A Comparative Analysis of Models for Predicting Airline Arrival Delays

Alaa Ibrahem
Computer Science Dept.,
Faculty of Computers and information
Menoufia University, Egypt
alaaibrahem1815@gmail.com

Heba Elbeh
Computer Science Dept.,
Faculty of Computers and information,
Menoufia University, Egypt
Heba_Elbeh@yahoo.com

Hamdy M. Mousa
Computer Science Dept.,
Faculty of Computers and information,
Menoufia University, Egypt
hamdimmm@hotmail.com

Abstract— **Flight data is a large source of big data. Million flights are delayed or canceled each year due to several factors. Study aviation systems are being significant to the economy which improves customer satisfaction, and saves time. Delay Prediction in aviation systems is somewhat complicated because of the large volume of data, the multiple causes of delays. The reasons vary from region to region and from company to another. In this paper, we compare the performance of different machine learning approaches (Random Forest Classifier, logistic regression, Gaussian Naive Bayes and Decision Tree Classifier) for predicting the arrival delay depending on the multiple characteristics and mention the features in each approach. Using machine-learning toolkit supported on the Splunk platform to make a comparison between them. The Airline On-Time Performance Data are used for evaluating the models. The results demonstrate that the Logistic regression is better than others and works well with discrete data.**

Keywords— Predicting - Airline Arrival Delays - Models - Splunk - Data analysis.

I. Introduction

Air transportation is a ticklish infrastructure. It is also a multifaceted system, with interactions among numerous components. The challenges of delay spread for other flights on the same day or the same airline company, which shows the importance of predicting delays to avoid its spread, increasing the number of flights at the same time and losing a lot of time and money. The U.S. Department of Transportation (DOT) released its March 2018 Air Travel Consumer Report (ATCR) on air carrier data compiled for the month of January 2018[1]. In January 2018, the reporting carriers posted an on-time arrival rate of 79.6 percent, up from the 76.0 percent on-time rate in January 2017, but slightly down from the 80.3 percent mark in December 2017. In the United States, when flights are canceled or delayed, passengers may be entitled to compensation due to rules obeyed by every flight company; this rule usually specifies that passengers may be entitled to certain reimbursements, where flight considered late after 15 minutes of flight time original. The most important thing for many passengers is accessing on time and knowing the flight delay helps to avoid wasting time and wasting their commitment opportunities. They are preferring early booking better than delay. The main objective of the paper is to compare the performance of different machine learning approaches for predicting the Airline arrival delay and explore the features of ML Toolkit, which provides difference classification algorithms, set up alarms and display dashboard. In this paper, section II discusses some related researches. Section III explains the prediction methodologies. Section IV Features selection. Section IV shows experiments and results. Finally, we make a conclusion on the project.

II. Background

Aircraft delay has a lot of reasons and many models were built to predict the spread of delays. Prediction of the arrival delay has also been an active topic. In [2], they tried to develop a model that aims to predict flight delay. They found that gradient boosting (XG Boost) gives good results with the computational cost. They have shown that the flight delay prediction is tractable. Decision trees and random forests are probably the best way to approach this problem.

i. Karthik, et al used National Airspace System (NAS) operational data to evaluate the models (Markov Jump Linear System (MJLS) model, Artificial Neural Network (ANN) model, classical machine learning techniques like Classification and Regression Trees (CART). Temporal, local and network features were used to make these predictions. They showed that the best prediction method depends on a) determination the problem classification or regression. b) Types of data balance. c) Prediction horizon [3].

ii. Jianmo Ni, et al study a public dataset to analyze the flight delay in the United States. Six classification tools were used to classify flight delayed or non-delayed and improved performance using the weight loss. Weight-loss experiment results show the importance of weighted loss function and short-term temporal data for great performance [4].

iii. Juan Jose and Hamsa presented new network-based air traffic delay prediction models that incorporated both temporal and network delay states as explanatory variables. The results obtained for the 100 most-delayed origin-destination (OD) pairs in the NAS showed an average test error of 19% when classifying delays as above or below 60 minutes, for a 2-hour forecast horizon. The median regression test error (averaged across the 100 OD pairs) only increased from 19.1 minutes to 27.4

minutes when the forecast horizon increased from 2 hours to 24 hours [5].

