Single, Bagging, Boosting Regression Algorithms in Predicting Flight Arrival Delay

Vikas Thoti Reddy*

Abstract

Delays in commercial airline flights are common, expensive, and are often unpredictable. It has been suggested that machine learning techniques can be utilised to predict delays in aircraft departure and arrival times. The new, more complex machine learning models include boosting and bagging algorithms. These newer models were compared to classical ones to explore the trade-off between model complexity, accuracy, and fit time. Rather than using classification algorithms, the study uses regression models that predict the exact delay of a flight using various input features. To create models for airline delay prediction, I explored boosting and bagging regression models in relation to more classical versions.

Keywords: bagging, boosting, cross validation, run time

1 Introduction

Domestic airplane delays have a significant impact on the US economy and cost billions of dollars each year. In 2007, 31.2 billion dollars was wasted as a result of the domestic flight delays in the US (Airlines for America, 2020). It has been shown that machine learning algorithms can predict arrival delays in advance; this would allow airport management to adjust schedules, minimize delays, and develop flight delay management procedures. Previous work in this regard has focused on the utilization of classical classifier machine learning algorithms; however, in this work, complex bagging and boosting algorithms were compared to classical models. The purpose of this work is to create regression models that predict the exact delay of a desired flight. Previous works have used classifier models that predict whether a flight will be delayed or not. In this specific case, simply determining whether an incoming flight will be delayed or not through a binary output will not considerably help airport traffic management. Rather, in this study, regression models were used to predict how delayed a flight will be as a continuous output. A model that predicts accurate arrival delay times as a continuous value will more substantially help airport traffic management than a classifier model. In turn, models focused on continuous outputs will help decrease the amount of money wasted from airline delays by bolstering airport traffic management. Before using machine learning algorithms to fix the issue, a deep exploration of the issues facing the air travel industry was done. It was found that the main cause of airline delays is the sheer volume of flights every year. In recent ages, the dependence on air travel has increased because of the comfort and speed of modern airplanes. The increase in the number of domestic and international flights puts more stress on air traffic control. This is important to note because the most common reason for delayed flights is poor air traffic management (Busson, 2019). Last year, 21% of flights had arrival delay, which is defined as arriving more than fifteen minutes after the scheduled arrival time (Chokshi, 2020). This is concerning because almost a quarter of all flights in 2019 were delayed. Moreover, from 2012 to 2018, the volume of carbon dioxide emitted by commercial airplanes has risen 32% (The Guardian, 2019). This increase is not directly related to airplane delays, but is an issue in the industry that is caused by the increased reliance on air travel. This work aims to help solve one issue facing the air travel industry, arrival delay, by applying machine learning models. More specifically, this study hopes to improve on previous work by using a larger sample size and utilize regression models which capture information from the output variable, arrival delay, better than classification models.

2 Related Works

There are several previous related works on machine learning models to predict airplane delays. In Kuhn and Jamadagni (2017), the most recent related work, decision tree, simple neural network, and logistic regression classification models were used to predict a binary value of whether an airplane would have a delayed arrival time. The authors used the definition that, in the US, a delayed arrival time is when an airplane that arrives 15 min after the scheduled arrival time for their classification models. Similarly, the authors used the US Bureau of Transportation Statistics for the domestic airplane data. Kuhn and Jamadagni (2017) subsetted the database to 50 thousand flights without arrival delay and 50

*Email: vikas.thoti@gmail.com.

thousand flights with arrival delay. Moreover, from the 13 features known in advance of the flight, the authors used a decision tree classifier model to determine the three most important features that affect whether a flight has a delayed arrival. These three features were then used to retrain a new decision tree classifier and train a simple neural network and logistic regression. A ten fold cross validation was used to train the models chosen. These models were tested using 30 thousand samples with an almost equal split between flights with and without arrival delay. The test results showed that all three models, decision tree, simple neural network, and logistic regression, had an F_1 score of 0.91. In contrast, in training the models, this study used cross validation to remove biases from overtraining and used a greater selection of features and samples to train and test the machine learning algorithms. Moreover, this work uniquely tested more complex versions of the classic decision tree model through bagging and boosting algorithms, and used regression models for continuous arrival delay predictions. Random forest and AdaBoost, examples of bagging and boosting algorithms, were designed as improvements to the classic decision tree model, but whether these more complicated models work better was explored in this study.

