INFO2048-1 Business Analytics

Group Work Report

Analysis of airport On-Time Performance

Laguardia, NY airport

GROUP 6

DUYSINX Antoine - S181411
HALLEUX Loïc - S181162
ORBAN Nicolas - S180176

HEC LIEGE - UNIVERSITY OF LIEGE

Master in Business Engineering

Academic Year 2021-2022 - Second Semester

Table of contents

1. Introduction & objectives **2**

2. Project developments **2**

3. Data collection **2**

3.1. Airline On-time Performance 3

3.2. Air Carrier Financial Reports: Schedule B-43 Inventory dataset 4

3.3. Weather data - NOAA Daily summaries 4

4. Data preparation with EXCEL **5**

5. Data preparation with Python **6**

5.1 Data cleaning 6

5.2 Feature selection 6

5.3 Transforming qualitative variables into dummies 7

6. Data exploration **7**

6.1 Data visualization (see Appendix for Power BI) 7

6.2 Correlation analysis 10

7. Predictive models for flight delays **10**

7.1 Decision tree 10

7.2 Logistic regression 12

7.2.1 Principal Component analysis 12

7.2.2 Results Logistic Regression 13

7.3 Neural Network 14

7.4 Random Forest 14

8. Model comparison & best model selection **15**

9. Final results and interpretation **16**

10. Appendixes **17**

11. Bibliography **21**

1. Introduction & objectives

In the context of this project, we deal with flight delays in the aircraft industry. This is a fundamental subject of matter in such a context because not being on time is not something desirable, but not only. Indeed, a delay at the airport location means money to be paid by the airport, as losing time is intrinsically linked to a loss of money due to opportunity cost, in addition to the compensation to be paid to the airport's customers in case of delay. Lastly, delays also generate stress for passengers, which may cause passengers to choose to travel with another airport.

This is why it is crucial for an airport company like the one of interest here, to predict if a flight might be delayed, and above all to understand the reasons behind such delays, in order to reduce them as much as possible; therefore lowering the overall costs of the company and keeping customers satisfied.

Objective:
This project aims at designing, explaining and detailing classification predictive methods that can predict flight delays before they are announced on the departure boards. The time frame used to build our models is the year 2019, between July and December. However, the model can be easily extended to realize forecasts on other time periods.
It should also be noted that the airport that will be studied here is the **LaGuardia Airport (LGA)**, located in the U.S. (in the state of New York).

2. Project developments

Our project follows very logical steps. Data usage requires several steps to be performed before actually being able to use them with our models. Such steps involved are: the **collection of data** (including the first data set we have been provided with), the **in-depth understanding of the variables**, permitting us then to perform the so-called "**data preparation”** and **“data cleaning**", before actually proceeding to the **data exploration** and **data visualization**. Finally, we apply some **predictive models**, select the best ones, and try to **interpret the results**.

3. Data collection

For the purpose of this work, we merged two additional datasets with the Airline On-time Performance dataset which was already provided. It is important to search for additional powerful predictors that can well explain the variance of the delay time. If it is the case, the quality of our classifier will be greatly improved.
In this section, we present these 3 datasets and their respective features.

3.1. Airline On-time Performance

Data table maintained by US certified air carriers on the U.S. Bureau of Transportation Statistics website. It contains scheduled and actual arrival and departure times for flights, which are essential to calculate the delays. The different variables that we have in this dataset are summarized in the table below:

YEAR	Time reference indicating the year of the data provided (here: 2019).
QUARTER	Time reference indicating the quarter of the data provided (so that QUARTER ∈ [1,4]).
MONTH	Time reference indicating the month of the data provided (so that MONTH ∈ [1,12]).
DAY_OF_MONTH	Time reference indicating the day of the data provided, on a monthly basis (so that DAY_OF_MONTH ∈ [1,31]).
DAY_OF_WEEK	Time reference indicating the day of the data provided, on a weekly basis (so that DAY_OF_WEEK ∈ [1,7]).
MKT_CARRIER_FL_NUM	Flight number of the carrier marketed to customers.
OP_UNIQUE_CARRIER	The two last characters of the tail number. Most airlines use an abbreviation of the company name, one of the airline's identification codes, or an abbreviation of the leasing company that owns the aircraft.
TAIL_NUM	Alphanumeric code between two and six characters in length used to identify a specific airplane. The alphabetical prefix of a tail number is indicative of an airplane's country of origin. All United-States-based tail numbers begin with "N", Canadian ones begin with "C", German ones with "D", and so on and so forth.
ORIGIN	Geographical reference indicating the airport of origin using an identification code which is assigned to the airport of origin.
DEST	Geographical reference indicating the airport of destination using an identification code which is assigned to the destination airport.
DEP_SCHED	Time reference indicating the scheduled time of departure.
DEP_OBS	Time reference indicating the observed time of departure.
DEP_DELAY	Time reference indicating the time difference between the scheduled departure time, and the actual departure time; measured from the origin airport gate.
TAXI_OUT	Time reference indicating the amount of time an

	aircraft spends in movement on the surface of an airport. Taxi-out time is the period of time between the time an aircraft leaves a terminal gate, and the time at which it actually takes off from an airport (this entire process is supervised by ATC (Air Traffic Control) where pilots are given clearance to depart from the gate, takeoff, and so on).
ARR_SCHED	Time reference indicating the scheduled arrival time.
ARR_OBS	Time reference indicating the observed arrival time.
ARR_DELAY	Time frame measured in minutes, such that arrival delay is equal to the difference of the actual arrival time, minus the scheduled arrival time.
ARR_DEL15	Binary variable; taking value 0 if the flight arrival is not delayed or delayed by less than 15 minutes (on-time), 1 otherwise (delayed).
DISTANCE	Distance in miles, between airports.

3.2. Air Carrier Financial Reports: Schedule B-43 Inventory dataset

This database contains information on the technical characteristics of the aircraft fleet. It was retrieved from the U.S. Bureau of Transportation Statistics website (*OST_R* | *BTS* | *Transtats*, 2022). Three chosen features were extracted from this dataset:

NUMBER_OF_SEATS	Numerical variable indicating the number of seats in the aircraft.
CAPACITY_IN_POUNDS	Numerical variable indicating the available capacity, measured in pounds.
MANUFACTURE_YEAR	Time reference indicating the year of manufacture of the aircraft.

3.3. Weather data - NOAA Daily summaries

The weather dataset was retrieved from The National Oceanic and Atmospheric Administration (NOAA) website (*Datasets* | *Climate Data Online (CDO)* | *National Climatic Data Center (NCDC)*, 2022). Since hourly weather data was not available for the Laguardia Airport, we decided to take the data from the daily weather summaries database instead. It contains **14 relevant variables** that will be used as inputs in our predictive models. They are the following:

AVG_TEMPERATURE	Average outside temperature during the day, measured in Fahrenheit.
PRCP	The amount of precipitation that has fallen during

	the day, measured in inches.
WIND_DIR	Geographical variable indicating the direction of the wind. It is normally measured in degrees from 0 degrees clockwise through 360 degrees. North is 360 degrees. A wind direction of 0 degrees is only used when the wind is calm.
WIND_SPEED	Speed of the wind variable measured in knots.
PRESSURE	Dummy variable: Atmospheric pressure, measured in bars.
WT01	Dummy variable: Fog, ice fog or freezing fog
WT02	Dummy variable: Heavy fog or heaving freezing fog
WT03	Dummy variable: Thunder
WT04	Dummy variable: Ice pellets, sleet, snow pellets
WT05	Dummy variable: Hail
WT06	Dummy variable: Glaze or rime
WT08	Dummy variable: Smoke or haze
WT09	Dummy variable: Blowing or drifting snow

4. Data preparation with EXCEL

A big part of our data preparation was done with the help of Excel. In this section, we will try to explain all the steps that we took to get our final dataset.

