2023 4th International Conference on Signal Processing and Communication (ICSPC) | 979-8-3503-0077-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSPC57692.2023.10125675

# Flight Delay Prediction Using Different Regression Algorithms in Machine Learning

A. Evangeline
*Department of Electronics and Communication Engineering*
*Karunya Institute of Technology and Sciences*
Coimbatore, India
evangelinea21@karunya.edu.in

R. Catherine Joy
*Department of Electronics and Communication Engineering*
*Karunya Institute of Technology and Sciences*
Coimbatore, India
catherinejoy@karunya.edu

A. Albert Rajan
*Department of Electrical and Electronics Engineering*
*Karunya Institute of Technology and Sciences*
Coimbatore, India
albertrajan.a@gmail.com

***Abstract*—The two forms of regression algorithms are investigated and compared in this Article Such as LASSO and RIDGE regression. For scheduled airlines to improve customer satisfaction and success, accurate forecasting of delay is essential. There is no way to stop a flight from being delayed, yet they significantly affect the profits and losses of carriers. This research investigates a larger spectrum of potential flight delay difficulties and compares two machine learning algorithms in defined extended flight delay time series forecasting. A dataset for the suggested technique is created by gathering, decoding, and linking automatic dependent surveillance-broadcast (ADS-B) signals with additional data including weather, flight schedules, and airport information. A regression approach is used in conjunction with a number of forecasting tasks as part of the defined prediction challenges. The accuracy of the suggested prediction model was examined and compared to current prediction approaches. The results of LASSO and RIDGE regression with the mean absolute errorMAE (0.2 and 0.1), mean squared error- MSE (0.1 and 0.04), root mean square error-RMSE (0.3 and 0.2) and Accuracy (99.7% and 99.8%) respectively.**

***Keywords— Machine Learning, ridge regression, Flight delay prediction, lasso regression, Air-Traffic, Data Mining.***

## I. INTRODUCTION

Because air travel contributes significantly to the economics of airlines and the airports, it is vital for them to improve the quality of their services. Flight delays are a significant modern-day concern for airports and airline companies. Furthermore, aircraft delays raise worry among passengers, resulting in additional costs for both the agency and the airport. Flight delays cost the US government between 31 and 40 billion dollars in 2007 [1]. In 2017, 76% of flights were on time. Whereas the proportion of on-time flights declined by 8.5% in comparison to 2016 [2]. Security, weather conditions, component shortages, technical and aviation equipment issues, and flight crew delays represent a few causes of aircraft delays [3, 4, 5]. Weather delays are unavoidable [6], and they have significant economic implications for passengers, airlines, and airports [7, 8, 9]. Delays also can negatively affect the planet by increasing fuel usage and emitting noxious gases [1]. Besides that, because product transportation is heavily reliant on consumer trust, which can boost or drop ticket sales, on-time flight adds to customer confidence. Hence, flight prediction can result in better informed choices and operations, which in turn can benefit clients of agencies and airports. Leading to the broad spectrum and intricate nature of causes for flight delays, we are presented with a massive volume of data that cannot be dealt by standard methods of data analysis such as classification [1] or trees [8] are used to performance metrics of public road systems. What individuals recollect the most about a delay. Notably, people involved in the commercial aviation defines delay as the amount of time that an aircraft is running late or delayed. The discrepancy between a plane's scheduled departure or arrival time and its time gone may therefore be utilized a represent a delay. A number of ways have grown over time to improve welfare and cosines in air transportation. The impact on the travel industry was important. Today, people are occasionally affected by delays or cancellations. Every time, over $20 million is wasted due to plane eliminations and suspensions, which also affects approximately 30% of flights. Because the country's economy has expanded fast, the need for air travel has risen dramatically. Flight delays are becoming increasingly problematic, harming the reputation of public aviation services. Flight delays have been a long-standing problem, charging the airline industry dollars. Frequent flight delays cost the business a significant amount of money. For the airport, the flight delay seems to have a significant impact on the airport's daily business activities. Avoiding flight cancellations has become a difficult task. Flight delays have had an impact on people, airlines and management. It is difficult to analyse the delays of flights in the aviation sector. With these kinds of huge quantities of information, it is feasible to predict cancelled flights in instantaneously. The development of a powerful and effective model for dealing with the delay prediction problem is crucial. Flights are cancelled for many causes. This could be caused to dramatic environment (for example, snowstorms, gusty winds, and so on), long flight times, security, the national aviation system (for example, air traffic control, high traffic volume, and so on), and the air carrier (e.g., luggage stocking, airliner vacuuming, etc.). As a result, airlines rely heavily on information analytics to estimate disruptions. Airways may try to overcome interruptions, eliminate steadily for the past rescheduling, and improve the flying experience if they can predict aircraft delays. A modelling approach is used in this paper. In this paper, lasso regression and ridge regression are used on flight control records. The LASSO and RIDGE REGRESSION performs a major role to predicting the flight delays. Both the models work well in independent dynamic data to forecasting the delays.

