

PROFESSIONAL TRAINING REPORT
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award
of Bachelor of Technology Degree in Information Technology

By

Jagadeesh V

40120038



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the Benefice work of Jagadeesh V,
40120038 who carried out the project entitled "PREDICTION OF FLIGHT

DELAYS" under my supervision from February 2023 to April 2023.

Internal Guide

DR.R. Jeberson Retna Raj MTech., Ph., D.,

Head of the Department

DR.R. SUBHASHINI M.E., PH.D.,

Submitted for Viva voce Examination held on 28/04/2023

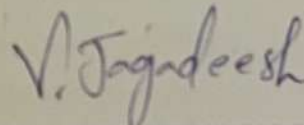
Internal Examiner
External Examiner

DECLARATION

I Jagadeesh V hereby declare that the Project Report entitled "**Prediction of Flight Delays**" done by me under the guidance of Dr. R. Jeberson Retna Raj M.Tech., Ph. d, at Sathyabama Institute of Science and Technology (Deemed to be University) Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600119 is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Information Technology.

DATE: 28/04/24

PLACE: Chennai



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. SASIKALA M.E., Ph.D., Dean**, School of Computing **Dr. R. Subhashini** Head of the Department of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. R. Jeberson Retna Raj M.Tech., Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project

TRAINING CERTIFICATE

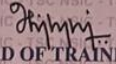
No. : 34342


एन एस आई सी
N S I C
राष्ट्रीय लघु उद्योग निगम लिमिटेड
NATIONAL SMALL INDUSTRIES CORPORATION LIMITED
(A Govt. of India Enterprise)
NSIC - TECHNICAL SERVICES CENTRE
Sector B-24, Guindy Industrial Estate, Ekkaduthangal, Chennai - 600 032.

CERTIFICATE

This is to certify that Mr. **JAGADEESH V. S/o.**
Shri. **VENKATRAJ A.** student of III year
B.Tech. (IT), Sathyabama Institute of Science and
Technology has undergone "**SKILL DEVELOPMENT
PROGRAM ON DATA ANALYTICS**" conducted by
us for a period of one week (30 Hours) from
09.03.2023 to 14.03.2023.




HEAD OF TRAINING


HEAD OF CENTRE

TABLE OF CONTENT

Page no.

Bonafide	ii
Declaration	iii
Acknowledgement	iv
Training Certificate	v
List of figures	x
Abstract	ix

CHAPTER

1.INTRODUCTION	1
1.1. Research Motivation	2
1.2. Problem Statement	2
1.3. Report Structure	3
1.4. Related Work	3
2. AIM AND SCOPE	4
2.1 Aim	4
2.2 Scope	4

3.METHODOLGY	5
3.1. SYSTEM REQUIREMENTS	5
3.1.1 Windows	5
3.1.2. Mac	5
3.1.3. Linux	5
3.2. BASIC CONCEPTS	6
3.2.1 Flight data	6
3.2.2. Delays	7
3.3. PROJECT DEVELOPMENT	8
3.4. DATA COLLECTION	9
3.5. DATA ANALYSIS	11
3.5.1. Data Collection	11
3.5.2. Data Cleaning	11
3.5.3. Data Analysis	11
3.5.4. Data Interpretation	11
3.5.5. Data Visualization	11
3.6. TYPES OF DATA ANALYSIS	12

3.6.1. Descriptive Analysis	12
3.6.2. Diagnostic Analysis	13
3.6.3. Predictive Analysis	14
3.3.4. Prescriptive Analysis	15
3.7. Big Data Analytics Tools	16
3.7.1. Apache Hadoop	17
3.7.2. Cassandra	17
3.7.3. Qubole	18
3.7.4. Xplenty	18
3.7.5. Spark	19
3.7.6. Mongo DB	20
3.7.7. Apache Storm	20
3.7.8. SAS	21
3.7.9. Data Pine	21
3.7.10. Rapid Miner	22
3.8 MICROSOFT POWERBI	23
3.8.1. Uses of Power BI	24
3.8.2. Features of Power BI	25
3.8.3 Power BI Embedded Analytics	26

4. PROJECT IMPLEMENTATION	27
4.1.DATA EXPLORATION OF DATA	29
4.1.1. Departure & Arrival Delay Distribution	29
4.1.2. Average departure & arrival by Month	31
4.1.3. Mean Delays mins by Date	33
4.1.4. Total Delayed by Mins	34
4.2. TYPES OF DELAY	35
4.2.1. Avg Delay of Luggage and tech issues by month	35
4.2.2. Avg of security and Weather Delay by Month	36
4.2.3. Average of Passenger and Propagation Delay by month	37
4.2.4. Average of number of delays by month	38
5.CONCLUSION AND FUTURE WORK	40
REFERENCES	41

LIST OF FIGURES

Chapter	Figures	Page
3.1	Data	7
3.2	Project Development	8
3.3	Data Collection	10
4.1	Arrival distribution	29
4.2	Departure distribution	30
4.3	Average Departure	31
4.4	Average Arrival	32
4.5	Means Delay	33
4.6	Total Delayed mins	34
4.7	Avg delay of luggage and tech issues	35
4.8	Avg of Security and weather delay	36
4.9	Avg Passenger and Propagation delay	37
4.10	Total number of Delay	38

ABSTRACT:

“Prediction of Flight Delays”

Flight delay is a significant problem that negatively impacts the aviation industry and costs billions of dollars each year. Most existing studies investigated this issue using various methods based on applying machine learning methods to predict the flight delay. However, due to the highly dynamic environments of the aviation industry, relying only on single route of airport may not be sufficient and applicable to forecast the future of flights. The purpose of this project is to analyze a broader scope of factors which may potentially influence the flight delay it compares several machine learning-based models in designed generalized flight delay prediction tasks. In this project we have used flight delay dataset from US Department of Transportation (DOT) to predict flight delays. We have used supervised learning algorithms to predict flight departure delay and then model evaluation is done to get best model and our model can identify which features were more important when predicting flight delays.

Key Words: departure delay; supervised learning; ensemble learning; prediction

CHAPTER 1. INTRODUCTION

In the present world, the major components of any transportation system include passenger airline, cargo airline, and air traffic control system. With the passage of time, nations around the world have tried to evolve numerous techniques of improving the airline transportation system. This has brought drastic change in the airline operations. Flight delays occasionally cause inconvenience to the modern passengers. Every year approximately 20% of airline flights are cancelled or delayed, costing passengers more than 20 billion dollars in money and their time. Notable reasons for commercially scheduled flights to delay are adverse weather conditions, air traffic congestion, late reaching aircraft to be used for the flight from previous flight, maintenance, and security issues. One of the key business issues that airlines face is that the vital prices that are related to flights being delayed because of natural occurrences and operational shortcomings that is an upscale affair for the airlines, making issues in scheduling and operations for the end users therefore inflicting unhealthy name and client discontent. As we all know that we tend to not get the flight delay before departure as customers of the Airline Company neither the airline company ground staff gets the airline delay prediction supported varied conditions

1.1. Research Motivation

Average aircraft delay is regularly referred to as an indication of airport capacity. Flight delay is a prevailing problem in this world. It's very tough to explain the reason for a delay. A few factors responsible for the flight delays like runway construction to excessive traffic is rare, but bad weather seems to be a common cause. Some flights are delayed because of the reactionary delays, due to the late arrival of the previous flight. It hurts airports, airlines, and affects a company's marketing strategies as companies rely on customer loyalty to support their frequent flying programs

1.2. Problem Statement

My case study was about LaGuardia Airport in New York, Logan International Airport in Boston, San Francisco International Airport in San Francisco, and O'Hare International Airport in Chicago, which are four major airports in the United States of America. But we focused the idea and research on LaGuardia International Airport. Compared with the data produced by all airports in USA, the data which we gathered was very limited, but it gave us a great direction on how weather plays a part in flight delays. In this project, the goal is to use exploratory analysis and to build models to predict airline departure and arrival delays.

1.3. Related Work

The main concern of the researchers and analysts is to predict the reasons for flight delays and for that they have put in their efforts on collecting data about flight and the weather. We have studied the pattern of arrival delay for non-stop domestic flights at the Orlando International Airport. They focused primarily on the cyclic variations that happen in the air travel demand and the weather at that particular airport. we have created a data mining model which enables the flight delays by observing the weather conditions. We have used Power BI to build their models by selecting different classifiers and choosing the one with the best results.

1.4. Report Structure

This master project report is organized into 5 chapters. The preface of the project, research motivation, and problem statement from chapter 1. Chapter 2 describes the Aim and scope of the project. Chapter 3 focuses on system requirements of the project and methodology of the project. Chapter 4 explain the project development, final chapter consists of chapter 5 speaks about the conclusion, future scope of the project and references

CHAPTER 2. AIM and SCOPE

2.1 Aim

Average aircraft delay is regularly referred to as an indication of airport capacity. Flight delay is a prevailing problem in this world. It's very tough to explain the reason for a delay. A few factors responsible for the flight delays like runway construction to excessive traffic is rare, but bad weather seems to be a common cause. Some flights are delayed because of the reactionary delays, due to the late arrival of the previous flight. It hurts airports, airlines, and affects a company's marketing strategies as companies rely on customer loyalty to support their frequent flying programs.