- **Correcting wrong data types**

At the opening of the "Airline On-time performance" dataset, there were some data type errors. It was mainly due to Excel because it did not understand the point as the decimal separator. We solved that problem by replacing the dot with a comma with the help of the tool "search & replace".

- **Creating the "DATE" variable**

We created the "DATE" variable using the "DATE()" function of Excel. This allows us to have a common variable with the weather dataset. Without it, we simply cannot merge the two datasets.

- **Creating the "DEP_HOUR" variable**

We also created the "DEP_HOUR" variable using the "ENT()" function of Excel to take the Integer value of the variable "DEP_SCHED". We thus end up with 24-time slots of one hour each for the departure time. This new variable will be helpful for data visualization.

- **Merging the "On-time performance" dataset with the "AirCarriersFinancialInventory dataset**

We merged the "On-time performance" dataset with the "AirCarriersFinancialInventory" dataset based on the common feature "TAIL_NUM". Three new variables were therefore added to the main dataset, namely the number of seats in each plane (NUMBER_OF_SEATS), and the capacity of the plane measured in pounds (CAPACITY_IN_POUNDS), and the year in which the aircraft was manufactured (MANUFACTURE_YEAR). From the technical point of view, the merge has been realised with the "RECHERCHEX()" function.

- **Merging the "On-time performance" dataset with the "Weather" dataset**

We merged the "On-time performance" dataset with the "Weather" dataset based on the date. This operation allows us to add 13 explanatory variables to the main dataset. Their description can be found in the variable description section of this report. We also used the "RECHERCHEX()" function to match the corresponding observations of the two datasets.

5. Data preparation with Python

The rest of the data preparation has been done using Python.

5.1 Data cleaning

- **Missing values**

We dropped all the 992 rows with missing values because it is a tiny proportion of the dataset.

- **Duplicated rows**

The only duplicated row was dropped.

5.2 Feature selection

We also dropped directly attributes that provide no useful information or are too similar to another one. The non-essential variables are presented in the table below:

YEAR	Always the same year. Contains no information.
QUARTER	Almost the same information as the MONTH variable.
DATE	Gives the same information as the DAY_OF_MONTH and MONTH variables.
TAIL_NUM	No real predictive power.
ORIGIN	Always the same origin airport. Contains no information.
DEP_OBS & DEP_SCHED	Too much redundant information. We decided to keep only a time slot of departure with DEP_HOUR as well as the difference between these two

	variables, that is DEP_DELAY. .
ARR_SCHED & ARR_OBS	Redundant information with ARR_DEL15.
WT05 & WT09	No recorded observations.
MKT_CARRIER_FL_NUM	No real predictive power.

5.3 Transforming qualitative variables into dummies

Two qualitative predictors, namely “OP_UNIQUE_CARRIER” and “DEST” were converted into dummy variables. Each level of these predictors is now represented by a dummy variable. To this end, the pandas method get_dummies() was used. It is our final dataset, before the split into a train set, a validation set and a test set. It now has a dimension of 82,765 rows and 114 variables, which is quite huge.

6. Data exploration

6.1 Data visualization (see Appendix for Power BI)

In order to better visualize the dataset, we created a Power BI dashboard. Inside, we analyzed the number of flights with arrival delays greater than 15 minutes, which accounts for 19,633 out of 83,759 flights (23.4%), with respect to other variables of the dataset.

1. **Percentage of flights with delays for each carrier**

We can see that the ***carriers with the most delays*** *are* ***C5*** *and* ***ZW*** *with respectively 44 and 42% of their flights having delays* respectively. It turns out that they are also *the ones with the lowest number of total flights with 99 and 50 flights respectively*. They both have one single destination which is **IAD** (Dulles International Airport). However, the **YV** carrier operator also offers flights to IAD, but only 28.2% of its flights had an arrival delay superior to 15 minutes. It is thus clearly a better performance than C5 and ZW carriers but it is still a high percentage. In conclusion, flying to IAD might result in a higher probability of arrival delay greater than 15 minutes.

For the last ***3 carriers with the lowest percentage of delays****, we observed* ***UA****,* ***AA*** *and* ***EV*** *with respectively 20, 20 and 19%*. One thing that is interesting to observe is that AA has a total of 12,298 flights and has one of the lowest percentages of delays. But by looking at the bigger picture, we can see that *carriers with the highest number of flights are not the ones with the highest percentage of delays, but on the contrary with a lower percentage*. Indeed, the carriers with a total number of flights higher than 10,000 all have a percentage of delays lower than 25%. Those carriers also serve more destinations in our dataset. Therefore, one possible relation is that the lower the number of total flights, the higher the delay percentage. The carrier operator EV, which has 889 flights with a delay percentage of 19.34%, is the only exception to that rule.

2. Top 6 Destinations with most delays

We observe that **ORD** (Chicago O'Hare International Airport), **BOS** (Boston Logan International Airport) and **ATL** (Atlanta Airport) are the ***destinations with the most arrival delays greater than 15 minutes***. Yet, they are also the ones with the highest number of total flights arriving. In terms of percentage, they respectively have 24.45, 26.42 and 21.78% of their arriving flights having delays, which is definitely not the highest.

A smarter metric would be the ***destinations with the highest percentage of arriving flights having delays superior to 15 minutes***. It shows that **CAK, MHT, DAY, LEX, CHA, IAD, DSM, and DAL are the destinations with the worst performances.** They all have a percentage of delayed arriving flights above 30% with CAK having 42.45% of their flights arriving with a delay greater than 15 minutes.

3. Percentage of flights with Arrival delay > 15 per Month

When we look at the graph, we can clearly observe that for July, August, October and December, we have a higher percentage of delays while for September and November, we observe a lower percentage. This can be related to the increase in demand for flights during those same months. One possible explanation for those higher delay percentages is that in July and August, people tend to go on holidays for summer which leads to an increase in demand and thus more delays. We can apply the same reasoning in December when people travel to see their families for Christmas and New Year's Eve.

4. Percentage of Arrival delays vs WT01

This graph analyzes the variable arrival delay greater than 15 minutes, versus the variable WT01, when there is fog. We can observe that 38.07% of the delays occur when there is some kind of fog. Thus, one possible relation is that when there is fog (WT01), there is a higher probability of having delays.

