Functional Requirements For Flight Delay Prediction

Data-Driven Probabilistic Flight Delay Predictions:

Mixture Density Networks and Random Forests Regression are two machine learning algorithms that are used in this section to provide probabilistic flight delay estimates.

Data description:

Flight schedules available at Rotterdam The Hague Airport (RTM) between January 1, 2017, and February 29, 2020, are taken into consideration for this analysis. There are 17,365 flights arriving and 17,336 flights departing in total. 42 airports in Europe and North Africa serve as both origin and destination airports for these flights. The map, which displays all airports to or from which aircraft depart or arrive, indicates that the shortest route featured is to London City Airport (LCY), while the longest route, with an average distance of 1300 km, is to Tenerife South Airport (TFS).The average absolute delay for aircraft leaving has a standard deviation of 25.1 minutes, while the average absolute delay for planes arriving has a standard deviation of 26.4 minutes. Here, the time difference between the schedule and the delay is referred to as the delay.

Weather Dataset:

All flights arriving at or departing from RTM between 2017 and 2020 are taken into account, as well as the weather conditions, including temperature, pressure, and wind speed, measured at the origin and destination airports. Every 30 minutes, measurements are accessible.

Feature selection:

The Pearson Correlation Coefficient is used for feature selection. For a specific training set, the correlation between any two features as well as the correlation between the features and the target (the flight delay) are computed. The characteristics are chosen in the following way: if any two features have correlations between them that are greater than 0.7, the feature with the lowest correlation to the target variable is eliminated. A description is given for each of the features that have been chosen for the flight delay prediction. The flight schedule dataset is used to get or calculate the features Airport, Airline, Season, Time of day, Day of week, Day of month, Day of year, Airport latitude and longitude, Distance, Month, Year, and Scheduled flights.The feature set is derived from the type of aircraft designated to carry out a flight. The features Pressure, Visibility, Dew-point, Temperature, and Wind Speed are derived from the meteorological dataset.

Prediction features:

a - This feature is target encoded
b - This feature is trigonometrically encoded
c - This feature is numerically encoded

Departure delay Airport a , Airline a , Season a , Time of day b , Day of week b , Day of month b , Day of year b , Airport latitude c , Airport longitude c , Day of month c , Seats c , Year c , Scheduled flights 2 h c , Scheduled flights day c , Dew point c , Visibility c , Pressure c , Wind speed c .Arrival delay ,Airport a , Airline a , Aircraft type a , Season a , Time of day b , Day of week b , Day of month b , Month b, Airport longitude c , Day of month c , Distance c , Seats c , Year c , Scheduled flights 2h c , Scheduled flights day c , Temperature c , Visibility c , Pressure c , Wind speed c.

Feature Description:

Airport - the airport of destination (departures) or origin (arrivals)
Airline - the airline operating the flight
Aircraft - type the aircraft type used for the flight
Season - the flight season (summer or winter schedule)
Time of day - scheduled time of day of the flight
Day of week - scheduled day of the week of the flight
Day of month - scheduled day of the month of the flight
Day of year - scheduled day of the year of the flight
Month - scheduled month number of the flight
Airport - latitude and longitude the latitude and longitude of the destination/origin airport
Distance - the distance between the origin and destination
Seats - the seat capacity of the used aircraft
Year - the year in which the flight was operated
Temperature - the air temperature at the destination/origin airport
Dew point - the dew point temperature at the destination/origin airport
Visibility - the prevailing visibility at the destination/origin airport
Pressure - pressure altimeter at the destination/origin airport
Wind speed - wind speed at the destination/origin airport
Scheduled flights day - the number of flights scheduled to depart/arrive during the day of flight
Scheduled flights 2h - the number of flights scheduled to depart/arrive during the period between one hour before and one hour after the scheduled time of the flight.

The characteristics can be category, temporal, or numerical. On the basis of a binary delay threshold of 15 minutes, the categorical features are target encoded. The delay rate of the category to which the sample belongs serves as the encoded value of the sample feature. For instance, all Tuesday flights would be encoded with the value 0.4 for the feature Day of the week if 8 out of 20 samples flying on Tuesdays had delays of greater than 15 minutes. To maintain the periodicity, the temporal features are encoded using trigonometric functions. Each time feature yields the sine and cosine of two characteristics. For instance, for a given month m, sin(2m12) and cos(2m12) are used to determine the characteristics Month sine and cosine.

The encoded value for the remaining features is identical to the value of the original feature because they are numerically encoded. It should be noted that the time aspects are encoded numerically and trigonometrically. For instance, the characteristics Day of the week sine, Day of the week cosine, and Day of the week are produced by the data field Day of the week. Every chosen feature's encoding technique is indicated. The trigonometrically encoded time features are chosen more frequently than the non-encoded time features after all feature values have been scaled to the range [0, 1] to prevent unwanted feature domination in neural network classifiers. Most features are chosen for at least one of the departure/arrival pair.

