National College of Ireland

# Performance based predictive analysis of divergent classifiers for United States flight delays

MSc Research Project
Data Analytics

Danish Bhatia
x16133137

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

**National College of Ireland**
**Project Submission Sheet – 2017/2018**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Danish Bhatia |
| **Student ID:** | x16133137 |
| **Programme:** | Data Analytics |
| **Year:** | 2016 |
| **Module:** | MSc Research Project |
| **Lecturer:** | Dr. Cristina Muntean |
| **Submission Due Date:** | 11/12/2017 |
| **Project Title:** | Performance based predictive analysis of divergent classifiers for United States flight delays |
| **Word Count:** | 6603 |

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 encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 11th December 2017 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. 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): | |

# Performance based predictive analysis of divergent classifiers for United States flight delays

Danish Bhatia
x16133137
MSc Research Project in Data Analytics

11th December 2017

**Abstract**

Flights are an imperative means of commutation in the transportation system for United States. Extensive reliability on them for Business or leisure, is making the flights inevitably delayed as a result of which passengers end up waiting for ages at the airport, because of which airports schedule and management gets disturbed, ultimately depleting the reputation of airlines. Accurately predicting the delay beforehand, could significantly reduce the impact of a late flight, if not completely alleviate it. Myriad researches have been done within this domain considering Machine Learning techniques and evaluating performance of a model based on Accuracy alone. However, considering only these algorithms and accuracy as a performance metric alone, is kind of biased. This research therefore, intended to propose a novel solution of analyzing the performance of four models including a Deep Learning model based upon Accuracy, Precision, Recall, F Measure and Kappa Statistics. The prime objective of this research is to get as accurate model as possible which could probably be used to assist customers to save their time. This research harnessed key insights informing that late arriving flights is the root cause of delay. Unstable weather also results in a lot of delay and eventually cancellations. This piece of information could be lucrative for the airlines as they can come up with ideas to mitigate these delays.

# 1 Introduction

In the contemporary world, where time plays an utmost imperative role in everyones life, using air transportation for travelling to very far off places have become a ritual. This ritual is further supported by the backsliding cost of fuel and impressive offers given by the airlines as a bait. This in turn has made people rely heavily on the air transportation. Also, air transportation today, is being considered as one of the vital means of commutation for personal and leisure activities (Khanmohammadi et al.; 2014). This heavy reliance on flight systems have escalated the number of flights to take off and has eventually caused air traffic congestion resulting in a delayed flight. Delay is one of the prominent performance metric for any transportation system. Factors causing delays like maintenance of aircraft, late arriving aircraft, security and weather other than air traffic congestion, greatly influence the economy of a country. These natural occurrences

and operational challenges results in exponential costs faced by the airlines companies (Thiagarajan et al.; 2017). This expensive deal to airline companies further creates problems for the airports in rescheduling the flights ultimately troubling the passengers for waiting till ages creating a level of dissatisfaction among the passengers resulting in the bad status for the airlines. To reduce the losses occurred as a result of flight delay, significant amount of research has been done for many years, in this domain (Zonglei et al.; 2009). However, the quality of analysis still remains limited (Kim et al.; 2016). Also, most of academic papers base their research solely on Accuracy as the metric of performance, comparison and evaluation between different models and algorithms. The motivation of this Research is to provide an accurate solution which could possibly support Federal Aviation Administration (FAA) to determine the key areas causing flight delays.

The layout of this Research is segmented as Section 2 would deeply describe the Research Question along with past works done within the sphere of this domain. Next, Section 3 would provide an overview of the Methodology used during this Research. Then, Section 4 would dive deep into how the Research was carried. Eventually Section 5 and Section 6 would smoothly end the Research defining the best results and future scope respectively.

## 1.1 HYPOTHESIS

Let ***"H"*** denote the Hypothesis, ***"I"***, ***"O"*** denote the Input and Output variables respectively where I = Airline, Scheduled Departure, Departure Time, Scheduled Arrival, Arrival Time O = Delay

$$H_0 = I \Rightarrow O$$

# 2 RELATED WORK

The concept of delay in air transportation is not new, hence, a lot of academic work has been published before. For the year 2007, it was estimated that the total loss for all flight delays was around $32.9 Billion (Ball et al.; 2010). Next, there was a huge drop of 65%, in the year 2009, in the market capitalization of major US flights, amounting to $26 Billion (Ball et al.; 2010). Also, as per (Manley and Sherry; 2010), due to the massive use of air transportation systems there is a disparity between the demand of the flights and the available capacity of the aircrafts and the cost arising as a result of this disparity is estimated to be roughly around $12 Billion. Currently, the Airlines of America showed that cost of delay due to crew members rose by 8.7% in 2016 than in 2015[1].

This section comprises the research findings from multiple pieces of work and would provide the brief comparison between them along with providing necessary opinion on the same. All the past work has been divided into 2 categories:

1. **RESEARCHES BASED ON TYPES OF FLIGHT DELAYS** This is further subdivided into 2 categories based on the model used
    - Delay Propagation
    - Delay Innovation

[1]http://airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/

2. **BASED ON THE PROCEDURES UNDERTAKEN** This section is bifurcated into 2 parts namely

    - Literature Reviews based on Machine Learning.
    - Literature Reviews based on Statistical Models.

## 2.1 RESEARCHES BASED ON TYPES OF FLIGHT DELAYS

- Delay Propagation

It is generally believed by most that one problem is the cause of second and second is the cause of third, thats indeed how the problem propagates. As per Roger Beatty and team, due to the imbalance between the demand and supply of aircrafts and crew members, once an initial delay takes place, it propagates perpetually (Beatty et al.; 1999). This is turn creates a chain reaction eventually resulting in long waiting periods. However, no specific approximate value was provided within this analysis. This problem was in-fact, solved by two researchers of MIT, SB Boswell and JE Evans who developed statistical models representing the downstream (propagated) delays and they indeed concluded that almost 80% of the propagated delays are because of the fact that an initial flight was delayed (Boswell and Evans; 1997). This piece of research was deep, crisp and concise as it not only provided with the round off value of flight delay being propagated but also laid the root cause of such delays. Now that the delay in air transportation is inevitable, it might lead to generation of some new delays eventually. This is called as Delay Innovation and thats exactly the focus area of next section.

