UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

London School of Economics

Data Science Report

Ibtihel Frini
02/04/2024

Table of Contents

PART 1

Part 1(a): Implementation of Random Walk Metropolis Algorithm for Sample Generation

The Random Walk Metropolis algorithm is a Markov chain Monte Carlo (MCMC) method used for generating samples from a probability distribution that may be difficult to sample from directly. In this report, we implement the Random Walk Metropolis algorithm to generate samples from a probability distribution given by equation: $f(x) = \frac{1}{2}e^{-|x|}$. Figure 1 illustrates the distribution obtained from sampling N=10000 data points and test size s=1, along with the kernel density plot with Sample Mean = 0.09092665049916825

and Sample Standard Deviation =

1.387768141761254

Part 1(b): Analysis of Convergence

In order to assess the convergence of the Metropolis-Hastings algorithm, we generated J=4 chain, we generated N=2000 samples and tested the convergence for different values of standard deviation ranging from s=0.001 until s=1. We then calculated and plotted R_hat values for each s value as outlined in figure 2. As may be noticed, the curve gradually decreases and stabilizes near a value close to 1 as s rises. This observation suggests that the algorithm effectively converges towards the same stationary distribution.

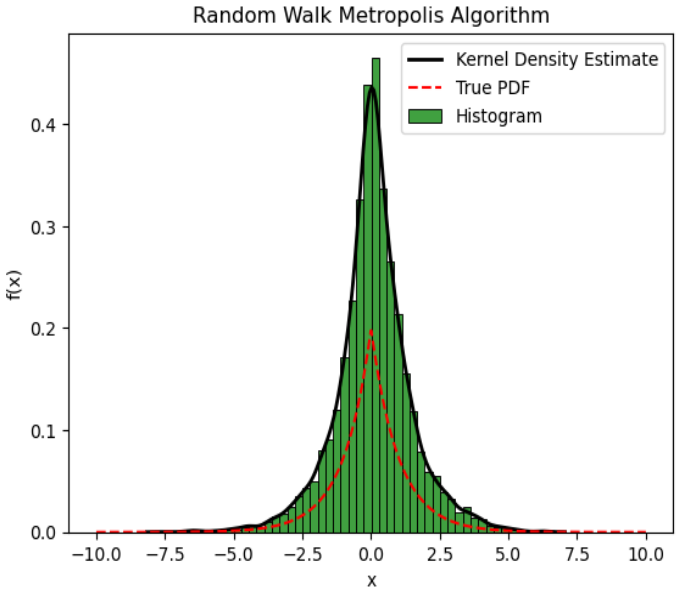

Figure 1: Random Walk Metropolis Algorithm

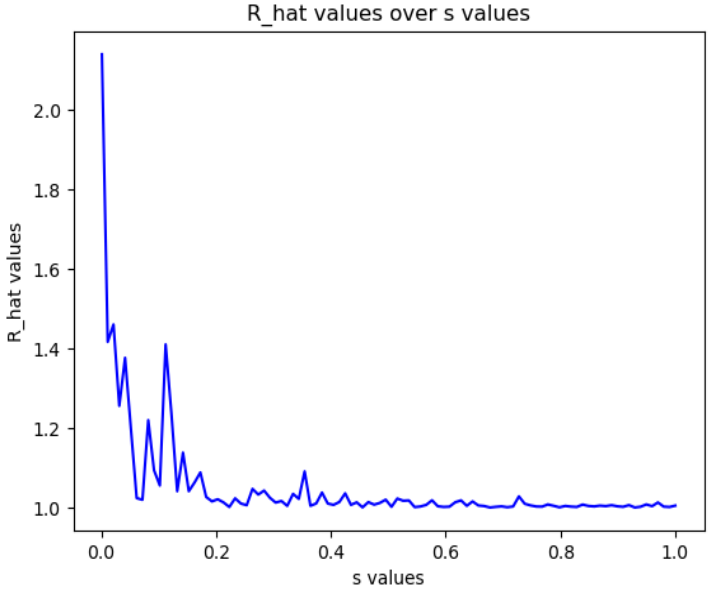

Figure 2: R_hat values over s_values

PART 2
Introduction

In this report, we will analyze historical flight data from the years 1999 to 2008 [1]. The data contains various attributes related to flight details such as departure and arrival times, delays, distances, and cancellation information. Our goal is to explore the data, perform descriptive analysis, and gain insights into factors affecting flight delays and cancellations.

