Airline Delay Prediction

Najeebuddin Ahmed, Khanjan Dabhi

***Abstract*—The following report is a literature review done to find different methods by which predictions for airline delay is made. This report also explores various problems which were tackled by the prediction models from a data science perspective and our proposed method for the solution.**

***Index Terms*—Delay, Literature review.**

I. Introduction

Airlines from the United States have been one of the most dependent and the world's biggest carriers regarding the number of destinations/arrivals served. Yet, with regards to homegrown flights, they have not satisfied the hopes with as far as promptness or on-time execution. Flight Delays additionally bring about carrier organizations working business trips to cause enormous loss. Along these lines, assessment of components influencing delays help in reducing loss in the flying industry on an everyday schedule. To better understand the entire flight ecosystems, vast volumes of data from commercial aviation are collected every moment and stored in databases. Submerged in this massive amount of data produced by sensors and IoT analysts and data scientists are intensifying their computational and data management skills to extract useful information from each datum. This report tries to summarize the most important trends in this field, showing how this problem is addressed and comparing all the various methods that have been used to build predictive models. Our approach is detailed in the block diagram shown in Fig. 1. Besides this, the rest of the paper is organized as follows. Section II presents the related works, Section III shows the proposed model, Section IV elaborates the results and discussion, and Section V concludes the paper with discussion on the future direction of the work.

II. Literature Review

Starting with the basic statistical analysis, many models have been proposed to use the regression models, correlation analysis, econometric models, parametric tests and non-parametric tests these models are usually used to identify the delay propagation through the network and to estimate the cost of delay [1,2]. Xiong et al [3] built an econometric model based on preexisting delays, potential delay savings, distance, characteristics of the destination airport and airline, frequency, aircraft size, occupancy rate and fare to understand which reasons lead airlines to cancel their flights. Hao et al [4] built a model to quantify how delays originating at New York are propagated to other airports. Finally, Abdel-Aty et al [5] calculated the daily average of delays to detect correlations to understand the principal causes of delays at Orlando International Airport.

III. The Proposed Model

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

1) If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of Y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of Y, the fitted model can be used to make a prediction of the value of Y.
2) Given a variable Y and a number of variables X_i and X_p that may be related to Y, linear regression analysis can be applied to quantify the strength of the relationship between Y and the X_j, to assess which X_j may have no relationship with Y at all, and to identify which subsets of the X_j contain redundant information about Y.

The decisions taken in the management of an airport are often based on common sense and influence several variables, such as flight delay. Reducing this delay presents the advantage of decreasing costs and increasing the quality of the service provided to the passengers. It is thus important to find which variables influence flight delay and use them to predict it. In this context, there are many studies. Some of them treat flight delay prediction as a regression problem, predicting the delay by the minute, and others as a classification problem, predicting a time interval where the delay will fall. The problem here considered is to predict flight arrival delay at a given airport.

Given information about a flight that will depart from this airport, the main objective in this scope is to predict its arrival delay by the minute. If the delay time falls into a predefined range, it indicates that there is no delay in the flight. If the delay time falls above the range it indicates the flight delay. So, two types of prediction mechanisms are considered: regression, where the continuous output is an estimate of the arrival delay, and classification, where the output is a binary prediction of whether the arrival delay is more or less than the predefined threshold.

Najeebuddin Ahmed and Khanjan Dabhi are with the Department of Software Engineering, Lakehead University, Thunder Bay, ON, Canada. (e-mail: {nahmed,kdabhi}@lakeheadu.ca).