- Delay Innovation

Certainly, todays delays in flights will give birth to new factors causing flights to be delayed which would for sure, further weaken the performance of air-transportation system. This is named as Delay Innovation. 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. A couple of academic pieces narrowed their focus on predicting and estimating the duration and the grade of flight delays. L Zonglei and W Jiandong developed an alarming system to flag the flight which will be having long delays (Zonglei et al.; 2008). The high-quality data used by them was collected from the airport and they successfully concluded the confidence of greater than 80%. On the other hand, Juan Jose Rebollo and Hamsa Balakrishnan, although, restricted their research for predicting the departure delays from 2 to 24 hours and not longer, the best part is that they extended their model by including delay due to network as a variable in their model (Rebollo and Balakrishnan; 2014). Moreover, some researches also highlighted multiple conditions responsible for causing delay like taxi-out time in US (Balakrishna et al.; 2010), customer demand (Hunter and Ramamoorthy; 2007). Overall, both the researches emphasized their work in a coherent fashion where the first one provided with the absolute estimate for delay of about 80% and the second piece work provided with greater depth.

## 2.2 RESEARCHES BASED ON THE PROCEDURES UNDERTAKEN

Prediction of flight delays has gained a lot of success and thus a lot of algorithms and methods have been implemented on the same, however, every algorithm differs based upon the objective which the academic work is trying to achieve. This section briefs about the different algorithms undertaken by many researchers over several years grouped by similar methods (procedures/technique).

- **Literature Reviews based on Machine Learning**

Machine Learning is the field of Data Science which predicts the output based upon examples and experience. In other words, when a machine is given a power to learn from the historical data in order to predict the future, i.e. called as Machine Learning. Because of its popularity, many researchers have their own version of its definition. As per (Assem and O'Sullivan; 2015), Machine Learning is learning from the data to discover patterns and trends along with getting a feel of how the real-world works. A machine learning practitioner, Michael, along with his team, provided an algorithm to calculate the delay in flights because of passengers delay (Ball et al.; 2010). They further followed a novel approach of sub-dividing their algorithm to consider itinerary of passengers, demand of passengers as per their itinerary and analysis of flight being delayed or cancelled. Moving ahead, researchers like Juan Jose Rebollo and Hamsa Balakrishnan, used a powerful Machine Learning Algorithm called Random Forest Algorithm to predict future departure delay on a particular link or at a particular airport (Rebollo and Balakrishnan; 2014). They extended their research by taking two mechanisms namely classification, for predicting if the departure delay is above or below the threshold (Binary) and regression, to predict if the output is an estimate of departure delay along the considered link or considered airport in United States. Not only this, Chen in his academic work presented an early warning flight delay model which was based on Fuzzy Support Vector Machine with weighted margin (WMSVM) by adjusting the margin between samples and hyperplane (Chen et al.; 2008). Their prime objective was determine the delay grades of the flights by comparative analysis between One-Against-One SVM and One-Against-One WMSVM, out of which One-Against-One WMSVM performed better. Balakrishna, on the other side used a non-parametric approach of Reinforcement Learning which was based upon a probabilistic framework to predict the accuracy of taxi-out timings. The objective of this research was different from the previous one as it primarily focused on backsliding the Downstream Congestion so that the flights could take-off into the air as early as possible (Balakrishna et al.; 2010). Overall, the positive point about Chens research is that it conspicuously provides the avenue in which the future work should progress, however, it didnt mention anywhere about the beneficiaries as a result of his research. This shortcoming was taken into consideration by Balakrishna as it was utmost clear that his research would assist the department responsible for planning departures of flights. This section gave an overview about some work done on US flight delay predictive analysis using various Machine Learning Techniques. The upcoming section will be a coherent demonstration about only those pieces of work which used Statistical Predictive Analytics Models to get fruitful results.

- **Literature Reviews based on Statistical Models**

As the name suggest, these models comprises of all the Statistical techniques like Regression, Multivariate Analysis, correlations etc. When it comes to Regression, it is the

favorite method to be implemented by many researchers. Researchers like Paul Wang and Roger Beatty, used Regression Techniques to figure out the delay factor in propagation through the flight system which in turn would give them the cost estimates associated with the delay (Wang et al.; 2003)(Beatty et al.; 1999). Some researchers even used the non-parametric approaches within the sphere of Statistics to calculate the efficiency of the US airports taking Delay as the baseline. One such example of this comprehensive research is the research by Pathomsiri (Pathomsiri et al.; 2008). It was concluded in her research that addition or deletion of flight delay in a model could greatly influence the output. She founded that on inclusion of flight delay metrics in the model, small and congested airports also become efficient, however, exclusion of the metric could do vice-versa. Eventually, Abdel-Aty, proposed a model to calculate the average of delay on a daily basis in order to detect the correlation between various variables and to understand the key reasons behind the delay (Abdel-Aty et al.; 2007). He kept his research only till Orlando International Airport. Overall, the researches certainly, figured out promising results but at the same time lacked future scope. The next section will describe the Methodology used.

## 2.3 RESEARCH QUESTION

Current Research Project addressed the Business Question as follows:

> ***"To what extent, the use of Deep Learning Algorithm would enhance the predictive analytics quality in terms of accuracy and precision of US flight delays as compared to already implemented Algorithms?"***

To answer this novel question, Artificial Neural Network was used and its calculated accuracy along with other metrics like Precision, Recall, Sensitivity, Specificity and Kappa Statistics were compared with 3 most common Machine Learning models namely: Nave Bayes, K-Nearest Neighbors, XGBoost Algorithm. Also, narrow focused graphs were generated through Tableau emphasizing key details and coherent factors which simultaneously answered few imperative queries like:

- What was the prominent factor behind delaying the flight, Diversion, Cancellation, Weather?
- Which airport got severely affected because of delay in flights?
- Which month and even day of the month did maximum delays took place?

The answers to these relevant Business queries should be rewarding for the Air transportation department if they consider the factors resulting in the delay. This would enable a chance for the airlines to notify the customers, beforehand, about the unfortunate circumstance and also providing the customers with complimentary arrangements like the accommodation, if the flight is international, or lodgment, if the flight is domestic. This service would enable the airlines companies to escalate the level of comfort zone in customers mind and would eventually improve the level of dignity. Not only it would be beneficial for the airlines, the solution to this predictive research analysis would give airports a chance to re-schedule the timings for the flights, clearing the runway and making avenues for the upcoming planes. Lastly, the solution would also be advantageous for the customers as they then could make necessary arrangements for themselves

like arranging a transit visa. This section threw a light on the Research Question which has been investigated here, also highlighting the beneficiaries from the solution and the possible advantages as a result of this Predictive Research. The following section would actively compare and contrast the previous and the related academic work in this domain while taking into account, various approaches and their research findings.

