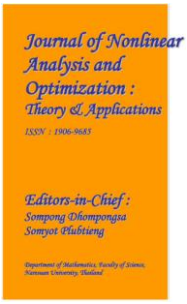
Journal of **N**onlinear **A**nalysis and **O**ptimization

Vol. 14, Issue. 2 : 2023

ISSN : **1906-9685**



ACHIEVING OPTIMAL FLIGHT DELAY PREDICTION AND ERROR CALCULATION USING MACHINE CLASSIFIERS

[1]Hanumanthappa Mahesh, [2]Modadugu Dimple Abhinaya Sri, [3]Vakala Naga Sindhu, [4]Mandala Koushitha Reddy, [5]Kummari Harika

[1]Assitsant Professor, [2,3,4,5]UG Students, Dept. of CSE (AI & ML), Malla Reddy Engineering College for Women (Autonomous), Hyderabad, India. E-Mail: maheshh118@gmail.com

ABSTRACT

Flight delays are one of the main issues facing the aviation industry. Increased air traffic as a consequence of the aviation industry's expansion throughout the last two decades has led to flight delays. Flight delays not only waste time and money, but they also have an adverse effect on the environment. Airlines that operate commercial flights suffer huge losses as a result of flight delays. As a result, they make all necessary steps to prevent or minimize flight disruptions and delays. In this research, we apply algorithms based on machine learning to forecast when a specific flight's arrival will be postponed or not, including logistic regression, decision tree regression, bayesian ridge, random forest regression, and gradient boosting regression.

Keywords: Machine learning, ML Models, error computation

INTRODUCTION

Flight is an essential mode of transportation in this century, allowing people to travel across far distances in a short amount of time. Several industries have been blooming along with airline industries, and tourism is one of the key players. However, as per recent observation, the COVID-19 pandemic has caused an adverse impact on businesses in terms of maintenance for several areas highly dependent on international tourists. Also, despite the growth of aviation industries, operational inefficiencies still need to be addressed, and one of the prominent ones

is flight schedule delays. According to Federal Aviation Administration, flights that are delayed for more than 15 minutes than the programmed time are measured delayed flights (NASEM, 2014).

In the event of a flight delay, the parties that are usually directly impacted are the airlines and passengers. The delay of one flight could propagate and impact the other subsequent flights. For airlines, higher counts of delays cause passengers' demands to decline. Also, airfares are higher for routes with higher delay counts. Thus, the buffer time is added to schedules to curb the delays, and aircraft can still arrive at their destinations as scheduled. However, this scenario is less likely to happen in crowded or busier airports. A longer buffer time translates to lesser scheduled flights for the day (NEXTOR, 2010).

A mathematical method for generating rough estimations from input data is statistical modelling. The significance of distance, date, and expected departure time in predicting flight delays has been demonstrated via a multiple regression model. Predictive modelling has several uses, such as predicting the likelihood of email spam and flight delays. Because they focus on the primary causes of aircraft delays, regression models were shown to be successful in forecasting flight delays during the examination of the effectiveness of various models in modelling flight delays.

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

However, they are unable to categorise complicated data. The models produce biassed and prejudiced results based on socioeconomic factors. The random forest model outperforms the competition. Depending on factors like prediction time and airline dynamics, the accuracy of predictions may vary.

Objectives

Flight delays are significant concerns in aviation industries, leading to revenue loss, fuel loss, and customer dissatisfaction. It creates fear among passengers taking a connecting flight, whereby the delay from the first flight could potentially cause them to miss the subsequent flight. Therefore, this scenario is a factor of motivation for this study. With a reliable method to predict flight delays, the event mentioned in the previous context could either be prevented or better managed.

The objectives of this study are:

To identify the attributes that affect flight delay.

To develop machine learning models that classify flight outcomes (either delayed or not delayed) with selected features.

To evaluate the presentation of different machine learning model

LITERATURE SURVEY

Flight delays have been the subject of extensive research. For air traffic control, identifying, analysing, and preventing flight delays has been a key challenge. An airfield manoeuvring area is a complicated structure that requires the use of a two-stage technique based on real-time simulation and fast simulation approaches to evaluate the air traffic flow. A baseline model developed to identify the bottleneck locations is evaluated in the first stage using quick and real-time simulations. The second stage involves the generation and evaluation of alternative scenarios that incorporate these upgrades in a fast-time simulation environment.

