IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

MACHINE LEARNING APPROACH TO PREDICT FLIGHT DELAYS

K.Ebenezer[1*], K.N. Brahmaji Rao[2]

[1,2]Dept. of Computer Science and Systems Engineering Andhra university college of engineering, Andhra University, Visakhapatnam, AP, India

Available online at: www.ijcseonline.org

Accepted: 20/Oct/2018, Published: 31/Oct/2018

ABSTRACT - Air transport provides efficient, well organized, and time effective services. Even though flights are the fastest way to transport, its delay leads to customer dissatisfaction. Many factors effect flight delays, some of them are weather, operational imperfection, baggage loading etc. In this paper, we are developing a predictive system which predicts flight delays based on weather data. Flightdata set taken from US DEPARTMENT OF TRANSPORTATION and weather data set from HOURLY LAND-BASED WEATHER OBSERVATIONS FROM NOAA. We have implemented Ensemble method, Decision tree and Random forest on the balanced data set. For balancing the data set we are using sampling techniques. The algorithms are applied on the combined flight and weather data set to predict the flight delays.

***Keywords*: ** Flights, Weather, Random Forest, Decision Tree, Ensemble

I. INTRODUCTION

Now a days, Airlines are playing crucial role in fastest mode of transport. But when flights get delayed, it not only effects customer convenience but also it certainly effects Flight Company's reputation. Flight companies have to pay huge compensation to the customers as they have to obey certain rules. In order to avoid all those losses, it is a must to predict flight delays. There have been introduced many predictive models to predict flight delays. The Federal Aviation System (FAA) decides a flight is delayed only when the flight is late by 15 minutes than its programmed time. Supervised Machine Learning algorithms are used to predict arrival flight delays. By using Machine learning it is capable to improve models with bulk amounts of datasets like flight dataset and weather dataset. Random Forest, Ada Boost, K-N-Neighbors, Decision Tree are applied to construct models to predict whether the flight will be delayed or not. Flight data and weather data are merged and fed in the model. By using this data, constructed model accomplished a binary classification to predict flight delays[1].

Gradient Boosted Decision Tree is applied to predict whether the flight will be delayed or not. This model attained the great coefficient of determination of above 92% in case of flight arrivals and above 94% in case of flight departures [2]. Artificial neural networks techniques are used for the benefit of application where the prediction model is built to prognosticate the flight delays. A type of ANN structure DMP-ANN is found which is worthy enough to predict delays. This is used to predict the flight delays with less mean square root error [3]. A two-stage predictive model is evolved making use of supervised machine learning algorithms to speculate the flight delays. In 1st stage binary classification is performed to predict the flight delays and in the 2nd stage regression is done to predict the value of delay in minutes. By using Gradient Boosting technique, accuracy of departure delay prediction is 84% and the accuracy of arrival delay prediction is 94% [4]. Early warning grading standards formed by the combination of flight operation and the data of flight delays. First of all we get the number of flights delayed on regular intervals of time from 7am to 10 pm. The section which contains the number of flights delayed in regular intervals of time as large emergence probability is treated as common condition, while the number of flight delayed in regular intervals of time as minimum emergence probability is treated as danger condition and then it is necessary to do early warning [6].

After surveying the composite and undetermined association between the flight delays and practicable influence components, Bayesian network change to replicate and determine about flight delay in busy hub-airport. Two learning methods were implemented with three models. 1st model is an approximation for arrival delay based on framework learning Bayesian network with Expectation-Maximization algorithm. 2nd model is an approximation for arrival delay based on construction learning of Bayesian network with K2 algorithm. The model well read by K2 is demonstrated more acceptable for modeling the estimation of flight delay, with a better approximation rate than 1st model [9]. A type of Bayesian Network Structure Learning algorithm called Target-fixed Stochastic-ordered which is used to build a predictive model to predict the flight delay [10].. A model established on the genetic algorithm to get the most out of the flight delays to lower significant air traffic flight delays. It holds two instances. First instance is to convey the delays to a number of delayed flights to stay away from delay, which can expand the flight well timed rate. Another instance is to pass around the delay losses to heterogeneous airlines to protect the civility and equilibrium the delay loss of the air service and travelers [12]. A prediction model of the flight delay generation is introduced in which the disapproving flight resources and the unfavorable airfield resources are examined to furnish a more productive technique for the prediction of flight delay generation. Simulation exhibit that the model and algorithm supply a constructive method for measuring prediction of flight delay transmission [13]. A new technique is introduced which is based on content based recommendation system. According to the transmission of the delay, this new technique vigilance the target airport by observing the status of corresponding airports. The discovered status is balanced with the previous data in order to predict the solemnity of the delay [14].