# 3 Methodology

From the Methodological perspective, Cross-industry Standard Process for Data Mining (CRISP-DM) was taken into consideration. Not only is CRISP-DM used by numerous industry data miners to handle multiple challenges, but also is a leading methodology and hence the entire Research Project is divided into 6 categories.

1. **Business Understanding** The Research objective, at this initial stage, was seen from a Business perspective. In this research, the Business was airlines companies operating in United States. Following this, the understanding developed, and the knowledge gained was, in turn, modelled into the problem definition which eventually resulted in the formulation of Data Understanding plan.

2. **Data Understanding** This phase of CRISP-DM was started with the collection of data. The dataset was collected from the authentic United States Department of Transportation website called Bureau of Transportation Statistics [2]. Since the website provided with the facility of attribute selection along with the time selection, datasets of 2014, 2015 and 2016 were chosen with the relevant attributes. Attributes were chosen while keeping the Business objective and problem definition in mind. The next step in this phase was to get acquainted with the dataset to discover the very basic insights and to get its feel. The next phase following the phase of Data Understanding was to preprocess the data.

3. **Data Preparation** From this phase onwards, the research was carried forward by Python Programming Language under Anacondas Jupyter Environment. The Anacondas Environment was chosen as it provides a couple of advantages over standard python environment like:

   (a) Anaconda is an incredible combination of tools and packages which provides the support of Python in it.

   (b) Anaconda Environment makes life easy by providing convenient package management, easy installation of other packages and add-ons.

   (c) Furthermore, Jupyter notebook was preferred over Spyder because of its dynamic nature as it provides the functionality of writing live code and viewing the output at the same time. There is no need to execute entire code in one go as Jupyter notebook facilitates line-by-line execution.

   (d) Lastly, Jupyter notebook provides visualizations within the notebook itself including Markdown as a further add-on.

   During this phase, general operations performed were data cleaning like removing Null Values (NaN), counting and removing duplicated rows, feature engineering i.e.

[2] https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

selecting the relevant columns according to the mindset of a passenger, selecting the mandatory subset from entire data, for example, all the rows and columns which contained year as 2015 and month as 12 was taken further because the dataset, as a whole, was consuming a lot of execution time and was out of the memory bounds of CPU limit, data types transformation like converting integer values to float values and vice-versa. These operations were performed in no particular order.

4. **Data Modelling:** After the data was defined and then prepared upto the satisfactory level, it was bifurcated into X (input matrix) and Y (response variable). Afterwards, the input matrix (X) was divided into training and testing data within the standard ratio of 75% training and 25% testing datasets using train_test_split method from cross_validation library of sklearn package. Then, a couple of models like Artificial Neural Network (an approach of Deep Learning), K Nearest Neighbors, XGBoostng and Nave Bayes (Machine Learning Approaches) were applied on the final version of produced data. This resulted in comparative analysis of performance of a model using various metrics like:

   (a) Accuracy Score

   (b) Precision Score

   (c) Specificity Score

   (d) Sensitivity Score

   (e) Recall Score

   (f) Kappa Statistics

   This stage was generally followed by Data Preprocessing phase a couple of times until a satisfactory result was obtained.

5. **Evaluation** The evaluation phase, for this research was divided into 2 segments namely:

   (a) Evaluation of the output produced by analysis.
   Comparative Analysis of multiple metrics were done on different models considered by this research work. According to the performance score KNN achieved the best prediction results even outperforming ANN, on the other hand Nave Bayes performed the worst. The magnitude of performance for Nave Bayes was so faulty that even its null accuracy was greater than the predicted accuracy.

   (b) Evaluation of the output produced by visualization.
   A supplementary piece of research was done through a data visualizing software called Tableau and comprehensive charts were prepared. As per the visualizations, Southwest Airlines was most delayed and Hawaiian Airlines was the least. Not only this, the months of February and June, experienced significantly much higher flight delays than any other month in the same year.

6. **Deployment** After careful analysis, the piece of work is ready to be deployed. Also, it provides an advantage of being general purpose in nature. Depending upon the requirements, any airline company can use this piece of work to analyze the key areas which are responsible for delay creation and delay propagation and this

work would even throw a light on the airports and the airlines which experience maximum delays.

## 3.1 Design Techniques used in Research

This research project provided the comparative analysis between Deep Learning and Machine Learning models. For the sake of clarity, this research used one Deep Learning classifier (model/technique) and three Machine Learning classifiers.

### 3.1.1 Deep Learning Model:

- Artificial Neural Network (ANN):
  Since Machine Learning Algorithms provides a non-linear distribution with the increase in demands of predictive analytics, Artificial Neural Networks provides a solution. ANN have the capacity to work well with those models as well, where other models even fail to perform (Das et al.; 2017).

  This prime objective of this research was to use Artificial Neural Network to figure out the performance of the predictive model and to compare it with the performance of other models generated as a part of this research.

  The motivation to use ANN came from the fact that it contains a combination of neurons, wherein, each neuron is responsible for performing some operation and some function. This in turn gives the power to ANN to perform outstanding parallel computation (Khanmohammadi et al.; 2016). The network with only one hidden layer is called as Shallow Learning and the prime requirement of a network to be a deep learning network is that it should have a minimum of two hidden layers (Schmidhuber; 2015). Thus, this piece of work took two hidden layers into consideration. Each neuron within the architecture generates a real valued sequence of activation functions. In the first layer, the neurons get activated when they receive an input signal either from sensors or from any external environment which follows the weighted activation of other neurons and finally the sequence generated is statistically computed according to the activation function and later passed on to the output layer (Schmidhuber; 2015). Basically, ANN are used to provide a non-linear mapping between the input and the output (Tewary; 2015). In the standard back-propagated Artificial Neural Network, the output of jth neuron in qth layer is given by: Figure 1

$$net_j = \sum_{i=1}^{n} (w_{ij} a_i^{(q-1)} - \theta_j),\ a_j^q = \frac{1}{1 + e^{-\tau \times net_j}}$$

Figure 1: Net Value of Neuron

Where,

netj = net value of neuron i(q-1) = output of ith neuron in layer (q-1) wij = the weight of ith neuron from the source layer to jth neuron in the target layer. j =

Threshold value jq = output value of jth neuron in qth layer = shape factor of sigmoid activation function

## ARCHITECTURE OF ANN

The multilayer, fully connected artificial neural networks comprises of input, hidden and output layers where in every single node in one layer is connected to every other node in the following layer. This high-level network of nodes (called as neurons) is shown in the Figure 2

