Journal of Engineering Sciences

Vol 12, Issue 7, July/ 2021
ISSN NO:0377-9254

FLIGHT DELAY PREDICTION BASED ON ACTIVATION BIG DATA AND MACHINE LEARNING

[1]M.MOHANA DEEPTHI , [2]KOLUKULURI GOWTHAMI,[3]SHAIK KAMARUNNISA,[4]DEEPTHI SREE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ANDHRA LOYOLA INSTITUTE OF ENGINEERING AND TCHNOLOGY-VIJAYAWADA

ABSTRACT:

flight delays detections are predicted using machine learning based QAR big data approach. It also utilizes the deep learning and machine learning techniques to predict the flight delays effectively with the neural nets. The QAR dataset along with meteorological data is obtained from the kaggle.com website to perform an EDA and spatio-temporal pattern analysis for prediction of flight delays. This QAR dataset set categorized into training and testing records that used to train and test the proposed machine learning based classifier models.

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.

I. INTRODUCTION

- In this present modern world, most significant type of transportation system is the Air transportation system. In the air traffic and passenger traffic, it is essential to sustain resilience and persistence by the increasing congestion.
- In the establishment of airport infrastructures, available land and resources are playing a major contribution. Maintaining the efficiency, safety and capacity are the basic idea to enhance the technology and procedures.
- This enhancing results in some of the environmental effects and hence the National Airspace System (NAS) is mainly intense to minimize these environmental effects. The present using technology offers the visualization of flight path, heading, altitude and further associated parameters to the passengers about their flights throughout the journey.

II. LITERATURE SURVEY

Taking into account of taxanomy of the flight delay & it's complications, one will scrutinize the scope of prediction.The replica originated during this system may be solicited to forecast the prevalence of flight delay at airports.This proceeding can be reduced by emerging the flight delay prediction apparatus which can be developed.

.

III. PROPOSED SYSTEM

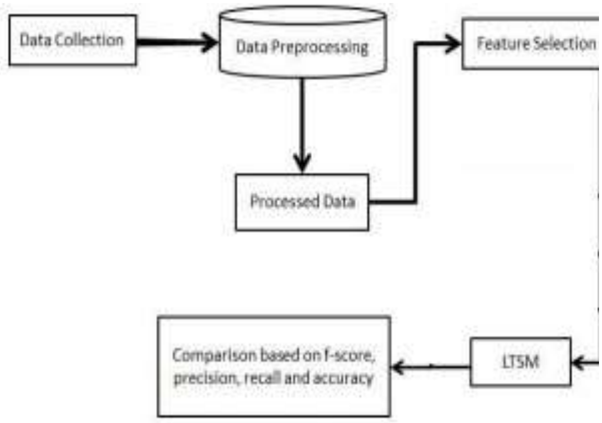
- As we all know that weather conditions are the major problems for delaying the flights.In this paper we mainly focuses on the weather conditions which causes delays in flight.
- For this prediction we are using different kinds of machine learning algorithms like random forest, k-nearest neighbours , svm, decision trees, logistic regression are used.By using these algorithms we can find out the accuracy results for our project.
- The accuracy results can also be find by using various classifications. By adding those we can improve our accuracy.
- All the algorithms which we are used is supervised machine

Vol 12, Issue 7, July/ 2021
ISSN NO:0377-9254

learning approaches .It means the supervised machine learning must contain the set of correct labelled data. To know the real time prediction we can use naive bayed model.

- Here, we are using the CSV file which can store flat file formats and can make easy to add, access and edit the data for calculations.Overall research is to predict the delays to minimize loses.

IV. SYSTEM ARCHITECTURE



V. DATA PREPROCESSING

- Before training the data, the data needs to be preprocessed so as to avoid any errors later . The data has been preprocessed using various Python programming and its various libraries. The techniques employed are as follows:
- Ignore Tuples With Null Values: In the beginning, the dataset contained 4 tuples with all values as „NaN‟, so these tuples had to be dropped since they were of no use in the analysis.
-
- Handling Missing Values: There were a few missing values in the „ArrDelay‟ column. This could be easily replaced by the mean of the delay of that column. But there were tuples with a missing value in the „Dest‟ column, so those columns had to be dropped.
-
- Dropping Irrelevant Attributes: Most of the attributes are relevant but not all are requires so some of them were dropped.
- ♣ „Year‟ has no relevance in our dataset since the entire dataset is for the same year.
- ♣ „Quarter‟ and „FlightDate‟ are also repetitive features so we have dropped those. Similarly, „Origin‟ and „OriginCityName‟ and „Dest‟ and „DestCityName‟ are also repetitive and have been dropped. Creating Dummy Variables: This involves converting the categorical variable into dummy/indicator variable. For a categorical variable that takes on more than one value, a dummy variable is created for each unique value that the categorical variable takes on. So now all the categorical data is converted into dummy variables with values 0 and 1 (0 if not present and 1 if present).
- 5. Finding Delay for each Tuple: After getting a cleaned dataset, we calculated the column „Delayed‟ and value 0 or 1, depending on the delay.