In order to locate congestion nodes in the maneuvering zones of huge airports and come up with strategies to reduce the congestion, this study combines fast- and real-time simulation approaches. It is demonstrated that the other strategy would enhance hourly operations while reducing overall ground delays. "Chakrabarty et al. [4] employed supervised automatic learning algorithms" (random forest, Gradient Boosting Classifier, Support Vector Machine, and the k-nearest Neighbour technique) to forecast delays in the expected arrival of operational planes, comprising the five busiest US airports. "Choi et al. [5] applied machine learning techniques" including decision tree, random forest, AdaBoost, and kNearest Neighbours to forecast delayed on particular flights, and found that gradient booster as a classification given limited data had the highest level of precision (70.7%). The model has been updated with data from aircraft schedules and weather forecasts. It has been demonstrated that the classification developed with no sampling significantly more precise than the classifiers taught using methods of sampling by balancing the data using sample A multiple linear regression model was used by Sruti Oza and Somya Sharma [6] to forecast weather-related flight delays in flight data, in addition to climatic conditions and risk of such delays. The forecasts depended on a few important variables, such as the airline, the arrival and departure times, and the origin and destination. Flight data was utilized by Anish M. Kalliguddi and Aera K. Leboulluec [7] to predict either arrival or departure delays using regression models such as Decision Tree Regressor, Multiple Linear Regression, and Random Forest Regressor. In order to increase random forest efficiency without reducing forecast error, larger forecast horizons are being shown to be advantageous.

Big Data Etani J Utilising flight and meteorological data, a supervised model [8]" of on-schedule arrival flights is used. Peach Aviation's pressure patterns and flight data are discovered to be related. Use of Random Forest as a Classifier allows for 77% accurate prediction of flights arriving on time.

The paper "SobhanAsian [9] Flight Delay prophecy for Commercial Air Transport" looks at high-dimensional data as of Beijing International Airport and provides a helpful model for predicting aircraft

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

delays. Following a multifactor approach, the internal patterns of flight delays are mined using a unique deep belief network technique. The created model employs support vector regression to perform supervised fine-tuning inside the proposed predicting structure. The suggested approach has been shown to be quite successful in addressing the difficulties posed by huge data sets and pinpointing the crucial elements generating delays. In order to minimize delay propagation throughout their network, connecting airports might cooperate by employing methods including synchronized delayed prediction. An article titled "Xiaotong Dou [10] Flight Arrival Delay Prediction and Analysis Using Ensemble Learning" The amount of civil aviation transportation have significantly expanded as a result of the recent growth of the civil aviation transportation business. Due to aircraft delays, increased carrier expenses and decreased airport operational efficiency have emerged as problems that require attention. How to increase airport transportation effectiveness, logical flight scheduling, and customer comfort while also increasing the accuracy of projecting flight arrival delay time.

EXISTING SYSTEM

The existing system proposed that the major contribution of the aviation industry to the American economy is highlighted by the anticipated augmentation in air travel stipulate and the positive association by means of economic indicators. In terms of airline performance and passenger happiness, on-time operations are crucial. Therefore, a thorough analysis of the factors that contribute to delays is crucial. In recent years, the use of machine learning methods in data mining has grown rapidly, attracting interest from a growing number of academic fields, including aviation.

Examining the possible application of SVM models for the study of flight delay causes and examination of flight delay patterns is the main contribution of the current work. The best level of precision was reached by gradient booster as a classifier by little data, coming in at 79.7%.

PROPOSED SYSTEM

By adopting certain actions, our suggested model tries everything within its power to prevent or avoid aircraft delays and cancellations. Machine learning models including Logistic Regression, Decision Tree Regression, and Random Forest Regression are worn in this design. We forecast whether a particular flight will arrive on time or not. We develop a system that predicts airline departure delays based on a number of variables. To train our forecasting model, we employ a range of flight-specific data, for instance arrival presentation, flight summary, origin/destination, etc.

