Predicting Flight Delays with Error Calculation using Machine Learned Classifiers

Revooru Siva[1], Sandeep Kumar[2]

Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract— *Flight delay is a major problem in the aviation sector. During the last two decades, the growth of the aviation sector has caused air traffic congestion, which has caused flight delays. Flight delays result not only in the loss of fortune also negatively impact the environment. Flight delays also cause significant losses for airlines operating commercial flights. Therefore, they do everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In Tree Regression this paper, using machine learning models such as Logistic Regression, Decision Bayesian, Ridge, Random Forest Regression and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.*

Keywords: Flight Prediction, Machine Learning, Error Calculation, Logistic Regression, Decision Tree, Bayesian Ridge, Random Forest, Gradient Boosting, Logistic Regression, U.S. Flight data.

I. INTRODUCTION

Flight delay is studied vigorously in various researches in recent years. The growing demand for air travel has led to an increase in flight delays. According to the Federal Aviation Administration (FAA), the aviation industry loses more than $3 billion in a year due to flight delays and, as per BTS, in 2016 there were 860,646 arrival delays. The reasons for the delay of commercial scheduled flights are air traffic congestion, passengers increasing per year, maintenance and safety problems, adverse weather conditions, the late arrival of plane to be used for next flight. In the United States, the FAA believes that a flight is delayed when the scheduled and actual arrival times differs by more than 15 minutes. Since it becomes a serious problem in the United States, analysis and prediction of flight delays are being studied to reduce large costs.

II. LITERATURE REVIEW

Development of a Predictive Model for On-Time Arrival Flight of Airliner by Discovering Correlation between Flight and Weather Data

Noriko Etani - 2019

An important business of airlines is to get customer satisfaction. Due to bad weather, a mechanical reason, and the late arrival of the aircraft to the point of departure, flights delay and lead to customer dissatisfaction. A predictive model of on-time arrival flight is proposed with using flight data and weather data. The key research in this paper is to discover the correlation between flight data and weather data. The relation between pressure pattern and flight data of Peach Aviation, which is LCC (low-cost carrier) in Japan, are clarified, and it is found that the sea-level pressures of 3 weather observation spots, which are Wakkanai as the most northern spot, Minami-Torishima as the most eastern spot, and Yonagunijima as the most western spot, can classify the pressure patterns. As a result, on-time arrival fight is predicted at 77% of the accuracy with using Random Forest Classifier of machine learning. Furthermore, feasibility of the predictive model is evaluated by developing a tool of on-time arrival flight prediction.

Flight Delay Prediction for Commercial Air Transport: A Deep Learning Approach

This study analyzes high-dimensional data from Beijing International Airport and presents a practical flight delay prediction model. Following a multifactor approach, a novel deep belief network method is employed to mine the inner patterns of flight delays. Support vector regression is embedded in the developed model to perform a supervised fine-tuning within the presented predictive architecture. The proposed method has proven to be highly capable of handling the challenges of large datasets and capturing the key factors influencing delays. This ultimately enables connected airports to collectively alleviate delay propagation within their network through collaborative efforts (e.g., delay prediction synchronization).

A Review on Flight Delay Prediction

Alice sternberg, Jorge Soares, Diego Carvalho, Eduardo Ogasawara – 2017

Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became

cumbersome due to the complexity of air transportation system, the number of methods for prediction, and the deluge of flight data. In this context, this paper presents a thorough literature review of approaches used to build flight delay prediction models from the Data Science perspective. We propose a taxonomy and summarize the initiatives used to address the flight delay prediction problem, according to scope, data, and computational methods, giving particular attention to an increased usage of machine learning methods. Besides, we also present a timeline of significant works that depicts relationships between flight delay prediction problems and research trends to address them.

