International Journal of Engineering in Computer Science

E-ISSN: 2663-3590
P-ISSN: 2663-3582
www.computersciencejournals.com/ijecs
IJECS 2024; 6(1): 14-21
Received: 04-11-2023
Accepted: 09-12-2023

G Sai Chaitanya
Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad, Telangana, India

Dr. Subhani Shaik
Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad, Telangana, India

P Visalakshi
Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad, Telangana, India

G Rakshitha
Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad, Telangana, India

Corresponding Author:
G Sai Chaitanya
Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Hyderabad, Telangana, India

Comparative study of ensemble machine learning techniques for airline delay prediction

G Sai Chaitanya, Dr. Subhani Shaik, P Visalakshi and G Rakshitha

DOI: https://doi.org/10.33545/26633582.2024.v6.i1a.105

Abstract
Airlines generate highest economy domain for most of the countries in the world. Air-traffic jamming causing flight delays, this problem face by aviation industry. Due to flight delay more impact on economic as well as environmental properties. Air traffic management is challenging to aviation business. Due to this challenge make huge loss as well as budget loss also. So many reasons impact in flight delay to weather conditions, security problems, airport traffic and mechanical problems etc. We propose hybrid machine learning models for prediction of airline delay and saves huge turnovers using ensemble machine learning models. Our research talk about few machine learning algorithms predicts the more than 90% accuracy rate. XGBoost algorithm provide more than 90% accuracy. But one thing talks about this airline delay prediction, it purely based on dataset with corresponding suitable algorithm.

Keywords: Ensemble machine learning, airline delay, prediction, accuracy

Introduction
Population increases extremely and time to time increase billionaires also. In 1960s most of the people not concentrated on flights because high cost and continuous delay of flights. Most of the governments help to airplane manufacturing companies for rapid development [1]. They also support for development of airports for more comfort and made control of airlines. Airlines generate highest economy domain for most of the countries in the world. Air-traffic jamming causing flight delays, this problem face by aviation industry. Due to flight delay more impact on economic as well as environmental properties [2]. Air traffic management is challenging to aviation business. Due to this challenge make huge loss as well as budget loss also. So many reasons impact in flight delay to weather conditions, security problems, airport traffic and mechanical problems etc. Airline business play a crucial role in country economy. Due to few problems like flight delay, there is huge loss occurred but recent technology of machine learning is a way to control flight delays. In traditional days mining techniques for avoid airline delays due to this reason move to machine learning models [3].

In airline industry so many areas included in this field. They are price of tickets, Airport maintenance, airline transportation, improve customer support, and promoting airlines. In real time to save public money and time [4]. Flight performance is based on time maintenance for customer convenience to show airlines excellence. To reduce airline delay can help airline continuously getting revenue and penalty costs. Flight arrival and departure had more impact on airlines, passengers and airports [5]. Flight delays is crucial impact for commercial airlines loss for passenger complaints. Petrol emissions through fuel usage negative influence on environment. To overcome these issues airline companies, adopt the novel technology for maintenance of airlines. Due to flight delay, airport authorities allocate space for passengers and provide customer requirements [6]. Flight delay impact on staff timing also, so plan schedule effectively. If you maintain properly and take strategic decisions then workforce and work space both reduced. Airport operation control centres monitoring different matters of airline schedule for help air traffic control, ground handling service, automation instead of manual based expertise and airport [7].

