JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

PREDICTING FLIGHT DELAYS WITH ERROR CALCULATION USING MACHINE LEARNING REGRESSION

[1]Mrs. G. Swetha [2]K.Raghavi, [3]Y.Bhargavi , [4]CH.Pavithra, [5]G.Shravani, ,

[1]Assistant Professor, Dept. of IT, CMR Engineering College, UGC Autonomous, Kandlakoya, Medchal Road, Medchal Dist, Hyderabad-501 401,e-mail : swetha.gottiparthy@cmrec.ac.in

[2,3,4,5]B.Tech Scholar, Dept. of IT, CMR Engineering College, UGC Autonomous, Kandlakoya, Medchal Road, Medchal Dist, Hyderabad-501 401

ABSTRACT: Flight delay is a major problem in the aviation sector. During the last two decades, the growth of the aviation sector has caused air traffic congestion, which has caused flight delays. Flight delays result not only in the loss of fortune also negatively impact the environment. Flight delays also cause significant losses for airlines operating commercial flights. Therefore, they do everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In this paper, using machine learning models such as Logistic Regression, 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.

Keywords: Flight Prediction, Machine Learning, Error Calculation, Logistic Regression, Decision Tree, Bayesian Ridge, Random Forest, Gradient Boosting, Logistic Regression, U.S. Flight data.

I. INTRODUCTION

Statistical modelling is a mathematical way of making approximations from input data. These approximations are then used to make predictions. Statistical models help in predicting the future probabilistic behavior of a system based on past statistical data. Predictive modelling has been used in many fields, for example in crime cases to detect the likeliness of an email being spam and flight delays. In evaluation of how different models perform in modelling of flight delays, regression models have been found efficient in predicting flight delays since they highlighted the various causes of flight delays. However, they could not categorize complex data. Econometric models have been used to model scheduled flight cancellation and to show how delays from one airport were propagated to other destinations. These models did not provide a complete vindication since they ignored variables that were difficult to quantify. When subjected to social-economic situations, the models showed discriminative and subjective results. Among the models used, random forest has been found to have superior performance. Prediction accuracy may vary due to factors such as time of forecast and airline dynamics.

A developed multiple regression model has shown that distance, day and scheduled departure are key factors in predicting flight delay. However, though the model gives flagged out the significant factors, its prediction accuracy was poor. Moreover, the model is limited to only one flight route. Comparison of other models, such as the K-means clustering Algorithms and Fourier fit model, have shown that Fourier fit model could predict flight delays with a high precision. However, the two models were found to be suitable a single airport, but not prediction applied to multiple airports. Probability models such as the normal distribution and the Poisson distribution have been used to Model flight departure and arrival delays. However, the prediction accuracy varied depending on variables such as time duration and the number of airports considered. However, these models are parametric and assume that the response takes a particular functional form. If this

form is not met by the training data set, the resulting model will not fit the data well and the estimates from this model will be poor. Logistic regression model has been used to model flight on-time performance.

The model showed good performance with the training data set and the testing data set. The variance of the model was also low. However, its parametric nature can be a weakness if the training data set will not meet the assumed functional form. Neural networks performed better than logistic regression model in prediction of death in patients with suspected sepsis in an emergency room. This was attributed to the neural networks having few features to be verified before model construction and its ability to fit non-linear relationship between dependent and independent variables. Support Vector Machine (SVM) model was fitted and it was observed to fit all the training data set correctly. In prediction of auto- ignition temperatures of organic compounds, SVM performed better than multiple linear regression and back propagation neural network. Random forests have been used to model delay innovation. Results from this study showed that more decision trees were better but up to a certain critical value. Prediction of new vehicle prediction approach in computational toxicology led to results with random forest performing better than decision tree.

II. LITERATURE SURVEY

Meel, P., Singhal, M., Tanwar, M., & Saini, N. et.al [1] Flight delay is a major problem in the aviation sector. During the last two decades, the growth of the aviation sector has caused air traffic congestion, which has caused flight delays. Flight delays result not only in the loss of fortune also negatively impact the environment. Flight delays also cause significant losses for airlines operating commercial flights. Therefore, they do everything possible in the prevention or avoidance of delays and cancellations of flights by taking some measures. In this paper, using machine learning models such as Logistic Regression, 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.

