

Model Card Version: 0.1_2022

License: Apache 2.0

Flight Delay Prediction

Model: [Microsoft R Client & ML Server](#)

Documentation: [GitHub](#)

Model Card Authors: Siddharth Mehrotra, s.mehrotra@tudelft.nl

The model analyzed in this card predicts whether the arrival of a scheduled passenger flight is delayed by more than 15 minutes. The model's goal is exclusively to predict whether the flight is delayed or on-time. It doesn't attempt to discover identities of the airlines.

In this card, we can learn more about how well the model performs on historical weather data combined with actual on-time performance data, and what kinds of situations one can expect the model to perform well and poorly on.

Model Snapshot

<https://github.com/PAIR-code/datacardsplaybook/find/main> Model Overview

MODEL ARCHITECTURE	INPUT(S)	OUTPUT(S)
<i>Describe the architecture of the model here.</i>	<i>Provide a description (with necessary specifications) of the input data provided to the model for outputs.</i>	<i>Provide a description (with necessary specifications) of the output data from the model for given inputs.</i>
Flight delay prediction model is pre-trained on historical on-time performance and weather datasets with logistic regression and decision tree learning algorithm.	<"Month", "Day_of_month", "Origin_Airport_ID", "Departure_time", "Destination_Airport_ID", "Visibility", "DryBulbCelsius", "DewPointCelsius", "RelativeHumidity", "WindSpeed", "Altimeter"> E.g., <14,02,AMS,0900,FRA,10,81,27.2,76,24.4>	<'0' for on-time flights and '1' for flights delayed longer than 15 minutes.> E.g., 1: "This flight is expected to be delayed longer than 15 minutes."

Usage		
APPLICATION	BENEFITS	KNOWN CAVEATS
<p><i>Where has this model been used, or where is it currently used? Include links for readers to learn more.</i></p> <p>This model can be used by Airlines to predict flight delays which plays an important role in financial losses. Furthermore, this model results can be applied to increase customer satisfaction and incomes of airline agencies. Another important usage of the model can be by flight ticketing websites such as Skyscanner and Google Flights. These websites can additionally include possible flight delay information to the search results.</p>	<p><i>Why might users choose to use this model, relative to others? Evidence your response with metrics or performance results</i></p> <p>This model displayed an AUC scores of 0.71 which is considered as a fair score by Thomas G. Tape (2019).</p>	<p><i>Are there any known and preventable failures about this model?</i></p> <p>This model is trained on historical dataset of flight performance and weather for the year 2013. However:</p> <ul style="list-style-type: none"> • Classification quality doesn't include any changes in weather conditions due to global warming. • The destination delay is highly dependent to arrival fights and the effective factors include; day, time, and airport capacity. • This model shows inefficiency in U.S.A but it is suitable for Europe. Where, only 1–4% of the Europe fights delayed due to weather condition this value for U.S.A is between 70 and 75% [Liou JS. Delay prediction models for departure fights. 2006.] • This model round down scheduled departure time to full hour.
MODEL CONTACT	MODEL AUTHOR(S)	CITATION
<p><i>How can model owners be contacted for questions about the model?</i></p> <p>D.Phasen, Microsoft Machine Learning Server (9.4), Quickstart: Run R code in R Client and Machine Learning Server. GitHub: https://github.com/microsoft/RTVS-docs, Web: https://docs.microsoft.com/en-us/previous-versions/machine-learning-server/r/quickstart-run-r-code</p>	<p><i>Write the names of all authors associated with the model. Provide the affiliation and year if different from publishing institutions or multiple affiliations, using the format Name, Title, Affiliation, YYYY:</i></p> <p>Microsoft R Tools for Visual Studio Team</p>	<p><i>If available, provide a citation to your model; else indicate unavailable.</i></p> <p>Unavailable.</p>

System Type		
SYSTEM DESCRIPTION	UPSTREAM DEPENDENCIES	DOWNSTREAM DEPENDENCIES
<i>Is this a standalone model, or intended to be used as part of a system with other models? Include links where necessary.</i>	<i>If the model requires specific inputs, where should they come from? Are there any specific preprocessing steps that should be applied? Include links where necessary.</i>	
Standalone use: Flight prediction model can be used as a web-service standalone model for predicting delay of flights. See detailed implementation for standalone use.	The model requires input of the Month as a numerical variable starting with a 0 for single digit Month. E.g., 04 for April and 11 for November.	The model requires mrsdeploy package to be deployed on the remote server if it is used as a web service or accessed through RESTful APIs .
HARDWARE & SOFTWARE FOR TRAINING		HARDWARE & SOFTWARE FOR DEPLOYMENT
<i>Describe the hardware and software used for training the model.</i>		<i>Describe the hardware and software used for deploying the model.</i>
<ul style="list-style-type: none"> • An installed instance of Microsoft R Client or Machine Learning Server. • R integrated development environment (IDE) 		<ul style="list-style-type: none"> • Microsoft R Client 3.x, R Server 9.x, Machine Learning Server 9.x.

