Predicting Flight Delays and Cancellations

(Final Project for CSYE7374)

Group 11 - Mengying Wang

Contents

Abstract 3

1. Background 4

2. Data Preprocessing 5

    2.1 Data Source 5

    2.2 Data Cleaning 5

    2.3 Data Format 6

3. Exploratory Data Analysis 9

4. Prediction 13

    4.1 Logistic Regression 13

    4.2 Random Forest 13

    4.3 CatBoost 14

    4.4 LightGBM 15

5. Conclusion 16

References 18

Abstract

When planning a trip, passengers should keep in mind that airlines do not guarantee their schedules, delayed and cancellation of flights always break our plans. Especially, airlines are cutting 20% to 50% of their domestic flight schedules by April 1, 2020, due to the coronavirus outbreak. So predicting the flight status before traveling is very necessary for passengers to plan a trip and prepare for the worst situation.

This project is predicting flight delayed by the data provided by The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics. The predicting is based on the airline companies, origin and destination, flight time and other factors in the dataset.

Delayed time was transferred to be a categorical variable, then I built some classifiers include Logistic Regression, Random Forest, LightGBM and CatBoost to train the data, LightGBM perform best in this project. Moreover, I tried to accelerate these classifiers by adding multiple CPU or GPU and got ideal results.

Keywords: *Flight Delay, Parallel, Machine Learning, GPU, Classifier*

1. Background

When planning a trip, passengers should keep in mind that airlines do not guarantee their schedules. While airlines want to get passengers to their destinations on time, there are many things that can – and sometimes do – make it difficult for flights to arrive on time. Some problems, like bad weather, air traffic delays, and mechanical issues, are hard to predict and often beyond the airlines' control. [1]

According to the Department of Transportation, 1.9% of scheduled flights were canceled and 21% were delayed by more than 15 minutes in 2019 – long enough to miss a tight connection. What's more, U.S. airlines are cutting 20% to 50% of their domestic flight schedules by April 1, 2020, due to the coronavirus outbreak.[2]

Especially during the coronavirus outbreak, many international travelers was stranded because of the flight delays and cancellations, so predicting the flight status before traveling is very necessary for passengers to plan a trip and prepare for the worst situation.

This project is predicting flight status include delay and cancellation by the data provided by The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics. The predicting is based on the airline companies, origin and destination and some other factors, this will be decided in Exploratory Analysis.

2. Data Preprocessing

(This Part in EAD.ipynb)

2.1 Data Source

Context:

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report.[3] It concludes 3 datasets:

(1). airlines.csv: information of airline company,

(2). airports.csv: Information of airports,

(3). flights.csv: Information for each individual flight.

Size:

The record in flights.csv include 5819079 flights and 31 underlying reasons.

Acknowledgements:

The flight delay and cancellation data were collected and published by the DOT's Bureau of Transportation Statistics.

2.2 Data Cleaning

Data cleaning is the first step for data analysis, cleaning the record with missing data and some useless columns, it will make the data analysis smoother and reduce the run time, especially for the huge database like the database I used in this project.

