A Machine Learning Approach Predicting Flight Arrival Delay Reduction for Delta Airlines

MSc Research Project
Data Analytics

ENWERE CHIBUIKE KENNETH
Student ID: X18178090

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Enwere Chibuike Kenneth …… …………………………………………………………………………………………………

Student ID: X18178090 ……………………………………………………………………………………………………..……

Programme: Data Analytics ……………………………………………………………… **Year:** …2019……………………..

Module: MSc Research Project ………………………………………………………………………………………………….………

Supervisor: Dr. Cristina Muntean …………………………………………………………………………………………….………

Submission Due Date: 12/12/2019 ………………………………………………………………………………………………….………

Project Title: A Machine Learning Approach Predicting Flight Arrival Delay Reduction for Delta Airlines ………………………………………………………………………………………………….………

Word Count: 8644 …………………………………… **Page Count**…………21……………………………………….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………………………………………………………………………………………………………

Date: ………………………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only

Signature:	
Date:	
Penalty Applied (if applicable):	

A Machine Learning Approach Predicting Flight Arrival Delay Reduction for Delta Airlines

Enwere Chibuike Kenneth
X18178090

Abstract

The efficiency of the air transport system has been attributed to it been the fastest means of transportation, which has over the years gained strong reliability from customers, however flight arrival delays have become imperative in the airline industry as delays in this phase of the flight depends on the organization of the airspace and runway availability. Cancellation and diversion of flights due to either air traffic operational short comings or unforeseen circumstances such as weather brings bad reputation to airlines and unnecessary expense due to reimbursement of customers. This research work was focused on analysing the arrival delay of domestic flights operated by Delta airlines in the United States using several supervised machine learning algorithms which includes Gradient Boosting Classifier, Decision trees, Naïve Bayes, Support Vector Machine, Random forest and logistic regression, while comparing their performance based on Accuracy, Precision, Recall and Specificity in order to find models with the best accuracy. Each predictive model was trained by collecting data from the Bureau of transport statistics (BTS), and the data contained Delta airlines operated flights for the year 2017 while using feature selection to get relevant attributes used in the analysis. The gradient boosting classifier showed the best accuracy of 70% as compared to other models. This research will provide insights to Delta airlines by shedding light on different aspects where flight arrival delays can be improved.

Keywords: Random Forest, Gradient boosting classifier, Flight Arrival Delay, Delta airlines, Support vector classifier, logistic regression, Naïve Bayes.

.

1 Introduction

The complexity involved in air travel is far from just the mechanics of take-off and landing the aircraft, flight delays inherently affects both customers and industry players alike. This heavy reliance on air travel has led to the strain of flight systems eventually leading to delays. Flight delay is a major problem in the United States with increasing delays in flight putting remarkable pressure on air travel system. These inefficiency in air transport caused by delays have indirect effects on the economy, affecting other businesses dependent on-air transport. Every minute a flight is delayed there is a large increase in the amount of fuel consumed which is detrimental to the environment and adds cost to the individual aircraft (Ison et al., 2015). The time of arrival of flights are mainly influenced by the timeliness of the departure, however the stability of flight operations is dependent on the punctuality of arriving flights. Late aircraft arrival even in a case of on time departure can be attributed to bad weather conditions which moves from one location to the other, airport congestion, long taxi-times and other external factors.

1.1 Background And Motivation

Delta Airlines, Inc. (DL) is one of Americas biggest and most influential airline ascertained by a net revenue of over 1 billion dollars. Carrying over 100 million passengers in a year in 1998, Delta airlines uses a vast number of flight networks and operates a flight system called hub and spoke (Schumacher 1999). The hub and spoke system make use of certain key airports to be a starting point for a substantial amount of flights providing passengers with the shortest flight distance to their destination. The key hubs used by Delta airlines includes Atlanta, Los Angeles, Boston, Seattle JFK and LaGuardia just to name a few. The On-time performance of arriving flights has long been affected by congested airports, flight schedules, time of day, season and most importantly weather. Flight delays associated to weather are as a result of bad visibility caused by fog, electric storms, heavy snow or even strong winds. This condition leads to the aircrafts spaced further apart reducing the number of aircraft that lands within a given timeframe, limitation in flight path which may lead to rerouting the plane to another airport or even spending so much time to de-ice the plane**.**

Several studies have described stages in flight where delays occur and the root causes. The Department of transportation DOT categorizes delay as airborne delay, taxi-in delay, gate delay and taxi out delay. From the data analysed, 84% of most delays happens on ground i.e. arrival at the gate, taxi out and taxi in, while about 76% occurs before the aircraft takes off i.e. leaving gate taxi out which implies that ground delay has more impact so therefore the arrival delay considered in this thesis is the value of delay counting at the gate (Mueller et al., 2002).

According to Bureau of Transportation Statistics BTS, the punctuality of an aircraft can be attributed to the time in minutes the aircraft get to the gate. There is a window for which flights are to arrive and anything outside that is considered as a late flight. Early or late arrival of flights can cause a ripple effect on things like airport capacity and gate availability. Machine learning involves getting data and automating models to find trends and patterns to aid business improvement, decisions and provide solution to problems (Assem et al.; 2015). Several research work on flight delays has been explored shedding light on both departure and arrival delays. (Etani, 2019) developed a predictive model for on-time arrival flight by finding the correlation between weather and flight data. Very little research has been carried out on the arrival delay associated with delta airlines, so this research is focused on using various supervised machine learning algorithm to better understand the factors that affect arrival delays and gain useful insight towards predicting flight arrival delay reduction in the major hubs used by delta airlines.

1.2 Research Question

The addressed research question sought to be answered is as follows

"*To what extent can the accuracy of models efficiently predict flight arrival delays in delta airlines and improvement made to flight efficiency* "

With this research question set, the following research objectives are addressed

- To find out the main airports used by Delta airlines that flight arrival delay affects the most.
- To find out the maximum flight arrival delay in the selected months
- To compare various machine learning algorithms such as (Gradient boosting classifier, Random forest, logistic regression, Support vector machine and Naïve Bayes) to find out the best performing algorithm which has more accuracy on predicting arrival delays.

1.3 Research Contribution

The project contribution clearly demonstrate how theory is put to practice in the analysis of flight arrival delay in Delta Airlines and find the effect of this delay in the respective airports used by Delta airlines. This research was drilled down this way to be innovative as several research has been carried out on flight delay. Furthermore, this research uses feature selection to improve prediction performance, a modified CRISP-DM methodology together with comparative analysis to find the best performing model that highlights airports used by Delta airlines which have high arrival delays and where improvement can be made to increase customer loyalty.

The structure of the research is organized in the following chapters: Chapter 2 comprises of the literature which shows past research work on flight delay and several machine learning algorithms used. Chapter 3 specifies and describes the proposed methodology used to bring the research to completion. Chapter 4 shows the design specifications which includes the design architecture and design techniques used. Chapter 5 shows the implementation which talks about the several methodologies employed the value of the implementation to stakeholders and technologies and tools used. Chapter 6 shows the proposed evaluation, highlighting the evaluation methods metrics and criteria. Chapter 7 is the research proposal conclusion. Chapter 8 shows the proposed project plan indicating the tasks, milestones, effort, resources and duration of the project.