2.2 Scope

Flight delays in air transportation are a major concern that has adverse effect on the economy, the passengers, and the aviation industry. This matter critically requires an accurate estimation for future flight delays that can be implemented to improve airport operations. Customer satisfaction. So that, flight prediction can cause a skillful decision and operation for agencies and airports, and also a good passenger information system can relatively satisfy the customer. Time consumes for customers.

CHAPTER 3. METHODOLOGY

3.1. SYSTEM REQUIREMENTS

3.1.1. WINDOWS

- Microsoft® Windows® 7/8/10 64-bit
- 4 GB RAM minimum, 8 GB RAM recommended
- 2 GB of available disk space minimum,
- 4 GB Recommended 500 MB for IDE + 1.5 GB for Android SDK and emulator system image
- 1280 x 800 minimum screen resolution

3.1.2. MAC

- Mac® OS X® 10.10 Yosemite or higher, up to 10.14 macOS Mojave
- 4 GB RAM minimum, 8 GB RAM recommended
- 2 GB of available disk space minimum,
- 4 GB Recommended 500 MB for IDE + 1.5 GB for Android SDK and emulator system image

3.1.3 LINUX

- GNOME or KDE desktop - Tested on gLinux based on Debian.
- 64-bit distribution capable of running 32-bit applications
- GNU C Library glibc 2.19 or later 17
- 4 GB RAM minimum, 8 GB RAM recommended
- 2 GB of available disk space minimum,

3.2. BASIC CONCEPTS

In this project, will be looking into some of the variables that play a vital role in determining the flight delay. Each variable has a different weight that is used while training the model. Prior knowledge of these variables should be present to understand the dataset.

3.2.1. Flight data

The flight data is from the year 2018, which is taken from Bureau of Transportation Statics. Every flight has many variables, which give detailed information about the specific flight.

1. Year, date, and month
2. Scheduled Departure time
3. Actual Departure Time
4. Scheduled Arrival Time
5. Actual Arrival Time
6. No. of Delays

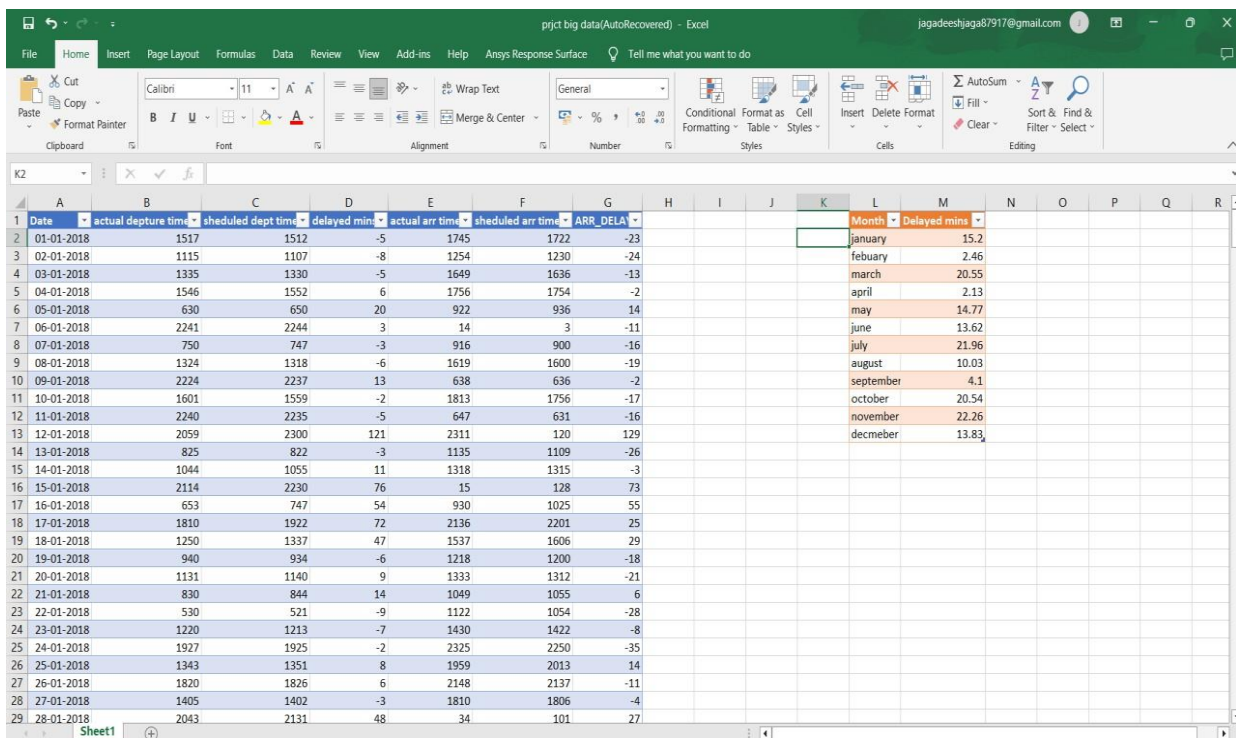
The 2018 on-time performance data contains 7 million records. The average number of daily flights is 169.7 included in the flight database.

3.2.1 Delays

Next one is the main objective of the project that is the cause of flight delays. Flight delays occurs for the varies of reason, in this project we explained about the delays namely:

1. Passenger Delay
2. Weather Delay
3. Propagation Delay
4. Security Delay
5. Luggage Delay
6. Technical Issues Delay

Now collect the dataset for these issues for particular over a year and import it to the excel shown in Figure (4.1).



Date	actual departure time	sheduled dept time	delayed min	actual arr time	sheduled arr time	ARR_DELA
01-01-2018	1517	1512	-5	1745	1722	-23
02-01-2018	1115	1107	-8	1254	1230	-24
03-01-2018	1335	1330	-5	1649	1636	-13
04-01-2018	1546	1552	6	1756	1754	-2
05-01-2018	630	650	20	922	936	14
06-01-2018	2241	2244	3	14	3	-11
07-01-2018	750	747	-3	916	900	-16
08-01-2018	1324	1318	-6	1619	1600	-19
09-01-2018	2224	2237	13	638	636	-2
10-01-2018	1601	1559	-2	1813	1756	-17
11-01-2018	2240	2235	-5	647	631	-16
12-01-2018	2059	2300	121	2311	120	129
13-01-2018	825	822	-3	1135	1109	-26
14-01-2018	1044	1055	11	1318	1315	-3
15-01-2018	2114	2230	76	15	128	73
16-01-2018	653	747	54	930	1025	55
17-01-2018	1810	1922	72	2136	2201	25
18-01-2018	1250	1337	47	1537	1606	29
19-01-2018	940	934	-6	1218	1200	-18
20-01-2018	1131	1140	9	1333	1312	-21
21-01-2018	830	844	14	1049	1055	6
22-01-2018	530	521	-9	1122	1054	-28
23-01-2018	1220	1213	-7	1430	1422	-8
24-01-2018	1927	1925	-2	2325	2250	-35
25-01-2018	1343	1351	8	1959	2013	14
26-01-2018	1820	1826	6	2148	2137	-11
27-01-2018	1405	1402	-3	1810	1806	-4
28-01-2018	2043	2131	48	34	101	27

Month	Delayed mins
january	15.2
february	2.46
march	20.55
april	2.13
may	14.77
june	13.62
july	21.96
august	10.03
september	4.1
october	20.54
november	22.26
decmeber	13.83

Figure (3.1):Data

3.3. PROJECT DEVELOPMENT

In this section, there is an overview of the process of data mining and data modeling, from collecting the data, through the data preparation and finally the data modeling. Data cleaning and formatting can be considered as the most critical part of the whole project according to several data scientists.

Figure (4.2) shows how the process of data mining works to extract knowledge using algorithms.

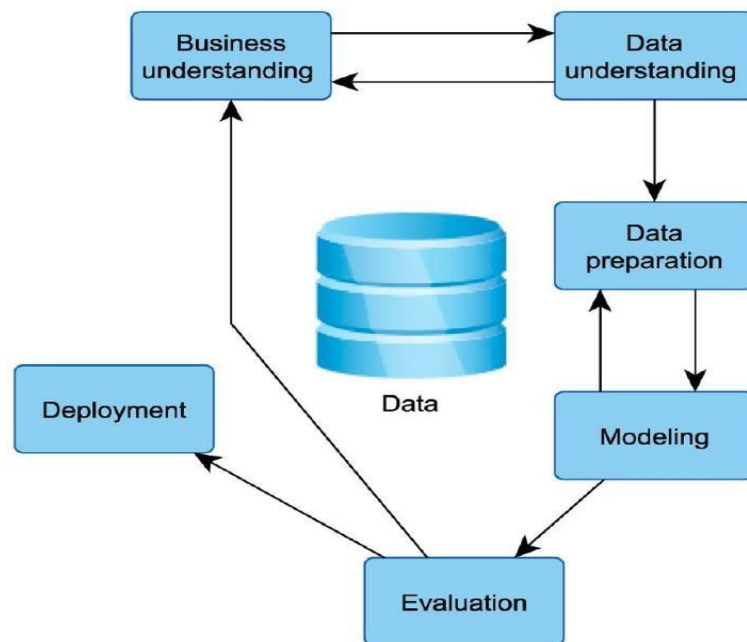


Figure (3.2): Project Development

3.4. DATA COLLECTION

Once the project undertaking is completely comprehended, our subsequent stage is to gather the information that is required for future model building. The information accumulation was an issue as data was not situated at a single source. The data was kept in unique information design. To achieve the end goal, it requires a clear understanding of the correct location of the data.

