such as date, territory, origin-destination, and passenger count to anticipate passenger demand to deliver the best outcomes for capacity utilization decisions.

II. Materials and Method

A. Data Preparation

1) Dataset Description: One of Malaysia's local airline firms provides the primary dataset. This dataset pertains to the records of the passengers, whether or not they arrive at the scheduled departure time. The dataset is a six-month information collection from July 1, 2022, to December 31, 2022. There are 1,046,486 rows and 8 columns in the dataset. The dataset includes information about each passenger's no-show status, the date and country of issuance, the departure date and time, the departure and arrival airports, the aircraft type, the type of cabin class, and the departure and arrival times. Since the dataset provided does not include nearly enough information, feature enrichment is being applied to the original dataset, resulting in additional columns. For instance, the IATA airport codes for each passenger's departure airport and arrival airport are checked to determine the kind of flight for each. IATA codes, often known as location identifiers, are made up of three-character alphanumeric geocodes. The International Air Transport Association uses it to designate numerous airports and metropolitan regions all over the world.

In addition to the original dataset, Wikipedia and other websites are crawled for additional datasets and information. "Open-Meteo" is a website where weather information is obtained. Selected hourly and daily meteorological and climatic data items are among the information made available by the website. Temperature, rain, snowfall, wind, and astronomical aspects like sunrise and sunset are all included in the list of weather variables. The goal of obtaining weather data is to determine whether or not the passenger will arrive if it is raining. Following that, the holidays of 56 nations contained inside the primary dataset are obtained by manually checking the countries' holiday calendars on Google. Since we all know that public holidays frequently result in traffic jams, this information can be utilized to forecast whether or not passengers will arrive at the airport during a holiday.

Based on the data provided in the original dataset, the aircraft type for each passenger is included. This allows for extracting additional aircraft details, such as the manufacturer, total number of seats, cargo capacity, fuel capacity, maximum take-off weight, and others. The purpose of analyzing these details is to determine if there is any correlation between the type of aircraft and the number of no-shows among passengers.

The last supplementary dataset being downloaded is data on airport foot traffic from the "BestTime" website. Foot traffic data is the term used to describe the gathering and analysis of information regarding the flow of people through a certain location, such as an airport. A variety of techniques, including sensors and cameras, are used to collect this data. By studying airport foot traffic statistics, we may learn more about how passengers navigate the airport and whether or not the foot traffic patterns have an impact on the passengers arriving at the boarding gate.

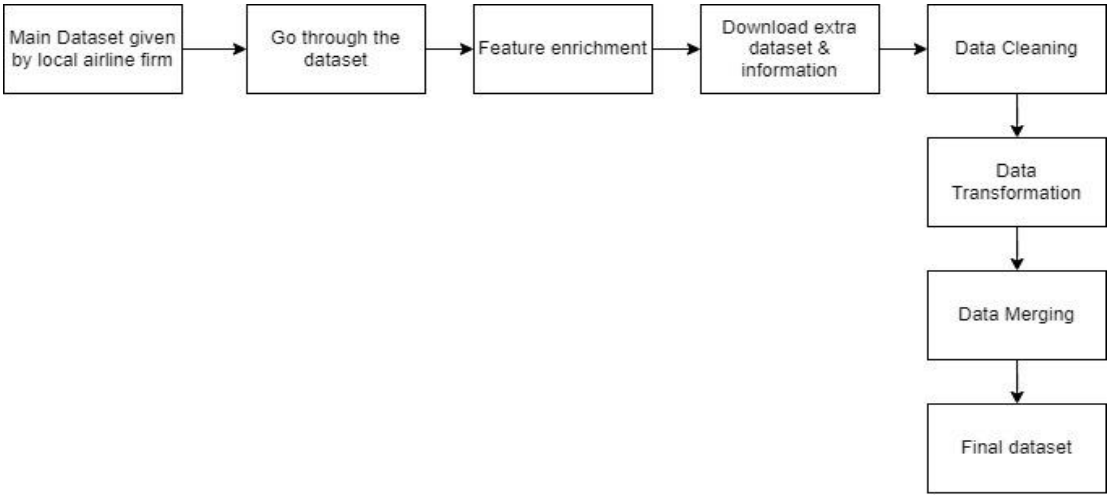

Fig. 1 Flow of obtaining the Final Version of the dataset.

