Journal of Current Research in Engineering and Science
JCRES
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 6 –Issue 2,August 2023
Paper:66

PREDICTING FLIGHT DELAYS WITH ERROR CALCULATION USING MACHINE LEARNEDCLASSIFIERS

Dr. K Lakshmi[1], K Arshad[2], O Deva Harsha[3], N Jabbar Khan[4], K Jaswanth[5]
drlakshmik@mits.ac.in
[1,2,3,4,5] Department of Computer Science and Engineering, Madanapalle Institute of Technology &Science(Affiliated by JNTUA), Andhra Pradesh, India

Abstract— Accurate flight delay prediction is fundamental to establish the more efficient airline business Increasing client happiness is a key component of the airline company. Participants in all capacities in commercial aviation must consider their prediction while making decisions. Due to bad weather, a mechanical problem, and the delayed arrival of the aircraft at the point of departure, flights are delayed and cause customer annoyance. Using flight data and weather data, a prediction model of on-time arrival flights is developed. In this paper, using machine learning models such as Decision Tree Regression, Bayesian Ridge, Using Random Forest Regression and Gradient Boosting Regression, we make predictions about whether a specific flight's arrival will be on time or not.

INTRODUCTION

When an airline lands or goes once its scheduled arrival or departure time, respectively, it is said have cell a flight delay. In recent years, AIR traffic load has escalated remarkably. Airline delays are one of many causes that cause substantial expenses for the aviation industry worldwide. Airports, airlines, and passengers all experience hardships as a result of aircraft delays. To reduce losses and boost customer satisfaction, precise and cautious airline delay prediction thinking all factors which have a significant impact is essential. Figure 1. Flight Delays and Cancellation Data in the United States From 2012to 2021[2]

When a flight's scheduled arrival time and actual arrival time differ by more than 15 minutes, the FAA in the United States views the flight as delayed.Since itbecomes a serious problem in the United States, analysis and prediction of flight delays are being studied to reduce large costs. Unfavourable weatherconditions, air traffic jams, late arriving aircraft from previous flights, repairs, and safety issues are notable reasons for delay for commercially scheduled flights.

II .RELATED WORK

[1] c. (2019). A A data mining strategy fo anticipating flight arrival delay for America Airlines.

There have been tonnes of instances concerning fligh delays and cancellations in current domestic airlin scenario in the USA. One of the most renowne airlines in the US and the largest airline in the world i terms of destinations served is American Airlines, In nevertheless AA has fallenshy of expectations when comes to punctuality or performance when goin domestic. Flight delays additionally lead airline firm that run commercial flights to suffer significant losse So as to prevent or avoid flight delays an cancellations, they thus put forth every effort. Th intended use of this study is for analysing fligh information for domestic flights in the United State that are The aim of this study is to summarized fligh logs for domestic American Airlines flights, coverin America's five busiest airports, and predict a nee arrival tardiness.employing methods from data minin and machine learning, flying. The Gradient Boosting Classifier

Model is executed with a maximum accuracy of 85.73%, after training and hyper-parameter adjustment. In order to predict flight on-time performance, such an intelligent system is very necessary.

2] G. Gui and F. Liu, "Aviation Big Data and

Machine Learning-Based Flight Delay Prediction," J. Sun, J. Yang, Z. Zhou, and D. Zhao, was published in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, in January 2020, pp. 140–150.

Predicting cancelled flights consistently serves as crucial to developing a more effective airline industry. Recent research has concentrated on using machine learning techniques to forecast aircraft delays. The majority of the earlier prediction techniques are limited to one route or airport. This study covers a greater number of things the could potentially affect To build a dataset for the proposed scheme, automatic dependent surveillance-broadcast (ADS-B) messages are received, pre-processed, connected with additional data like weather, travel information, and airport schedules. Different classification tasks and a regression task are included in the intended prediction tasks. Long short-term memory (LSTM) is capable of managing the acquired aviation sequence data, according to experimental results, although overfitting issues arise in our small dataset. The suggested random forest-based model can address the overfitting issueand possibly achieve higher prediction accuracy (90.2% for the binary classification) than the preceding systems.

[3] Himani Sharma and Sunil Kumar (2016). a review of categorization decision tree methods used in data mining. The International Journal of Science and Research is referred to as IJSR.

