Report

Airline on-time performance

Subject : Business Intelligence
Name : Suja John
ID : a22sujjo
Date : 7-1-2023

Part 1

Introduction

In the US, flight delays are a significant and pervasive issue. Almost 40% of airline flights in 2007 arrived at their destination more than 15 minutes late (Ball et al., 2010). Flight delays are unavoidable and have a significant impact on airline profits and losses and airlines need to estimate fight delays accurately since the findings can be used to boost customer happiness and airline agency profits (Yazdi et al., 2020). Roughly one-third of these delayed arrivals were directly caused by the aviation system's incapacity to manage the traffic demands made on it, while the remaining third were the consequence of internal airline issues (Ball et al., 2010). The majority of the remaining amount was brought on by an aircraft that arrived late and had to depart late for its subsequent journey (Ball et al., 2010). Ball et al. (2010) describes the different kinds of flight delays, their causes, and the effects they have on travelers, airlines, and the US economy. The final estimates of the direct and indirect costs of delays for 2007 are also presented, and they are based on several delay components and models.

In this project, comparison of airline on-time data of the year 1991 and 2001 and analysis has done to find new knowledge from the data. The data consists of flight arrival and departure details for all commercial flights within the USA and the dataset contain 29 features and 11044705 records about the airline performance. For identifying the most affected features of on-time flight, Power BI and PySpark have used in this project. Flight on-time data is available from the US Department of Transportation. If a flight departs from the destination fewer than fifteen minutes after the scheduled departure time indicated in the airline's computerized reservation system (CRS), it is deemed to be on time (Khaksar & Sheikholeslami, 2017).

Most affecting features by EDA

The exploratory data analysis and visualizations help to find the most important features that affect the on-time flights.

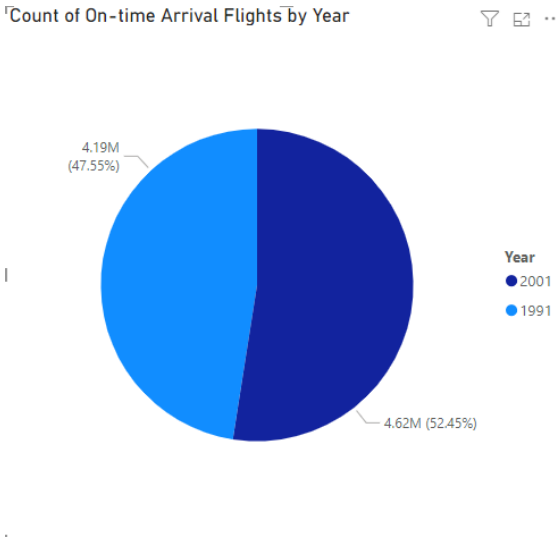


Figure 1. Total count of on-time arrival flights in the years 1991 and 2001.

We can see in figure 1 there is a 15% difference in percentage of number of on-time arrival flights from the year 1991 to 2001.

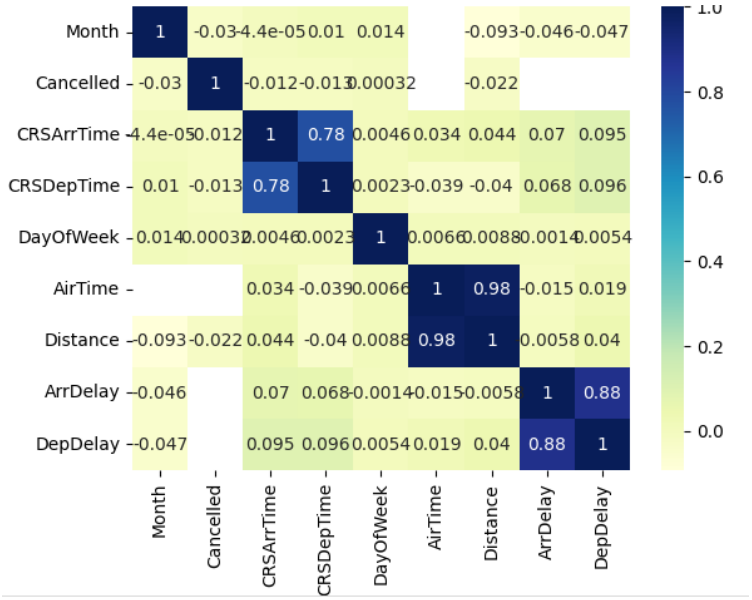


Figure 2. Correlation graph.