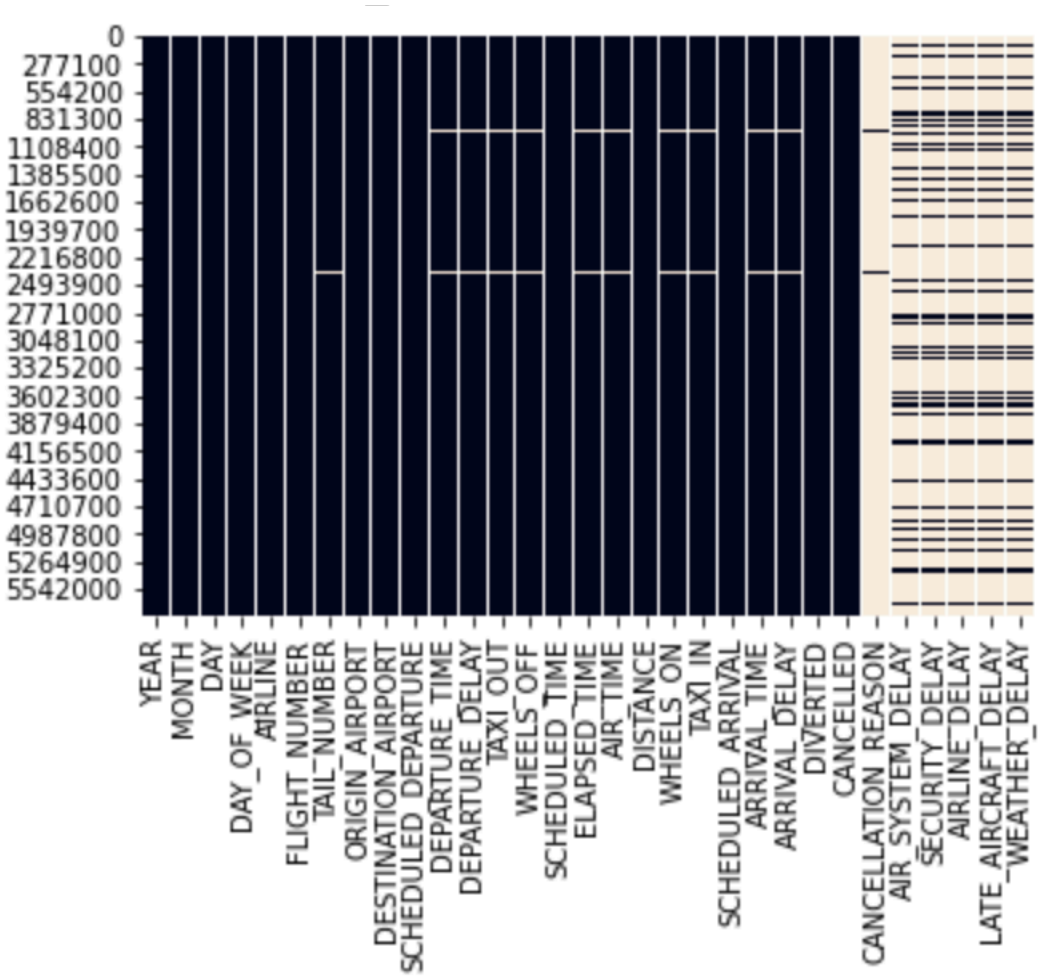

Figure 1 – Missing Data in the flights.csv

In Figure 1, I can see that there are several columns have too many missing data, such as AIR_SYSTEM_DELAY and CANCELLATION_REASON, so I dropped these columns.

Some other columns have the similar meaning, I will delete them optionally. For the rest columns, I will drop the lines with null data.

2.3 Data Format

Date and Time

This database include many date and time data, the major thing I have done can be divided into two parts：

The first part is merging YEAR, MONTH and DAY into a single value DATE, which is the date of the flight, then transfer it to type ‘datetime’.

The second part is deal with the column SCHEDULED_DEPARTURE. This variable means the hour of the take-off, it is coded as a float where the two first digits indicate the hour and the two last, the minutes. This format is not convenient for analysis and I thus convert it to date format. Finally, I merge the take-off hour with the flight date.

To proceed with these transformations, I define a few functions in preprocess.py.

Transforming

I transferred column DEPARTURE_DELAY from numerical variable to categorical variable as the prediction variable in this project.

As we always regarded the fight which is delayed more than 15 minutes as ‘Delayed’, so I added this variable, made the flight with more than 15 minutes delayed as 1, with means that this flight is delayed, others are 0, which means that this flight is on time.

Moreover, I made some changes for other two variables: ORIGIN_AIRPORT and DESTINATION_AIRPORT. For the 100 busiest airports, I maintain the original name, and changed others to ‘Other’. I have to loss some accuracy due to the huge database and limited device.

Getting dummy variables

In statistics and econometrics, particularly in regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. [4][5]

They can be thought of as numeric stand-ins for qualitative facts in a regression model, sorting data into mutually exclusive categories (such as smoker and non-smoker). [6]

In this project, I made dummy variables for categories variables to prepare for data analysis, it is very convenient with the Library 'Pandas', just use:

```
df = pd.get_dummies(df)
```

Then dummy variables will be created for every categorical variables in data frame df.

3. Exploratory Data Analysis

(This Part in EAD.ipynb)

Exploratory Data Analysis will be done before prediction.
In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.[7]

During EDA, I exploratory the columns in the dataset and their relationship to understand the whole database and chose the factors for prediction. For example:

Canceled Flights

I calculated the rate for cancellation, then got the result:

(1). Total flights: 5819079,
(2). Cancelled flights 89884,
(3). % of cancelled flights: 1.54 %.

