International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

FLIGHT DELAY PREDICTION USING SUPERVISED LEARNING

SP Lakshmi Narayanan[1], S Sharbini[2], R Priyanka[3], R Kamalendran[4], Mohit kumar[5], C Murale[*]

[1,2,3,4,5]*Student, Department of IT, Coimbatore Institute of Technology, Tamilnadu, India*
[*]*Assistant Professor, Department of IT, Coimbatore Institute of Technology, Tamilnadu, India*

Abstract - *Flight delay is a significant problem that negatively impacts the aviation industry and costs billions of dollars each year. Most existing studies investigated this issue using various methods based on applying machine learning methods to predict the flight delay. However, due to the highly dynamic environments of the aviation industry, relying only on single route of airport may not be sufficient and applicable to forecast the future of flights. The purpose of this project is to analyze a broader scope of factors which may potentially influence the flight delay it compares several machine learning-based models in designed generalized flight delay prediction tasks. In this project we have used flight delay dataset from US Department of Transportation (DOT) to predict flight delays. We have used supervised learning algorithms to predict flight departure delay and then model evaluation is done to get best model and our model can identify which features were more important when predicting flight delays.*

***Key Words*: departure delay; supervised learing; ensemble learning; prediction**

1. INTRODUCTION

A flight delay is said to occur when an airline lands or takes off later than its scheduled arrival or departure time respectively. Conventionally if a flight's departure time or arrival time is greater than 15 minutes than its scheduled departure and arrival times respectively, then it is considered that there is a departure or arrival delay with respect to corresponding airports. Notable reasons for commercially scheduled flights to delay are adverse weather conditions, air traffic congestion, late reaching aircraft to be used for the flight from previous flight, maintenance and security issues. One of the key business issues that airlines face is that the vital prices that are related to flights being delayed because of natural occurrences and operational shortcomings that is an upscale affair for the airlines, making issues in scheduling and operations for the end users therefore inflicting unhealthy name and client discontent. As we all know that we have a tendency to not get the flight delay before departure as customers of the Airline Company neither the airline company ground staff gets the airline delay prediction supported varied conditions. We can use it to predict if a flight carrier will have a departure delay and hence try to avoid that from happening. This will prevent the customers as well as the airlines to avoid any losses, whether in their time or business.

1.1. FLIGHT DELAY PREDICTION

Flight delays lead to negative impacts, mainly economical for commuters, airline industries and airport authorities. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions. Hence, these factors indicate how necessary and relevant it has become to predict the delays no matter the wide-range of airline meshes. To carry out the predictive analysis, which encompasses a range of statistical techniques from supervised machine learning and, data mining, that studies current and historical data to make predictions or just analyze about the future delays, with help of Regression Analysis using regularization technique in Python 3.This prediction will be helpful for giving a detailed analysis of the performance of individual airlines, airports, and then making a well-assessed decision. Moreover, apart from the assessment related to the passengers, delay prediction analysis will also help in important decision-making procedures necessary for every pivotal player in the air transportation system.

1.2. PREDICTION OF FLIGHT DELAY USING SUPERVISED LEARNING ALGORITHM

Flight delay is inevitable and it plays an important role in both profits and loss of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and income of airline agencies. A fundamental sub-area of artificial intelligence has come into notice, called as Machine Learning, which enables computers to get into a mode of self-learning without being explicitly programmed. With the concept of machine learning, we have been able to apply complex mathematical computations to big data iteratively and automatically, that too with efficient speed, this phenomenon has been encompassing momentum over the last several years. There have been many researches on modelling and predicting flight delays, where most of them have been trying to predict the delay through extracting important characteristics and most related features. However, most of the proposed methods are not accurate enough because of massive volume data, dependencies and extreme number of parameters. It is a machine learning task where the dataset inputs and outputs are clearly recognized and already given, then several type of algorithms are trained using labelled examples. A supervised learning algorithm contains an

entire dataset, which is further divided into training and test data. The algorithm examines the training dataset and produces an inferred function, which is then used for mapping new examples. In case of the aviation industry, commercialized aviation is a type of transportation system that is complexly distributed. It tends to deal with several important resources, demand fluctuations, and various other kinds of stages. Moreover, apart from the assessment related to the passengers, delay prediction analysis will also help in important decision-making procedures necessary for every pivotal player in the air transportation system.