To do the analysis of airline on-time data of 1991 and 2001 years, I have used Power BI and PySpark and did the descriptive and predictive analysis. To find the insights from the raw data I created some visualizations using Power BI and find out the features that affecting the most on on-time flights. According to the correlation graph of the dataset in figure 2, distance, airtime, unique carrier, month, day of week, destination and arrival are the features determining whether the flight is on-time or not.

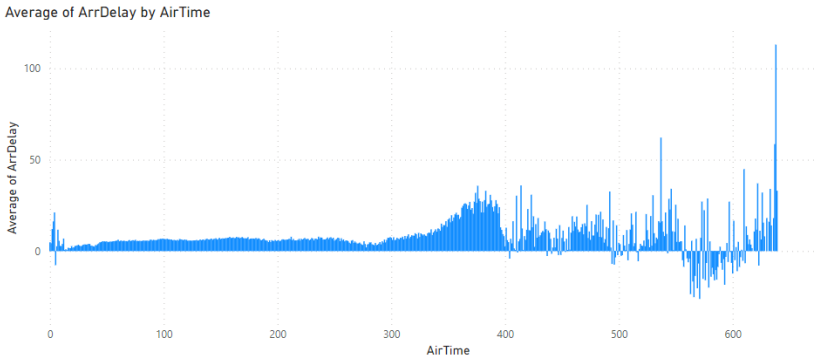
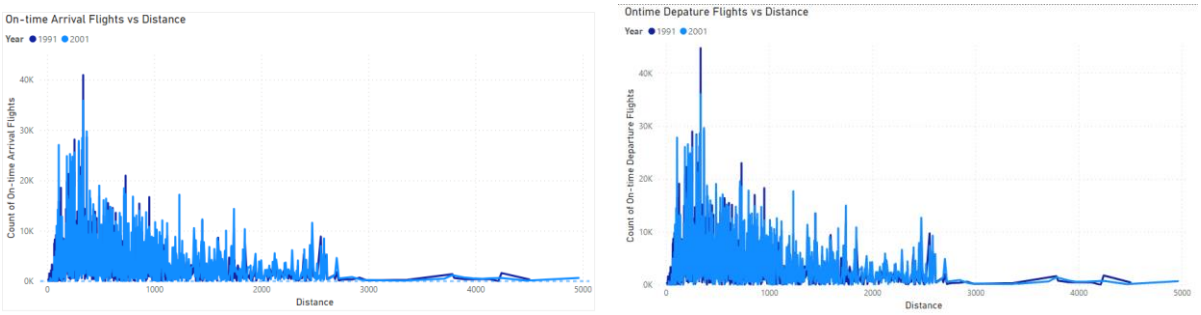


Figure 3. Average arrival delay vs Airtime in the year 2001.

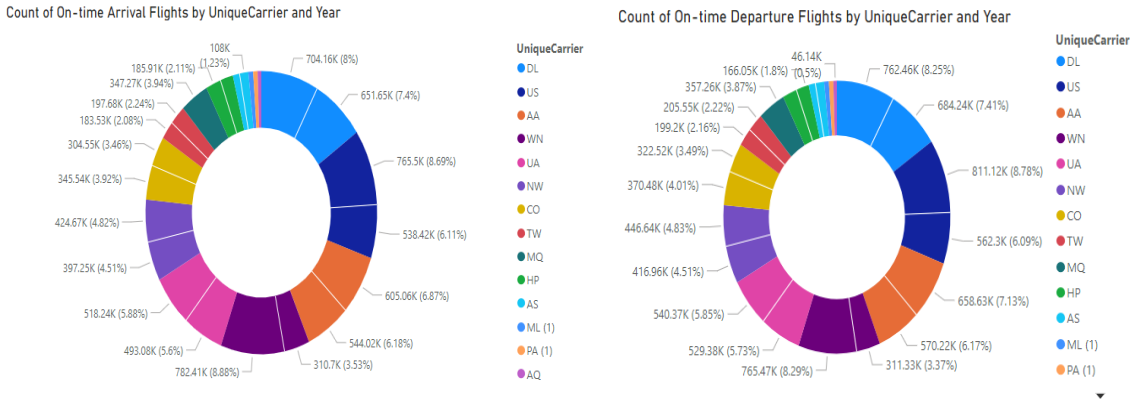
From figure 3 we can see that average arrival delay is considerably low until around 340 minutes and a sudden variation in arrival delay can be observable after airtime more than 340 minutes.