The sheer value of data in the information the industry is increasing easily as computer and computer network technology increase.. This massive volume of data needs to be analyses in order to gain relevant knowledge. Data mining is the process of obtaining valuable information from a large collection of imperfect, noisy, imprecise, and unpredictable data. One of the most used data mining approaches is the decision tree classification method. Divide and conquer is a basic learning mechanism used in decision trees. Branches, a rootnode, and leaf nodes make up the geometrical structure of a decision tree. Each leaf node contains a class label, eac internal node represents a test on an property, and eac branch stands for a test's outcome. The root node is th topmost node in the tree., branches, and leaf nodes .Eac leaf node contains a class label, each internal noc represents a test on an property, and each branch stanc for a test's outcome. The root node of a tree is the node a the top. This essay focuses on the several decision-tree algorithms (ID3, C4.5, CART), as well as their traits, difficulties, benefits, and drawbacks. **Summary:** In this paper, we learn about Decision Tree, types of Decision tree (ID3, C4.5, CART etc). It also discusses about the advantages and disadvantages of Decision Tree.

[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

using sequentially fitting a basic parameterized function (base learner) to the present "pseudo"-residuals" using least squares at each iteration, gradient boosting creates additive regression models. The gradient of the loss functional that is being minimised with respect to the model values at each training data point assessed at this stage is the pseudo-residuals. It is demonstrated that by adding randomness to the process, gradient boosting's approximation accuracy and execution speed may both be significantly increased. In more detail, a subsample of the training data is randomly selected (without replacement) from the entire training data set at each cycle. The base learner is then fitted to this randomly chosen subsample, and the model update for the current iteration is computed instead of using the entire sample. This placed the plan also enhances robustness against the base learner's overcapacity.

Volume 6 –Issue 2,August 2023
Paper:66

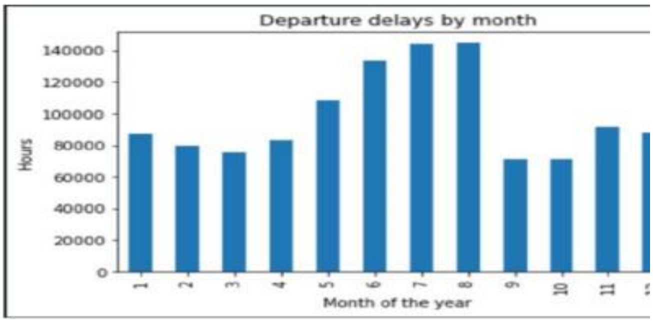


Figure 2. Departure delays by month

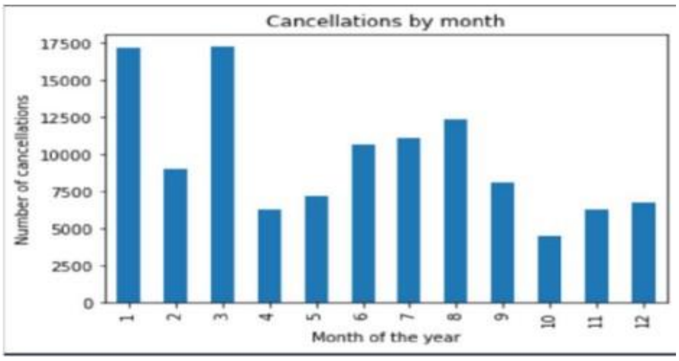


Figure 3. Cancellations by month

III. METHODOLOGY

Algorithms that allow a computer to analyse data, find probable patterns, and make predictions are referred to as machine learning. Learning algorithms can shed lighton how challenging learning is in various contexts [8]. The two most popular types of machine learning algorithms are supervised learning and unsupervised learning, which are classified into various groups. A function was created by supervised learning algorithms that converts inputs into desired outputs. Regression andclassification are the two main types of supervised learning algorithms. Absent labelled examples, unsupervised learning can model a set of inputs.

3.1

odels for Classification

In this study, seven algorithms were used to choose and train classification models. Decision Tree, Gaussian Naive Bayes, K-Nearest Neighbour (KNN), Logistic Regression, Support Vector Machine (SVM), GradientBoosted Tree and Random Forest. Due to the fact thatonly one classifier instance is trained for each of the first five algorithms, they are referred to as basis classifiers. Because more than one instance of the basic classifiers is trained, and their combined judgement is provided as the final prediction, the last two techniques are known as ensemble classifiers [5]. Since Random Forest and Gradient Boosted Tree, two of the most well-liked ensemble algorithms, combine various independent models to enhance performance by increasing accuracy and decreasing variation.