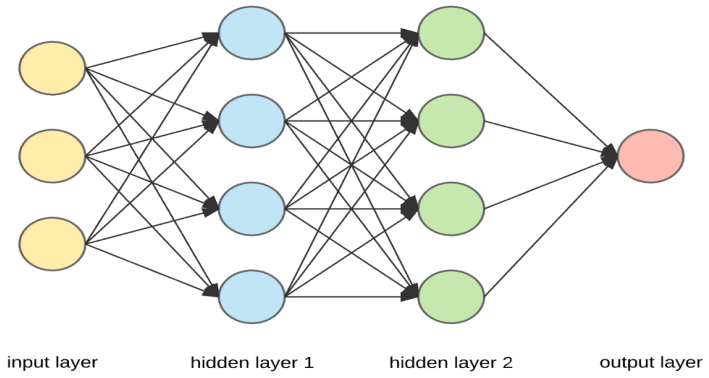


Figure 2: Architecture of ANN

From statistical point of view, an in-depth look for ANN is shown in the Figure 3

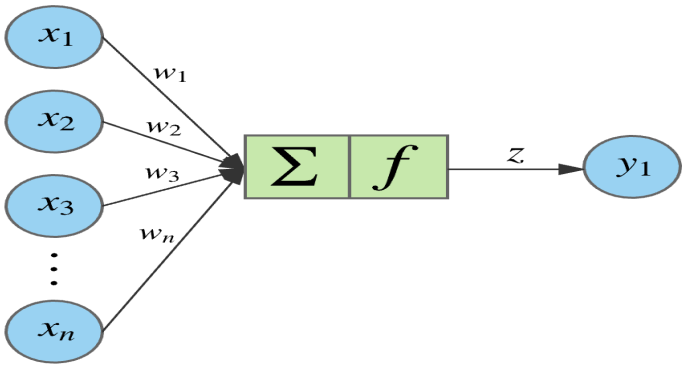


Figure 3: Statistical Representation of ANN

This representation of ANN shows that the nodes in the hidden layer take the weighted sum of all inputs which are eventually passed through an activation function. The training of the network is done by first randomly initializing weights for all the nodes, then feeding the output forward to next layers, comparing the calculated output (called predicted output) with the actual output and eventually updating the weights according to the error. For the scope of this research, it used Rectified Linear Unit (ReLU) as is hidden layered activation function and Sigmoid activation function for output layer.

### 3.1.2 Machine Learning Models:

- K Nearest Neighbors:
  KNN is a simple machine Learning classifier that is intended to make predictions according to the closest data points it finds for every single test data. Being a Lazy model, it does not make any generalizations based upon the training dataset (Devi et al.; 2007).

  In general, the KNN works by selecting the Nearest (as the name suggest) data points from the training dataset and assigns its class as the predicted class to testing data point. The variation in the number of nearest neighbors (K) varies the accuracy score and the decision boundary between the training and testing data set. Variation in decision boundary as a result of variation in K is shown in the Figure 4 For the scope of this research, a graph was made showing the variation in

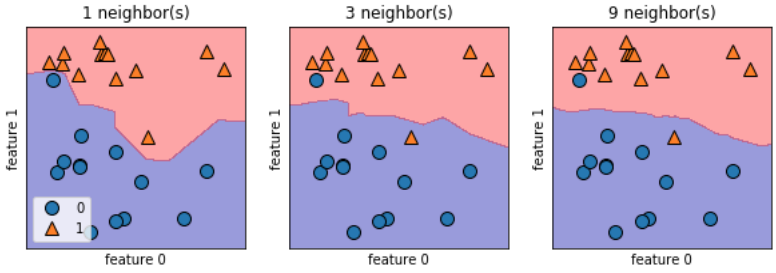


Figure 4: Change of Decision Boundary vs K Value

the accuracy score of prediction for delay in US flights according the various values of K. this graph is shows in the Figure 5

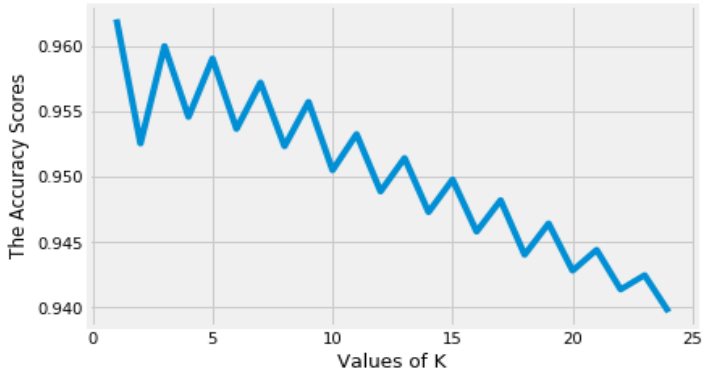


Figure 5: Accuracy vs K Value

- Nave Bayes:
  The Nave Bayes classifier of Machine Learning is a statistical approach based upon the Bayes Theorem (as the name suggests) and it deals with the probability distribution of the variables in the dataset. As an outp
  ut, Nave Bayes gives the probability of a variable belonging to a particular class. One extreme advantage of Nave Bayes is that it allows previous logic and facts to be

applied for those statements which are not certain (Stolfo et al.; 1997). Nave Bayes classifier has an assumption of conditional independence between the features in the dataset which was indeed taken care of for this research. For the scope of this research, since the outcome variable was Binary, hence the conditional probabilities for Binary Classes as per Nave Bayes classifier is (Stolfo et al.; 1997): Figure 6

$$P(c_i \mid f_k) = \frac{P(f_k \mid c_i) * P(c_i)}{P(f_k)}$$

$$P(f_k \mid \mathbf{c_i}) = \prod_{i=1}^{n} P(f_k | c_i) \; k = 1,\ldots,n; i = 1,2$$

Figure 6: Bayes Theorem (Stolfo et al.; 1997)

- XGBoost:
  A relatively new Machine Learning Algorithm proposed by Chen in the year 2014 is named as e**X**treme **G**radient **B**oosting, is particularly used for Binary Classification in supervised machine learning systems (Gumani et al.; 2017) and because it is based on gradient boosting framework, XGBoost formalizes to control overfitting ultimately escalating the models performance. From statistical point of view, It generally works by determining the step by directly solving the equation shown in Figure 7

$$\frac{\partial L(y, f^{(m-1)}(x) + f_m(x))}{\partial f_m(x)} = 0$$

Figure 7: Equation to calculate Step Size (Gumani et al.; 2017)

Where, x = data point in the data set.
This piece of research considered XGBoost for the following advantages:

1. It is faster than the standard Gradient Boosting Model (GBM).
2. It provides more options for regularization.
3. It provides some randomness alleviating the chance of overfitting and making the dataset to be remembered by the classifier.