As we can see in Figure (4.3) the US Bureau of Transportation Statistics gives detailed information on every single household flight, which incorporates their booking and take off circumstances and real takeoff, origin, destination, date, and carrier. We consolidated a portion of the information properties with Local Climatological Data from National Oceanic and Atmospheric Administration (NOAA) to shape a join data set.

Since the datasets for every year are very massive, we decrease our concentration to one-year, i.e., 2018, which as of now contains 1 million records for the most significant airplane terminals. In this venture, we have taken 2018 as our preparation set and 2019 as our test set.

Handling speed is a noteworthy thought since the machine learning methodology that functions admirably on smaller datasets cause issues with the Jupyter Notebook establishments on our PCs.

Month ▾	no of d ▾	passeng ▾	weathe ▾	propag ▾	security ▾	luggage ▾	technic ▾
january	15	10	5	7	9	2	4
february	18	5	17	11	12	10	15
march	15	10	2	3	5	6	9
april	20	5	10	11	18	19	15
may	19	5	18	17	15	12	11
june	24	11	12	15	19	20	22
july	16	15	11	10	5	12	13
august	12	2	11	10	9	2	5
september	10	4	5	6	7	8	9
october	21	10	11	12	13	12	20
november	17	10	11	12	13	14	15
december	11	10	5	4	3	2	1

Figure (3.3): Data Collection

3.5. DATA ANALYSIS

The process of data analysis, or alternately, data analysis steps, involves gathering all the information, processing it, exploring the data, and using it to find patterns and other insights. The process of data analysis consists of: Data Requirement Gathering: Ask yourself why you're doing this analysis, what type of data you want to use, and what data you plan to analyze.

3.5.1. Data Collection: Guided by your identified requirements, it's time to collect the data from your sources. Sources include case studies, surveys, interviews, questionnaires, direct observation, and focus groups. Make sure to organize the collected data for analysis.

3.5.2. Data Cleaning: Not all the data you collect will be useful, so it's time to clean it up. This process is where you remove white spaces, duplicate records, and basic errors. Data cleaning is mandatory before sending the information on for analysis.

3.5.3. Data Analysis: Here is where you use data analysis software and other tools to help you interpret and understand the data and arrive at conclusions. Data analysis tools include Excel, Python, R, Looker, Rapid Miner, Charito, Meta base, Rehash, and Microsoft Power BI.

3.5.4. Data Interpretation: Now that you have your results, you need to interpret them and come up with the best courses of action based on your findings.

3.5.5. Data Visualization: Data visualization is a fancy way of saying, "graphically show your information in a way that people can read and understand it." You can use charts, graphs, maps, bullet points, or a host of other methods. Visualization helps you derive valuable insights by helping you compare datasets and observe relationships.

If you want to make informed decisions, you need data, but there's more to it. The data in question must be accurate. Data analysis helps businesses acquire relevant, accurate information, suitable for developing future marketing strategies, business plans, and realigning the company's vision or mission. A simple example of data analysis can be seen whenever we make a decision in our daily lives by evaluating what has happened in the past or what will happen if we make that decision. Basically, this is the process of analyzing the past or future and making a decision based on that analysis.

3.6. TYPES OF DATA ANALYSIS

The four types of data analysis are:

1. Descriptive Analysis
2. Diagnostic Analysis
3. Predictive Analysis
4. Prescriptive Analysis

3.6.1. Descriptive Analysis

Descriptive analytics answer the question, "What happened?". This type of analytics is by far the most commonly used by customers, providing reporting and analysis centered on past events. It helps companies understand things such as:

- How much did we sell as a company?
- What was our overall productivity?
- How many customers churned in the last quarter?

Descriptive analytics is used to understand the overall performance at an aggregate level and is by far the easiest place for a company to start as data tends to be readily available to build reports and applications. It's extremely important to build core competencies first in descriptive analytics before attempting to advance upward in the data analytics maturity model.

Core competencies include things such as data modeling fundamentals, communicating data with right visualization and basic dashboard skills.

It's likely you've adopted some form of descriptive analytics internally, whether that be static P&L statements, PDF reports, or reporting within an analytics tool. For a true descriptive analytics program to be implemented, the concepts of repeatability and automation of tasks must be top of mind. Repeatability in that a data process is standardized and can be regularly applied with minimal effort (think a weekly sales report), and automation in that complex tasks (VLOOKUPS, merging of excel spreadsheets, etc.) are automated—requiring little to no manual intervention.

3.6.2. Diagnostic analytics

Diagnostic analytics, just like descriptive analytics, uses historical data to answer a question. But instead of focusing on “what”, diagnostic analytics addresses the critical question of why an occurrence or anomaly occurred within your data. Diagnostic analytics also happens to be the most overlooked and skipped step within the analytics maturity model. Anecdotally, I see most customers attempting to go from “what happened” to “what will happen” without ever taking the time to address the “why did it happen” step. This type of analytics helps companies answer questions such as:

- Why did our company sales decrease in the previous quarter?
- Why are we seeing an increase in customer churn?
- Why is a specific basket of products vastly outperforming their prior year sales figures?

Diagnostic analytics tends to be more accessible and fit a wider range of use cases than machine learning/predictive analytics. You might even find that it solves some business problems you earmarked for predictive analytics use cases. Being at the diagnostic analytics phase likely means you've adopted a modern analytics tool. Most modern analytics tools contain a variety of search-based, or lightweight artificial intelligence capabilities.

These features allow for detailed insights a layer deeper (for example: the Key Drivers visualization in Power BI, or Qlik's search-based insight functionality). To be clear, these are an effective lightweight means to address diagnostic analytics use cases but are not a means to a full-scale implementation. Software vendors like Sisuu have built their core business around addressing diagnostic analytics use cases (what they call "augmented analytics") and are a great bet. Diagnostic analytics is an important step in the maturity model that unfortunately tends to get skipped or obscured. If you cannot infer why your sales decreased 20% in 2020, then jumping to predictive analytics and trying to answer "what will happen to sales in 2021" is a stretch in advancing upward in the analytics maturity model.

3.6.3. Predictive Analytics

Predictive analytics is a form of advanced analytics that determines what is likely to happen based on historical data using machine learning. Historical data that comprises the bulk of descriptive and diagnostic analytics is used as the basis of building predictive analytics models. Predictive analytics helps companies address use cases such as:

- Predicting maintenance issues and part breakdown in machines.
- Determining credit risk and identifying potential fraud.
- Predict and avoid customer churn by identifying signs of customer dissatisfaction.

At the outset of any predictive analytics build, three core elements need to be established:

- Identify a problem to solve,
- Define what is you want to predict, and
- State what you will achieve by doing so.

To start you should collect existing data, organize data in a useful way to allow for data modeling, cleanse your data and review overall quality, and finally determine your modeling objective.

While modeling takes up the spotlight in predictive analytics, data prep is a crucial step that needs to happen first. This is why organizations with a rock-solid foundation in descriptive and diagnostic analytics are better equipped to handle predictive analytics. Simply put, the time and effort to prep, transform, and ensure data quality for retrospective reporting has already taken place. The groundwork should be relatively well laid to quickly identify and leverage data for the modeling phase. I always encourage customers with well-defined KPIs and business logic in a specific business reporting area (think sales reporting for example) to use that as the first predictive analytics use case. The goal is to derive value quickly, and there is no better place to start than an area where you know data is well defined and of high quality.

3.6.4. Prescriptive Analytics

Prescriptive analytics is the fourth and final pillar of modern analytics. Prescriptive analytics pertains to true guided analytics where your analytics is prescribing or guiding you toward a specific action to take. It is effectively the merging of descriptive, diagnostic, and predictive analytics to drive decision making. Existing scenarios or conditions (think your current fleet of freight trains) and the ramifications of a decision or occurrence (parts breakdown on the freight trains) are applied to create a guided decision or action for the user to take (proactively buy more parts for preventative maintenance).

Prescriptive analytics requires strong competencies in descriptive, diagnostic, and predictive analytics which is why it tends to be found in highly specialized industries (oil and gas, clinical healthcare, finance, and insurance to name a few) where use cases are well defined. Prescriptive analytics help to address use cases such as:

- Automatic adjustment of product pricing based on anticipated customer demand and external factors.
- Flagging select employees for additional training based on incident reports in the field.

Prescriptive analytics' primary aim is to take the educated guess or assessment out of data analytics and streamline the decision-making process. Prescriptive analytics is commonly considered the merging of descriptive, diagnostic, and predictive analytics. Getting started isn't so much a step-by-step list but rather the time and effort up front to build your competencies within the analytics maturity curve.