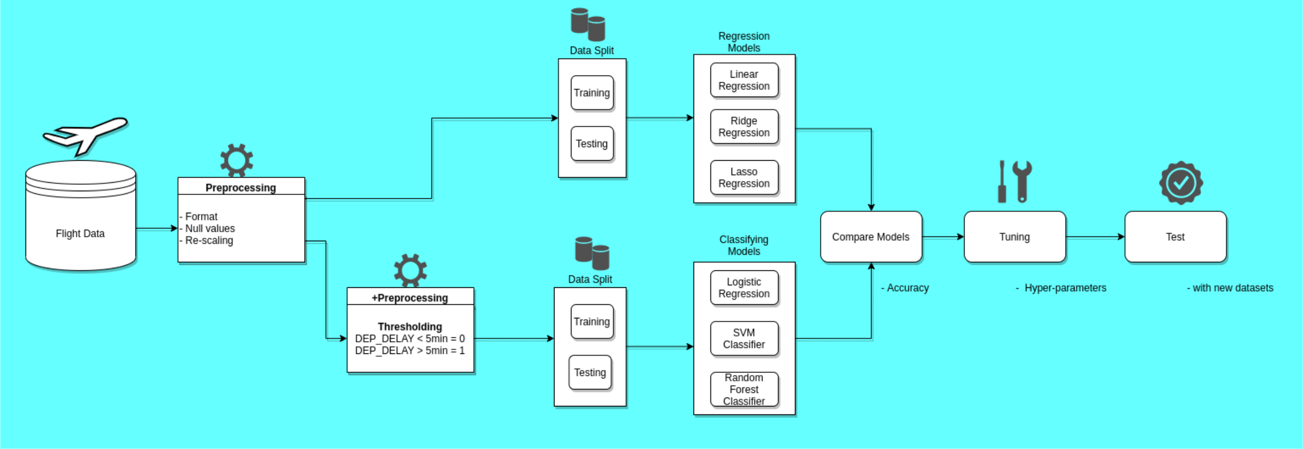

Fig. 1: The Proposed Approach.

It is important to find which variables influence flight delay and use them to predict it. After careful analysis of the data, we found that there is a close relation between arrival delay and departure delay. So, we can use departure delay to predict arrival delay. Using the formula as given by Eq. 1.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 \tag{1}$$

We establish a multiple linear regression model , where the target variable Y is arrival delay, the predictor variables X_1 is departure delay and X_2 is route distance.

At first, we add a field of departure delay subtracting the planned departure time with the actual departure time. In the training phase, we use departure delay and route distance training model in order to learn the three parameters. In the predicting phase, given information about a flight that will depart from one airport, route distance is already known. If we know the departure delay of the flight, we could use the model to predict its arrival delay, so as to determine whether the flight delay. Given information about a flight, at first, we look for the similar data according to the departure airport, the aircraft type and the weather of departure airport in the dataset, and then we cluster the similar data using departure delay and set up a threshold. If the amount of data exceeds the threshold, we calculate the average of departure delay, and use it as the departure delay of the given flight.

After removing the outliers we ran the regression models again and we found out that the linear regression is the second best in terms of errors compared to the other regression.

IV. Experimental Analysis

A. Dataset

i. The original dataset contained 28 attributes out of which 3 were integers, 5 were strings, and 20 were floats. There were 5819079 entries. It contained dates, places, times and multiple types of delays.

ii. First and foremost, we made sure that the attributes had a general format and scale. Unwanted attributes such as ORIGIN and DEST, as well as all the attributes which had more than 80 percent of their entries missing such as the cancelled flights and diverted flights were removed. One of the main problems while getting the dataset ready for the model was that the time formats in the original dataset, they were not properly formatted. To overcome this we had to create a function which would take the dates which were in float format and convert them into proper datetime format. Initially, for the final dataset after cleaning and preprocessing and using the correlation heatmap for the best attributes, we have:

a) DEP_DELAY: Datetime format calculated as a difference between scheduled and actual delay

b) ARR_DELAY: Datetime format calculated as a difference between scheduled and actual delay

c) SCHEDULED_ELAPSED_TIME: Datetime format which show the elapsed time for the flight taking the scheduled time into consideration

d) ACTUAL_ELAPSED_TIME: Datetime format which show the elapsed time for the flight that happened taking the delays into consideration

e) DISTANCE: An int datatype which shows total distance travelled in given elapsed time.

