Flight delays prediction using Machine Learning

Andrey Kulagin
Innopolis University
Innopolis, Russia
a.kulagin@innopolis.ru

ABSTRACT

In this project we will build a framework for gathering and preprocessing historical meteorological data for airports, combine it with information about Airline On-Time Performance from U.S. Department of Transportation and test Random Forest and Gradient Boosting techniques for flight delays prediction.

KEYWORDS

flight delays, delays prediction, meteorology, gradient boosting, random forest

1 INTRODUCTION

Today we cannot imagine our lives without air communication. Planes cut travel time significantly comparing to land transport and allow us to reach previously inaccessible places. More than 100,000 flights served every day in the world to transfer people and goods[10].

But with increase of air traffic there is an increase of flight delays which occur because of various reasons: weather conditions, aircraft technical issues, airline internal problems... To properly reschedule takeoffs and landings airport dispatchers may rely on statistical data and current situation at airport. But it is often too hard for humans to quickly analyze all incoming information to make optimal decisions. Here various Machine Learning techniques can help.

The goal will be to predict delays for flights scheduled within next 3 hours.

2 PROJECT IDEA

This project started with a request from "Aeroflot" to develop a software which helps a dispatcher optimally rearrange flights in Sheremetyevo International airport. "Aeroflot" is the largest airline company in Russia[1]. Sheremetyevo is one of the 3 major airports in Moscow, and the busiest in Russian Federation with 40,093,000 passengers and 308,090 aircraft movements handled in 2017[9]. This is the central "Aeroflot"'s hub for passenger operations and it is very important that real-time disruption management is organized effectively. Every minute of a flight delay costs real money.

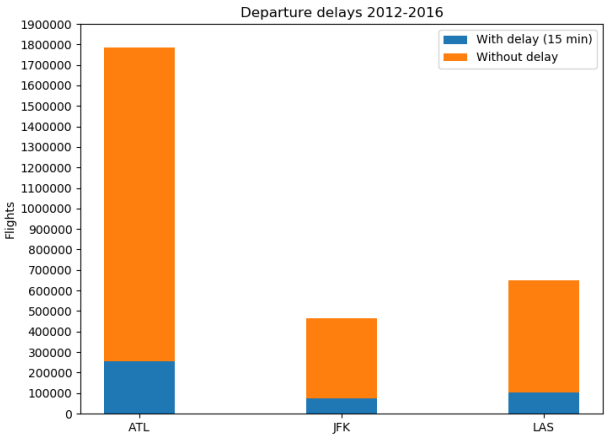
This project is a part of a larger solution which offers a flight dispatcher how to rearrange takeoffs and landings optimally (given that some disruptive events have occurred).

3 PERSONAL MOTIVATION

It will be good experience to work with real data and practice skills related to statistics and machine learning. Also it is interesting to discover more insights for airline industry.

Table 1: Overall departure delay statistics for 2012-2016

Airport	total flights	with delay	delay fraction
ATL	1785878	255439	0.143
JFK	465307	76041	0.163
LAS	648959	102859	0.159



4 FLIGHT DELAYS DATA

Unfortunately due to bureaucracy we were not able to get data about flights from Sheremetyevo. So it was decided to start working with different dataset: "Airline On-Time Performance and Causes of Flight Delays"[2] from American Bureau of Transportation Statistics (BTS). This database contains scheduled and actual departure and arrival times for American domestic flights. We will consider flights from 2012 to 2016 from 3 particular airports:

- John F Kennedy International Airport (JFK) in New York
- Hartsfield Jackson Atlanta International Airport (ATL) in Atlanta
- McCarran International Airport (LAS) in Las Vegas

We will treat flight as **delayed** if its actual departure was 15 minutes later or more. You can find overall statics for departure delays in table 1 and on figure. It is interesting that we have so few flights from JFK airport, after all it is the busiest international air passenger gateway in North America[4]. But probably the reason is that we have only domestic flights in our dataset and the majority of flights to/from JFK are international.