2 Related Work

2.1 Introduction

The idea of flight delay is not a new topic which is why a lot of research and analysis has been carried out to provide and assist in good decision making to improve customer satisfaction. The purpose of the literature review is to put the meaning of this work into perspective. This chapter has two sections. The first section talks about important analysis on past research done on flight delay, which includes the background of flight delay, its effect on the U.S. economy and reviews based on statistical models. The second section focuses on machine learning algorithms and methodologies used in the analysis of flight delays most especially arrival delays and their impact to both airlines and passengers, finally summary and conclusion of the chapter.

2.2 Background Of Flight Delay And Its Effects On U.S. Economy

Flight scheduling is a cumbersome process as it deals with using valuable resources to tackle fluctuations in demand made by arriving aircrafts. In the United states of America, almost one in four flights arrived 15 minutes late at their destination. These late arrivals where a direct result of aviation system failure to meet traffic demands BTS (2009).

In 2007, a report presented by the Joint Economic Committee estimated the overall cost of flight delay on the U.S. economy to be $41 billion. This comprises of $19 billion on airline operating cost and $12 billion on passenger delay (Schumer and Maloney 2008). Statistics from the U.S. Department of Transportation (DOT) showed that they were 7 million flight operation in America in 2007 out which 2 million were arrival delays. The Federal Aviation Administration (FAA) was obliged to introduce ground delay program (GDP) to deal with arrival delay reduction by ensuring that aircrafts are not allowed to take off until there is high assurance of flight completion with little or no delay.

There are fundamentally two types of delay these are arrival and departure delay (BTS, 2011). Arrival delays are acknowledged when a scheduled flight that is airborne lands 15 minutes after its schedule time of arrival and this delay is calculated by subtracting the scheduled arrival time from the actual time of arrival. On the other hand, departure delays are disruptions experienced when a scheduled flight fails to take-off after 15 minutes of scheduled time. There are several hourly arrivals for each flight which includes the departures allotted per airport however, since the focus is on arrival delay, it is essential to understand that a flight arrival delay is the responsibility of an airline. To eradicate this type of delay from the system, researches have been done with the intention to accurately predict place and time where the delay will occur along with the delay justification. Further explanation on flight arrival delay have been related to several sources which includes the hub and spoke model design for U.S. airports, slot control and runway and airspace congestion.

The Hub and spoke network were first used by Delta airlines, to improve the routing capacity of passengers and the movement of goods from one point to another. Its benefits according to (Bryan et al.; 1999) can be seen in the reduction of underutilized routes, improving profit made by carriers by using larger aircrafts to move more passengers, however the hub and spoke method did not solve delay rather increased passenger anxiety due to long haul flights caused by indirect routes.

Next was the slot control method used in airport depending on their size and location to determine the number of departure and arrival of an air carrier. Over the years, selling and leasing of slots was considered an alternative to make more money (Donohue 2004). This in turn leads to airspace congestion, hence causing delays.

For the runway and airspace congestion, (Schaefer and Millner 2001) stated that the main cause of flight arrival delay was due to overcrowding and insufficient airport. He further explained that 94 percent of the 450 airports were handled by the top 60 U.S. airport out of which 20 of those airports handled half of the air traffic.

In a work proposed by (Tandale et al., 2009) results from queuing theory was employed to find the relationship between trajectory uncertainties and the efficiency of traffic flow in the airspace. This research work used approximate queuing network analysis to model arrival and service processes using first and second moments. Another researcher talked about queuing delay analysis by the characterization of traffic flow. This was done by analysing metering delays in the cruise phase instead of the descent phase (Kimura et al., 2011).

According to (Mueller et al., 2002) analysis on departure and arrival delay were carried out using statistical method focusing on traffic management systems. Further research work by (Tu et al., 2008) using statistical data were used to develop models, predicting flight delays with short- and long-term patterns. The major drawback to this was that relying solely on statistical methods gave a biased result. Other researchers paid more focus to predicting and knowing the duration and grade of flight delays, (Zonglei et al., 2008) created an alarming system that notifies if a flight will have long delays or not.

Furthermore, (Rebello et al., 2014) analysed the taxi in and out as well as wheel on/off time to predict delays in departure to see patterns and dependencies of airport within the air traffic network systems.

An econometric model by (Hansen et al., 2006) was developed to predict the average daily delay which combines the effects of arrival queuing, seasonal effects and weather condition. This analysis was conducted to find the effect of arrival queuing on a particular time of day and the effect on scheduled arrivals. The result showed that queuing has a significant effect in the morning as compared to afternoon and evenings. A research model proposed by (Abdel-Aty et al., 2007) had an analysis on detecting the periodic patterns of arrival delay of domestic flights at the Orlando International airport. Repeating patterns in flight demand and weather were identified and their effect on flight delay were further investigated. A two-stage approach using mathematical frequency and statistical analysis was used. The result of both approaches showed daily, weekly and seasonal patterns and also showed how time of day, the days of the week and scheduled time intervals between completed flights correlated with arrival delay.

Further investigation by (Rosen 2002) studied the amount of change in flight times resulting from the constant changes in passenger demand in terms of infrastructure. From this research, the result showed that as demand to get better infrastructure increased, delays increased as well.

Conclusively the background of flight delays and its effect to the economy has been discussed in this section

2.3 Different Algorithms Used In Flight Delay

In the last few years, researchers have used machine learning together with several algorithms, deep learning and big data techniques to predict flight delay. This section talks about various algorithms used by researchers.

Chakrabarty in his research work used four supervised machine learning algorithm which includes Random forest, Support vector machine (SVM) and Gradient Boosting Classier (GBC) to predict flight arrival delay for American airlines, highlighting the busiest 5 airports in America (Chakrabarty et al., 2019).Results were that gradient boosting classifier gave the best prediction of 79.7% for arrival delay. The fuzzy support vector machine with weighted margins was used by (Chen et al., 2008) to build a model that detects and gives an early warning on flight delay and as well determine delay grades of flights. The weighted margin SVM performed better as compared to the one against one SVM.

A research conducted by (Liu et al., 2008) showed the Estimation of arrival delays by taking one busy hub airport to examine the impact of propagation into and out the airport. It was found arrival delay is the source of other delay. In another study, the optimization of flight arrival together with sequencing, and assignment of runways was proposed by (Wang et al.,2014) and was intended to reduce flight arrival delays caused by air traffic congestion. Furthermore, genetic algorithm was verified, and it showed that it can improve the efficiency of air traffic control by reducing congestion in busy airport terminals. Some researchers like Klein looked in the direction of the integration of airport weather forecast, and scheduled flight, carrying out analysis to predict airport delay time using a metric called Weather impacted traffic index (Klein et al., 2010). Similarly, research done by (Choi et al., 2016) proposed a model to classify airline delay caused by weather conditions. The study pointed at using weather and traffic data and supervised machine learning algorithms such as decision tree, random forest, AdaBoost and k-Nearest Neighbour to build models to predict individual flight delay. The researchers came to conclusion by saying the model can be further improved and the performance classifiers achieved if the false positive and negative are considered. Gaps in flight delay has been explored by analysing air traffic data to find patterns in air traffic delay. (Manna et al., 2017) combined regression model to analysis patterns in air traffic delays, Gradient boosting decision tree showed high accuracy in modelling sequential data which involved predicting efficiently daily sequence in arrival and departure flights.

