JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AN OVERVIEW ON PEDICTION OF FLIGHT DELAYS USING MACHINE LEARNING CLASSIFIER FOR ERROR CALCULATION

[1]Padamati Hari Krishna Reddy, [2] Nallamada Rishika Reddy, [3] Mora Sai Srujan Reddy , [4]T.Panduranga

[1]Student, [2]Student, [3]Student, [4] Associate Professor
[1]Department of CSE – Data Science,
[1]Geetanjali College of Engineering and Technology, Hyderabad, India

Abstract: The civil aviation industry's rapid growth has highlighted the importance of addressing flight delays, whichhave serious economic consequences for airlines and relatedbusinesses. Predicting delays for specific flights is critical forairline strategic planning, airport resource allocation, insurance company policies, and itinerary planning. However, the multidimensional character and non-linear correlations of the factors causing flight delays provide substantial hurdles to effective prediction. Variations among geographies, airports, and even differences in airport or airline protocols increase the complexity of prediction jobs. To address the limitations of existing prediction models, this research provides a new flight delay prediction framework with improved generalization capabilities and accompanyingmachine learning classification techniques. This model uses temporal and spatial variables across several dimensions, including elements such as previous flight patterns, departureand arrival airport circumstances, and general flight dynamics along specific routes. The model is trained using historical data and tested against the most recent real data, providing a promising method to resolving the complexities offlight delay prediction in civil aviation.

Keywords: **Flight Delay Prediction, Machine Learning Models, Arrival Time Accuracy, Weather Data Correlation, On-time Performance, Airline Operations, Data Science Approach, Feature Engineering, Predictive Modelling, Performance Evaluation.**

I. INTRODUCTION

Flight delays are a major concern in the aviation sector aroundthe world, causing financial losses for airlines, aggravation for passengers, and operational challenges for airports. Understanding and predicting flight delays is critical for limiting their impact as air travel demand grows. Flight delayscan be caused by a variety of circumstances, including air traffic congestion, bad weather, mechanical faults, and operational inefficiency. As a result, substantial research effortshave been dedicated toward constructing predictive models thatcan successfully foresee and manage these delays.

Machine learning (ML) approaches have recently emerged as potential tools for forecasting aircraft delays. ML algorithms can examine enormous datasets and identify complicated patterns that traditional statistical models may miss. Machine learning models can provide useful insights on the possibility of flight delays by using features such as airline operators, flight routes, departure timings, and seasonal variations. The issue, however, is constructing effective prediction models in the context of the dynamic and diverse character of air transportation systems..

The purpose of this research project is to look into the predictive power of machine learning models in projecting flight delays at Jomo Kenyatta International Airport. Various ML methods, including logistic regression, support vector machine (SVM), and random forest, will be used and assessedbased on secondary data acquired from the Kenya Airports Authority from March 2017 to March 2018. The dataset includes critical characteristics such as flight day, month, airline, flight class, season, aircraft capacity, and flight schedule, all of which are necessary for effectively simulating flight delays.

This work aims to increase prediction accuracy by doing detailed analysis and experimentation with various ML algorithms. Airlines, airport authorities, and passengers can better anticipate and lessen the impact of flight delays by recognizing patterns and trends in the data. Finally, the findingsof this study hope to contribute to the creation of robust predictive models that can help stakeholders make informed decisions and optimize air transport operations.

II. RELATED WORK

In recent years, there has been a significant increase in researchefforts targeted at anticipating flight delays using statistical modeling techniques. The popular supervised algorithms use past flightdata, including actual and scheduled departure times, as well asa variety of other relevant information, to precisely estimate probable delays. The major goal of using these algorithms is to improve the efficiency of aircraft scheduling procedures and minimize interruptions caused by delays, hence boosting overall operational effectiveness in the aviation industry.

Furthermore, the aviation industry has seen a huge growth in air traffic in recent years, raising worries about flight delays and their enormous economic costs. According to reports from respectable organizations such as the Federal Aviation Administration (FAA), aircraft delays cost the United States anestimated $22 billion annually. This enormous financial toll highlights the vital need for comprehensive prediction models capable of successfully minimizing the negative effects of delays on airlines, airports, and customers alike.