3 Data Set

The data set used to train and test the machine learning models came from the US Bureau of Transportation Statistics. The data set contained information about US domestic flights that were scheduled throughout 2015 with many features. Features chosen to train the models were ones that are typically known ahead of the departure of the flight: month, day of the week, airline, origin airport, destination airport, departure time, scheduled departure, departure delay, scheduled time, and scheduled arrival. These features were the parameters that would help the models predict the delay of an airplane. Additional features also included in the data set, but not used for the study were flight distance, elapsed time, air time, wheels off, taxi out, day of the month, flight number, tail number, wheels on, taxi in, diverted, cancelled, and the cancellation reason. The final feature looked at was the arrival delay of each flight; this variable was the chosen output feature that the models predict.

To clean the data, the flights in the data set that were cancelled were deleted. A cancelled flight had no substantial information recorded because the flight was cancelled. For the purpose of this work, deleting cancelled flights did not have an adverse effect on the models. Additionally, in the real world, there is no purpose in predicting whether a cancelled flight will have a delayed arrival because it will never reach its destination; inherently, it made sense to delete these flights because without a recorded arrival delay, cancelled flights cannot contribute to model creation.

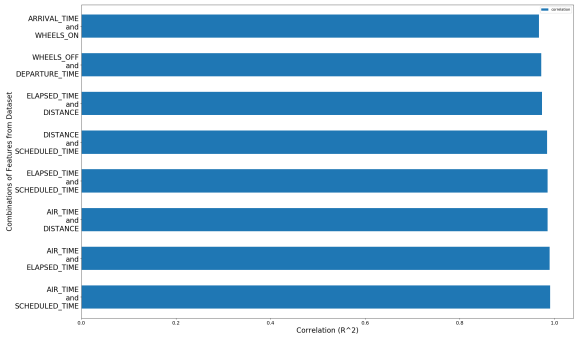

FIGURE 1: Correlations Between Variables in Data Set. The top eight correlations from the correlation matrix of the features from the data set are shown in the figure. Correlations are measured using the R^2 metric.

3.1 Categorical Values Replaced with Dummy Data

The python sklearn library for machine learning regression models can only take in features that can be converted into float format; as a reulst, categorical features have to be recoded. They cannot take in non-numerical categorical data as features that contribute to the output variable prediction. In response to this issue, the categorical variables in the data set, airline, origin airport, and destination airport, were converted into dummy variables. The essential idea behind this concept is to take string values for categorical data and assign numerical values to them. For example, a number was assigned to each airline and replaced the string value representing the specific airline of each flight. Now, categorical data can still be incorporated into the model because the data type has been changed to an integer which can be changed to a float. Even though, to a human, the airline column has been reduced to a bunch of numbers, the sklearn models can now incorporate this feature into its output prediction. Unfortunately, the flights from the month of October had to be removed because the origin and destination airports were given in a different format compared to the other months. As a result, dummy variables could not be assigned, so they were removed. Because deleting a two features, origin and destination airport, from the data set is much worse than just deleting the month of October, the latter was done.

4 Methods

The full data set was first subsetted to one million samples because of the computational limits of the computer used and the time constraints. The one million samples were then split into a training and testing set with a 80-20 split respectively. The Models, decision tree regressor, random forest

regressor and AdaBoost regressor, came from the sklearn python library. Using the training data, a five-fold cross validation was run on three models selected to initially train the models with various hyperparameters. Then, after training the data, the combination of hyperparameters that had the highest cross validation test score for each model was selected for the testing set that was split off earlier. The models took in the input features in the testing set and predicted a continuous value for the output variable: arrival delay. The model predictions for the testing set were compared to the actual arrival delays of the flights in the testing set and and scored using the R^2 metric. The R^2 metric was used to score the three regression models selected.