5. Arrival delay greater than 15 w.r.t. day precipitation in inches

This sector chart shows us that 83.84% of arrival delays greater than 15 minutes happen when there is no rain or less than 0.5 inch of rain during the day. If we take a closer look at delays in our dataset, there are 1,247 delays greater than 15 minutes out of 3,533 flights when there is more than 1 inch of precipitation per day. Meaning that 35.3% of flights are delayed by more than 15 minutes during precipitation with more than 1 inch per day. When precipitation is more than 2 inches, it becomes 35.2% (144 out of 409). But when there is no rain during the day, the percentage of flights delayed is 19.13% (10,785 out of 56,370). Therefore, we can conclude that, the more it rains, the higher the probability of observing flights delayed by more than 15 minutes.

6. Percentage of flights with delay > 15 per day of the week

This bar chart displays the **percentage of flights that have an arrival delay superior to 15 minutes for each weekday**. Please note that on the horizontal axis, the first day of the week corresponds to Sunday. We can observe that between 15% and 20% of flights are delayed from Sunday to Thursday, but for Friday and Saturday we observe a lower percentage around 5% and 9%. This might be explained by the fact that business travels occur more between Sunday and Thursday.

7. Map

The map displays all **flight destinations** with a bubble. The **size of the bubble** corresponds to the number of flights to that destination (the bigger the bubble, the higher the number of flights). Regarding the **color of the bubble**, it determines the percentage of flights with an arrival delay greater than 15 minutes for that destination. We can see that the most served destination airports are not the ones with the highest percentage of delays. Indeed, when the number of total flights is above 1,000, the percentage of delays does not exceed 29.78%. On the contrary, all destinations that have a percentage above 30% have a total number of flights below 1,000. However, it does not mean that when the number of flights is below 1,000 there is a percentage of delay superior to 30% because, for example, there is a destination with 15 flights that has a percentage of delays of 20%.

8. Type of Arrival delay

This chart displays the type of delay. In the caption, “On time” includes flights arriving in advance. We notice that the higher the number of flights, the higher the number of delays, which is obvious. We can also observe that large delays occur less often than small delays but still represent a significant portion for bigger carriers. In terms of proportion, we can see that they are moving around the same value. There is no carrier with a high number of flights and no small or large delay.

9. Total flight delays per hour

On that chart, we plot the **number of flights with an arrival delay greater than 15 minutes with respect to the departure hour**. It should be noticed that the total number of delayed airlines is rising all over the day. One rational explanation might be that each delay generates a domino effect that causes other delays throughout the day. It is also worth seeing that there are no flights between 11:00 PM and 5:00 PM. To conclude, we can state that there is a higher probability of having arrival delays greater than 15 minutes when taking a flight between 3 PM and 9 PM.

6.2 Correlation analysis

Even though the linear correlation coefficient completely misses out on nonlinear relationships, it is still a powerful indicator of a potential relationship between different variables. Let's apply it to our dataset (OP_UNIQUE_CARRIER and DESTINATION variables excluded) to see which variable can significantly influence the delay:

```
ARR_DEL15              1.000000
DEP_DELAY              0.544038
TAXI_OUT               0.409832
DEP_HOUR               0.234376
WT01                   0.156828
WT03                   0.130719
PRCP                   0.128860
WT08                   0.116215
WIND_SPEED             0.088107
WT04                   0.056027
AVG_TEMPERATURE        0.042208
WT06                   0.034304
WT02                   0.020355
MANUFACTURE_YEAR      -0.010352
DISTANCE              -0.016343
CAPACITY_IN_POUNDS    -0.019461
NUMBER_OF_SEATS       -0.028189
MONTH                 -0.054981
DAY_OF_MONTH          -0.064631
DAY_OF_WEEK           -0.070924
WIND_DIR              -0.101798
PRESSURE              -0.109972
Name: ARR_DEL15, dtype: float64
```

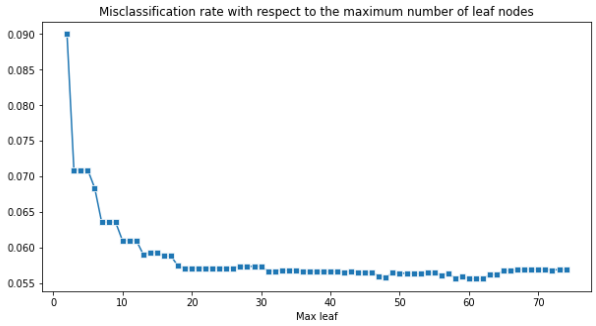
Without any surprise, a **delay at departure** implies very often a delay at the arrival.
The **taxi-out time** also heavily influences the delay. It seems logical because the higher the average departure runway queuing time, the higher the probability of being delayed.
One of our previous observations is also confirmed. It was that the **delay increases linearly as the day progresses**, and the departure hour. A rational explanation behind this phenomenon is that the total delay accumulates throughout the day and leads to further delays. Each flight delay generates some domino effects, that's why we need to prevent it.

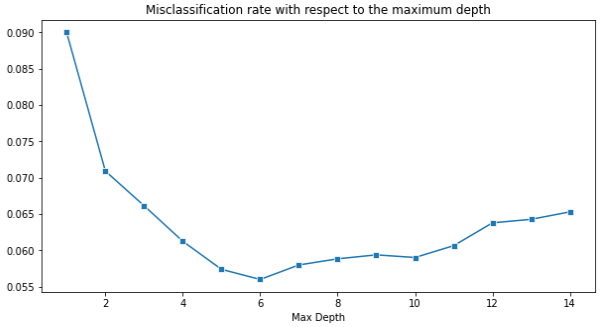
We also see that **meteorological variables** such as the pressure, the wind direction, the fog, the rain, or also a thunderstorm can really lead to some delays.
This is not surprising since we know that **airline companies are heavily dependent on the weather quality**. Unfortunately, there is nothing we can do about it. Finally, for the rest of the variables, either the linear relationship with the delay is less significant, or the relationship is non-linear.

7. Predictive models for flight delays

7.1 Decision tree

The first classification model that we decided to use is a decision tree. It is capable of fitting very complex datasets, this is why we decided to train the model on our full set of features. However, they are prone to overfitting the training dataset, which means that they do not generalize the data well, so predictions are of poor quality. To counter this tendency, we decided to optimize the tree by limiting its depth and its number of leaf nodes. We set them to the value minimizing the misclassification rate of the tree in the validation dataset.
Training an optimized decision tree with these parameters, gave us the following results:





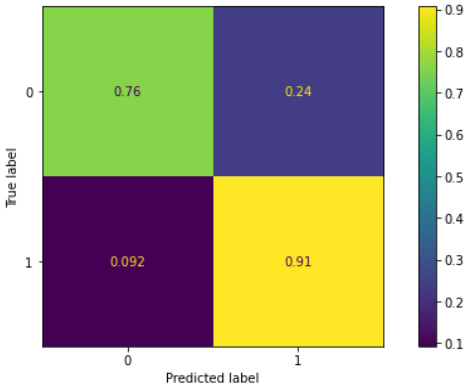
- **Most significant variables**

If we inspect the representation of our decision tree, there is no doubt; the best predictors for the decision tree are the **departure delay (DEP_DELAY)** and the **aircraft taxi-out time (TAXI_OUT)**. Most of the tree nodes' conditions are based on these two variables. In a lesser extent, the model used other predictors such as:

- The departure hour (DEP_HOUR)
- The manufacture year of the plane (MANUFACTURE_YEAR)
- The day of the week (DAY_OF_WEEK)
- The average temperature (AVG_TEMPERATURE)
- The number of seats (NUMBER_OF_SEATS)
- The pressure (PRESSURE)
- The distance between origin airport and destination (DISTANCE)

- **Confusion matrix**

The first row of the matrix represents the **negative class**, that is on-time airlines. While the second row represents the **positive class**, that is the flights considered as late.



