Accuracy in third flight*

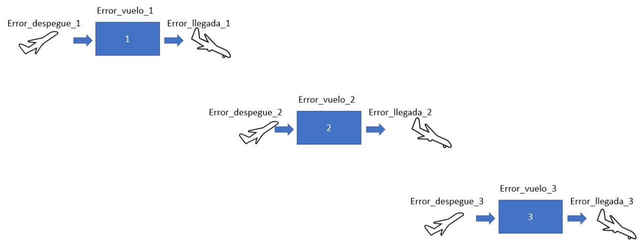
Jose Maria Matas Larrucea
Barcelona,Spain
jmmatas08@gmail.com

***Abstract*—Punctuality and price are the two things that passengers value the most when evaluating an Air company. In this case, We have to figure out what the timeliness of an airline is in what they call the third jump, using a database provided by them.We will also show that the delay does not depend on other variables than their own delay departure and error in the flight time of this particular time. And we will demonstrate that the process of delay is not capable and work only with the average giving a false sensation that the process is ok. We are going to start defining the meaning of Error, which will be the difference between the estimated time versus the real-time and it will be positive if there is a delay, and negative if the time is shorter than was planned**

***Index Terms*—Data Science; Delay; logistic regression; Machine Learning**

I. Introduction

We will use 2 database (BBDDs), both are similar and has the information of estimated times planned and the real-time but from 2 different airlines. This will help us to do benchmarking Database has several columns with the dates of takeoff and landing of each aircraft. A plane in one day makes several ”jumps”.



As example, here you can see three jumps an aeroplane will do in a day.

First We have to adapt the time fields to a date format to be able to do the time calculations. We have to identify flights of each aircraft per day, sort them by departure time to be able to identify each jump and create a sequence. This will force you to work in a line format so that the calculated time of each jump is on a single line and then you can filter the file to have all the information of one aircraft per day, the three jumps. With this, we will reduce the calculation time by simplifying the size of the file by 80 % accelerating the calculation process. We weigh from 500,000 rows to 80,000. Once the file is assembled we can make the main calculations and be able to carry out capacity studies, calculating the Cp and Cpk and obtaining the value of % of flights whose punctuality is greater than 0 or 10 minutes.

Definitions of Cp and Cpk:

Both Cp (1) and Cpk (2) are Indexes of the Potential Capacity of a process and are shown on a histogram that collects the availability of offering the benefits required of it.

Formulas are:

$$Cp = \frac{(USL - LSl)}{6\sigma} \tag{1}$$

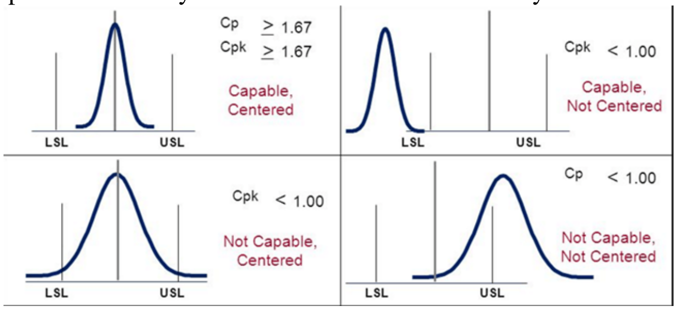
Where USL is the Upper Specified limit, and LSL is Lower Specified Limit. In this case, we will say that arrive on time when the Error is higher than -30 minutes and the Error is lower than 10 minutes. We know that in a pure Gauss curve we can say 6*σ is 99.7% of the total population, then we want to count how many curves you can put in between both limits, and as much the number is higher, the meaning the process is better controlled. We will use Cpk to quantify if the process is centred. Equations for Cpk are:

$$CpU = \frac{()USL - \bar{x})}{3\sigma}$$

$$CpL = \frac{(\bar{x} - LSL)}{3\sigma}$$

$$Cpk = Min(CpU, CpL) \tag{2}$$

We will choose the minimum value, which means the centre is nearest to this limit. Of course, never the Cpk will be higher than Cp, for it, it is so important first to minimise the process variation. These indexes are part of tools for statistical control of processes widely used in the automotive industry.