## II. LITERATURE REVIEW

Meel and their team published an article titled "Predicting flight delays with error computation using machine trained classifiers" (2020). The Random Forest Regressor were chosen to predict the best forecast for departure delay in all categories, with the lowest Mean Squared Error (2261.8) and Mean

Absolute Error (24.1). The Randomized Forest Regression model remained the preferred model in the Arrival of flight Delays, with the lowest Mean Squared Error 3019.3 and Mean Absolute Error 30.8 values for each of these variables. Besides that, while this method may be adjusted to estimate cancellations of the flights at other ports, input from certain airport terminals must be incorporated in the design of that kind of analysis. Ding, Yi, "Predicting flight delay using multiple linear regression" (2017). These Under informed as well as C4.5 systems are compared against a technique of predicting coming airlines and a multiple linear regression strategy for estimating delay in this research.

Based to research utilizing one real collection of collections in the major airports, the designed system closely resembles an efficiency of 80%, a significant superior towards the Naive-Bayes and C4.5 methods. The screening findings show that this method is functionality and able to provide accurate predictions of airbus cancellations. It may likely bolster the company's position. Handling air traffic congestion is getting increasingly complex. It has previously been proved that all three of the aforementioned classifiers can obtain a precision of about 90% by utilizing only three features. Human flight data and weather parameters was used to provide five kinds of airplane aggregate features for forecasting. The Four key supervised learning methods are employed to improve the method in different models' prediction and exactness: linear regression, SVM, many randomized trees, and moderate GBM. The assessment goal is to predict airport departure delays. The suggested model for China's Nanjing Lukou International Airport was trained and validated using statistics from March 2017 to Jan 2018. These paper reveals that the gradient boosting ensemble method yields the greatest outcomes for a 60-minute timescale, including an average error of 0.865 and an average square error of 6.6 minutes, that is 1.83 minutes smaller than the findings of the previous articles. According to the statistics, this model surpasses the new research for a 1-hour simulation period, with a higher accuracy of 0.865 and an average square inaccuracy of 6.6 min, which became 1.83 min shorter. Operational dataset from March 2017 to Feb 2018 were used in the preparation and evaluate the proposed model for China's Nanjing Lukou Airport Terminal. Figure 1 Represents the supervised Machine Learning prediction described here should employ supervised machine learning techniques to recognize plane schedule cancellations.

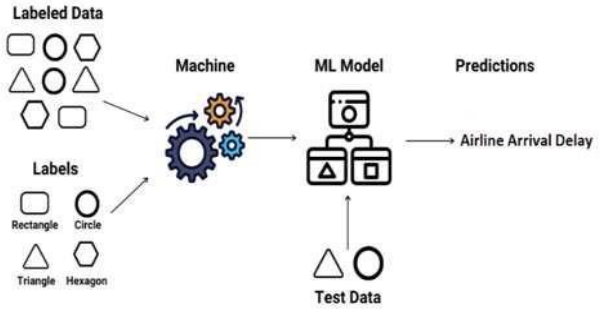



Fig. 1. Importance of Supervised Learning

### *A. Xgboost Model*

A gradient boosted branching approach, termed as a XG Boost, is widely used and well tested in a publicly available dataset. XG boost is a supervised learning approach which forecasts a value by combining overall predictions of numerous smaller regression models with linear regression. The work "Machine learning model-based prediction of flight delay" by Samanvitha and their teammates is available online (2020, October). For build a forecasting model, information on domestic travel in the United States and also weather parameters from June to December 2019 were collected.