Navoneel, Chakrabarty et. al [2] presented a Flight Arrival Delay Prediction Using Gradient Boosting Classifier. The basic objective of the proposed work is to analyze arrival delay of the flights using data mining and four supervised machine learning algorithms: random forest, Support Vector Machine (SVM), Gradient Boosting Classifier (GBC) and k-nearest neighbor algorithm, and compare their performances to obtain the bestperforming classifier. To train each predictive model, data has been collected from BTS, United States Department of Transportation. The data included all the flights operated by American Airlines, connecting the top five busiest airports of United States, located in Atlanta, Los Angeles, Chicago, Dallas/Fort Worth, and New York, in the years 2015 and 2016. Aforesaid supervised machine learning algorithms were evaluated to predict the arrival delay of individual scheduled flights. All the algorithms were used to build the predictive models and comparedto each other to accurately find out whether a given flight will be delayed more than 15 min or not. The result is that the gradient boosting classifier gives the bestpredictive arrival delay performance of 79.7% of total scheduled American Airlines' flights in comparison to KNN, SVM and random forest. Such a predictive model based on the GBC potentially can save huge losses; the commercial airlines suffer due to arrival delays of their scheduled flights.

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

Noriko, Etani et. al [3] provided Development Of A Predictive Model For On-Time Arrival Fight Of Airliner By Discovering Correlation Between Fight And Weather Data. An important business of airlines is to get customer satisfaction. Due to bad weather, a mechanical reason, and the late arrival of the aircraft to the point of departure, flights delay and lead to customer dissatisfaction. A predictive model of on-time arrival flight is proposed with using flight data and weather data. The key research in this paper is to discover the correlation between flight data and weather data. The relation between pressure pattern and flight data of Peach Aviation, which is LCC (Low-Cost Carrier) in Japan, are clarified, and it is found that the sea-level pressures of 3 weather observation spots, which are Wakkanai as the most northern spot, Minami-Torishima as the most eastern spot, and Yonagunijima as the most western spot, can classify the pressure patterns. As a result, on-time arrival fight is predicted at 77% of the accuracy with usingRandom Forest Classifier of machine learning. Furthermore, feasibility of the predictive model is evaluated by developing a tool of on time arrival flight prediction.

A. M. Kalliguddi, Area K., Leboulluec et.al [4] Flight delay has been one of the major issues in the airline industry. A study by Frankfurt-based consulting company 'Aviation Experts', presented that costs of $25 billion were incurred in 2014 due to flight delays worldwide. Domestic flight delays have an indirect negative impact on the US economy, reducing the US Gross Domestic Product (GDP) by $4 billion. This project investigates the significant factors responsible for flight delays in the year 2016. The data set extracted from Bureau of Transportation Statistics (BTS) containing one million instances each having 8 attributes is used for the analysis. We describe a predictive modeling engine using machine learning techniques and statistical models to identify delays in advance. The data set is cleaned and imputed and techniques such as decision trees, random forest and multiple linear regressions are used. We attempt to put forth a solution to the delay losses incurred by the airline industry by identifying the critical parameters responsible for flight delay. Not only airlines incur a huge amount of cost per year, airport authorities and its operations are also affected adversely. This leads to inconvenience to the travelers. Predictive modeling developed in this study can lead to better management decisions allowing for effective flight scheduling. In addition, the highlighted significant factors can give an insight into the root cause of aircraft delays.

Y. J. Kim, S. Briceno, D. Mavris, Sun Choi et. al [5] presented a Prediction Of Weather induced Airline Delays Based OnMachine Learning Algorithms. The primary goal of the model proposed in this paper is to predict airline delays caused by inclement weather conditions using data mining and supervised machine learning algorithms. US domestic flight data and the weather data from 2005 to 2015 were extracted and used to train the model. To overcome the effects of imbalanced training data, sampling techniques are applied. Decision trees, random forest, the AdaBoost and the k Nearest Neighbors were implemented to build models which can predict delays of individual flights. Then, each of the algorithms' prediction accuracy and the receiver operating characteristic (ROC) curve were compared. In the prediction step, flight schedule and weather forecast were gathered and fed into the model. Using those data, the trained model performed a binary classification to predicted whether a scheduled flight will be delayed or on-time.

