Predicting Flight Arrival Delays

Doğancan Gemici, Mayasa Dablan, Samer Kanakri
Istanbul Sehir University
34865 Dragos, Istanbul, Turkey
{dogancangemici,mayasadablan,samerkanakri}@std.sehir.edu.tr

24/12/2018

Abstract

With the increasing population around the world and the growing aviation industry which leads to an increase in air-traffic causing many flight delays. Flight delays don't have only an impact on the economy but they also have some bad effects on airlines, airports, and passengers. In addition, being able to create accurate prediction models for arrival delays became problematic because of the complexity of the airline systems, and the flooding flights data. In this project, we applied multiple machine learning algorithms, such as linear regression, logistic regression, SVR, and decision tree, in order to predict whether the arrival of a flight would be delayed or not. We used features obtained after applying feature importance and grid search to determine the best parameters for each model, to eventually apply best acquired model, and we were able to accomplish prediction with an approximate R squared score of 0.37 and an root mean squared error of 37.1.

1 Introduction

As many travelers around the world are preferring to choose air flights specifically because it consumes less time to travel from one place to another. Moreover, because it is more comfortable and faster. Nowadays many of passengers are suffering to choose a reliable flights or airport and the passengers' decision of air travel can be expected to be affected by the delay information. On the other hand, flight delays became a significant problem not only for passengers but also for flight scheduling and airports managing. According to [1], In Europe each year there are more than 2.4 million flights which are delayed or cancelled due to a variety of factors such as, weather conditions, air traffic controls, airlines. Although many airports in different countries spent a lot of efforts to reduce flight delays but still in some airports flight delays are not avoidable. Delay has become an essential problem for the air transportation systems. Despite the fact that flight delays have negative effects on passengers, airports, and airlines. They also can have a huge impact on environmental and economical sides; where fuel consumption can be increased, plus, according to Federal Aviation Administrative statistics which showed that flight delays cost the airline industry each year approximately $8 billion, and cost passengers around $17 billion [2]. In this project, we have selected multiple column features from the dataset such as Departure Delay, Scheduled Arrival Hour, Scheduled Departure Hour, etc. as an input to our Machine Learning algorithms, in order to predict the flights' arrival delay and whether a flight is going to be delayed or not. We have used Decision Tree classifiers to apply the prediction, as well as we compared results with other methods, like logistic regression and linear regression.

2 Literature Review

Through the last decade there were many literatures which were made for predicting the flight delays, and how to reduce them. In [3], the authors used a decision tree classifier and compared it with logistic regression and a neural network. Back-propagation along loss- function was used to get the parameters of the network. Moreover, L2 regularization was used to prevent their model to over-fit for the logistic regression and neural network. The authors of the paper were able to predict the flight delay, show the classifiers which did the prediction, and get a satisfying accuracy results. In [4], the main focus of the author in this paper was to show how flight cancellation and delays can affect the airports and passengers in small cities. Another approach of calculating the arrival delay of flights in small cities' airports depending on the flights which were cancelled, by comparing the original itineraries with the itineraries that changed due to the cancellations. Furthermore, the author clarified that the departure delays and cancellations played a major role in arrival delays and with huge rebooking limitations' options. Whereas, in [1], the authors aimed in their paper to focus on the interdependency over a sequence of the flight delays which were due to various factors such as weather, operations in airports, and air traffic control conditions. This paper's main goal was to analyze the propagation effects of flight delays through using copula-based approach, to get the correlation between flight delays regarding the delay factors. Appropriate scenarios were created to test the possibilities of reducing the delays propagation through improving the flights' schedules.

3 Data Exploration

Since we were supposed to train our models, we have used the given dataset which is randomly collected from thousands of flights all over the world during the year of 2017. The provided dataset contains 90,000 rows where each row represents a flight, and 18 columns representing the features of these flights which include:

i. Date information (Month, Day of the month, Day of the week, Days to Holiday).

ii. An identification number assigned by IATA to identify a unique airline of the flight (Unique Carrier).

iii. Flights arrival information (Scheduled Arrival Time, Actual Arrival Time, Scheduled Arrival Hour, Actual Arrival Hour, Arrival Delay).

iv. Flights departure information (Scheduled Departure Time, Departure Time, Scheduled Departure Hour, Departure Hour, Departure Delay).

v. The flight's origin and destination.

vi. The flight's distance.

The Dataset was so is messy, and missing a lot of features that have more effect on the flight delay which could have made our job easier, but at the same time it also included many unnecessary features which made our job harder to check which set of features are the most relevant to our problem, and work with it. For Null values we filled them with the mean of the specified column. Then we split the data into training and testing with a ratio of 20% to 80% each respectively, and prepared them for training. Starting with detecting the best features to use, first by checking for correlation between features, but correlation didn't help.