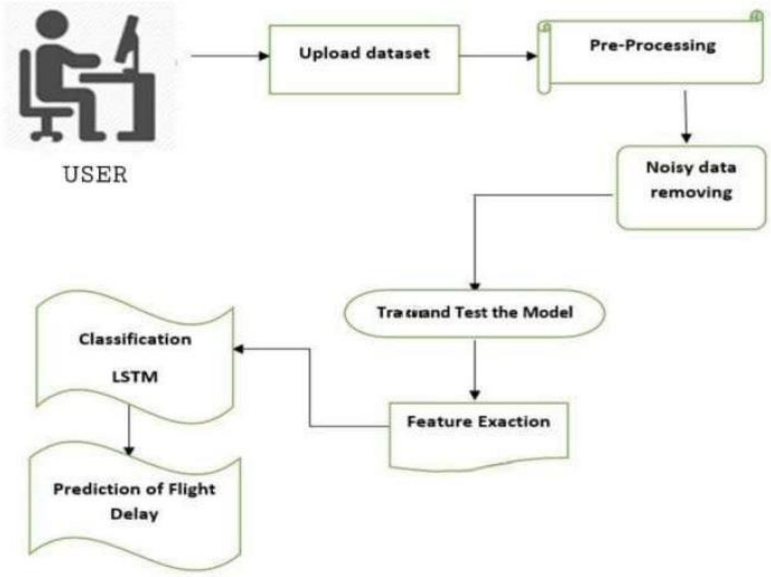


Fig.1. Proposed model

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

MODULES AND FUNCTIONALITIES

Date Time module

Classes for working with date and time are provided by the Python Date time package. Numerous functions that deal to dates, times, and time spans are provided by these classes. Python treats Date and DateTime as objects, therefore when you work with them, you're really working with objects rather than strings or timestamps.

Six major classes make up the Date Time module:

Date: An idealised naive date based on the premise that the Gregorian calendar used now is constantly in use and continually will be. The year, month, and day are its characteristics.

Time is an idealized period of time, untainted by any specific day, where each day contains approximately 24 hours and 60 minutes. Hour, minute, second, microsecond, and tzinfo are some of its properties.

date time is an amalgamation of the elements year, date, and time,month, day, hour, minute, second, microsecond, and tzinfo.

time delta – A duration expressing the difference between two date, time, or datetime instances to microsecond resolution.

tzinfo – It gives time zone in sequence items.

time zone – A group that executes the tzinfo abstract base class as a permanent offset as of the UTC (New in version 3.2).

The cautions Module Warnings are notifications of defects or anomalies that might not be critical enough to warrant interfering with programmed control flow (as would occur by raising a typical exception).

Fine-grained control over which warnings are output and what happens to them is possible by the warnings module. By calling the function warns in the module warnings, you can conditionally emit a warning. You can format warnings anyway you like, specify where they should go, and conditionally suppress some warnings (or turn some warnings into exceptions) using the module's other functions.

Several exception classes that represent warnings are provide by the Classes-Module warnings module. Class The basic class for all warnings is called Warning, which is a subclass of Exception. You can create your own warning classes, but they must directly or indirectly subclass Warning by one of its other subclasses, which are:

Deprecation Uses out-of-date features that are solely provided for backward compatibility

Sci Module

A scientific computation package known as SciPy is built on top of NumPy. Scientific Python is a common abbreviation. It offer added supportive functions for signal processing, statistics, and optimization. Algebraic equations, differential equations, statistics, eigen value issues, integration, interpolation, optimization, and many other classes of problems are all addressed by algorithms in SciPy. The SciPy data structures and algorithms have broad domain compatibility. provides new array computation tools and specialised data structures, such as sparse matrices and k-dimensional trees, to extend NumPy. High-performance implementations created in low-level languages are wrapped in SciPy.

Data collection

In order to forecast flight delays and train models, we used information acquired by the Bureau of Transportation, U.S. Statistics of all domestic flights taken in 2015.

Statistics on arrival and departure, such as wheels-off time, departure delay, and taxi-out time per airport, are provided by the US Bureau of Transport Statistics. It also gives the scheduled elapsed time, the planned elapsed time, and the scheduled departure time. Along with the date, time, flight identity, and airline airborne time, the airport and the airline also provide postponement and rerouting

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

information. The data collection consists of 27 columns and 1048576 rows. Figure 2 displays a few of the fields from the original dataset.