In the top-left of the matrix, we can see that 76% of the on-time flights are classified as such by the model (**True negatives**). While 24% of the on-time flights are wrongly classified as late (**False Positives**). In the bottom-left, 9% of the delayed flights are wrongly classified as on-time (**False Negatives**). Finally, 91% of the delayed flights are correctly identified by the model (**True Positives**). Therefore, it is not unreasonable to say that the model performs quite well.
In addition, we believe that it is better to have a higher false positive rate rather than a false negative rate, as we prefer a model that falsely warns us of a flight that has all the characteristics of a delay, rather than a model that does not sufficiently filter out truly late flights.

Although the confusion matrix is an interesting tool, there exists some more concise and equally relevant metrics. First, we can measure the accuracy of the positive predictions with **Specificity**. This is really the precision rate of the classifier. Secondly, we can look at the **Sensitivity**, which measures the proportion of positive instances that are correctly detected by the classifier. Lastly, the relation between the specificity and the sensitivity is the **misclassification rate**, that is the proportion of observations in the dataset that are classified in the wrong class by the model.

It is also worth noting that there is a **tradeoff between the specificity and the sensitivity**. Unfortunately, we cannot have it both ways. Indeed, increasing one reduces the other at a certain point. Therefore, the higher the sensitivity, the more false positives we get.

Decision tree's performance		
Specificity	TP/(TP+FP)	0.9768
Sensitivity	TP/(TP+FN)	0.8333
Misclassification rate	1 - (TP+TN)/Observations	0.0569

In the context of Airline delays classification, we mostly care about sensitivity, because we prefer a classifier that identifies all the flights that might be potentially late. In other words, we prefer to be warned for nothing. For us, it is preferable to be warned that a flight has to be monitored for its delay, even if it is not late at the end (false positive) than a flight that is really late but for which we don't have the information (false negative). In this configuration, the airport will be able to take dispositions in advance, and to adapt the organization.

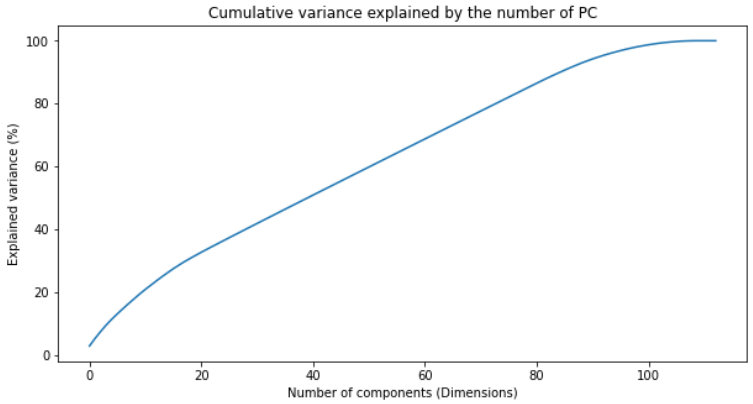
7.2 Logistic regression

The next classification model that we used is the logistic regression one. It estimates the probability that a flight will be late. If the probability is greater than 50%, then the flight is considered as being late.

7.2.1 Principal Component analysis

Unfortunately, due to the large number of covariates, it is really difficult to train the model with the whole set of features. As a consequence, we need to narrow down the feature dimension of our dataset. To this extent, we chose to perform a **principal component analysis (PCA)**. On the one hand, it will help to greatly speed up the training process, and on the other hand, it will also remove the multicollinearity between the variables. PCA identifies orthogonal axes (principal components) that account for the largest amount of variance in the training set. It identifies as many axes as the number of dimensions in the dataset. Once all the principal components have been found, the dimension of the dataset can be downed to *k-dimensions* by projecting it onto the hyperplane defined by the first *k-principal components*. PCA tries to preserve as much variance as possible, but there is still always a loss in information. Therefore, it will, on the one hand, slightly influence the quality of the predictions, and on the other hand reduce the interpretability of our results.

Results can seen on the following graph:

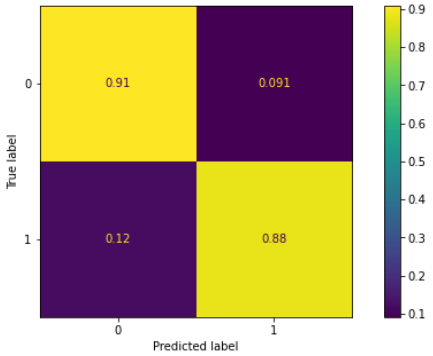


We must admit that our results are a little bit disappointing (as shown by the graph). In fact, the first components explain very little variance (e.g.: the 1st PC explains only 2.95% of the total variance), hence we were not able to substantially reduce the dimension of our initial data. We chose to reduce the dimensionality down to about 95 dimensions in order to not lose too much explained variance. In our opinion, PCA is not performing well because we have a lot of dummy variables. However, it helps anyway to reduce the noise and eliminate less informative variables.

7.2.2 Results Logistic Regression

If we train the logistic regression model on the dataset returned by the PCA, it performs better.

- **Confusion matrix**



Logistic regression's performance		
Specificity	TP/(TP+FP)	0.9089
Sensitivity	TP/(TP+FN)	0.8810
Misclassification rate	1 - (TP+TN)/Observations	0.0976

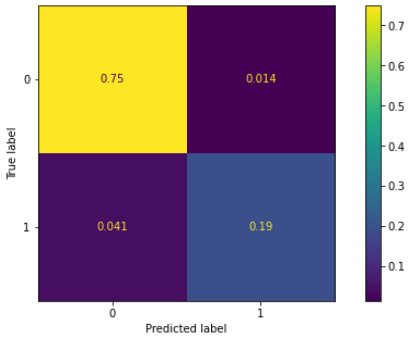
In the case of logistic regression, specificity and sensitivity are more balanced. It means that the model has a high ability to correctly identify flights with, as well as without delays.

7.3 Neural Network

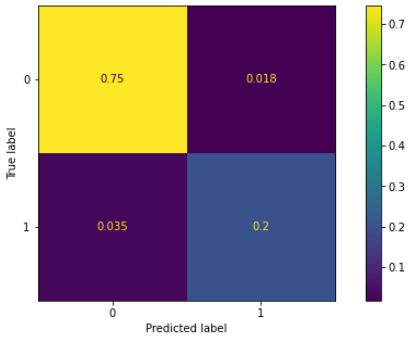
We pursue our predictive-modeling task by constructing a Neural Network model. It is one of the most powerful models available today. However, *its interpretability is almost impossible since it acts as a black box*. As we have a huge quantity of data, we decided to train our data on the full set of features. Nonetheless, we also tried to use the results of our PCA as the input for the NN, and it gave slightly better results. It may be explained by the absence of multicollinearity and by the data standardization which gives less weights to outliers.