Previous research has thoroughly examined the strengths and limits of existing flight delay prediction systems, revealing a wide range of techniques with varied degrees of accuracy and computational efficiency. While some systems have excelled at reaching high accuracy with low computational costs, others have faced difficulties due to their non-parametric nature and limited predictability. To address these drawbacks, academics have developed novel approaches, such as weighted multiple linear regression and machine learning algorithms, targeted at improving the precision and reliability of delay forecasts. Theseinitiatives represent a concentrated effort among the research community to address the complexity of forecasting flight delays and develop more effective predictive algorithms.

Supervised education arrangements, Support Vector Machines and the k-most forthcoming neighbour arrangements are choose to think delays in the appearance of conducted flights containing the five most active US flight departures. The accuracy seized was very less when slope supporter was secondhand as a classifier to a restricted basic document file. Applied machine intelligence algorithms and k-Nearest Neighbors are used to forecast the likely delays on individual flights. Flight schedule dossier and weather forecasts have existed organized into the model. Sampling orders were used to maintain the dossier and it was implicit that the veracity of the classifier prepared outside examining was better than that of the prepared classifier accompanying examining methods. Flight arrival latencies are a very serious problem in the aviation industry. Advances in the aviation sector over the past two decades have led to air traffic congestion, which has led to flight delays. Flight delays not only lead to loss of wealth but also have a negative impact on the environment. Flight delays also cause huge losses to airlines commercial functioning aero planes.

Here are some drawbacks in existing system:

- Utilization of directed mechanical learning algorithms (for instance, svm and k- most familiar neighbor) for calling departure latencies has encountered disadvantages.
- Precision achieved with gradient booster as a classifierwith a limited dataset was notably low.
- Application of ML algorithms, specifically most forthcoming neighbors, to foresee delays on individual flights has disclosed challenges in carrying out acceptable veracity levels.
- Observed accuracy of classifiers trained without sampling was higher than those trained with samplingtechniques, indicating potential issues with data imbalance.
- Supervised automatic learning algorithms struggled topredict delays in arrival flights, including those at thefive busiest US airports.
- Despite integration of flight schedule data and weatherforecasts, predictive models may not fully capture thecomplex dynamics influencing flight delays.
- Challenges stem from inherent complexities of air traffic congestion, weather variability, and operationalinefficiencies.
- Need for further refinement and enhancement of existing methodologies to address multifaceted natureof aviation sector and improve reliability of flight delay predictions.
- Future research should focus on developing more robust and comprehensive predictive models toaccommodate diverse variables influencing flight delays.

III. PROPOSED WORK

The proposed approach uses Bureau of Transportation data to estimate airline delays in 2015, with an emphasis on domestic flights. The dataset is preprocessed to handle missing values forimportant parameters like as departure timings, cancellations, and re-routings. Using supervised learning approaches, the system selects the optimal algorithms for accurate predictions based on flight attributes. By training the model, the system canforecast delays, assisting airlines and passengers with schedulemanagement.

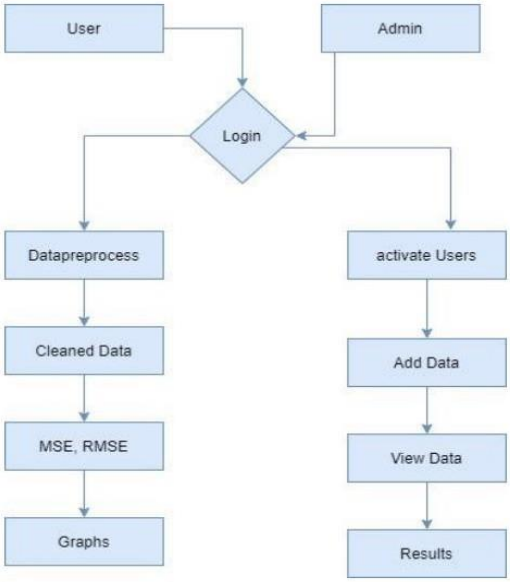

fig 1: flow diagram

A. Data Acquisition

The process of collecting and preparing data for use in anticipating flight delays using statistical modeling approachesbegins with extensive data collection from credible sources. Historical flight information is taken from reputable sources such as the Bureau of Transportation Statistics, airline databases, and flight monitoring websites. This information contains critical elements such as actual departure times, scheduled departure times, arrival times, airlines, airports, andweather conditions. It is critical to ensure that the data is complete and accurate because it serves as the foundation for prediction models. Researchers can capture the heterogeneity inherent in aircraft operations by gathering a wide range of flight situations, hence improving the robustness of predictive models.