We have to figure out if the takeoff data landing, boarding time, and flight time are related between jumps. We will create a logistic regression where the output is ’0’ that the plane arrives on time or ’1’ which arrives 10 minutes late than the scheduled time.

Identify applicable funding agency here. If none, delete this.

The database is very unbalanced [] because 9% of the aircraft do not arrive on time so if we want to have a credible result it is enough to swing. We will use the method resample, from sklearn.utils, to have a result of more or less than 50% for 0 and 1.

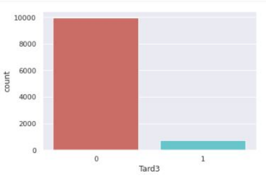


Fig. 1: Unbalanced

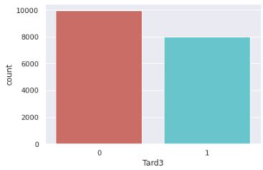


Fig. 2: Balanced

II. State of Art

A. Maintaining the Integrity of the Specifications

There are similar DDBBs to this. Although the information is very similar to other ones that you can find on the internet. But these have pretty confidential data from 2022 and 2019. Surely companies do internal analysis to find out their behaviour with punctuality, but I'll try to get a little deeper and with updated in recent months, by comparing punctuality with another company, for example.

Mainly the idea is to know if the punctuality they have is due to a reason and find out what that reason is.

Aviation databases in Europe are very closed, while in the USA can be consulted easily, and almost online.

B. Abbreviations and Acronyms

Here are the different columns we are going to work and their descriptions.

III. Methodology:

We will use the dataBase

$$VFLT2022$$

as main file We will divide the file into train data and test data at 80/20% respectively. We will create a logistic regression model.

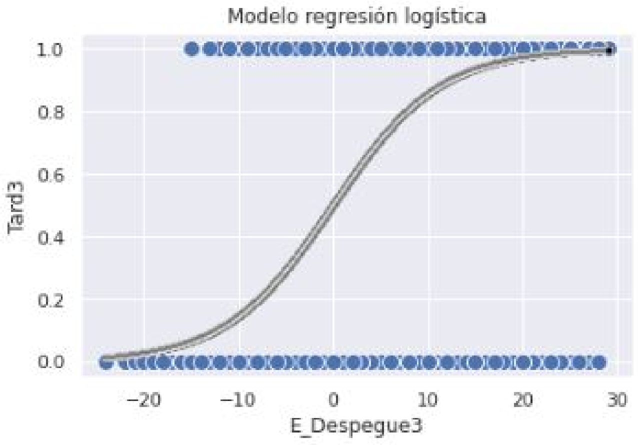
Multiple logistic regression, developed by David Cox in 1958, is an extension of simple logistic regression. It is a regression method that allows estimating the probability of a binary qualitative variable as a function of a quantitative variable. One of the main applications of logistic regression is binary classification, in which observations are classified into one group or another depending on the value taken by the variable used as a predictor. Predictors can be both continuous and categorical.

Why logistical and nonlinear regression? If a qualitative variable with two levels is encoded as 1 and 0, mathematically it is possible to adjust a linear regression model for least squares 0+1x. The problem with this approximation is that, since it is a line, for extreme values of the predictor, values of Y less than 0 or greater than 1 are obtained, which contradicts the fact that the probabilities are always within the range [0,1].