2. METHODOLOGY

2.1. PROPOSED SYSTEM

This proposed system helps Airline passengers to know whether the flight will get delayed or not. To make the system more scalable it is necessary to choose an algorithm which considers all the parameters to be independent. Supervised learning as the name indicates a presence of supervisor as teacher. Essentially supervised learning could be a learning that within which we tend to teach or train the machine exploitation data which is well tagged which means some data is already labelled with correct answer. After that, machine is given new set of examples(data) so supervised learning algorithm rule analyses the coaching knowledge(set of training examples) and produces an correct outcome from tagged data Using supervised machine learning approach, the labeled data gives it authenticity.

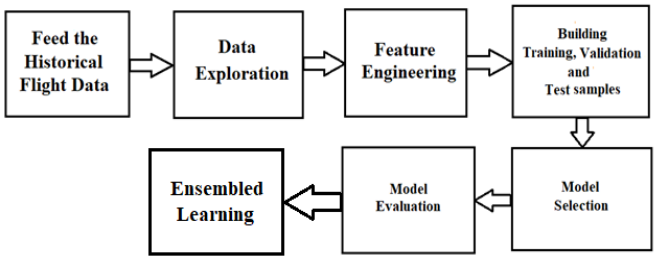

Fig1. Prediction model

2.1.1. DATA EXPLORATION AND FEATURE ENGINEERING

The Dataset is taken from The U.S. Department of Transportation's (DOT). They provide a summary of the arrival delays, departure delays, on-time arrival etc in their monthly report. We can use it to predict if a flight carrier will have a departure delay and hence try to avoid that from happening. This will prevent the customers as well as the airlines to avoid any losses. We have used the flights and airports data set for this project. The flights data set contains 1049000 rows of flights data of 14 different carriers in the year 2015. We have used the first 100000 rows to build my model and predict the delay. I have also used the airport data set to use the origin and destination airport. The most important column in the table is DEPARTURE_DELAY. This will be used to predict if there will be a delay in the departure of any flight or not. The negative values indicate an early departure of the flights and the positive values indicate that there was a delay in the departure of the flights. Other than this column, another column CANCELLED plays a major role. This tells us if the flight was cancelled or not and we can filter out the rows depending on that because in case that happened there won't be a delay in the departure of the flight. The cancelled flights are assigned 1 and the rest are 0. Then we removed unwanted columns that were not necessary for delay prediction. We will break our datset into numerical and categorical data and check for null values. Then we performed one-hot encoding on categorical columns.

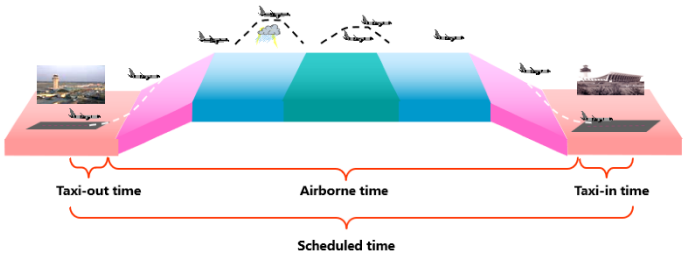

Fig2. Attributes used in dataset

2.1.2. BUILDING TRAINING, VALIDATION AND TEST SETS

- Training: This is used to train our model. The major portion of the data set is assigned as training samples. We have used 70% of data for training.
- Validation: These are separate from the training set and are used to evaluate how our model will perform and hence make changes to improve it. We have used 15% of data for validation.
- Test: These samples are kept separate and after selecting our best model are used for final evaluation. We have used 15% of data for test.

2.1.3. MODEL SELECTION

In this module the models are built and checked its performance on the validation tests and hence select the best model.

The models will be evaluated on different performance metrics as shown below.

- AUC
- Accuracy
- Recall

- Precision
- Specificity

Comparing the performance of the following machine learning algorithms using hyper parameters.

- Logistic regression
- Naive Bayes
- Decision tree
- Random forest
- Gradient boosting classifier

2.1.4. MODEL EVALUATION

In this module the models are selected based on their performance and its performance is checked by using the test set. Using the model, the most import feature that plays an important role in flight delay prediction is identified.

2.1.5. ENSEMBLED LEARNING

Blending is also an ensemble technique that can help us to improve performance and increase accuracy. It follows the same approach as stacking but uses only a holdout (validation) set from the train set to make predictions. In other words, unlike stacking, the predictions are made on the holdout set only. The holdout set and the predictions are used to build a model which is run on the test set. Here is a detailed explanation of the blending process:

- The train set is split into two parts, viz-training and validation sets.
- Model(s) are fit on the training set.
- The predictions are made on the validation set and the test set.
- The validation set and its predictions are used as features to build a new model.
- This model is used to make final predictions on the test and meta-features.