B. Data Pre-processing

1) Data Cleaning: Extensive data cleaning and checks were conducted on all datasets used in the project to ensure the accuracy and reliability of the predictive model. The main dataset provided by the corporation had missing values, specifically in the "Issuing_Country" column. To handle this, the missing data was filled with the value "None" as a workaround. Aside from this, no further missing data was found in the dataset, and all columns were considered meaningful, requiring no added data-cleaning tasks.

Moving on to the IATA code dataset, unnecessary columns such as "Airport ID", "ICAO", and "Altitude" were removed to streamline the dataset. Furthermore, any null values present were filled with the value "None" to ensure data completeness. In the case of the foot traffic and weather dataset, no columns needed to be eliminated since the required information was obtained through an online API, and only the relevant data was retrieved. Moreover, this dataset was found to be free from null values, eliminating the need for additional data-cleaning steps. Finally, for the public holiday dataset, all information was manually acquired by referring to various

sources, including Wikipedia. The only data cleaning step performed on this dataset was filling any null values with the value "No", indicating that it was not a public holiday.

By meticulously addressing missing values, removing unnecessary columns, and ensuring data completeness, the datasets were prepared for further analysis and the development of a robust predictive model.

Algorithm 1: Data Cleaning (Main dataset)
1. Import panda's library.
2. Read the Main dataset.
3. Check null value.
4. Fill in null values in "Issuing_Country" with "None"
5. Export the cleaned dataset from the data frame into a CSV file

Algorithm 2: Data Cleaning (IATA dataset)
1. Import panda's library.
2. Read the IATA dataset.
3. Remove unwanted columns.
4. Check null value.
5. Fill in null values in "City" with "None."
6. Export the cleaned dataset from the data frame into a CSV file

2) Data Transformation: Three datasets are undergoing data transformation: the main, foot traffic, and weather datasets. To prevent errors from occurring in other sections, the first step is to modify the data type of the "Acft_Type", "Issue_Date" and "Dep_Date_Time" columns in the main dataset. Since each value inside the column should be a name of an aircraft type rather than a number, the value in the "Acft_Type" column is transformed from float to string. While being read, the initial data type of "Issue_Date" and "Dep_Date_Time" is a string; after that, both are converted to DateTime data types. Following that, the date and time are retrieved from the "Dep_Date_Time" column and placed in new columns called "Dep_Date" and "Dep_Time". The goal of adding these two new columns is to save time during the data merging process. Based on the "Dep_Date" column, "dt.day_name()" is used to determine the day of the departure date. In addition, departure times are divided into four categories based on the "Dep_Date_Time" column: early in the morning, during the day, during the afternoon, and in the evening.

Algorithm 4: Data Transformation (Main dataset)
1. Import panda's library.
2. Read the main dataset.
3. Change "Acft_Type" from float to string data type
4. Change "Issue_Date" from string to DateTime data type
5. Change "Dep_Date_Time" from string to DateTime data type
6. Retrieve the date and time from "Dep_Date_Time" and place them in new columns called "Dep_Date" and "Dep_Time"
7. Determine the day of each departure date.
8. Divide departure time in "Dep_Date_Time" into four categories: early in the morning, during the day, during the afternoon, and in the evening

Since the online API does not offer the necessary data for some airports, the foot traffic data of a site close to the airport is utilized for that particular airport. To minimize confusion when integrating the data, the name of the airport is then used to replace the names of these locations. In the weather dataset, the "Time" column applies the same logic as the "Dep_Date_Time" column in the main dataset. Besides that, the values in the "WeatherCode" column are merely numbers, which are quite perplexing to people as to what they signify, so the meaning of each number is put into a Python dictionary to replace all of the numbers with their respective meanings.

Algorithm 5: Data Transformation (Weather dataset)
1. Import panda's library.
2. Read the weather dataset.
3. Change "Time" from string to DateTime data type
4. Create a dictionary which inside contains the respective meanings of each value in "WeatherCode"
5. Replace the value in "WeatherCode" by using the dictionary that has been created

3) Data Merging: Data merging is a time-consuming task since three additional datasets, including the details of each IATA code, weather, foot traffic, and public holiday datasets, as well as some aircraft details, must be merged with the main dataset to conduct data analysis and uncover what the data can genuinely tell us.