4.1 Regression Model

Since this study focused on predicting a flight's arrival delay as a continuous output, decision tree, random forests and AdaBoost regressors were used. Regression models yield a continuous variable while classification models have a binary output. Both types of models have their uses based the purpose of the machine learning algorithm. In this specific situation with flight delays, a regressor offers more specific information about the output variable, arrival delay. A classification model will predict whether there will be arrival delay, but it collapses a continuous variable into a binary output, which hides important information. In relation to airport traffic management, an estimated time of a flight's arrival delay, as a continuous variable, has more importance and value compared to a binary value of delay. Knowing approximately how late a flight will arrive can help airport traffic management make more efficient decisions to address the delay; however, simply knowing whether a flight will arrive late or not does not offer the same utility to airport traffic management. As a result, regressor versions of the decision tree, random forest, and AdaBoost models were used for the study versus the classifier counterparts.

4.2 Decision Tree

Decision trees are straightforward models that focus on creating statistically significant splits to eventually predict the outcome. This model resembles a tree and starts off with a root node that focuses on a specific feature of the data set. Based on the value of the feature selected for that specific data point, the model will move onto a system of internal nodes that will focus on different features of the data point. Eventually, the internal nodes will lead to a leaf node where the outcome variable is predicted. From the perspective of this specific data set, at the root node, the model might focus on the month of the flight. If the month that the flight took place was in winter, the data point will be taken along a specific route along the decision tree. However, if the month that the flight took place was in summer, the data point will go through a different set of internal nodes. Since the season has an important impact on flight arrival delay, splitting the data based on the month feature can help predict the arrival delay. Regardless of the path of internal nodes taken, the data point will eventually reach a leaf node where the delay time is predicted. The leaf nodes predict the outcome variable based on the internal and root nodes. Data points with very similar features will take similar paths through the decision tree and end at similar leaf nodes. A decision tree uses one large model that focuses on a variety of features in a data set. Compared to other machine learning models, decision trees are simple, easy to understand, and do not have any data type constraints. However, being such a simple model, there are inherent problems like overfitting and poor performance when introducing new data and outliers. When new data and outliers are applied to very convoluted decision trees, the model will not perform well since it has been overfitted to the training data. When decision trees are overtrained, they can become huge and very complicated with many unnecessary parts that do not contribute to the outcome prediction. This is a heavy disadvantage of decision trees. Regardless, decision trees are simple and very useful.

The specific hyperparameters that were manipulated for the decision tree model were the number of leaf nodes and max depth. The number of lead nodes was varied because these are the final outputs to the decision tree. Changing how many leaf nodes a decision tree has will impact its model performance. Additionally, the max depth of the decision tree will change the max amount of levels the decision tree is allowed to have. Making the max depth too large can over complicate the model, creating areas that do not contribute at all to the overall decision tree. Making the max depth too small can limit the amount of features used to predict the outcome variable. Therefore, changing max depth will also influence the performance of the decision tree model. The range for the leaf nodes was from 23 to 26 and the max depth varied between 15 and 19.

4.3 Random Forest

To fix the inherent problems with decision trees, an updated model, random forest, was created based on the principle ideas behind a decision tree. Instead of using one large decision tree, random forests utilize a technique called bagging that uses multiple smaller decision trees called weak learners. Random forests, a bagging algorithm, create weak learners in a parallel fashion: one after the other. Each weak learner produces an output prediction for a data point, but the final outcome prediction is based on the results of all of the smaller learners. For example, in a random forest regressor, the output prediction from the multiple weak learners will be averaged to produce a more accurate result. Because multiple weaker decision trees are used in a random forest, one larger convoluted model will be avoided. In turn, this would

prevent overfitting and poor performances when introduced to outliers and new data. When using a significant amount of weak learners to predict a specific outcome based on a variety of features, the disadvantages related to regular decision trees can be avoided. However, the main key drawback of random forests and other more complex models is that they are very slow and take a significant amount of time to run. In the real world, dealing with large data sets, creating multiple decision trees at the same time, and using all of them to predict the outcome of one data point in a database will take an extended period of time to complete.

For the random forest model, the number of weak learners and max depth of each learner was varied. Since random forest uses multiple weak learners to predict an outcome variable, this hyperparameter is important to vary. The number of weak learners ranged from 42 to 50 across all of the random forest models trained. Similar to the decision trees, the max depth was varied for the random forests; however, the values used were much lower because random forests use much smaller weaker learners. As a result, the max depth for the random forest weak learners was either three of four.