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

S. Sharma, H. Sangoi, R. Raut, V. C. Kotak, S. Oza et. al [6] presented a Flight Delay Prediction System Using Weighted Multiple Linear Regression. Flight delays hurt airlines, airports, and passengers. Their prediction is crucial during the decision-making process for all players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became cumbersome due to the complexity of air transportation system, the number of methods for prediction, and the deluge of flight data. In this context, this paper presents a thorough literature review of approaches used to build flight delay prediction models from the Data Science perspective. We propose a taxonomy and summarize the initiatives used to address the flight delay prediction problem, according to scope, data, and computational methods, giving particular attention to an increased usage of machine learning methods. Besides, we also present a timeline of significant works that depicts relationships between flight delay prediction problems and research trends to address them.

E. Cinar, F. Aybek, A. Caycar, C. Cetek et. al [7] recommended Capacity and Delay Analysis for Airport manoeuvring areas using Simulation. To investigate the air traffic flow in a highly complex system such as an airport manoeuvring area, a two-stage method based on fast- and real-time simulation techniques is applied. The first stage involves the analysis with fast- and real-time simulations of a baseline model created to determine the congestion points. Based on the analysis, improvements to be performed in the layout of the manoeuvring area are proposed. In the second stage, alternative scenarios implementing these improvements are generated and evaluated in a fasttime simulation environment. Based on the results of simulations of different runway configurations, the main areas of congestion in the baseline airport model are determined. Congestion nodes are identified in the departure queue points and in the taxiway system. To mitigate congestion at these points, three alternative models comprising taxiway and fast-exit taxiway reconfigurations are tested using the fast-time simulation technique. The alternative solution found to be the best in these tests is selected for further testing in real-time simulations. It is shown that the solution would result in an increase in the number of hourly operations and a significant decrease in total ground delays. When conducting the studies needed to identify congestion and design improvements, simulation techniques save both expense and time. Although fast-time simulations are usually adequate for identifying solutions, when critical configurations for the airport are considered, it is shown that it is necessary to also test the results of the fast-time simulations in real-time simulations. The effects of meteorological events, such as rain, fog and snow, etc.are ignored in the simulations. Ground movements in manoeuvring areas are significantly affected by the runways used. Consequently, to enable a comprehensive evaluation in the study, three alternative runway use scenarios are examined. This study utilizes a combination of fast- and real-time simulation techniques to identify the points where congestion occurs in the manoeuvring areas of large-scale airports and to find solutions to minimize the congestion. This approach attempts to combine advantages of both techniques while reducing their short comings. No study is found in the literature using both of these techniques together for the capacity analysis of airport manoeuvring areas.

III. PROPOSED SYSTEM

The purpose of this project is to analyze a broader scope of factors which may

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

potentially influence the flight delay. It compares several machine learning-based models in designed generalized flight delay prediction tasks. We have used Regression algorithms like Decision Tree Regression, Ridge Regression, Logistic Regression, Random Forest Regression, Gradient Boosting Regression to predict flight departure delay and then model evaluation is done to get the best model and our model can identify which features are more important when predicting flight delays. We have implemented these algorithms to predict the best algorithm which is suitable for our dataset. Based on our Project Random Forest Regression seems to fit our model and can predict the delay of 80%. Previously used Supervised learning algorithms managed to achieve 77% of accuracy, we have improved the accuracy rate through regression technique.

IV. SOFTWARE DESIGN

4.1 System Architecture:

The Architecture of the interface given below shows how the interface works. It shows the detailed view of project.