1. Importing data and required Libraries

We start by importing the necessary Python libraries, for data processing, modeling, and visualization. These include Pandas for data structure manipulation, NumPy for numerical computing, scikit-learn for modeling machine learning algorithms and adjunct operations (example: splitting data), and Matplotlib (more particularily the pyplot module) , and Seaborn for data visualization.

1.1. Reading Data

Next, we read the historical flight data for each year from 1999 to 2008 into separate dataframes using the method **read_table() of pandas.**

2. Data Analysis

Throughout this study, we thoroughly analyze historical flight data in an effort to extract insights that can assist airlines in streamlining their processes, increasing productivity, and offering passengers better services. More precisely, the analysis is focused on figuring out what causes flight cancellations and delays and developing a machine learning prediction model to foresee these problems.

Having explored the historical flight data, we aim to answer the following questions:

Best Day and Time that minimizes Delays

2.1. Identifying the best times and days of the week to minimize delays each year

We examined the historical flight data for every year to identify the optimal hours and days of the week to reduce delays. We determined the precise day of the week (DOW) and time of day (TOW) with the least amount of average delay by examining both arrival and departure times. For that purpose, we used a number of methods such as .query() method

Table 1 provides a summary of the acquired results. Consequently, it may be said that Saturday is typically the greatest day of the week to reduce airline delays, however this

fluctuates from year to year. The ideal times of day to arrive and depart also vary, although they usually fall in the early morning hours.
Through careful consideration of these facts, airlines may optimize resource allocation and schedule adjustments to reduce delays and enhance passenger satisfaction.

Table 1: Days and times for which minimal delays are recorded from 1999 to 2008

Best DOW and TOW to minimize delays				
Year	Day		Time	
	Arrival	Departure	Arrival	Departure
1999	Tuesday		07:55	06:56
2000	Saturday	Wednesday	07:52	05:56
2001	Saturday		07:01	05:55
2002	Saturday	Tuesday	09:48	06:55
2003	Saturday		07:30	06:55
2004	Wednesday		07:50	05:55
2005	Tuesday		08:05	05:55
2006	Wednesday		07:55	05:55
2007	Saturday		08:00	05:55
2008	Saturday		07:21	05:55

2.2.Impact of the age of planes on delays

In this section, we evaluate whether older planes suffer more delays on a year-to-year basis. For that purpose, we analyze the existence of any between the age of the aircraft and the frequency of delays.

To evaluate the impact of aircraft age on delays, we used the following methodology:

Step 1: Define a function to categorize planes as "Older" or "Younger" based on a prespecified threshold age.

Step 2: Set a threshold age of 30 years to classify planes.

Step 3: Create a new column indicating whether each plane is older or younger. To do this a categorization function, labelled: **categorize_plane_age()** is implemented

Step 4: Group the data by year and age category and calculate the average delay for each group.

Step 5: Calculate the difference in average delay between older and younger planes for each year.

The results thus obtained are given in table 2. Overall, the data suggests that older planes tend to have fewer delays, particularly in terms of arrival delays. However, the difference in delay frequency between older and younger planes is not consistent across all years for both arrival and departure delays. Airlines should consider both the age and maintenance of their aircraft fleet to effectively manage and reduce delays. Since 1999 is the first year, there is no previous year's data used to compare with, thus the NaN values notation

Table 2: impact of older planes on delays

Year	ArrDelay	DepDelay
	Older Planes	Younger Planes
1999	NaN	NaN
2000	-3.157557	-0.557287
2001	-2.208560	-0.747054
2002	9.382446	-4.507209
2003	-8.507993	1.014182
2004	4.654698	2.965372
2005	0.095664	1.436713
2006	-1.603769	-2.040260
2007	10.009732	4.549437
2008	-2.939007	0.469617

3. Investigating features that influence flight diversion

3.1.Selecting predictive features

In this part, the logistic regression model is used to analyze flight diversion. For that purpose, we started by pre-selecting an initial set of potential features based on their correlation with this target variable

The correlation matrix was generated using the Pearson correlation coefficient, for which we provide an illustration in figure 3.

