Developing a Flight Delay Prediction Model using Machine Learning

*Submitted in partial fulfilment for the award of the degree*

BE.ELECTRONIC & COMMUNICATION ENGINEERING

BY

AJAY KUMAR.P(511319106002)

SANJAY.S(511319106027)

SUBASH.C(511319106030)

SUNDER MOORTHY.GP(511319106031)

CONTENTS

1. INTRODUCTION

1.1 Project Overview

1.2 Purpose

2. LITERATURE SURVEY

2.1 Existing problem

2.2 References

2.3 Problem Statement Definition

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming

3.3 Proposed Solution

3.4 Problem Solution fit

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

4.2 Non-Functional requirements

5. PROJECT DESIGN

5.1 Data Flow Diagrams

5.2 Solution & Technical Architecture

5.3 User Stories

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

6.2 Sprint Delivery Schedule

6.3 Reports from JIRA

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature

1 7.2 Feature   
2 7.3 Database Schema (if Applicable)

8. TESTING

8.1 Test Cases

8.2 User Acceptance Testing

9. RESULTS 9.1 Performance Metrics

10. ADVANTAGES & DISADVANTAGES

11. CONCLUSION

12. FUTURE SCOPE

13. APPENDIX Source Code GitHub & Project Demo Link

INTRODUCTION

1. INTRODUCTION
2. A Review on Flight Delay Prediction
3. Alice Sternberg
4. CEFET/RJ
5. Jorge Soares
6. CEFET/RJ
7. jorge.soares@cefet-rj.br
8. Diego Carvalho
9. CEFET/RJ
10. dcarvalho@ieee.org
11. Eduardo Ogasawara
12. CEFET/RJ
13. eogasawara@ieee.org
14. Abstract
15. Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-
16. making process for all players of commercial aviation. Moreover, the development of accurate
17. prediction models for ﬂight delays became cumbersome due to the complexity of air transportation
18. system, the number of methods for prediction, and the deluge of ﬂight data. In this context, this paper
19. presents a thorough literature review of approaches used to build ﬂight delay prediction models from
20. the Data Science perspective. We propose a taxonomy and summarize the initiatives used to address
21. the ﬂight delay prediction problem, according to scope, data, and computational methods, giving
22. particular attention to an increased usage of machine learning methods. Besides, we also present a
23. timeline of signiﬁcant works that depicts relationships between ﬂight delay prediction problems and
24. research trends to address them.
25. The published version of this paper is made available at https://doi.org/10.1080/01441647.
26. 2020.1861123.
27. Please cite as:
28. L. Carvalho, A. Sternberg, L. Maia Gonc¸alves, A. Beatriz Cruz, J.A. Soares, D. Brand˜
29. ao, D. Car-
30. valho, e E. Ogasawara, 2020, On the relevance of data science for ﬂight delay research: a systematic
31. review, Transport Reviews
32. Keywords Flight delays ·Commercial aviation ·Brazilian system
33. 1 Introduction
34. Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial avia-
35. tion players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by
36. the diﬀerence between scheduled and real times of departure or arrival of a plane [117]. Country regulator authorities
37. have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential sub-
38. ject in the context of air transportation systems. In 2013, 36% of ﬂights delayed by more than ﬁve minutes in Europe,
39. 31.1% of ﬂights delayed by more than 15 minutes in the United States, and 16.3% of ﬂights were canceled or suﬀered
40. delays greater than 30 minutes in Brazil [45, 5]. This indicates how relevant this indicator is and how it aﬀects no
41. matter the scale of airline meshes.
42. Flight delays have negative impacts, mainly economic, for passengers, airlines, and airports. Given the uncertainty
43. of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip
44. costs, to ensure their arrival on time [11, 55]. On the other hand, airlines suﬀer penalties, ﬁnes and additional operation
45. costs, such as crew and aircrafts retentions in airports [25, 112, 51, 62]. Furthermore, from the sustainability point of
46. view, delays may also cause environmental damage by increasing fuel consumption and gas emissions [95, 105, 102,
47. 75, 8, 125].
48. A Review on Flight Delay Prediction