This work "A unique work: Aircraft lag forecast via machine learning" by Natarajan and their team is available online (2018, December). This research further considers the significance of collective features with the example of cross validation. According to statistics, 19% of domestic flights in the United States land with either an average time delay of 900 seconds. Etani, N.'s article "Development of a prediction model for aircraft on-time arrival flight by detecting association between flight and meteorological data" was published (2019). User's approval is essential in the transport sector. Weather-related problems, a design flaw, as well as the craft's late arrival at the drop - off point, planes get detained and riders are disgruntled. A proposed methodology for the planes being on time is constructed using airframe and atmospheric datasets. This original study purpose is to create a connection among meteorological data and flight data. Relative sea-level concentrations of different climatic monitoring stations, the most northernmost station, MinamiTorishima, the most eastern position, and Yonagunijima, the much more west location can be used to characterize force variations. This publication discusses the relationship between altitude variations and aircraft systems from Apricot Aircraft, a Japanese minimal carrier (low-cost airline). As a consequence, utilizing machine learning's Random Forest Classifier, perfect time entrance rates could well be forecasted with 77 %. Furthermore, the software for estimating aircraft arrival times is being created to examine the efficacy of the classification algorithm.

### *B. Problem Identification*

Flight delays and cancellations can cause significant inconvenience and disruption to passenger's travel plans and have a considerable impact on the airline industry. Some of the problems associated with flight delay and cancellation include Passenger inconvenience, economic impact, safety concerns, legal and regulatory issues, operational inefficiencies Overall, flight delay and cancellation can have significant consequences for both passengers and airlines, highlighting the importance of addressing these issues through effective management and decision-making. Given the possibility of further delays (flight delays), these flight delays impede the functioning of the transportation network. To address flight delays, researchers develop prediction models that forecast when and where they will occur, as well as the causes and sources of the delays. Algorithms along this category try to calculate the total amount of instances, possibility, or overall level in lag for such a direct flight.

## III. PROPOSED METHODOLOGY

### *A. Dataset Information*

In this paper datasets are collected from the source https://www.kaggle.com/.The dataset contains the information about as flight such as DAY_OF_MONTH,

DAY_OF_WEEK, OP_UNIQUE_CARRIER, OP-_ CARRIER_AIRLINE_ID, OP_CARRIER, TAIL_NUM, OP_CARRIER_FL_NUM, ORIGIN_AIRPORT_ID, ORIGIN_AIRPORT_SEQDORIGIN, DEST_AIRPORT_ID, DEST_AIRPORT_SEQ_IDDEST, DEP_TIME, DEP_DEL15, DEP_TIME_B, ARR_TIME, ARR_DEL15, CANCELLED, DIVERTED and DISTANCE.

Lasso regression is such technique of smoothing. It is favored over regression approaches for higher accurate prediction. This method takes advantage of shrink. This kind of regression is ideal for models with high levels of collinearity or possible you schedule to automate certain phases of model selection, such as variable selection and parameter removal. The lasso method encourages basic, sparse models (i.e., models with fewer parameters). Shrinkage is the process by which data values are shrunk towards a central point known as the mean. Regularization is a significant idea which helps to prevent data overfitting, especially when the learned and test sets of data differ. In order to achieve less variance with the tested datasets, method is performed by adding a "penalty" term to the best option derived from the training data. By compressing their values, it also reduces the influence of the response variable on the output variable. Ridge regression is a technique for analyzing multi collinearity in multiple regression data sets. It is most appropriate whenever the amount of response variable in a collection of data exceeds the number of observations. In this paper we have performed two categories of regression procedures such as lasso and ridge regression with preprocessing the input flight delay datasets. This idea is mainly to predict the flight delays with the forecasting report for a particular flight. This kind of approach using machine learning is helpful for major airports in India.