We will use the following features from BTS dataset for delay prediction:

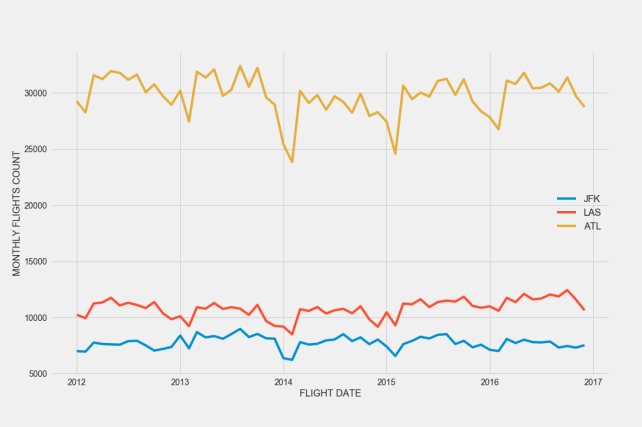


Figure 1: Monthly total flights count

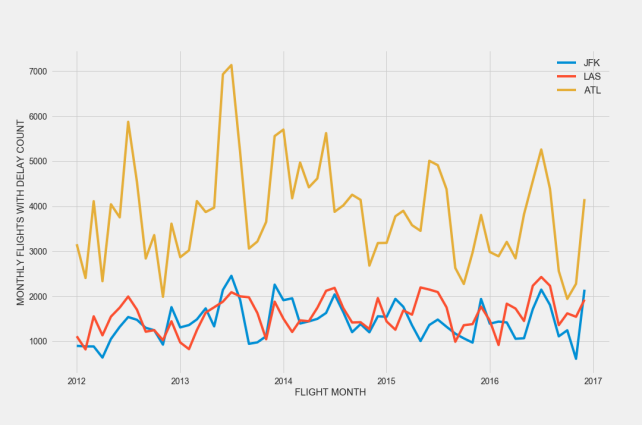


Figure 2: Monthly delayed flights count

- Quarter
- Month
- Day of month
- Day of week
- Time to midnight in hours
- Origin airport
- Airline company
- Actual time of last landing for current aircraft
- Late arrival at Origin airport for current aircraft

For *"Time from last actual landing for current aircraft"* and *"Late arrival"* features we will use information about flights to JFK, ATL and LAS.

Relation between month and flight delay statistics you can find on figures 1 and 2.

5 METEOROLOGICAL DATA

It was much harder to find proper meteorological data. The majority of Internet sources ask big money for detailed historical weather. But we found a resource [7] which allows us to retrieve free METAR reports for airports around the world. METAR[5] is a format for reporting comprehensive weather information designed for airline industry.

Special downloader and HTML parser was written to get all METAR reports. Due to rate limits it took about 15 hours to download data for 3 airports for 5 years (hourly reports).

The next step was parsing METAR reports. They look like

METAR LBBG 041600Z 12012MPS 090V150 1400 R04/P1500N R22/P1500U +SN BKN022 OVC050 M04/M07 Q1020 NOSIG 8849//91=

and readable by technicians, but they are just strings and should be parsed to become informative for ML models.

The most advanced existing implementation of METAR parser is the Python-METAR library[8]. But its functionality was significantly extended to meet our needs: parsing separate weather conditions and exporting to CSV files. The resulting implementation can be found here [6].

We will use the following meteorological features to predict delays:

- Wind speed
- Wind direction
- Gust speed
- Horizontal visibility
- Runway visual range
- Height and coverage of clouds
- Weather conditions: present/absence of 42 different conditions with intensities (light freezing rain, light snow grains, blowing snow, snow and rain, heavy snow, snow, light snow, shallow fog, freezing rain, snow and snow pellets showers, light rain and snow pellets showers, light drizzle, light snow and rain showers, snow showers, thunderstorm, mist, light snow showers, light thunderstorm with rain and hail, light rain and snow, light thunderstorm with rain, ice pellets, freezing rain and snow, light rain, heavy snow showers, haze, light snow and snow pellets showers, patches of fog, light ice pellets, smoke, fog, light snow and rain, light rain showers, light snow pellets showers, rain, light freezing drizzle, snow and snow pellets, light rain and snow showers, light thunderstorm with snow, low drifting snow, ice crystals, freezing fog, drizzle)

6 LITERATURE REVIEW

Though problem of predicting flight delays has been existing for many years, majority of proposed solutions were probabilistic based. Machine Learning models started to be applied quite recently. Probably, it was because of absence of open datasets.