The on-time performance of flights by researchers (Thiagarajan et al., 2017) used a two-stage predictive model (Classification and Regression) and supervised machine learning algorithms to predict arrival delays and departure delays using the schedule for flights and weather features. The study revealed that in the two stages, Gradient Boosting Classifier performed the best for classification while Extra-Tress Regressor also had the best performance for regression. On the other hand, a broader aspect of machine learning has been used in analysing flight delay this can be seen in the use of deep learning approach. The effectiveness of deep learning due to remarkable results in several machine learning task has seen researchers drawn to using this approach. A detailed analysis of patterns in air traffic delays using deep learning models was proposed by (Kim et al., 2016). This research was based on using Recurrent Neural Networks to predict flight delay and the increase in its accuracy was as a result of deeper architectures.

On the other hand, a systems approach for scheduling flight arrival delay by (Khanmohammadi et al., 2014) was considered using General systems problem solving (GSPS), Adaptive network based fuzzy inference system (ANFIS) and GSPS-CI framework to schedule inbound flight to JFK airport. Fuzzy decision-making procedure was applied to the arriving flights while GSPS methodology was tested on the JFK airport.

In conclusion, this section threw light on the background of flight delay and its effect on the U.S. economy, the different algorithm used in flight delays by several researchers and investigated specific papers on arrival delay. The gaps found showed that there is little research in the area of flight arrival delays especially in relation to Delta airlines so therefore carrying out analysis here using supervised machine learning algorithm will efficiently give recommendation on how flight arrivals delays can be minimized for Delta airlines.

3 Research Methodology

3.1 Introduction

This chapter includes step by step procedure taken in the research, the techniques and tools used as well as the design process flow. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is the approach taken for this research. Furthermore, a two (2) tier design was used to show the process it takes from getting the data to algorithms applied, tools used and returned visualized results.

3.2 Modified Methodology

CRISP-DM established in 1996 with a six-stage process flow (Chapman et al., 2000) is a proven way used by the industry to guide data mining efforts, however it is famous, and a leading methodology used by numerous researchers to address different problems. This research work follows the Cross-industry Standard Process for data mining (CRISP-DM) because its targeted towards getting business insight for Delta airlines with a few modified phases.

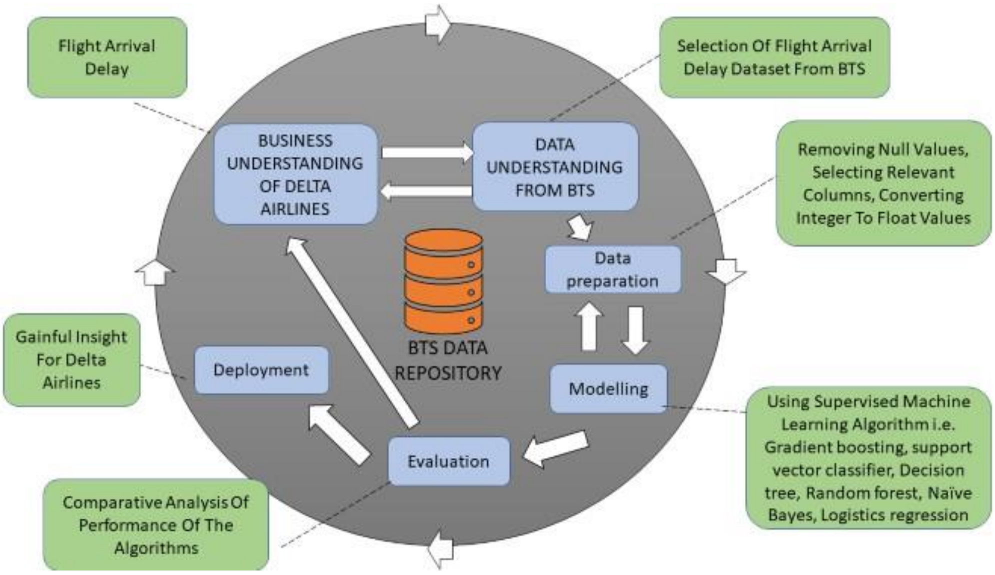

Flight Arrival Delay
Selection Of Flight Arrival Delay Dataset From BTS
BUSINESS UNDERSTANDING OF DELTA AIRLINES
DATA UNDERSTANDING FROM BTS
Removing Null Values, Selecting Relevant Columns, Converting Integer To Float Values
Data preparation
BTS DATA REPOSITORY
Gainful Insight For Delta Airlines
Deployment
Modelling
Using Supervised Machine Learning Algorithm i.e. Gradient boosting, support vector classifier, Decision tree, Random forest, Naïve Bayes, Logistics regression
Evaluation
Comparative Analysis Of Performance Of The Algorithms

Figure 1: Modified CRISP-DM Methodology

The modified CRISP-DM fundamentally has 6 phases with arrows pointing out how each phase rely on each other and provides descriptions and the task associated with phase

1. **Business Understanding**: The business understanding in this research is for the airline industry with focus on flight arrival delay in Delta airlines. From the basis of this understanding, useful knowledge was acquired and subsequently a gap was found which lead to the data understanding.

2. **Data Understanding:** The dataset used in this research was sourced from the Bureau of Transportation Statistics (BTS) website. The website provided the ability to highlight different attributes to include in the dataset and this formed the foundation of the research. Data containing all flights arriving at Hartsfield Jackson International airport ATL in the year 2017 was selected for analysis, while the summer period which includes the months of June, July, august and September were as well highlighted due to the fact people fly more during those months. Finally, Delta airline was the airline chosen to be analysed. All these were done, with the business understanding in mind. The next phase is the data preparation.

3. **Data Preparation:** In this phase, the dataset was prepared by carrying out data cleaning operations such as removing duplicated rows and removing Null values. Further to this, feature engineering was applied to get relevant features to build the model and also dropping irrelevant columns that adds no meaning to the analysis i.e. a column in the dataset named delay national aviation system etc. Data transformation which includes converting integers to floats was also done. The Anaconda package manager was used in a windows environment with the Python programming Language version 3 and Jupyter notebook embedded in it. The reason for using anaconda as follows
   a) Anaconda provides robust tools and packages and interactive environments that supports easy project deployment

b) With Jupyter notebook embedded in it not just writing codes visualizations can also be carried out

4. **Data Modelling:** In the modelling phase, the scikit-learn package was helpful in dividing the data i.e. train_test_split as the dataset was split into 70% for train and 30% for testing and this was done to avoid over fitting. Furthermore, models such as Naïve Bayes, Random Forest, Support Vector Machine, K Nearest Neighbour and Gradient boosting classifier were trained and their performance was compared based on these metrics Accuracy, Recall, Specificity and Precision

5. **Evaluation:** In this phase, the evaluation was categorized into two parts this are Assessment of the model performance and assessment the visualized output. The assessment of model performance had the output of each model compared over one another to find the best performing model. From the performance of these algorithm the Gradient boosting classifier had the best prediction result as compared to the rest of the models while Naïve Bayes had the lowest percentage. However, for the assessment of the visualized output the matplotlib was used to derive important insights to achieve greater understanding by the creation of detailed graphs. This was achieved by plotting a histogram stating origins of flights delayed on arrival, a scatter plot showing if distance affects delays and a line cart chart to determine which months had more arrival delays. The months of June and July had the most arrival delay during the summer season as compared to the other months.