In this paper, we are proposing Ensemble method to improve the prediction flight delays.

II. PROPOSED METHODOLOGY

A. S*ampling Technique:* OVERSAMPLING

Oversampling is the technique used in this model to modify the class diffusion of a data set. The dataset is not balanced when the classification groupings are not equally constituted. To balance the dataset we have applied oversampling. Performance of classifiers is just made better by using this oversampling. Minority and majority class are balanced using this sampling technique.

B. A*lgorithms Implemented*

Ensemble method, Random forest algorithm and Decision tree algorithms are utilized to advance the predictive model for flight delay survey. Classification based approach is applied in this model.

1. Ensemble Method

Data split samples are drawn in selecting variables from the complete training set alternative of bootstrap sample of instruction set. From range of values, splits are chosen absolutely at random. Extra tree classifier is normally economical to train from a measurement point of view but can improve much larger. Extra tree classifier can some time conclude superior to Random forest. In this paper, this method is taken to show good accuracy in predicting flight delays.

Algorithm:

```
  Split a node(S)
   Input: the local learning subset S corresponding to the
node we want to split
  Output: a split [a < ac] or nothing
    – If Stop split(S) is TRUE then return nothing.
– Otherwise select K attributes {a1,..., $a_K$ } among all
non constant (in S) candidate attributes;
– Draw K splits {s1,...,$s_K$ }, where $s_i$ = Pick a random
split(S, $a_i$), $\forall_i$ = 1,..., K;
– Return a split s∗ such that Score(s∗, S) = $\max_i$=1,...,K
Score($s_i$, S).

 Pick a random split(S,a)
  Inputs: a subset S and an attribute a
  Output: a split
   – Let $a^S_{max}$ and $a^S_{min}$ denote the maximal and minimal
value of a in S;
   – Draw a random cut-point ac uniformly in [$a^S_{min}$,  $a^S_{max}$];
   – Return the split [a < ac].
 Stop split(S)
  Input: a subset S
  Output: a boolean
  – If |S| < nmin, then return TRUE;
   – If all attributes are constant in S, then return TRUE;
– If the output is constant in S, then return TRUE;
   – Otherwise, return FALSE.
```

2. Random Forest Algorithm

Random Forest Algorithm is a supervised classification algorithm. It can be used for both classification and regression problems. It is used for handling the missing values. It is used to model the categorical values. We have implemented this algorithm on balanced dataset to get good accuracy.

3. Decision Tree

Decision tree is a popular tool in machine learning. It is one of the types of Supervised Machine Learning where the data is constant split to a definite framework. Decision nodes and leaves are existed. These leaves are the final outcomes and the decision nodes are the place where data is spitted.

III. DATA ANALYSIS

A. Data Set Description

Flight on time performance Transportation Statistics collection of data from U.S. Department of Transportation. And weather data collected from Hourly land based weather observations taken from NOAA. And these two are combined together and processed using predictive models to predict flight delays. This flight data set contains data of 70 airports in United States. It is taken that flight is considered to be delayed only if it is delayed by more than fifteen minutes only. Also the flights which are diverted are eliminated from the flight data set. In the flight data set we have taken column headers like month, year, day of week, carrier, day of month, origin and destination airport id, departure delays, and arrival delays and finally cancelled. In weather data set we have taken columns like year, adjusted month, adjusted delay, adjusted hour, time zone, visibility, dry bulb Fahrenheit, dry bulb Celsius, dew point Fahrenheit, dew point Celsius, relative humidity, wind speed. After taking these two data sets unnecessary columns are dropped from both flight data and weather data. In weather data, date column is spitted into year, month and day columns.