Simply put, there is no starting point in prescriptive analytics without the requisite first three pillars of modern analytics being established first. If you're ready for prescriptive analytics, then quantifying your call to action and the underlying criteria will be the first requirement. For example: if the use case is to call corrective action for an employee (i.e. – additional training based on poor performance) then the factors that necessitate this action must be firmly established and the action itself must be clearly defined. Knowing how each kind of analytics helps you better understand your data and how to use it.

3.7 Big Data Analytics Tools

As we're growing with the pace of technology, the demand to track data is increasing rapidly. Today, almost 2.5quintillion bytes of data are generated globally and it's useless until that data is segregated in a proper structure. It has become crucial for businesses to maintain consistency in the business by collecting meaningful data from the market today and for that, all it takes is the right data analytic tool and a professional data analyst to segregate a huge amount of raw data by which then a company can make the right approach. There are hundreds of *data analytics tools* out there in the market today but the selection of the right tool will depend upon your business NEED, GOALS, and VARIETY to get business in the right direction. Now, let's check out the top 10 analytics tools in big data.

3.7.1. APACHE Hadoop

It is a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers. Hadoop also offers cross-platform support for its users. Today, it is the best big data analytic tool and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

Features of Apache Hadoop:

- Free to use and offers an efficient storage solution for businesses.
- Offers quick access via HDFS (Hadoop Distributed File System).
- Highly flexible and can be easily implemented with MySQL, and JSON.
- Highly scalable as it can distribute a large amount of data in small segments.
- It works on small commodity hardware like JBOD or a bunch of disks.

3.7.2. Cassandra

APACHE_Cassandra is an open-source NoSQL distributed database that is used to fetch large amounts of data. It's one of the most popular tools for data analytics and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance. It can deliver thousands of operations every second and can handle petabytes of resources with almost zero downtime. It was created by Facebook back in 2008 and was published publicly.

Features of APACHE Cassandra:

- Data Storage Flexibility: It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
- Data Distribution System: Easy to distribute data with the help of replicating data on multiple data centers.

- **Fast Processing:** Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.
- **Fault-tolerance:** The moment, if any node fails, it will be replaced without any delay.

3.7.3 Qubole

It's an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines. It is capable of configuring multi-cloud services such as AWS, Azure, and Google Cloud. Besides, it also helps in lowering the cost of cloud computing by 50%.

Features of Qubole:

- *Supports ETL process:* It allows companies to migrate data from multiple sources in one place.
- *Real-time Insight:* It monitors user's systems and allows them to view real-time insights
- *Predictive Analysis:* Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.
- *Advanced Security System:* To protect users' data in the cloud, Qubole uses an advanced security system and also ensures to protect any future breaches. Besides, it also allows encrypting cloud data from any potential threat.

7.3.4. Xplenty

It is a data analytic tool for building a data pipeline by using minimal codes in it. It offers a wide range of solutions for sales, marketing, and support. With the help of its interactive graphical interface, it provides solutions for *ETL*, *ELT*, etc. The best part of

using Xplenty is its low investment in hardware & software and it offers support via email, chat, telephonic and virtual meetings.

Xplenty is a platform to process data for analytics over the cloud and segregates all the data together.

Features of Xplenty:

- Rest API: A user can possibly do anything by implementing Rest API
- Flexibility: Data can be sent, and pulled to databases, warehouses, and salesforce.
- Data Security: It offers SSL/TSL encryption and the platform is capable of verifying algorithms and certificates regularly.
- Deployment: It offers integration apps for both cloud & in-house and supports deployment to integrate apps over the cloud.

3.7.5. Spark

APACHE_Spark is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is capable of handling multi-petabytes of data as well. Recently, Spark made a record of processing 100 terabytes of data in just 23 minutes which broke the previous world record of Hadoop (71 minutes). This is the reason why big tech giants are moving towards spark now and is highly suitable for ML and AI today.

Features of APACHE Spark:

- Ease of use: It allows users to run in their preferred language. (JAVA, Python, etc.)
- Real-time Processing: Spark can handle real-time streaming via Spark Streaming
- Flexible: It can run on, Mesos, Kubernetes, or the cloud.

3.7.6. Mongo DB

Came in limelight in 2010, is a free, open-source platform and a document-oriented (NoSQL) database that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of Mongo DB. It is so popular among developers due to its availability for multi-programming languages such as Python, Jscript, and Ruby.

Features of Mongo DB:

- Written in C++: It's a schema-less DB and can hold varieties of documents inside.
- Simplifies Stack: With the help of mongo, a user can easily store files without any disturbance in the stack.
- Master-Slave Replication: It can write/read data from the master and can be called back for backup.

3.7.7. Apache Storm

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to handle a pool of large data in fault-tolerance and horizontally scalable methods. When we talk about real-time data processing, Storm leads the chart because of its distributed real-time big data processing system, due to which today many tech giants are using APACHE Storm in their system. Some of the most notable names are Twitter, Zendesk, NaviSite, etc.

Features of Storm:

- Data Processing: Storm process the data even if the node gets disconnected
- Highly Scalable: It keeps the momentum of performance even if the load increases
- Fast: The speed of APACHE Storm is impeccable and can process up to 1 million messages of 100 bytes on a single node.

3.7.8. SAS

Today it is one of the best tools for creating statistical modeling used by data analysts. By using SAS, a data scientist can mine, manage, extract or update data in different variants from different sources. Statistical Analytical System or SAS allows a user to access the data in any format (SAS tables or Excel worksheets). Besides that it also offers a cloud platform for business analytics called **SAS Viya** and also to get a strong grip on AI & ML, they have introduced new tools and products.

Features of SAS:

- Flexible Programming Language: It offers easy-to-learn syntax and has also vast libraries which make it suitable for non-programmers
- Vast Data Format: It provides support for many programming languages which also include SQL and carries the ability to read data from any format.
- It provides end-to-end security with a feature called SAS/SECURE.

3.7.9. Data Pine

Datapine is an analytical used for BI and was founded back in 2012 (Berlin, Germany). In a short period of time, it has gained much popularity in a number of countries and it's mainly used for data extraction (for small-medium companies fetching data for close monitoring). With the help of its enhanced UI design, anyone can visit and check the data as per their requirement and offer in 4 different price brackets,

starting from \$249 per month. They do offer dashboards by functions, industry, and platform.

Features of Datapine:

- Automation: To cut down the manual chase, datapine offers a wide array of AI assistant and BI tools.
- Predictive Tool: datapine provides forecasting/predictive analytics by using historical and current data, it derives the future outcome.
- Add on: It also offers intuitive widgets, visual analytics & discovery, ad hoc reporting, etc.

3.7.10. Rapid Miner

It's a fully automated visual workflow design tool used for data analytics. It's a no-code platform and users aren't required to code for segregating data. Today, it is being heavily used in many industries such as ed-tech, training, research, etc. Though it's an open-source platform but has a limitation of adding 10000 data rows and a single logical **processor**. With the help of Rapid Miner, one can easily deploy their ML models to the web or mobile (only when the user interface is ready to collect real-time figures).

Features of Rapid Miner:

- Accessibility: It allows users to access 40+ types of files (SAS, ARFF, etc.) via URL
- Storage: Users can access cloud storage facilities such as AWS and dropbox
- Data validation: Rapid miner enables the visual display of multiple results in history for better evaluation.

3.8 MICROSOFT POWERBI

Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence. It is part of the Microsoft Power Platform. Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. Data may be input by reading directly from a database, webpage, or structured files such as spreadsheets, CSV, XML, and JSON. Microsoft Power BI is a business intelligence (BI) platform that provides nontechnical business users with tools for aggregating, analyzing, visualizing and sharing data. Power BI's user interface is fairly intuitive for users familiar with Excel, and its deep integration with other Microsoft products makes it a versatile self-service tool that requires little upfront training. Users can download an application for Windows 10, called Power BI Desktop, and native mobile apps for Windows, Android and iOS devices. There is also Power BI Report Server for companies that must maintain their data and reports on premises.

That version of Power BI requires a special version of the desktop app -- aptly called Power BI Desktop for Power BI Report Server. Came in limelight in 2010, is a free, open-source platform and a document-oriented (NoSQL) database that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of MongoDB. It is so popular among developers due to its availability for multi-programming languages such as Python, Javascript, and Ruby.

3.8.1 Uses of Power bi

Microsoft Power BI is used to find insights within an organization's data. Power BI can help connect disparate data sets, transform and clean the data into a data model and create charts or graphs to provide visuals of the data. All of this can be shared with other Power BI users within the organization.

The data models created from Power BI can be used in several ways for organizations, including the following:

- telling stories through charts and data visualizations;
- examining "what if" scenarios within the data; and
- creating reports that can answer questions in real time and help with forecasting to make sure departments meet business metrics.

Power BI can also provide executive dashboards for administrators or managers, giving management more insight into how departments are doing.

- It's a schema-less DB and can hold varieties of documents inside.
- With the help of mongo, a user can easily store files without any disturbance in the stack.
- It can write/read data from the master and can be called back for backup.

3.8.2 FEATURES OF POWER BI

Microsoft has added a number of data analytics features to Power BI since its inception and continues to do so. Some of the most important features are the following:

Artificial intelligence. Users can access image recognition and text analytics in Power BI, create machine learning models using automated ML capabilities and integrate with Azure Machine Learning.