6. **Deployment:** For this phase gainful insights got from the analysis of flight arrival delay will provide Delta airlines with useful information to improve key areas that creates arrival delay

4 Design Specification

4.1 Design Architecture

In the design architecture**,** a two-tier design is chosen and comprises of two parts namely the Client Tier (1st Tier) and the Business Logic Tier (2nd Tier). According to Mohammed (2007) there is a direct communication between the client and the database. This research work follows the two-tier design architecture, and this is because the model requires no backend operation**.**

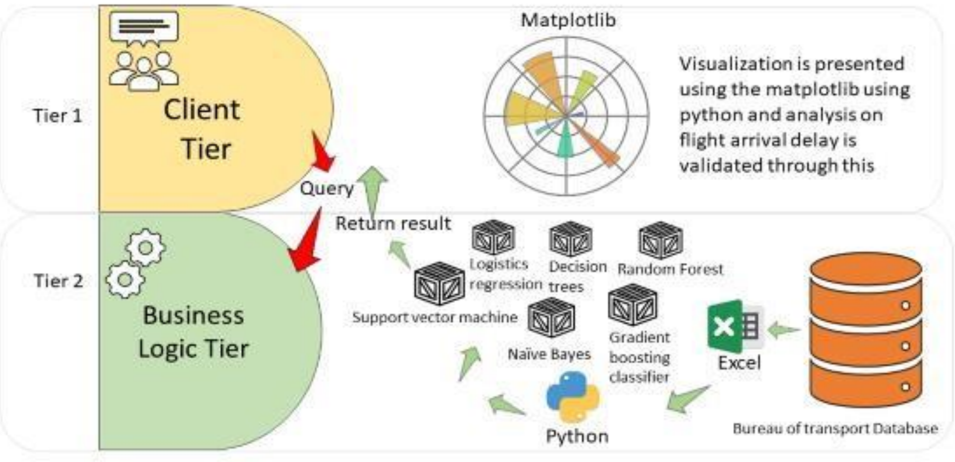

Tier 1
Client Tier
Matplotlib
Visualization is presented using the matplotlib using python and analysis on flight arrival delay is validated through this
Query
Return result
Tier 2
Business Logic Tier
Logistics regression
Decision trees
Random Forest
Support vector machine
Naïve Bayes
Gradient boosting classifier
Excel
Python
Bureau of transport Database

Figure 2: Two-tier Design Architecture

Figure 2 above summarizes how the client and business logic tier communicate, the movement of data using tools such Microsoft excel and python, different algorithms used in modelling and a returned visualized result. The architecture of the 2-tier design is made up of the client tier and the business logic tier. The Client tier handles both presentation and application whereas the business logic tier has the database. Communication between these two layers happens if the user sends a request for information in the form of a query to the business logic tier. The information in the Bureau of transport database (BTS) has dataset downloaded into Microsoft excel imported to python and undergoes pre-processing which involves cleaning and transformation of the raw data i.e. removing irrelevant columns, removing null values and feature engineering. Then the 6 different algorithms are modelled using this cleaned data and from the metric used in evaluation a returned result is taken back to the client tier and presented using the matplotlib visualization tool in python.

4.2 Design Techniques

For the design techniques the justification of using different machine learning models has been highlighted. The models used include Gradient boosting classifier, Support vector machine, Naïve Bayes, Decision tree, Random Forest and Logistics regression.

The Gradient boosting classifier was used because it is effective in handling regression and classification task. However, by building ensembles in an incremental manner improvement in the learning process of the model was established by making use of weak learners to reduce bias.

$$F(x)\sum_{m=1}^{M} y_{mh_m(x)}$$

Figure 3: Equation showing the collection of weak learners

From figure 3 above, y_m is the weight of each learner, h_m(x) is a function in boosting called weak learners. Furthermore, series of predictor values are produced sequentially when using the gradient boosting method, so the measured average of these predictor values are calculated to generate the final predictor.

The next model is the Support vector machine (SVM) which are used through variable rankings and selection to reduce dimensionality (Bi et al., 2003). As a pattern classification technique, the idea behind using the SVM was to transform the data which in its previous space could not be split up linearly into a higher dimensional area for which a hyperplane can be easily separated. This classifier was aimed at developing a model to tackle class labels of validation data comprising of attributes only. The kernel function used in SVM was radial function.

The next model Naïve Bayes which has produced fair result spam filtering and document classification was used to find the variables distribution in the dataset by showing if a given variable is part of a given class. The Naïve Bayes can be very fast as it takes little amount of training data to estimate the variables needed. In this research the Gaussian Naïve Bayes was used .

When using the decision tree, the whole dataset depended on constructed classification models separating repeated data. The data is then divided into several subset as a result of one or more attributes and then subsequent split of subset into smaller subset until the appropriate size is reached. In a tree structure, models can be represented across the modelling procedure as set of “if-then” rules. While representing the tree structure, the complete dataset is represented as a root node so when there is a split to the data, multiple child nodes are formed, and they correspond to the split data subsets. In a case were a node cannot be broken further it is called a leaf. Furthermore, parameter values were set to control the size of the trees e.g. min_sample_leaf was set to 1 while max_depth was set to 10. This was done to decrease memory consumption as an unpruned tree can be too large for a given dataset.

Next the was the Random forest. In this project, the random forests were meta estimators that was consistent with several decision tree classifications for different sub-samples of the data set and showed average predictive accuracy and help in controlling over-fitting. Finally, logistic regression in the project uses a multiclass case were the training algorithm takes the one vs rest scheme. This is to determine the expected outcome which has discrete possible values. The probability of high delay occurring or not can be seen in this equation

$$g(x) = \text{In}\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \ldots.\ \beta_n\, x_n$$

Figure 4: Logit multiple logistic regression model

where, x_n are independent variables, $\pi(x)$ is the probability of a delayed flight divided by the total amount of flight. β_n in the equation, is the model coefficient which determines the odd ratio that is in simple terms the probability of arrival delay occurring or not.

5 Implementation

5.1 Introduction

The implementation chapter talks about the operations carried out in analysing flight arrival delays showing how the data was cleaned and transformed, written codes, models developed, and programming language used in the research.

5.2 Data Description

The analysis of this research work involved developing a research question and setting relevant objectives to be accomplished. This was achieved by getting the on-time performance data from Bureau of transportation statistics website (BTS). The BTS database contained delay data for both departure and arrival but for this research, arrival delay was the most important

The dataset was for the year 2017 consisting of domestic flights operated by Delta airlines, investigating and carrying out analysis on flight arrival delay during selected months. This month's include June, July, August and September and are the summer holiday period with travelling at its peak. The dataset had the total number of origin airport to be 150 and destination airport to be 150 as well, however the shape of the data was seen in its size with the number of rows to be 322199 and that of columns to be 16 however, BTS enabled the ability to manually select relevant features even before download and this formed the first phase of feature engineering. The initial dataset had 32 columns from which 16 significant columns were selected this is described below

Table 1: Description of attributes used in the research

Attributes	Description of Attributes	Data type
MONTH, DAY_OF_MONTH, DAY_OF_WEEK	Flight dates (Time period)	int
ORIGIN	Origin airport	object
DEST	Destination airport	object
CRS_DEP_TIME	CRS departure time	int
DEP_DEL15	Departure delay indicator	float
ARR_TIME, ARR_DELAY, ARR_DELAY_NEW	Actual arrival time, Difference in minute between scheduled and actual time of arrival	float
ARR_DEL15	Arrival delay indicator	float
CANCELLED	Cancelled flight indicator	int
DIVERTED	Diverted flight indicator	int

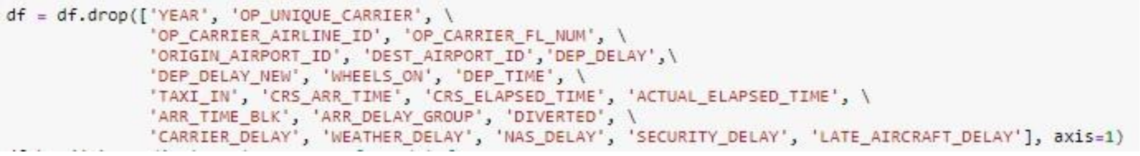
AIR_TIME	Flight time in minutes	float
FLIGHTS	Number of flights	int
DISTANCE	Distance between airports	int

Data types show the compiler how the data is to be used i.e. what tasks are to be carried out and data stored. Data type object are the categorical features in the dataset and are strings pandas. Int and float are both numeric with float having decimal point and int having no decimal point.