The following 2 features were later added after using Random Forest Feature Selection:

a) TAXI_IN: A float datatype that shows the time taken for the plane to reach the gate from the tarmac.

b) TAXI_OUT: A float datatype that shows the time taken for the plane to reach the tarmac from the gate.

The correlation Heatmap with the features and the predict attribute can be seen in Fig. 2.

B. Experimental Setup

In order to perform the project, the following were the specifications of the software environment:

i. Operating System: Windows 10 / Arch Linux
Development Language: Python3
Development Tool: Google Colab
Hardware: Python 3 Google Compute Backend

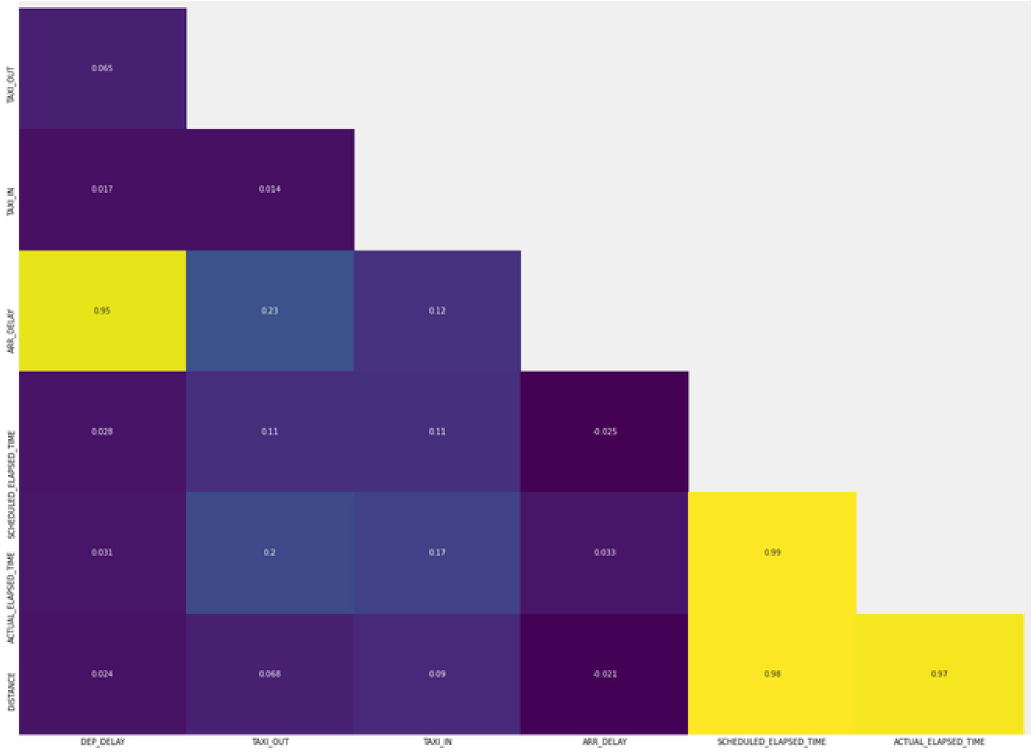

Fig. 2: Correlation Heatmap with selected features and predict value.

RAM: 12.72GB
Disk: 107.77GB

ii. The initial testing with regression models was done without removing the outliers and also all the numeric features of the dataset were used. However, the later regression models were done with removal of outliers. Various trials were done taking features which were least correlated to highly correlated into consideration for better output of the regression model.

C. *Results and Discussion*

We compared Linear Regression, Ridge Regression, Lasso Regression, Logistic Regression, SVM and Random Forest Classification.