The merging process begins by combining the IATA code dataset with the main dataset using an inner merger. Due to website restrictions, some of the data are being obtained separately, therefore additional data merging tasks need to be conducted during or after the data crawling process. As the weather data will be gathered from more than 50 nations, an empty data frame is generated before the data is crawled. The weather data is then combined using the "concat" function to create a single CSV file. The same approach is also used for the foot traffic data after it is crawled.

The first step is to perform a left join to merge the public holiday data with the main dataset. This join combines the two data frames using the "Dep_Date" and "Dep_city" columns as the common keys. A left join ensures that all the rows from the main dataset are retained in the merged data frame, and only the matching rows from the public holiday data are included. Before the merging process, a data type checking task is conducted to ensure no data type mismatches could lead to errors.

Algorithm 6: Data Merging (IATA code dataset)
1. Import panda's library.
2. Read the IATA dataset.
3. Merge the IATA dataset with the main dataset using the inner merger.
4. Rename the columns after merging the process.
5. Iterate through the dataset.
6. Determine the type of flight of each passenger by checking both values in "Dep_country" and "Arr_country" If equals to "Malaysia" then assign the string "Domestic" to "Type_of_flight" else "International"

Similarly, the other additional datasets, such as weather, foot traffic, and aircraft data, are merged with the main dataset using left-join operations. After the merging process, some columns are renamed to improve readability. To ensure that the final dataset is accurate, complete, and reliable, cleaning tasks such as checking for and removing null values and

unnecessary columns are performed. These steps are taken to enhance the quality of the data for analysis and decision-making purposes. After the merging and cleaning processes, a comprehensive version of the main dataset is generated.

Algorithm 7: Data Merging (Weather dataset)

1. Import panda's library.
2. Create an empty data frame.
3. Crawl weather data through API
4. Store weather data in a new data frame
5. Merge the empty data frame created at the beginning with the data frame containing crawled weather data.

Export the crawled weather data from the data frame into a CSV file.

Algorithm 8: Data Merging (Main dataset with Public Holiday data, Weather data, Foot Traffic data, and aircraft data)

1. Import panda's library.
2. Read the main dataset.
3. Read the public holiday dataset.
4. Check the data type of each column.
5. Rename the column's name.
6. Merge the data with the main dataset using the left join.
7. Check null value.
8. Remove unwanted columns.
9. Repeat steps 3 to 8 by replacing the public holiday dataset with the next dataset.
10. Export the final version of the dataset from the data frame into a CSV file.

C. Feature Selection

The final version of the main dataset consists of 52 columns, which can be considered quite extensive. To optimize the performance and efficiency of a machine learning model, it is crucial to perform feature selection on the dataset. Besides that, we are also aiming to identify the most relevant and informative features of this dataset. Before proceeding to the feature selection part, the data is split into predictor and target variables. The target variable represents the column that indicates the presence of passengers. Subsequently, the predictor variables are stored in a variable called X, while the target variable is stored in a variable called y. This study employed the Correlation-based Feature Selection (CFS) technique for feature selection. CFS evaluates feature subsets solely based on their intrinsic properties within the data. The top 20 features were selected using this technique, as shown in Table 3 below.

TABLE III
TABLE OF TOP 20 FEATURES

No.	Feature	Score
1	Dep_Date	580371.66
2	Issuing_Country	58433.70
3	Type_of_flight	56373.87
4	Temperature	49009.35
5	SeaLevelPressure	37920.54
6	Business_Class	33773.07
7	Cargo_Cap(kg)	30938.80
8	Max_TOWeight(kg)	29982.75
9	Max_LandWeight(kg)	29908.23
10	WindDirection	29069.74
11	Max_Speed(km/h)	28625.15
12	Fuel_Cap(kg)	28344.75
13	Aircraft_Manu	28233.18
14	Total_Seats	28141.63
15	Acft_Type	27580.18
16	Econ_Class	26896.28
17	EconELR_Class	17740.58
18	Aircraft_Name	17346.74
19	Dep_Date_Time	13775.21
20	BusinessSuite	12574.52

D. Model Construction

This study involves the construction of four classification models using different machine learning algorithms: Neural Networks, Decision Tree, Gradient Boosting, and Random Forest. The Neural Network classifier is implemented using Keras. It begins by specifying the input shape as (20,), indicating that only the top 20 features are fed into the model. The architecture comprises two hidden layers with 32 and 20 units respectively, utilizing the ReLU activation function. The output layer consists of a single unit with sigmoid activation. To compile the model, the Adam optimizer and binary cross-entropy loss and accuracy metrics are chosen. The model is then trained on the training set for 10 epochs using a batch size 32.

