Airline on-time performance

Most affecting features by EDA

The exploratory data analysis and visualizations help to find the most important features that affect the on-time flights. We can see that there is a 15% difference in percentage of number of on-time arrival flights from the year 1991 to 2001.

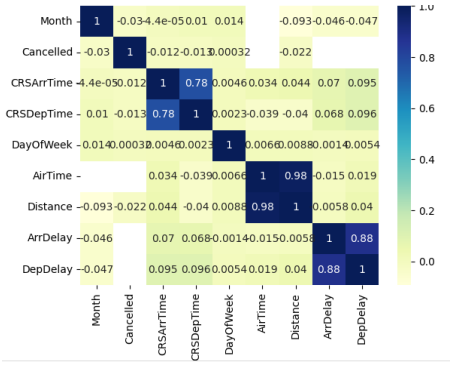


Figure 2. Correlation graph.

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.

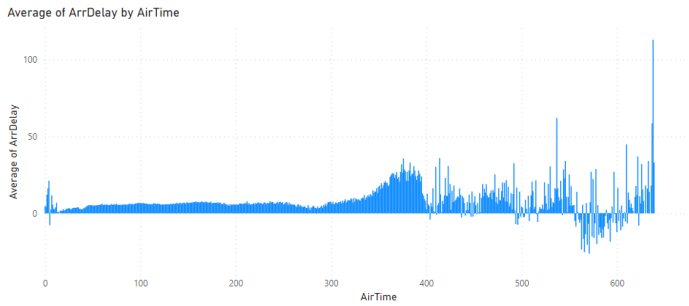
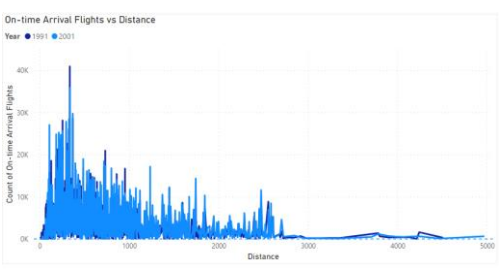
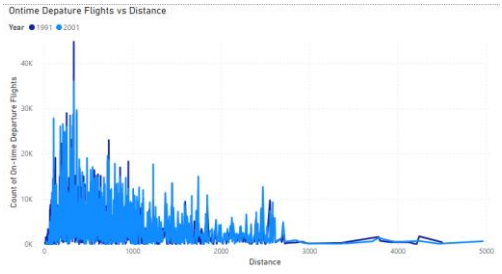


Figure 3. Average arrival delay vs Airtime in the year 2001.

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.



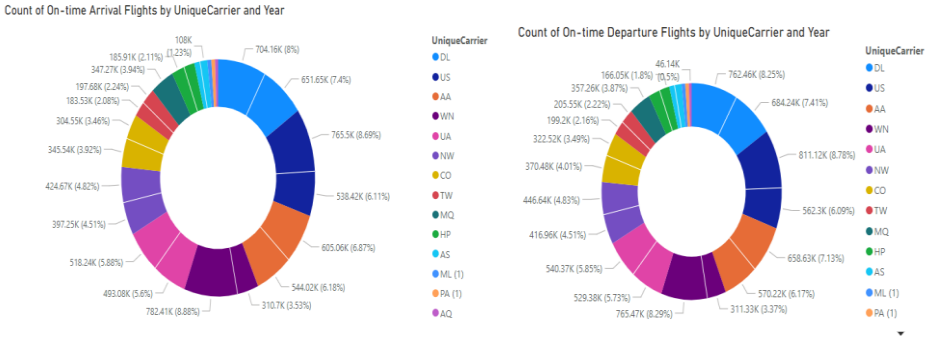
(a)



(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.

	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 6. 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 7. 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 8. 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 9. 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.

Part 2

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

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