International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Novel Approach for Flight Delay Prediction Using AI

T. Vasanth Kumar Reddy[1], Dr. Srinivasan Jagannathan[2], Mr. Suresh[3]
(PG Scholar)[1,2] (Assistant Professor)[2] (Professor)[3]
Department of Computer Applications, Madanapalle Institute of Technology and Science, India

Article Info

Publication Issue :
Volume 8, Issue 4
July-August-2022

Page Number : 260-265

Article History
Accepted: 10 August 2022
Published: 28 August 2022

ABSTRACT

Predicting flight delays accurately is essential for building a more effective airline industry. Increasing client happiness is a key component of the airline company. All participants in commercial aviation must consider their prediction while making decisions. Flights are delayed and cause consumer displeasure due to inclement weather, a mechanical issue, and the delayed arrival of the aircraft at the place of departure. With the aid of weather and flight data, a predictive model for flights arriving on time is put forth. In this study, we forecast whether a specific flight's arrival will be delayed or not using machine learning models such Decision Tree Regression, Bayesian Ridge, Random Forest Regression, and Gradient Boosting Regression.

Keywords: Decision Tree Regression, Bayesian Ridge, Random Forest Regression, and Gradient Boosting Regression.

I. INTRODUCTION

A flight delay is said to occur when an airline lands or takes off later than its scheduled arrival or departure time respectively. AIR traffic load has experienced rapid growth in recent years. The aviation industry around the globe incur huge losses due to various factors, one of these factors is Airline Delay. Airline delay tends to be onerous for every entity involved i.e. airports, airlines and passengers. Precise and meticulous prediction of Airline delay using the factors which play prodigious role will be the key to minimize the losses and increase customer satisfaction. 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. Notable reasons for commercially scheduled flights to delay are adverse weather conditions, air traffic congestion, late reaching aircraft to be used for the flight from previous flight, maintenance and security issues.

In the paper, several machine learning algorithms have been employed to produce a comparative study with respect to the accuracy of each algorithm. we are using machine learning models such as 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.

II. RELATED WORKS

[1] Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.

In the present scenario of domestic flights in USA, there have been numerous instances of flight delays and cancellations. In the United States, the American Airlines, Inc. have been one of the most entrusted and the world's largest airline in terms of number of destinations served. But when it comes to domestic flights, AA has not lived up to the expectations in terms of punctuality or on-time performance. Flight Delays also result in airline companies operating commercial flights to incur huge losses. So, they are trying their best to prevent or avoid Flight Delays and Cancellations by taking certain measures. This study aims at analyzing flight information of US domestic flights operated by American Airlines, covering top 5 busiest airports of US and predicting possible arrival delay of the flight using Data Mining and Machine Learning Approaches. The Gradient Boosting Classifier Model is deployed by training and hyper-parameter tuning it, achieving a maximum accuracy of 85.73%. Such an Intelligent System is very essential in foretelling flights'on-time performance.

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

Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. To build a dataset for the proposed scheme, automatic dependent surveillance-broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as weather condition, flight schedule, and airport information. The designed prediction tasks contain different classification tasks and a regression task. Experimental results show that long short-term memory (LSTM) is capable of handling the obtained aviation sequence data, but overfitting problem occurs in our limited dataset. Compared with the previous schemes, the proposed random forest-based model can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.

[3] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher. It is necessary to analyze this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. Decision tree classification technique is one of the most popular data mining techniques. In decision tree divide and conquer technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. This paper focus on the various algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges, advantage and disadvantage.

III. Methodology

Proposed system:

Accurate flight delay prediction is fundamental to establish the more efficient airline business. An important business of airlines is to get customer satisfaction. The existing methods requires highly skilled people and hence is costly to implement as it requires manually selecting features for prediction. In this paper, using machine learning models such as 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.