The results shown for XGBoost in this research were indeed promising

# 4 Implementation

For the research to accomplish its objective, the dataset for 3 years (2014, 2015, 2016) was collected from the official United States Department of Transportation website. The

reason why current years data (2017) was not considered is because at the point of this research, the year is not yet finished thus, it would not be having the data of the months August, September, October, November and December. Since, the website provided with the functionality of manually selecting the features prior to download, hence 1st phase of feature engineering was done before actually downloading the dataset. The shape of dataset was (11664018, 31). This means that the dataset consisted of 11664018 rows and 31 columns.

After the dataset was downloaded, the phase of data cleaning, data transforming was started. This was the most time-consuming task in preparing the dataset for analysis purpose. Some examples of data cleaning and transformation include: thinking from the passengers perspective who wants to book a flight and is unaware of the fact whether his flight will be delayed or not, this research kept only those columns into picture which a passenger would think of at the time he is booking a flight. Moreover, the dataset was not having the response variable and hence the variable must be created manually. The Response variable or the outcome called as **Delay** was introduced with the condition in mind that if the difference between flights arrival delay and flight's departure delay is greater than zero minutes, the flight will be counted as DELAYED = 1 otherwise the flight is NOT DELAYED = 0. This was done with pandas [3] package in Python.

A correlation matrix was made to check the dependencies of one variable on another and all those variables which were dependent upon the relevant variables (according to problem definition) were removed so that the result is unbiased.

After that, the dataset was divided into training and testing dataset manually in the ratio of 75:25. This was done with the train_test_split library of sklearn.cross_validation package[4] and the random_state was set to 0 so that every time same result computed on the same system.

Eventually, before subjecting the data into machine learning and deep learning classifiers, the standard scaling was done to remove any Bias. This would ensure that the entire data is on same scale. The scaling was done using StandardScaler[5] library of sklearn.preprocessing package.

Since the column named AIRLINE was categorical, label encoding was performed using LabelEncoder library of sklearn.preprocessing package to convert the strings into integers which in turn was followed by one-hot-encoding using OneHotEncoder[6] library of sklearn.preprocessing package. Eventually, first column was dropped to avoid the Dummy Variable trap.

After preparing the data to the best possible state, 4 algorithms namely Artificial Neural Network (ANN), K Nearest Neighbors (KNN), Nave Bayes and XGBoost were tested to evaluate the best performing one with special attention on the accuracy, precision, specificity, sensitivity, F-Measure and Kappa value of Artificial Neural Network.

For Artificial Neural Network, Keras library build on top of tensorflow was used as the state of the art (Vidnerová and Neruda; 2017) because as keras in build on top of tensorflow and theano, it provides better results (accuracy) and in less time. Two modules namely Sequential and Dense were used to define the sequential and dense architecture of the classifier. Next, since, deep learning methods depends upon the relationship between the input variable to extract the features, activation function was used for the same

[3] https://pandas.pydata.org/

[4] http://scikit-learn.org/stable/modules/cross_validation.html

[5] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[6] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

(Godfrey and Gashler; 2017). For hidden layers, it was ReLU and for output layer it was Sigmoid. Finally, after hit and trial, batch_size of 500 and epochs of 7 was used as the reliable option.

For Nave Bayes, the same splitting, as of ANN, was done i.e. in the ratio of 75:25. GaussianNB[7] library was used from the sklearn.naive_bayes package and metrics like accuracy and null accuracy were computed which showed that the model was not optimal.

Similar method was performed at the time of evaluating K Nearest Neighbors (KNN). This was made possible with KNeighborsClassifier[8] library from sklearn.neighbors package. For the value of K, a graph was generated (accuracy vs value of K) and the optimal value of K was considered.

Last but not least, the XGBoost algorithm was considered with the prime intention to either support Nave Bayes (as it did not perform well) or KNN (as it performed the best). For this XGBClassifier[9] library was used from xgboost package in python. Again, all the performance metrics were calculated and compared against each other.

Finally, data visualization tool called Tableau Desktop 10.4 was used to generate **beyond the data** visualizations deeply focusing the attention towards the most prevalent factors causing delays in US flights. This was done with the intention to make the airlines companies realize where exactly the problem persists.

All in all, all the tests were performed under Anaconda distribution called Jupyter Notebook with python 3.6 (64-Bit) and Tableau Desktop version. The performance measures calculated for each model, were compared with amongst themselves and the tabular representation along with some visualizations would be shown in Evaluation Section coming next.

# 5 Evaluation

As there are many metrics used to evaluate the performance of classification models (A collaborative filtering algorithm and evaluation metric that accurately model the user experience), accuracy, precision, recall, F Measure and specificity are predominant. **Accuracy:** is defined as the total number of correct predictions over all the cases to be predicted. However, accuracy is not always a reliable performance measure specially in real life datasets. **Precision:** shows the fraction of cases correctly identified over all the cases identified in that category. An example could be True Positives divided by all positives. **Recall:** in laymans terms is defined as the fraction of relevant cases divided by all the relevant cases. **F Score:** is simply the weighted average of Precision and Recall (Sokolova et al.; 2006). It gives a value between 0 and 1, where 1 is the best value describing 100%. **Specificity:** is just the classifiers ability to figure out negative outcomes. The performance in terms of several metrics is tabularized in Table 1.

After critically evaluating the performance metrics of all four models, KNN and ANN were the best predictive models as their Accuracy, F-Measure and Recall are better which is further confirmed by Kappa Statistics. Also, this analysis answered the Business question where the Deep Learning Approach (Artificial Neural Network) performed significantly better than already implemented Machine Learning Approaches like Nave Bayes and XGBoost in terms of accuracy.

[7] http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
[8] http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
[9] http://xgboost.readthedocs.io/en/latest/python/python_api.html

| Metric | ANN | Naive Bayes | KNN | XGBoost |
|---|---|---|---|---|
| Accuracy | 0.9072 | 0.6057 | 0.9413 | 0.7885 |
| Null Accuracy | 0.6189 | 0.6189 | 0.5869 | 0.2114 |
| Precision | 0.9395 | 0.4795 | 0.9652 | 0.9848 |
| Recall | 0.8084 | 0.4068 | 0.8775 | 0.4521 |
| F Measure | 0.8690 | 0.4402 | 0.9193 | 0.6197 |
| Specificity | 0.9679 | 0.7281 | 0.9805 | 0.9957 |
| Kappa Statistics | 0.7978 | 0.1390 | 0.8733 | 0.4997 |

Table 1: Performance Comparison of various Models

This research not only took customers perception but also took airline companies into consideration. The data used for predictive analytics was only restricted till the month of December however, this research was scaled further to consider the data for entire year to visualize the dataset for finding key elements causing delays and the airports which are suffering adversely due to this delay in United States.