5.3 Data Preparation/ Transformation

Right after downloading the dataset the data had to pass through cleaning and transformation. As a result of the dataset filled with inconsistent data, in the cleaning process a quick feel of the data was carried out by looking at the rows by importing libraries such as pandas and NumPy to make data manipulation easy. Next was to load the data. Furthermore, close attention was paid to the features of the data, the expected data type and missing data that were evident. Since data cleaning helps to eliminate inconsistencies the cleaning of the data was carried out as follows

a. **Label Encoding**: In this research the label encoding involved changing strings to numbers. Using the scikit-learn label encoder, the categorical variable in the origin and destination were convert to numeric values. The datatype for DEP_DEL15 and ARR_DEL15 was converted from float to int
b. **Removing irrelevant columns:** The dataset was however filtered, and the removal of less important columns was achieved.



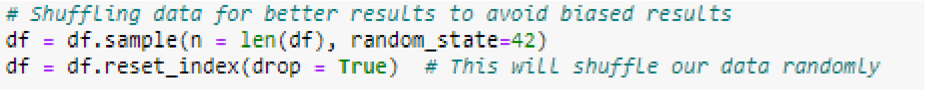
```
df = df.drop(['YEAR', 'OP_UNIQUE_CARRIER', \
              'OP_CARRIER_AIRLINE_ID', 'OP_CARRIER_FL_NUM', \
              'ORIGIN_AIRPORT_ID', 'DEST_AIRPORT_ID','DEP_DELAY',\
              'DEP_DELAY_NEW', 'WHEELS_ON', 'DEP_TIME', \
              'TAXI_IN', 'CRS_ARR_TIME', 'CRS_ELAPSED_TIME', 'ACTUAL_ELAPSED_TIME', \
              'ARR_TIME_BLK', 'ARR_DELAY_GROUP', 'DIVERTED', \
              'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY', 'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY'], axis=1)
```

Figure 5: Code to remove duplicate column

c. **Fixing null values**: In order to clean the data, it is important to either fill null values with some data or remove them from the dataset. In this case we will drop the missing values. Hence, we are not considering null values for ARR_DEL15 as it is our output. df.isnull() and dropna() were used respectively. To see columns with the number of missing values the df.isnull().sum code was executed. This returned numeric values of 1 when an index is missing and 0 if everything is intact while the dropna() function takes out the data missing in the dataset. For data transformation, variables such as the Month, Day_of_Month and Day_of_Week columns were left in their original order as they are ordinal categorical variables and not nominal.

5.3.1 Preparing Training Data

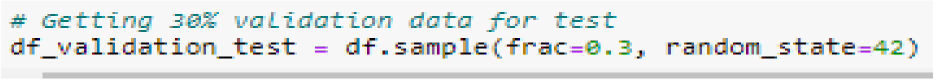
To train the model, a balanced dataset for training was created in order to get better accuracy on test data. Below are codes to prepare the dataset by dividing to training and testing dataset



```
# Shuffling data for better results to avoid biased results
df = df.sample(n = len(df), random_state=42)
df = df.reset_index(drop = True)  # This will shuffle our data randomly
```

Figure 6: Code showing shuffling of data

From figure 6, by shuffling the data bias was avoided in the data by even distribution of positive and negative values. Next was to split the data into train and test dataset respectively.



```
# Getting 30% validation data for test
df_validation_test = df.sample(frac=0.3, random_state=42)
```

Figure 7: Code showing how data was split

From figure 7, the code above removed 30% data from our data frame to create a validation data plus test data leaving the remaining 70% as the train data. Furthermore, the test and validation data were splitted equally that is 15% each.

5.4 Exploratory Analysis

In the exploratory analysis important insights to help understand flight delays better was provided showing the amount of late flights based on different categories like finding out the total number of unique airports in the data frame, flights from origin airport with the most volume, destination airport with most flight volume, how distance affects arrival delay and how selected month of travel affect arrival delay. We also checked statistical insights to understand flight delays better

Table 2: Table showing statistics of flight arrival delay

Departure on time count	274685
Departure on time arrival count	264788
Departure on time arrival late count	9897
Departure on time arrival late count*100/ departure on time count (in percent)	3.60303
Departure on time arrival late count *100/rows (in percent)	3.07171

From table 2, 274685 were flight that departed on time, 264788 is the frequency of flights arrival delay, 9897 is the value of flight that departed late but arrived early, for the last two values in order to find flights with operated without delay the percentage was found between departure on time arrival late count and departure on time count to give 3.6% and flight that departed on time but arrived late gave 3.0% .

According to the unique airports presented in the data, there was a total of 300 airport incorporating both origin and destination. However, these airport were further drilled down to get the top 20 origin and destination airport by their volume with arrival delay focusing on destination airports.

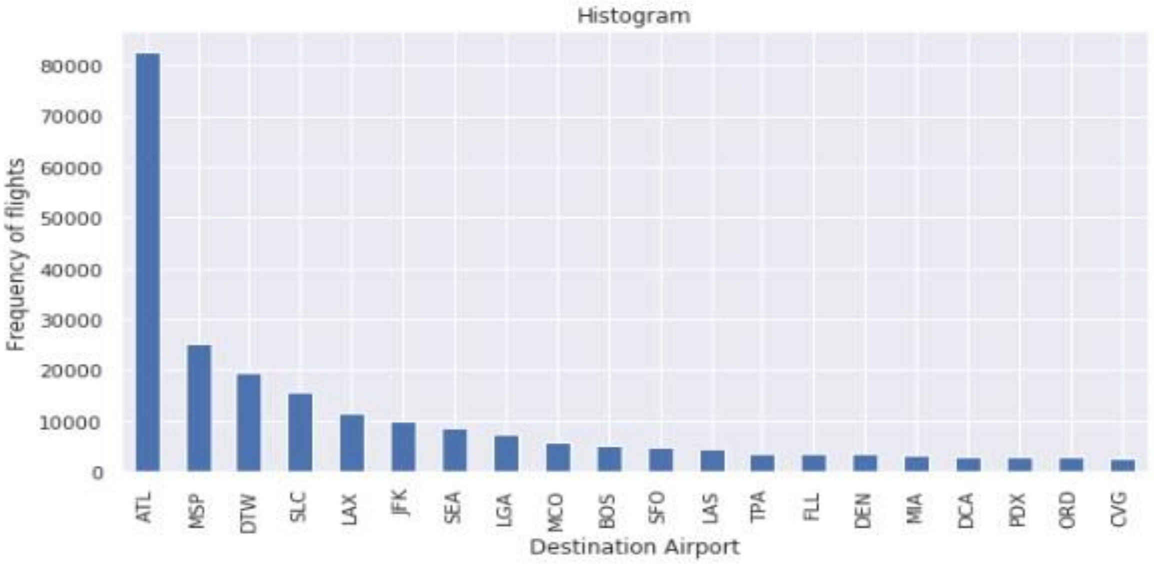


Figure 8: Top 20 Destination Airport Used by Delta airlines by Flight volume

From figure 8 we observed that the Hartsfield-Jackson Atlanta airport (ATL) at destination had the most flight volume with a value of 82519 and Cincinnati/Northern Kentucky international airport (CVG) was the least airport in terms of flight volume with a value of 2605 For the third visualization the role of distance in flight arrival delay was considered by creating a data frame to store two variables (distance and arrival delay)