While searching through related researches we spotted 3 points:

(1) Features used for prediction
(2) Used ML models (decision trees, kNN, neural networks...)
(3) Results

The following related researches were found:

(1) *Machine learning approach for prediction of on-time performance of flights*[13] The authors predict occurrence and try to estimate the time of flight delays. Close set of features to our dataset is used, but we have more comprehensive meteorological data and also information about late arrival. For classification (presence/absence of delay) Gradient Boosting,

Table 2: Using SMOTE to improve performance

	accuracy	recall
Random Forest without sampling	0.8	0.6
Random Forest with sampling	0.83	0.75

Table 3: Final results for different models

Model (with sampling)	accuracy	precision	recall
Random Forest (100 trees, max_depth=5)	0.82	0.85	0.73
Random Forest (100 trees, max_depth=10)	0.83	0.89	0.75
Gradient Boosting (100 trees, max_leaves=31)	0.91	0.93	0.86
Gradient Boosting (100 trees, max_leaves=63)	0.92	0.95	0.88

Extra-Trees and AdaBoost classifier models are used with best results for Gradient Boosting. For regression (prediction of delay time) Extra-Trees, Random Forest Regressor, Gradient Boosting, Multilayer Perceptron (MLP) regressors are used. Best results for regression showed Extra-Trees and again Random Forest. Authors pay great attention to feature scaling, hyper-parameter tuning.

(2) *A Deep Learning Approach to Flight Delay Prediction* [12] Here another approach was used - Recurrent Neural Network, in particular Long-Short Term Memory. RNN are very good at dealing with sequential data like speech or text recognition. So it was natural idea to try them for flight delays propagating through time. Authors consider day-level time slots and this approach also showed good performance.

7 PREDICTION OF FLIGHT DELAYS

As was mentioned above our problem is a classification problem for more than 15 minutes delays. The dataset was splitted into train (80%) and test sets (20%). We will test Random Forest (sklearn implementation) and Gradient Boosting (LightGBM implementation) techniques.

Also, we are dealing with imbalanced problem (only 15% of flights are delayed). So as suggested here [13] and here [11] we will use Synthetic Minority Over-sampling TEchnique (SMOTE) to add more minority class data for training, it will improve performance as you can see in the table 2

The final results are presented in the table 3. Gradient Boosting Decision Trees performed better than Random Forest.

8 CONCLUSION

The results are quite good and comparable with other researches. The next possible steps are:

(1) Test Recurrent Neural Networks as suggested in [12]
(2) Applying regression techniques to predict time of delay.

You can find all the source code here [3]

REFERENCES

[1] [n. d.]. Aeroflot. ([n. d.]). https://en.wikipedia.org/wiki/Aeroflot
[2] [n. d.]. Airline On-Time Performance and Causes of Flight Delays. ([n. d.]). https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-delays-on-time-data
[3] [n. d.]. GitHub: Flight delays prediction using Machine Learning. ([n. d.]). https://github.com/and-kul/flight_delays
[4] [n. d.]. John F. Kennedy International Airport. ([n. d.]). https://en.wikipedia.org/wiki/John_F._Kennedy_International_Airport
[5] [n. d.]. METAR. ([n. d.]). https://ru.wikipedia.org/wiki/METAR
[6] [n. d.]. METAR downloader and parser. ([n. d.]). https://github.com/and-kul/flight_delays/tree/master/METAR
[7] [n. d.]. Professional information about meteorological conditions in the world. ([n. d.]). https://www.ogimet.com/metars.phtml.en
[8] [n. d.]. Python-metar. ([n. d.]). https://github.com/tomp/python-metar
[9] [n. d.]. Sheremetyevo International Airport. ([n. d.]). https://en.wikipedia.org/wiki/Sheremetyevo_International_Airport
[10] 2014. 100,000 Flights a Day. (2014). http://garfors.com/2014/06/100000-flights-day.html/
[11] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. 2016. Prediction of weather-induced airline delays based on machine learning algorithms. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*. IEEE, 1–6.
[12] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. 2016. A deep learning approach to flight delay prediction. In *Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th*. IEEE, 1–6.
[13] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, and Vineeth Vijayaraghavan. 2017. A machine learning approach for prediction of on-time performance of flights. In *Digital Avionics Systems Conference (DASC), 2017 IEEE/AIAA 36th*. IEEE, 1–6.