present be a large number of lines with empty or blank values. To benefit from have the agenda and actual arrival time, the methodology in this situation employs the supervised learning technique. For later use, the data needs to be pre-processed. After exploring a few novel monitoring techniques with minimal computation costs, the best contestant be finally refined for the final model. We enlarge a structure that forecasts airline departure delays based on a number of variables. To train our forecasting model, we employ a range of flight-specific data, for example arrival presentation, flight summaries, origin/destination, etc.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	OP_CARRIE	OP_CARRIE	ORIGIN	DEST	CRS_DEP_T	DEP_TIME	DEP_DELA	TAXI_OUT	WHEELS_O	WHEELS_O	TAXI_IN	CRS_ARR_T	ARR_TIME	ARR_DELA	CANCELLED	CANCELLA	DIVERTED	CRS_ELAPS
2	UA	2429	EWR	DEN	1517	1512	-5	15	1527	1712	10	1745	1722	-23	0		0	268
3	UA	2427	LAS	SFO	1115	1107	-8	11	1118	1223	7	1254	1230	-24	0		0	99
4	UA	2426	SNA	DEN	1335	1330	-5	15	1345	1631	5	1649	1636	-13	0		0	134
5	UA	2425	RSW	ORD	1546	1552	6	19	1611	1748	6	1756	1754	-2	0		0	190
6	UA	2424	ORD	ALB	630	650	20	13	703	926	10	922	936	14	0		0	112
7	UA	2422	ORD	OMA	2241	2244	3	15	2259	1	2	14	3	-11	0		0	93
8	UA	2421	IAH	LAS	750	747	-3	14	801	854	6	916	900	-16	0		0	206
9	UA	2420	DEN	CID	1324	1318	-6	11	1329	1554	6	1619	1600	-19	0		0	115
10	UA	2419	SMF	EWR	2224	2237	13	10	2247	627	9	638	636	-2	0		0	314
11	UA	2418	RIC	DEN	1601	1559	-2	12	1611	1748	8	1813	1756	-17	0		0	252
12	UA	2417	PDX	EWR	2240	2235	-5	9	2244	624	7	647	631	-16	0		0	307
13	UA	2416	ORD	CLE	2059	2300	121	24	2324	112	8	2311	120	129	0		0	72
14	UA	2415	EWR	PDX	825	822	-3	15	837	1104	5	1135	1109	-26	0		0	370
15	UA	2414	EWR	ATL	1044	1055	11	11	1106	1310	5	1318	1315	-3	0		0	154
16	UA	2413	ORD	BTV	2114	2230	76	14	2244	123	5	15	128	73	0		0	121
17	UA	2412	MCO	LAX	653	747	54	14	801	1003	22	930	1025	55	0		0	337
18	UA	2411	EWR	SMF	1810	1922	72	16	1938	2157	4	2136	2201	25	0		0	386
19	UA	2410	RSW	EWR	1250	1337	47	12	1349	1600	6	1537	1606	29	0		0	167
20	UA	2409	IAH	JAC	940	934	-6	18	952	1156	4	1218	1200	-18	0		0	218

Fig.2. Snapshot of Dataset

Pre-Processing

The classifier must be given the datasets after the data has been taken from the twitter source. Before doing the analysis, the classifier purges the dataset of unnecessary information such as stop words and emoticons. This ensures that non-textual stuff is recognized and eliminated. To facilitate the modeling process, the only flight data that was considered and included was the data from the busiest airports since they contained the most significant number of schedules for arrival flights in the U.S. Data cleansing was performed on the name of flight carrier, origin airport and destination airport as the abbreviation of IATA code was used. Attributes with more than 50% of missing values that did not provide helpful information to this analysis were dropped—unrelated attributes such as attributes that recorded the outcome of canceled flights and diverted flights were also removed. Since our main objective was to predict flight delay, attributes relating to canceled flights were eliminated.

For classification purposes, a binary attribute, namely "flight delay," was added to the record status of the flight. The duration between the flights taking off and the wheels off the ground, as well as flight on land and wheels on land, were derived as this provided information about the actual duration of these activities. Information about a month, day, and day of the week was transformed from the actual flight date. Before modeling, all categorical attributes such as destination airports, day of the week, flight carrier, and flight delay factors were converted to numerical variables via one hot encoding method.

Feature Selection:

The constant variable was removed as it did not provide helpful information to the model. Attributes highly correlated to each other were examined to avoid the multicollinearity effect on the model by selecting the most predictive one. Planned elapsed time, airtime, distance, and actual elapsed time correlate higher than 0.8. In this group, several attributes were highly correlated. To select which attributes to remove, a random forest algorithm was utilized to determine their feature importance. Thus, the actual elapsed time was not removed as it gave the greatest importance compared to other attributes (shown in Table below).