iv. Brett Naul used large historical datasets to make predictions. For the purposes of this paper, only departure delays are studied. They tried to solve the problem by using a number of classifiers including logistic regression, Naive Bayes classifiers, and Support Vector Machines (SVM) classifiers. Simple, uncomplicated algorithms outperformed more sophisticated methods. The simple algorithms actually performed rather impressively: in particular, the Naive Bayes probability estimates proved to be more accurate than the historical estimates used by many travel sites [6].

v. Alice Sternberg et al publish a comprehensive review of all scientific research and classifications in building predictive models of flight delays [7].

This paper differs from the previous papers in two points.
1- We used very large and varied data, and this is a point that the previous papers didn't care about.
2 - Training was conducted on four different algorithms to study the differences between them. Feature selection algorism helpful for select most related features.

In my view, researches in this field are still full of challenges and to be exploited and developed by machine tools. The data is constantly increasing, analysts should focus on developing data cleaning and feature selection methods.

III. Toolkit and Prediction Methodologies

There are set of techniques for forecasting and predicting (e.g., Traditional Time Series Prediction Methods, Chaos theory, Statistical analysis, Probabilistic models, Comparative methods, Network Representation, Operational Research and Machine learning methods) to handle and model the problem of the flight delay based on the research's objectives. In this section, the brief description of the used methodologies is elucidated.

A. Splunk Machine Learning Toolkit:

When we cannot directly solve a given problem by writing a computer program and past experience or samples of data are available, in these cases, we need learning. Machine learning is computer's programs for modeling and optimizing a performance criterion using these past experience or example data as shown in figure 1. Machine learning methods are commonly used for study flight systems. The exploration of information from data using learning algorithms and building models. There are two major types of learning. The first one is called supervised learning which trains with known data samples of inputs and outputs. When there is training set of input without target output of them, this type of learning calls unsupervised learning.

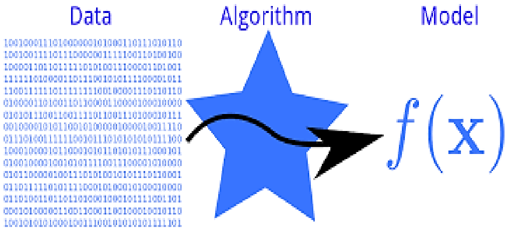

Figure 1: Machine learning workflow.

- Splunk is a general-purpose search, analysis, reporting, dynamic and scalable IT data engine for time-series text data, typically machine Data. Splunk makes it easy to use the power of machine learning to optimize your data [8]. Splunk Machine learning Tool Kit supports many algorithms that can be used to classify the data e.g. Random Forest Classifier, Decision Tree Classifier, Logistic regression and etc.
- Supervised Learning: We mention before that Supervised learning is the first type of learning which construct algorithms to produce patterns and hypothesis using past information from external examples are given with known labels (correct output) to predict future cases .Supervised machine learning has several algorithms like (Decision Table, Random Forest (RF) , Naïve Bayes (NB) , Support Vector Machine (SVM), Neural Networks (Perceptron),Decision Tree, etc.) as well as determines the most efficient classification algorithm based on the data set, the number of instances and variables.

B. Random Forest Classifier:

Random forest is a classifier that consists of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. Random Forest uses an ensemble of decision trees as a basis and therefore has all advantages of decision trees, such as high accuracy, easy usage, and no necessity of scaling data [9].

C. Decision Tree Classifier:

Decision Trees are a type of Supervised Machine Learning. A decision tree consists of internal nodes that represent the decisions corresponding to the hyper-planes or split-points (i.e., which half-space a given point lies in), and leaf nodes that represent regions or partitions of the data space, which are labeled with the majority class. A region is characterized by the subset of data points that lie in that region. Where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split [10].