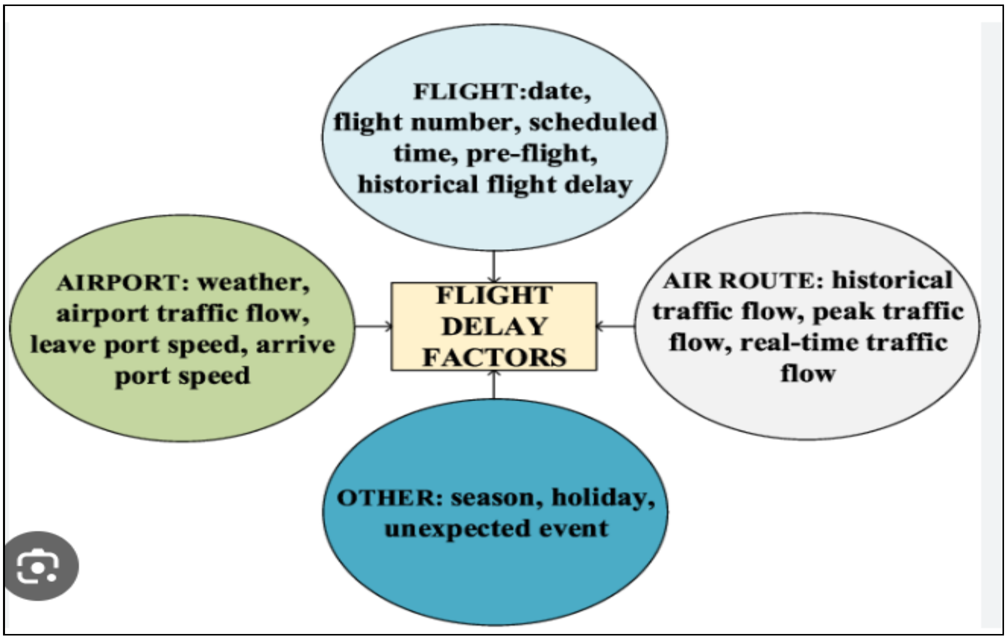

Fig 1: Factors of Flight delay [8].

The following paper continuous with proposed work and architecture of the research in section two. Section three deal with comparative study. Section four deal with results and analysis and final section conclude the paper.

Proposed Work and Architecture

The goal of research to estimate the probability of flight, the total connection time and time to departure. The recent technology to determine the flight delays. Due to machine learning models predicting accurate outcome, promoting effective airlines, reducing cancellation cost, improve customer support and reduce time delay with complete their work smooth [9]. We propose hybrid machine learning models for prediction of airline delay and saves huge turnovers using ensemble machine learning models. Our research talk about few machine learning algorithms predicts the more than 90% accuracy rate.

The following figure 2 shows the architecture of our proposed system. This diagram follows the normal procedure of machine learning model starting from input data to visualization of results with different phases of project. The following phases or modules included in our project.

1. Input data.
2. Pre-process the data.
3. Feature selection.
4. Data test and train.
5. Visualization.

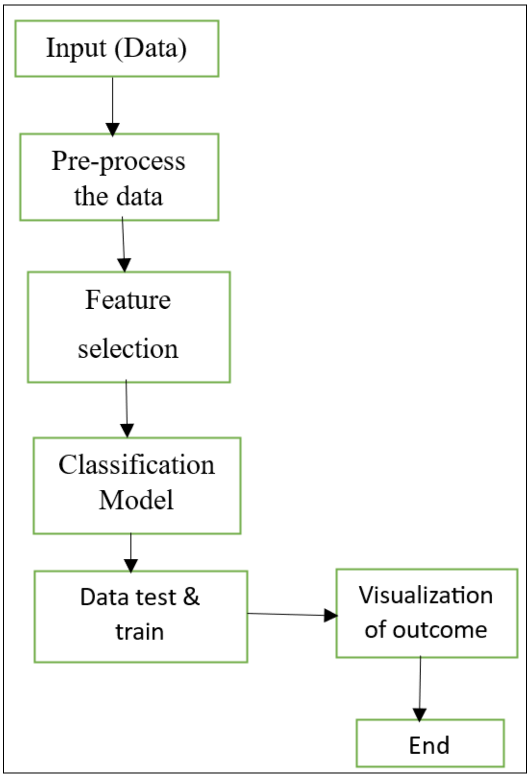

Fig 2: Architecture of the system

Creating algorithms and statistical models that enable computers to automatically learn from data and improve their performance on a particular activity without being explicitly taught is known as machine learning, and it is a subset of artificial intelligence [10]. The ability to learn from examples and experience is what machine learning enables computers to do, which improves their ability to predict the future or take action in response to that knowledge. This is often done by training a model on a dataset. In this process, the model learns to recognize relationships and patterns in the data and then uses this understanding to predict or make decisions about brand-new, unseen data [13].