2.2. PREDICTION ALGORITHMS

2.2.1 LOGISTIC REGRESSION

Logistic regression is executed when the dependent variable is binary. It is a predictive analysis technique. It describes data and explains the relationship between one dependent binary variable and one or more independent variables. The model is usually easy to interpret, and we

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

can know which feature is important for us. We need to scale our data before applying this model to it.

2.2.2 STOCHASTIC GRADIENT DESCENT

Stochastic gradient descent and logistic regression are similar in nature. Though they both use coefficient of the liner function, Stochastic Gradient Descent works in small batches hence making the execution much faster than the former. We can use the following code to fit this model from scikit-learn.

2.2.3 NAIVE BAYES

Naive Bayes is a machine learning model that uses the Bayes rule. It is called so because it assumes that all the variables are independent of each other. This is a good assumption in case of natural language processing.

2.2.4 DECISION TREE

Machine learning models can also be tree-based. We will first look at Decision Tree. In this model, we divide samples by splitting them based on a threshold. In each question, you ask if the samples have a specific variable greater than some threshold and then split the samples. The final prediction is the fraction of positive samples in the final split of the tree. We need to figure out what threshold to select for each split. They usually do not have any presumptions and we need to provide sufficient depth for the split.

2.2.5 RANDOM FOREST

Decision trees tend to result in overfitting because they memorize the training data. Random forest helps overcome this disadvantage. Multiple trees are generated, and results are then aggregated. Random forests tend to perform better than decision trees because they can generalize easily.

2.2.6 GRADIENT BOOSTING CLASSIFIER

Gradient boosting classifier is another tree-based method which is used to improve the problem of overfitting in decision trees. A bunch of shallow trees are created, and they optimize the error occurred previously. It is paired with the gradient descent classifier.

3. PERFORMANCE ANALYSIS

3.1 COMPARISION OF ALGORITHMS:

The comparison of models developed by various algorithms is done by following factors

3.1.1 PRECISION

Precision is the ratio between the True Positives and all the Positives.in our model, it is the measure of flights that we correctly identify having a delay out of all the fights actually having it.

3.1.2 RECALL

The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart disease.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

Mathematically:

3.1.3 AUC-ROC CURVE

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s.

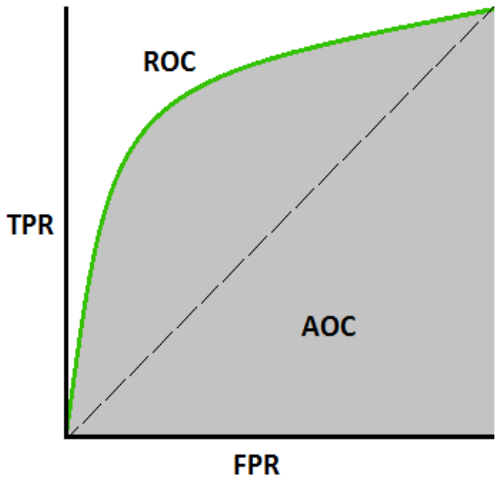

Fig 3 AUC-Roc Curve

3.1.4 ACCURACY

Accuracy is a statistical measure which is defined as the quotient of correct predictions (both True positives (TP) and True negatives (TN)) made by a classifier divided by the sum of all predictions made by the classifier, including False positves (FP) and False negatives (FN).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3.1.5 SPECIFICITY

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negative, which got predicted as positive and could be termed as false positives.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

3.1.6 PREVALANCE

Prevalance is the fraction of positives in the total prediction

$$Prevalence = \frac{\#\ of\ people\ in\ sample\ with\ characteristic}{Total\ \#\ of\ people\ in\ sample}$$

In this project we have used various algorithms to predict flight delay. In that the model developed using Naïve Bayes has the lowest accuracy and the model using logistic regression has the highest accuracy. The models using decision tree and Gradient Boosting classifier has similar accuracy with a minute difference. The models using stochastic gradient descendant has similar accuracy with random forest.