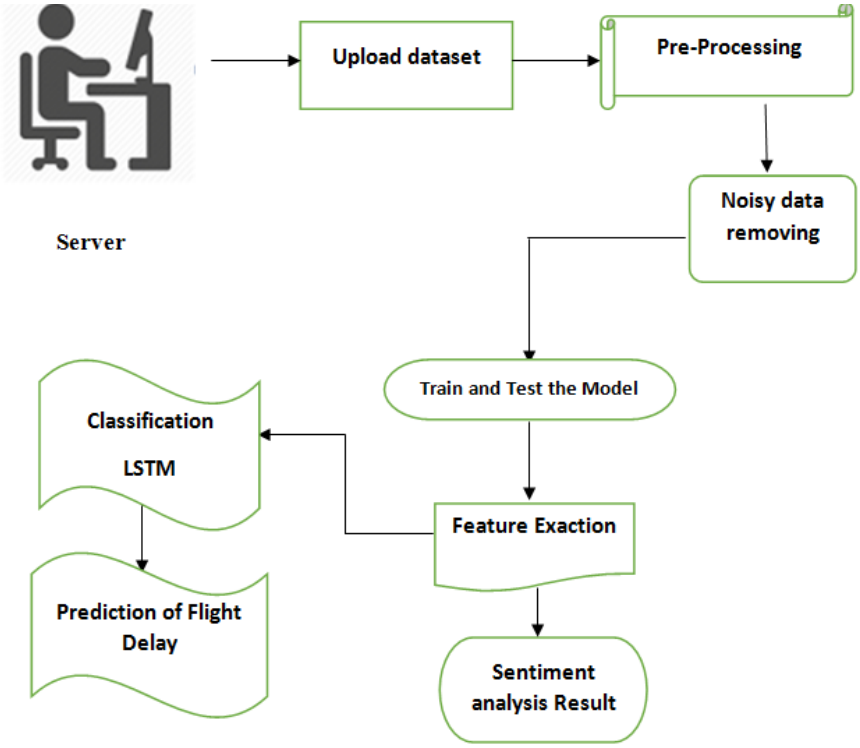
Flight Arrival Delay Prediction and Analysis Using Ensemble Learning

Xiaotong Dou - 2020

With the development of the civil aviation transportation industry in recent years, the volume of civil aviation transportation has increased rapidly. Increased carrier costs and reduced airport operating efficiency caused by flight delays have become issues that need to be addressed. How to improve the accuracy of predicting flight arrival delay time is of great significance for improving airport transportation efficiency, rationally scheduling flights and improving passenger comfort. In this paper, the Cat-boost model is utilized on the U.S Domestic airline on-time performance data from U.S. Transportation Administration, combined with the characteristics of the model to determine the influencing factors, and to predict the arrival delays of flights within the United States. The accuracy; precision and some other criterion of the model are given to evaluate the performance on the data. A better effect is obtained: the accuracy reaches 80.44% in this case. Finally, the specific delay time is predicted, we found that the support vector machine has the best prediction result for the flight delay time, the average prediction error is 9.733 min, which has a certain reference value for flight operation and airport scheduling.

A Statistical Approach To Predict Flight Delay Using Gradient Boosted Decision Tree

Supervised machine learning algorithms have been used extensively in different domains of machine learning like pattern recognition, data mining and machine translation. Similarly, there has been several attempts to apply the various supervised or unsupervised machine learning algorithms to the analysis of air traffic data. However, no attempts have been made to apply Gradient Boosted Decision Tree, one of the famous machine learning tools to analyse those air traffic data. This paper investigates the effectiveness of this successful paradigm in the air traffic delay prediction tasks. By combining this regression model based on the machine learning paradigm, an accurate and sturdy prediction model has been built which enables an elaborated analysis of the patterns in air traffic delays. Gradient Boosted Decision Tree has shown a great accuracy in modelling sequential data. With the help of this model, day-to-day sequences of the departure and arrival flight delays of an individual airport can be predicted efficiently. In this paper, the model has been implemented on the Passenger Flight on-time Performance data taken from U.S. Department of Transportation to predict the arrival and departure delays in flights. It shows better accuracy as compared to other methods.