This section aims to describe a method by which we attempted to address issues with large data sets and processing challenges for flight delay prediction in such case we use algorithms like

1 decision tree alogoritgm
2 Bayesian Ridge Regression
3 Random Forest Regression:
4 Gradient Boosting regression
5 Support vector regression

Both the bagging technique and distinct decision tree models are applied. Each subset of the trained data has its own decision tree, and the subsets are chosen at random. All trees in the forest are given the data in parallel, and the class that most trees predicted has new information

The decision trees that make up Random Forest choose the appropriate attribute for each node starting at the root and divide the data into subsets according to the chosen attribute. Both the bagging technique and distinct decision tree models are applied. Each subset of the trained data has its own decision tree, and the subsets are chosen at random.All trees in the forest are given the data in parallel, and the class that most trees predicted has the new data [7]. Boosted gradient The initial prediction foreach training set is represented by a single decisiontree at the beginning of the tree. Since it employs aboosting technique, each model is trained in turn. The forecast of a tree is assessed based on its residual errors. Each tree model thus gains knowledge from the errors committed by the previous model. When a new tree cannot make the prediction more accurate, construction of new treeswill cease. A single root node tree is used to presentthe data [7].

3.2 Methods of Evaluation

The confusion matrix can be used to determine how well the classifier performs. The classifier results can produce one of four values after being compared to theactual result:

- True Positive (TP): both the expected and actualvalues are positive.
- True Negative (TN): When a value is projected tobe negative but actually turns out to be negative.
- False Positive (FP): the actual value is negative yetthe projected value is positive.
- Negative (FN): the predicted value is negative; theactual value is positive.
-

ccuracy, precision, recall, and F1 score were the four metrics that were utilized to assess the performance of the chosen algorithms. The effectiveness of algorithms is positively correlated with each of these metrics. As a result, the better these measurements perform for a certain algorithm, the higher their values are. Calculations employing the following parameters can be used to get the values of the four measures:

$$F1-Score=2\times Recall\times PrecisionRecall+Precision$$

Decision tree

The primary concept behind the decision tree algorithm is to create a tree-like structure and obtain answers in the form of true or false. The decision marks the end of the model, which starts at the root node. Each node is given a Yes/No question, and the response is then forwarded to the following node. The training dataset is all input to the root node. The difficulty in building such a tree is determining when to ask a query at each node. To do this, a decision tree algorithmic programme measures an uncertainty or impurity associated to an explicit node using recognised indices like entropy or Gini- impurity. Equations (1) and (2) demonstrate how to calculate entropy and Gini impurity separately for a SVM regression

Volume 6 –Issue 2,August 2023
Paper:66

Featured snippet from the web

To predict discrete values, support vector regression, a method for supervised learning, is utilized. The SVMs and Support Vector Regression both operate on the same theory. The cornerstone of SVR is locating the best fit line. The SVR best- fitting line is the hyperplane with the most points.

DISCUSSION

The ultimate goal of this study is flight predictions of the flights of the current date that is entering the flight

information into the system, such as flight number, flight carrier, destination, etc. and the system will tell the user how long the delay of the flight will be and whether the flight will likely be canceled or not. The scope of future work involves the application of more advanced and novel pre- processing techniques and deeplearning or unsupervised learning models. With these models, it will be possible to reach a higher accuracy with fewer data, increasing the efficiency of the prediction[8]. Furthermore, a dataset with flight information from2020 could be used, which could also help to increase accuracy when predicting flight delay and cancellations of the current date. However, when severe natural disasters take place, the model might not be so accurate. For example, during the COVID-19 period, most of flights are canceled and delayed. In this case, a smaller model might be more accurate. A dataset with the flight dataof the past week or past month could.

FUTURE SCOPE

We should consider flight delay prediction using boosting techniques like Boost which involves extreme gradient boosting. We can also simulate complicated neural networks that provide improved accuracy and automated feature selection.