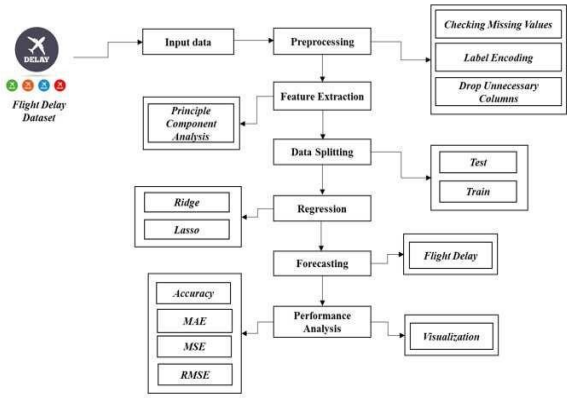
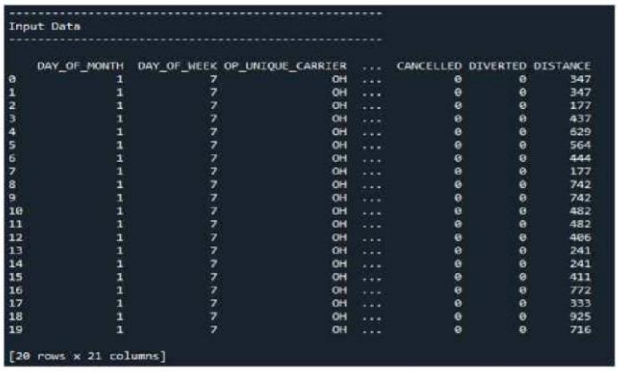



Fig. 2. Proposed System Architecture

The above block diagram (Figure 2) shows the overall work flow for the proposed method. The input is pre-processed by set of protocols that should be trained and predict the flights delays by using LASSO and RIDGE regressors. For each model, performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and accuracy must be calculated.

## *I. Input Data*

The process of picking data for forecasting flight delays is known as data selection. The time series dataset is utilised in this method to forecast flight delays. The dataset including information on passengers, states, and so on is shown in the Figure 3. We must read the dataset in Python using Panda's packages. Our dataset has a file extension of '.csv.'



Input Data

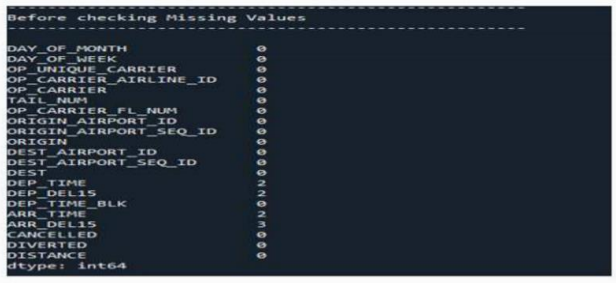
| | DAY_OF_MONTH | DAY_OF_WEEK | OP_UNIQUE_CARRIER | ... | CANCELLED | DIVERTED | DISTANCE |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 7 | OH | ... | 0 | 0 | 347 |
| 1 | 1 | 7 | OH | ... | 0 | 0 | 347 |
| 2 | 1 | 7 | OH | ... | 0 | 0 | 177 |
| 3 | 1 | 7 | OH | ... | 0 | 0 | 437 |
| 4 | 1 | 7 | OH | ... | 0 | 0 | 629 |
| 5 | 1 | 7 | OH | ... | 0 | 0 | 564 |
| 6 | 1 | 7 | OH | ... | 0 | 0 | 444 |
| 7 | 1 | 7 | OH | ... | 0 | 0 | 177 |
| 8 | 1 | 7 | OH | ... | 0 | 0 | 742 |
| 9 | 1 | 7 | OH | ... | 0 | 0 | 742 |
| 10 | 1 | 7 | OH | ... | 0 | 0 | 482 |
| 11 | 1 | 7 | OH | ... | 0 | 0 | 482 |
| 12 | 1 | 7 | OH | ... | 0 | 0 | 406 |
| 13 | 1 | 7 | OH | ... | 0 | 0 | 241 |
| 14 | 1 | 7 | OH | ... | 0 | 0 | 241 |
| 15 | 1 | 7 | OH | ... | 0 | 0 | 411 |
| 16 | 1 | 7 | OH | ... | 0 | 0 | 772 |
| 17 | 1 | 7 | OH | ... | 0 | 0 | 333 |
| 18 | 1 | 7 | OH | ... | 0 | 0 | 925 |
| 19 | 1 | 7 | OH | ... | 0 | 0 | 716 |