This section will visualize the graphs based upon 8 factors due to which delay could be caused:

1. Arrival Delay
2. Departure Delay
3. Air System Delay
4. Airline Delay
5. Security Delay
6. Weather Delay
7. Cancelled
8. Diverted

## 5.1 Experiment / Case Study 1

**Which airline company is not at all able to cope up with any sort of delay?** According to the represented Figure 8 Southwest Airlines Co. is the worst flight having very high arrival and departure delays (91,786 hours, 209264 hours respectively). This delay is the aggregation of all the Southwest Airlines flights over the period of 1 year travelling from any source to any destination within US. Moreover, American airlines, seems to have a very strict security ultimately causing delay called as Security delay. This security factor estimated a flight on an average to be delayed by 251.4 hours. From the other angle, the Delta Airlines is usually affected by the weather conditions. This in turn made the flight delay by a total of 9917 hours. Southwest Airlines again tops the list of all airlines when it comes to the delay resulted by Airlines called Airlines delay like technical issue in a flight, late arrival of crew members etc. This resulted in Southwest Airlines to be delayed by 60,496 hours as a whole throughout the year. On the other side, Delta Airlines experienced huge delay (of 28,476 hours) due to Air system. One imperative

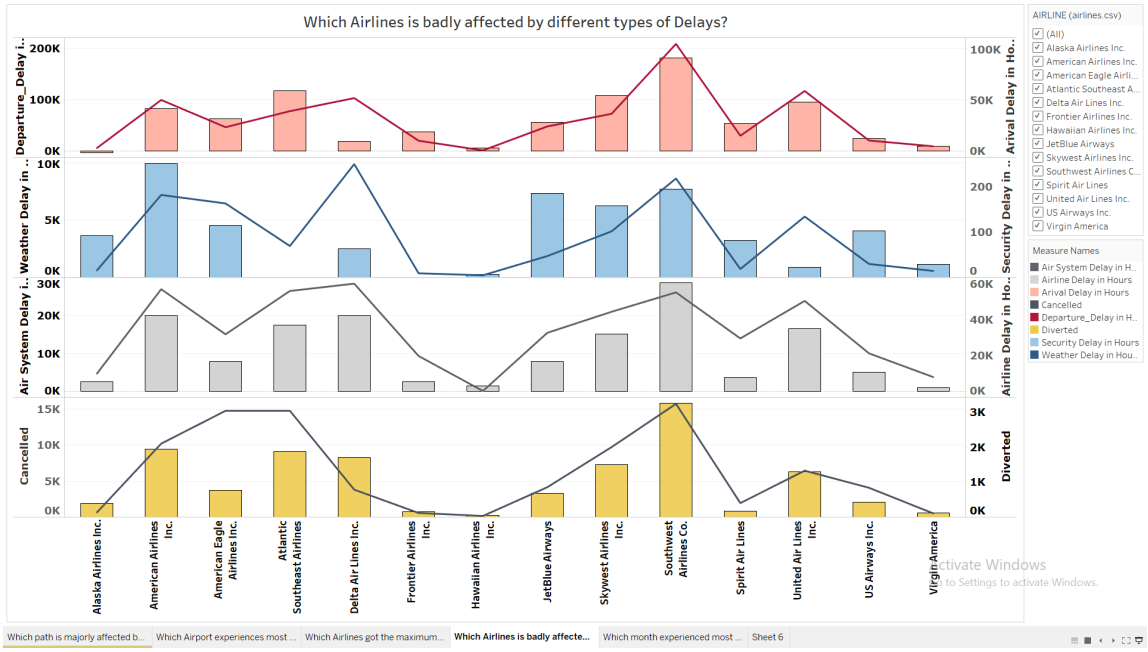


Figure 8: Graphical Representation of Airlines performing worst

cause of this type of delay could be network congestion. Finally, Southwest Airlines, once again, proved to be inevitably bad flight as the quantity of flights being cancelled or diverted skyrocketed. 3268 flights got Diverted and 15,721 flights got Cancelled.

On the contrary, airlines like Alaska Airlines, Frontier Airlines, Hawaiian Airlines and Virgin America are the best flights because they have the lowest amount of delay throughout and hence a passenger can certainly build a trust level on these companies. Overall, this graphical representation contributed to harness the insights as to which airlines company should a passenger rely on. Southwest Airlines is definitely out of the list as it is worst performing under almost all conditions. This company as a whole, can re-evaluate their air crafts and make changes so that they can escalate their performance.

## 5.2 Experiment / Case Study 2

### Which Airports got severely affected due to long delays by Southwest Airlines Co.?

Figure 9 narrowed down its focus to only consider Southwest Airlines Co. with the intention to extract out the path (origin and destination) which is adversely affected by the delay caused in this airline company.

The graph shows that the avenue from Dallas (DAL) to Houston (HOU) received the maximum departure delay of 1129 hours throughout the year. Just because the planes of Southwest Airlines takes longer than the scheduled time the delay is generated and hence this delay is propagated throughout the network resulting in disturbance not only to passengers, as they have to wait till ages, but also to airports as they have to re-schedule entire system for incoming and outgoing flights, costing a lot of money, backsliding economy of United States.