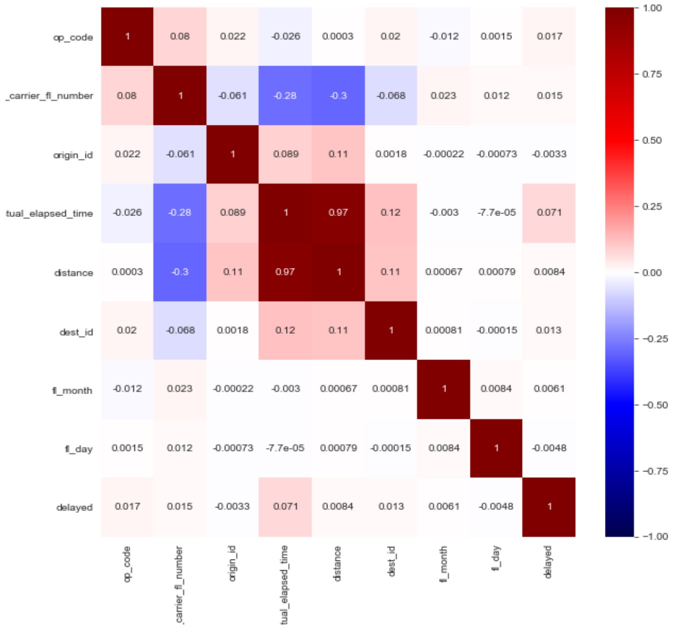


Figure 4: Heatmap

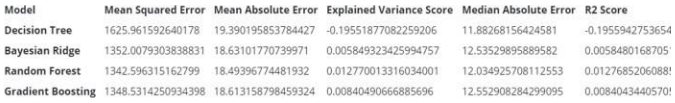
V. CONCLUSION

In this application, we have pre processed the data by removing the null values and encoding all the variables. we have also scaled all the predictor variables.

We have used Decision Tree, Bayesian Ridge, Random Forest and Gradient Boosting regression.

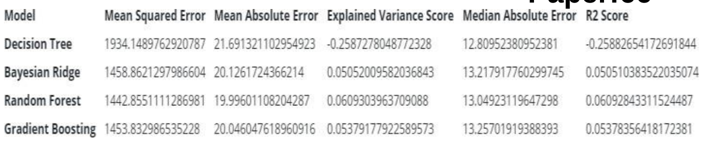
The best model was the Random Forest (by a small margin) model with Hyper parameters tuning. The dataset used was the 2015 FAA Flight dataset.

The performance for all models are shown below:
For Departure Delay as Target variable:



Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1625.961592640178	19.390195853784427	-0.19551877082259206	11.88268156424581	-0.1955942753654
Bayesian Ridge	1352.0079303838831	18.63101770739971	0.005849323425994757	12.53529895889582	0.0058480168705
Random Forest	1342.596315162799	18.49396774481932	0.012770013316034001	12.034925708112553	0.0127685206088
Gradient Boosting	1348.5314250934398	18.613158798459324	0.00840490666885696	12.552908284299095	0.0084043440570

For Arrival Delay as Target variable:



Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1934.1489762920787	21.691321102954923	-0.2587278048772328	12.80952380952381	-0.25882654172691844
Bayesian Ridge	1458.8621297986604	20.1261724366214	0.05052009582036843	13.217917760299745	0.050510383522035074
Random Forest	1442.855111286981	19.99601108204287	0.0609303963709088	13.04923119647298	0.06092843311524487
Gradient Boosting	1453.832986535228	20.046047618960916	0.05379177922589573	13.25701919388393	0.05378356418172381

VI. REFERENCES

[1] Chakrabarty, Navoneel. (2019). An approach using data mining to predict flight arrival delay for American Airlines.

[2] "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," IEEE Transactions on Vehicular Technology, vol. 69, no. 1, Jan. 2020, pp. 140–150. G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao.

[3] Sunil Kumar and Himani Sharma. (2016). A survey of data mining's classification decision tree algorithms. The International Journal of Science and Research is referred to as IJSR.

[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

[5] N. G. Rupp, at the Department of Economics at East Carolina University, "Further Investigation into the Causes of Flight Delays," 2007.

[6] In "Emerging Technologies in Data Mining and Information Security," Singapore, 2019, Navoneel et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier,"

[7] Area K. Leboulluec and A. M. Kalliguddi's article, "Predictive Modelling of Aircraft Flight Delay," appeared in the Universal Journal of Management in 2017. Pages 485–490.

[8] "Development of an airliner on-time arrival predictive model by determining the correlation between flight and weather data," By Etani Noriko in 2019.

[9] Navoneel Chakrabarty, "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines." IEMECON, the 9th annual conference on information technology, electromechanical engineering, and microelectronics, was held in 2019.

[10] Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. (2017). An analysis of flight delay forecasting.

[11] "Iterative machine and deep learning approach for aviation delay prediction," 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, pp., by V. Venkatesh, A. Arya, P. Agarwal, S. Lakshmi, and S. Balana. 562-567, doi: 10.1109/UPCON.2017.8251111.

[12] Yogita Borse , Dhruv in Jain , Shreyas Sharma , Viral Vora, Aakash Zaveri, 2020, Flight Delay Prediction System, International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 03 (March 2020).