[20 rows x 21 columns]

Fig. 3. Input data

## *II. Preprocessing*

The process of removing undesirable data from a dataset is known as "data pre-processing." Pre-processing data transformation techniques are used to turn the dataset into a machine-learning-friendly structure. Removal of missing data Null values, such as missing values and Nan values, are replaced by 0 throughout this operation. To represent categorical data, categorical variables with a finite number of label values are used.



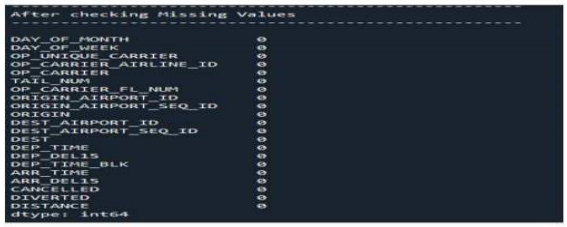
Before checking Missing Values

| | |
|---|---|
| DAY_OF_MONTH | 0 |
| DAY_OF_WEEK | 0 |
| OP_UNIQUE_CARRIER | 0 |
| OP_CARRIER_AIRLINE_ID | 0 |
| OP_CARRIER | 0 |
| TAIL_NUM | 0 |
| OP_CARRIER_FL_NUM | 0 |
| ORIGIN_AIRPORT_ID | 0 |
| ORIGIN_AIRPORT_SEQ_ID | 0 |
| ORIGIN | 0 |
| DEST_AIRPORT_ID | 0 |
| DEST_AIRPORT_SEQ_ID | 0 |
| DEST | 0 |
| DEP_TIME | 2 |
| DEP_DEL15 | 2 |
| DEP_TIME_BLK | 0 |
| ARR_TIME | 2 |
| ARR_DEL15 | 3 |
| CANCELLED | 0 |
| DIVERTED | 0 |
| DISTANCE | 0 |

dtype: int64

Fig. 4. Before checking missing values

Figure 4 shows the missing values before checking from the input dataset. This is the first step of the pre-processing methods in the given datasets.



After checking Missing Values

| | |
|---|---|
| DAY_OF_MONTH | 0 |
| DAY_OF_WEEK | 0 |
| OP_UNIQUE_CARRIER | 0 |
| OP_CARRIER_AIRLINE_ID | 0 |
| OP_CARRIER | 0 |
| TAIL_NUM | 0 |
| OP_CARRIER_FL_NUM | 0 |
| ORIGIN_AIRPORT_ID | 0 |
| ORIGIN_AIRPORT_SEQ_ID | 0 |
| ORIGIN | 0 |
| DEST_AIRPORT_ID | 0 |
| DEST_AIRPORT_SEQ_ID | 0 |
| DEST | 0 |
| DEP_TIME | 0 |
| DEP_DEL15 | 0 |
| DEP_TIME_BLK | 0 |
| ARR_TIME | 0 |
| ARR_DEL15 | 0 |
| CANCELLED | 0 |
| DIVERTED | 0 |
| DISTANCE | 0 |

dtype: int64

Fig. 5. After checking missing values

Figure 5 shows the missing values before checking from the input dataset. This is the first step of the pre-processing methods in the given datasets.

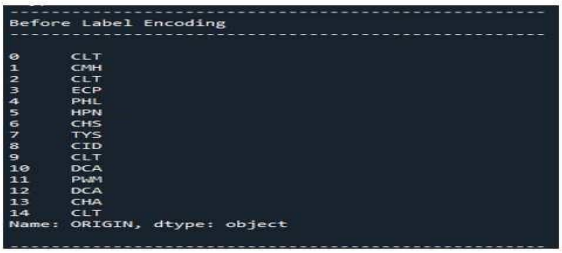



Fig. 6. Before label encoding

In figure 6 the original format of the label which the machine or a computer cannot able to understand the labels.
That need to be encoded in the further steps.