Hybrid deployment support. This feature provides built-in connectors that allow Power BI tools to connect with a number of different data sources from Microsoft, Salesforce and other vendors.

Quick Insights. This feature allows users to create subsets of data and automatically apply analytics to that information.

Common data model support. Power BI's support for the common data model allows the use of a standardized and extensible collection of data schemas (entities, attributes and relationships).

Cortana integration. This feature, which is especially popular on mobile devices, allows users to verbally query data using natural language and access results using Cortana, Microsoft's digital assistant.

Customization. This feature allows developers to change the appearance of default visualization and reporting tools and import new tools into the platform.

APIs for integration. This feature provides developers with sample code and application program interfaces (APIs) for embedding the Power BI dashboard in other software products.

Self-service data prep. Using Power Query, business analysts can ingest, transform, integrate and enrich big data into the Power BI web service

3.8.3 Power BI Embedded Analytics

Power BI embedded analytics allows you to embed your Power BI items such as reports, dashboards and tiles, in a web application or in a website. Power BI embedded analytics gives you additional benefits over secure embed. It offers a rich, fully integrated experience with full API support, automatic authentication, and the reports can be hosted in apps as well as web pages. Embedded analytics allows you to automate the monitoring, management, and deployment of analytics, while getting full control of Power BI features and intelligent analytics.

Power BI Embedded has basically the same features as Power BI Premium.

You can:

- Deliver compelling data experiences for your end users, enabling them to take action based on insights from your solutions data.
- Quickly and easily provide exceptional customer-facing reports, dashboards, and analytics in your own apps by using and branding Power BI as your own.

Secure embed

Secure embed is the simplest no-code way to embed a report into any portal that accepts a URL or iFrame. The viewer of the report must have the proper Power BI license. The viewer can interact with the report, but not edit, save, or make any changes to it. Secure embed is available in the Power BI service. For more advanced solutions that give your users more flexibility and control, use one of the Power BI embedded analytics solutions described in this article.

CHAPTER 4. PROJECT IMPLEMENTATION

The project is implemented in Microsoft excel and Microsoft power Bi. We need to collect the data set for particular flight delay, over an year or 365 days.

That includes the flight's :

1. Scheduled departure time
2. Actual departure time
3. Total departure Delayed mins .
4. Scheduled Arrival time
5. Actual Arrival time
6. Total Arrival delayed mins.

Now enter the dataset in Microsoft excel sheet as per the date and month shown in figure(i). Next one is the main objective of the project that is cause of flight delays. Flight delays occurs varies of reasons ,in this project we explained about theses delays namely:

1. Passenger delay
2. Weather delay
3. Propagation delay
4. Security delay
5. Luggage delay
6. Technical issues

Now collect the dataset for these issues for particular flight over an year.

Import to the excel sheet. Then import ot the MICROSOFT POWERBI.

4.1. DATA EXPLORATION FOR FLIGHT DATA

As a first exploratory examination, we consider the watched likelihood of deferral in minutes on the whole dataset. The best path is through a histogram, taking a departureat flight and arrival delays independently.

4.1.1 Departure & Arrival Delay Distribution

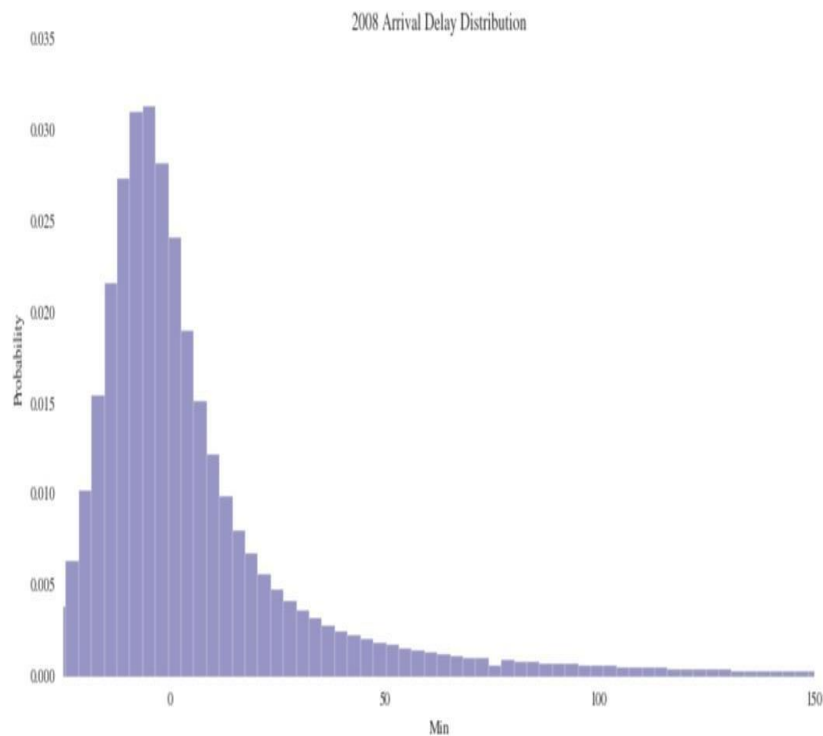


Figure (4.1): Arrival Distribution

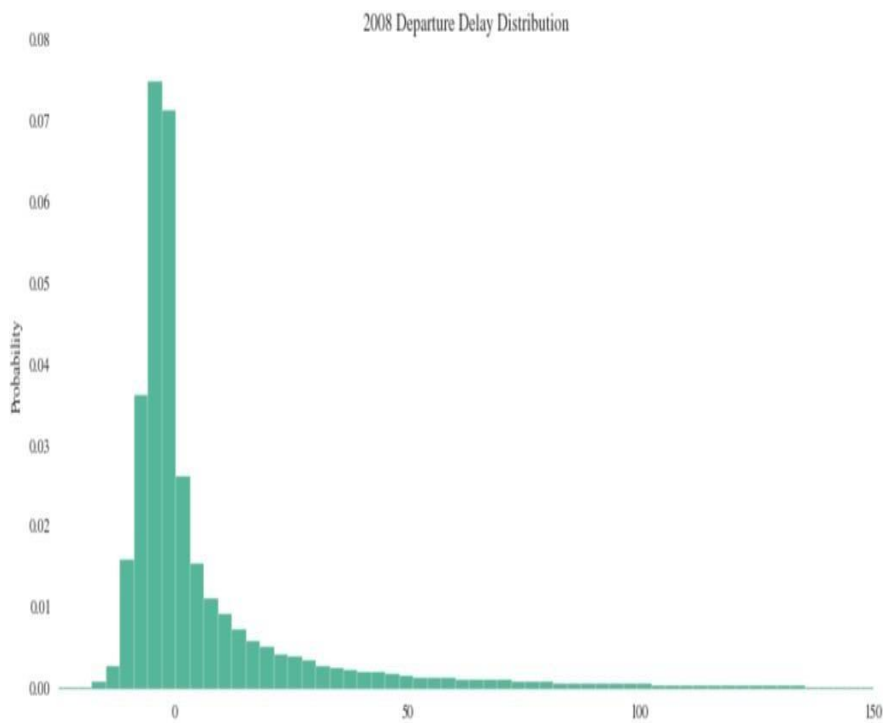


Figure (4.2): Delay Distribution

In Figure (4.1) and Figure (4.2), we notice a much higher probability of short delays. Notice the long right-hand tails from Figure(4.2). Some flights are delayed for very long time, e.g., over two hours.

On the other hand, the delays were centered around just below zero. In both cases, the mode of the distribution is less than zero, meaning most of the flights leave from the gate and arrive at the gate even before the published schedule time of departure and arrival.

As we will show below, the more extended delays cancel out, the shorter negative delays (advances), leading to average delays that are above zero.

The x-axis for the two plots is to scale. As a result, we can see that the arrival delay distribution, compared with the departure delay distribution, leans toward left. The scheduled time of an event defines a flight delay compared to the actual time of the event. Airlines usually put extra buffer time on a flight to ensure on-time arrival. Therefore, the departure delay and arrival delay distributions difference indicate that some departure delays were recovered during the flights due to the extra amount of time embedded in the flight time between two airports.

4.1.2. Average Departure and Arrival by Month

Next, we consider the impact of the months on the delays. We would expect that winter months have the most extended delay. A column chart with departure and arrival delay in minutes plotted by month is the most effective way to see the potential effects of the months.