For the Decision Tree classifier, the scikit-learn library is employed. The classifier is initialized as a decision tree object, labeled as clf, and subsequently trained on the training data. Moving on to the Gradient Boosting Classifier, scikit-learn is also used for implementation. The classifier is initialized with 100 estimators (decision trees) and a learning rate of 0.1. It is assigned a maximum depth of 3, which controls the complexity of the individual decision trees within the ensemble. The Random Forest Classifier is implemented using scikit-learn as well. Firstly, an instance of the Random Forest Classifier is created with 100 estimators (decision trees) and a random state of 42. The random state ensures the reproducibility of the results across different runs.

All four models are then fitted to the training data, utilizing the top 20 selected features based on CFS (Correlation-based Feature Selection). Subsequently, predictions are made on the testing data, and various evaluation metrics, including accuracy, F1 score, precision, and recall, are calculated. Finally, these metrics are printed to the console, assessing each model's performance.

III. RESULTS AND DISCUSSION

The results obtained from evaluating the different classification models provide valuable insights into their performance. By comparing the models, it is evident that the Neural Networks model achieved an accuracy of 67.6%, which is the lowest compared to the other models. Additionally, its precision value of 45.6% indicates that it struggles to accurately identify positive instances, while the recall value of 67.6% implies that it captures only a moderate proportion of the actual positive instances. Overall, the F1 score of 54.5% suggests a mediocre performance for this model as shown in Table 4.

On the other hand, the Decision Tree model demonstrates superior performance. With an accuracy of 90.2%, precision of 80.8%, recall of 91.7%, and an impressive F1 score of 85.9%, it outperforms the Neural Networks model in all evaluation metrics. These results indicate that the Decision

Tree model accurately identifies positive instances and captures a significant proportion of the actual positives, resulting in a balanced and robust performance. Similarly, the gradient-boosting model exhibits favorable results. It achieves an accuracy of 86.5%, a precision value of 75.4%, a recall value of 86.8%, and an F1 score of 80.7%. Although slightly lower than the Decision Tree model, the Gradient Boosting model still demonstrates a strong ability to identify positive instances correctly and captures a good proportion of the actual positives.

TABLE IV
TABLE OF MODEL RESULTS

Model	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
Neural Network	67.6	54.5	45.6	67.6
Decision Tree	90.2	85.9	80.8	91.7
Gradient Boosting	86.5	80.7	75.4	86.8
Random Forest	90.4	86.2	80.7	92.4

Lastly, the Random Forest model showcases the highest accuracy among the models, with a value of 90.4%. It also attains a precision of 80.7%, a recall of 92.4%, and an F1 score of 86.2%, reflecting a balanced and commendable performance. These results highlight the Random Forest model's capability to accurately identify positive instances while capturing a substantial proportion of the actual positives.

In summary, the evaluation of the classification models reveals that the Decision Tree, Gradient Boosting, and Random Forest models consistently outperform the Neural Networks model in terms of accuracy, precision, recall, and F1 score. Among them, the Random Forest model has the highest accuracy and recall. These results suggest that the Random Forest and Decision Tree models are particularly well-suited for the given classification task, while the Neural Networks model may benefit from further improvements or adjustments.

IV. CONCLUSION

In conclusion, this study intended to thoroughly investigate the issues brought on by no-shows across several industries, focusing on the aviation sector. The study aimed to uncover the factors influencing the occurrence of no-shows and its ramifications for the industry by analyzing a real-life dataset provided by a local airline company. The findings of this study contribute to the existing body of knowledge regarding the detrimental effects of no-shows. Numerous insights into the underlying reasons and their effects on operational effectiveness, revenue management, and customer satisfaction have been achieved by examining the causes and effects of no-shows in numerous industries. The study emphasizes the importance of resolving the issue as soon as possible to offset its negative consequences on many industries, particularly the airline industry, which relies largely on efficient scheduling and capacity utilization.