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

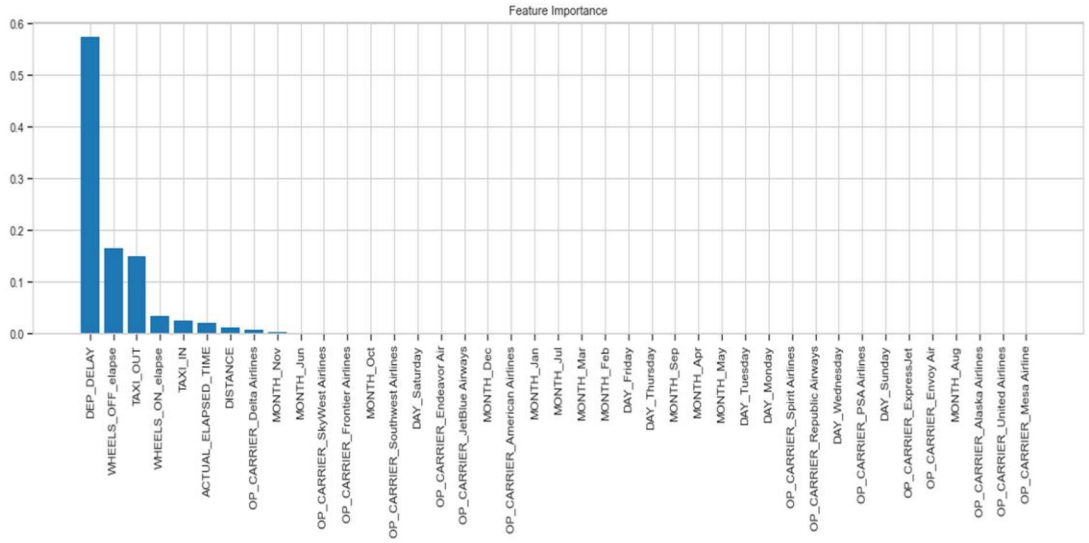


Fig.3. Feature Selection

Evaluation

Following pre-processing and characteristic taking out on our dataset, 60% of the dataset was chosen for training and 40% for testing. Planned for calculating errors, we're by means of Scikit-learn metrics. The consequences are reported in Departure Delay(A) and Arrival Delay(B), the two sections.

Departure Delay Based on multiple assessment measures, our results for departure delay compare numerous Machine Learning models, including Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor, and Gradient Boosting Regressor. Additionally, we evaluate each model against a single assessment metric. B. Arrival Delay Based on multiple assessment metrics, Several Machine Learning simulations, such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression, and Gradient Boosting Regression are compared in our findings for arriving delay. Additionally, we evaluate each model against a single assessment metric.

Model	Precision (P.R.)		Recall (R.C.)		F1-score (F1)	
	On-time	Delayed[1]	On-time	Delayed[1]	On-time	Delayed[1]
Base Algorithms						
Naïve Bayes (NB)	97.7%	69.6%	95.1%	83.6%	96.4%	75.9%
Logistic Regression (L.R.)	97.6%	89.6%	98.7%	81.6%	98.1%	**85.4%**
Decision Tree (D.T.)[2]	97.4%	86.9%	98.4%	80.6%	97.9%	83.6%
Random Forest (R.F.)[3]	97.0%	85.4%	98.2%	77.2%	97.6%	81.1%

Table.1. Data Evaluation

The F1 score of models trained with resampled data did not change much compared to the models trained with imbalanced data. However, the recall metric for N.B., L.R., D.T., and R.F. have increased respectively after applying resampling techniques.

Implementation

LIBRARIES USED:

NumPy

A general-purpose programme for handling arrays is called NumPy. It offers a multidimensional array objects via outstanding speed as well as capabilities for interacting with such arrays. It is the cornerstone Python module for scientific computing. It has a number of characteristics, especially the following crucial ones:

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

a. An effective N-dimensional array object
b. Complex (broadcasting) operations
c. Functionality for effective linear algebra, the Fourier transform, and random numbers. c.
d. Tools for integrating C/C++ and Fortran programmers

Pandas

Pandas is an open-source Python toolkit so as to provides high-performance data manipulation and analysis tools by means of its authoritative data structures. Python was mostly worn for data mugging and preprocessing. On data analysis, it had little of an effect. Pandas discovered the answer. No matter where the data came from, we may use Pandas to complete the five typical steps of data processing and analysis: prepare, modify, model, and analyse. Many academic and professional fields, including finance, economics, statistics, analytics, etc., use Python with Pandas.

Matplotlib

Publication-quality graphics are produced in a variety of physical formats and cross-platform interactive settings using the Python 2D plotting package Matplotlib. Matplotlib can be used by four graphical user interface toolkits, Web-based application servers, Jupyter Notebooks, the Python and Ipython shells, and Python scripts. Matplotlib aims at rendering both challenging and basic tasks achievable. Using just a couple of lines of code, you can create graphs, histograms, power spectra, bar charts, error charts, scatter plots, and more. Explore sampling plots and thumbnail gallery for samples.

The Scikit-Learn

Utilising a common Python interface, Scikit-learn offers a variety of supervised and unsupervised learning techniques. It is offered beneath a liberal basic BSD license that encourages both academic and commercial use and is provided under numerous Linux distributions.