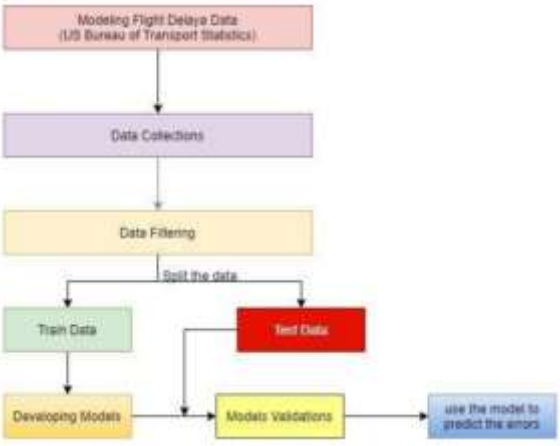

Fig.1: System Architecture

4.2 Data Flow Diagram:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The Data Flow Diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

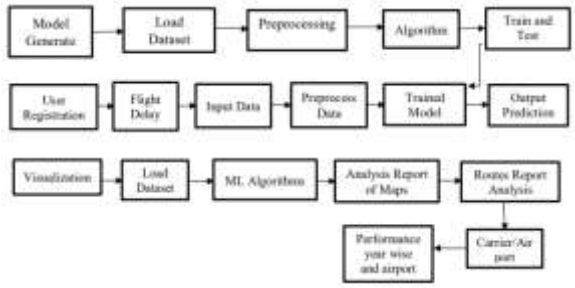

Fig.2: Data Flow Diagram

4.3 UML Diagrams:

UML stands for Unified Modeling Language. UML is a standardized general purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

Goals: The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

4.3.1 Use Case Diagram: A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what System functions are performed for which actor. Roles of the actors in the system can be depicted.