Furthermore, by examining the approaches taken by researchers in other sectors to overcome the challenges posed by no-shows, this study has identified potential strategies and techniques that can be adapted and implemented within the airline industry. Understanding and applying successful practices from other sectors can contribute to developing effective countermeasures, such as proactive communication, dynamic pricing, and overbooking management, to minimize the impact of no-shows on operational planning and resource allocation.

In addition to investigating the causes and consequences of no-shows, this research has constructed and evaluated four classification models: Neural Networks, Decision Trees, Gradient Boosting, and Random Forest. The accuracy rates achieved by each model provide insights into their effectiveness in predicting no-show incidents. Notably, the Decision Tree and Random Forest models exhibited high accuracy rates of 90.2% and 90.4%, respectively, suggesting their potential applicability in developing predictive models for identifying passengers at a higher risk of no-shows.

However, it is important to acknowledge that the accuracy of the models could be further improved by incorporating additional passenger information. This study recommends obtaining more comprehensive data, including variables such as age, home address, gender, travel history, and booking patterns, to enhance the predictive power of the models. By considering a broader range of factors that influence no-show incidents, the models can better capture the complexity of passenger behavior and provide more accurate predictions, enabling airlines to implement targeted measures to minimize the occurrence of no-shows.

In conclusion, this research paper has comprehensively examined the problems posed by no-shows in different industries, particularly the airline sector. By exploring the causes and consequences of no-show incidents and analyzing existing strategies from various sectors, valuable insights have been gained. Additionally, the construction and evaluation of classification models have demonstrated promising results, indicating the potential for accurate prediction of no-show incidents. However, further research is needed to incorporate additional passenger information and improve the models' accuracy. By doing so, the airline industry can better understand and address the challenges of no-shows, ultimately enhancing operational efficiency, revenue management, and customer satisfaction.

REFERENCES

[1] S. AlMuhaideb, O. Alswailem, N. Alsubaie, I. Ferwana, and A. Alnajem, "Prediction of hospital no-show appointments through artificial intelligence algorithms," *Ann Saudi Med*, vol. 39, no. 6, pp. 373–381, Dec. 2019, doi: 10.5144/0256-4947.2019.373.

[2] A. Alshammari, R. Almalki, and R. Alshammari, "Developing a Predictive Model of Predicting Appointment No-Show by Using Machine Learning Algorithms," *Journal of Advances in Information Technology*, vol. 12, no. 3, 2021, doi: 10.12720/jait.12.3.234-239.

[3] C. Amberger and D. Schreyer, "What do we know about no-show behavior? A systematic, interdisciplinary literature review," *J Econ Surv*, Sep. 2022, doi: 10.1111/joes.12534.

[4] D. Carreras-García, D. Delgado-Gómez, F. Llorente-Fernández, and A. Arribas-Gil, "Patient No-Show Prediction: A Systematic Literature Review," *Entropy*, vol. 22, no. 6, p. 675, Jun. 2020, doi: 10.3390/e22060675.

[5] T. Daghistani, H. AlGhamdi, R. Alshammari, and R. H. AlHazme, "Predictors of outpatients' no-show: big data analytics using apache spark," *J Big Data*, vol. 7, no. 1, p. 108, Dec. 2020, doi: 10.1186/s40537-020-00384-9.

[6] G. Fan, Z. Deng, Q. Ye, and B. Wang, "Machine learning-based prediction models for patients no-show in online outpatient appointments," *Data Science and Management*, vol. 2, pp. 45–52, Jun. 2021, doi: 10.1016/j.dsm.2021.06.002.

[7] S. L. Harris and M. Samorani, "On selecting a probabilistic classifier for appointment no-show prediction," *Decis Support Syst*, vol. 142, p. 113472, Mar. 2021, doi: 10.1016/j.dss.2020.113472.

[8] D. Marbouh *et al.*, "Evaluating the Impact of Patient No-Shows on Service Quality," *Risk Manag Healthc Policy*, vol. Volume 13, pp. 509–517, Jun. 2020, doi: 10.2147/RMHP.S232114.

[9] I. Mohammadi, H. Wu, A. Turkcan, T. Toscos, and B. N. Doebbeling, "Data Analytics and Modeling for Appointment No-show in Community Health Centers," *J Prim Care Community Health*, vol. 9, p. 215013271881169, Jan. 2018, doi: 10.1177/2150132718811692.