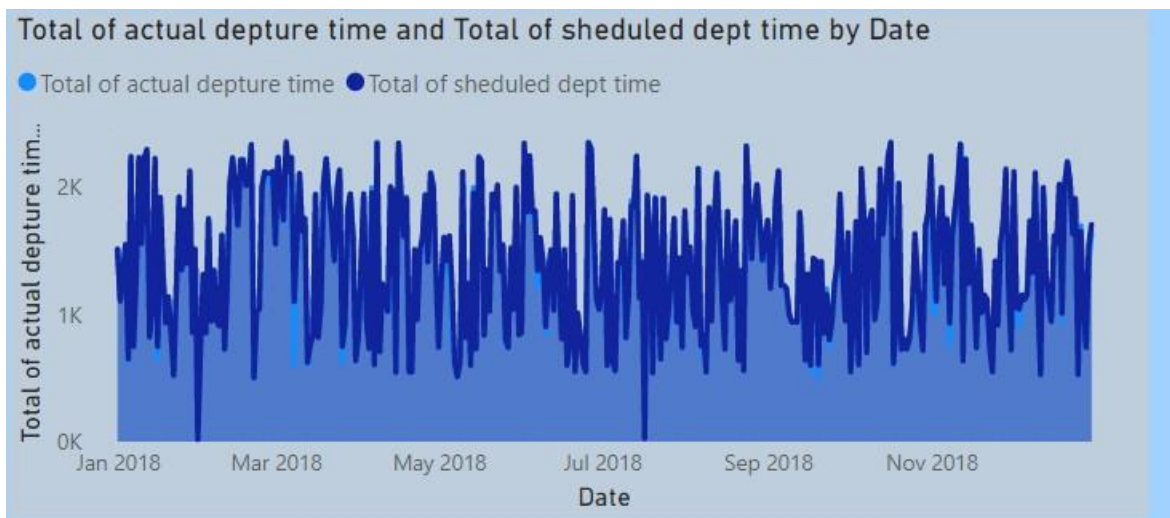


Figure (4.3): Avg Departure

Figure (vi) and (vii) clearly show that for both departures and arrivals, the impact of December is clear - the highest delays are in that month. On the other hand, September, October, and November are the months with the least amount of delay. For the summer, June and July are marked by higher delays. Also, February posts high delay values as well.

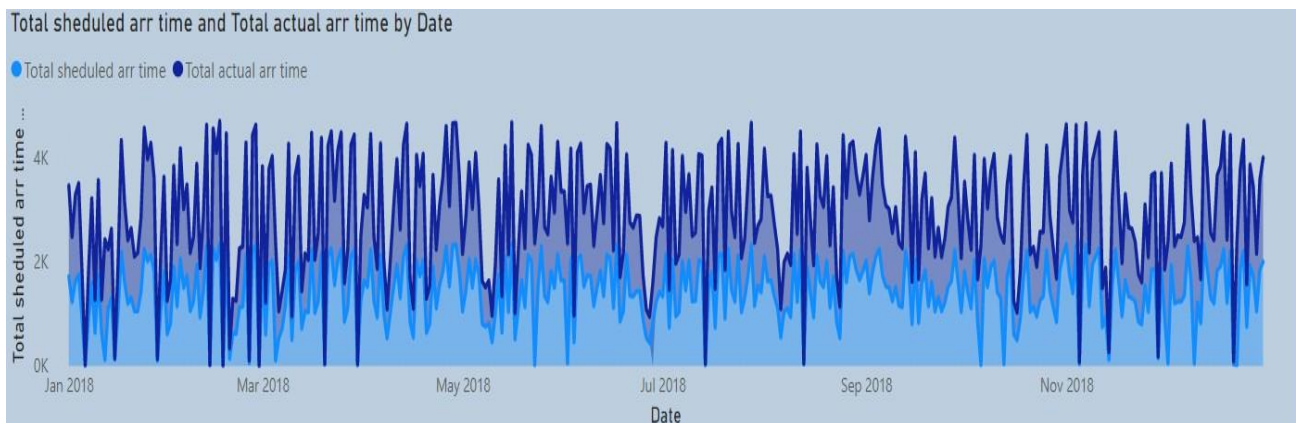


Figure (4.4): Avg arrival

Also, in Summer, thunderstorms i areas can cause high delay impact to the rest of the country. A snowstorm/storm may only affect operations at an airport or two. However, delay propagation, which marks as the significant contributor to the flight delay, can cause ripple effects on delay to downstream flight operations. The time of day should have an impact. Normally, flight delays cumulate throughout the day through a knock-on effect, where delayed flights provoke other delays because of tight schedules and runway congestion. We plot the mean delay by an hour of the day in a column chart.

4.1.3. Mean Delays mins by date

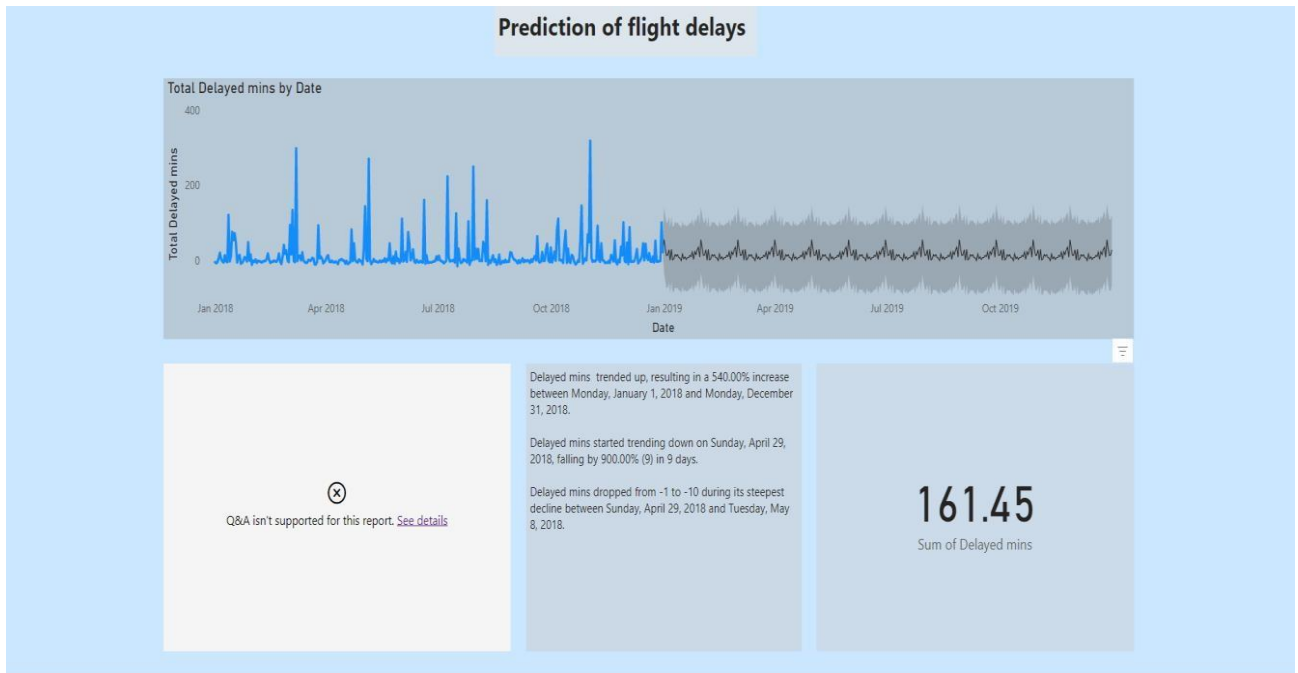


Figure (4.5.): Means Delay

In this figure, we can notice that x-axis is year by month and y-axis is date and the blue color waves are representing the current year delayed minutes.

And by using forecast method in power Bi, we can predict the future flight delays of upcoming year, represents in black color waves.

Question and answer model used in the *figure (4.5)*, is the easy way to convey the message, you can ask preferred questions to the model it will the appropriate answer to your question. For example you need the delayed mins of particular month ,you no need to see the graph you see in the q & a column itself.

By this project the sum of delayed mins is 161.45 mins.

4.1.4. Total Delayed by Months

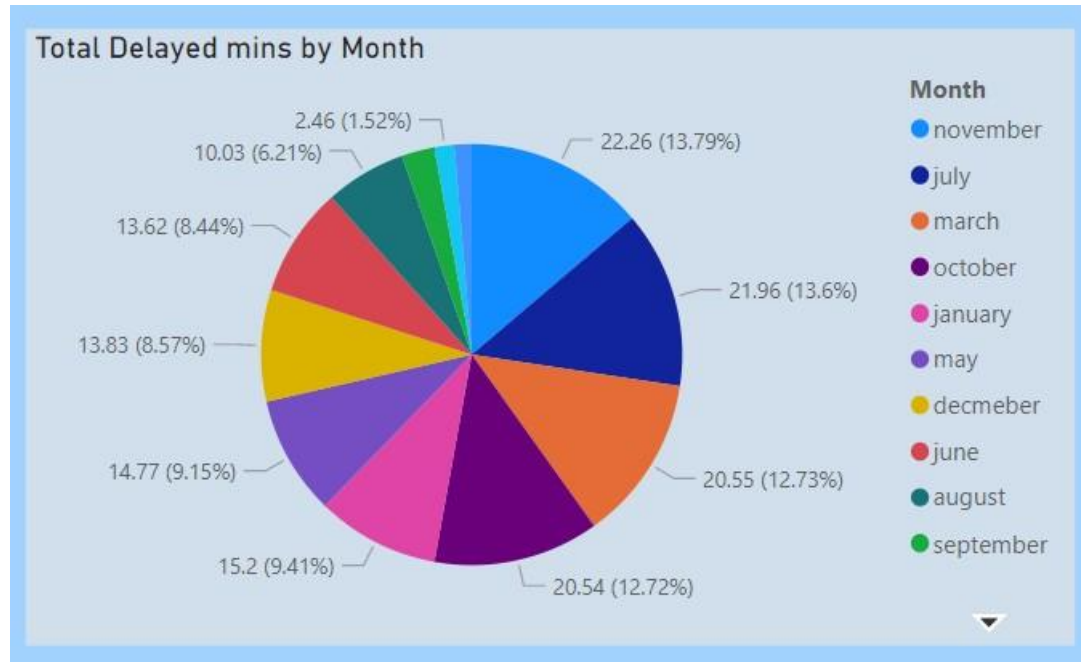


Figure (4.6): Total Delayed Mins

This is a pie chart model for total delayed mins by month. Here the different colors denote the months January to December. Each month is calculated and denoted by average mins.