DATE	Date
FLT	Flight
REG	Register
AC	aircraft type
DEP	Departure Airport Code
ARR	Arrival Airport Code
STD	Estimated departure time
STA	Estimated time of arrival
TKof	Take-off time at the runway threshold
TDwn	H. landing on the pita
ATD	actual departure time
ATA	actual arrival time
BLOCK	estimated flight time
FLThr	actual flight time
ACT PAX	Number of passengers on the plane
Taxi-out	Departure taxi time
Taxi-In	Arrival taxi time
SLOT	H. theoretical by the controller
C1	Reject Code 1
DLY1	minutes with that code
Sub1	criticality
C2	Rejection Code2
DLY2	minutes with that code
Sub2	criticality
C3	Reject Code 3
DLY3	minutes with that code
Sub3	criticality
C4	Reject Code 4
DLY4	minutes with that code
Sub4	criticality
Close Pax Door	H. passenger door closing
Close Cargo Door	H. aircraft hold closure
Open Cargo/Pax Door	H. passenger door opening
close pax door2	H. airplane hold opening

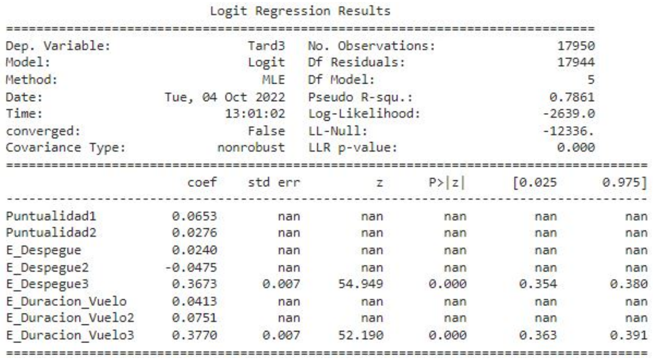
In this case, linear regression could have been used but we have as an output variable a continuous value since a plane arrives late regardless of the delay time if it exceeds 10 minutes it is better to see that predictors are the most influential.

To demonstrate that this is a logistic regression out we are going to plot the relation between time to delay takeoff versus if the flight will arrive late o not



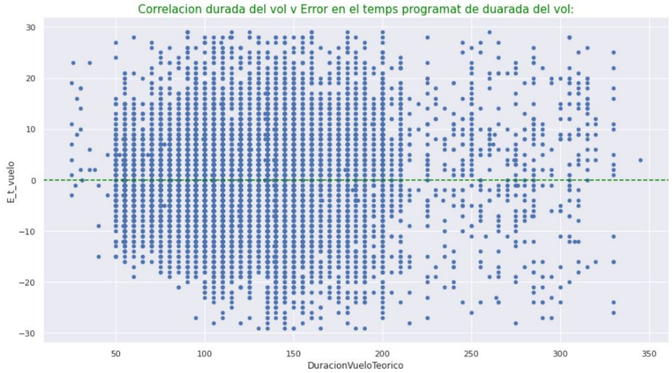
IV. Results:

As a result, we see that the third jump is influenced solely by its takeoff time, its flight duration and not time on land. It does not affect the delays of the previous jumps and the reason is that between jumping and jumping, excess time is calculated to perform a 'spring' function that prevents the noise that can influence one flight/jump from affecting the next.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  Tard3   No. Observations:                17950
Model:                          Logit   Df Residuals:                    17944
Method:                           MLE   Df Model:                            5
Date:                Tue, 04 Oct 2022   Pseudo R-squ.:                  0.7861
Time:                        13:01:02   Log-Likelihood:                -2639.0
converged:                      False   LL-Null:                       -12336.
Covariance Type:            nonrobust   LLR p-value:                     0.000
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Puntualidad1          0.0653        nan        nan        nan         nan         nan
Puntualidad2          0.0276        nan        nan        nan         nan         nan
E_Despegue            0.0240        nan        nan        nan         nan         nan
E_Despegue2          -0.0475        nan        nan        nan         nan         nan
E_Despegue3           0.3673      0.007     54.949      0.000       0.354       0.380
E_Duracion_Vuelo      0.0413        nan        nan        nan         nan         nan
E_Duracion_Vuelo2     0.0751        nan        nan        nan         nan         nan
E_Duracion_Vuelo3     0.3770      0.007     52.190      0.000       0.363       0.391
=====================================================================================
```

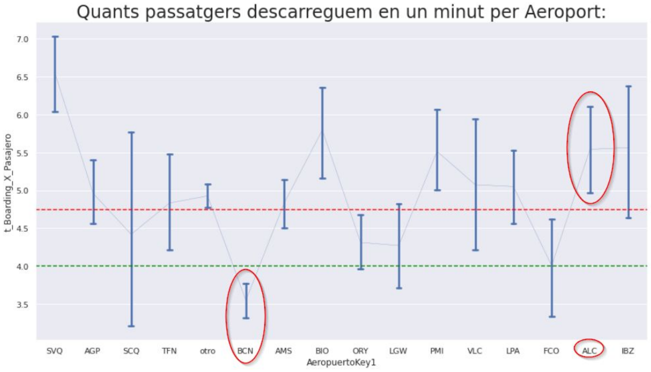
We can see that the influence of the pilot to reduce delay on the flight is difficult because in case he can, he will reduce as much as possible in long flights and we can see that this is not true. In this graph we can see that there is not a correlation between how long the flight is and the error to arrive.



It is important to say that this assertion is reached when an all-time greater than 30 minutes is eliminated because 30 minutes we can consider that it is a time within normal and that if a plane leaves later it is for a reason that will influence the flight of the whole day. Examples, are breakdown planes, controllers strike, a closed airport, etc, and these outliers are out of the studio.Remove these outliers was approved by the experts in aviation analysis and only represents less than 3% of the total amount of fligths. Note: When a flight is delayed more than the "slot" approved in the flight plan, then this plane must wait till will be a gap that let the ATC give permission to departure.

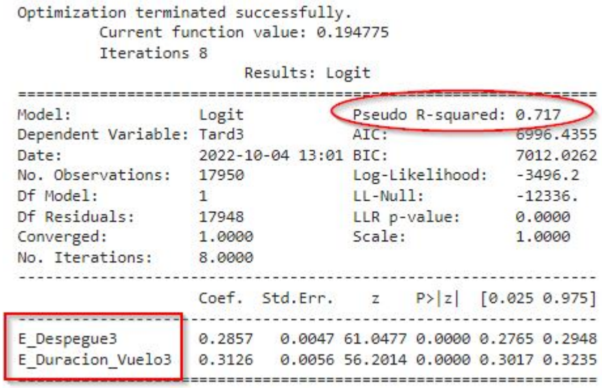
During onboarding time, we can see that the size of the airport and its infrastructure to help the passengers transit and the number of doors to proceed to aircraft, affect a lot in this time. This graph we can evidence that for ALC (Alicante) that boarding is for 2 doors is faster because can board more people per minute (5,5 passengers per minute) than Barcelona for example, which is 3,5. That represents that can do it in half of the time boarding the same plane.

The solution taken is to give more time in this kind of huge airports and in function of the plane size.



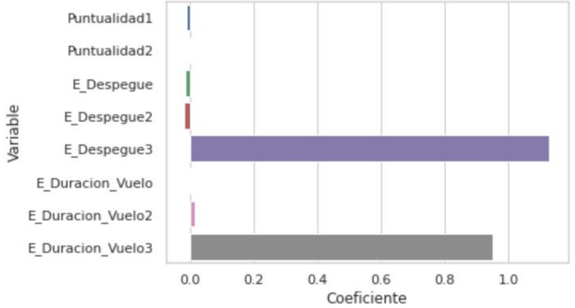
V. Conclusion:

- Between jumps there is a spring that cushions the impact between jump delays. Only the previous Error in take-off and Error in how long the flight is will affect 70% if the time to arrive. No, previous jumps affect.