Various machine learning applications such as fraud detection, autonomous driving, language comprehension, speech and image recognition, and recommendation systems utilize a diverse range of algorithms and methodologies to address increasingly intricate challenges in the field [11].

Results and Analysis

We propose hybrid machine learning models for prediction of airline delay and saves huge turnovers using ensemble machine learning models. The following table 1 represents the dataset of our research.

Table 1: Dataset

Rows of data: 539383								
	Airline	**Flight**	**Airport From**	**Airport To**	**Day of Week**	**Time**	**Length**	**Delay**
0	CO	269	SFO	IAH	3	15	205	1
1	US	1558	PHX	CLT	3	15	222	1
2	AA	2400	LAX	DFW	3	20	165	1
3	AA	2466	SFO	DFW	3	20	195	1
4	AS	108	ANC	SEA	3	30	202	0

Attribute Description

ID - Unique row identifier ID.
Airline - Abbreviated name of different commercial airlines.
Flight - Talk about the type of aircraft used.
Airport from - The airport origin for the flight.
Airport to - The airport destination for the flight.
Day Of Week - Day of the week when the flight took place.
Time - Time of flight.
Length - Length of flight.
Delay - Whether or not there was a delay.

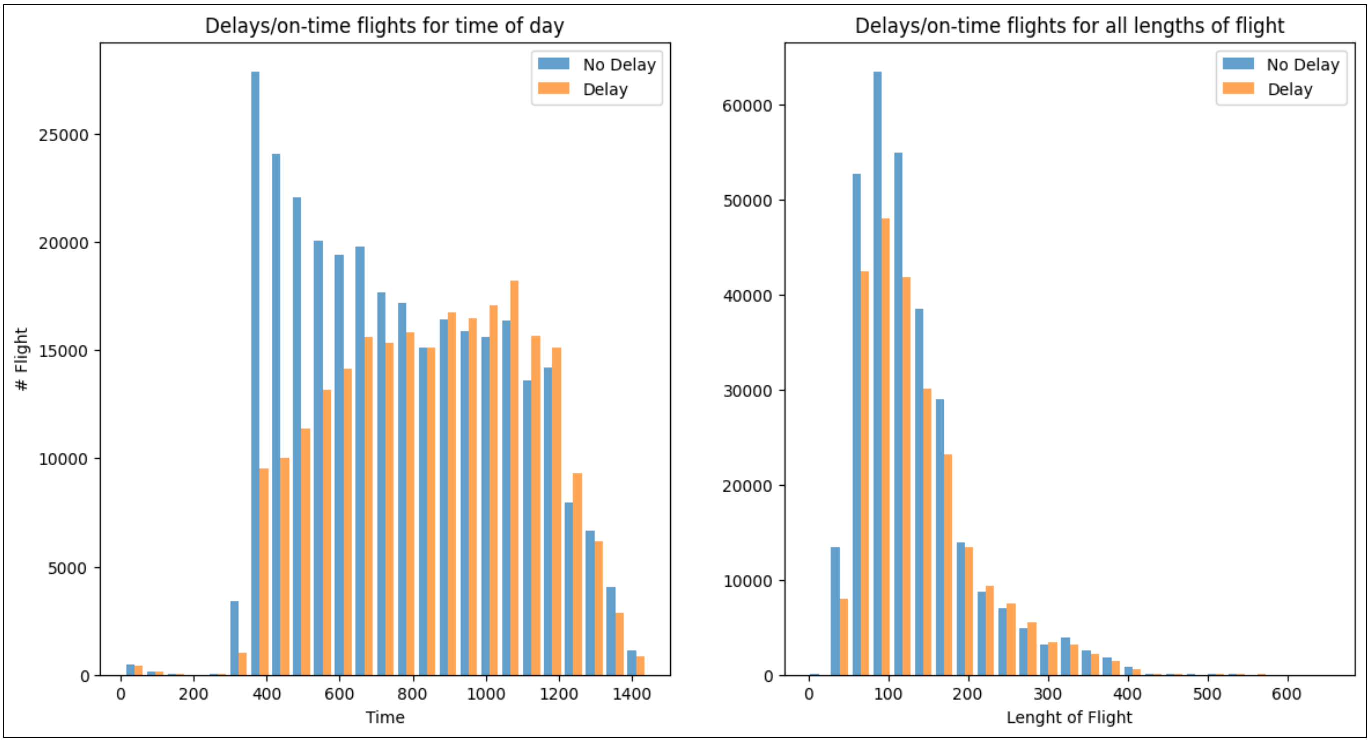

Fig 3: Visualization of delay of flights

In the plot above, the only two continuous features (time and length), are examined in histograms. The histogram on the left reveals important information about the predictive quality of the "Time" feature. It's clear flights that are earlier in a day are significantly less likely to incur a delay compared to flights later in the day. At around time 800, the ratio of delays to no delays comes back to about 1:1 and stays that way for the remainder of the day. The histogram on the right shows that the length of the flight generally doesn't provide very good insight on whether there will be a delay.

The fraction of overall delays stands at around 45% for this data set. With no built model, the baseline prediction of no delay would be about 55% accurate. Let's see how much predictive value the features given can provide.

We use train_test_split method of sklearn. Pre-processing module to split the dataset. The dividing ratio has a huge impact on the accuracy of prediction. Usually the training data size should be greater than the testing data size for achieving good accuracy. We have divided the dataset in the ratio 80-20 for good accuracy of the model.