Proposed System

- Our proposed model does everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures.
- In this model, using machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression and Gradient Boosting Regression
- We predict whether the arrival of a particular flight will be delayed or not.
- We develop a system that predicts for a delay in flight departure based on certain parameters. We train our model for forecasting using various attributes of a particular flight, such as arrival performances, flight summaries, origin/destination, etc.

Advantages

- The system collects huge number of datasets to train to the model and predict the flight delay error calculation.
- Speed and accuracy score is high.
- Prediction rate is high.

III. METHODOLOGY

3.1 Data collection

To predict flight delays to train models, we have collected data accumulated by the Bureau of Transportation; U.S. Statistics of all the domestic flights taken in 2015 was used. The US Bureau of Transport Statistics provides statistics of arrival and departure that includes actual departure time, scheduled departure time, and scheduled elapsed time, wheels-off time, departure delay and taxi-out time per airport. Cancellation and Rerouting by the airport and the airline with the date and time and flight labelling along with airline airborne time are also provided. The data set consists of 25 columns and 59986 rows. Fig. 1 shows some of the fields of the original dataset. There were many lines with missing and null values. The data must be pre-processed for later use

The methodology here uses the supervised learning technique to gather the advantages of having the schedule and real arrival time. Initially, some specific monitoring algorithms with a light computation cost were considered candidates and therefore the best candidate was perfected for the final model.

We develop a system that predicts for a delay in flight departure based on certain parameters. We train our model for forecasting using various attributes of a particular flight, such as arrival performances, flight summaries, origin/destination, etc

3.2 Pre-Processing

Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

Text pre-processing is an essential a part of any NLP method and the significance of the NLP pre-processing are

- To minimize indexing (or knowledge) records dimension of the textual content records.
    1. Stop words bills 20-30% of total phrase counts in a special textual content record.
    2. Stemming may just diminish indexing size as much as forty- 50%.
- To make stronger the efficiency and effectiveness of the IR method.
    1. Stop words aren't valuable for shopping or textual content mining.
    2. Stemming used for matching the similar words in a text record.

3.2.1 Tokenization:

Tokenization is the process of breaking a circulate of textual content into phrases, phrases, symbols, or different significant factors called tokens .The aim of the tokenization is the exploration of the phrases in a sentence. The list of tokens turns into input for further processing akin to parsing or textual content mining. Tokenization is valuable both in linguistics (where it's a form of textual content segmentation), and in laptop science, the place it forms a part of lexical analysis. Textual knowledge is simplest a block of characters at the starting.

All strategies in know-how retrieval require the words of the data set. For that reason, the requirement for a parser is a tokenization of records. This might be sound trivial because the text is already saved in computing device-readable codecs. However, some problems are nonetheless left, like the removing of punctuation marks. Different characters like brackets, hyphens, and so on require processing as well.

3.2.2 Stop word Removal:

Stop phrases are very more often than not used fashioned phrases like ‘and’, ‘are’, ‘this’ etc. They don't seem to be useful in classification of records. So they must be removed. However, the development of such stop phrases record is problematic and inconsistent between textual sources. This process also reduces the text knowledge and improves the approach performance. Each textual content report offers with these phrases which are not vital for text mining applications.

3.2.3 Stemming and Lemmatization:

The aim of both stemming as well as lemmatization is to scale down inflectional types & mostly derivationally associated varieties of a phrase to a fashioned base kind.

Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of accomplishing this goal accurately more often than not, and quite often involves the removal of derivational affixes.

Lemmatization often refers to doing matters competently with the usage of a vocabulary and morphological analysis of phrases, in most cases aiming to eliminate inflectional endings only and to come back the base or dictionary type of a word, which is often called the lemma.

3.3 Feature Extraction

We have studied from various sources to find out which parameters will be most appropriate to predict the departure and arrival delays. After several searches, we conclude the following parameters:

- Day Departure
- Delay Airline
- Flight Number
- Destination Airport
- Origin Airport
- Day of Week
- Taxi out

3.4 Evaluation

After pre-processing and feature extraction of our dataset, 60% of the dataset was selected for training and 40% of the dataset was selected for testing. For error calculation, we are using scikit-learn metrices. Results are divided between two sections, Departure Delay (A) and Arrival Delay (B).