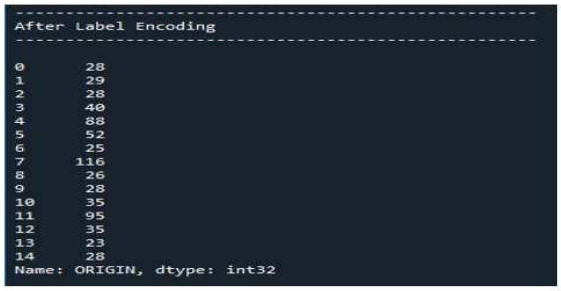



Fig. 7. After label encoding

The above picture (figure 7) represents the after encoding the labels which are available in the dataset. This step is mainly to understand to the machine language we are converting the label into numerical representations. Every feature from the dataset is converted into the numerical format for the further pre-processing stages.

### *III. Feature Extraction*

The purpose of feature extraction is to supply valuable characteristics for data mining operations that will increase the accuracy or explanatory power of the model.
Many things impact arrival flight time. We must employ methods for extracting features such as principal component analysis in our approach (PCA). The PCA is used to reducing dimensionality in machine learning using unsupervised learning. PCA is a technique for calculating the principal components and using them to modify the mechanism of data by using only the top few major features and disregarding the rest. PCA is used for analysing the data for both predictive and exploratory purposes. A collection of linearly predictor variables is combined into a set of correlated observations using the statistical technique of modified algorithm.

### *IV. Data Splitting*

The data is required throughout the machine learning process in order for learning to occur. In addition to training data, test data are needed to assess the algorithm's performance and determine how well it performs. In our procedure, we categorised 70% of the dataset as training data and 30% as testing data. "Data splitting" means the practice of dividing accessible data into two sections, more typically for cross-validator reasons. One set of data is used to build a predictive model, while the other is used to assess the model's performance.

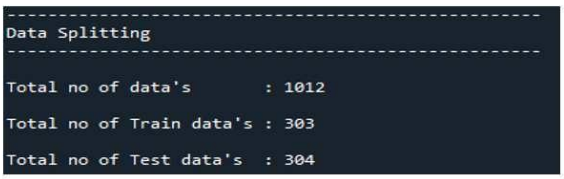



Fig. 8. Data splitting

In every machine learning algorithm, we need to divide the datasets into two.one is for testing and another one is for training from the datasets. Maximum samples from the training data will leads to good accuracy. That outcome has shown in the figure 8.

## IV. RESULTS AND DISCUSSION

The term LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a deformation linear regression technique. Shrinkage is the process of reducing data values towards a central point, such as the mean. The lasso method can improve basic, sparse models. This type of regression is ideal for models with a high level of multicollinearity or for automating certain aspects of model selection, such as variable selection/parameter removal.

Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients) Where,

- λ represents the amount of shrinkage.
- λ represents - 0 indicates that all features are taken into account, and it is comparable to linear regression, which uses only the residual sum of squares to generate a prediction model.
- λ represents - ∞ signifies no feature is considered i.e., as λ closes to infinity it eliminates more and more features.
- When the bias increases, then the λ is also increases.
- When the variance increases, then the λ is decreased.
  Residual Sum of squares + λ * (Sum of the absolute value of the magnitude of coefficients)

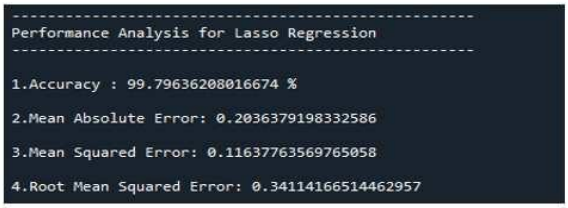



Fig. 9. Performance analysis for LASSO regression

The above figure 9 shows the overall Accuracy of LASSO regression 99.79% of performance in machine learning models. When predictors are associated, RR was first proposed as a method of estimating regression coefficients with lower mean-square error than their least squares equivalents. RR is part of a family of penalised regression algorithms, which is also include Lasso regression and the Elastic Net. The latter is a weighted average of the Lasso and ridge penalties. RR has been proven to have high predictive accuracy among penalised regression techniques. The selection of the shrinkage parameter, or parameters, that regulate the amount of shrinkage of the regression coefficients is one issue in using penalised regression techniques. The below figure 10 shows the overall Accuracy of RIDGE regression and some other parameters

such as MAE, MSE and RMSE. It gives 99.89% performance compared to LASSO it gives more Accuracy it predicts the delays.