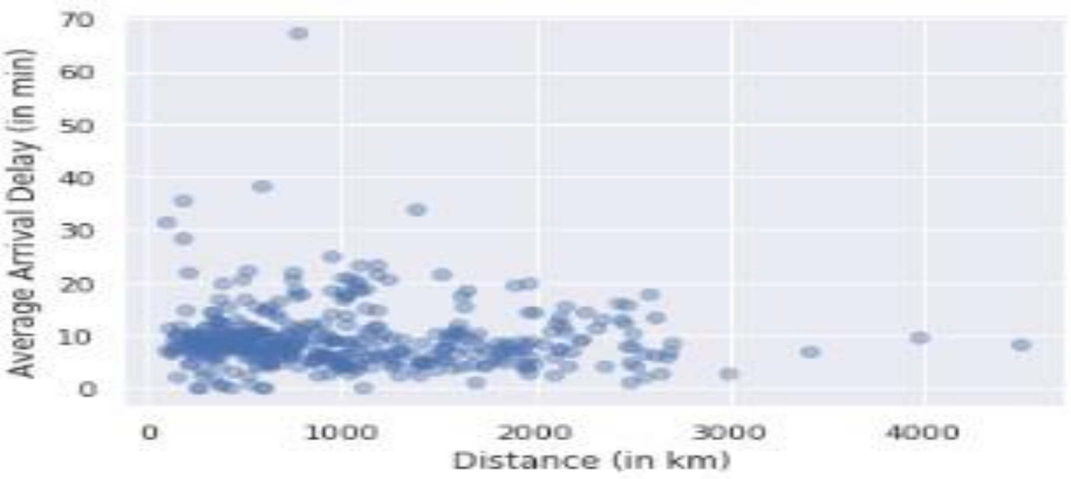


Figure 9: The role of distance in arrival delay

From figure 9 the x-axis had distance in km and y-axis arrival delay in mins. A scatter plot was plotted taking distance as predictor and arrival delay as the response. The scatter plot shows outliers indicating that flights with shorter distance have higher arrival delays. Lastly, we look at how departure time affect flight arrival delay

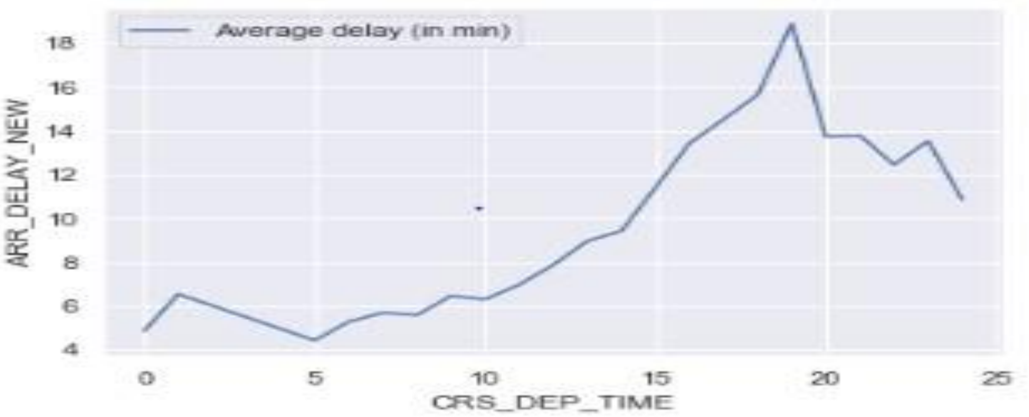


Figure 10: The role of departure time in flight arrival delay

From figure 10 the x-axis displays departure time in mins and y-axis shows average arrival delay in mins as well. it was deduced that late flight departure leads to flights arriving late as well.

Conclusively, from this visual analysis performed using the matplotlib the exploratory analysis reveals how frequency of flights and distance is a factor that affects arrival delay not forgetting flight arrival delay in Delta airlines changes across the selected months.

5.5 Training Machine learning Models

In this phase the machine learning algorithms are trained to using our cleaned dataset. As this is a binary classification problem the two group here will be whether a flight arrived early or is delayed. In order to train our model successfully, it is important to make sure that the training data we are providing is balanced. By a balanced data we mean that no of positive values (ARR_DEL15 = 1) should be approximately same as that of negative value. There were approximately 14% positive values in our training data which means that majority of the data were for the negative values (ARR_DEL15 = 0)i.e. flights arriving on time. In-order avoid the classifiers learning one class better than the other, it was important to balance negative and positive in the training data this was because negatives values were more than the positive value so therefore the model may be biased towards the negatives i.e. (ARR_DEL15=0). Here are the steps followed to make our training data balanced with same positive and negative values

- Taking out all positive values from the training data
- Taking out all the negative values by removing the positive values from the complete training dataset
- Calculating how many positive values present using len function
- Taking out the same number of negative data randomly from the negative dataset
- Using df_train= pd.concat the same number of negative values as positive were taken and the rest were removed in this way the training data was balanced with equal positives and negative value
- The train dataset was shuffled by randomly mixing the values to achieve values that are representatives of the entire data distribution which will produce good performance of the algorithm and avoid bias
- Calculating prevalence to show the total number of cases where arrival delay happened in the test and validation data

After getting the data ready, the best performance with close attention to accuracy, precision, specificity and recall was tested in 6 different algorithms namely Gradient boosting classifier, Support vector machine, Logistic regression, Naïve Bayes, Decision tree and Random forest.

For the Gradient boosting classifiers, it creates more models in a forward stage wise fashion, the sklearn.ensemble method was used they were built with the following hyper parameter 100 estimators and had a maximum depth of 3 while the learning rate was 1.0 the gradient boosting had the best result which outperformed other model. The support vector classifier was implemented based on libsvm using the sklearn.svm on the python programming language where the multi class support handled one vs one scheme. The radial kernel function was used with a cache size of 200 as its mapped input space to infinite dimensional space. Logistic regression was performed using the sklearn.linear_model package in python. Further to this, the weight of each feature was easy to interpret and was indicated as none which means all the classes had one weight. The liblinear library uses class that implements regularized logistic regression. So, since the liblinear solver support both L1 and L2 penalty, the normalization used for the penalty was l2 indicating that regularization was performed.

The Naïve Bayes used the GaussianNB library which was imported from the sklearn.naive Bayes and metric such as accuracy showed a fair prediction. For the decision tree classifiers certain parameters were used . For instance, the Gini function was used to ascertain the quality of split, there were no maximum depth for the tress which implies that the nodes had less than the minimum sample slip which was 2 when they were expanded.