B. Data Preprocessing

In preprocessing redundant attributes are removed from data sets and column names are renamed as the data sets are having some column headers as same names. Numerical values like origin airport id, destination airport id are converted to categorical values as they are not actually numerical values built they represent some identity. And later the columns which are not necessary in this prediction process are removed. Later the two tables are merged which is necessary for the prediction model. Extra tree classifier is applied on the obtained data set. Random forest and decision tree algorithms are also applied to improve the performance when compared to previous papers. Later the categorical values are gained converted into numerical as machine learning algorithms disclose good performance by with numerical variables only.

IV. RESULT

Table. 1 Metric values of applied algorithms

MODEL	ACCURACY	PRECISION	RECALL
Ensemble	97.83	95.04	95.34
Decision Tree	95.62	86.22	95.82
Random Forest	96.89	91.78	94.64

ROC curves

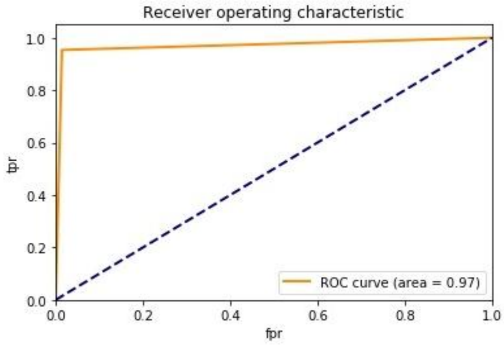

Fig.1.Ensemble

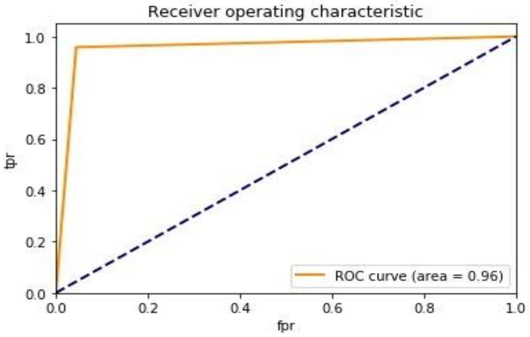

Fig.2. Random Forest

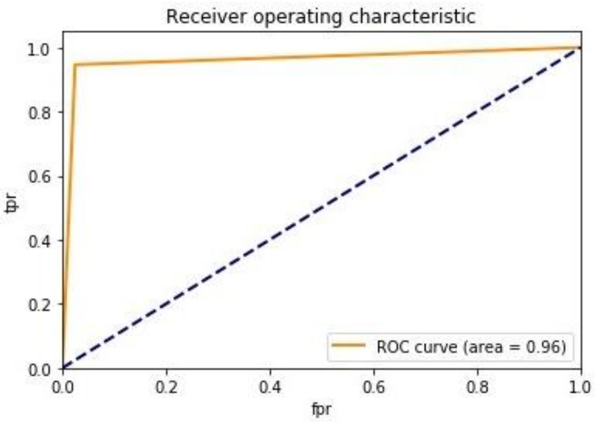

Fig.3. Decision Tree

V. CONCLUSION AND FUTURE EXTENSION

In this paper, a prediction model is permitted to classify flight delays influenced by bleak weather scenario. A model is constructed on the data sets of both flight delays and weather data sets and sampling technique is applied to balance the data. Extra tree classifier, Decision tree and Random forest algorithms are applied on the balanced data to predict flight delays with better accuracy.

Flight delay prediction can be useful to many people if its prediction is more accurate, which can be attained by applying Deep Neural Network based algorithms like CNN, RNN etc.