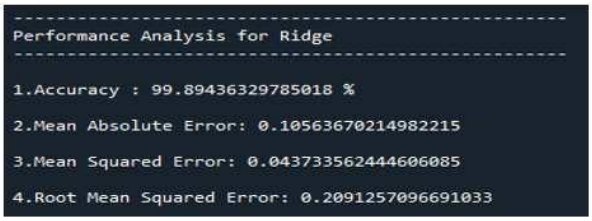



Fig. 10. Performance analysis for ridge regression

For flight time prediction in this section, 1000 arrival aircraft timing at Indian airports will be used. First step is the dataset splitting. Second, utilising the evaluation indicators, the prediction effects of the lasso regression model and the ridge regression model are compared.

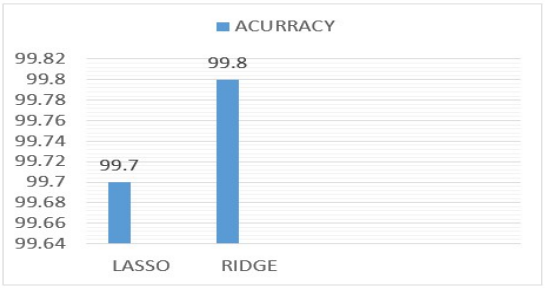



Fig. 11. Comparison of lasso and ridge regression

The above figure 11 shows the comparison of two different models in regression. Compared to Lasso regression RIDGE performs high for predicting the flight delays.

## V. CONCLUSION

This study suggested a plan for forecasting total flight delays at airports through machine learning techniques. Apart from decisions that directly affect passengers, delay prediction is important for all the parties involved in the air transportation system. By investigating these models from a data physical perspective, this paper contributes. We have made two different machine learning regression techniques for Regression. Finally, the result reveals that some performance metrics, such Accuracy, MAE, MSE, and RMSE. The LASSO and RIDGE REGRESSION occurs the accuracy of 99.7% and 99.8% respectively. The flight delay is then forecast or examined via visualizations. Predicting aircraft delays is an intriguing scientific issue that has attracted a lot of attention recently. In order to improve the precision and accuracy of forecasting flight delays, most researches have tried to improve and extend their models. Since on-time flight arrival is critical, flight delay prediction models must be exceedingly accurate and exact. In order to enhance the precision and accuracy of forecasting flight delays, most studies have attempted to improve and widen their models. Because on-time flight arrival is critical, flight delay prediction models must be highly accurate and exact.

## REFERENCES


[1] AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. J Air Transp Manag. 2008;14(5):221–36.

[2] Balaban E, et al. Dynamic routing of aircraft in the presence of adverse weather using a POMDP framework. In 17th AIAA aviation technology, integration, and operations conference. 2017.

[3] Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and societal welfare. Transp Res Part E Logist Transp Rev. 2012;48(2):460–9.

[4] Choi S, et al. Prediction of weather-induced airline delays based on machine learning algorithms. In 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). 2016. New York: IEEE.

[5] D'Ariano A, Pistelli M, Pacciarelli D. Aircraft retiming and rerouting in vicinity of airports. IET Intel Transp Syst. 2012;6(4):433–43.

[6] Deepudev, S., Palanisamy, P., Gopi, V. P., & Nelli, M. K. (2021). A machine learning based approach for prediction of actual landing time of scheduled flights. In *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications* (pp. 755-766). Springer, Singapore.

[7] Ding, Y. (2017, August). Predicting flight delay based on multiple linear regression. In *IOP Conference Series: Earth and Environmental Science* (Vol. 81, No. 1, p. 012198). IOP Publishing.

[8] Etani, N. (2019). Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. *Journal of big data*, *6*(1), 1-17.

[9] Evans JE, Allan S, Robinson M. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Proceedings of the 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis. 2004.