(a) (b)

Figure 4. On-time arrival (a) and departure (b) flights vs distance.

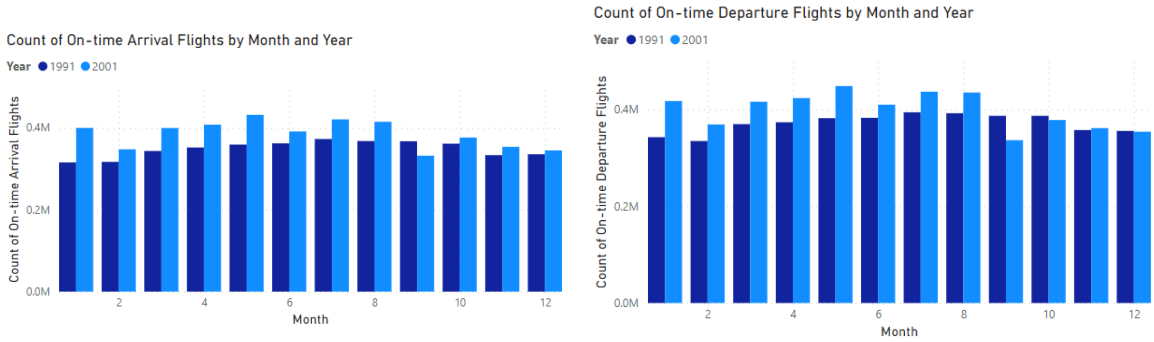
In figure 4, around below 1000miles the number of on-time arrival and departure fights are higher in both the year 2001 and 1991. So increase in distance is one of the reason of flight delay.



(a) (b)

Figure 5. On-time arrival (a) and departure (b) flights by uniquecarrier and year.

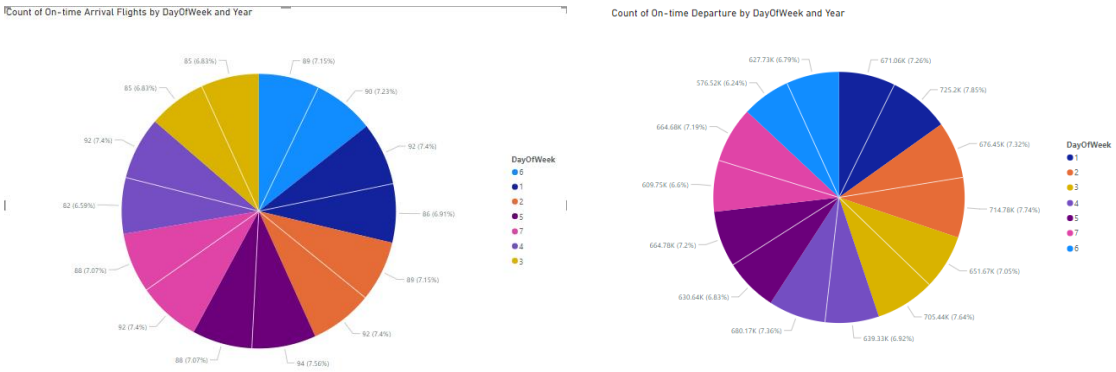
In figure 5, we can see the significance of airline carriers on on-time flight. The carriers like DL, US, AA, and WN has the hight number of on-time arrival flights than others. We can see the similar pattern in both the years 1991 and 2001.