D. Logistic regression:

Logistic regression is the most famous machine learning algorithm after linear regression. In a lot of ways, linear regression and logistic regression are similar. But the biggest difference lies in what they are used for. Linear regression algorithms are used to predict/forecast values but logistic regression is used for classification tasks. Each of the features also has a label of only 0 or 1. Logistic regression is a linear classifier and therefore used when there is some sort of linear relationship between the data [11].

E. Gaussian Naive Bayes (NB):

Naive Bayes (NB) is a simple technique for constructing classifiers A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It’s specifically used when the features have continuous values. It’s also assumed that all the features are following a Gaussian distribution [12].

F. Feature selection

Feature selection (fs) is a key aspect of machine learning problems, and greatly influences the performance of the algorithms. Using Field Selector algorithm to choose feature variables related to the target variable. This Field Selector algorithm uses the sci-kit learn Generic Univariate Select for selecting the best predictor fields based on univariate statistical tests [13]. It chooses feature variables that have the strongest relationship with the output variable.

IV. DATASET AND FEATURES

The US Bureau of Transport Statistics data set for the year 2021 & 2020 is chosen. This data set consists of about 500000 samples with 30 features. We will focus only on the Airline On-Time Performance data for June 2021 & December 2020, provides on-time data for flights. They operate on-time arrival and departure data for non-stop domestic flights by month, year, carrier, and origin. It includes scheduled actual, departure and arrival times, canceled and diverted flights, taxi-out and taxi-in times, causes of delay and cancellation, air time, and non-stop distance. We should be familiar with the data to understand its features there are many factors caused delays. Let's identify major features that cause delay:

1. *Weather*: This is the most important feature. It is one of the biggest causes of delaying and disrupting flights. Many researchers in this subject have addressed this factor because of its impact. It is influenced by mild rain, heavy rain, mild wind, high wind and precipitation of rainfall and knowledge of the possibility of vision.
2. *Day of week and Months*: The factor of the day is also effective in predicting delays. It is found that a certain day of the week is busy and there are many delays, and this also spreads over months. The months of holidays or months that have bad weather are susceptible more than the remaining months of the year for delays. Temporal features affect the prediction of flight delays (Day of Week, Month, and Day of the month).
3. *National Aviation System (NAS):* Delays and cancellations attributable to the national aviation system that refers to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
4. *CARRIER:* Code assigned by IATA and commonly used to identify a carrier. When studying airlines and their knowledge of the possibilities of their impact on the delays, it is very useful to expect that this company is more prone to delay than others.
5. *ArrDel15:* Arrival Delay Indicator, it is true if arrival delay is 15 Minutes or more.
6. *DepDel15:* Departure Delay Indicator, it is true if departure delay is 15 Minutes or more.

When studying airlines and their knowledge of the possibilities of their impact on the delays, it is very useful to expect that this company more prone to delay than others.

Figure 2 show delay rate in 2017, 2018.

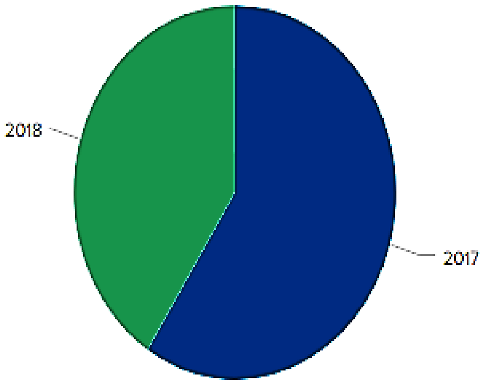

Figure 2: The delay rate between 2017 and 2018 which shows an increase in access rate of 15%.