4.4 AdaBoost

Furthermore, instead of creating weak learners in a sequential fashion where all trees have equal weights in the final outcome prediction, AdaBoost varies the weights of the trees based on its ability to predict the outcome feature. AdaBoost is a sequential algorithm that uses a method called boosting. Boosting is when iterations are made to the weak learners based on the previous one's flaws and mistakes. Therefore, each weak learner benefits from the previous ones that have already been created. Unlike random forests, each weak learner is given a different weight when predicting the final outcome. The weights for each weak learner are determined based on the iterations and mistakes each learner makes. Through boosting, a more accurate model can be created. When compared to random forests which average the results from the multiple weak learners, AdaBoost builds on each weak learner through an iterative process and assigns weights to each specific learner; as a result, AdaBoost uses a theoretically more accurate process compared to its predecessors.

Finally, the hyperparameters that were varied for the AdaBoost model were the number of weak learners, max depth of each learner, and learning rate. Because AdaBoost is more complex compared to random forest, in the interest of time, a range of lower numbers for the number of weak learners was used: between 25 and 29. Increasing the number of weak learners used will increase the amount of time needed to create and train the models. Similar to the random forest models, the max depth for each weak learner was either three of four. Finally, the learning rate for the AdaBoost varied between 0.088, 0.089, and 0.090. The learning rate manipulates the weights of each new weak learner to the overall prediction.

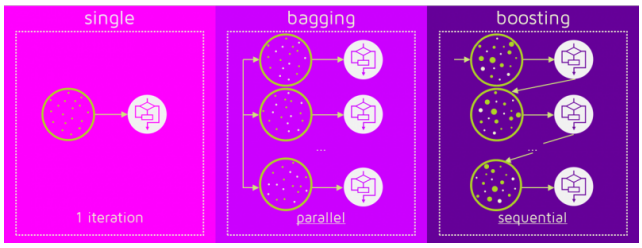

FIGURE 2: From XRISTICA. (2016). What is the difference between Bagging and Boosting?. Quantdare. Image downloaded from Quantdare. Single, Bagging, and Boosting Comparison Diagram. Figure shows the difference between the three types of algorithm. Decision tree is a single algorithm with one learner. Random forest is a bagging algorithm and AdaBoost is a boosting algorithms.

4.5 Cross Validation

The data set must be used to both train and test the selected machine learning models. However, the approach taken to train and test the models must avoid overtraining. Using a full set of data to both train and test the models can cause major issues. If the model is tested using the same data that was used to train it, the recorded model performance score from testing can be misleading. When these models are applied to new data in the real world, the algorithms may not perform nearly as well as they did in the testing phase. Instead, the model performance score from the testing phase should be an accurate indicator of how the models will do in the real world with new data that it was not trained on. This goal can be accomplished by doing a cross validation where the large data set is split up into pieces that each play a separate role in creating the model. Cross validations are run to prevent the models from being tested on the same data set they were trained on.

A cross validation was performed on the models using the training set that contained 80 % of the subsetted data meaning a total of 800 thousand flights were used to train the models. The 800 thousand flights were used to perform a five-fold cross validation. This means the training set was split into five groups. One of these groups would be left from training the model and used to test the model after; this would yield a training and testing score. The model would be trained five separate times where each time one of the five groups would be left out to test the model. As a result, there will be five different training and testing scores. To measure overall performance of a model on the training set, the mean training and testing score was found from the five-fold cross validation. This process was done for each set of hyperparameters for each of the three types of models.

After running all of the cross validations, the set of hyperparameters with the highest cross validation test score for each of the three models was selected. Moreover, the difference between the train and test score was also looked at. A large disparity between train and test scores in a cross validation indicates that the model will not do well when presented with new data that it was not trained on. In short, the best combination of hyperparameters from the range of values inputted will be determined by using the cross validation mean test score.

TABLE 1: Decision Tree Cross Validation Results. This table shows the results of the cross validation for the top three sets of hyperparameters for the decision tree models. The scoring metric used was R^2.

Max Leaf Nodes	Max Depth	Test	Train
26	19	0.887	0.887
26	18	0.887	0.887
26	17	0.887	0.887

TABLE 2: Random Forest Cross Validation Results This table shows the results of the cross validation for the top three sets of hyperparameters for the Random Forest models. The scoring metric used was R^2.