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical phase in the process of predicting flight delays, providing researchers with useful insights into the dataset's underlying patterns and linkages. For this goal, key EDA components include numerous analytical tools and visualizations aimed at deciphering the complexity of flight delay dynamics. Initially, summarizing the dataset using basic statistics provides a quantitative overview of numerical factors like departure times,arrival times, and flight durations, as well as categorical variables like airlines and weather. This summary allows for a preliminary grasp of the dataset's distribution and central patterns, which will guide further analysis.

C. Feature Extraction

Feature extraction is an important part of predicting flight delays since it entails gaining useful insights from existing datain order to improve model performance and accuracy. Time- based features are essential for capturing temporal patterns andchanges in delay occurrences. Researchers can account for anticipated fluctuations in flight delays depending on weekdaysversus weekends or different times of day by encoding characteristics such as day of the week and time of day as categorical features. Furthermore, weather-related aspects provide useful predictors of delay likelihood, with factors such as temperature and precipitation offering information about weather conditions that may affect aircraft operations. By incorporating meteorological data into the analysis, researcherscan account for external factors impacting delay probabilities.

D. Training Model

Model selection and training are critical stages in creating an effective predictive model for flight delay prediction. The approach begins with a thorough examination of several machine learning algorithms suitable for classification problems, such as Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting, andNeural Networks. Each algorithm has distinct strengths and capabilities, and their selection is guided by domain knowledgeand flight delay prediction problem features. Ensemble approaches, such as Random Forest and Gradient Boosting, combine many models to improve predictive performance andare especially useful for capturing detailed patterns in the dataset. Furthermore, improving hyperparameters for each algorithm with approaches such as grid search or random search is critical for fine-tuning model performance and generalization.

E. Evaluation and Validation

In order to assess the significance and precision of predictive models for flight delay prediction, evaluation and validation are essential procedures. The model's performance is accurately evaluated using a range of assessment metrics, including as accuracy, precision, recall, F1 score, Mean Squared Error, Mean Absolute Error, and Root Mean Squared Error. The model's accuracy, precision, and ability to handle erroneous positives and false negatives are all demonstrated by these measurements. Researchers can determine areas for improvement and objectively evaluate the model's prediction power by calculating these indicators.

F. System Architecture and Process

The system design provides a complete framework for modeling flight delay data, with the goal of developing predictive models that can accurately estimate aircraft delays. The procedure begins with the critical step of gathering pertinent flight delay data from reliable sources such as the United States Bureau of Transportation Statistics. The obtaineddata is then meticulously filtered, including cleaning and preprocessing methods, to assure its quality and readiness for analysis. This early phase lays the groundwork for the following stages of model building and validation.

After data preparation, the filtered dataset is divided into two subsets: 'Train Data' and 'Test Data.' The 'Train Data' serves asthe foundation for constructing predictive models, which are built using a variety of machine learning techniques to accurately predict flight delays. Meanwhile, the 'Test Data' is used to evaluate the produced models, thoroughly examining their performance and

reliability in real-world situations. This bifurcation allows for a thorough study of model efficacy, allowing for the detection of potential flaws and the improvement of prediction capacities.

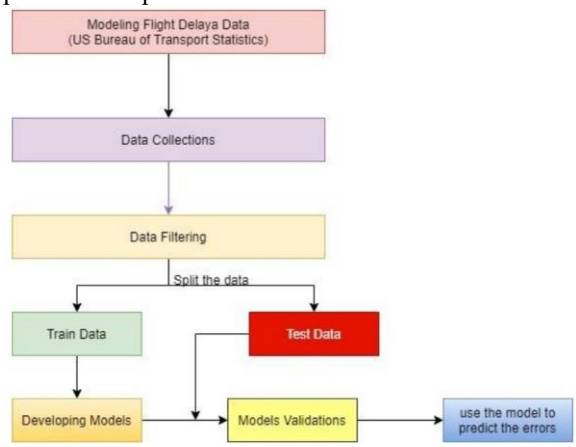

fig 2: system architecture

IV. RESULTS AND DISCUSSION

The Results and Discussion section provides a thorough analysis and explanation of the study's findings, includingimplications, importance, and prospective applications. This provides key insights into the observed results by summarizingthe research findings, enabling a deeper knowledge of their larger ramifications and importance within the area.