(a) (b)

Figure 6. On-time arrival and departure flights by month and year.

The above figure 6 plots the total number of on-time flights arrived and departed in each month in the year 1991 and 2001. We can see that in year 1991 there is a slight decrease in number of on-time flights in the beginning and end of the year. In 2001, in February and the end of year shows decrease in on-time flights. So year feature also has a slight impact on on-time flights.



(a) (b)

Figure 7. On-time arrival flights by day of week and year

In figure 7 we can see that the percentage of count of on-time flights by day of week. There is not a significant variation we can identify here. But still there are some variations in the total on-time flight count in both years.

	Origin	Dest	N
0	SFO	LAX	25290
1	LAX	SFO	24984
2	LAX	LAS	20841
3	PHX	LAX	20568
4	LAX	PHX	20238
5	LAS	LAX	19759

(a)

	Origin	Dest	N
0	SFO	LAX	27806
1	LAX	SFO	26256
2	LAX	LAS	21154
3	LAS	LAX	20819
4	PHX	LAX	20467
5	LAX	PHX	20082

(b)

Figure 8. Top five on-time arrival (a) and departed (b) flight by the origin and destination in descending order. The airports SFO, LAX, PHX, and LAS have the highest number of on-time arrival and departure flights.

From this analysis we can say that, distance, airtime, unique carrier, arrdelay, depdelay, origin, destination, month, and day of week are the crucial features affecting on-time flights in USA in the descending feature importance. Weather delay, NSA delay, security delay are the other most affecting features in airline on-time prediction but unfortunately our dataset does not contain the details of it.

Predictive Analysis

In predictive analysis, I used logistic model and decision tree model to predict whether the flight will arrive on-time or not. For predictive analysis I used the data of the 2001 only, because airtime feature has null value for the year 1991. Then created temporary table using PySpark and did the exploratory analysis using SQL queries. After the exploratory analysis the most important non-correlated features that are used for training the model, that are airtime, unique carrier, arrdelay, destination, month, tailnum and day of week. Removed the null values present in airtime, arrdelay features. Then type cast airtime, arrdelay, month, and day of week from string to integer. Then created a column called label with 0 or 1 values to show the particular flight is on-time or not. For that the data separated as the flights that arrived before 15 minutes of its schedules time as on-time flights. The flights arrived on the airport after 15 minutes of its schedules arrival time considered as delayed flights.

```
data_model.groupBy('label').count().show()
```

```
+-----+-------+
|label|  count|
+-----+-------+
|    1|1104439|
|    0|4619234|
+-----+-------+
```

Figure 9. On-time and delayed flight count represented in the column named label and 1 represents the delayed flight count and 0 represents the on-time flight count.

Created a pipeline to specify the work flow and created string indexer and did the one hot encoding for the string type features such as unique carrier, destination and tailnum. Then train the logistic regression model using 75% and kept the rest of the data as test data. The accuracy on test data using logistic regression is 0.7862.

```
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    1|       0.0|22463|
|    0|       0.0|83236|
|    1|       1.0|  182|
|    0|       1.0|  210|
+-----+----------+-----+
```

Figure 10. Confusion matrix (Logistic regression model).

Using decision tree model trained with the same training data and got the accuracy on test data 0.8072.

label	prediction	count
1	0.0	330497
0	0.0	1385310
1	1.0	558
0	1.0	473

Figure 11. Confusion matrix (Decision tree model).

We can see that decision tree model outperformed logistic regression model. But in the figure 9 we can see that the data is not well balanced. The count of 1's and 0's is not equally distributed. So there may be chances of bias and an inclination towards 0.

Prescriptive Analysis

Why on-time arrival flight prediction is important and what are the procedures or precautions we can take in order to avoid the issues related to the arrival delay of flights? This is the main examination I would like to introduce in this predictive analysis.