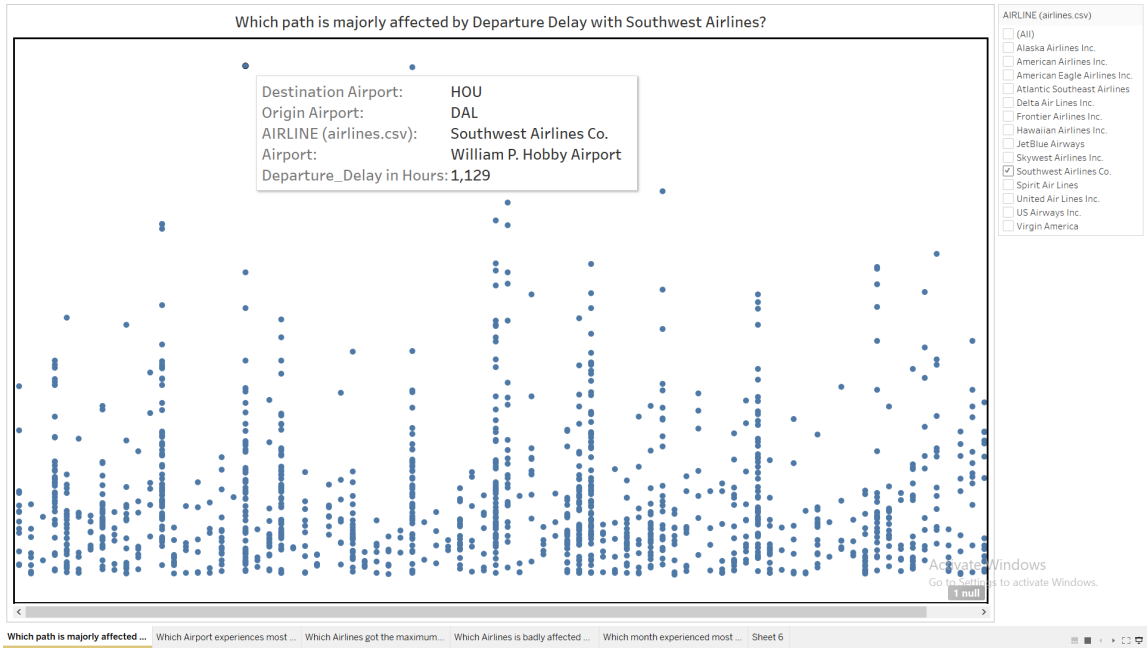


Figure 9: Graphical Representation of the Airport facing most delays

## 5.3 Experiment / Case Study 3

**Which month in the year 2015 faced maximum delay of any sort?**
According to the represented Figure 10 the Research was scaled up further considering all 12 months for the year 2015. According to graph 3, the month of June experienced maximum delay due to incoming and outgoing flights (arrival and departure delays) which was 78869 hours and 115431 hours respectively, maximum delay due to Airlines malfunctioning like faulty behavior of aircraft, maintenance issues or late crew arrival and Air system congestion which is calculated to be 38272 hours and 27181 hours respectively and due to this malfunctioning of airrafts maximum number of diversions i.e. 1930 diversions took place. August 2015 saw most delays (around 200 hours) due to security issues, however, weather significantly affected the month of February creating a delay of 6854 hours in the entire US air transportation network and due to this bad weather 20517 flights got cancelled in the same month. The month of September, however, faced a lot less delays. In fact, it proved to be the best month in the year 2015 by arriving 5951 hours prior to the combined scheduled time.

In a nutshell, the reason due to which a flight is delayed somehow triggers another factor like cancellation or diversion because of which delay is further propagated as cancellation and diversion hurts airports the most and the responsibility to accommodate the delayed aircraft increases. Secondly, during the months of June and February, special precaution must be taken to ensure smooth flow of air transportation throughout US

## 5.4 Discussion

This research project was made with the novel intention to be as generic and simple to use as possible without hampering the quality of analysis, performance metrics and the results. Phase 1 of this Research conspicuously answers the Business Question as Deep Learning Algorithm (ANN) showed a minimum of about 30% improvement in Accuracy

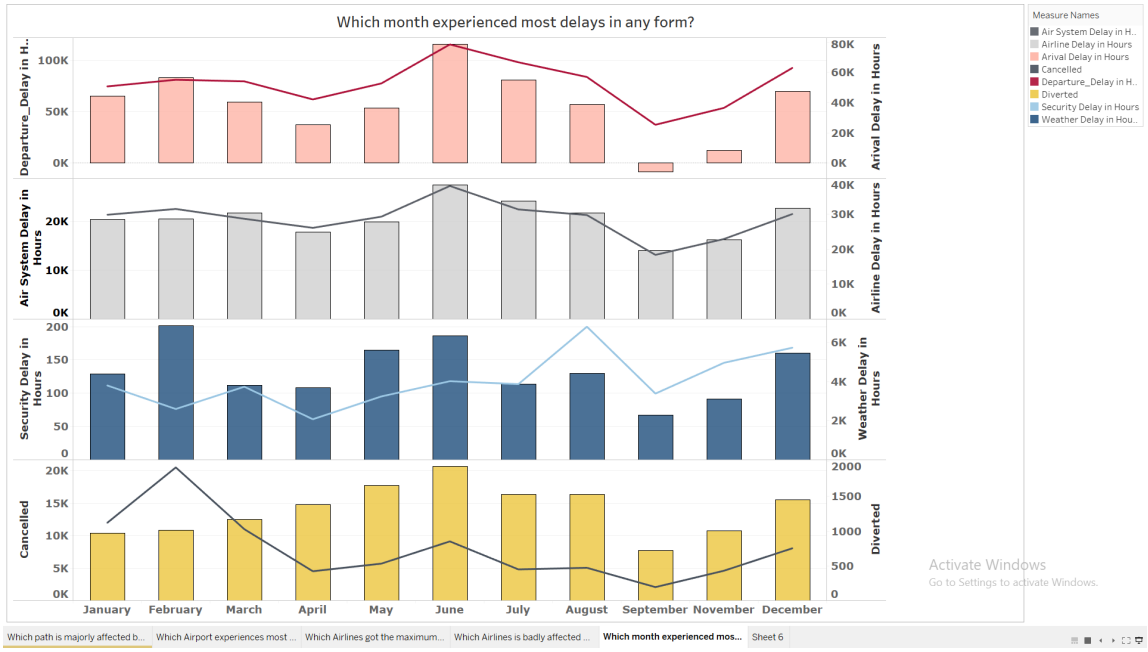


Figure 10: Graphical Representation of Month facing maximum delays

over Naive Bayes and about 12% over XGBoost. Furthermore, ANN is about 45% more precise than Naive Bayes. As per the Phase 2, since, Southwest Airlines is delayed majorly irrespective of the conditions, it should be the last choice for the passengers. Also, during the months of June and February, special arrangements should be implemented to minimize the flight delay.

As a suggestion, the airline companies should strictly monitor the factors for which they solely are responsible for causing delays, like, proper crew timings, proper fueling, proper maintenance. If this is taken care of, flight delays will be significantly reduced.

This piece of research can certainly be used by any air transportation organization in any country with minimal changes of **dataset** and **duration** upon which analysis has to be done. There is no need to make changes in any performance metrics as everything will be calculated automatically by simply compiling the code and executing it in a suitable IDE.