49. Alice Sternberg
50. CEFET/RJ
51. Jorge Soares
52. CEFET/RJ
53. jorge.soares@cefet-rj.br
54. Diego Carvalho
55. CEFET/RJ
56. dcarvalho@ieee.org
57. Eduardo Ogasawara
58. CEFET/RJ
59. eogasawara@ieee.org
60. Abstract
61. Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-
62. making process for all players of commercial aviation. Moreover, the development of accurate
63. prediction models for ﬂight delays became cumbersome due to the complexity of air transportation
64. system, the number of methods for prediction, and the deluge of ﬂight data. In this context, this paper
65. presents a thorough literature review of approaches used to build ﬂight delay prediction models from
66. the Data Science perspective. We propose a taxonomy and summarize the initiatives used to address
67. the ﬂight delay prediction problem, according to scope, data, and computational methods, giving
68. particular attention to an increased usage of machine learning methods. Besides, we also present a
69. timeline of signiﬁcant works that depicts relationships between ﬂight delay prediction problems and
70. research trends to address them.
71. The published version of this paper is made available at https://doi.org/10.1080/01441647.
72. 2020.1861123.
73. Please cite as:
74. L. Carvalho, A. Sternberg, L. Maia Gonc¸alves, A. Beatriz Cruz, J.A. Soares, D. Brand˜
75. ao, D. Car-
76. valho, e E. Ogasawara, 2020, On the relevance of data science for ﬂight delay research: a systematic
77. review, Transport Reviews
78. Keywords Flight delays ·Commercial aviation ·Brazilian system
79. 1 Introduction
80. Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial avia-
81. tion players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by
82. the diﬀerence between scheduled and real times of departure or arrival of a plane [117]. Country regulator authorities
83. have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential sub-
84. ject in the context of air transportation systems. In 2013, 36% of ﬂights delayed by more than ﬁve minutes in Europe,
85. 31.1% of ﬂights delayed by more than 15 minutes in the United States, and 16.3% of ﬂights were canceled or suﬀered
86. delays greater than 30 minutes in Brazil [45, 5]. This indicates how relevant this indicator is and how it aﬀects no
87. matter the scale of airline meshes.
88. Flight delays have negative impacts, mainly economic, for passengers, airlines, and airports. Given the uncertainty
89. of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip
90. costs, to ensure their arrival on time [11, 55]. On the other hand, airlines suﬀer penalties, ﬁnes and additional operation
91. costs, such as crew and aircrafts retentions in airports [25, 112, 51, 62]. Furthermore, from the sustainability point of
92. view, delays may also cause environmental damage by increasing fuel consumption and gas emissions [95, 105, 102,
93. 75, 8, 125].
94. A Review on Flight Delay Prediction
95. Alice Sternberg
96. CEFET/RJ
97. Jorge Soares
98. CEFET/RJ
99. jorge.soares@cefet-rj.br
100. Diego Carvalho
101. CEFET/RJ
102. dcarvalho@ieee.org
103. Eduardo Ogasawara
104. CEFET/RJ
105. eogasawara@ieee.org
106. Abstract
107. Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-
108. making process for all players of commercial aviation. Moreover, the development of accurate
109. prediction models for ﬂight delays became cumbersome due to the complexity of air transportation
110. system, the number of methods for prediction, and the deluge of ﬂight data. In this context, this paper
111. presents a thorough literature review of approaches used to build ﬂight delay prediction models from
112. the Data Science perspective. We propose a taxonomy and summarize the initiatives used to address
113. the ﬂight delay prediction problem, according to scope, data, and computational methods, giving
114. particular attention to an increased usage of machine learning methods. Besides, we also present a
115. timeline of signiﬁcant works that depicts relationships between ﬂight delay prediction problems and
116. research trends to address them.
117. The published version of this paper is made available at https://doi.org/10.1080/01441647.