Neural Network Models

For this project, I use FastAi's deep learning library (an extension of PyTorch) to create three fully connected neural networks with varying sizes.

Model 1 will be a fully connected neural network containing three hidden layers with 500, 250, and 10 neurons respectively. The activation function for each hidden layer is ReLU (Rectified Linear Unit), and the activation function for the output layer is the sigmoid function. There is no special reason for using ReLU here, but Sigmoid serves to compress the output layer to numbers between 0 and 1. Each output neuron (here 2 neurons, delay or no delay) will sum together to 1, keeping consistent with the law of total probability which states that the sum of probabilities of all events in a given space must sum to 1.

The activation function definitions are as follows:

ReLu (x) = max (0, x) (1)

Sigmoid (x) = 11+e−x (2)

FastAi's fit_one_cycle() function uses fluctuating learning rates to improve speed and accuracy over its generic fit() function. In addition, the learner will be using CrossEntropyLoss in each model. For this problem, a binary classification problem, the cross-entropy loss will look like

CE_Loss=−ylog(p)+(1−y) log(1−p) (3)

Where y is the true label value (either 0 or 1) and p is the probability that y=1

Table 2: Neural network model prediction

Epoch	Train_loss	Valid_loss	Accuracy	Time
0	0.62369	0.614684	0.660202	0:31
1	0.606615	0.606686	0.66747	0:30
2	0.592889	0.606685	0.667489	0:30
3	0.57645	0.609003	0.667248	0:31
4	0.550825	0.619873	0.662242	0:30
Better model found at epoch 0 with valid_loss value: 0.6146835684776306. Better model found at epoch 1 with valid_loss value: 0.6066862344741821. Better model found at epoch 2 with valid_loss value: 0.60668534				

In this model 1 predicts the accuracy rate is more than 60%. In the same way remaining two models (Model 2 and Model 3) also predicts the same type of accuracy.

After training all three neural networks, it is clear that each network is very similar in both loss and accuracy. For the sake of precision, I will use the network with the lowest loss for the rest of this document, network 2.

Decision Tree

Here, the decision tree prediction will be evaluated. Seen below are two decision tree graphs of the same 4 leaf node decision tree generated on the data.