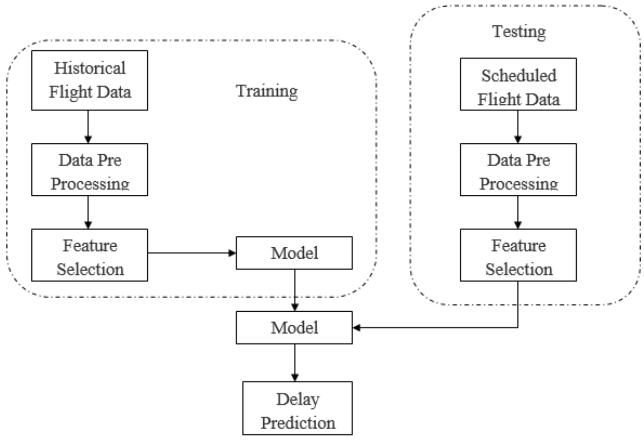


Figure 1: Block diagram

IV. Implementation

The algorithms listed below were used to complete the project.

Decision Tree:

Decision trees are non-parametric supervised learning Method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Bayesian Ridge Regression:

Ridge Regression is the name usually given to Linear Regression with an L2 regularizer. The regularizer penalizes model complexity by adding the sum of the parameter squares to the error function. You can get there using Maximum Likelihood estimation on a Gaussian likelihood model and then applying the rationale of structural risk minimization (think of an SVM).

From the Bayesian side of things, if you start with your Gaussian likelihood model, a Gaussian prior on the model parameters with mean zero and standard deviation 1, and then apply the Bayes Rule to find the posterior distribution of the model parameters given your dataset, you will find that said posterior distribution is also a Gaussian whose mean is equivalent to the Ridge Regression estimate of the model coefficients. This is known as the Maximum A Posteriori estimate of the regression model.

Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. One of the most useful type of Bayesian regression is Bayesian Ridge regression which estimates a probabilistic model of the regression problem.

Random Forest Regression:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low.

In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. A Random Forest is an ensemble technique capable of performing both regression and classification tasks.

Gradient Boosting Regression:

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

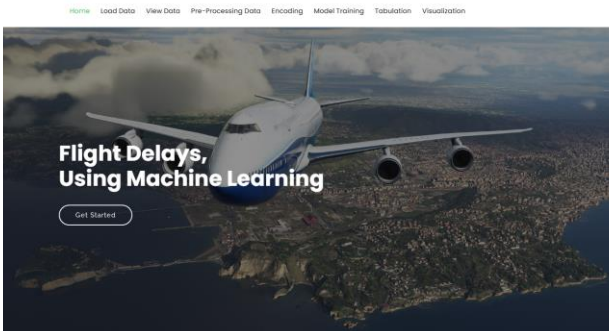
The idea of gradient boosting originated in the observation that boosting can be interpreted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed simultaneously with the more general functional gradient boosting.

The boosting can be viewed as iterative functional gradient descent algorithms. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction. This functional gradient view of boosting has led to the development of boosting algorithms in many areas of machine learning and statistics beyond regression and classification.

V. Results and Discussion

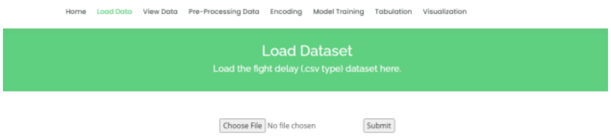
The following screenshots are depicted the flow and working process of project.

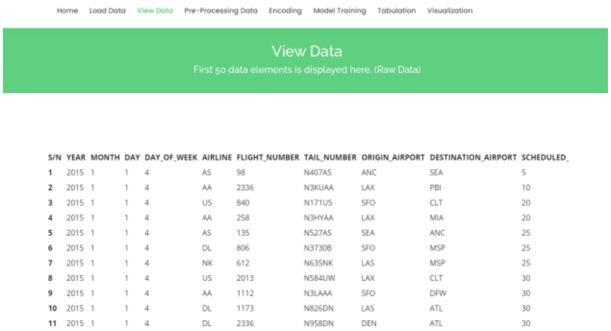
Home Page: Here user view the home page of A Machine Learning Methodology for flight delay web appellation.



About page:

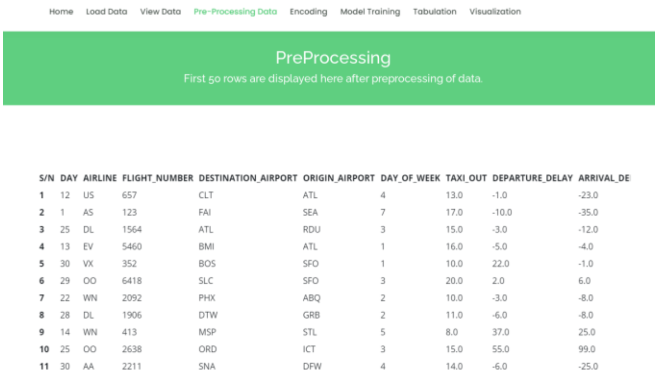
In the about page, users can learn more about loading the dataset and view dataset.