V. V Experiments Results and Discussion

The main objective of this paper is to compare and evaluate the performance of different machine learning approaches (Random Forest Classifier, logistic regression, Gaussian Naive Bayes and Decision Tree Classifier) for predicting the arrival delay. Using Splunk 8.0.1, Splunk machine learning toolkit version 5.0.0 to execute comparison among them. The Airline On-Time Performance Data are used for evaluating those models. First, we split the dataset 70% for the training set and 30% for testing set with 13 features. We perform these experiments on Samsung core i3 with 12GB RAM and Windows 7 ultimate edition operating system 64-bit. Consider prediction problem (i.e., there are only correct prediction and n samples that need to be predicted), so a given sample either be located into right category (**ArrDel15 or not)** (i.e., positive example) or does not be located into right category (**ArrDel15 or not)** (i.e., negative example). Assume that the target output of samples data and predicted output of machine learning approach are carried out. Then recall, precision, fallout, and error rate are defined as:

$$Non - Delay\ Recall = \frac{A}{A + C} \qquad (1)$$

$$Delay\ \ Recall\ = \frac{B}{B + D} \qquad (2)$$

$$Non - Delay\ Precision = \frac{A}{A + D} \qquad (3)$$

$$Delay\ \ Precision = \frac{B}{B + C} \qquad (4)$$

$$Accuracy = \frac{A + B}{N} \qquad (5)$$

$$F1 - score\ = \frac{2\ x\ Recall\ x\ Precision}{Recall + Precision} \qquad (6)$$

$$Fallout = \frac{D}{B + D} \qquad (7)$$

$$Error\ Rate = \frac{C + D}{N} \qquad (8)$$

$$N\ =\ A\ +\ B\ +\ C\ +\ D \qquad (9)$$

Where *A* is number of samples that both the target output and predicted are non-delay (True positive), *B* is number of samples that both the target output and predicted are delay (True negative). *C* is number of samples that the target output classifies as non-delay but the predicted output classifies as delay(False negative), *D* is number of samples that the target output classifies as delay but the predicted output classifies as non-delay (False positive), and N is total number of test samples. Figure 3 illustrates the confusion matrix of real target output and predicted output that classify as non-delay or delay.

	Real Non-Delay	Real Delay
Predicted Non-Delay	A	D
Predicted Delay	C	B

Figure 3**:** Confusion matrix of real target output and predicted output

Experiment 1**:** for testing the effect of feature selection into the performance of four machine learning approaches and quality of prediction. December 2020 data was using. Experiment was setup by using Field selector algorithm in Splunk machine learning toolkit to select features that have strong relation to target output of each approach. ("CRS_DEP_TIME","DAY_OF_MONTH","FIRST_DEP_TIME"). Figure 4 shows recall, precision, Accuracy, fall out, Error rate, and F1-score for every machine learning approach.

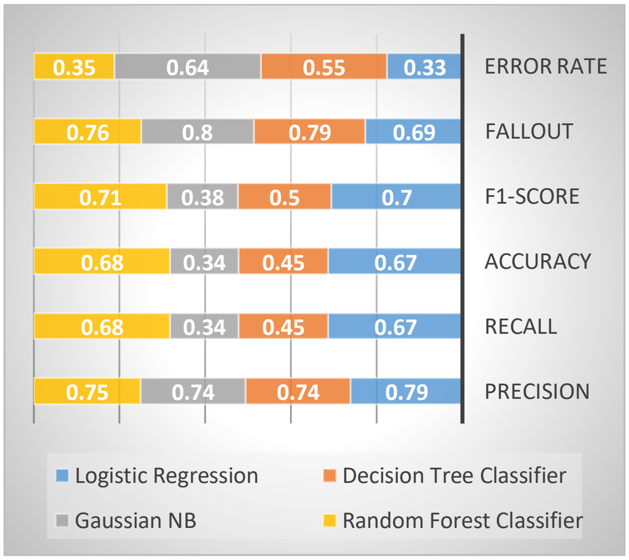

Figure 4**:** shows experiment 1 results.

Experiment 2: In this experiment, we will use all parameters that chosen by field selector, but from another time of the year, June 2021 to show if there another features that affect the delays will change or not.

Evaluating the performance of four machine learning approaches (Random Forest Classifier, logistic regression, Gaussian Naive Bayes and Decision Tree Classifier**).**

Table 2: Second Experiment's Precision, Recall, Accuracy, fall out, Error rate and F1-score of machine learning approaches.