In the aviation industry, on-time arrival flight prediction is essential for a number of reasons. Passengers may experience discomfort, annoyance, and discontent due to flight delays, particularly if they miss their appointments or connections. Airlines that anticipate delays might improve passenger communications, present alternate plans, and give incentives or compensation (Yazdi et al., 2020). First of all, Precise forecasts aid in flight schedule management, lowering the possibility of traffic jams, runway incursions, and other safety problems (Khaksar & Sheikholeslami, 2017). Delay in flights may lead to economic losses and it will adversely affect the operational efficiency too. So predicting a flight is on-time or not using the previous data is crucial.

If we can efficiently predict flight is on-time or not, it will be a huge aid for robust approach and disruption management that have been implemented in aviation. Robust preparation is aimed to reduce delays and avoid delays from affecting other aircraft (Khaksar & Sheikholeslami, 2017). As a result, the first stage entails determining delay and disruption management should also focus on delay prediction. Government authorities or important personalities can schedule their travels according to the accurate predictions given by machine learning algorithms.

Part 2

A larger airport will collect data from passengers' smart phones, flight plans, passport control, security control, and trips to stores and restaurants. The goal is to collect data and then categorize travelers, this arises the privacy concerns. According to the Solove (2006), privacy is a tangled concept with no clear definition or logical understanding of its costs and benefits. It presents a new taxonomy to comprehensively and concretely identify and classify privacy problems. This article organizes privacy problems into four basic groups of harmful activities information collection, such as surveillance and interrogation. Next one is information processing, such as aggregation and exclusion. Then information dissemination, such as disclosure and distortion and invasion, such as intrusion and decisional interference. Airports collect data through various ways such as passports, interrogations, Wi-Fi connections, purchase details etc., then categorizing the passengers connecting to their travel

details like delayed fights and on-time flights will predict the business like shops, restaurants etc., inside the airport. According to Solove (2006), customer segmentation, which is the process of categorizing customers based on their qualities, habits, or needs. client segmentation enables organizations to adjust their products, services, and marketing methods to distinct client segments, thereby increasing customer happiness, loyalty, and retention.

Accessing personal information such as age, gender, ethnicity, habits, pattern of purchase and socioeconomic status is required while collecting demographic data and segmentation. It is vital to ensure informed consent and to respect individuals' privacy rights (Solove, 2006). Researchers must properly convey the goal of data gathering to participants and acquire their agreement. Demographic data, if not collected and examined appropriately, can perpetuate prejudices (Solove, 2006). Biased sampling or discriminatory techniques may result in unfavorable results. Researchers should aim for representative samples and overcome any biases that may exist. Certain demographic characteristics (for example, sexual orientation and health issues) can be stigmatized. Such data collection may unwittingly hurt individuals or communities. Researchers must use caution while handling sensitive information in order to prevent propagating preconceptions (Solove, 2006). It is critical to protect demographic data. Individuals can be harmed by unauthorized access, data breaches, or misuse. Strong security measures and data anonymization are done properly as the first steps. Researchers should be open and honest regarding data collection, storage, and use. Accountability guarantees that data is handled appropriately and ethically (Solove, 2006).

The data collecting to understand customer preferences and communication patterns may lead privacy issues so businesses should make certain that their data gathering and analysis practices do not discriminate against, exclude, or hurt any customer groups. They should also avoid exploiting data to influence or exploit the vulnerabilities or prejudices of their customers. They should also be held accountable for any data misuse or breach and give compensation to affected customers (Solove, 2006).

Reference

Airline On-Time Performance and causes of flight delays. (n.d.). Bureau of Transportation Statistics. https://www.bts.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays

Ball, M. O., Barnhart, C., Dresner, M., & Voltes, A. (2010). Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the. . . *ResearchGate*. https://www.researchgate.net/publication/272202358_Total_Delay_Impact_Study_A_Comprehensive_Assessment_of_the_Costs_and_Impacts_of_Flight_Delay_in_the_United_States

Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. *Scientia Iranica*, *0*(0), 0. https://doi.org/10.24200/sci.2017.20020

Solove, D. J. (2006). *A taxonomy of privacy*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=667622

Yazdi, M. F., Tabbakh, S. R. K., Chabok, S. J. S. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00380-z