Python.

Python is Interpreted the interpreter processes Python while it is being used. Your programme does not need to be compiled before running. This is comparable to PHP and PERL.

Python is Interactive − you can actually sit

ALGORITHMS

MACHINE LEARNING

A cutting-edge area of science called machine learning gives computers the ability to acquire knowledge on their own using historical data. Machine learning uses a variety of approaches to develop mathematical models and produce recommendations that utilize previously gathered data or information. Nowadays, it is employed for a wide range of tasks, including recognizing images, Facebook auto-tagging, speech recognition, recommender systems, and email filtering. Some claim that the field of artificial intelligence known as "machine learning" focuses mostly on developing algorithms that let computers independently study data and prior experiences to learn. Arthur Samuel coined the phrase "machine learning" for the first time in 1959. A succinct explanation of machine learning is provided below: "Without having to be explicitly programmed, machine learning permits a machine to learn on its own from data, enhance effectiveness through situations, and anticipate things."

A machine learning structure uses the prediction models it has created using historical data to predict the outcome when it gets new data. The amount of information is analyzed determines the precision with which the result can be anticipated, as a larger data set makes it easier to develop a model which more accurately forecasts the result. Consider a difficult circumstance that calls for some forecasts. We could just provide the data to generic algorithms, which would then use the input to create logic and forecast the consequences, rather than writing specific code for it. Machine learning has changed our perspective on the problem.

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

LOGISTIC REGRESSION ALGORITHM

Logistic regression is one of the Machine Learning techniques that is usually employed in the Supervised Learning category. It is employed to forecast the dependent variable that is categorical using a specified set of independent variables. Logistic regression can be used to forecast the outcome of a dependent variable that is categorised. The outcome must therefore be a discrete or classifying value. It provides probabilistic values that fall in the range of 0 to 1, rather than the exact numbers ranging from 0 to 1. moreover True or False, 0 or 1, or Yes or No, are possible outcomes.

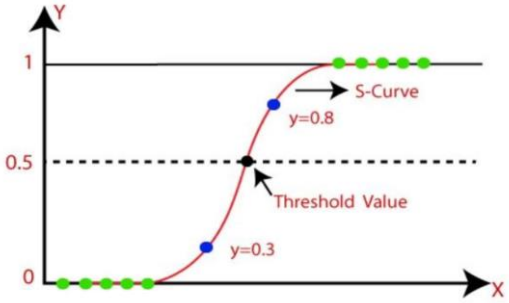


Fig.4. logistic regression

DECISION TREE ALGORITHM

Regression and classification problems are able to be resolved using the supervised learning technique known as a decision tree, however this approach is frequently preferred. It is a classifier with a tree structure, wherein every leaf node represents the classification outcome and inside nodes represent the features of a dataset. The two nodes in a decision tree are the Decision Node and Leaf Node. Decision nodes are utilized to make judgments and have many branches, whereas Leaf nodes are the outcomes generated by choices and are devoid of any more branching.

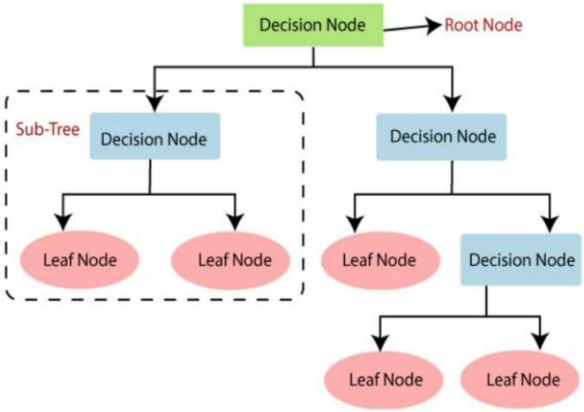


Fig.5.Decision tree algorithm

RANDOM FOREST ALGORITHM

Recommended machine learning technique Random Forest is a part of the controlled learning strategy. It can be applied to ML issues involving regression as well as classification. Its foundation is the idea of ensemble learning, which is a method for mixing several classifiers to solve complex issues and enhance the accuracy of the models. A classification algorithm is referred to as "Random Forest" if it "includes an assortment of decision trees based on different subsets of the data at hand and chooses average into account to enhance the predictive power of that dataset."

Instead of depending exclusively on one decision tree, the random forest incorporates forecasts across every tree of choices and anticipates the result according to the responses of the greatest number of projections. The forest has more trees because of this accuracy is higher and the over fitting issue is avoided.