Learners	Max Depth	Test	Train
50	4	0.884	0.885
48	4	0.884	0.885
49	4	0.884	0.885

TABLE 3: AdaBoost Cross Validation Results This table shows the results of the cross validation for the top three sets of hyperparameters for the AdaBoost models. The scoring metric used was R^2.

Learners	Learn Rate	Max Depth	Test	Train
27	0.09	4	0.889	0.889
28	0.09	4	0.889	0.889
29	0.09	4	0.889	0.889

Train and test columns in Tables 1, 2 and 3 refer to the mean train and test scores from the five fold cross validation. These tables show the results of the top three sets of hyperparameters in terms of test scores for the three machine learning models. The cross validation train and test scores for all three models are similar and not significantly different.

5 Results and Discussion

The testing set that was initially split off and set aside had 200 thousand flights as per the 80-20 split of the original 1 million. This testing set was used to test the final three models that were chosen. The models were chosen by look-

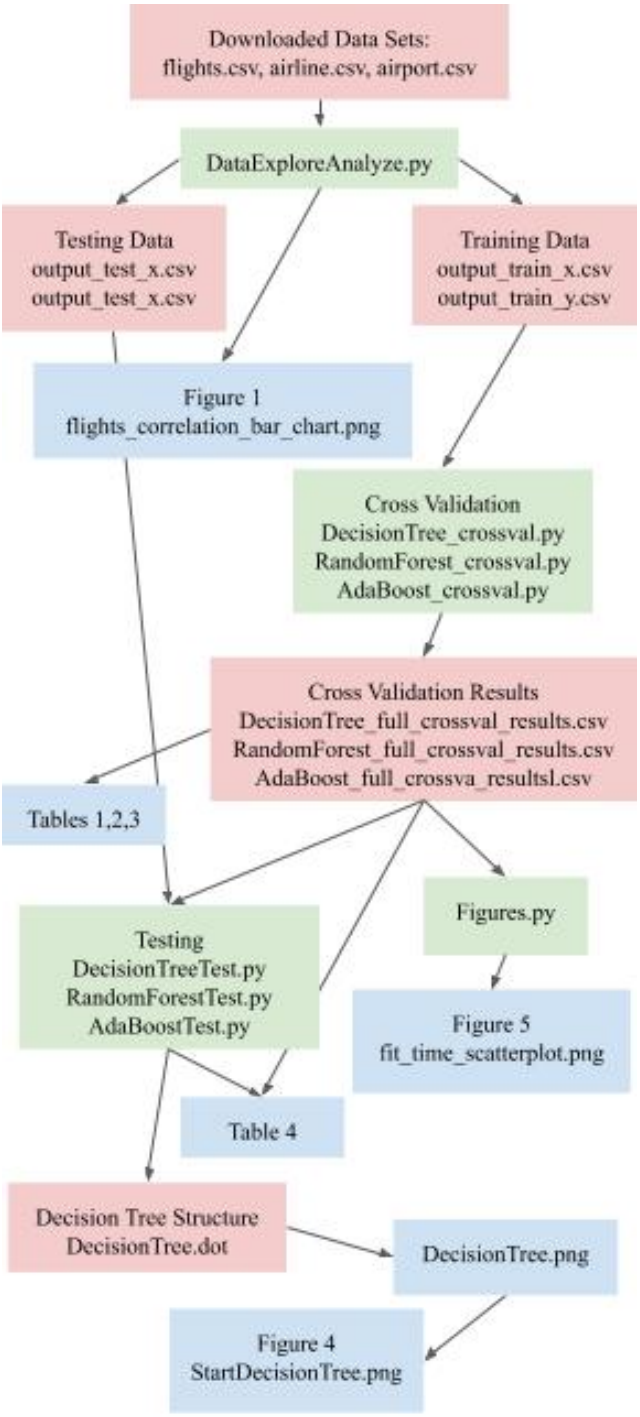

FIGURE 3: Work Flow Diagram. This diagram shows the various data files, python files and charts/tables used for the study. Data files are red, python files are green, and charts/tables are blue. Arrows show what files were used to create the subsequent file. All files except extremely large data files are available on the Github repository for this work. https://github.com/vikasreddy17/AirplaneDelaysMLAnalysis

