Final Report for Biostat 625

Airline on-time performance

Group 3

12/17/2021

1 Introduction

With air traffic restrictions and the rapid increase in the number of flights, flight delays are becoming more frequent and has become is a critical issue to airlines and customers associated with significant costs to the society. High costs of delay motivate scientists to analyze the manner of delay and use models to predict delay time. Flight delay can be driven by many factors; weather, distance, carriers, security, origin airports, etc. Previous studies have studied the propagation of delays, estimated distributions of the flight departure delay, and impacts of delays on passengers. Here we use the public data of all commercial flights within the USA from 2003 to 2008 with flight arrival and departure details from expo2009, to investigate the prediction of departure time, with consideration of roles of departure time, arrival time, origin airports, carrier and other related covariates contributing to flight delay. We use random forest method and Gradient Boosted Decision Trees(GBDT) method to predict flight delay and find the association between flight delay and interested factors within our data set.

2 Computational Challenge and Data Preproccesing

We face tremendous computational challenge when analyzing this data set. First, the data set is extremely large which takes up 4.87 gigabytes when uncompressed. Data from 2001 and 2002 are also problematic that it is not correctly coded. To ensure continuity in Year, we use data from 2003 to 2008. It need to be mentioned that choosing more recent data could give us a better model since number of flights has greatly increased since 2003 hence leads to a different delay pattern. After cleaning data, there are more than 41 millions observations in total. Such a large data set brings us a lot of difficulties in data reading and fitting model. Hence, we use tidyverse package in R and do most of our data cleaning and modeling with python which has nice packages equipped with parallel computing. Also, most of code for modeling are submitted to biostat cluster, which has larger memory and allow parallel computing to improve efficiency.

Based on literature review, we select follow variables to fit our model.
Outcome: In our project, we focus on the *DepDelay* which is departure delay time and treat it as categorical outcome *depdelayC*. It is coded as 1 if DepDelay is more than 15 and 0 otherwise.
Covariates: Covariates include several time covariates, flight carriers, origin and destination airports and distance. Time variables include year, month, day of month, day of week, and estimated time of departure and arrival *CRSDepTime* and *CRSArrTime*. Time variables are numeric variables in the original data set and we treat them as factors. Month has twelve levels and day of month has 31 levels. *Origin* and *Dest* indicate origin and destination airport of each flight. *Distance* refers to the distance between rigin and destination airport of each flight which the unit is in kilometers. Since most of these covariates are categorical, python automatically formulate 733 dummy variables when modeling, which also add up the computational challenge.

We Also plot simplified delayed flight route in USA from 2003 to 2008, which only flight routes with over 10 fights per day is considered. It shows a significant pattern that LAX, ORD, ATL and ERW is four delay center in USA. Actually, this are four busiest airport in USA

Figure 1: Delayed Flight Route in USA from 2003 to 2008

3 Methods

We fit Random Forest and GBDT model on 10 selected variables to predict airline on time performance. The data set is split to training set and testing set by 7:3. Performance of the two models are compared on testing set.

Random Forest

Random Forest is an ensemble learning method for classification by constructing a multitude of decision trees at training time. Compared to original bagging method for decision trees which repeatedly select random samples from training set, random forest take the procedure of feature bagging. It selects a random subset of the features at each split in the learning process. It successfully fixes over fitting problem in decision tree model and generate variable importance automatically but it's still a black box model lack of interpretation. Parameters in the models are selected via 5-fold cross validation. Weights are added to solve imbalance problem between delay and un-delay flights.

Gradient Boosting Decision Tree

Gradient boosting combines weak learner such as decision tree to strong learner by iteration. In GBDT, it aims to minimize the loss function by applying steepest descent to this minimization problem in each iteration of decision tree. Comparing to Random Forest, it has better performance in terms of generalization and precision accuracy. However, GBDT is more sensitive to outliers. Learning rate and number of estimators are two important parameters in GBDT which are associated with complexity of the model. It also adjust generalizability and over fitting problem. All parameters in the models are selected via 5-fold cross validation. Here, we point out that learning rate is 0.8 and number of estimators is 200 in our model.

4 Results

Feature Importances

Our random forest and GBDT classifier uses a total of 733 features(coverting every categorical features to binary one-hot coding).In order to find out which feature has the most impact on predict results, we use the feature importance plot.

