Airline Delay Prediction Competition Report

Oleksii Renov

May 25, 2016

Introduction

In the United States, the Federal Aviation Administration estimates that flight delays cost approx. $20 billion yearly (2010 study). Knowledge of the factors leading to specific flight delays can help Aviation authorities and Airlines in taking necessary actions to ensure smooth operations.

The goal is to build a model to predict arrival delays of flights – Delay/No-Delay.

These investigations can save lots of money for travellers to don't buy tickets on delayed flights.

Variable Importance

```
##                      Feature        Gain       Cover   Frequence
##  1:             CRS_DEP_TIME 0.067538955 0.009750510 0.046050672
##  2:                  TAXI_IN 0.060080493 0.016965822 0.019583752
##  3: mean_arr_del_by_tail_num 0.047777669 0.016505330 0.045115231
##  4:     mean_dest_orig_delay 0.045376736 0.015047969 0.029631595
##  5:                 ARR_HOUR 0.034485232 0.004128841 0.017028743
##  6:                 DEP_HOUR 0.032647393 0.003703050 0.013729104
##  7:                    HDAYS 0.027175630 0.006474805 0.016107264
##  8:                    MONTH 0.024292481 0.007041030 0.011532447
##  9: mean_arr_del_by_orig_num 0.024278613 0.005945723 0.022748427
## 10:           num_arr_origin 0.023020047 0.012352359 0.029920139
## 11:             num_arr_dest 0.022220532 0.014089561 0.032670613
## 12:             CRS_ARR_TIME 0.021473028 0.003520279 0.044575375
## 13:             DAY_OF_MONTH 0.019677907 0.005230142 0.012370155
## 14:   mean_arr_del_by_fl_num 0.017441235 0.007825958 0.032703191
## 15: mean_arr_del_by_dest_num 0.016059724 0.006052748 0.022385420
## 16:         UNIQUE_CARRIERWN 0.015480914 0.003817961 0.004695819
## 17:         CRS_ELAPSED_TIME 0.014411298 0.019240104 0.033266317
## 18:         NUM_FLIGHTS_TAIL 0.014295335 0.011585818 0.018262035
## 19:                 DISTANCE 0.014010104 0.013946163 0.031674671
## 20:              num_dep_all 0.013027535 0.004836114 0.022506422
## 21:              DAY_OF_WEEK 0.012913049 0.005445104 0.006808705
## 22:              dew_point_f 0.011124020 0.005704627 0.017945568
## 23:              num_dep_ori 0.010706764 0.005175936 0.024591385
## 24:        ORIGIN_AIRPORT_ID 0.010286268 0.003407063 0.017926952
## 25:         num_arrival_dest 0.010209894 0.004207939 0.023641982
## 26:          num_arrival_all 0.009924553 0.003387032 0.023502364
## 27:          DEST_AIRPORT_ID 0.009677023 0.001564888 0.021063703
## 28:                 humidity 0.008458275 0.005585257 0.016516810
## 29:                  QUARTER 0.008347896 0.002004959 0.003946536
## 30:         wind_dir_degrees 0.006547475 0.001926476 0.018280651
##                      Feature        Gain       Cover   Frequence
```

Modelling Report

All code is available on github : https://github.com/dataminders/airline-delay-prediction.

The goal was to find models wich performs differently on hold out dataset and to combine them in the best way. Basically I created 1 analogue of fm machine and 4 xgboosts with different subsets of parameters. Surprisingly I got very high weight for fm model.

Another trick which helped me to to bean benchmark score (0.70) is correcting output probabilities by delay rate in exactly this months last year. It looks like the problem of delay flights is very seasonal.

Some quick summary of results:

Best single model result - 0.6957.

Blending used precomputed values from out of sample validation data - 1.5 millions of points(4 months I guess).

Public score - 0.70355

Private score - 0.70257

Feature Engineering

Here all list of features which I created from raw data including weather with some background why exatly this features were created. Added lots of simillar weather data for origin and destination.

To reduce number of features I don't use one hot encoding of Tail num or Flight num features in all models, only some simple basic features like average delay rate on features.

Here the complete table of created features and their description

Feature Name	Description
num arrival all	Number of arrival flights planned in this day in overall airports in time blk
num dep all	Number of departure flights planned in this day in overall airports in time blk
num arrival dest	Number of arrival flights planned in this day in destination ariport in time blk
num arrival org	Number of arrival flights planned in this day in origin ariport in time blk
ARR HOUR	Arrival hour
DEP HOUR	Departure hour
NUM FLIGHTS TAIL	Number of flights in a day by tail num
same state	is local flight
Holiday days	Days before/after nearest public holiday
mean arr del by fl num	Average delay rate by flight number
mean arr del by tail num	Average delay rate by tail num
mean arr del by dest num	Average delay rate by destination
mean arr del by orig num	Average delay rate by origin
mean dest orig delay	Average delay rate by origin:destination
is fl num delay	is delay ever for flight num
is tail num delay	is delay ever for tail num
is dest orig delay	is delay ever for origin:destination

Replacing Missing Values

Missing Values were replaced as zeros (or -999) or most probable category for factor variables. (especially for weather data)

Here is an example:

```
weather[is.na(precipitation_in), precipitation_in := 0]
weather[is.na(events), events := "NoAny"]
weather[is.na(temperature_f), temperature_f := mean(weather[['temperature_f']], na.rm = TRUE)]
weather[is.na(dew_point_f), dew_point_f := -999]
weather[is.na(humidity), humidity := -999]
weather[is.na(sea_level_pressure_in), sea_level_pressure_in := -999]
weather[is.na(visibility_mph), visibility_mph := -999]
weather[is.na(gust_speed_mph), gust_speed_mph := 0]
weather[is.na(wind_dir_degrees), wind_dir_degrees := 0]
weather[is.na(wind_direction), wind_direction := 'Calm']
weather[is.na(wind_speed_mph), wind_speed_mph := 'Calm']
```

Applicability of model for future flights

Due to finding some really interesting results of combining models it can help.

So my second important model in the blend is linear model. It scores only 0.667 on public leaderboard, but contributes so good for final ensemble. Combination of few different xgboosts showed that different subset of features with different parameters play the role.

But even this strong combination of models can't catch some time trends. And hopefully time series analysis can also help, exspecially some SARIMA models for predicting priors.

One thing that dissapointed me a lot is that simple ordered counts of events by flight number, tail number of something else, simply don't work. They showed that there is no any linear trend between delay rate and flight number, it just gives me more confidence that all this events have very periodic nature.

Conclusion

Thank you for this competition. That was great to learn about so important problem.