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

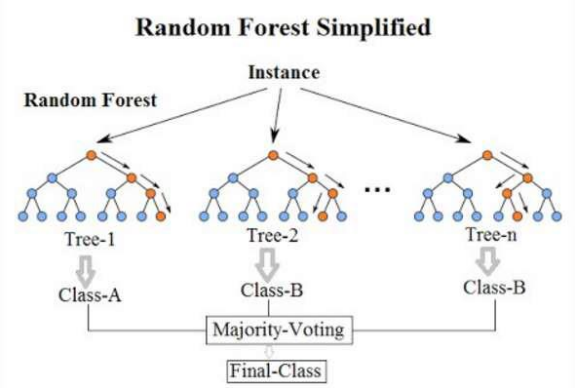


Fig.6. Random forest algorithm

Testing Scenarios

S.NO	TEST SCENARIO	TEST STEPS	EXPECTED RESULT	RESULT
1	Check whether the flight is delay	Test with delay dataset	Successful in displaying the result	Success
2	Check whether the flight is delay	Test without delay dataset	Successful in displaying the result	Success

Table.2. Testing Scenarios

Results

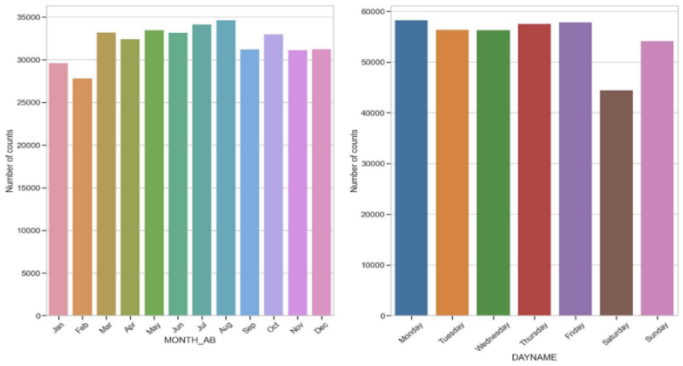


Fig.7. Categorical features

```
Logistic regression
Roc-auc Logistic Regression roc-auc: 0.9687186016187245

Accuracy Logistic Regression test: 0.9478190818964958

Precision Logistic Regression test: 0.8554251915058664

Recall Logistic Regression test: 0.7774742222613906

F-measure Logistic Regression test: 0.8145891043397968

Summary Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97     65616
           1       0.86      0.78      0.81     11347

    accuracy                           0.95     76963
   macro avg       0.91      0.88      0.89     76963
weighted avg       0.95      0.95      0.95     76963
```

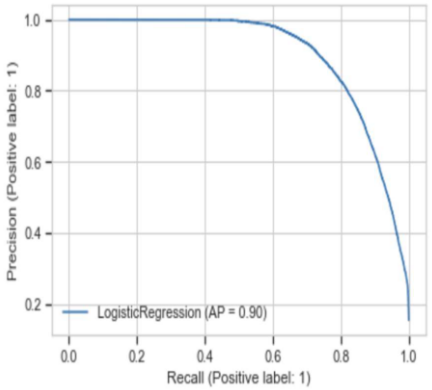


Fig.8. Logistic regression summary report and graph

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

Test set
Roc-auc Decision Tree roc-auc: 0.911810166784978

Accuracy Decision Tree: 0.9553291841534244

Precision Decision Tree: 0.8474040235438812

Recall Decision Tree: 0.8500925354719309

F-measure Decision Tree: 0.8487461504619446

Summary Report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	65616
1	0.85	0.85	0.85	11347
accuracy			0.96	76963
macro avg	0.91	0.91	0.91	76963
weighted avg	0.96	0.96	0.96	76963

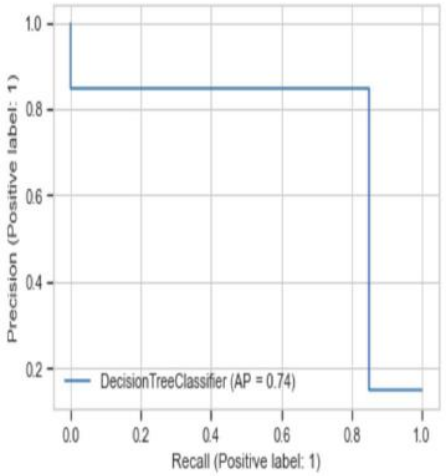


Fig.9. Decision tree summary report and graph

Random Forest
Roc-auc Random Forests roc-auc: 0.967824946807227

Accuracy Random Forest test: 0.8826449072931148

Precision Random Forest test: 0.9463941380640185

Recall Random Forest test: 0.2162686172556623

F-measure Random Forest test: 0.3520803443328507

Summary Report:

	precision	recall	f1-score	support
0	0.88	1.00	0.94	65616
1	0.95	0.22	0.35	11347
accuracy			0.88	76963
macro avg	0.91	0.61	0.64	76963
weighted avg	0.89	0.88	0.85	76963