Finally, for the random forest classifier which is a meta-estimator that uses the average number of decision tress classifiers on sub-samples to get better predictive accuracy, had n_estimator for the number of trees in the forest to be 100. To find the best split, the number of features were considered that is max_feature=sqrt(n_features) however the number of split points was considered as the minimum number of samples at the leaf node was 1

5.6 Proposed tools

The tools used for the analysis which involves the programming language and platforms were structured into software and hardware

Software - The analysis was carried out using the Jupyter Notebook under the Anaconda distribution with python 3.6 (64-Bit). Visualization was carried out using the matplotlib in python to give graphically insights on factors causing arrival delay. This was carried out with the aim of drawing the attention of Delta airlines to recognize areas where the problem exists.

Hardware - Processor: Intel (R) Core(TM) i5-8265U CPU @ 1.80GHz, Installed memory: 8.00GB System type: 64-bit operating system, x64-based processor

6 Evaluation

To carry out the evaluation, the performance of the different classification models was assessed using various metrics. This includes Accuracy, Precision, Recall, Specificity also using four methods to achieve it which are True positives (TP), True Negatives (TN), False Negative (FN) and False positive (FP).For the True Positives (TP) arrival delays that are positives and predicted as positives. Next is the True Negative (TN) which are arrival delays that are negative and predicted as negatives. For the False Negative (FN), arrival delays that are positive but predicted as negative and lastly the False Positives they are arrival delay that are negative but predicted positive.

The evaluation measurements are explained as follows

- Accuracy : is the ratio or fraction of correct prediction got by the model.

$$accuracy = \frac{\text{true positive (TP) + true negative (TN)}}{\text{true positive (TP)+true negative (TN)+false positive (FP)+false negative (FN)}}$$

(Hossin et al.,2015)

- Precision : This is the amount of positive results that are truly positive

$$precision\ (p) = \frac{\text{true positive (TP)}}{\text{true positive (TP)+false positive (FP)}}$$ (Hossin et al.,2015)

- Recall: This is number of actual positives that were recognized correctly

$$recall\ (r) = \frac{\text{true positive (TP)}}{\text{true positive (TP)+false negative (FN)}}$$ (Hossin et al.,2015)

- Specificity: The proportion of negative patterns rightly classified

$$Specificity\ (sp) = \frac{\text{true negative (TN)}}{\text{true negative (TN) +false positive (FP)}}$$ (Hossin et al.,2015)

The next step was to use this metrics to measure the performance of the 5 models used as can be seen in the table below

Table 3: Comparative analysis of model performance (CASE STUDY 3)

Metric	Gradient boosting classifier	Support vector Classifier	Naïve Bayes	Logistics Regression	Decision Trees	Random forest
Accuracy	0.7062	0.6215	0.6049	0.6410	0.6891	0.6462
Precision	0.2711	0.2012	0.2051	0.2233	0.2469	0.2279
Recall	0.6574	0.5788	0.6390	0.6373	0.6015	0.6457
Specificity	0.7141	0.6284	0.5994	0.6415	0.7032	0.6462
AUC	0.6922	0.5916	0.6208	0.6407	0.6646	0.6542

From table 3 above there was comparative analysis between the different models which showed the gradient boosting classifier and decision tree to be the best predictive models with gradient boosting having the highest value of 70%. This answers the research objective 3. Furthermore, the data was only restricted for the months of June July August and September hence not too high accuracy indicated. The next section uses graphs to visualize factors that influence arrival delay.

6.1 Experiment / Case Study 1

Which Airport used by Delta airlines experienced the most delays?

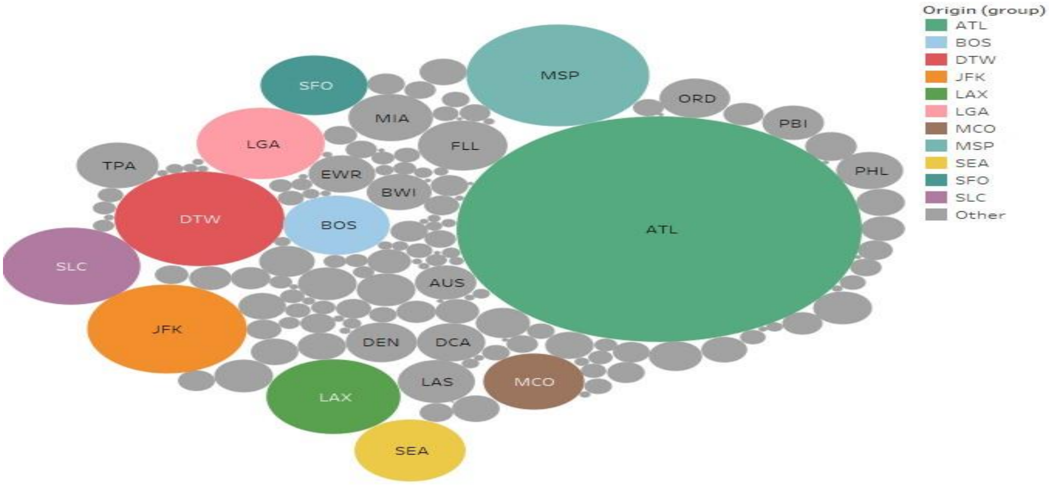


Figure 11: Visualization of Airport used by Delta airlines with most delays

From figure 11 above, the airports with the most delay is the Hartsfield Jackson Atlanta international airport (ATL) due to the high volume of flights arriving and departing from the airport. The graph illustrates that the ATL airport has a flight volume of 82519 arriving flights for the selected months of June, July august and September in 2017. For Delta airlines having Hartsfield Jackson Atlanta international airport (ATL) as a major hub will lead to flight congestions impacting on scheduled time and propagating across the flight network. This affects flight operators most especially as it reduces flight efficiency, loyalty of customers and the economy .This answers the research objective 1

6.2 Experiment / Case Study 2

What summer month in 2017 had the most arrival delay?