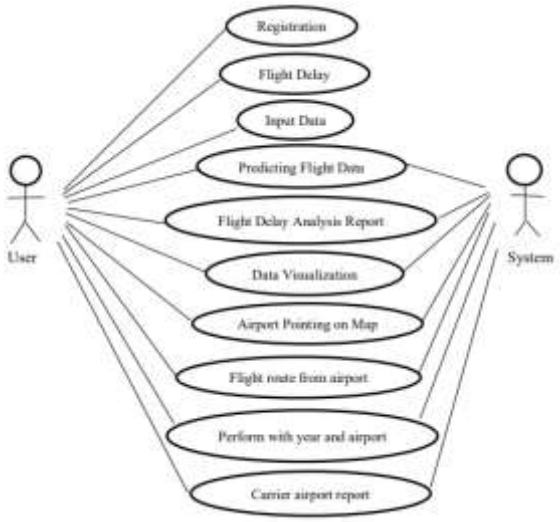

Fig.3: Use Case Diagram

4.3.2 Class Diagram: In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

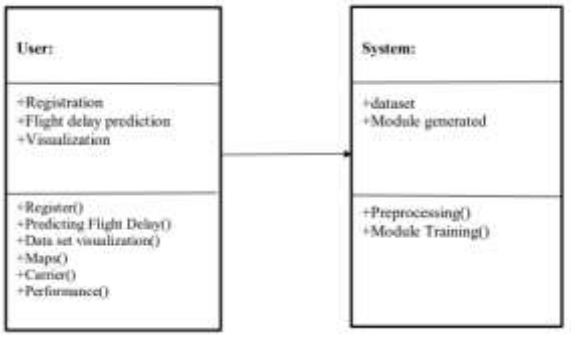

Fig.4: Class Diagram

4.3.3 Sequence Diagram: A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

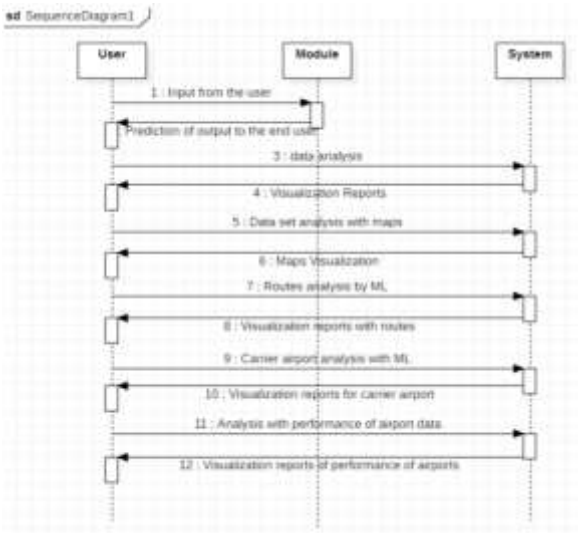

Fig.5: Sequence Diagram

4.3.4 Activity Diagram: Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

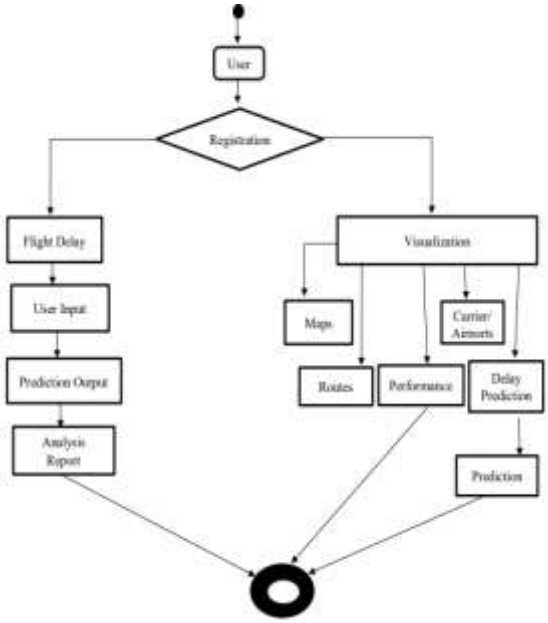

Fig.6: Activity Diagram

V. OUTPUT SCREENS

Fig.7: User Login

Fig.8: User Registration

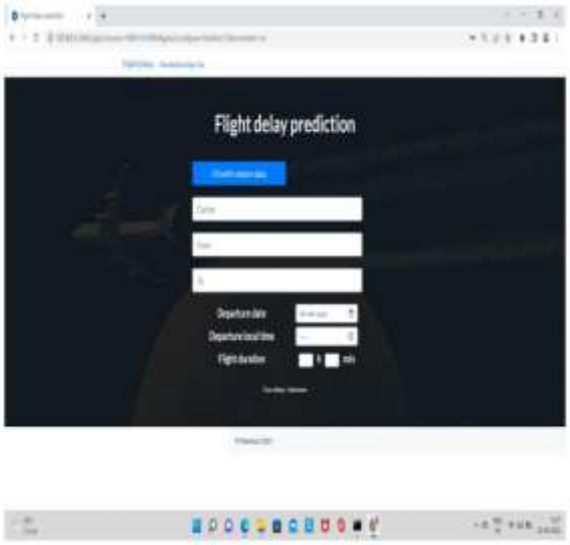

Fig.9: Flight Delay Prediction with Random Data

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

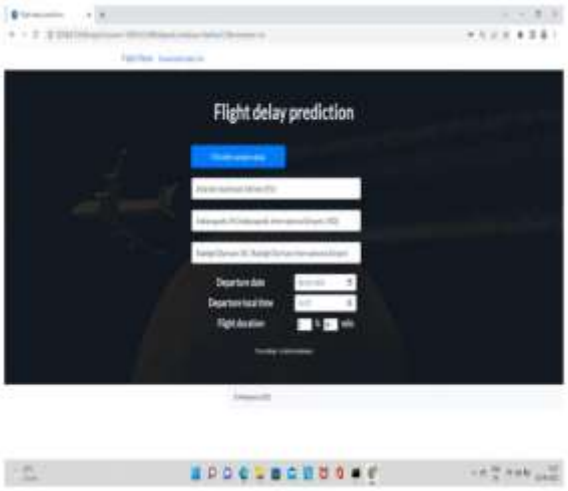

Fig.10: Output of Flight Delay Prediction with Random Data

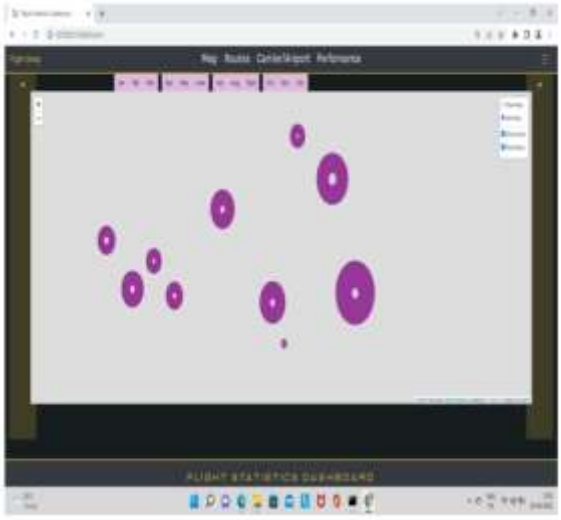

Fig.11: Visualization of Maps

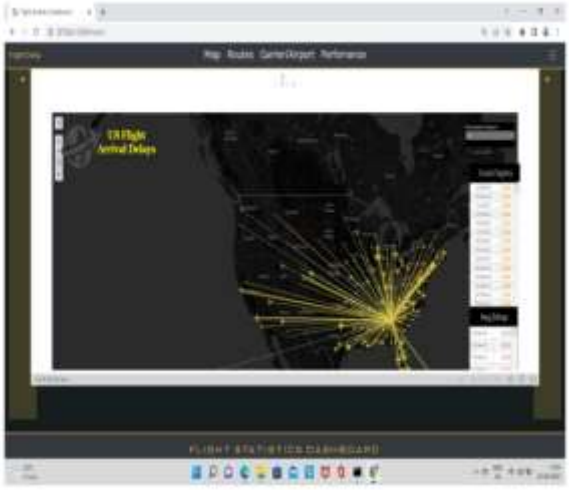

Fig.12: Visualization of Routes

Fig.13: Visualization of Airports and Carriers by Graphs

Fig.14: Visualizing Performance of Flights

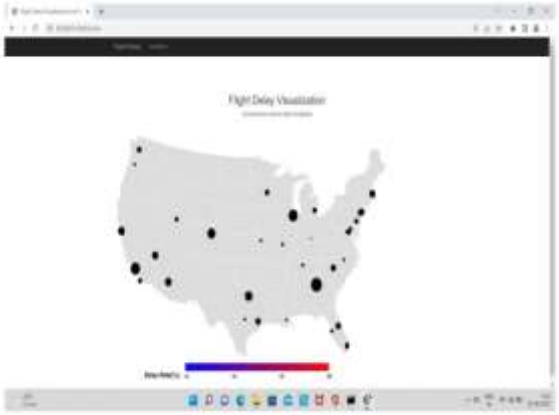

Fig.15: Flight Delay Visualization

ISSN- 2394-5125 VOL 8, ISSUE 05, 2021

Fig.16: Flight Delay Prediction

Fig.17: Flight Delay Predicted Output

VI. CONCLUSION & FUTURE SCOPE

Machine learning algorithms were applied progressively and successively to predict flight arrival & delay. We built five models out of this. We saw for each evaluation metric considered the values of the models and compared them. We found out that In Departure Delay, Random Forest Regressor was observed as the best model with Mean Squared Error 2261.8 and Mean Absolute Error 24.1, which are the minimum value found in these respective metrics. In Arrival Delay, Random Forest Regressor was the best model observed with Mean Squared Error 3019.3 and Mean Absolute Error 30.8, which are the minimum value found in these respective metrics. In the rest of the metrics, the value of the error of Random Forest Regressor although is not minimum but still gives a low value comparatively. In maximum metrics, we found out that Random Forest Regressor gives us the best value and thus should be the model selected.

FUTURE SCOPE:

The future scope of this paper can include the application of more advanced, modern and innovative preprocessing techniques, automated hybrid learning and sampling algorithms and deep learning models adjusted to achieve better performance. To evolve a predictive model, additional variables can be introduced. e.g., a model where meteorological statistics are utilized in developing error-free models for flight delays. In this paper we used data from the US only, therefore in future, the model can be trained with data from other countries as well. With the use of models that are complex and hybrid of many other models provided with appropriate processing power and with the use of larger detailed datasets, more accurate predictive models can be developed. Additionally, the model can be configured for other airports to predict their flight delays as well and for that data from these airports would be required to incorporate into this research.

VII. REFERENCES

[1] Meel, P., Singhal, M., Tanwar, M., & Saini, N. (2020). “Predicting Flight Delays with Error Calculation using Machine Learned Classifiers”, 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN). doi:10.1109/spin48934.2020.9071159

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

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

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

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.