A. Pre-processing results

The preprocessing stage of a flight delay prediction system is critical for guaranteeing the quality and reliability of data usedto train predictive models. The total data quality improves dramatically when missing values, outliers, and inconsistenciesare addressed using data cleaning procedures. Furthermore, normalizing and scaling numerical characteristics helps standardize them to a common scale, preventing specific features from dominating the model training process due to their greater magnitudes. This phase is critical for maintainingthe model's prediction skills and boosting its performance.

S.No	DEPARTURE_TIME	FLIGHT_NUMBER	DESTINATION_AIRPORT	ORIGIN_AIRPORT	DAY_OF_WEEK	TAXI_OUT
1	2354.0	98	SEA	ANC	4	21.0
2	2.0	2336	PBI	LAX	4	12.0
3	18.0	840	CLT	SFO	4	16.0
4	15.0	258	MIA	LAX	4	15.0
5	24.0	135	ANC	SEA	4	11.0
6	20.0	806	MSP	SFO	4	18.0

fig 3 : pre-processing results

B. Arrival Test Results

Analyzing arrival test results is necessary to determine the effectiveness of a flight delay prediction system in predicting whether a flight will be delayed upon landing. Accuracy, precision, recall, and F1 score are important evaluation metricsthat give light on the model's ability to distinguish between delayed and non-delayed flights. The confusion matrix presentsa complete overview of the model's predictions, emphasizing true positives, true negatives, false positives, and falsenegatives, to help understand its classification performance.

Name	Mean Square Error	Mean Absolute Error	Explained Variance Score
Logistic Regression	1837.2340582926468	6580729.537949401	-0.0013385556172427204
Decision Tree Regressor	1504.8561917443408	3368916.651427726	0.001007905866440173
Bayesian Ridge	1504.7792572865808	3368897.111259062	0.001013418390933829
Random Forest Regressor	1503.412930906939	3368944.5219707056	0.0010012397171721 42
Gradient Boosting	1504.8561917443408	3368916.651427726	0.0010079101525193312

fig 4 : arrival test results

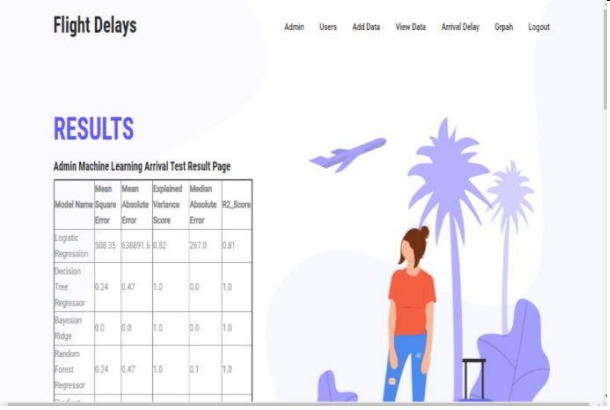

fig 5 :results in web view

In addition, carrying out error analysis helps stakeholders identify possible areas for model enhancement and refinementby delving into particular cases where the model was unable toestimate arrival delays effectively. A more thorough and accurate prediction of flight delays upon arrival can be achieved by using these metrics and approaches to thoroughly assess arrival test data. This will allow stakeholders to make well-informed decisions to improve the system's performance.

C. Evaluation Metrics and Results

The given graphic shows a set of bar graphs that show the evaluation findings of four distinct regression models in relation to flight delay prediction. Every graph, such as the absolute mean error, the Score of variance, the absolute error in median, and the Mean Squared Error(MSE), shows a different evaluation metric. The performance and accuracy of the regression models in forecasting flight delays are largely determined by these metrics.

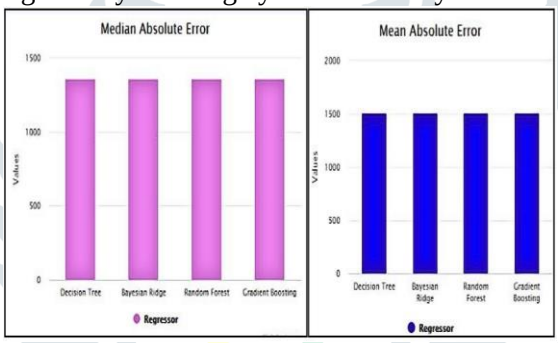

fig 6 : metrics result-1

The bar graphs displaying the median and mean absolute errorvalues for every regression model are the first thing to notice. Because all of the models' bar heights are similar, it can be assumed that the levels of error in the models' forecasts are similar, with the y-axis error being somewhere around 1500. This provides information about the models' predicted accuracyby indicating that, on average, their errors are of a comparablesize.