3.4.1 Departure Delay:

Results for departure delay which compares different Machine Learning models, i.e. Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor and Gradient Boosting Regressor, based on various evaluation metrics. Further, we compare each model concerning one evaluation metric at a time.

3.4.2 Arrival Delay

Results for arrival delay which compares different Machine Learning models, i.e. Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor and Gradient Boosting Regressor, based on various evaluation metrics. Further, we compare each model concerning one evaluation metric at a time.

IV. IMPLEMENTATION

Sentiment Analysis

```
tweet_id                      0
airline_sentiment             0
airline_sentiment_confidence  0
negativereason             5462
negativereason_confidence  4118
airline                       0
airline_sentiment_gold    14600
name                          0
negativereason_gold       14608
retweet_count                 0
text                          0
tweet_coord               13621
tweet_created                 0
tweet_location             4733
user_timezone              4820
```

dtype: int64

	text	airline_sentiment
3	virginamericaits really aggressive to blast o...	negative
4	virginamerica and its a really big bad thing a...	negative
5	virginamerica seriously would pay a flight fo...	negative
6	virginamerica yes nearly every time i fly vx t...	positive
8	virginamerica well ididntbut now i do d	positive

Positive: 2292
Negative 9113
Neutral 0

	text	airline_sentiment
3	virginamericaits really aggressive to blast o...	negative
4	virginamerica and its a really big bad thing a...	negative
5	virginamerica seriously would pay a flight fo...	negative
6	virginamerica yes nearly every time i fly vx t...	positive
8	virginamerica well ididntbut now i do d	positive
...	...	...
14631	americanair thx for nothing on getting us out ...	negative
14633	americanair my flight was cancelled flightled ...	negative
14634	americanair right on cue with the delays	negative
14636	americanair leaving over minutes late flight ...	negative
14638	americanair you have my money you change my fl...	negative

11405 rows × 2 columns

```
MONTH                574268
DAY_OF_MONTH         574268
DAY_OF_WEEK          574268
OP_UNIQUE_CARRIER    574268
ORIGIN               574268
DEST                 574268
DEP_TIME             569330
DEP_DEL15            569317
DISTANCE             574268
Unnamed: 9                0
dtype: int64
```