[10] A. Perez, "Models for Fitting Correlated Non-identical Bernoulli Random Variables with Applications to an Airline Data Problem," Doctoral Dissertation, Temple University, 2021.

[11] K. Topuz, H. Uner, A. Oztekin, and M. B. Yildirim, "Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network," *Ann Oper Res*, vol. 263, no. 1–2, pp. 479–499, Apr. 2018, doi: 10.1007/s10479-017-2489-0.

[12] C. Wang, R. Wu, L. Deng, Y. Chen, Y. Li, and Y. Wan, "A Bibliometric Analysis on No-Show Research: Status, Hotspots, Trends and Outlook," *Sustainability*, vol. 12, no. 10, p. 3997, May 2020, doi: 10.3390/su12103997.

[13] Syed Arbab Mohd Shihab, Caleb Logemann, Deepak-George Thomas, and Peng Wei, "Autonomous Airline Revenue Management: A Deep Reinforcement Learning Approach to Seat Inventory Control and Overbooking," Cornell University, 2019.

[14] B. P. Berg *et al.*, "Estimating the Cost of No-Shows and Evaluating the Effects of Mitigation Strategies," Medical *Decision Making*, vol. 33, no. 8, pp. 976–985, Nov. 2013, doi: 10.1177/0272989X13478194.

[15] M. Z. I. Chowdhury and T. C. Turin, "Variable selection strategies and its importance in clinical prediction modelling," *Fam Med Community Health*, vol. 8, no. 1, p. e000262, Feb. 2020, doi: 10.1136/fmch-2019-000262.

[16] Y. Ding, "Predicting flight delay based on multiple linear regression," *IOP Conf Ser Earth Environ Sci*, vol. 81, p. 012198, Aug. 2017, doi: 10.1088/1755-1315/81/1/012198.

[17] N. Idrus and N. Mohamed, "Forecasting The Number of Airplane Passengers Using Box-Jenkins And Artificial Neural Network in Malaysia," *Universiti Malaysia Terengganu Journal of Undergraduate Research*, vol. 2, no. 4, pp. 89–100, Oct. 2020, doi: 10.46754/umtjur.v2i4.183.

[18] S. T. Lim, J. Y. Yuan, K. W. Khaw, and X. Chew, "Predicting Travel Insurance Purchases in an Insurance Firm through Machine Learning Methods after COVID-19," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 43–58, Sep. 2023, doi: 10.33093/jiwe.2023.2.2.4.

[19] X. Xu, M. Hu, and X. Li, "Coping with no-show behaviour in appointment services: a multistage perspective," *Journal of Service Theory and Practice*, vol. 32, no. 3, pp. 452–474, Apr. 2022, doi: 10.1108/JSTP-08-2020-0196.

[20] R. J. Mieloszyk, J. I. Rosenbaum, C. S. Hall, D. S. Hippe, M. L. Gunn, and P. Bhargava, "Environmental Factors Predictive of No-Show Visits in Radiology: Observations of Three Million Outpatient Imaging Visits Over 16 Years," *Journal of the American College of Radiology*, vol. 16, no. 4, pp. 554–559, Apr. 2019, doi: 10.1016/j.jacr.2018.12.046.

[21] D. Samano, S. Saha, T. C. Kot, J. E. Potter, and L. M. Duthely, "Impact of Extreme Weather on Healthcare Utilization by People with HIV in Metropolitan Miami," *Int J Environ Res Public Health*, vol. 18, no. 5, p. 2442, Mar. 2021, doi: 10.3390/ijerph18052442.

[22] S. Alodhaibi, R. L. Burdett, and P. KDV. Yarlagadda, "Framework for Airport Outbound Passenger Flow Modelling," *Procedia Eng*, vol. 174, pp. 1100–1109, 2017, doi: 10.1016/j.proeng.2017.01.263.

[23] Y. Li, X. Gao, Z. Xu, and X. Zhou, "Network-based queuing model for simulating passenger throughput at an airport security checkpoint," *J Air Transp Manag*, vol. 66, pp. 13–24, Jan. 2018, doi: 10.1016/j.jairtraman.2017.09.013.