ing at the mean test score from the five fold cross validation performed. One model was picked from each of the three different machine learning algorithms. The set of hyperparameters used to yield such models were used to create the final algorithms that would be tested on the testing set. As shown in Table 1, the hyperparameters that yielded the best decision tree model were a max leaf nodes of 26 and a max depth of 19. Furthermore, as shown in Table 2, the random forest model with the best mean cross validation test score was the one with 50 weak learners which each had a max depth of four. Lastly, as displayed in table 3, the AdaBoost model with the best mean cross validation test score was the one with 27 estimators, a learning rat of .09 and a max depth of 4 for each weak learner.

TABLE 4: Results of the Decision Tree, Random Forest, AdaBoost Models Chosen for Testing. The table shows testing and cross validation performance of the models selected for the final testing set. Fit time for the models is also included. Cross validation is represented as CV in the table. . Scoring used is R^2 metric.

Model	Decision Tree	Random Forest	AdaBoost
Testing	0.874	0.867	0.876
Fit Time	2.00	69.92	55.15
CV Train	0.887	0.885	0.889
CV Test	0.887	0.884	0.889

The decision tree, random forest, and AdaBoost models with these specific hyperparameters used the features from the testing set to predict the continuous arrival delay output. These predictions were then compared to the actual arrival delay of the flights in the testing set; then, a score was calculated using the R^2 scoring metric. Because regression models were used, R^2 seemed an appropriate scoring metric to measure model performance. The testing scores of the three algorithms are shown in Table 4.

5.1 Final Model Selection

The AdaBoost model did the best among the three models, followed by the decision tree and the random forest model. As shown in Table 4, it is very important to point out that the testing scores for the three models are not significantly different from each other, so the small differences in the r-squared metric could be explained by the randomization of the selection of flights in the testing set; however, the results from the study are still significant. The bagging and boosting algorithms in this case did not perform significantly better than the base decision tree algorithm. With fit time in mind, the decision tree model is the best option. The decision tree model did slightly worse than the AdaBoost model but took significantly less time to run; Therefore, the decision tree model chosen for the testing set is the best of the three regression models when taking into consideration accuracy as well as the time it takes to train and fit the model.

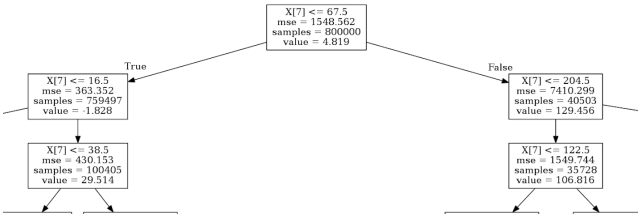

FIGURE 4: Peliminary Nodes of Final Decision Tree Model. This diagram shows the decision tree structure for the model that had a max leaf nodes of 26 and a max depth of 19. This model had the highest cross validation R^2 value among all other decision tree models with different sets of hyperparameters. This model was also chosen has the final model to predict the arrival delay of flights. Full decision tree is very large with 16 leaf nodes and a depth of 7.

5.2 Complexity and Run Time Trade-off

The bagging and boosting algorithms, theoretically supposed to perform better than the base algorithm, did not have a higher testing score. This study is evidence for the claim that bagging and boosting algorithms may not perform significantly better than simpler models 100 percent of the time. With model fit time in consideration, a simpler model that performs slightly worse or maybe even better than more complex models in significantly less time may be of better value. Bagging and boosting algorithms create multiple decision trees and use iterative processes requiring a significant amount of time to run; comparatively, the base algorithm, decision tree, takes significantly less time to run. In a world which emphasizes quick results, simpler models may be more useful in certain situations.

To expand on the tradeoff between complexity and time, Figure 1 shows how, in this case, the decision tree models were in the optimal place with a low fit time and a similar cross validation test performance when compared to the other more complex models. In this case, keeping this tradeoff in mind, the best model to predict a flight's arrival delay is the decision tree model. Even though the decision tree model with the best set of hyperparameters does not significantly out-perform the bagging and boosting algorithms on the testing set, the fit time is significantly less; therefore, the decision tree model with the best hyperparameters is more useful in this case because of its relatively quick run time compared to its more complex successors. As shown in this case, the simplest model, the decision tree, did not do much worse than the other two more complex models; with model fit and run time in mind, the answer to which model to choose, in this scenario, becomes overwhelmingly easy: decision tree.