This is a very small probability, then I drop these flights to explore the delay status.

Delayed Flights

I explored the overall situation of delayed flights and get that result:

(1). On time: 72.82271918415235%,
(2). Small delay: 19.481071780088513%,

(3). Large delay: 7.696209035759138%.

I set the flights delayed less than 15 minutes as ‘On time’, less than 45 minutes and more than 15 minutes as ‘Small Delayed’, more than 45 minutes as ‘Large Delay’.

Delay and Month

I draw a figure to explore the relationship between delayed flights and month:

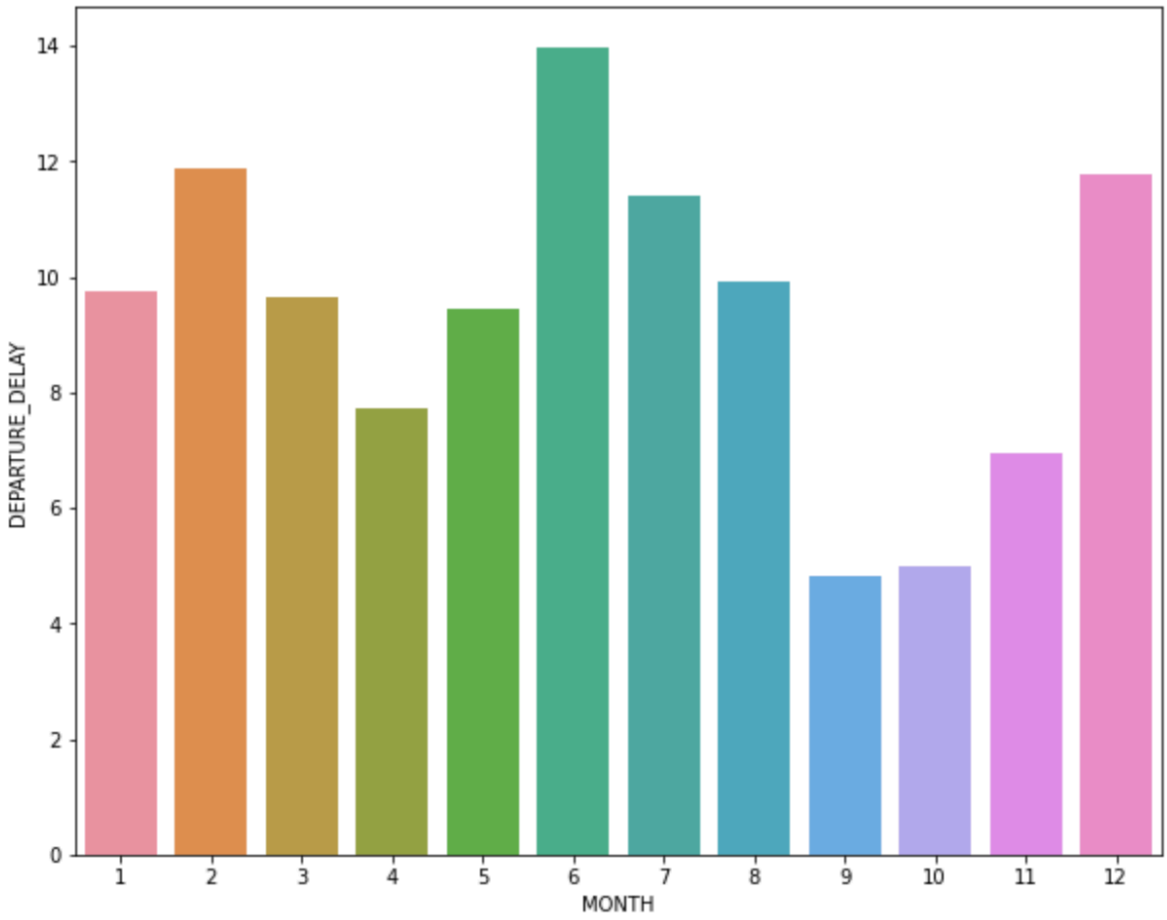

Figure 2 – Delayed flights each month

As you can see in Figure 2, the flights in summer have the highest possibility of ‘on time’, and the flights during spring and autumn always be delayed. I think the

influences of the month should be combined with the departure airport and arrival airport, due to the different weather in different place, so I maintain the MONTH variable.