If the „ArrDelay‟ is less than 10 minutes, then we assign 0, which means the flight arrived on time. If the flight is delayed more than 10 minutes, then we assign 1, which indicates that the flight is delayed.

```
1: Set ip units, lstm units, op units and optimizer to define LSTM Network (L)

2: Normalize the dataset (Di) into values from 0 to 1 using
$Xnorm = X − min(X) / max(X) − min(X)$

3: Select training window size (tw) and organize Di accordingly

4: for n epochs and batch size do
5: Train the Network (L)
6: end for
7: Run Predictions using L
8: Calculate the loss function using (7)
```

DELAY PREDICTION MODELS

The flight delay prediction models have been significantly improved since from the early 1990’s as flight delay causes many issues over the world. The market strategy quality is reduced based on the delay amount of flights. The operations of international flights are affected by the arrival and/or departure delays of domestic flights. In an airport sector, a little amount of change in delay can lead to a huge amount of success. The availability of airline data is the first and most important thing. The Federal Aviation Administration (FAA) and other boards of aviation authorities openly lease the information because of the flooding of data in order to facilitate the researches throughout the world in a huge amount. The data regarding to the air traffic is analyzed using the deep Recurrent Neural Networks (RNN) proposed in [2]. Here, flight delays are predicted using the RNN architecture of long short term memory. Also it discusses four different types of deep RNN in delay prediction

The National Climate Data Center (NCDC) is helpful for collecting the weather data. In addition, other details on flight delays were obtained from Tran Statistics of [7]. The Airport Id used by FAA, longitude and latitude information, place, city and state of airports has been considered as the major factors. In addition to these factors, the following factors such as carrier, equivalent rain water, minimum and maximum temperatures are also play an important role. The Neural network, Decision tree and Regression models were tested but Logistic regression model is proved as a best model based on the Misclassification rate. Raj Bandyopadhyay in [8] have grabbed data from Bureau of Transportation Statistics for a subset of commercial flights in the United States regional goals in mind were:

- Identifying the factors causing the delay
- Delay prediction of individual delays
- Estimating amount of flight delay

predicting the changing errors related to the size of training dataset, the primary analysis of linear regression model was performed.

features which play most significant role in causing of flight delay are derived.

Following are those important features.

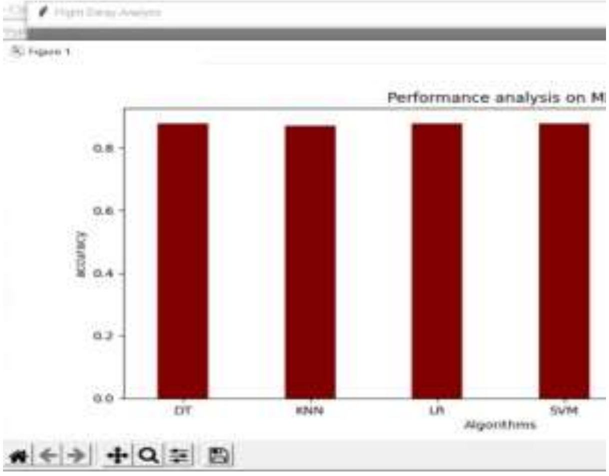
1. Distance
2. Departure Time
3. Arrival Time
4. Elapsed Time
5. Day of the Week (DOW)
6. Total Delay Time

- There is an immense of importance for every one of these attributes. That means, as airline delays are almost occurred on the Mondays, the feature of Day of Week (DOW) must necessary.
- The interval between the actual and scheduled times gives the features of Departure Time, Arrival Time and Elapsed Time.
- Similarly Total Delay Time is also a derived feature that means sum of Security, TaxiIn and TaxiOut and Weather Delays.
- In addition, dataset contains 10 lakh examples in it. Among those 10 lakh of example of records, 8 lakh records are used to train the proposed models where as the remaining 2 lakh examples of records are used to test those trained models. The satisfied sampling random method can be utilized to select the trained data set because of processing power limitation of the system which conducts the experiment.
- The complete dataset is divided into various stratums. Training examples of dataset from every one of the stratum are randomly selected and are grouped as a one training dataset.
- It is a non-biased representative of the complete dataset. From this analysis there is a reduction of training records from 8 into 2 lakhs of examples and also there is a reduction in testing records from 2 to 50000 examples.
- The actual representation of complete dataset and reduction in selecting sample bias are the main advantage of stratified sampling. In a dataset different categories or different classes are represented by different stratum. The training records are reduced to two lakh examples by twice using the concept of stratified sampling.