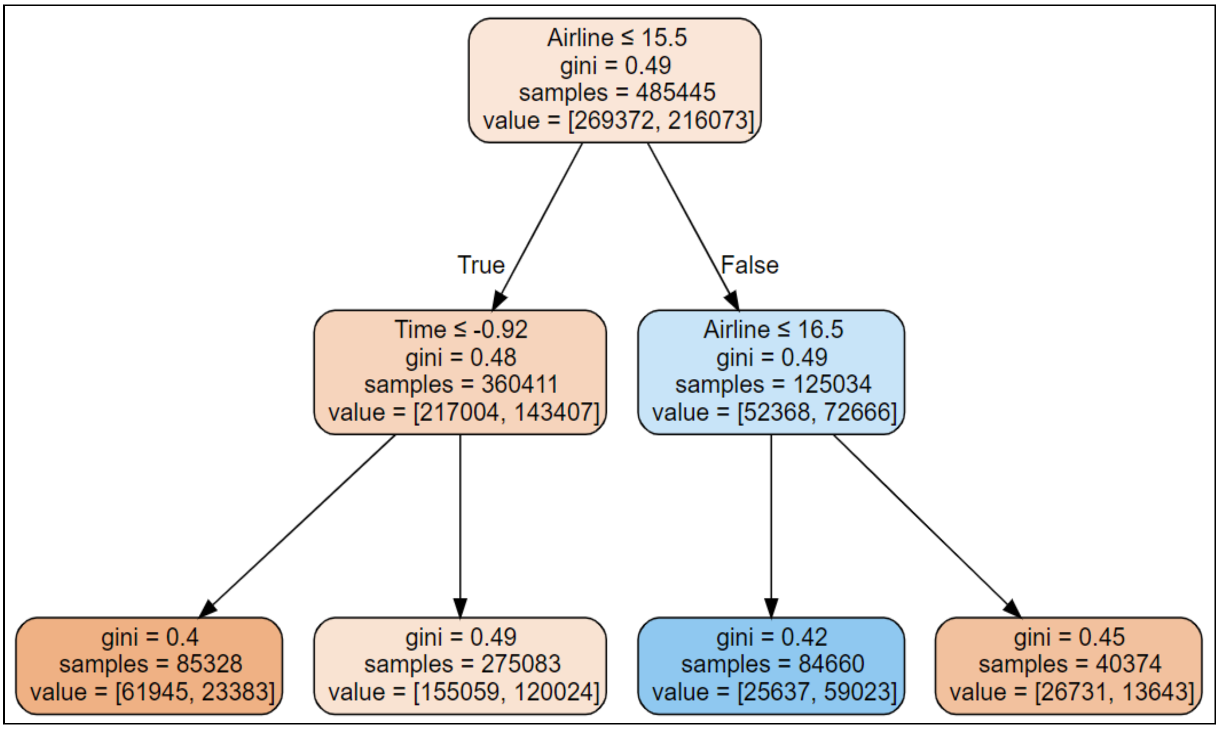

Fig 4: 4 leaf node decision tree generation

The test accuracy for the four-leaf decision tree generated above is 0.623 (62.3%). This is noticeably worse than the neural network from earlier. Now, I will look at the accuracy for a much larger tree. The test accuracy has dramatically improved when the minimum number of leaves was forced to be 25, now sitting at 0.651 (65.1%).

Random Forest

The random forest utilise a technique called bagging that puts together an ensemble of decision trees and takes the average prediction of all of them. The decision trees in the random forest will use at most 2/3 (hyper-parameter) of the data in their construction. This amazingly ensures that as the number of trees in the random forest increases, the loss decreases, as seen in a plot below.

The accuracy for the random forest sits at 0.665 (66.5%), right around where the neural networks were at. The plot below shows the depiction of the random forest converging to a minimum loss as the number of trees in the forest increases.