118. 2020.1861123.
119. Please cite as:
120. L. Carvalho, A. Sternberg, L. Maia Gonc¸alves, A. Beatriz Cruz, J.A. Soares, D. Brand˜
121. ao, D. Car-
122. valho, e E. Ogasawara, 2020, On the relevance of data science for ﬂight delay research: a systematic
123. review, Transport Reviews
124. Keywords Flight delays ·Commercial aviation ·Brazilian system
125. 1 Introduction
126. Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial avia-
127. tion players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by
128. the diﬀerence between scheduled and real times of departure or arrival of a plane [117]. Country regulator authorities
129. have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential sub-
130. ject in the context of air transportation systems. In 2013, 36% of ﬂights delayed by more than ﬁve minutes in Europe,
131. 31.1% of ﬂights delayed by more than 15 minutes in the United States, and 16.3% of ﬂights were canceled or suﬀered
132. delays greater than 30 minutes in Brazil [45, 5]. This indicates how relevant this indicator is and how it aﬀects no
133. matter the scale of airline meshes.
134. Flight delays have negative impacts, mainly economic, for passengers, airlines, and airports. Given the uncertainty
135. of their occurrence, passengers usually plan to travel many hours earlier for their appointments, increasing their trip
136. costs, to ensure their arrival on time [11, 55]. On the other hand, airlines suﬀer penalties, ﬁnes and additional operation
137. costs, such as crew and aircrafts retentions in airports [25, 112, 51, 62]. Furthermore, from the sustainability point of
138. view, delays may also cause environmental damage by increasing fuel consumption and gas emissions [95, 105, 102,
139. 75, 8, 125].
140. Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial avia-
141. tion players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by
142. the diﬀerence between scheduled and real times of departure or arrival of a plane [117]. Country regulator authorities
143. have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential sub-
144. ject in the context of air transportation systems. In 2013, 36% of ﬂights delayed by more than ﬁve minutes in Europe,
145. 31.1% of ﬂights delayed by more than 15 minutes in the United States, and 16.3% of ﬂights were canceled or suﬀered
146. delays greater than 30 minutes in Brazil [45, 5]. This indicates how relevant this indicator is and how it aﬀects no
147. matter the scale of airline meshes
148. Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial avia-
149. tion players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by
150. the diﬀerence between scheduled and real times of departure or arrival of a plane [117]. Country regulator authorities
151. have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential sub-
152. ject in the context of air transportation systems. In 2013, 36% of ﬂights delayed by more than ﬁve minutes in Europe,
153. 31.1% of ﬂights delayed by more than 15 minutes in the United States, and 16.3% of ﬂights were canceled or suﬀered
154. delays greater than 30 minutes in Brazil [45, 5]. This indicates how relevant this indicator is and how it aﬀects no
155. matter the scale of airline meshes
156. Delay is one of the most remembered performance indicators of any transportation system. Notably, commercial avia-
157. tion players understand delay as the period by which a ﬂight is late or postponed. Thus, a delay may be represented by
158. the diﬀerence between scheduled and real times of departure or arrival of a plane [117]. Country regulator authorities
159. have a multitude of indicators related to tolerance thresholds for ﬂight delays. Indeed, ﬂight delay is an essential sub-
160. ject in the context of air transportation systems. In 2013, 36% of ﬂights delayed by more than ﬁve minutes in Europe,
161. 31.1% of ﬂights delayed by more than 15 minutes in the United States, and 16.3% of ﬂights were canceled or suﬀered
162. delays greater than 30 minutes in Brazil [45, 5]. This indicates how relevant this indicator is and how it aﬀects no
163. matter the scale of airline meshes.