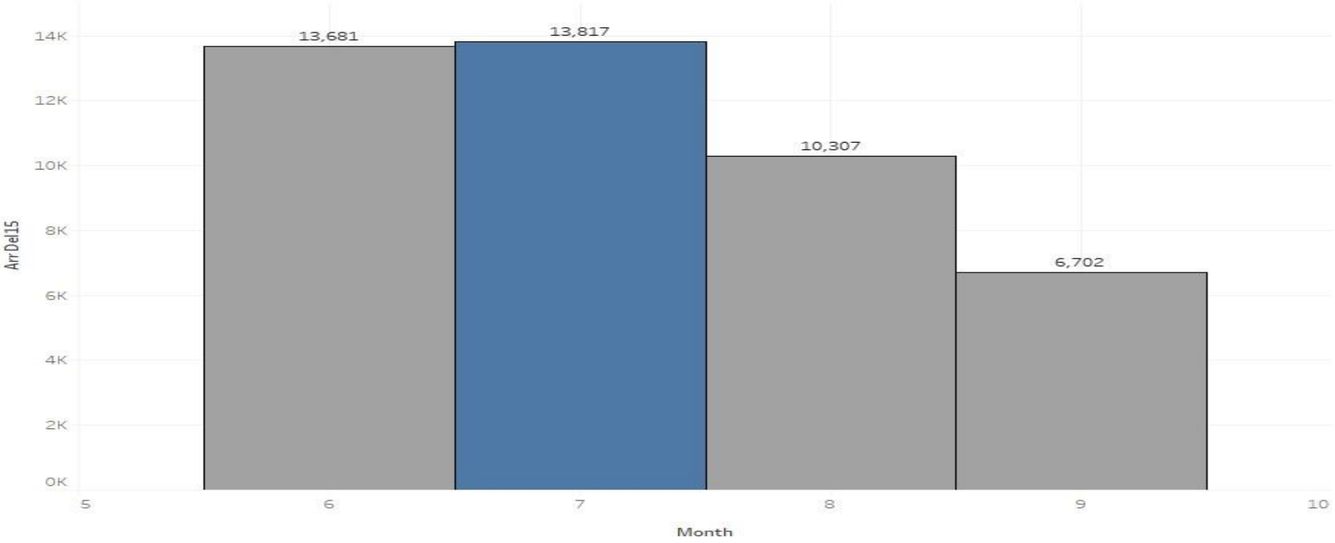


Figure 12: Bar chart visualization of months with the highest delays

In this research the visualization was for the summer months and from figure 12 above the month of June and July had the highest arrival delay in 2017 with a value of 13601 for June and 13817 for July while comparatively we notice a decrease in arrival delay from 10,307 to 6,702 by flights operated by delta airlines. This suggest that there are more travellers during summer and factors such as airport congestion, cancellation and diversions would cause increase in delays This answers the research objective 2

6.3 Discussion

For this project, a new way to analyse flight delay was achieved by looking at arrival delays for Delta airlines without reducing the value of analysis. The research objectives were met accordingly, The first objective talks about major airports of delta airlines that flights delay affects the most this delays can be mitigated by supporting air traffic control through boosting organization between airlines and airports. In the Second objective which looks at months with highest delays, better flight scheduling operations should be set in place in order to improve flight efficiency. Phase three met the third objective of comparative analysis between algorithms, with gradient boosting classifier having the best accuracy and specificity as compared to other algorithms.

7 Conclusion and Future work

This research made an impact by the predictive analysis of flight arrival delay reduction for Delta airlines by carrying out a comparative analysis between 6 different machine learning model used this are Naïve Bayes, Decision trees, Support vector classifier, Gradient Boosting classifier, Logistics regression and Random forest. Worst performing models were Naïve Bayes and Support vector classifier with 60 and 62% respectively. Gradient boosting classifier out performed other models slightly with 70% accuracy.

The research was in two parts and they worked consistently to comparatively analyse the best performing model that predict flight arrival delay in Delta airlines and produce a visualize insights on areas were Delta airlines could shed light on to reduce flight arrival delay and improve efficiency to its flights.

Furthermore, the limitation to this research can be drawn from the nature of the analysis which used few months in the year 2017. This research can further be improved in the future by using real time flight data for Delta airlines, carrying out analysis on more than just one year, including factors such as weather and using unsupervised machine learning approach.

Acknowledgement

Glory to God for the grace of completing this project. A big thank you to my family, and appreciation to my supervisor Dr. Cristina Muntean who showed support patience and understanding during the course of this project.

References

Abdel-Aty, M, Lee, C., Bai, Y., Li, X. and Michalak, M. (2007). Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, 13(6), pp.355-361.

Assem, H. and O'Sullivan, D. (2015). Towards bridging the gap between machine learning researchers and practitioners, *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*, IEEE, pp. 702–708.

Barnhart, C., Fearing, D., Odoni, A. and Vaze, V. (2012). Demand and capacity management in air transportation. *EURO Journal on Transportation and Logistics*, 1(1-2), pp.135-155.

Bi, J., Bennett, K., Embrechts, M., Breneman, C. and Song, M., 2003. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, *3*(Mar), pp.1229-1243.

Bryan, D.L. and O'kelly, M.E., 1999. Hub-and-spoke networks in air transportation: an analytical review. *Journal of regional science*, *39*(2), pp.275-295.

Chakrabarty, N., Kundu, T., Dandapat, S., Sarkar, A. and Kole, D.K., 2019. Flight Arrival Delay Prediction Using Gradient Boosting Classifier. In *Emerging Technologies in Data Mining and Information Security* (pp. 651-659). Springer, Singapore.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide. 20

Chen, H., Wang, J., & Yan, X. (2008). A Fuzzy Support Vector Machine with Weighted Margin for Flight Delay Early Warning. *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 3*, 331-335

Etani, N., 2019. Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. *Journal of Big Data*, *6*(1), p.85.

Gwiggner, C., Fujita, M., Fukuda, Y., Nagaoka, S. and Nikoleris, T., 2011. Trade-offs and Issues in Traffic Synchronization. In *Proceedings of 9th USA/Europe ATM 2005 R&D Seminar* (pp. 1-6). USA/Europe: FAA/Eurocontrol.

Hiagarajan, B., Srinivasan, L., Sharma, A.V., Sreekanthan, D. and Vijayaraghavan, V., 2017, September. A machine learning approach for prediction of on-time performance of flights. In *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)* (pp. 1-6). IEEE.

Hsiao, C.Y. and Hansen, M., 2006. Econometric analysis of US airline flight delays with time-of-day effects. *transportation research Record*, *1951*(1), pp.104-112.

Khanmohammadi, S., Chou, C.A., Lewis, H.W. and Elias, D., 2014, July. A systems approach for scheduling aircraft landings in JFK airport. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1578-1585). IEEE.

Ison, D.C., Weiland, L., McAndrew, I. and Moran, K., 2015. Identification of air traffic management principles influential in the development of an airport arrival delay prediction model. *Journal of Aviation/Aerospace Education & Research*, *24*(2), pp.39-53.

Klein, A., Kavoussi, S. and Lee, R.S., 2009, June. Weather forecast accuracy: Study of impact on airport capacity and estimation of avoidable costs. In *Eighth USA/Europe Air Traffic Management Research and Development Seminar*.

Le, L, Donohue, G. and Chen, C.H., 2004. Auction-based slot allocation for traffic demand management at Hartsfield Atlanta international airport: A case study. *Transportation research record*, *1888*(1), pp.50-58.

Morrison, S. and Winston, C., 1997. Airline deregulation and fares at dominated hubs and slot-controlled airports. *Statement by Steven A. Morrison before the Committee on the Judiciary, United States House of Representatives*.

Mueller, E. and Chatterji, G., 2002. Analysis of aircraft arrival and departure delay characteristics. In *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum* (p. 5866).

Mohammed, A., 2007. Qualitative & Quantitative analysis of tiered Architecture of WebApplications.

Rebollo, J.J. and Balakrishnan, H., 2014. Characterization and prediction of air traffic delays. *Transportation research part C: Emerging technologies*, *44*, pp.231-241.

Schaefer, L. and Millner, D., 2001, October. Flight delay propagation analysis with the detailed policy assessment tool. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)* (Vol. 2, pp. 1299-1303). IEEE.

Schumacher, B., 1999, December. Proactive flight schedule evaluation at Delta Air Lines. In *WSC'99. 1999 Winter Simulation Conference Proceedings.'Simulation-A Bridge to the Future'(Cat. No. 99CH37038)* (Vol. 2, pp. 1232-1237). IEEE.

Tu, Y., Ball, M.O. and Jank, W.S., 2008. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, *103*(481), pp.112-125.

Zonglei, L., Jiandong, W. and Guansheng, Z., 2008, December. A new method to alarm large scale of flights delay based on machine learning. In *2008 International Symposium on Knowledge Acquisition and Modeling* (pp. 589-592). IEEE.