From this *figure(4.6)* we can say November month has most delayed mins and most average of 13.79% and second most average contains 13.6 % by July month and 12.73% by march month and 12.72% by October month.

By this way you can calculate the delayed mins by month

4.2. TYPES OF DELAY

4.2.1. Average delay of luggage and technical issues by month

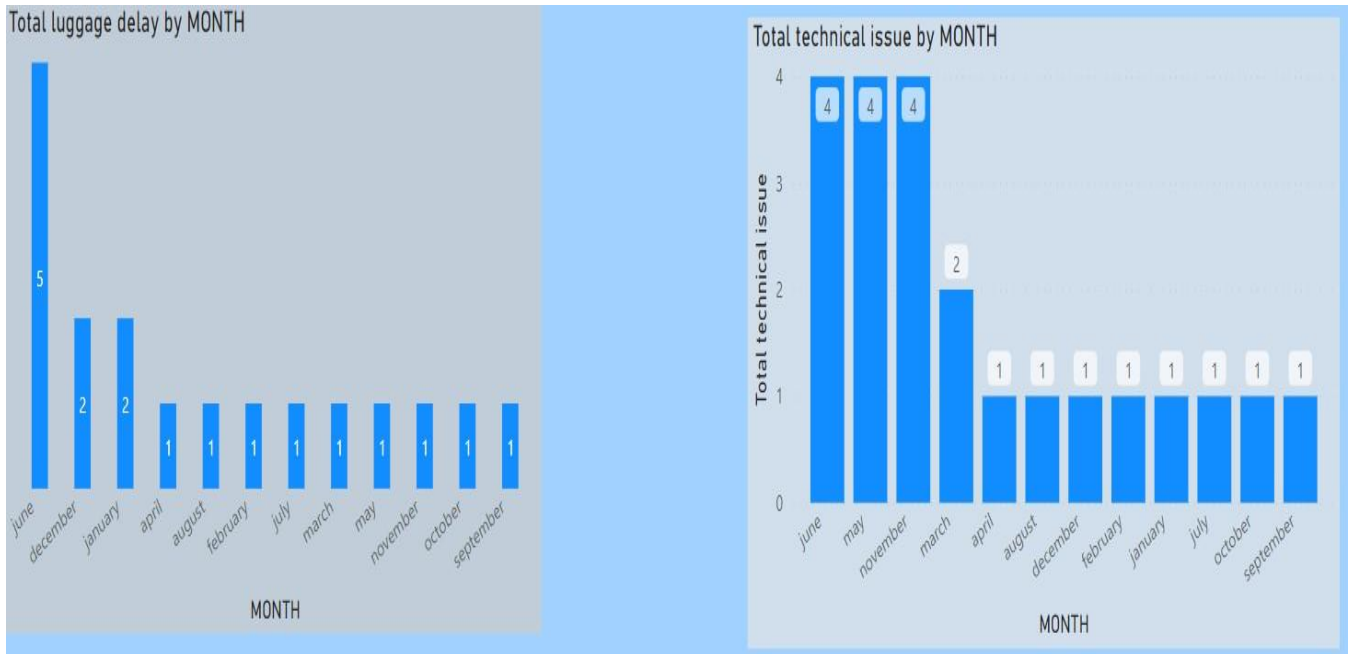


Figure (4.7): Avg delay of luggage and tech issues

Figure (4.7.) shows the bar graph model of average delay of luggage and technical issues by month.

The x axis denotes the month and y axis denotes the total number of delays for both diagrams.

From diagram 1, June month arrives the largest no of delay from luggage delay.

From diagram 2, June, may, November achieves the largest number of delays from technical issues.

4.2.2. Average of security and weather delay by month

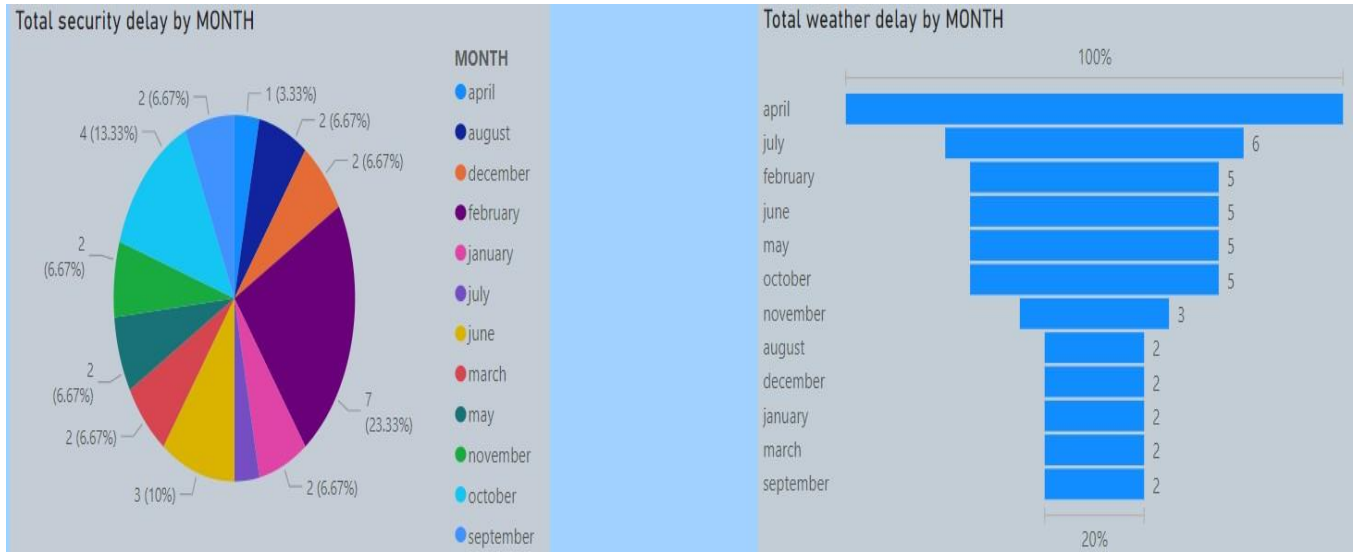


Figure (4.8): Avg of Security and weather delay

From the *figure(4.8)*, the pie chart model and bar graph model denote the total security delay and weather delay by month. In pie chart diagram, the major delay is happened in February month of percentage 23.33. the different colors denote the month of a year over a year

And the bar chart represents the total weather delay here the scale is varying from 20% to 100 % by month. From the chart we can say April month achieves high percentage of other months 100% and July comes next of 60%, February, June , may, October of 50% and August, December , January, march , and September of 20% delays of a year.

