

IML HACKATHON 2020

Group members:

Nir Vaknin, David Ponarovsky, Amir Levy, Nitsan Wojakovsky

Task chosen : Flight Delay Prediction

Our solution was divided up to a few main segments that each had their own important role as part of our final solution.

Our first task was to preprocess all the data. Naturally dealing with this amount of data was quite challenging and we had to "play" with it until we had a datasheet that was compatible to our learning goals.

The alterations we performed on the data were the following:

- 1) Extracting the FlightDate feature into "day", "month" and "year" features, in order to achieve more information that would be helpful, unlike a date format that would not contribute at all.
- 2) Binning the "CRSArrTime", "CRSDepTime" and "Distance" features:

The granularity of the numbers in these columns contains a lot of unique values — which could have a negative impact on accuracy in a machine-learning model. We resolved this by dividing each number in these columns by 100/10 accordingly and rounded down to the nearest integer. 1030 would become 10, 1925 would become 19, and so on, resulting in a maximum of 24 discrete values in the time columns. Intuitively, it makes sense, because it probably doesn't matter much whether a flight leaves at 10:30 a.m. or 10:40 a.m. It matters a great deal whether it leaves at 10:30 a.m. or 5:30 p.m.

- 3) In addition, the dataset's Origin, Dest and Reporting_Airline columns contain airport codes that represent categorical machine-learning values. These columns were converted into discrete columns containing indicator variables, known as "dummy".
- 4) We dropped out a few features that we thought had no meaningful value to contribute, such as Flight_Number_Reporting_Airline, Tail_Number, DelayFactor, OriginCityName, OriginState, DestCityName, DestState, FlightDate, and CRSElapsedTime.


In addition, we preprocessed slightly differently for the second part of our classification that dealt with possible delay reasons, assuming there was a delay in the flight. In that specific classification we dropped out all the samples of flights with no delay whatsoever.

We also dropped out features with high correlation to others which optimized the prediction even more.

The second main part of our project was to predict the delay of flights, we did this by building for different time thresholds a classifier (by taking a max/min error over various other options) and running over many different combinations of features and on that running an adaboost model.

Since our heap size is limited, we train over a fixed number of subsets of features, and then going over each iteration we keep only those subsets with a pleasing result, and from them we build new subsets.

In the end we hold an object that keeps the data set and 2 ranges and runs a binary search, and then return the prediction for the delay.



The Third part was to predict in case of a delay what would be the delay reason. We chose to deal with this multiple classification problem with a RandomForest model. After doing some research we got to the understanding that this was the right model for the job. Random forests are ensembles of decision trees, they consist of a bunch of independent decision trees, each of which is trained using only a subset of the features in our training set to ensure that they're learning to make their predictions in different ways. Their outputs are then pooled together using simple voting.

We were able to get to an average score of 43% for a correct delay reason, which isn't very high but is much better than a random guess of 25%.