4 Methodology

As we have mentioned previously, we have used multiple machine learning algorithms to train our models and predict flight arrival delays, starting with weak learners such as Linear Regression, moving to more complex and adaptive ones like Logistic Regression, SVR, and Decision Tree, ending with boosting models. In this section of the report, we go through the methodology we applied including a general description of each method we used. After splitting the dataset into training and testing by a ratio of 20:80 respectively, using only 6 features out of 18 which are the most important for our purpose. Starting with Linear Regression Algorithm which tries to fit the best line possible that represents the data, and since the data is not linear we had to move to more complex algorithms. Moving to Support Vector Regressor algorithm which is used to work with continuous values unlike other methods working with classification such as SVC. In this algorithm, the error is fit within a certain threshold, however the resulted scores were not promising.

4.1 Decision Trees

We turned our focus to Decision Trees, The main purpose of using decision trees is that it builds a model just like a tree from the root node till the leaves which represent the prediction we aim for, in each node of the tree it asks a Boolean question and according to the answer it splits the data into two subsets, after that these subsets become the input to the next child nodes, and these nodes ask another question again to one of the other features, this process maintains the interpretability of the model unlike other methods used before, but we still need stronger learners.

4.2 Random Forest

Since the data is big and algorithms like SVR and random forest requires time and consume hardware resources, we moved our work to Google colaboratory, which provided us with the necessary speed and resources to apply more complex algorithms. We picked Random Forest as our new learner, random forests is an ensemble model of many decision trees, a Random Forest Regressor instantly resulted with better r2_score of 0.33 and RMSE score of 38.1.

4.3 Adaptive Boosting

Now our new aim is to make this score better, by going back to the base of a random forest, a decision tree. We applied a grid search on a decision tree regressor to get the best parameters as follows random_state=0, max_depth=90, max_features=3, min_samples_leaf=5, min_samples_split=5, also applied feature_importance to get the most important features as well. Although the results were not promising for a single tree model, even with best parameters we got r2_score of

0.09 and RMSE score of 44.8, we applied Adaptive boosting (also known as AdaBoosting) to the best parameter decision tree to get even better results. Boosting originally called hypothesis boosting, refers to any ensemble method that can combine several weak learners into a strong learner. The general idea of boosting methods is to train predictors sequentially, each trying to correct its predecessor.The trick is to pay a bit more attention to the training instances that the predecessor under fitted [5]. after searching for the best number of estimators for our boosting model, the scores were actually promising and even higher than the random forest scores with best parameters, ending with an r2_score of 0.39, RMSE score of 36.7 with number of estimators 78 as shown in Figures 1, and 2.

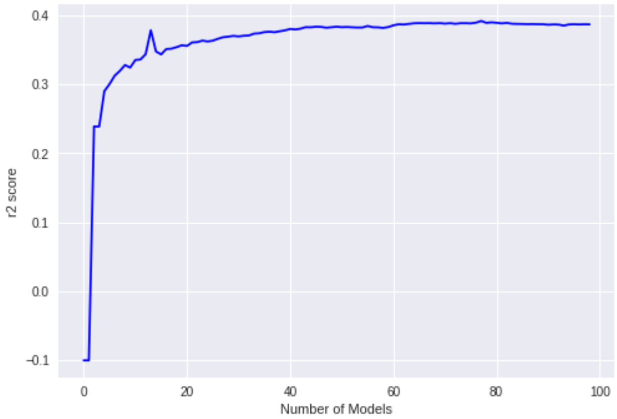

Figure 1: number of estimators against R squared score

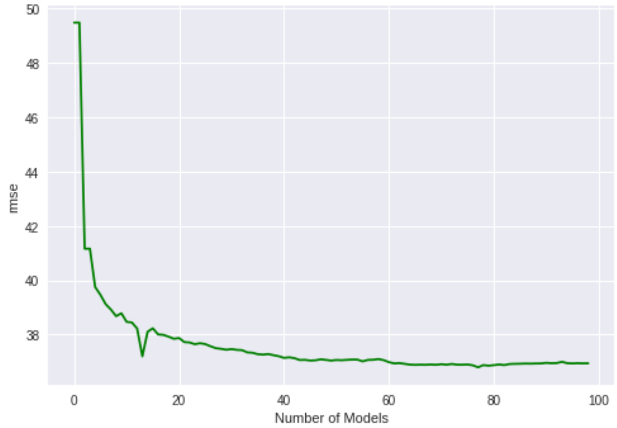

Figure 2: number of estimators against RMSE score

5 Conclusions

In conclusion, given the dataset provided, we were able to find a good estimation of the flight arrival delay, getting advantage of boosting methods to get a better score out of a simple decision tree model to get better predictions. As a future work we would like to extend our methods for prediction on more representative features which may lead to a better results on predicting the delays, features such as Taxi delays, flight Cancelation, and weather delay, etc.

References

[1] T. F. H. Z. Weiwei Wu, Cheng-Lung Wu and S. Qiu, "Comparative analysis on propagation effects of flight delays: A case study of china airlines," 2012.

[2] A. Press, "Flight delays are costing airlines serious money," 2014.

[3] N. Kuhn and N. Jamadagni, "Application of machine learning algorithms to predict flight arrival delays," 2017.

[4] M. J. Stone, "Impacts of flight delays and cancellations on travel from small community airports," 2015.

[5] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly, 2017.