4.2.3. Average of Passenger and Propagation Delay by month

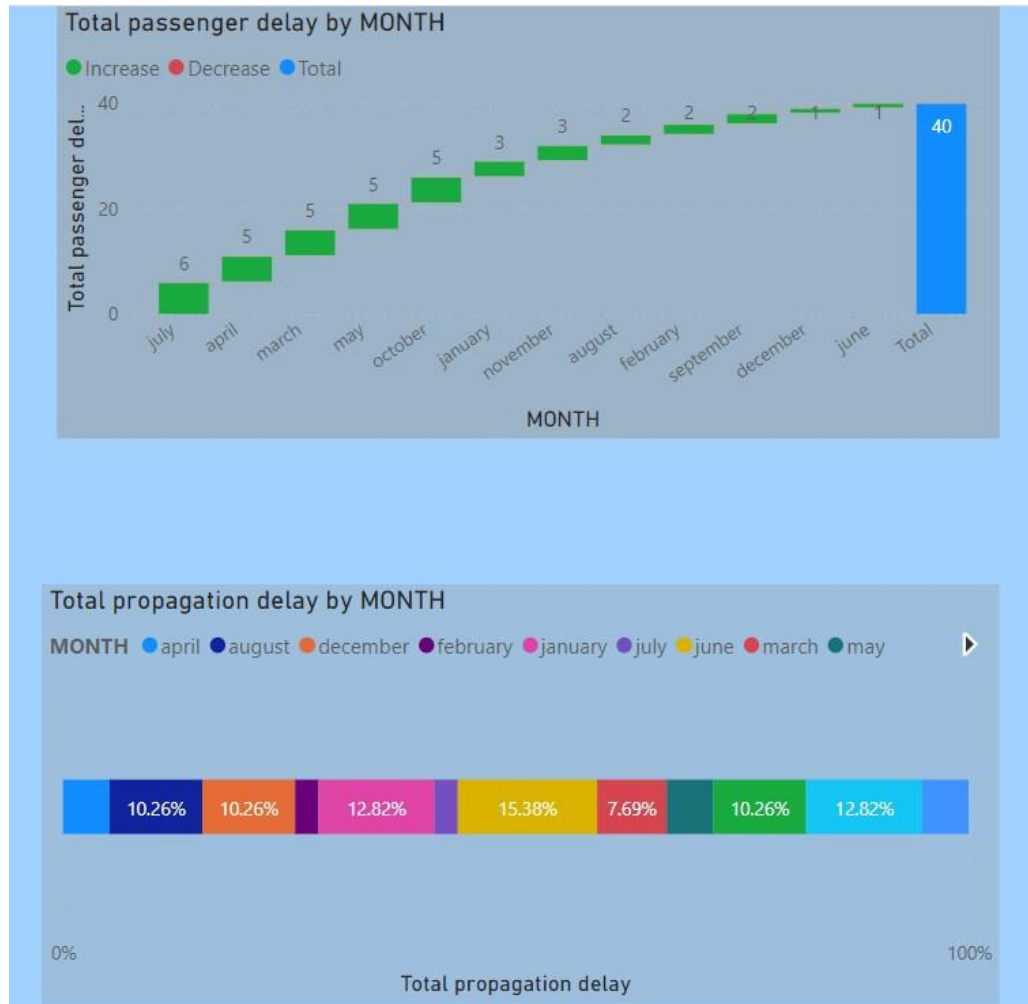


Figure (4.9): Avg Passenger and Propagation delay

Figure (4.9) represents two diagrams top one is the total passenger delay and bottom denotes the total propagation delay.

In first diagram x-axis denote the months and y -axis denote number of passengers. The total passenger's delays varying from 0 to 40 %, so from the diagram we can conclude that July highest numbers of passenger's delays.

In second diagram the months are varying from 0 to 100%, from the bottom diagram June has the highest signal delays of 15.38%

4.2.4. Average of number of delays by month

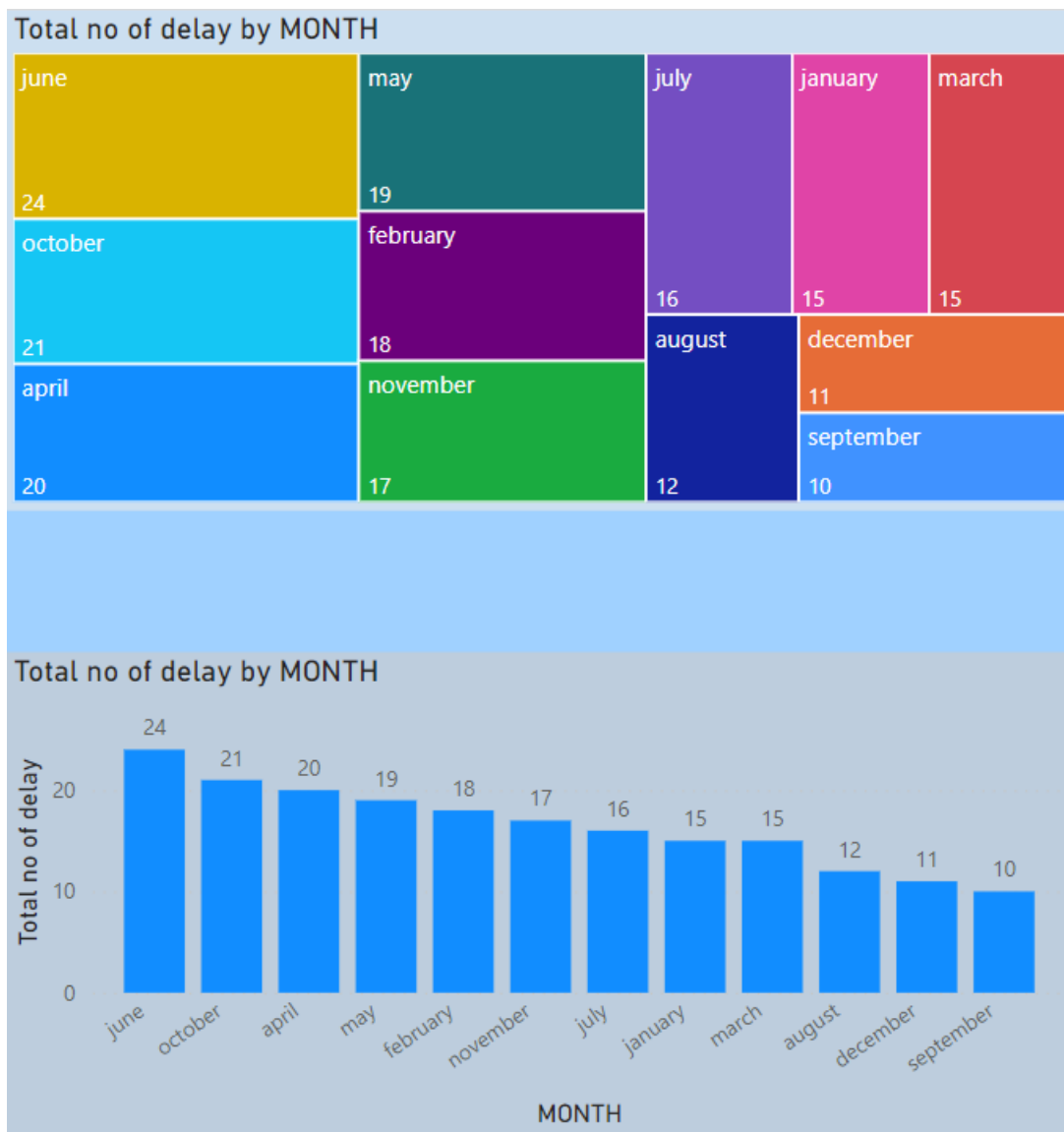


Figure (4.10): Total no.of Delay

Here in the *figure (4.10)* consists of two diagrams, both are same data but different models namely total no of delays by month.

From 1st diagram we can say that June as the highest number of delays of 24%. From 2nd diagram x-axis as months and y-axis as number of delays, by the chart it implies the June as the highest percentage of 24% which ranges from 0 to 20%.

5. CONCLUSION and FUTURE WORK

5.1. CONCLUSION

In this project, we use flight data, weather, and demand data to predict flight departure delay. In the end, our model correctly predicts 91% of the non-delayed flights. However, the delayed flights are only correctly predicted 41% of time. As a result, there can be additional features related to the causes of flight delay that are not yet discovered using our existing data sources.

In the second part of the project, we can see that it is possible to predict flight delay patterns from power bi models. Without more data, we cannot make a robust model and find out the role of related factors and chance on these results. However, as a proof of concept, there is potential for these results. It may be possible to routinely use to ascertain an understanding of concurrent airline delays and traffic patterns, which could be useful in a variety of circumstances

5.2. FUTURE WORK

This project is based on data analysis from year 2018. A large dataset is available from 1987-2008 but handling a bigger dataset requires a great amount of preprocessing and cleaning of the data.

Therefore, the future work of this project includes incorporating a larger dataset. There are many ways to preprocess a larger dataset like running a Spark cluster over a server or using a cloud-based services like AWS and Azure to process the data. With the new advancement in the field of deep learning, we can use Neural Networks algorithm on the flight and weather data. Neural Network works on the pattern matching methodology. It is divided into three basic parts for data modelling that includes feed forward networks, feedback networks, and self-organization network. Feed-forward and feedback networks are generally used in the areas of prediction, pattern recognition, associative memory, and optimization calculation, whereas self-organization networks are generally used in cluster analysis. Neural Network offers distributed computer architecture with important learning abilities to represent nonlinear relationships.

Also, the scope of this project is very much confined to flight and weather data of United States, but we can include more countries like China, India, and Russia. Expanding the scope of this project, we can also add the flight data from international flights and not just restrict our self to the domestic flights.

REFERENCES

- [1] A. B. Guy, "Flight delays cost \$32.9 billion, passengers foot half the bill".
[Online] Available : https://news.berkeley.edu/2010/10/18/flight_delays/3/.
[Accessed on
June 2017].
- [2] M. Abdel-Aty, C. Lee, Y. Bai, X. Li and M. Michalak, "Detecting periodic patterns of arrival delay", *Journal of Air Transport Management*,, Volume 13(6), pp. 355– 361, November, 2007.
- [3] S. AhmadBeygi, A. Cohn and M. Lapp, "Decreasing Airline Delay Propagation By Re-Allocating Scheduled Slack", *Annual Conference*, Boston, 2008.
- [4] A. A. Simmons, "Flight Delay Forecast due to Weather Using Data Mining", M.S. Disseration, University of the Basque Country, Department of Computer Science, 2015.
- [5] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms", *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, Sacramento, CA, USA, 2016.
- [6] L. Schaefer and D. Millner, "Flight Delay Propagation Analysis With The Detailed Policy Assessment Tool", *Man and Cybernetics Conference*, Tucson, AZ, 2001.
- [7] B. Liu "Sentiment Analysis and Opinion Mining Synthesis", *Morgan & ClaypoolPublishers*, p. 167, 2012.