This piece of work contributed to an extent as digging out imperative and deep information with greatest quality without this minimalistic and easy-going system might not be possible this fast.

# 6 Conclusion and Future Work

The research project instigated the performance based predictive analysis of divergent classifiers for United States flight delays. Four models namely ANN, Nave Bayes, KNN and XGBoost were compared against each other grounding performance metrics where KNN followed by ANN proved to be the best overall predictors giving 94% and 90% accuracy respectively while Nave Bayes behaved counter-productively by being 60% accurate.

The two-part Research worked coherently upon investigating the lucrative classifier for predicting United States flight delays on one stage, then, scaling up the scope to harness

the critical factors unpleasantly delaying the flights severely affecting US economy.

Since it is irrevocably vital for airlines companies to resolve any challenge resulting in delayed performance, the visualization phase of this Research provides an apt solution. This would not only benefit the airlines to resolve any shortcoming but would assist customers to make necessary lodgments or arrangements and airports to become sufficiently more productive by effectively managing the arrival and departure times and gates.

However, this research project is only restricted to the dataset of the year 2015 and the month of December, still, it can be extended to use the latest dataset as well. Furthermore, the course of this project uses only Supervised Learning Techniques of Machine Learning and Deep Learning but can be enhanced to use Unsupervised Learning Techniques as is currently being done by PwC, whose prime objective is to predict the flight delays using Unsupervised Learning Methods on the data collected from thousands of sensors from within the plane and on the airports.

The potential of this Research can be expanded by adding a flight recommendation system developed through Deep Learning Approach of Stacked Autoencoders. The future layout could precisely be like, automatically recommending the best flight (causing least delay) after predicting that current flight will be delayed. This would be a breakthrough and would give passengers enormous power to switch flights on the go. Not only this, it would enable airlines to self-judge their level of dignity by comparing themselves to their competitors.

# Acknowledgement

I would like to extend a vote of appreciation to my Supervisor Dr. Cristina Muntean, who has been constantly supporting me and providing her valuable feedback. She has not only guided me in the right avenue but also motivated me during every phase of the Research.
Secondly, I would like to thank Dr. Jason Roche who helped me on-campus and off-campus to rectify any technical issue encountered during this Research.
Eventually, I would love to thank my parents for always standing with me and continuously encouraging me to achieve the goal.

# 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): 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.

Balakrishna, P., Ganesan, R. and Sherry, L. (2010). Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of tampa bay departures, *Transportation Research Part C: Emerging Technologies* **18**(6): 950–962.

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A., Zou, B. et al. (2010). Total delay impact study, *NEXTOR Research Symposium, Washington DC. http://www. nextor. org*.

Beatty, R., Hsu, R., Berry, L. and Rome, J. (1999). Preliminary evaluation of flight delay propagation through an airline schedule, *Air Traffic Control Quarterly* **7**(4): 259–270.

Boswell, S. B. and Evans, J. E. (1997). *Analysis of downstream impacts of air traffic delay*, Lincoln Laboratory, Massachusetts Institute of Technology.

Chen, H., Wang, J. and Yan, X. (2008). A fuzzy support vector machine with weighted margin for flight delay early warning, *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*, Vol. 3, IEEE, pp. 331–335.

Das, N., Kalita, K., Boruah, P. and Sarma, U. (2017). Prediction of moisture loss in withering process of tea manufacturing using artificial neural network, *IEEE Transactions on Instrumentation and Measurement* .

Devi, M. I., Rajaram, R. and Selvakuberan, K. (2007). Automatic web page classification by combining feature selection techniques and lazy learners, *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*, Vol. 2, IEEE, pp. 33–37.

Godfrey, L. B. and Gashler, M. S. (2017). A parameterized activation function for learning fuzzy logic operations in deep neural networks, *arXiv preprint arXiv:1708.08557* .

Gumani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V. and Bhirud, S. (2017). Forecasting of sales by using fusion of machine learning techniques, *Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on*, IEEE, pp. 93–101.

Hunter, G. and Ramamoorthy, K. (2007). Evaluation of the national airspace system aggregate performance sensitivity, *Digital Avionics Systems Conference, 2007. DASC'07. IEEE/AIAA 26th*, IEEE, pp. 1–E.

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

Khanmohammadi, S., Tutun, S. and Kucuk, Y. (2016). A new multilevel input layer artificial neural network for predicting flight delays at jfk airport, *Procedia Computer Science* **95**: 237–244.

Kim, Y. J., Choi, S., Briceno, S. and Mavris, D. (2016). A deep learning approach to flight delay prediction, *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*, IEEE, pp. 1–6.

Manley, B. and Sherry, L. (2010). Analysis of performance and equity in ground delay programs, *Transportation Research Part C: Emerging Technologies* **18**(6): 910–920.

Pathomsiri, S., Haghani, A., Dresner, M. and Windle, R. J. (2008). Impact of undesirable outputs on the productivity of us airports, *Transportation Research Part E: Logistics and Transportation Review* **44**(2): 235–259.

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

Schmidhuber, J. (2015). Deep learning in neural networks: An overview, *Neural networks* **61**: 85–117.

Sokolova, M., Japkowicz, N. and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, *Australian conference on artificial intelligence*, Vol. 4304, pp. 1015–1021.

Stolfo, S., Fan, D. W., Lee, W., Prodromidis, A. and Chan, P. (1997). Credit card fraud detection using meta-learning: Issues and initial results, *AAAI-97 Workshop on Fraud Detection and Risk Management*.

Tewary, G. (2015). Data mining using neural networks, *International Journal of Data Mining & Knowledge Management Process* **5**(2): 65.

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

Vidnerová, P. and Neruda, R. (2017). Evolving keras architectures for sensor data analysis, *Annals of Computer Science and Information Systems* **11**: 109–112.

Wang, P. T., Schaefer, L. A. and Wojcik, L. A. (2003). Flight connections and their impacts on delay propagation, *Digital Avionics Systems Conference, 2003. DASC'03. The 22nd*, Vol. 1, IEEE, pp. 5–B.

Zonglei, L., Jiandong, W. and Guansheng, Z. (2008). A new method to alarm large scale of flights delay based on machine learning, *Knowledge Acquisition and Modeling, 2008. KAM'08. International Symposium on*, IEEE, pp. 589–592.

Zonglei, L., Jiandong, W. and Tao, X. (2009). A new method for flight delays forecast based on the recommendation system, *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, Vol. 1, IEEE, pp. 46–49.