- **Confusion matrix**

Neural Network : all features



Neural Network: dimensionality reduction



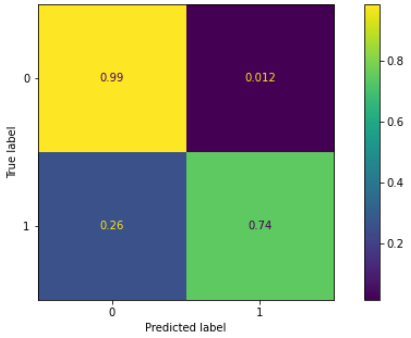
In both cases, results are extremely good but it is true that the neural network based on dimensionality reduction slightly outperforms the one based on all features. It has a lower false negative rate, a higher true positive rate and it generalizes a little bit better. At the end, the misclassification rate is marginally improved. However, based on hundreds of thousands of flights every month, even a small difference in model performance can have the impact of saving a lot of money. In conclusion, we can say that these two classifiers are excellent!

	NN : all features	**NN: dimensionality reduction**
Specificity	0.9812	0.9764
Sensitivity	0.8241	0.8534
Misclassification rate	0.0557	0.05261

7.4 Random Forest

The last classification method that we decided to implement is a Random Forest. This method aggregates the predictions of a group of Decision Tree classifiers, and predicts the class that gets the most votes. It is one of the most powerful Machine Learning algorithms available today, but its interpretation is extremely difficult.

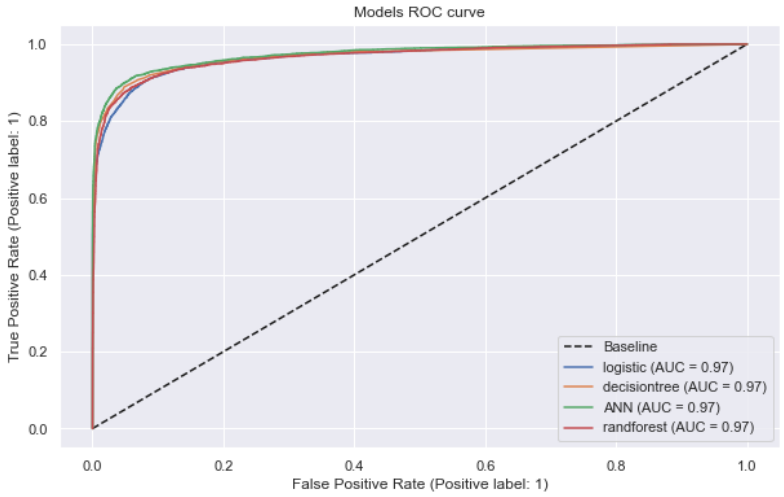
- **Confusion matrix**



Random Forest's performance		
Specificity	TP/(TP+FP)	0.9899
Sensitivity	TP/(TP+FN)	0.7421
Misclassification rate	1 - (TP+TN)/Observations	0.0684

8. Model comparison & best model selection

As a final step in this report, we propose to carry out a comparison of the implemented classification models. One popular tool to compare models' performance is the **Receiver Operating Characteristic (ROC) curve**. It evaluates the performance of a classifier by *plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity)*. Moreover, it also represents a baseline, which is the ROC curve of a purely random classifier and for which we hope our model to be as far as possible from it. Then to classify the classifier, we looked at Area Under the Curve (AUC). The best one should have the highest AUC score.



After having drawn the ROC curve of each model and computed their respective AUC score, we can state that they all have almost a comparable performance since their AUC are extremely close in value. However, since we are mostly focused on having a good sensitivity rate and a low false negative rate, the neural network with dimension reduction is perhaps the best model to rely on because on average only 3.5% of the late flights are

not classified as such. Not only has this model the best false negative, but also an excellent misclassification rate. On the contrary, random forest is certainly the worst model for our problem since it does not detect 26% of the planes actually late, even if it has an acceptable misclassification rate. Finally, we should not forget to mention that the decision tree also merits honors, because it also has a very low false negative rate.

9. Final results and interpretation

As detailed in the previous section, the predictive method that we advised to use is an artificial neural network based on data with dimension reduction, or a decision tree.

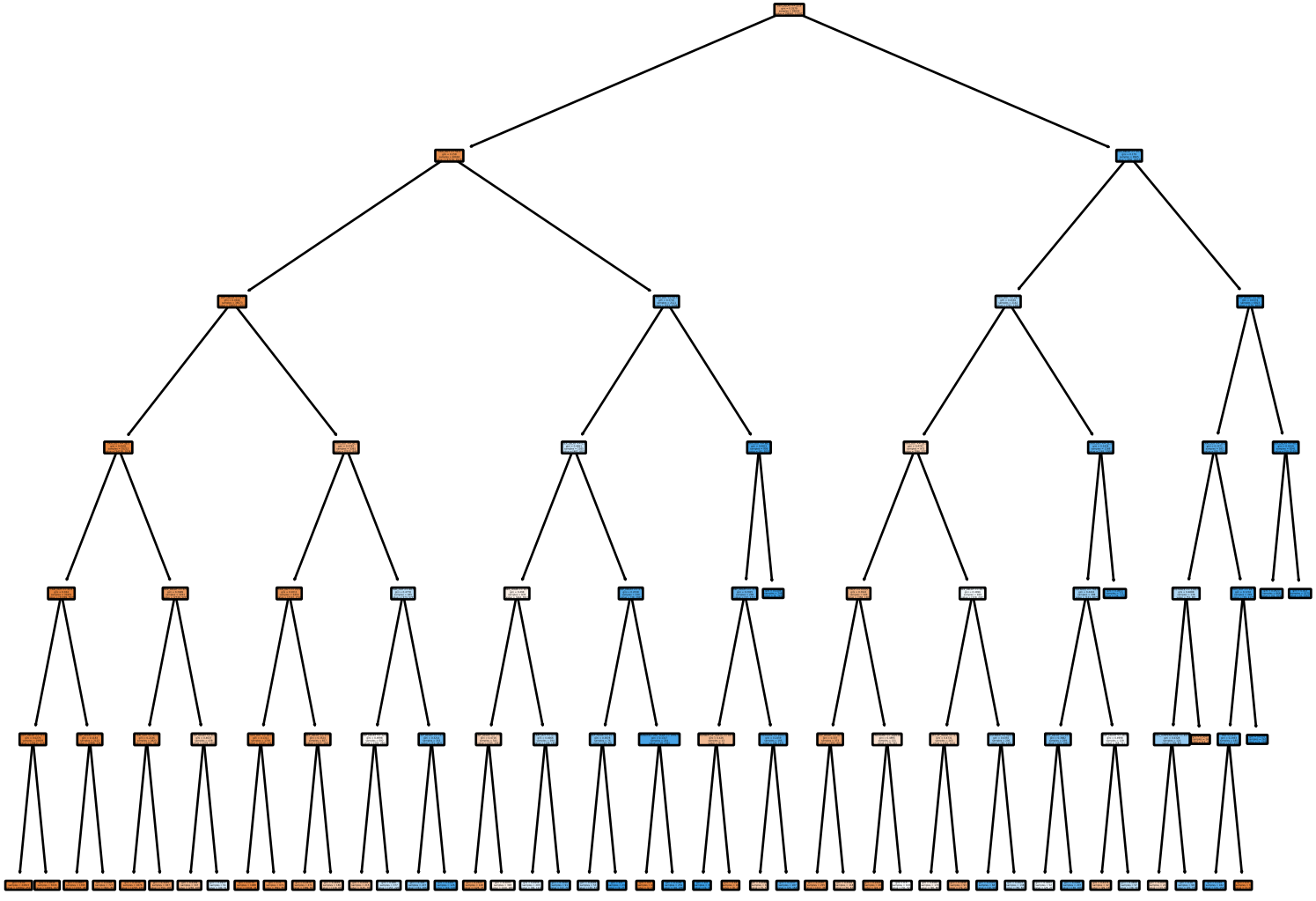
Based on these models, we could really improve the service quality for the customer, or the operation efficiency of the airport and of the carrier operators. As soon as the weather forecasts are available, delays predictions can be carried out several days before the departure so that a potential delay can be handled in the best manner. For example, schedules can be adjusted, the workforce size can be enlarged during congesting times or cut during breathing space. All these good examples make the predictive models of this kind extremely valuable for all companies having stakes and links with aircraft transportation.