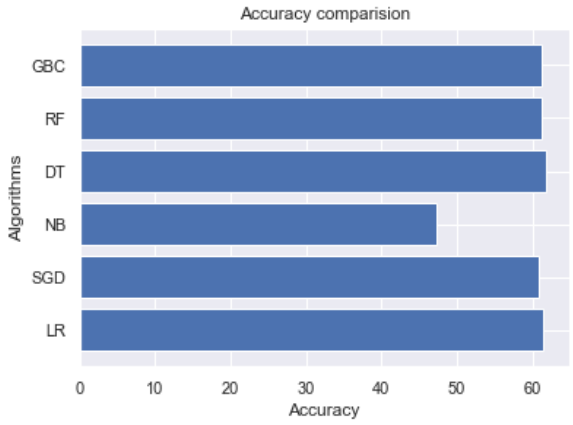

Fig 4. Accuracy comparision

When comparing these models based on training time Gradient boosting classifier takes more time to train than all other algorithms combined as it is based on boosting technique so it needs various iterations to predict. Naïve Bayes has the lowest training time compared to other models. Logistic regression model takes second highest training time as it also involves iterations. The model using Stochastic gradient descendant Takes second lowest training time followed by Decision tree and random Forest.

Based on these comparisons Stochastic Gradient model has the lower training time and better accuracy compared with other models for the dataset used.

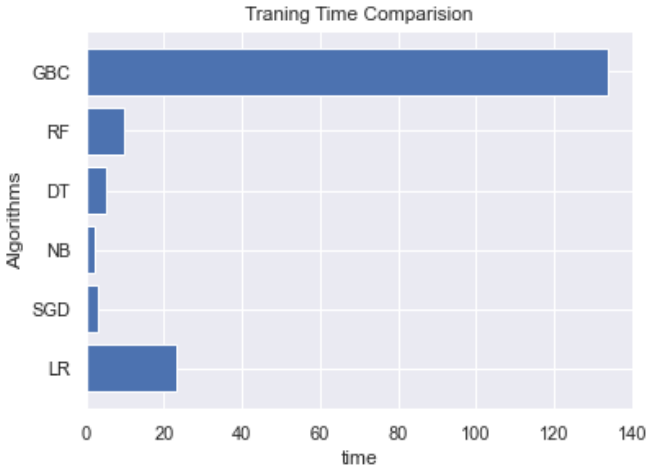

Fig 5. Training time comparision

4. EXPERIMENTAL RESULTS

We have implemented application in number of phases. We would like to describe our results and discuss about results in following manner.

In the Initial phases we perfrmed data exploration, Feature engineering and Train test splitting.

Then we performed the prediction algorithms and the results are analysed using Confusion matric and ROC-AUC curve