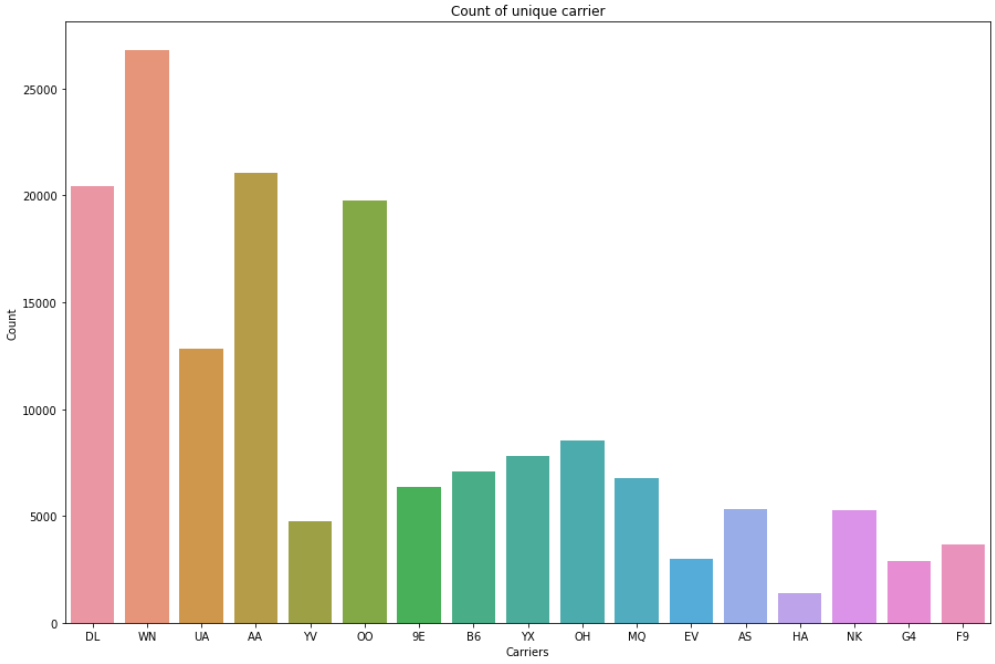


FIGURE 1: Count of Unique Carrier

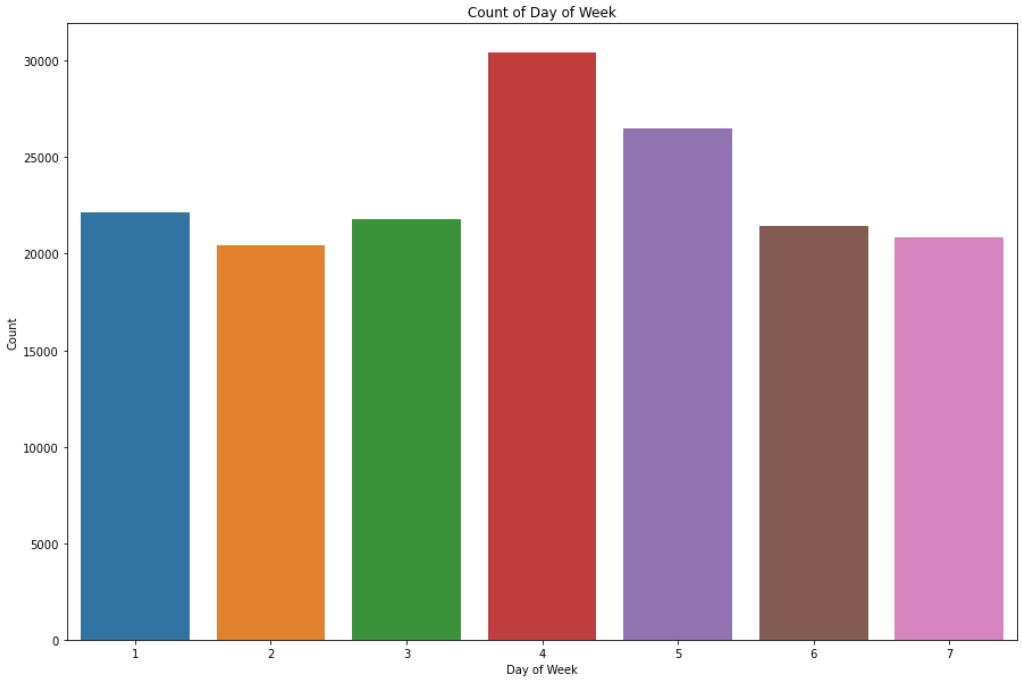


FIGURE 2: Count Day by Week

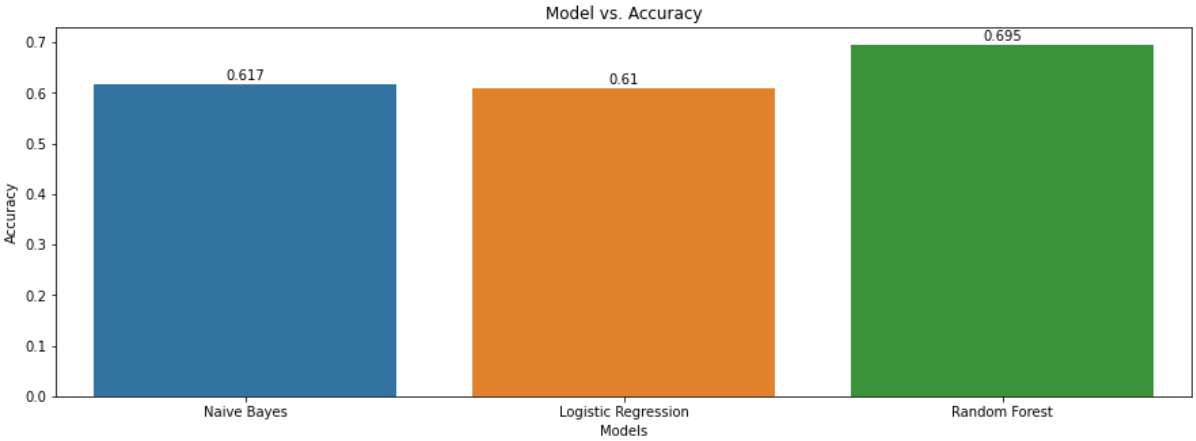


FIGURE 3: Model vs Accuracy

V. CONCLUSION AND FUTURE WORK

Machine learning algorithms were applied progressively and successively to predict flight arrival & delay. We built five models out of this. We saw for each evaluation metric considered the values of the models and compared them. We found out that: -

In Departure Delay, Random Forest Regressor was observed as the best model with Mean Squared Error 2261.8 and Mean Absolute Error 24.1, which are the minimum value found in these respective metrics. In Arrival Delay, Random Forest Regressor was the best model observed with Mean Squared Error 3019.3 and Mean Absolute Error 30.8, which are the minimum value found in these respective metrics.

In the rest of the metrics, the value of the error of Random Forest Regressor although is not minimum but still gives a low value comparatively. In maximum metrics, we found out that Random Forest Regressor gives us the best value and thus should be the model selected.

The future scope of this paper can include the application of more advanced, modern and innovative pre-processing techniques, automated hybrid learning and sampling algorithms, and deep learning models adjusted to achieve better performance. To evolve a predictive model, additional variables can be introduced. e.g., a model where meteorological statistics are utilized in developing error-free models for flight delays. In this paper we used data from the US only, therefore in future, the model can be trained with data from other countries as well. With the use of models that are complex and hybrid of many other models provided with appropriate processing power and with the use of larger detailed datasets, more accurate predictive models can be developed. Additionally, the model can be configured for other airports to predict their flight delays as well and for that data from these airports would be required to incorporate into this research.