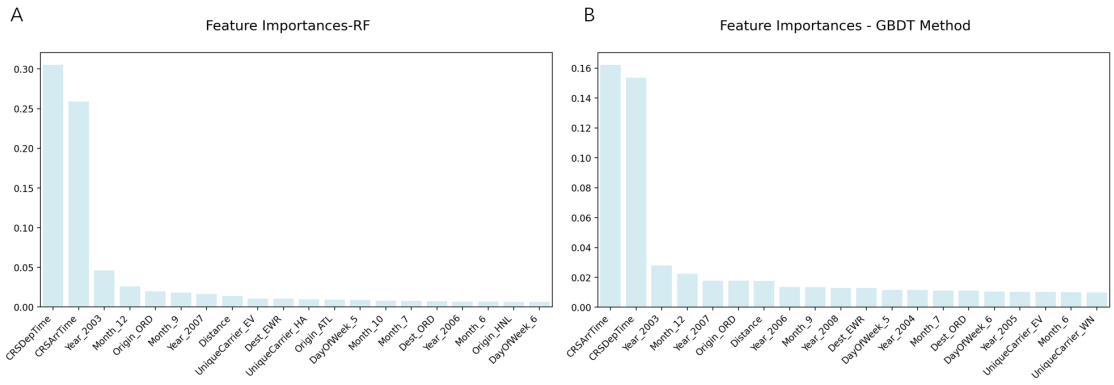

Figure 2: Feature Importance for RF and GBDT

From the plot above, we can see that, the most important features for both model is Departure time and Arriving time, month, year, and distance. Combined with the descriptive analysis above, we can know that flights between afternoon and midnight have a higher probability of being delayed, and flights at dinner time have the highest chance of being delayed. Also, flights have higher delays from June to August and with Christmas holidays.

RF	**Origin**	ORD	ATL	HNL	EWR	SLC
	Destination	EWR	ORD	IAH	SLC	SFO
GBDT	**Origin**	ORD	ATL	EWR	PHL	DFW
	Destination	EWR	ORD	ATL	LGA	SFO

Figure 3: Top 5 Important Airport in RF and GBDT

We list top five important airports in Origin and Destination, red for delay and green for non delay. Two models both show ORD, ATL and EWR to be airports with higher delay importance. It should be noticed that one of the largest airport as while as delay center LAX is not listed.

	Most Influencial Carriers						
RF	EV	HA	WN	DL	OO	NW	MQ
GBDT	EV	WN	OO	DL	US	MQ	NW

Figure 4: Top 5 Important Carrier in RF and GBDT

We list top seven important carrier for two models, red for delay and green for non delay. Two models both show EV, WN and MQ to be carrier with higher delay importance while DL, OO and NW has largest non-delay importance.

Classifier performence

The accuracy of the Random Forest model and GBDT model on predicting delays on the test set is 0.612 and 0.637. We want to be able to visualize how the classifier scores against delay or no delay, so we plotted the histogram of the classifier scores. The histograms for RF model are centered around 50% because we set the parameter "class_weight" to "balanced_subsample"(Figure 5). The scikit-learn GBDT implement cannot directly get balanced sample, so the histogram centered around 0.2 which is the fraction of positive(delay) in the dataset. We can see that the two groups are not completely distinguished from each other, perhaps because some key influencing factors are missing in our model.