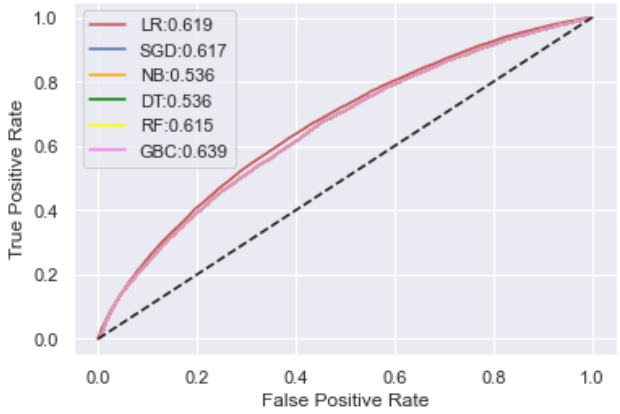

Fig 6. ROC-AUC Curve of all models

Stochastic Gradient Descent

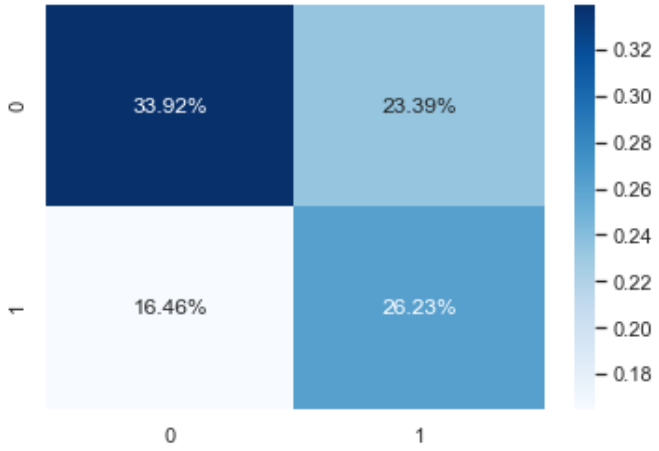

Fig 7. Confusion Matrix stochastic gradient descent

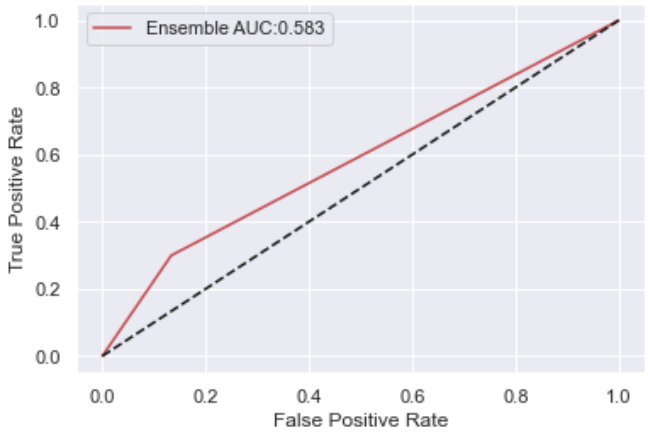

Fig 8. ROC-AUC Curve of all models

From the evaluation we found that Gradient Boosting Classifier gave more accuracy but it takes more time to train. So when comparing the training time and accuracy stochastic gradient descent is the best performing algorithm for the dataset used.

5. CONCLUSION

This paper and the analysis retrieved are useful not only for passengers point of view, but for every decision maker in the aviation industry. Apart from the financial losses incurred by the industry, flight delay also portray a negative reputation of the airlines, and decreases their reliability. It causes various sustainability issues, for example, increase in fuel consumption and gas emissions. The analysis carried here not only predicts delays based on the previous available data, but also give statistical description of airlines, their rankings based on their on-time performance, and delays with respect to time, showing the peak hours of delay. This project can be used as a prototype by any aviation authority for their benefit,

in the Indian Scenario too, it can work as an efficient model or a proper prototype to study delay analysis, based on the real dataset provided.

ACKNOWLEDGEMENT

We are immensely grateful to Mr.Murale C, Assistant Professor, Dept. of Information Technology, Coimbatore Institution of Technology, Coimbatore for his constant guidance and unending support.

REFERENCES

[1] M. Leonardi, "Ads-b anomalies and intrusions detection by sensor clocks tracking," IEEE Trans. Aerosp. Electron. Syst., to be published, doi: 10.1109/TAES.2018.2886616.

[2] Y. A. Nijsure, G. Kaddoum, G. Gagnon, F. Gagnon, C. Yuen, and R. Mahapatra, "Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration," IEEE Trans. Veh. Technol., vol. 65, no. 5, pp. 3150–3165, 2015.

[3] J. A. F. Zuluaga, J. F. V. Bonilla, J. D. O. Pabon, and C. M. S. Rios, "Radar error calculation and correction system based on ads-b and business intelligent tools," in Proc. Int. Carnahan Conf. Secur. Technol., pp. 1–5, IEEE, 2018.

[4] D. A. Pamplona, L. Weigang, A. G. de Barros, E. H. Shiguemori, and C. J. P. Alves, "Supervised neural network with multilevel input layers for predicting of air traffic delays," in Proc. Int. Jt. Conf. Neural Networks, pp. 1–6, IEEE, 2018.

[5] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," in Proc. Int. Conf. Comput. Intell. Data Sci., pp. 1–5, IEEE, 2017.

[6] L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara, "On evaluating data preprocessing methods for machine learning models for flight delays," in Proc. Int. Jt. Conf. Neural Networks, pp. 1–8, IEEE, 2018.

[7] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," Transp. Res. Part C Emerg. Technol., vol. 44, pp. 231– 241, 2014.

[8] L. Hao, M. Hansen, Y. Zhang, and J. Post, "New york, new york: Two ways of estimating the delay impact of new york airports," Transp. Res. Part ELogist. Transp. Rev., vol. 70, pp. 245–260, 2014.

[9] ANAC, "The Brazilian National Civil Aviation Agency." anac.gov, 2017. [online] Available:http://www.anac.gov.br/.

[10] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient knn classification with different numbers of nearest neighbors," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2017

[11] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, "A deep learning approach to flight delay prediction," in Proc. Digit. Avion. Syst. Conf., pp. 1–6, IEEE, 2016.

[12] https://medium.com/@banikanusheela/flight-departure-delay-prediction-417240a72ea4