REFERENCES

[1] Chakrabarty, Navoneel, Tuhin Kundu, Sudipta Dandapat, Apurba Sarkar, and Dipak Kumar Kole. "Flight arrival delay prediction using gradient boosting classifier." In Emerging Technologies in Data Mining and Information Security, pp. 651-659. Springer, Singapore, 2019.

[2] Chakrabarty, Navoneel. "A data mining approach to flight arrival delay prediction for american airlines." In 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), pp. 102-107. IEEE, 2019.

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

[4] Sternberg, A., Soares, J., Carvalho, D. and Ogasawara, E., 2017. A review on flight delay prediction. arXiv preprint arXiv:1703.06118.

[5] Ding, Y., 2017, August. Predicting flight delay based on multiple linear regression. In IOP conference series: Earth and environmental science (Vol. 81, No. 1, p. 012198). IOP Publishing.

[6] Manna, Suvojit, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, and Subhas Barman. "A statistical approach to predict flight delay using gradient boosted decision tree." In 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1-5. IEEE, 2017.

[7] Dou, Xiaotong. "Flight arrival delay prediction and analysis using ensemble learning." In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 1, pp. 836-840. IEEE, 2020.

[8] Chen, Jun, and Meng Li. "Chained predictions of flight delay using machine learning." In AIAA Scitech 2019 forum, p. 1661. 2019.

[9] Rodríguez-Sanz, Álvaro, Fernando Gómez Comendador, Rosa Arnaldo Valdés, Javier Pérez-Castán, Rocío Barragán Montes, and Sergio Cámara Serrano. "Assessment of airport arrival congestion and delay: Prediction and reliability." Transportation Research Part C: Emerging Technologies 98 (2019): 255-283.

[10] Kuhn, Nathalie, and Navaneeth Jamadagni. "Application of machine learning algorithms to predict flight arrival delays." CS229 (2017).