VI. RESULTS AND OBSERVATIONS:-

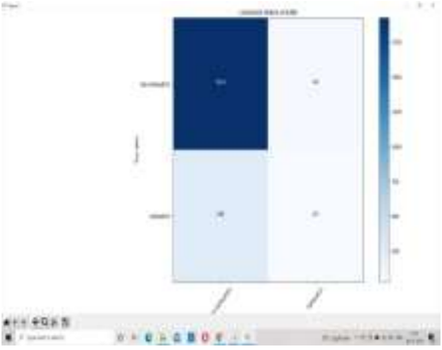
The delays in flight may cause many problems to passengers and airlines, recognising them through prediction may improve marketing decisions.

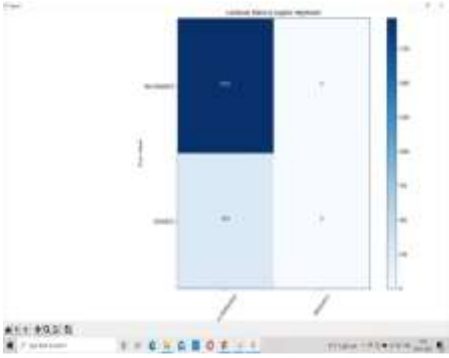
For this,several models have to sought to understand how delays occur and predicate the root delay or comprehend the cancellation process.



Results:- This is the gui for our project here, first we need to click on perform ml then it will import all the ml files

- It will displays the predicated delayed flights data with respective to the seconds.
- In this we can see the training sample size and test sample size.
- It will also give the performance analysis on the algorithms like decision tree, KNN, logistic regression ,SVM, random forest, XGboost and it will gives the confusion matrix.

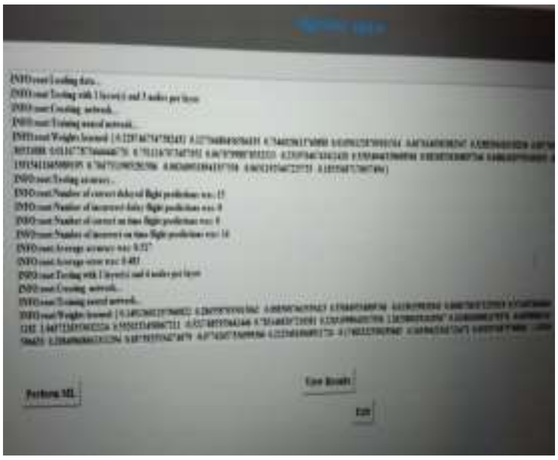




- Confusion matrix which gives the no. of correct and incorrect predictions with count values by each class.
- Here, we get the various confusion matrix such as KNN, decision tree ,SVM,logistic regression

This are the visualization of the accuracy results.

- Finally,we can click on the view results which was displayed on gui window then it will show the no. of flights delayed.



```
INFO:root:Loading data...
INFO:root:Testing with 1 layer(s) and 3 nodes per layer
INFO:root:Creating network...
INFO:root:Training neural network...
INFO:root:Weights learned: [illegible]
INFO:root:Testing accuracy...
INFO:root:Number of correct delayed flight predictions was: 15
INFO:root:Number of incorrect delay flight predictions was: 8
INFO:root:Number of correct on time flight predictions was: 8
INFO:root:Number of incorrect on time flight predictions was: 14
INFO:root:Average accuracy was: 0.57
INFO:root:Average error was: 0.483
INFO:root:Testing with 1 layer(s) and 4 nodes per layer
INFO:root:Creating network...
INFO:root:Training neural network...
INFO:root:Weights learned: [illegible]
```

VII. CONCLUSION

1.Flight delay is an important subject in the literature due to their economic & environmental impacts.

2. They may increase costs to customers and operational costs to airlines.

3. Apart from outcomes directly added to Passengers, delay prediction is crucial during the decision-making process for every player in air transportation system.

4.We developed a taxanomy scheme and classified models in respect of detailed components.

VIII. FUTURE SCOPE:

1.The scope of this project is very much confined to flight & weather data of US, but we can include more countries like India,China & Russia.

2.Expanding the scope of this project , we can also add the flight data from international flights & not just restrict ourself to domestic flights.

IX. ACKNOWLEDGEMENT

This research project was partially supported by the Department of Computer Science and Engineering, Andhra Loyola Institute of Engineering and Technology, Jawaharlal Nehru Technological University. We are grateful to Associate Professor

M.MOHAN DEEPTHI for leading us to develop and contribute a paper to the conference.

X. REFERENCES

- BTS. bureau of transportation statistics databases technical report http://www.rita.dot.gov/bts/home
- Mofokeng TJ, Marne wick A. Factors contributing to delays regarding aircraft during A-check maintenance.
- Yang C, Marshall ZA, Mott JH. A novel integration platform to reduce flight delays in the National Airspace System. In 2020 Systems and Information Engineering Design Symposium (SIEDS).