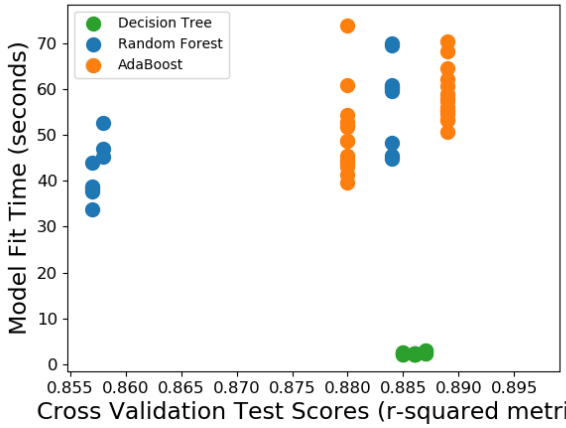

FIGURE 5: Comparison Between Model Fit Time and Cross Validation Test Scores. This scatter plot shows the three models selected and the relationship between their cross validation test scores and fit time.

5.3 Improvement from Previous Works

This work uses a very large sample size of one million total flights which is a significant boost to the amount used by previous works. This increased sample size in conjunction with the regression models will offer more information to air traffic management. Because an increased sample size was used for training and testing, the model performance from the study more accurately represents how well the model will do in the real world. Additionally, because regression algorithms were used, the final decision tree model will yield a prediction of how late an airplane will arrive based on the ten input variables used to train the model. By expanding arrival delay from a binary output to a continuous variable, the models created in this study offer more unique predictions compared to previous works. Air traffic control benefits from a continuous prediction of arrival delay because it allows them to more effectively manage the delay. With a specific and accurate prediction of how late a flight will arrive, airplane delay management at airports can more effectively adjust schedules to accompany an arrival delay. The models in this study did not perform as well as previous works, but a more harsh scoring metric, R^2, was used to score the models. R^2 is a harsher scoring metric compared to most classification metrics because it takes into consideration specific continuous values and predictions. Since the models created in this study were trained/tested on a greater sample of data and predict arrival delay as a continuous variable, this study improves on previous works done on this topic.

6 Conclusion and Future Works

This study focused on comparing different levels of complex machine learning algorithms (single, bagging, and boosting) and using a different approach through regression models to predict arrival delay for flights. The benchmark that bagging and boosting algorithms will always do better than their more simple counterparts is not true in all cases. In the future, more features such as weather patterns and distance of flight can be included when training and testing the models. Additionally, the structure of each model could be looked at more deeply to understand how exactly the machine learning models are forming. Lastly, a different set of regression models could be tested to see if they perform better than the three selected for this study.

References

Airlines for America. (2020). Annual U.S. Impact of Flight Delays (NEXTOR report). Airlines For America. https://www.airlines.org/data/annual-u-s-impact-of-flight-delays-nextor-report/.

Bureau of Transportation Statistics (2015). https://www.transtats.bts.gov/ONTIME/Departures.aspx.

Busson, T. (2020, January 9). Why is My Flight Delayed? The 20 Main Reasons for Flight Delays. The Claim-Compass Blog. https://www.claimcompass.eu/blog/why-is-my-flight-delayed/.

Chokshi, N. (2020, February 19). Airline Flight Delays Got Worse in 2019. Here's a Scorecard. The New York Times. https://www.nytimes.com/2020/02/19/business/air-travel-delays-airlines.html.

The Guardian. (2019, September 19). Airlines' CO2 emissions rising up to 70% faster than predicted. The Guardian. https://www.theguardian.com/business/2019/sep/19/airlines-co2-emissions-rising-up-to-70-faster-than-predicted.

Kuhn, N., Jamadagni, N. (2017). Application of Machine Learning Algorithms to Predict Flight Arrival Delays. Project Posters and Reports, Fall 2017. http://cs229.stanford.edu/proj2017/final-reports/5243248.pdf.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

XRISTICA. (2016). What is the difference between Bagging and Boosting?. Quantdare. https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/.