[24] H. Yamada *et al.*, "Modeling and Managing Airport Passenger Flow Under Uncertainty: A Case of Fukuoka Airport in Japan," 2017, pp. 419–430. doi: 10.1007/978-3-319-67256-4_33.

[25] O. Perdikaki, S. Kesavan, and J. M. Swaminathan, "Effect of Traffic on Sales and Conversion Rates of Retail Stores," *Manufacturing & Service Operations Management*, vol. 14, no. 1, pp. 145–162, Jan. 2012, doi: 10.1287/msom.1110.0356.

[26] Y. Zhou, D. Dong, and W. Jiang, "Influence Factors of Patient No Show in a Outpatient Department," *IOP Conf Ser Mater Sci Eng*, vol. 439, p. 032047, Nov. 2018, doi: 10.1088/1757-899X/439/3/032047.

[27] A. R. Teo, C. W. Forsberg, H. E. Marsh, S. Saha, and S. K. Dobscha, "No-Show Rates When Phone Appointment Reminders Are Not Directly Delivered," *Psychiatric Services*, vol. 68, no. 11, pp. 1098–1100, Nov. 2017, doi: 10.1176/appi.ps.201700128.

[28] A. Brieden and P. Gritzmann, "Predicting show rates in air cargo transport," in *2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT)*, IEEE, Feb. 2020, pp. 1–9. doi: 10.1109/AIDA-AT48540.2020.9049209.

[29] D. Dalalah, U. Ojiako, and M. Chipulu, "Voluntary overbooking in commercial airline reservations," *J Air Transp Manag*, vol. 86, p. 101835, Jul. 2020, doi: 10.1016/j.jairtraman.2020.101835.

[30] Shinta Saylindra, Nurul Islami, Tito Warsito, Ira Rachman, and Imam Ozali, "The Understanding of Airlines Overbooking by Some Airlines at The Soekarno Hatta International Airport," *Advances in Transportation and Logistics Research*, vol. 1, pp. 71–87, 2018.

[31] D. Zenkert, "No-show forecast using passenger booking data," Lund University, 2017.

[32] O. A. C. Dewi, "Revenue management model based on capacity sharing and overbooking in the airline," *Journal of Engineering and Management in Industrial System*, vol. 6, no. 2, pp. 86–94, Dec. 2018, doi: 10.21776/ub.jemis.2018.006.02.3.

[33] D. Victor and M. Stevens, "United Airlines passenger is dragged from an overbooked flight," The New York Times.

[34] M. Somboon and K. Amaruchkul, "Applied Two-Class Overbooking Model in Thailand's Passenger Airline Data," *The Asian Journal of Shipping and Logistics*, vol. 33, no. 4, pp. 189–198, Dec. 2017, doi: 10.1016/j.ajsl.2017.12.002.

[35] J. An, A. Mikhaylov, and S.-U. Jung, "A Linear Programming approach for robust network revenue management in the airline industry," *J Air Transp Manag*, vol. 91, p. 101979, Mar. 2021, doi: 10.1016/j.jairtraman.2020.101979.

[36] L. Lin, X. Liu, X. Liu, T. Zhang, and Y. Cao, "A prediction model to forecast passenger flow based on flight arrangement in airport terminals," *Energy and Built Environment*, vol. 4, no. 6, pp. 680–688, Dec. 2023, doi: 10.1016/j.enbenv.2022.06.006.

[37] N. Vojtek, B. Petrović, and P. Milošević, "Decision Support System for Predicting the Number of No-Show Passengers in Airline Industry," *Tehnicki vjesnik - Technical Gazette*, vol. 28, no. 1, Feb. 2021, doi: 10.17559/TV-20191215144655.

[38] N. M. Asrah, M. E. Nor, S. N. A. Rahim, and W. K. Leng, "Time Series Forecasting of the Number of Malaysia Airlines and AirAsia Passengers," *J Phys Conf Ser*, vol. 995, p. 012006, Apr. 2018, doi: 10.1088/1742-6596/995/1/012006.

[39] P. H. K Tissera, A. N. M. R. S. P. llwana, K. T. Waduge, M. A. l. Perera, D. P. Nawinna, and D. Kasthurirathna, "Predictive Analytics Platform for Airline Industry," in *2020 2nd International Conference on Advancements in Computing (ICAC)*, IEEE, Dec. 2020, pp. 108–113. doi: 10.1109/ICAC51239.2020.9357244