Delay is one of the most remembered performance indicators of any transportation system. Notably commercial aviation players understand delay as the period by which a flight is late or postponed. Thus delay may be represented by the difference between scheduled and real times of departure or arrival of a plane . Country regulator authorities have a multitude of indicators related to tolerance thresholds for flight delays. Indeed, flight delay is an essential subject in the context of air transportation systems. In 2013, 36% of flights delayed by more than five minutes in Europe, 31.1% of flights delayed by more than 15 minutes in the United States, and 16.3% of flights were canceled or suffered delays greater than 30 minutes in Brazil . This indicates how relevant this indicator is and how it affects no matter the scale of airline meshes. “Flight Delay Predictions” is a supervised machine learning project. The primary goal of this project is to predict airline delays caused by various factors. Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions. Here we train a regression model to predict if a flight will be delayed by more than 15 minutes. The model was trained using features of the flights known at the time of booking such as the airline, month, week, and hour of departure. This report will help to understand basic concepts of supervised machine learning and the concepts used in making this project.

* 1. Project Overview

This dataset reports flights in the United States, including carriers, arrival and departure delays, and reasons for delays, for 2008. This dataset was obtained from the **RITA** (Research and Technology Bureau of Transportation Statistics)

A flight is considered delayed when it arrived 15 or more minutes than the schedule . Delayed minutes are calculated for delayed flights only.  
When multiple causes are assigned to one delayed flight, each cause is prorated based on delayed minutes it is responsible for. The displayed numbers are rounded and may not add up to the total.

1.2 Purpose

Therefore, predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the economy. In this study, the main goal is **to compare the performance of machine learning classification algorithms when predicting flight delays**.

LITERATURE SURVEY

Nowadays, service quality plays an important role in attracting customers.

Among these, air travels have their special customers and the most important matter

in these travels is the flight time, on time arrival at destination for passengers such

those who have an important meeting, that has been leading to high expenses for

the passenger until they get to their destination on time. Flight delay has negative

economic effects on the passenger, agencies and airport. Therefore, any reduction of

these effect requires decreasing postponed flight price, so that prediction or

estimation has a great significance and numerous studies has been to dedicated this

subject. Correspondingly, all the scientists have tried to design a model that

understands effective factors and computes effect of each factor and their relation.

Overall, the prediction methods are classified into five groups including Statistical

Methods, Probability methods, network-based methods, operational methods and

machine learning methods .

In one of the best studies that has been performed based on statistics delay

time has been considered to be reduced. Their study has investigated important

factors before fly and those which occur on the ground. In the next step, it has

predicted the delay at destination based on factors that occur in the vicinity of arrival

time at destination. Eventually, results have shown that whenever, the delay is

correctly predicted, passenger disaffection and fuel consumption decrease and

consequently number of flight increases. Moreover, it is possible to increase the

agencies benefits through reducing number of passengers who wrongly selected

their routes or specifying the probabilities for some flights and optimizing delay time

prediction.

Another prominent investigation based on Probability has been done and the

author believes that huge storm in U.S.A has highly affected the flight delay. This

study has been devoted to predict delay based on mathematical calculations and

through considering delay time duration of the flights that had been engaged to

storm in the same day. Metrological reports have shown the effect of storm one hour

before and after event cause ephemeral climate at the region. In the next step,

Monte-Carlo simulation has been used to estimate the airport runway capacity, so

that traffic of each runway would have been estimated. As the research has

employed only one factor, the model has not enough accuracy, but it is possible to

increase region air capacity path structure.

**CONCLUSION**

Predicting flight delays is on interesting research topic and required many

attentions these years. Majority of research have tried to develop and expand their

models in order to increase the precision and accuracy of predicting flight delays.

Since the issue of flights being on-time is very important, flight delay prediction

1. models must have high precision and accuracy.

2.1 Existing problem

* **Air Carrier:** The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
* **Extreme Weather:** Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
* **National Aviation System (NAS):** Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
* **Late-arriving aircraft:** A previous flight with same aircraft arrived late, causing the present flight to depart late.
* **Security:** Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

2.2 References

Paper1 : **DOI:**[**https://doi.org/10.1145/3497701.3497725**](https://doi.org/10.1145/3497701.3497725) **ICEBI 2021:**[**2021 5th International Conference on E-Business and Internet**](https://doi.org/10.1145/3497701)**, Singapore, Singapore, October 2021**

**Paper2:** [2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)](https://ieeexplore.ieee.org/xpl/conhome/9633338/proceeding)

Paper3: [IOP Conference Series: Earth and Environmental Science](https://iopscience.iop.org/journal/1755-1315), [Volume 81](https://iopscience.iop.org/volume/1755-1315/81), [2nd International Conference on Materials Science, Energy Technology and Environmental Engineering (MSETEE 2017) 28–30 April 2017, Zhuhai, China](https://iopscience.iop.org/issue/1755-1315/81/1)**Citation** Yi Ding 2017 IOP Conf. Ser.: Earth Environ. Sci. **81** 012198**DOI** 10.1088/1755-1315/81/1/012198

2.3 Problem Statement Definition

Problem is the core feature in domain taxonomy. As seen in Section 2, there are three major concerns regarding the

ﬂight delay prediction problem: delay propagation, root delay and cancellation. Depending on the emphasis of the

research, authors select one of these lines to develop their models

Problem is the core feature in domain taxonomy. There are three major concerns regarding the flight delay prediction problem

1.**delay propagation**

**we**

**2.root delay and**

**3.cancellation.**

Depending on the emphasis of the research, authors select one of these lines to develop their models.