Moreover, thanks to this data analysis, we were able to identify some clear relationships between airline delays and certain variables. In summary, we saw that:

- Holidays months were clearly bad months for delays.
- Fog and heavy rain were real hindrances for smooth airline traffic. More generally, flights are completely dependent on weather.
- Some air carriers can also be blamed for causing more delays than others.
- Total delay per hour is rising throughout the day.
- Business weekdays tend to be more congested.
- Some destinations generate more delays than others.

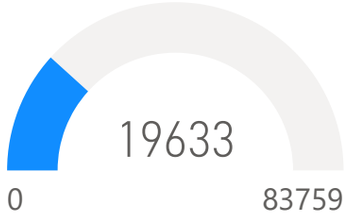
Yet, our model is only an early result. It can be, and should be improved. Many potential predictors can still be found to add predictive power to the model.

10. Appendixes



Data Report of "LGA_JulyDecember" dataset

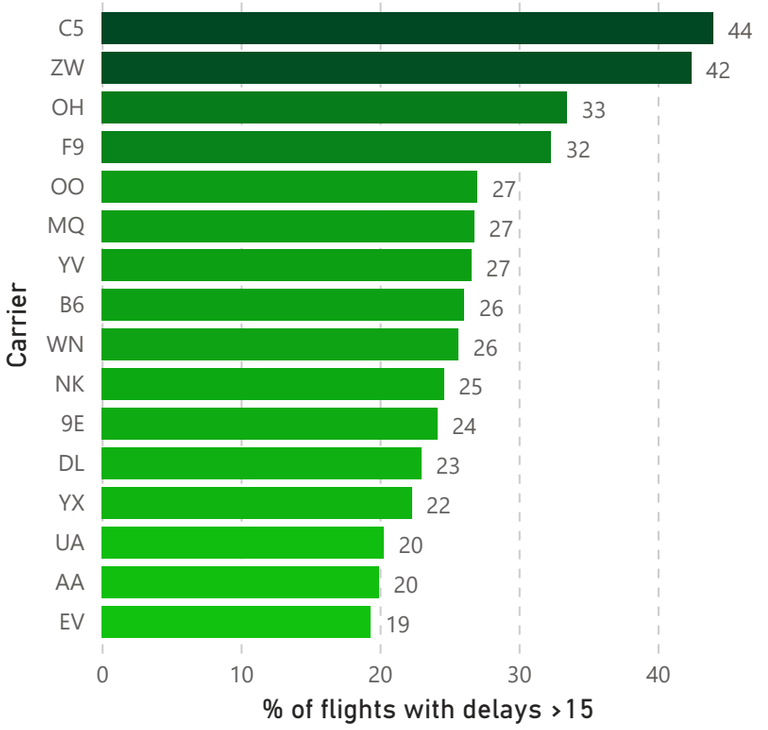
Nb of flights with arrival delay > 15



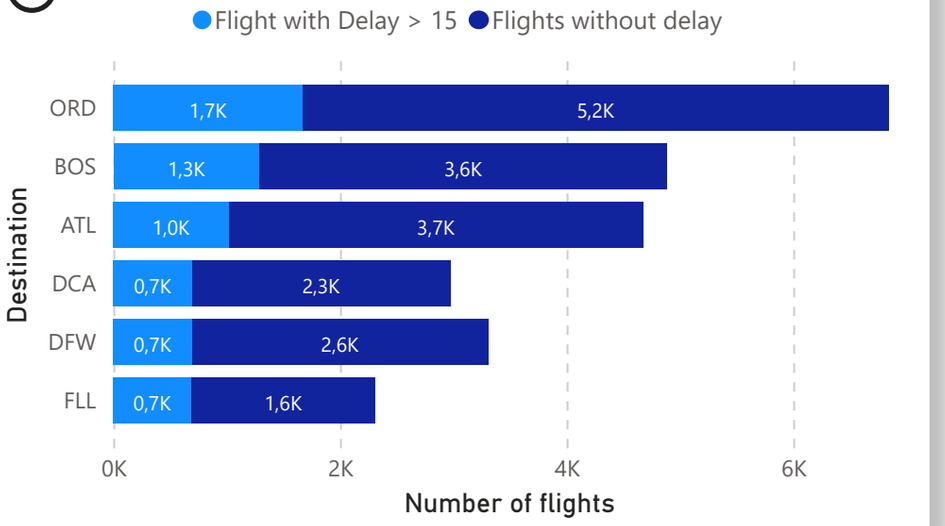
% of flights with arrival delay > 15

23,44

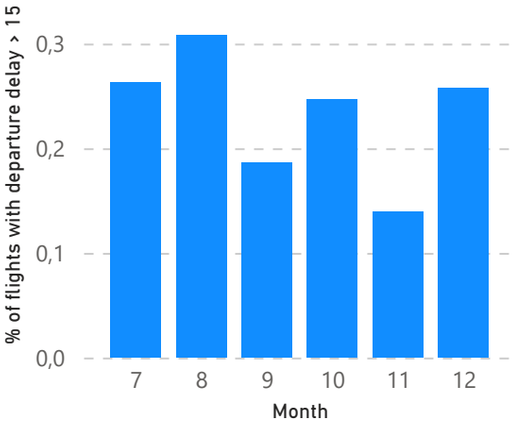
1 % of flights with delays for each carrier



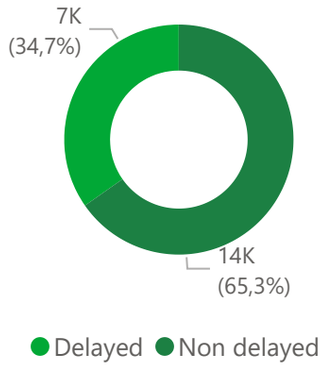
2 Top 6 Destination with most delays



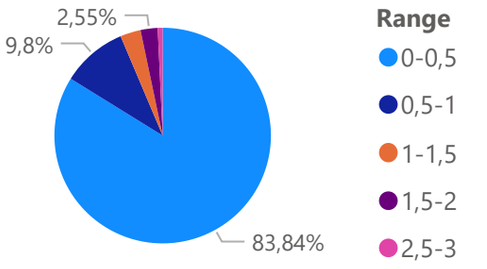
3 % of flights with Arrival delay > 15 per Month



