

# Airline Departure Data Analysis and Regression

Lucchi Manuele & Tricella Davide

August 31, 2022

Instructors: Professor CESA-BIANCHI & Professor MALCHIODI

*We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.*

## Contents

<b>1</b>	<b>Definitions</b>	<b>2</b>
<b>2</b>	<b>Dataset</b>	<b>2</b>
<b>3</b>	<b>Preprocessing Techniques</b>	<b>3</b>
3.1	Algorithms and Techniques . . . . .	3
3.2	Parallelization . . . . .	3
<b>4</b>	<b>Model</b>	<b>3</b>
4.1	Parameters initialization . . . . .	3
4.2	Iterations . . . . .	3
4.3	Algorithm . . . . .	3
4.4	. . . . .	3
<b>5</b>	<b>Performances</b>	<b>4</b>
<b>6</b>	<b>Experiments</b>	<b>4</b>
6.1	Canceled Flights . . . . .	4
6.2	Diverted Flights . . . . .	4

<b>7 Results and Conclusions</b>	<b>4</b>
7.1 Space and Time . . . . .	4

### Abstract

The purpose of this paper is to evaluate the usage of a Logistic Regression model on a airlines dataset to predict flight cancellation or diversion, in a scalable and time/space efficient implementation.

## 1 Definitions

**Label**

**Model**

## 2 Dataset

The initial dataset, [Airline Delay and Cancellation Data] CITAZIONE is made of 9 years of airlines flights data, composed by 10 files (one for each year from 2009 to 2018) of around 6 millions records each. The files presents 28 columns, of which we only took the 9 more relevant

**FL\_DATE** The flight date.

**OP\_CARRIER** The carrier code.

**ORIGIN** The departure place.

**DEST** The destination place.

**CRS\_DEP\_TIME** The.

**CRS\_ARR\_TIME** The.

**CANCELLED** If the flight has been canceled.

**DIVERTED** If the flight has been diverted.

**CRS\_ELAPSED\_TIME**

**DISTANCE** The distance the flight has to cover.

In the case the prediction is about the cancellation, the DIVERTED column will be ignored, while if the prediction is on if the flight would be diverted or not, the CANCELLED column will be ignored.

The carrier code is a two characters alphanumeric code, the origin and destination places are a three characters alphanumeric code.

Flight date, departure time and arrival time are dates, while the elapsed time and the distance are real numbers.

Cancelled and diverted are either 0 or 1.

One million of records equally distributed between the files were taken to perform the training.

## 3 Preprocessing Techniques

### 3.1 Algorithms and Techniques

Multiple preprocessing techniques were used.

First, the dataset has been balanced in regard of the evaluated property, be it being canceled or diverted, so that there are an equal number of uniformly drawn positives and negatives. MIGLIORARE

Then the data not already represented as real numbers has been converted; places and carriers, that were alphanumeric codes, had a number assigned based on the code, dates were splitted between the year and the rest, with the latter being hashed MIGLIORARE.

The data (now completely composed of real numbers) was then normalized between 0 and 1, to avoid exploding values.

Lastly, the data was splitted between the training set (75%) and the test set (25%).

DIVISO PER MEDIA E VARIANZA???

### 3.2 Parallelization

## 4 Model

### 4.1 Parameters initialization

Parameters such as Weights and Bias are initialized using a uniform distribution between 0 and 1, with the first one having the same length as the number of columns and the second being a scalar value.

### 4.2 Iterations

### 4.3 Algorithm

differenze con sklearn

### 4.4

- descrizione inizializzazione parametri - descrizione batching e iterazioni - descrizione forward propagation - descrizione loss - descrizione calcolo gradiente - descrizione update

## 5 Performances

- come scalano le operazioni di numpy effettuate

## 6 Experiments

### 6.1 Canceled Flights

### 6.2 Diverted Flights

Iterations	LR	L2	Loss
100	0	0	0
500	0	0	0
1000	0	0	0

- nostro modello dopo v1 iterazioni con learning rate v1 - nostro modello dopo v1 iterazioni con learning rate v2 - nostro modello dopo v2 iterazioni con learning rate v1 - nostro modello dopo v2 iterazioni con learning rate v2 - confronto modello di sklearn con i valori migliori

## 7 Results and Conclusions

### 7.1 Space and Time