Machine-Learning Algorithms to Estimate the Probability Distribution of Flight Delays:

The distribution of flight delays is to be estimated by Mixture Density Networks (MDN) and Random Forests regression (RFR). Neural networks and decision trees, respectively, are two different kinds of machine learning algorithms that these methods fall under. Mixture Density

Networks (MDNs). A Gaussian mixture model and a neural network are combined to form a mixture density network. An MDN produces the weight, mean, and standard deviation for each Gaussian in the mixture given the feature values xi of the flight i.

The probability density function p(yi|xi) of the target variable yi, the flight delay, is calculated using these parameters. The MDN is generally a great tool for estimating multimodal probability distributions. As a result, it may forecast a distribution that has peaks at, say, two different anticipated delay values.

The flight delay probability distribution is constructed as the weighted sum of Gaussian distributions as follows:

p(yi|xi) = m∑j=1 αj(xi)ϕj(yi|xi), (1) ϕj(yi|xi) = 1 q 2πσj(xi)2 exp –(yi – μj(i))2σj(x)2 ! (2) where p(yi|xi) is the probability distribution of delay value yi given feature values xi from flight sample i, while αj(xi), μj(xi) and σj(xi) are the weight, mean, and standard deviation of the jth Gaussian component, 1 components considered for the mixture. ≤ j ≤ m with m the total number of Gaussian while the parameters αj, μj , and σj are the output of the MDN. Thus, there are 3m outputs of the MDN. The weights use a softmax activation function, and the standard deviations use an exponential activation function, while the means are unrestricted. The neural network is trained using back-propagation, i.e, the network parameters, the weights and biases of each node are updated using an error function E, which is the negative logarithm of the likelihood that the model derived from the output of the current network gives rise to the training data . This likelihood is the product of the likelihood of every data point, given the current network parameters.

E = Nf ∑i=1– ln m∑=1 αj(xi)ϕj(yi|xi)! = Nf i=1 – ln p(yi|xi), (3) where Nf is the total number of samples in the training set. For every data point fed to the neural network, the derivatives of the error with respect to all network parameters are used to update the weights and biases of the network. Following training, the MDN is applied to a test set and multimodal probability distributions for the delay of each flight in the test set are estimated. The MDN method illustrated. Schematic representation of a Mixture Density Network: parameters for a multimodalGaussian distribution are obtained using a Neural Network. Random Forests Regression and Kernel Density Estimation Random Forests regression (RFR) is a class of decision tree-based machine learning algorithms . The regular RFR algorithm is an ensemble method that combines the results of a number of decision trees. When building each tree, a random subset of the feature values of each training data point is used to make branches. The algorithm outputs a point estimate for the target variable (flight delay) of every test sample by averaging the output values of all considered decision trees. However, for our analysis, we are interested in estimating the probability distribution for the delay of the given flight, rather than a point estimate. In order to obtain the flight delay distribution of a flight in the test phase, the output values of the decision trees are not averaged, but collected, and a kernel density estimation (KDE) is performed . A KDE results in a normalised probability density function. Two settings of the KDE are the kernel type and the bandwidth. In our analysis, a bandwidth of 1.5 is used to render the estimated distribution smooth. Gaussian kernels have been selected for their generality. Random Forests regression is a well -established technique that has been applied in many research areas. However, there are very few examples of studies utilising the algorithm to obtain probability distributions. For startle use quantile values, obtained from Quantile Random Forests, to construct a right-continuous cumulative distribution function of aircraft's time-to-fly from the turn onto the final approach course to the runway threshold. Schlosseretal and Rahmanetal use Random Forests algorithms to obtain probability distributions for precipitation forecasts and drug sensitivity, respectively.

Both studies make use of feature probability distributions estimated via maximum likelihood to make splitting decisions when constructing the decision trees. In contrast, in this study, the feature values and splitting decisions are kept deterministic throughout the Random Forests algorithm. In this way, the probability density function is estimated from deterministic feature values without the need for stochastic variables. Furthermore, the working of the original Random Forests regression algorithm need not be changed

Hyperparameter Tuning:

The hyper-parameters of the MDN and the RFR prediction algorithms have been optimised using a grid search. The hyper-parameters leading to the lowest mean CRPS scores have been selected shows the selected hyper-parameters and their search range. For MDN, a network with three hidden layers of 50 nodes is selected. The output layer of the network consists of 24 nodes, with which an 8-modal Gaussian distribution function is constructed. For RFR, 200 decision trees with a maximum depth of 10 layers are constructed. For every branch split, three out of four features are considered of at least seven training samples.

Mixture Density Network:

Number of modes m = 8 = [3, 5, 8, 10, 15]
Number of hidden layers = 3 = [1, 2, 3]
Number of nodes per hidden layer = 50 = [25, 50, 75, 100]
Number of epochs = 1000 = [500, 750, 1000, 1250, 1500]

Random Forest Regression:

Number of estimators = 200 = [100, 150, 200, 300]
Split criterion = Mean-squared error = [MSE, MAE]
Maximum tree depth = 20 = [4, 6, 8, 10, 12, 15, 20, 30]
Minimum samples per leaf node = 7 = [0, 3, 5, 7, 9]
Fraction of features considered for split = 0.75 = [0.25, 0.50, 0.75, 1.00]
KDE Bandwidth h = 1.5 = [0.5, 1, 1.5, 2]