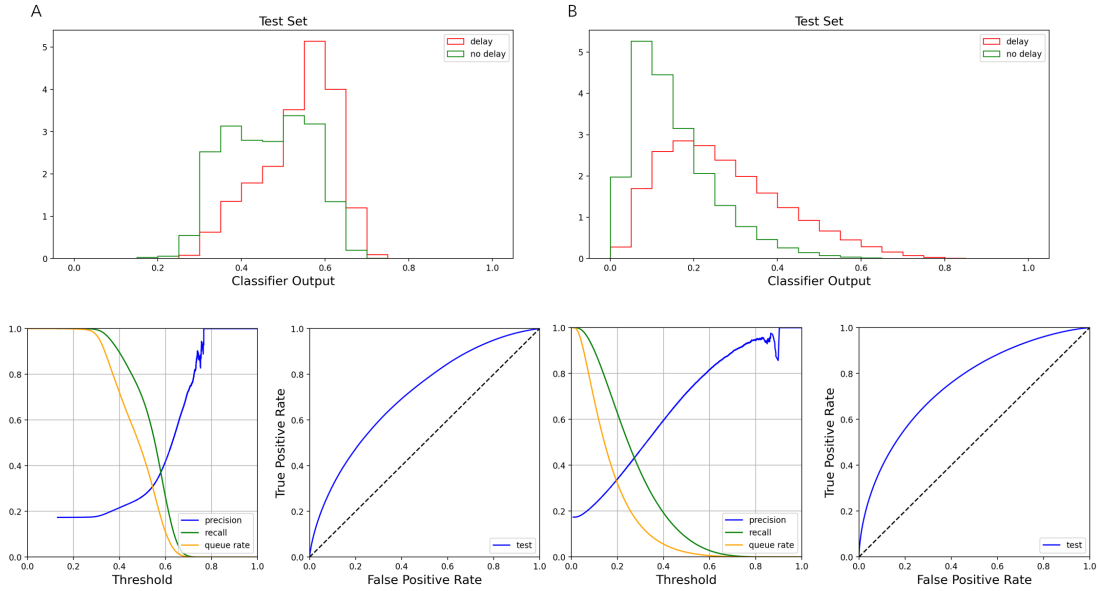

Figure 5: Performance of RF and GBDT

The overall precision and recall vs threshold plot are shown. In this plot we want to tell how the threshold change will effect the model precision preference. The queue rate is the fraction of datapoint that pass the threshold cut. As the threshold increases, fewer true positives pass the cut, leading to a decrease in recall that follows the decrease in queue rate and a increase in precision. Notice that at very high thresholds, the precision curves are not very smooth in both models.This may be due to the fact that at high thresholds, only a small number of true values will pass the screening, when the variation of true values can significantly affect the Precision value.

The ROC curve is a curve that measures the diagnostic ability(True positive and False Positive) when the threshold limit is changing. The area under the ROC curve is a good estimator for model performance and it is equivalent to the two sample Wilcoxon rank-sum statistic. That area of Random Forest model is 0.708 and 0.754 for GBDT model, indicating GBDT model has a better prediction performance than Random Forest Model in our project.

Model	Accuracy	Area under test ROC	Time Consumption	Peak Memory Usage
Random Forest	0.612	0.707	2:33:30	298.0G
GBDT	0.637	0.754	17:10:30	147.5G

Figure 6: Speed of RF and GBDT

Finally, we compare speed of two models on test set. It is shown by literature that GBDT is a less complex

algorithm than Random Forest. However, even Random Forest takes up much more memory space than GBDT, it is less time consuming. Actually, since Random Forest is a bagging algorithm, it allows parallel computing which is not possible for GBDT which makes the whole process faster.

5 Conclusion and Discussion

In this project we download the flight data set from Expo2009 website. Data are selected from the year 2003 to 2008 and 10 variables: *Year*, *Month*, *DayofMonth*, *DayofWeek*, *CRSDepTime*, *CRSArrTime*, *UniqueCarrier*, *Origin*, *Dest* and *Distance* are used to predict airline delay status *depDelayC*.

We train Random Forest and GBDT separately on training set. We only got an accuracy of 61%(RF) and 63%(GBDT) in the test data set and a ROC score of around 0.73. However, Random Forest is much more faster as it allows parallel computing.

Both Random Forest and GBDT model shows great importance of delay in ORD, ATL and EWR airport. It is not surprising that ORD, ATL and EWR are transportation hub in USA. Greater flights every day causes larger delay rate. However, one of the delay center in USA LAX is not listed. This is possibly because the relative low delay rate in LAX. It indicates that LAX airport may have done a excellent job in flight scheduling. We should pay more attention to flight delay in large airports such as ORD, ATL and EWR.

Also, EV, WN and MQ are carriers that have large importance in delay while OO, DL and NW are more important carriers in terms of non-delay. If we expect to experience less delay in flights during travel, we should choose OO, DL or NW as our carrier.

There are several limitations in our project.

1. Most of the characteristics in the data set are relatively static, and there is a comparative lack of dynamic information such as weather, which has a relatively large impact on flight delays. We could introduce this kind of information from other data sources in the model to improve accuracy.
2. Even with cluster computing and parallel computing, our data set is still very large, making it very difficult to reconcile the model. This may lead to over fitting problems and make the model less accurate in the test set.

Reference

[1] Data Expo 2009 - Airline on-time performance (https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009)

[2] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[3] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[4] Zonglei, Lu, Wang Jiandong, and Zheng Guansheng. "A new method to alarm large scale of flights delay based on machine learning." 2008 International Symposium on Knowledge Acquisition and Modeling. IEEE, 2008.

[5] Ball, Michael, et al. "Total delay impact study." NEXTOR Research Symposium, Washington DC. 2010.

[6] Rebollo, Juan Jose, and Hamsa Balakrishnan. "Characterization and prediction of air traffic delays." Transportation research part C: Emerging technologies 44 (2014): 231-241.

[7] Yazdi, Maryam Farshchian, et al. "Flight delay prediction based on deep learning and Levenberg-Marquart algorithm." Journal of Big Data 7.1 (2020): 1-28.

[8] Xu, Ning, et al. "Estimation of delay propagation in the national aviation system using Bayesian networks." 6th USA/Europe Air Traffic Management Research and Development Seminar. FAA and Eurocontrol Baltimore, 2005.