Delay and Airlines

Many passengers complain they always meet delayed when they choose some specific airlines, so I explored the relationship between delayed flights and airlines, the results shown in Figure 3.

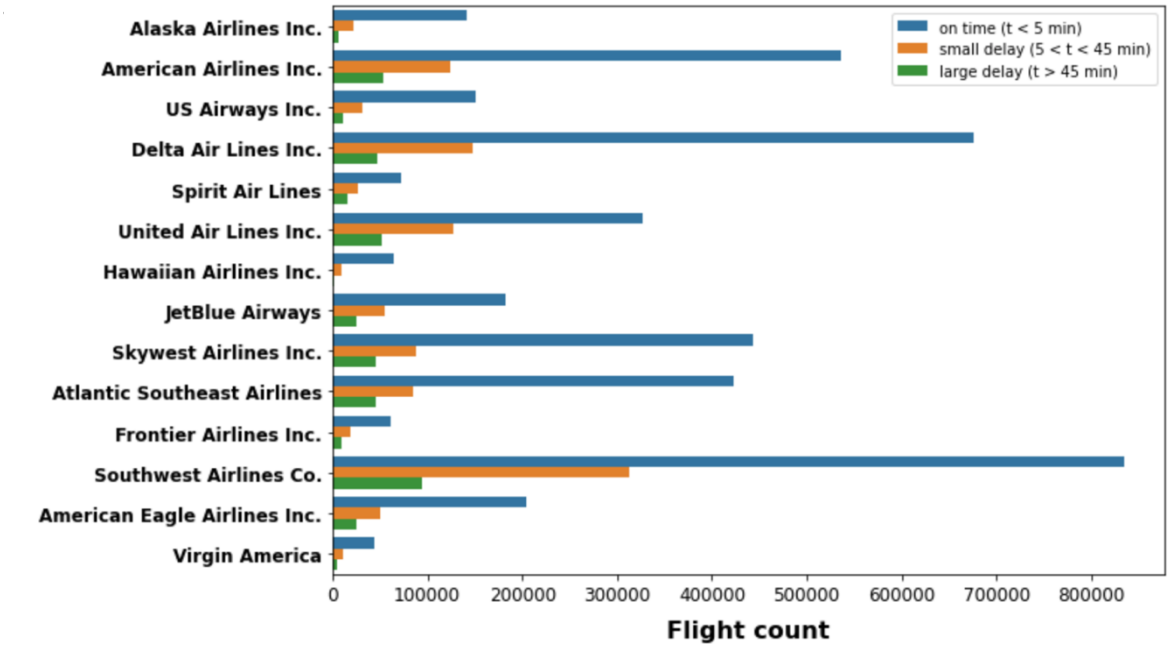

Figure 3 - relationship between delayed flights and airlines

Others

I draw a heatmap of the data frame (Figure 4) to avoid omitting factors which have strong relationship with delay status.