REFERENCES

[1] Sun Choi, Young Jin Kim, Simon Briceno and Dimitri Mavris, "*Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms*, "978-1-5090-2523-7/16/2016 IEEE

[2] Suvojit Manna, Sanket Biswas, Riyanka Kundu Somnath Rakshit, Priti Gupta and Subhas Barman,"*A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree*," in International Conference on Computational Intelligence in Data Science (ICCIDS), 978-1-5090-5595-1/17/2017 IEEE

[3] Sina Khanmohammadi, Salih Tutun, Yunus Kucuk, *"A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport, "* in Conference Organized by Missouri University of Science and Technology-Los Angeles, CA, Sina Khanmohammadi et al. / Procedia Computer Science 95 (2016) 237 – 244

[4] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, Vineeth Vijayaraghavan, "*A Machine Learning Approach for Prediction of On-time Performance of Flights",* 978-1-5386-0365-9/17/2017 IEEE

[5] Rong Yao, Wang Jiandong, Xu Tao , "*A Flight Delay Prediction Model with consideration of Cross-Flight Plan Awaiting Resources, "*978-1-4244-5848-6/10/2010 IEEE

[6] Guansheng Tong, Jianli Ding " *Real time Sub time Early Warning of Airport Scheduled Flight Delay Based on Immune Algorithm*, " in Second International Symposium on Intelligent Information Technology Application, 978-0-7695-3497-8/08/2008 IEEE

[7] Ding Jianli , Yu Yuecheng , Wang Jiandong, "*A Model for Predicting Flight Delay and Delay Propagation Based on Parallel Cellular Automata*, " in ISECS International Colloquium on Computing, Communication, Control, and Management , 978-1-4244-4246-1/09/2009 IEEE

[8] Jianli Ding , Xuesen Li ,Guansheng Tong , "*The dynamic immune forecasting method of the airdrome flight delay under considering the stochastic factors*, " in Ninth International Conference on Hybrid Intelligent Systems, 978-0-7695-3745-0/09/2009 IEEE

[9] Yujie Liu, Song Ma," *The Multimode Estimation Modeling for Flight Delay of a Busy Hub-Airport in Flight Chain*," in IITA International Conference on Services Science, Management and Engineering, 978-0-7695-3729-0/09/2009 IEEE

[10] Yu-Jie Liu, Fan Yang , "*Initial Flight Delay Modeling and Estimating Based on an Improved Bayesian Network Structure Learning Algorithm*, " in Fifth International Conference on Natural Computation, 978-0-7695-3736-8/09/2009 IEEE

[11] Yujie Liu, Song Ma, "*Modeling and Estimating for Flight Delay Propagation in a Reduced Flight Chain Based on a Mixed Learning Method*," in International Symposium on Knowledge Acquisition and Modeling, 978-0-7695-3488-6/08/2008 IEEE

[12] Zhiwei Xing, Yunxiao Tang, "The *model for optimizing airport flight delays allocation,*" in 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, 978-1-5090-0768-4/16/2016 IEEE

[13] Rong Yao, Wang Jiandong, "*Prediction Model and Algorithm of Flight Delay Propagation Based on Integrated Consideration of Critical Flight Resources,*" in ISECS International Colloquium on Computing, Communication, Control, and Management, 978-1-4244-4246-1/09/2009 IEEE

[14] Lu Zonglei, Wang Jiandong, Xu Tao, "*A New Method for Flight Delays Forecast Based on the Recommendation System*, "in ISECS International Colloquium on Computing, Communication, Control, and Management, 978-1-4244-4246-1/09/2009 IEEE

Authors Profile

Mr. K.Ebenezer done his Bachelor of Technology in computer science from affiliated college of Jawaharlal Nehru Technological University Kakinada, in 2016 and pursuing Master of Technology in computer science from Andhra University Visakhapatnam in 2018.

Mr. K.N.Brahmaji Rao obtained his M.Sc. in Mathematics from Andhra University, M.Phil in Mathematics from Madhurai Kamaraj University, M.Tech in Computer Science and Technology with specialization in Artificial intelligence and Robotics, Andhra University. He is Pursuing Ph.D. in Computer Science and Engineering, Andhra University. He is currently working as guest faculty in the department of Computer Science & Systems Engineering, AUCE (A), Andhra University since 2016. His main research work focuses on Text Based Mining and Machine Learning. He has 18 years of teaching experience.