Models	Testing Score	RMSE	MAE
Linear Regression	0.499	0.313	0.207
Ridge Regression	0.323	0.364	0.307
Lasso Regression	-2.67	0.442	0.391

TABLE I: Performance Comparison: Linear Regression vs Ridge Regression vs Lasso Regression w.r.t. Testing score, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)

As we can see from the Table I regression models did not produce satisfactory results. They had very low testing scores. To get better results we used the other regression models. Initially, running Logistic Regression on our model with 4 features and 1623633 data points, gives us an accuracy of around 0.6918. The stratified k fold cross validation was done with 5 folds. Also, initially running SVM gave much low accuracy of around 0.48 which is not good enough. Additionally, Random Forest Classifier received an accuracy of 0.69. i.e. the prediction of aircraft with delay is about 0.69 times. However, this accuracy will improve after some fine-tuning.

After adding two more features into the models namely ’TAXI_IN’ and ’TAXI_OUT’ and increasing the dataset to 6000000 data-points the accuracy for each model was greatly improved(Table II). For the Random Forest Classifier we also did rigorous hyper parameter tuning using RandomizedSearchCV() and GridSearchCV() to get the optimal hyperparameters. Doing so resulted in an accuracy of 0.862015 for Random Forest Classifier. The table shows new accuracy after making those changes.

Comparing the accuracy of the three methods accuracy rates(table II), Random Forest Classifier proves to be the leading model. Also, the confusion matrices of Logistic Regression (Table III), SVM (Table IV), and Random Forest (Table V) are given below.

Model	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**
LR	85.72%.	75.40%.	88.90%.	87.16%.
SVM	71.77%.	7.90%.	93.10%.	53.37%.
RF	86.20%.	75.80%.	86.50%.	84.88%.

TABLE II: Performance Comparison: Logistic Regression (LR) vs SVM Classifier vs Random Forest Classifier (RF) w.r.t. Accuracy, Sensitivity and Positive Predictive Value (PPV)

True Negative	False Positive
88.9%.	11.1%.
False Negative	True Positive
24.6%.	75.4%.

TABLE III: Logistic Regression results w.r.t. Confusion Matrix

True Negative	False Positive
93.1%.	6.9%.
False Negative	True Positive
92.1%.	7.9%.

TABLE IV: SVM results w.r.t. Confusion Matrix

True Negative	False Positive
86.5%.	13.5%.
False Negative	True Positive
24.2%.	75.8%.

TABLE V: Random Forest Classifier results w.r.t. Confusion Matrix

If there were more influencing factors such as weather and wind direction our model would have potentially improved its accuracy. These types of impact changing factors are discussed in “Predicting Airline Delays”[11] and “Multi-Factor Model for Predicting Delay at U.S. Airports”[10].

Threshold of 5 minutes was defined for DEP_DELAY and ARR_DELAY attributes. If the DEP_DELAY was less than or equal to 5 minutes then it was considered ‘0’ and if it was more than 5 minutes it was considered as ‘1’. Doing so resulted in making our problem a classification problem.

However, after performing the operation it was seen that there was a vast difference between no. of attributes that were ‘0’ and the number of attributes that were ‘1’. Thus, to overcome this skewed attribute problem we selected only an equal number of ‘1’ and ‘0’ attributes as it has in the main data-set with the same ration(75:25) and merged into one data-set and used that to split our data-set into training and testing. Thus eliminating the skewed data-set problem.

Our model can also be used to make predictions for specific kinds of delays. For example there are different types of delays like weather delay, delay due to incoming air traffic. We can also run these classification models for predicting those delays.

Overfitting happened for SVM and was resolved by tuning hyperparameter MAX_Iteration. At first the Max_Iteration was set to default which generated very low accuracy. After which there increments were made in 50 steps which resulted in overfitting at Max_Iteration = 150. However, at lower iteration levels the SVM generated better True Negative and True Positives.

V. Conclusion

This paper proposes an airline-delay prediction model using random forest classifier. From the examination performed on the data-set, we can presume that the factors which were incorporated in airline delay are ultimately affected via air-traffic. Thus, if there is a shift in air-traffic, delay times would be influenced and be changed. Airline Carriers have most traffic when they have high-recurrence of flights.