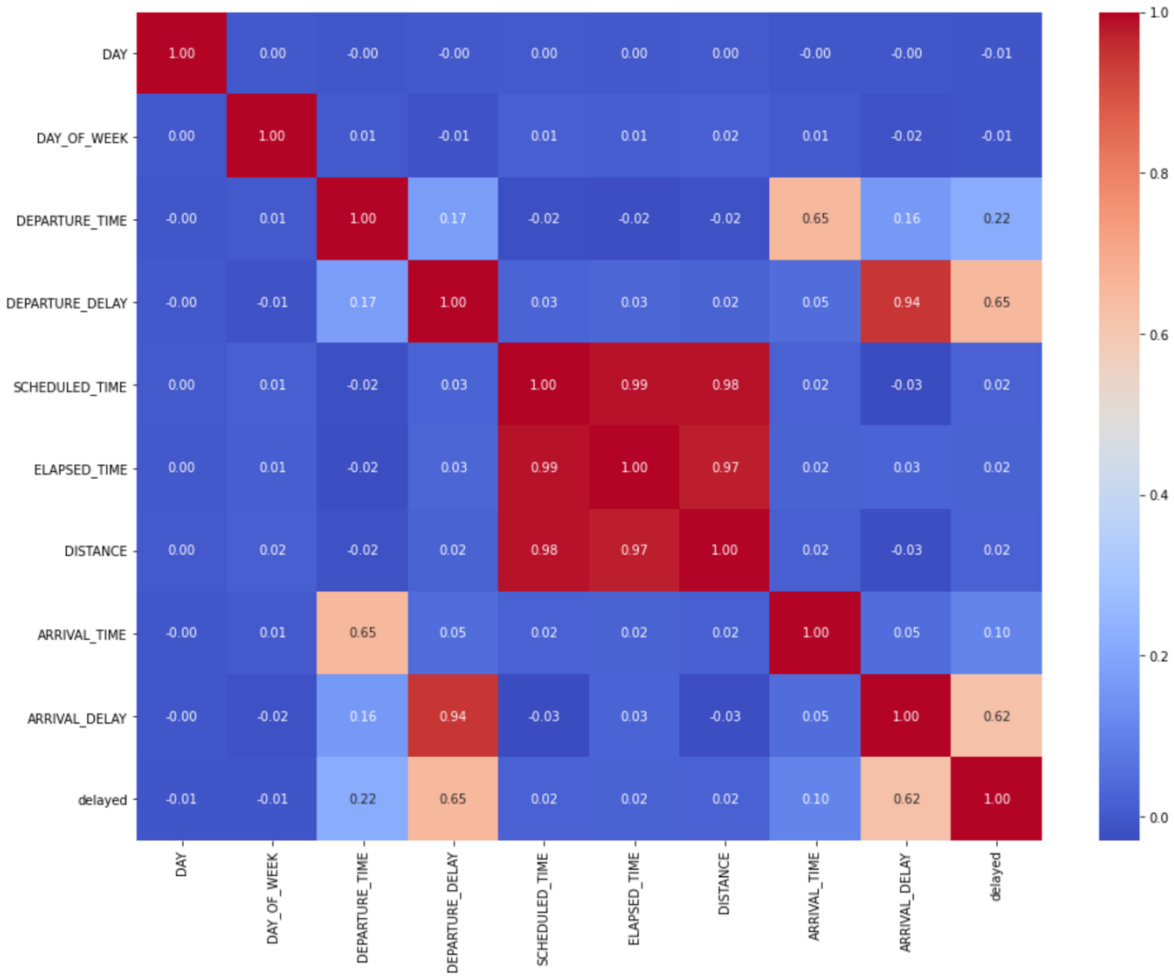

Figure 4 - heatmap for all columns

Make Training Dataset

At the end of Exploratory Data Analysis, I picked up the factors for prediction and made the final training dataset based on the result I got in EDA.

4. Prediction

In this section, I built 4 kinds of classifiers for training data: Logistic Regression, RandomForest, CatBoost and LightGBM, and evaluated their performance according to AUC, which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').[8]

Moreover, for Random Forest, I accelerated it by parallel model, for CatBoost and LightGBM, I added GPU to speedup, all of those accelerations deduced the run time obviously.

For training, I made dummy variables for categories variables, ‘delayed’ column which was transferred to binary variable was set as the dependent variable, other variables were set as independent variables. Then the training dataset was split into train part and test part, on 2:1 ratio.

4.1 Logistic Regression

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.[9]

I built Logistic Regression on H2O frame to make it faster, but the AUC for this model is only 0.658, it is not suitable for this project, so I give up more exploration on this model.

4.2 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[10][11]

I built 4 Random Forest models on my local Jupyter Notebook, the number of trees in the forest is 100 and the maximum depth of the tree is 10.

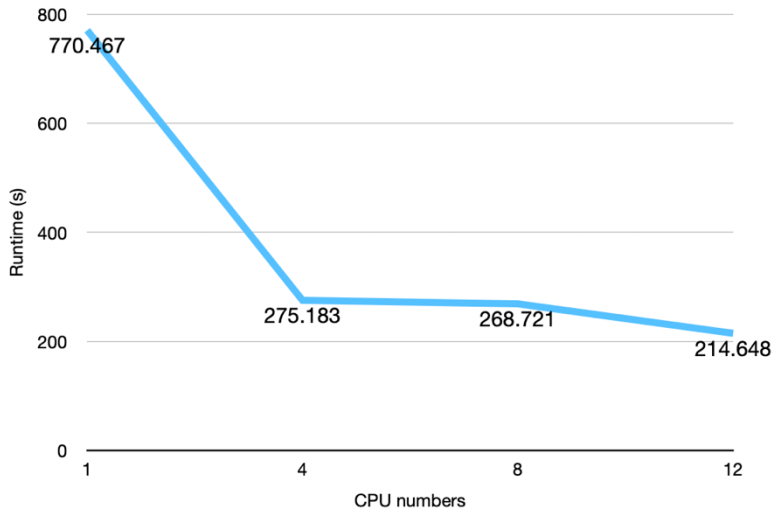

Figure 5 - Runtime with different numbers of CPU by Random Forest

4.3 CatBoost

CatBoost is a machine learning method based on gradient boosting over decision trees. It has many advantages:

(1). Superior quality when compared with other GBDT libraries on many datasets.