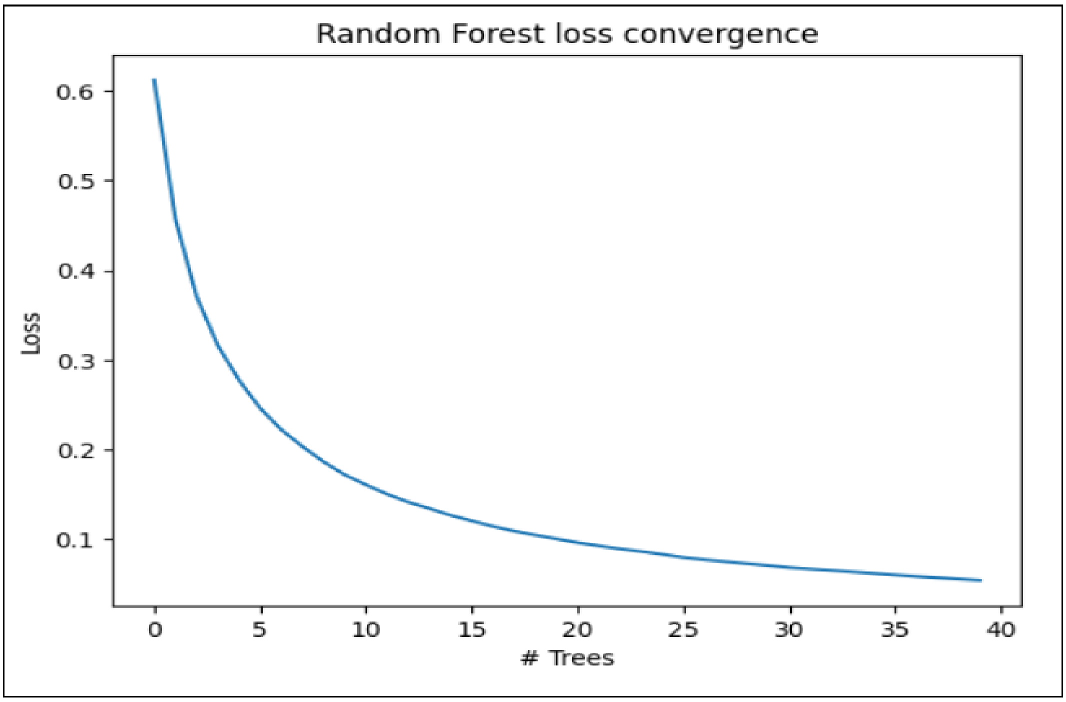

Fig 5: Random forest loss convergence

Variable Importance

One easy metric to look at with random forests and decision trees is variable importance. A data set can have thousands of features with only a few of them actually making a meaningful impact on the outcome. Variable importance shows which variables are the most important in predicting the outcome. To calculate this, each decision tree computes how much loss is reduced from each variable that was split. In a random forest, this is looked at as an average from all decision trees in the ensemble.