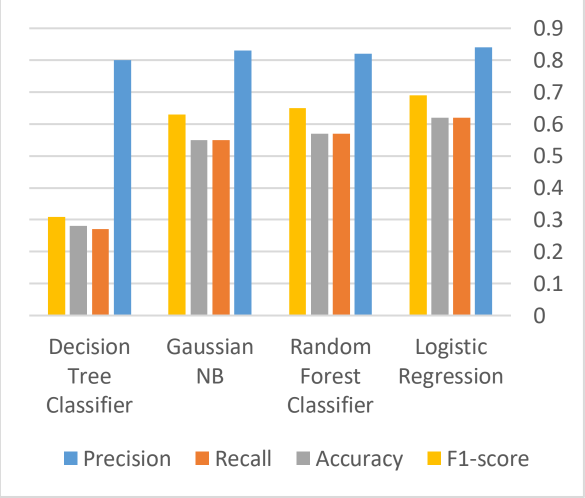

Figure 5**:** shows experiment 2 results.

It is apparent that the best choice of delay prediction method depends on the time-of-day was the most important factor driving the accuracy of delay classification. We noticed that the feature selection noticed that origin city & destination cities, although related to predicting arrival delay. For the classification problem in figure 4&5 logistic regression achieved higher rates that show the impact of the logistic regression algorithm which performs better when the dataset has features that are linearly separable. Naive Bayes classifiers have high accuracy and speed on large datasets, but some problems happen due to data lack. For any possible value of a feature, you need to estimate a probability value by a frequents approach. This can result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results. Random Forest Classifier achieved less accuracy Due to their complexity; they require much more time to train than other comparable algorithms but uses Ensemble Learning technique which creates as many trees on the subset of the data and combines the output of all the trees. In this way it reduces over fitting problem in decision trees and also reduces the variance and therefore improves the accuracy. Decision tree algorithm a slight change in the data can lead to a significant change in the structure of the decision tree, causing instability although Missing values in the data does not affect the process of building decision tree to any huge extent

.

CONCLUSIONS

We compared the performance of several algorithms (Logistic Regression, Random Forest Classifier, Gaussian NB, and Decision Tree Classifier) for flight delay prediction. Temporal factors like [Day of week, day of the month] were the most important factors used to make these predictions. Gaussian NB and Decision Tree Classifiers algorithms gave least accuracy. Logistic regression, despite its simplicity, it produced better results than the rest. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features. The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.

REFERENCES

[1] Office of Aviation Enforcement and Proceedings, "Air Travel Consumer Reports," pp. 1–53, 2016.

[2] N. Movva and S. Menon, "Predicting flight delays and cancellations using weather as a feature," pp. 1–6, 2016.

[3] K. Gopalakrishnan and H. Balakrishnan, "A comparative analysis of models for predicting delays in air traffic networks," 12th USA/Europe Air Traffic Manag. R D Semin., 2017..

[4] J. Ni, X. Wang, and Z. Li, "Flight Delay Prediction using Temporal and Geographical Information," pp. 1–4, 2017..

[5] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," Transp. Res. Part C Emerg. Technol., vol. 44, pp. 231–241, 2014

[6] B. Naul, "Airline Departure Delay Prediction," Bernoulli, pp. 1–5, 2008.

[7] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, "A Review on Flight Delay Prediction," pp. 1–21, 2017.

[8] S. Sujitparapitaya, A. Shirani, and M. Roldan, "Issues in Information Systems," Issues Inf. Syst., vol. 13, no. 2, pp. 112–122, 2012.

[9] F. Baumann, F. Li, A. Ehlers, and B. Rosenhahn, "Thresholding a Random Forest classifier," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8888, pp. 95–106.

[10] M. J. Zaki and W. Meira, Jr, "Decision Tree Classifier," in Data Mining and Analysis, 2018, pp. 481–497.

[11] O. L. Regression, S. Data, and A. Examples, "Ordered Logistic Regression | Stata Data Analysis Examples," UCLA Stat. Consult. Gr., pp. 1–17, 2018.

[12] R. Saxena, "Gaussian Naive Bayes Classifier implementation in Python," Dataaspirant. 2017.

[13] https://docs.splunk.com/Documentation/MLApp/4.4.1/User/ Algorithms . October 2019