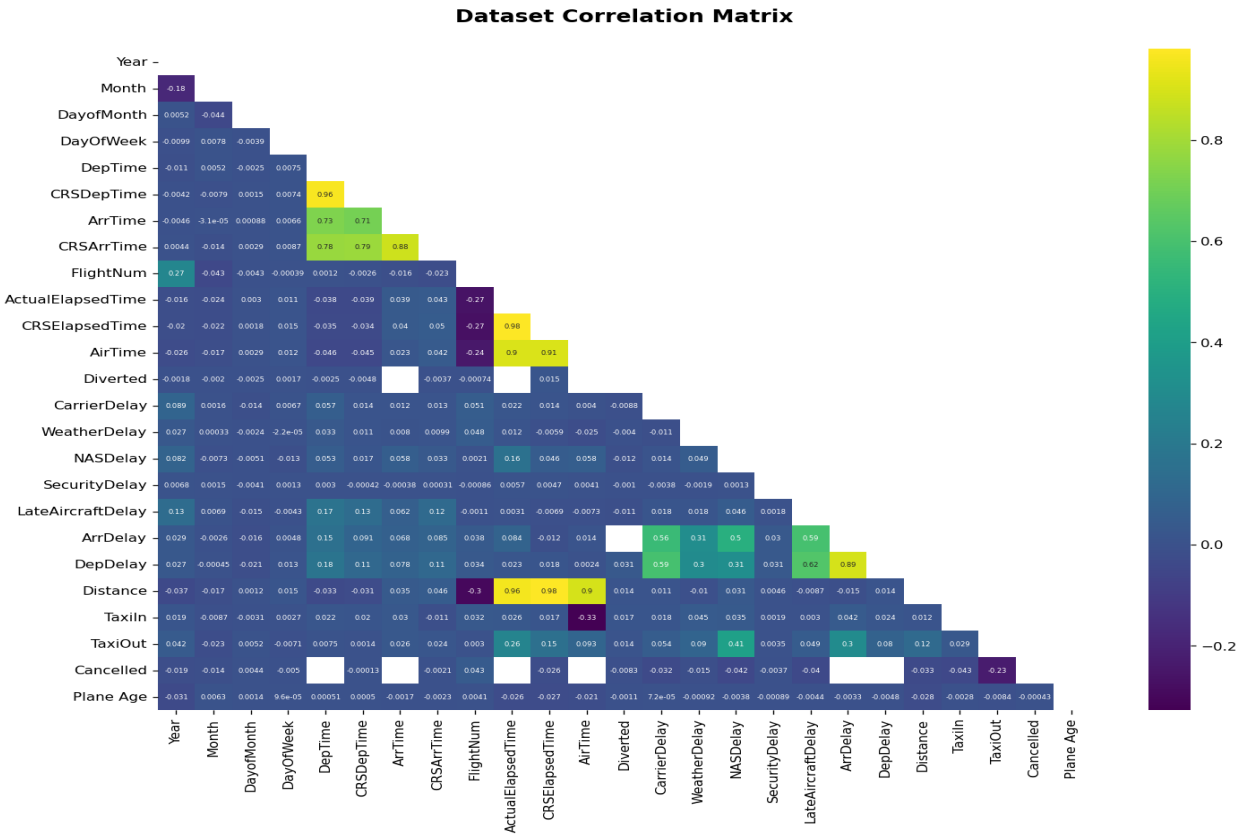

Figure 3: Correlation Matrix

We noticed that, features such as departure delay, arrival delay, taxi-out time, distance, and NAS delay showed the highest correlation with flight diversion. Accordingly, we selected them to model the logistic regression algorithm.