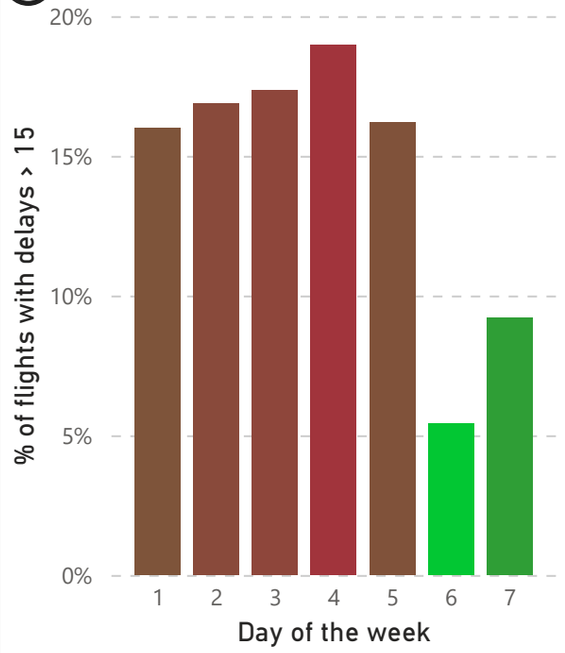
4 % of Arrival delays vs WT01 (Fog, ice fog, freezing fog)