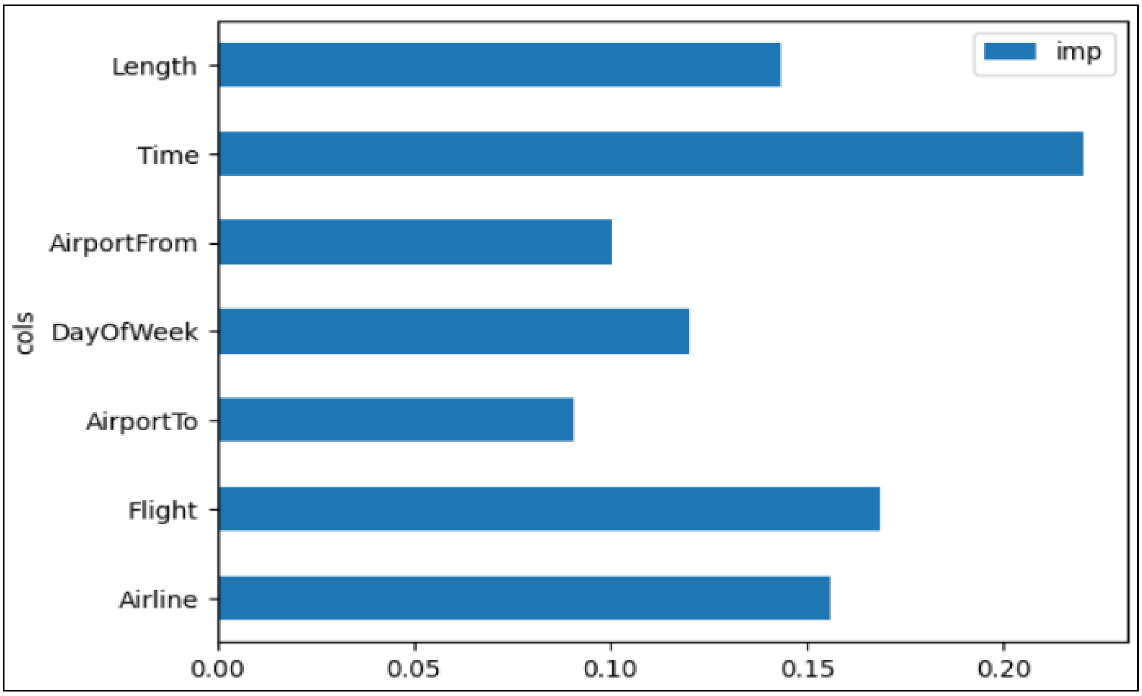

Fig 6: Visualization of variable importance

According to the random forest, the most important feature is the "Time" feature. This makes sense according to the plot of the time and length features seen in the beginning of the notebook. "Time" clearly had more predictive value than "Length", and the variable importance echos this realization.

3.5 Ensemble the neural network and random forest predictions

```
[ ] rf_preds = m.predict(valid_xs)
    nn_preds,_ = learn.get_preds()
    ens_preds = (np.transpose(to_np(nn_preds.squeeze()))[1] + rf_preds) /2

    ens_preds

    array([0.8089541 , 0.10249221, 0.7542748 , ..., 0.68490934, 0.17556119,
           0.8531245 ], dtype=float32)

[ ] def get_preds(preds): return [0 if i < .5 else 1 for i in preds]

[ ] # Accuracy of ensemble of NN and Random Forest
    round(accuracy(get_preds(ens_preds), valid_y), 3)

    0.663
```

The accuracy of the ensemble of the neural network and the random forest sits at 0.665 (66.5%). This is still very slightly worse than the neural network, so the best model to use is still the neural network.

Logistic Regression: Here I am using logistic regression to predict delays. The two most popular machine learning models are linear regression and logistic regression. In our case, because the dependant variable is categorical, logistic regression will be used. Linear regression is used for continuous dependant variables.

Table 3: List of variables

Airline	18
Flight	6585
Air Port from	293
Air Port to	293
Day of week	7
Time	1131
Length	426
Delay	2
Dtype: int64	

Directly above you can see the number of unique values for each feature. In order to perform standard linear regression, we must one-hot encode the categorical variables. As you can see above, the number of unique items in each category can be large, especially in the "Flight" category. To accommodate for this, the category is either left as is, or removed from the data frame entirely, because the machine would endure memory overload if this category was one-hot encoded. For this problem, it makes more sense to remove it entirely than to leave it as is due to the fact that there is no predictive value in the ordering of the ambiguous ID assignment.

Now, the number of features has risen from 8 to 613 with the one-hot encoded features. These new features are called dummy variables because each one represents just one unique value from all of the categorical features [12].

```
[ ] reg = LogisticRegression().fit(xs_linear, y_train)

[ ] preds = reg.predict(xs_valid_linear)

[ ] round(accuracy(preds, y_valid), 3)

    0.638
```

The accuracy from the logistic regression is 0.638 (63.8%), the lowest of all models looked at today, apart from the four-leaf decision tree.

Ensemble all three models: Logistic Regression, NN, and RF

```
rf_preds = m.predict(valid_xs)
nn_preds,_ = learn.get_preds()
ens_preds = (np.transpose(to_np(nn_preds.squeeze()))[1] + rf_preds + preds) /3

ac = round(accuracy(get_preds(ens_preds), valid_y),3)
print(f"Triple ensemble accuracy: {ac}")

Triple ensemble accuracy: 0.669
```

The ensemble of all three models gives quite a large accuracy, slightly improving that of our best neural network.

Models Comparison

In our research, applied different types of machine learning algorithms for classification purpose on particular airline dataset. These algorithms used individually and ensemble learning with two or three models [14]. We observe in our research any model can't give higher accuracy with this dataset. We are also trying ensemble learning with two or three models but no use to get higher accuracy. Almost all models predict the same type of accuracy rate.