The accuracy of 86%. for the model was achieved by optimizing the hyper parameters for Random Forest Classifier, using randomized search to get an approximate of best parameters. Afterwards we get that we an exhaustive grid search which then tries every single possible combination of the hyper parameters. The model parameters that were used to achieve the accuracy are bootstrap=True, max_.depth = 10, max_.features= 2, min_.samples_.leaf= 4, min_.samples_.split= 2, n_.estimators= 400. A disadvantage for using this is the trade off with time with every additional parameters and cross-validation that we add. The functions that we ran took a long amount of time to compute since it was a large dataset.

Acknowledgements

This work acknowledges The U.S. Department of Transportation’s (DOT) Bureau of Transportation Statistics, the organizer of the Airline Data-set and Dr. Thangarajah Akilan for the technical comments and suggestions.

References

[1] Beatty, R. Hsu, L. Berry, &. J. Rome, “Preliminary evaluation of flight delay propagation through an airline schedule. 2nd USA/Europe Air Traffic Management R&.D Seminar,” 7(4):259–270, 1998.

[2] D. Markovic, T. Hauf, P. Rohner, &. U. Spehr.,”A statistical study of the weather impact ¨ on punctuality at Frankfurt airport. Meteorological Applications,”15(2):293–303, 2008.

[3] J. Xiong &. M. Hansen,”Modelling airline flight cancellation decisions. Transportation Research Part E: Logistics &.Transportation Review,” 56(Supplement C):64–80, Sept. 2013.

[4] L. Hao, M. Hansen, Y. Zhang, &. J. Post,”New York, New York: Two ways of estimating the delay impact of New York airports. Transportation Research Part E: Logistics and Transportation Review”,70(Supplement C):245–260, Oct. 2014.

[5] M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak, ”Detecting periodic patterns of arrival delay. Journal of Air Transport Management”, 13(6):355–361, Nov. 2007.

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

[7] T. Kotegawa, D. De Laurentis, K. Noonan, and J. Post. ”Impact of commercial airline network evolution on the U.S. air transportation system. In Proceedings of the 9th USA/Europe Air Traffic Management Research and Development Seminar”, ATM 2011, pages 572–580, 2011.

[8] S. AhmadBeigi, A. Cohn, Y. Guan, and P. Belobaba,” Analysis of the potential for delay propagation in passenger airline networks. Journal of Air Transport Management”, 14(5):221–236, Sept. 2008.

[9] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen, ”Estimation of delay propagation in the national aviation system using Bayesian networks. In 6th USA/Europe Air Traffic Management Research and Development Seminar”, Citeseer, 2005.

[10] Xu, Sherry, & Laskey, 2008, ”Multi-Factor Model for Predicting Delays at U.S. Airports”, ”http://catsr.ite.gmu.edu/pubs/XuMultiFactorModelAirpor tDelays TRBv6.pdf”

[11] Bandyopadhyay &. Guerrero, 2012,”Predicting Airline Delays”, http://cs229.stanford.edu/proj2012/BandyopadhyayGuerrero-PredictingFlightDelaysṗdf

[12] W.-W. Wu, T.-T. Meng, and H.-Y. Zhang, ”Flight plan optimization based on airport delay prediction. Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology”, 16(6):189–195, 2016.

[13] J. J. Rebollo and H. Balakrishnan,” Characterization and prediction of air traffic delays. Transportation Research Part C: Emerging Technologies”, 44(Supplement C):231–241, July 2014.

[14] P. Balakrishna, R. Ganesan, and L. Sherry,” 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, Dec. 2010.

[15] P. Balakrishna, R. Ganesan, L. Sherry, and B. S. Levy,”Estimating Taxi-out times with a reinforcement learning algorithm. In 2008 IEEE/AIAA 27th Digital Avionics Systems Conference”, pages 3.D.3–1–3.D.3–12, Oct. 2008.

[16] L. Zonglei, W. Jiandong, and Z. Guansheng, ”A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. In 2008 International Symposium on Knowledge Acquisition and Modeling”, pages 589–592, Dec. 2008