```
Optimization terminated successfully.
         Current function value: 0.194775
         Iterations 8
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.717
Dependent Variable: Tard3            AIC:              6996.4355
Date:               2022-10-04 13:01 BIC:              7012.0262
No. Observations:   17950            Log-Likelihood:   -3496.2
Df Model:           1                LL-Null:          -12336.
Df Residuals:       17948            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     8.0000
-----------------------------------------------------------------
                    Coef.  Std.Err.    z    P>|z|  [0.025 0.975]
-----------------------------------------------------------------
E_Despegue3         0.2857   0.0047 61.0477 0.0000 0.2765 0.2948
E_Duracion_Vuelo3   0.3126   0.0056 56.2014 0.0000 0.3017 0.3235
=================================================================
```

With all factors from previous jumps, and with data normalizes, we can see that the variables from actual flight (E_DespegueE_3 and E_Duracion_Vuelo3) are in proportion more important as you can see in this graph:



- Having given more boarding time at large airports like Barcelona, helps to minimize the impact of delays.
- The concept of punctuality should change and it is not to arrive late 10 minutes but to leave on time. This is demonstrated by the first jump because it is the jump with fewer exogenous elements to leave late and we see that 31% of the planes are already leaving late. Figura 3 In the rest of flights, there are already more exogenous elements such as the previous ones, more passengers who

arrive late for the transfer of another flight, a passenger who does not get on the plane and is the suitcase that is to get off, or people who do not respect the size of the suitcases, climbing large suitcases in the cabin, which greatly delays the boarding time.

- Benchmarking with other companies is a very good technique because here we saw that the behaviour in the first jump as referenced by the Barcelona airport is better for the competition.... And if they can (competitors), why not this company?.
- A deep LEAN process analysis is recommended to apply in the boarding process and identify opportunities for improvement. For example, minimising the passenger amount that carries with the luggage in, can minimize the time necessary for boarding. It is faster to board in the taxi way than by finger. Or involve as in Japan the passengers to keep clean the cabin and remove garbage by them.
- **Conclusion:** The aviation world requires great coordination, as well as a choreography full of variables that affect this coordination, and in which all the elements that exist.... Airplane, breakdowns, the weather, the air traffic controllers (ATC), the passengers, the ground staff, the one who manages the bags, etc... Everyone must participate in the "aeronautical philharmonic" playing rhythmically, because the moment an element goes out of tune, the delay occurs. And starting with the first jump, accuracy in closing gates before time departure could be the key and take the habit

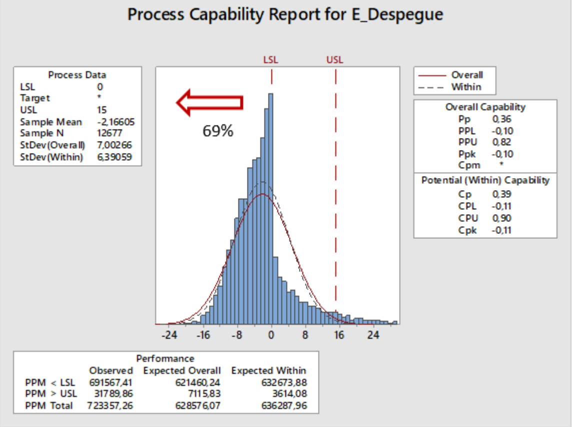


Fig. 3: Error in the first take off

Acknowledgment

I want to thank Aina Palacios for her patience in my way of being and for this course. And I also want to thank the IT Academy for this course because it has allowed me to take a huge leap forward in my knowledge of Data Science. When I see myself one year ago, and the knowledge I have now about Data Science I can say I achieved the level I wanted. Thank you!!!

References

[1] Capacidad del Proceso (Cp – Cpk), link

[2] Regresión logística con Python by Joaquín Amat Rodrigo, available under a Attribution 4.0 International (CC BY 4.0) at link

[3] How to write a great looking research article using LaTeX on Overleaf, by Data Professor link

[4] Train-Test Split for Evaluating Machine Learning Algorithms, by Jason Brownlee link

[5] Cross-validation: evaluating estimator performance, link