3.2.Implementing the logistic regression model

Table 3 displays the coefficients of the logistic regression models across years. As may be noticed :

Table 3: Flight diversion prediction using logistic regression coefficients

	'DepDelay'	'TaxiIn'	'CRSElapsedTime'	'Distance'	'TaxiOut'	Intercept
1999	2.55032044e-02	-3.77342496e+01	-6.67379308e-03	7.71123775e-04	5.17826886e-01	-4.49853157
2000	2.60156095e-02	-3.46176888e+01	-2.33935850e-02	2.85075552e-03	5.43438228e-01	-4.04039564
2001	3.18923388e-02	-3.28326065e+01	-1.90416032e-02	2.50436550e-03	5.03977737e-01	-4.26333571
2002	3.25443663e-02	-2.79847476e+01	-7.82787319e-03	1.56216334e-03	3.14170871e-01	-3.92173749
2003	0.0065642 9	0.016826 77	0.00588575	-0.0006773 5	0.00975729	-6.93315205
2004	0.0049931 1	0.002925 48	0.01202734	-0.0010303 3	0.00143049	-7.46884852
2005	1.50732348e-02	-2.81526026e+01	-1.65290625e-02	2.49309717e-03	3.75111119e-01	-3.84942449
2006	5.34028593e-03	-1.23876154e+00	1.07895446e-02	-5.05246721e-04	4.29603784e-02	-4.15404271
2007	7.37251838e-03	-2.64674215e+01	-7.36598983e-03	1.79325722e-03	3.26939743e-01	-4.52747839
2008	0.0049347 2	-0.006709 23	-0.00143252	0.0005138 2	0.01280351	-6.43897945

In this analysis, we first preprocessed the data by imputing missing values using the mean of each feature. Subsequently, we split the data into training and testing sets using an 80-20 split, with 80% of the data used for training and 20% for testing. The logistic regression models were then trained using the training subset, and the coefficients presented in Table 3 were estimated using this subset. Following the training phase, the accuracy of the models was evaluated using the test subset. This process allowed us to assess the performance of the models and the accuracy of the coefficient estimates. Table 3 presents the coefficients of the logistic regression models for a range of years, shedding light on the associations between key attributes and the likelihood of a flight divert. For example, a positive coefficient suggests that the chance of aircraft diversion increases as the associated characteristic value increases, whereas a negative coefficient suggests the reverse. The intercept represents the base probability of flight diversion when all other features are zero.

After training the model, it was evaluated using the test subset. The accuracy of the model varied across different years, with an average accuracy of 0.98.

General Conclusion

We addressed three main topics in this research by performing a thorough examination of US flight data from 1999 to 2008. To start, we determined which days and times of the week each year throughout this time frame would be ideal for minimizing flight delays. Through the examination of past flight data, we were able to ascertain the best day-of-week and time-of-week for both arrival and departure in order to reduce delays. Secondly, we looked into whether planes that are older than others have higher delays. Using historical flight data, we categorized planes as older or younger and analyzed their average delay for each year. Lastly, we fitted logistic regression models to predict the probability of flight diversion for each year from 1999 to 2008. Utilizing features such as departure delay, arrival delay, taxi-out time, distance, and NAS delay, we trained logistic regression models and visualized the coefficients across years. Our analysis revealed that the best days and times to minimize flight delays varied across different years. Logistic regression models provided insights into the probability of flight diversion, highlighting the most influential features such as departure delay, arrival delay, and distance. Our data reveals that airlines may reduce delays and maximize flight schedules by steering clear of peak hours and days. Airline companies may make more informed judgments about fleet management and maintenance by having a better understanding of how jet age affects delays. All things considered, our analysis sheds light on the variables affecting flight delays and diversions, allowing airlines to increase operational effectiveness and customer happiness.

[1] Harvard Dataverse. (2008). Data Expo 2009: Airline on time data (Version V1). https://doi.org/10.7910/DVN/HG7NV7