Our research talk about few machine learning algorithms predicts the more than 90% accuracy rate. For example, XGBoost algorithm provide more than 90% accuracy. But one thing talks about this airline delay prediction, it purely based on dataset with corresponding suitable algorithm.

Table 4: Models comparison

S. No.	Name of the Model	Accuracy Rate (%)
1.	Neural networks	61.4
2.	Decision tree	65.1
3.	Random Forest tree	66.5
4.	Logistic regression	63.8
5.	Neural networks+ random forest tree	66.3
6.	Logistic regression+ Neural networks + Random Forest Tree	66.9

Ensemble learning is not suitable for airline delay prediction problem. So better to avoid these models. Most of the algorithms used up to now for this problem. In future research search for other algorithms for best accuracy.

Conclusion

Predicting whether a flight will incur a delay is difficult. Most of the time, a delay will occur for seemingly random reasons, like weather. In this data set, the only information given about the flight is that which is created when the flight itself is made. Things weather and status of prior flights are not included in this data, but certainly are big components in influencing flight delays. This creates an upper bound of predictive power we can achieve with this data that cannot be crossed because of this lack of crucial information.

With the base prediction of no delay giving 55% accuracy, we improved around 11% in total accuracy with our best model, the neural network model. This is a good improvement that would likely require features like weather or other information that is unknown until the day of flight to improve further.

These algorithms used individually and ensemble learning with two or three models. We observe in our research any model can't give higher accuracy with this dataset. We are also trying ensemble learning with two or three models but no use to get higher accuracy. Almost all models predict the same type of accuracy rate.

References

1. Balakrishnan H. Control and optimization algorithms for air transportation systems. Annual Reviews in Control; c2014.
2. Surya Teja KP, Reddy V, Shaik S. Flight Delay Prediction Using Machine Learning Algorithm XGBoost. Journal of Advanced Research in Dynamical & Control Systems, 2019, 11(5).
3. [Anonymous]; c2021. https://doi.org/10.1145/3497701.3497725
4. Khaksar H, Sheik Holeslami A. Airline delay prediction by machine learning algorithms. Scientia Iranica; c2017.
5. Esmaeilzadeh E, Mokhtarimousavi S. Machine learning approach for flight departure delay prediction and analysis. Transportation Research Record: Journal of the Transportation Research Board. 2020;2674(8):145-159.
6. Al-Tabbakh MS, Mohamed HM, El ZH. Machine learning techniques for analysis of Egyptian flight delay. International Journal of Data Mining & Knowledge Management Process. 2018;8(3):01-14.
7. Khan WA, Ma HL, Chung SH, Wen X. Hierarchical integrated machine learning model for predicting flight

departure delays and duration in series. Transportation Research Part C: Emerging Technologies. 2021;129:103225.
8. Etani N. Development of a predictive model for on-time arrival flight of airliner by discovering correlation between flight and weather data. Journal of Big Data. 2019;6(1):85.
9. Zhu X, Li L. Flight time prediction for fuel loading decisions with a deep learning approach. Transportation Research Part C: Emerging Technologies. 2021;128:103179.
10. Mamdouh M, Ezzat M, Hefny H. Optimized planning of resources demand curve in ground handling based on machine learning prediction. International Journal of Intelligent Systems and Applications. 2021;13(1):1-16.
11. Evler J, Asadi E, Preis H, Fricke H. Airline ground operations: Schedule recovery optimization approach with constrained resources. Transportation Research Part C: Emerging Technologies. 2021;128:103129.
12. Abdel-Aty M, Lee C, Bai Y, Li X, Michalak M. Detecting periodic patterns of arrival delay. Journal of Air Transportation Management. 2007;13(6):355-361.
13. Shaik S, Ravibabu U. Detection and Classification of Power Quality Disturbances Using Curvelet Transform and Support Vector Machines. In: Proceedings of the 5th IEEE International Conference on Information Communication and Embedded System (ICICES-2016); 2016 Feb 25-26. Chennai, India.