(2). Best in class prediction speed.

(3). Support for both numerical and categorical features.

(4). Fast GPU and multi-GPU support for training out of the box.

(5). Visualization tools included.[12]

I built two CatBoost models on Google Colaboratory, the learning rate is 0.1 and the maximum depth of the tree is 5.

The first model ran without GPU cost 9m13s, another one ran with GPU only cost 22s. Same as the result of Random Forest, the AUC of the accelerated model is slightly below the first model.

4.4 LightGBM

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

(1). Faster training speed and higher efficiency.

(2). Lower memory usage.

(3). Better accuracy.

(4). Support of parallel and GPU learning.

(5). Capable of handling large-scale data.[13]

I installed the GPU version of LightGBM on Google Colaboratory and added CMake option '-DUSE_GPU=1' to make it available, I tried to run it without GPU, but failed due to limited environment.

GridSearchCV was used to generates candidates from a grid of parameter values specified with the param_grid parameter and LGBMClassifier was used to build LightGBM classifier. This model got the highest AUC value in this project.

5. Conclusion

Results in above training are shown in the Table below:

	Run Time	AUC
Logistic Regression	5.612s	0.658
Random Forest (1 CPU)	12m24s	0.7026835834131947
Random Forest (12 CPU)	3m30s	0.7014089947897176
CatBoost (no GPU)	9m 13s	0.7669619882652702
CatBoost (with GPU)	22s	0.7660014629988159
LightGBM (with GPU)	21m 45s	0.8585378889100973

Running Time:

(1). When we ran Random Forest, it costed 12m24s with one core and 3m30s with 12 cores, so the training was accelerated in parallel mode.

(2). When we ran CatBoost, it costed only 22s with GPU and 9m13s without GPU, so the training was accelerated with GPU.

(3). I have tried LightGBM without GPU in this project, but it cannot be completed due to the limited resource.

(4). There is no comparison of run time between different methods, because I ran Logistic Regression based on H2O Frame, ran Random Forest on local Jupyter Notebook, ran CatBoost and LightGBM on Google Colaboratory.

AUC:

I use AUC to evaluate my classifiers.

(1). For the prediction in this project, rank classifiers according to their performance: LightGBM > CatBoost > Random Forest > Logistic Regression.

(2). According to the performance of Random Forest and CatBoost models, there is a very minor slight decrease on AUC after acceleration, it needs follow-up experiment to research on it.

(3). I didn't try to accelerate Logistic Regression model in this project, because the performance is not very ideal, so I think it is not suitable for this problem.

(4). These models will be more accurate if we have the data of weather or other events such as air traffic control.

References

[1]. https://www.transportation.gov/individuals/aviation-consumer-protection/flight- delays-cancellations

[2]. https://travel.usnews.com/features/things-to-do-when-your-flight-is-canceled-or-delayed

[3]. https://www.kaggle.com/usdot/flight-delays#airports.csv

[4]. Draper, N. R.; Smith, H. (1998). "'Dummy' Variables". Applied Regression Analysis. Wiley. pp. 299–326. ISBN 0-471-17082-8.

[5]. Interpreting the Coefficients on Dummy Variables" (PDF). Archived from the original (PDF) on August 18, 2003.

[6]. Gujarati, Damodar N. (2003). Basic Econometrics. McGraw Hill. ISBN 0-07-233542-4.

[7]. https://en.wikipedia.org/wiki/Exploratory_data_analysis

[8]. Fawcett, Tom (2006); An introduction to ROC analysis, Pattern Recognition Letters, 27, 861–874.

[9]. https://en.wikipedia.org/wiki/Logistic_regression

[10]. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

[11]. Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844. doi:10.1109/34.709601

[12]. https://github.com/catboost/catboost

[13]. https://lightgbm.readthedocs.io/en/latest/