5 Arrival delay > 15 w.r.t. day precipitation in inches

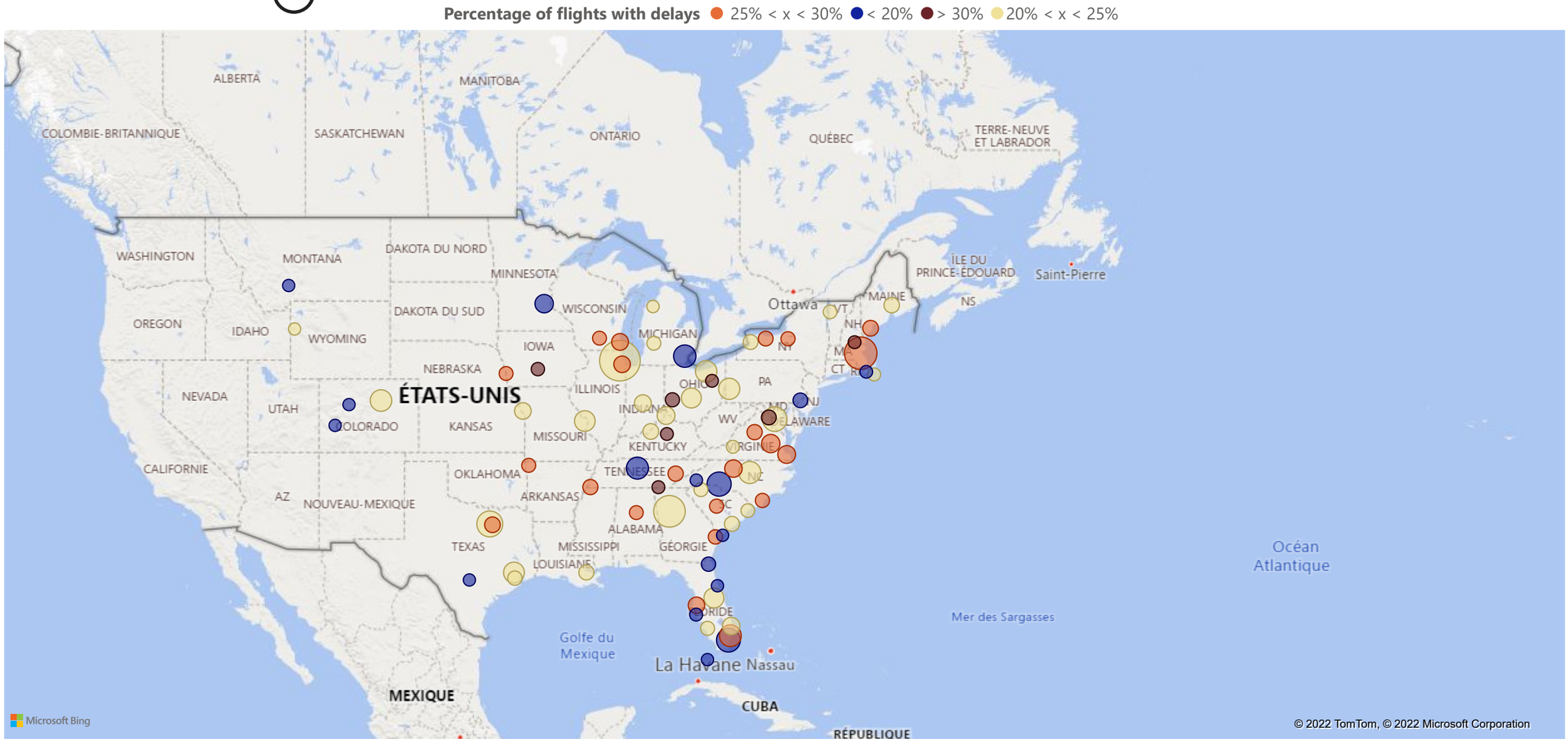


6 % of flights with delay > 15 per day of the week



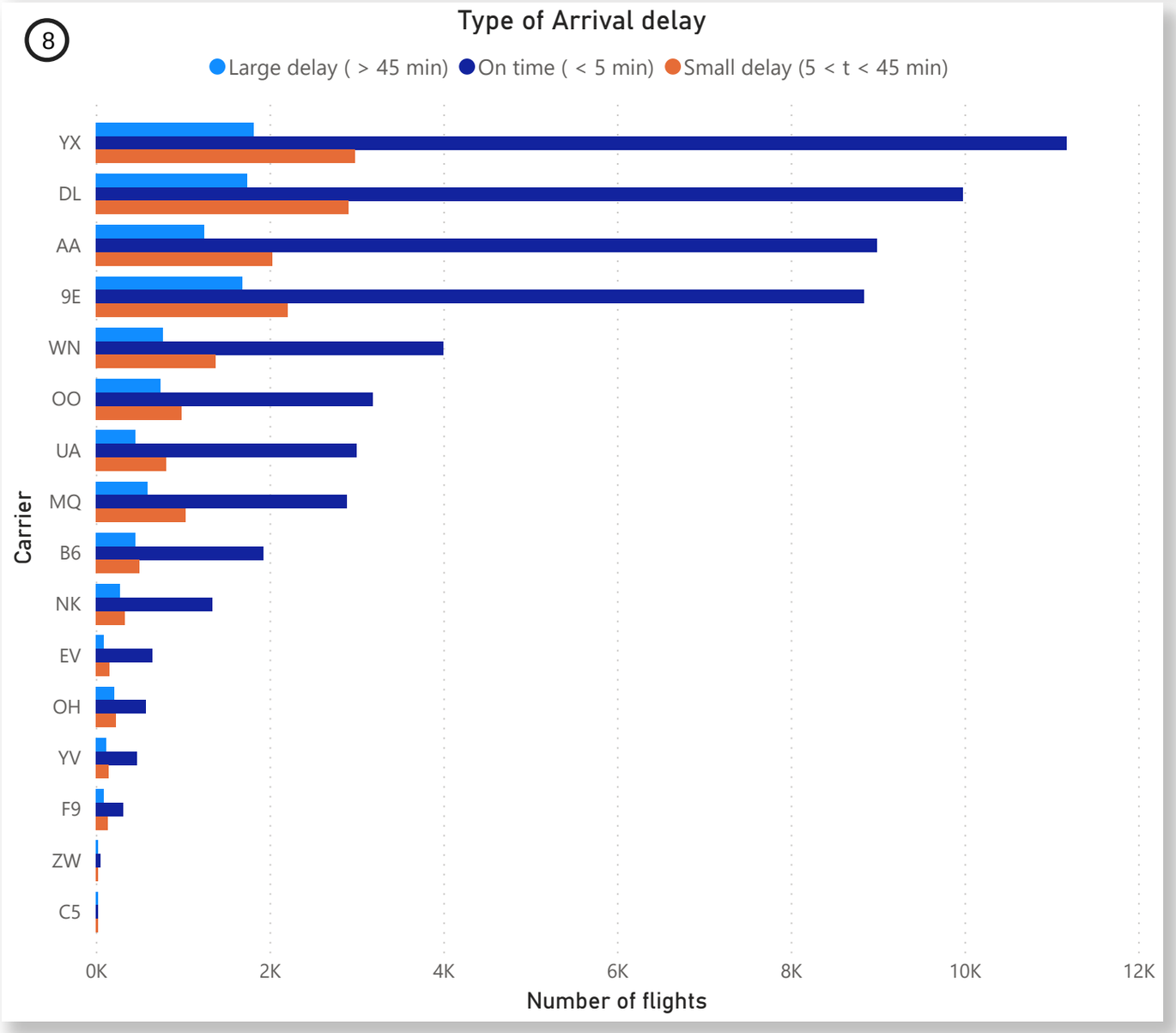
Data Report of "LGA_JulyDecember" dataset

7 Destinations w.r.t. the number of flights (bubble size) and the percentage of delays (bubble color)

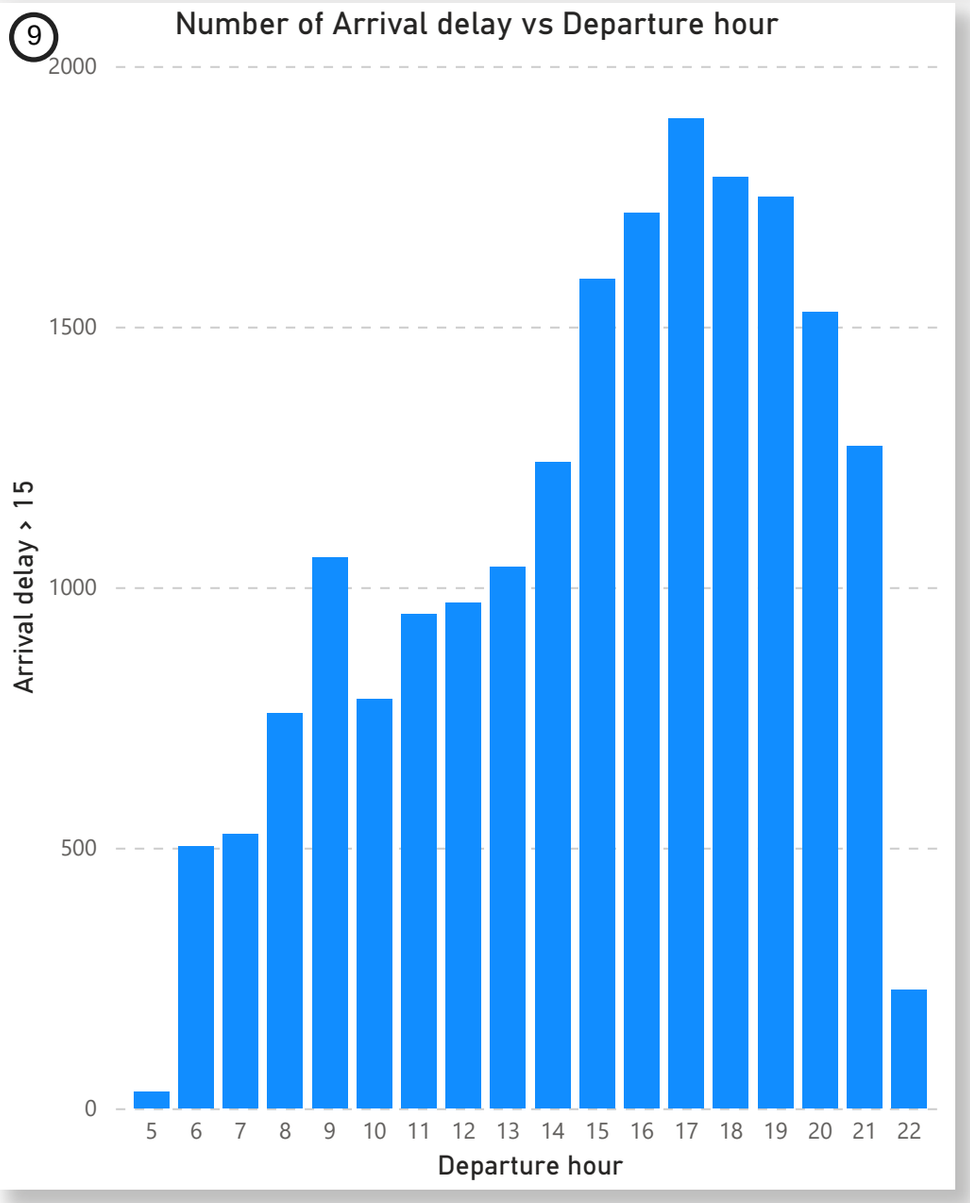


Data Report of "LGA_JulyDecember" dataset

8



9



11. Bibliography

References

Datasets | Climate Data Online (CDO) | National Climatic Data Center (NCDC). (2022). National Centers for Environmental Information. Retrieved May 13, 2022, from https://www.ncdc.noaa.gov/cdo-web/datasets

Géron, A. (2019). *Hands-On Machine Learning with Scikit-learn, Keras & TensorFlow* (2nd edition ed.). O'Reilly.

OST_R | BTS | Transtats. (2022). OST_R | BTS | Transtats. Retrieved May 10, 2022, from https://www.transtats.bts.gov/Tables.asp?QO_VQ=EGI&QO_anzr=Nv4%FDPn44vr4%FDSv0n0pvny%FDer21465%FD%FLS14z%FDHE%FDSv0n0pvny%FDQn6n%FM&QO_fu146_anzr=Nv4%FDPn44vr4%FDSv0n0pvny

Predicting flight delays [Tutorial]. (n.d.). Kaggle. Retrieved May 20, 2022, from https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial/notebook