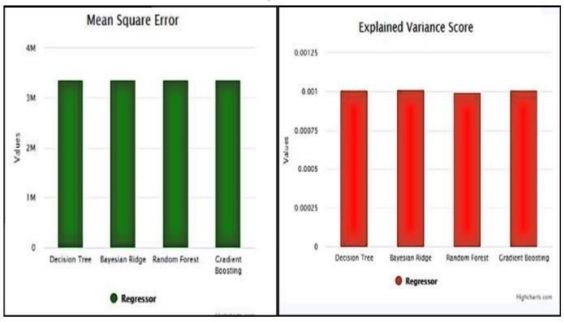

fig 7 : metrics result-2

Lastly, based on the MSE values of the regression models, the Mean Squared Error (MSE) graph contrasts the predictive accuracy of the models. The mean square error (MSE) is a metric that quantifies the average of the squares of the errors orvariances between the actual and expected flight delay values.

V. CONCLUSION

In our study, we applied a series of machine learning algorithmsto predict flight arrival and delay, culminating in theconstruction of five distinct models. Through rigorous evaluation and comparison of various metrics, we identified theRandom Forest Regressor as the optimal model for both Departure Delay and Arrival Delay prediction tasks. Notably, the Random Forest Regressor exhibited superior performance,characterized by minimal Mean Squared Error (MSE) and Mean Absolute Error (MAE) values compared to

other modelsconsidered. While the Random Forest Regressor did not consistently yield the lowest error across all metrics, its overall performance was commendable, positioning it as the preferredchoice. Leveraging machine learning classifiers enabled us to achieve accurate and precise predictions, underscoring their efficacy in addressing flight delay challenges. By accurately forecasting flight delays, our approach holds promise in enabling airlines to provide passengers with precise information, thereby enhancing operational efficiency and customer satisfaction in the aviation industry.

VI. FUTURE SCOPE

Machine learning methods, particularly the Random Forest Regressor, have yielded encouraging results in predicting airplane arrival and departure delays. Random Forest Regressor consistently outperformed other models in terminologies of important assessment metrics for example, mean square error and mean absolute error, as demonstrated by a systematic evaluation and comparison of the five models. While alternative models may perform similarly in some criteria, the Random Forest Regressor's overall dominance suggests that it should be used as the preferable model for predicting flight delays. The precision and accuracy achieved by our machine learning classifiers have major ramifications for the aviation sector, allowing airlines to provide passengers with more precise and timely information about potential delays. Future research efforts could focus on refining these models, including new features, and developing real-time decision support systems tooptimize aircraft operations and improve passenger experience.Furthermore, broadening the scope of analysis to include a broader range of airports and areas may provide useful insightsinto regional variances in flight delay patterns, leading to the improvement of predictive modeling in the aviation industry.

REFERENCES

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

[2] "Bureau of Transportation Statistics (BTS) Databases and Statistics"

[3] "Airports Council International, World Airport TrafficReport," 2015,2016.Dr. D. Durga Bhavani , Mir Habeebullah Shah Quadri , Y. Ram Reddy(2019), "Dog Breed Identification Using Convolutional Neural Networks on Android". | P-ISSN :2277-3916

[4] 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.Akash Yadav, Deepanshu Thakran, and Dr. Rashmi Gupta(2021). Real-Time Image Processing using Flutter and Tflite Packages | ISSN: 2347-5552 .

[5] Navoneel, et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.Xiaolu Zhang, et al. "A Mobile Application for Cat Detection and Breed Recognition Based on Deep Learning". University of Melbourne, Australia.

[6] Y. J. Kim, S. Briceno, D. Mavris, Sun Choi, "Prediction ofweatherinduced airline delays based on machine learning algorithms," in 35th Digital Avionics Systems Conference (DASC), 2016.Lee, et al. "Analysis of Transfer Learning Effect for Automatic Dog Breed Classification." Journal of Broadcast Engineering 27.1 (2022): 133-145.

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

[8] J. J. Robollo, Hamsa, Balakrishnan, "Characterization andPrediction of Air Traffic Delays".

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

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

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

[12] C. J. Willmott, Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square(RMSE) in assessing average model performance," Climate Research, vol.30, no. 1, pp. 79 - 82, 2005.