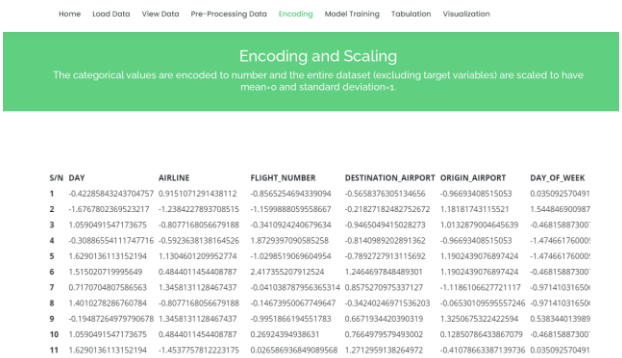
S/N	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_
1	2015	1	1	4	AS	98	N407AS	ANC	SEA	5
2	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10
3	2015	1	1	4	US	840	N171US	SFO	CLT	20
4	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20
5	2015	1	1	4	AS	135	N527AS	SEA	ANC	25
6	2015	1	1	4	DL	806	N3730B	SFO	MSP	25
7	2015	1	1	4	NK	612	N635NK	LAS	MSP	25
8	2015	1	1	4	US	2013	N584UW	LAX	CLT	30
9	2015	1	1	4	AA	1112	N3LAAA	SFO	DFW	30
10	2015	1	1	4	DL	1173	N826DN	LAS	ATL	30
11	2015	1	1	4	DL	2336	N958DN	DEN	ATL	30

Preprocessing Page



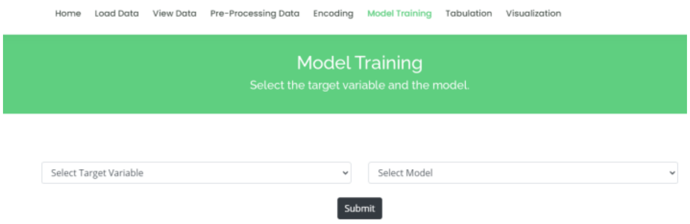
S/N	DAY	AIRLINE	FLIGHT_NUMBER	DESTINATION_AIRPORT	ORIGIN_AIRPORT	DAY_OF_WEEK	TAXI_OUT	DEPARTURE_DELAY	ARRIVAL_DE
1	12	US	657	CLT	ATL	4	13.0	-1.0	-23.0
2	1	AS	123	FAI	SEA	7	17.0	-10.0	-35.0
3	25	DL	1564	ATL	RDU	3	15.0	-3.0	-12.0
4	13	EV	5460	BMI	ATL	1	16.0	-5.0	-4.0
5	30	VX	352	BOS	SFO	1	10.0	22.0	-1.0
6	29	OO	6418	SLC	SFO	3	20.0	2.0	6.0
7	22	WN	2092	PHX	ABQ	2	10.0	-3.0	-8.0
8	28	DL	1906	DTW	GRB	2	11.0	-6.0	-8.0
9	14	WN	413	MSP	STL	5	8.0	37.0	25.0
10	25	OO	2638	ORD	ICT	3	15.0	55.0	99.0
11	30	AA	2211	SNA	DFW	4	14.0	-6.0	-25.0

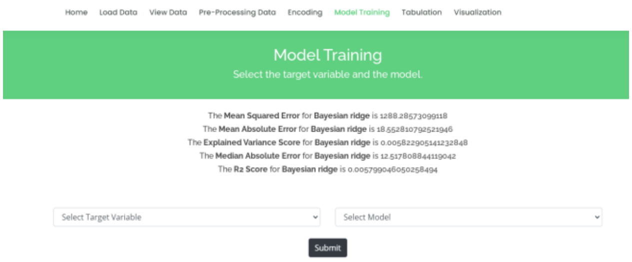
Encoding Page:



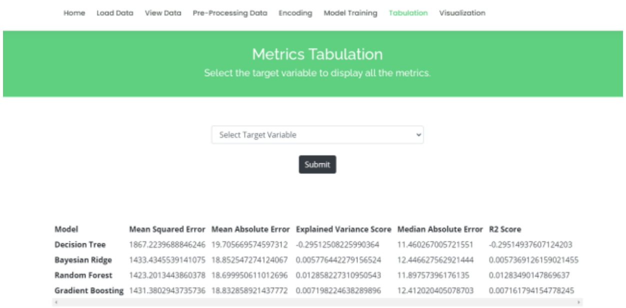
S/N	DAY	AIRLINE	FLIGHT_NUMBER	DESTINATION_AIRPORT	ORIGIN_AIRPORT	DAY_OF_WEEK
1	-0.42285843243704757	0.9151071291438112	-0.8565254694339094	-0.5658376305134656	-0.96693408515053	0.035092570491
2	-1.6767802369523217	-1.2384227893708515	-1.1599888059558667	-0.21827182482752672	1.18181743115521	1.544846900987
3	1.0590491547173675	-0.8077168056679188	-0.3410924240679634	-0.9465049415028273	1.0132879004645639	-0.46815887300
4	-0.30886554111747716	-0.5923638138164526	1.8729397090585258	-0.8140989202891362	-0.96693408515053	-1.47466176000
5	1.6290136113152194	1.1304601209952774	-1.0298519069604954	-0.7892727913115692	1.1902439076897424	-1.47466176000
6	1.515020719995649	0.4844011454408787	2.417355207912524	1.2464697848489301	1.1902439076897424	-0.46815887300
7	0.7170704807586563	1.3458131128467437	-0.041038787956365314	0.8575270975337127	-1.1186106627721117	-0.97141031650
8	1.4010278286760784	-0.8077168056679188	-0.14673950067749647	-0.34240246971536203	-0.06530109595557246	-0.97141031650
9	-0.19487264979790678	1.3458131128467437	-0.9951866194551783	0.6671934420390319	1.3250675322422594	0.538344013989
10	1.0590491547173675	0.4844011454408787	0.26924394938631	0.7664979579493002	0.12850786433867079	-0.46815887300
11	1.6290136113152194	-1.4537757812223175	0.026586936849089568	1.2712959138264972	-0.41078663387139736	0.035092570491

Train Model Page:





Performance Tabulation Page:



Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1867.2239688846246	19.705669574597312	-0.29512508225990364	11.460267005721551	-0.29514937607124203
Bayesian Ridge	1433.4345539141075	18.852547274124067	0.005776442279156524	12.446627562921444	0.005736912615902145‌5
Random Forest	1423.2013443860378	18.699950611012696	0.012858227310950543	11.89757396176135	0.01283490147869637
Gradient Boosting	1431.3802943735736	18.832858921437772	0.007198224638289896	12.412020405078703	0.007161794154778245

Visualization Page:

VI. Conclusion

Here proposed method deals with consider flight delay prediction using boosting techniques like XgBoost which involves extreme gradient boosting. We may also model a neural network which are high in complexities but offers higher accuracy and automation of feature selection.

VII. REFERENCES

[1]. Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.

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

[3]. Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

[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, "Further Investigation into the Causes of Flight Delays," in Department of Economics, East Carolina University, 2007.

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

[7]. A. M. Kalliguddi, Area K., Leboulluec, "Predictive Modelling of Aircraft Flight Delay," Universal Journal of Management, pp. 485 - 491, 2017.

[8]. Noriko, Etani, "Development of a predictive model for on-time arrival fight of airliner by discovering correlation between fight and weather data," 2019.

[9]. Chakrabarty, Navoneel. "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines." 2019 9th Annual Information Technology, Electromechanical

Engineering and Microelectronics Conference (IEMECON) (2019): 102-107.

[10]. Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. (2017). A Review on Flight Delay Prediction.

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

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

Cite this article as :

T. Vasanth Kumar Reddy, Dr. Srinivasan Jagannathan, Mr. Suresh, "Formers Products Online", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 4, pp. 266-271, July-August 2022.
Journal URL : https://ijsrcseit.com/CSEIT228446