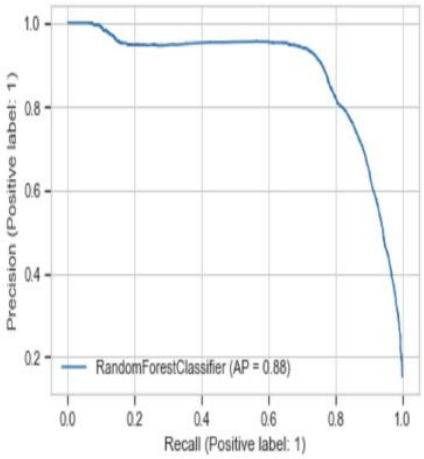
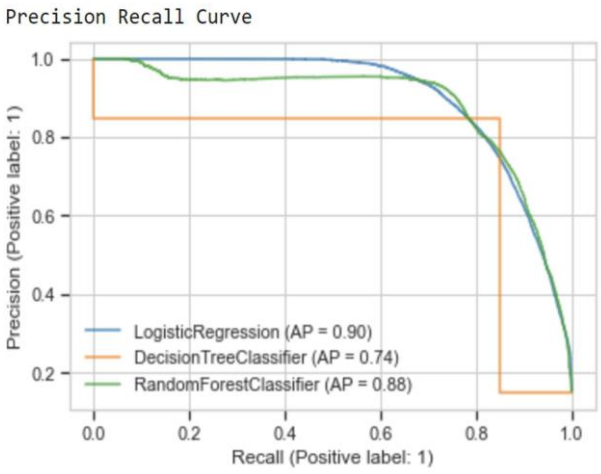


Fig.10. Random forest summary report and graph



http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181

CONCLUSION

The research aims to increase the flight delay prediction model's predicted value accuracy in comparison to other created models. As a result, both travelers and the airline industry will be able to make wiser decisions and expand their businesses. Given how crucial it is to fly on time, flight delay prediction models must be extremely precise. In this paper, we developed a new optimized forecasting model depending on machine learning classifiers that employs the Linear Regression, Decision Tree, and Random Forest algorithms.

FUTURE SCOPE

The scope of the project be able to be extended by training the model with the Neural Network algorithm. To handle imbalance data, there are more options of oversampling technique for example Adaptive Synthetic (ADASYN), which prevents the overlapping of synthetic observations, and under sampling techniques, which employ data cleaning concept using Tomek-link (T.L.) and Condensed Nearest Neighbour (CNN). Other than the re sampling techniques, we can also apply Cost-Sensitive Learning, which considers misclassification costs by applying penalties on the wrongly classified results. We can also employ a hybrid method such as SMOTE Boost to handle the imbalanced data.

REFERENCES

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

[2] "Airports Council International, World Airport Traffic Report," 2015,2016.

[3] E. Cinar, F. Aybek, A. Caycar, C. Cetek, "Capacity and delay analysis for airport manoeuvring areas using simulation," Aircraft Engineering and Aerospace Technology, vol. 86, no. No. 1,pp. 43-55, 2013.

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

[5] Y. J. Kim, S. Briceno, D. Mavris, Sun Choi, "Prediction of weatherinduced airline delays based on machine learning algorithms," in 35th Digital Avionics Systems Conference (DASC), 2016.

[6] W.-d. Cao. a. X.-y. Lin, "Flight turnaround time analysis and delay prediction based on Bayesian Network," Computer Engineering and Design, vol. 5, pp. 1770-1772, 2011.

[7] J. J. Robollo, Hamsa, Balakrishnan, "Characterization and Prediction of Air Traffic Delays".

[8] S. Sharma, H. Sangoi, R. Raut, V. C. Kotak, S. Oza, "Flight Delay Prediction System Using Weighted Multiple Linear Regression," International Journal of Engineering and Computer Science, vol. 4, no. 4, pp. 11668 - 11677, April 2015.

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

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

[11] Available: https://towardsdatascience.com/metrics-toevaluate-your-machinelearning- algorithm-f10ba6e38234.

http://doi.org/10.36893/JNAO.2023.V14I2.0171-0181