1. IDEATION & PROPOSED SOLUTION

**Here are some tips to help you handle a cancelled or delayed flight in a manner that's more likely to get you some satisfying results.**

1. Be Polite, But Firm. ...
2. Check For Reimbursements. ...
3. Agree to A New Connection. ...
4. Call the Airline. ...
5. Don't Take It Out On The Flight Crew. ...
6. Have Insurance. ...
7. Arrive Early

3.1 Empathy Map Canvas

Diagram

Description automatically generated

3.2 Ideation & Brainstorming

3.3 Proposed Solution

3.4 Problem Solution fit

**Root delay and cancellation**

Considering that new delay (root delay) may happen eventually, these root delays impair the performance of transportation network. Researchers create prediction models to tackle root delay, predicting when and where a delay will occur and what are its reasons and sources.

**Delay propagation**

In delay propagation, the primary objective is to understand how delay propagates through airlines and airports based on the assumption that an initial delay has already occurred in the transportation system.

**Scope**

Delays can be induced by different sources and affect airports, airlines, en route airspace or an ensemble of them. For analysis purposes, one may assume a simplified system where only one of these actors or any combination of them is considered.

4. REQUIREMENT ANALYSIS

The delay ratio is calculated by summing all the flights that have been delayed at the origin, and dividing by the total number of flights made at the origin.

The trick is **narrowing your scope by location and time**.

4.1 Functional requirement

4.2 Non-Functional requirements

5. PROJECT DESIGN

5.1 Data Flow Diagrams

5.2 Solution & Technical Architecture

5.3 User Stories

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

6.2 Sprint Delivery Schedule

6.3 Reports from JIRA

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature

1 7.2 Feature   
2 7.3 Database Schema (if Applicable)

8. TESTING

8.1 Test Cases

8.2 User Acceptance Testing

9. RESULTS

9.1 Performance Metrics

10. ADVANTAGES & DISADVANTAGES

ADVANTAGE

Therefore, predicting flight delays can **improve airline operations and passenger satisfaction**, which will result in a positive impact on the economy. In this study, the main goal is to compare the performance of machine learning classification algorithms when predicting flight delays.

DISADVANTAGE

Delay in flight is inevitable which has **too much negative economic effects on passengers, agencies and airpor**

11. CONCLUSION

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system. In this context, researchers created flight delay models for delay prediction over the last years, and this work contributes with an analysis of these models from a Data Science perspective. We developed a taxonomy scheme and classified models in respect of detailed components. 10 preprint Figure 8: Trends in machine learning for flight delay prediction Mainly, the taxonomy includes domain and Data Science branches. The former branch categorizes the problem (flight delay prediction) and the scope. The last branch groups methods and data handling. It was observed that the flight delay prediction is classified into two main categories, such as delay propagation and root delay and cancellation. Besides, the scope determines one of the three specific extents: airline, airport, en-route airspace or an ensemble of them. Additionally, considering Data Science branch, we aimed at the datum, by categorizing data sources, dimensions that can be used in the models, and data management techniques to preprocess data and improve prediction models efficiency. We also studied and divided the main methods into five categories: statistical analysis, probabilistic models, network representation, operations research, and machine learning. Those categories have been grouped as their use on specific forecast models for flight delays. Besides the taxonomic scheme, we also presented a timeline with all articles to spot trends and relationships involving the main elements in the taxonomy. In the light of the domain-problem classification, this timeline showed a dominance of delay propagation and root delay over cancellation analysis. Researchers used to focus on statistical analysis and operational research approaches in the past. However, as the data volume grows, we noticed the use of machine learning and data management is increasing significantly. This clearly characterizes a Data Science trend. Researchers from airlines, airports, and academia will require a combination of skills of both domain specialists and data scientists to enable knowledge discovery from flight Big Data.

12. FUTURE SCOPE

Flight delay is inevitable and it plays an important role in both profits and loss of the airlines. An accurate estimation of flight delay is critical for airlines because **the results can be applied to increase customer satisfaction and incomes of airline agencies**

13. APPENDIX Source Code GitHub & Project Demo Link