Compute Requirements					
COMPUTE REQUIREMENTS FOR FINE-TUNING*			COMPUTE REQUIREMENTS FOR INFERENCE*		
<i>Describe the following compute requirements. Indicate unavailable if necessary. Do not delete any choices.</i>			<i>Describe the following compute requirements. Indicate unavailable if necessary. Do not delete any choices.</i>		
Number of Chips	unavailable		Number of Chips	unavailable	
Training Time (days)	unavailable		Training Time (days)	unavailable	
Total Computation (floating pt operations)	unavailable		Total Computation (floating pt operations)	unavailable	
Measured Performance (TFLOPS/s)	unavailable		Measured Performance (TFLOPS/s)	unavailable	
Energy Consumption (MWh)	unavailable		Energy Consumption (MWh)	unavailable	
*Modeled after Patterson, David, et al. " Carbon emissions and large neural network training ." arXiv preprint arXiv:2104.10350 (2021).					

Model Characteristics					
MODEL INITIALIZATION		MODEL STATUS		MODEL STATS	
<i>Describe how the model has been initialized. Include information about if the model trained from random initialization, or fine-tuned from a pre-trained model?</i>		<i>Is the model static, or retraining on online data? If this model is trained and retrained, please include the update cadence, and the release date for the latest version.</i>		<i>What is the size of the model? Include attributes like number of weights and layers.</i>	
This model imports the data sets from GitHub and creates a temporary directory to store the intermediate XDF files. Next, it joins flight records and weather data using the origin and destination of the flight `OriginAirportID` and `DestAirportID`. After pre-processing, it randomly splits data (80% for training, 20% for testing).		Static model trained on an offline dataset.		This is a relatively small model, designed for on-device use.	
Training Epochs	Not applicable	Dataset Name	Weather_Sample & Flight_Delays_Sample	Size	7.42 MB

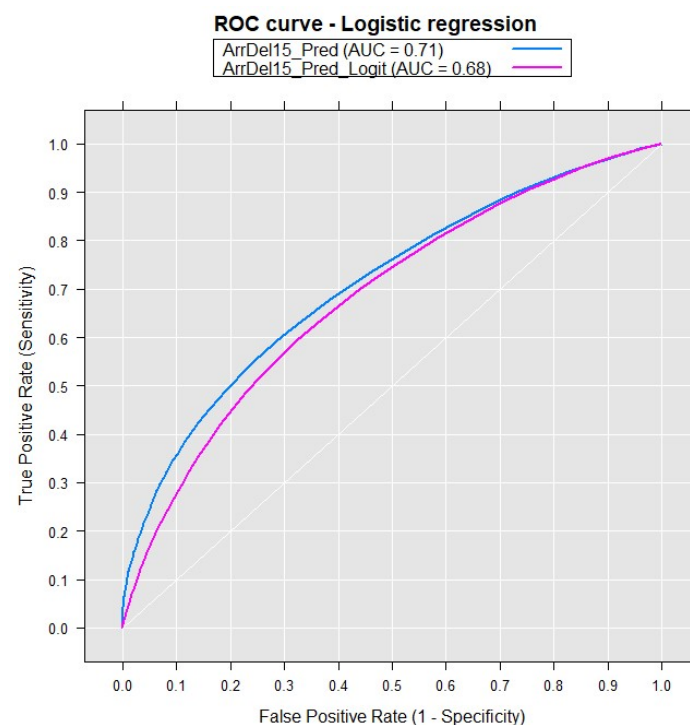
Base Learning Rate	unavailable	Version	0.1	Weights	Not applicable
Method	Decision Trees	Release Date	March 10, 2016	Layers	Not applicable
Loss	unavailable	Update Cadence	Never	Latency	Total Chunk Time: 0.270 seconds
PRUNING		QUANTIZATION		DIFFERENTIAL PRIVACY	
Is your model pruned? If so, what is the level of sparsity of the deployed model?		Is your model quantized? If so, what is the bit representation of the deployed model?		If any, describe the techniques implemented to preserve privacy?	
Yes. If the prediction accuracy is not affected then the change is kept following the pruning.		No		Certain columns are removed that are possible target leakers. E.g., 'Cancelled_flight', 'Year', 'Airline_code'. These columns are directly removed before deploying the model.	
Methods	Reduced error pruning	Methods	Not applicable		
Structuring	Structural pruning	Pre-quantized Representation	Not applicable		
Sparsity Level	unavailable	End Bit Representation	Not applicable		
Number of Params at Sparsity	unavailable	Hardware	Not applicable		
Accuracy at Final Sparsity after Training	unavailable				
Perplexity at Final Sparsity after Training	unavailable				



Data Overview					
TRAINING DATASET SNAPSHOT		DATASET MAINTENANCE & VERSIONS		INSTRUMENTATION	
Describe the dataset used to train the model. If a requested detail is inapplicable, following guidance on N/A. Include links to additional table(s) with more detailed breakdowns in the caption.		Is the training data static, or updated/expanded? If so, what is the frequency with which this data is updated?		What instruments were used to collect or process the data? Describe any notable instrumentation requirements in the collection or preprocessing of data by customizing the table.	
The datasets used to train the model is publicly available as a .csv file on Microsoft R Tools for Visual Studio GitHub page. The two datasets namely Weather_Sample & Flight_Delays_Sample includes weather history of the year 2013 for the location GMT -4 and flight arrival departure history in 2013 for the location GMT -4, respectively.		The training data is static.		Unavailable	
Dataset Size	Weather_Sample: 5.45 MB Flight_Delays_Sample: 36.1 MB	Current Version	0.1	Instrumentation Criteria	
Number of Instances	113332	Update Cadence for Online Data	Unavailable	Focal spot size	Unavailable
Number of Fields	16	Sampling methods	Unavailable	Cooling method	Unavailable
Labeled Classes	2 [on time/delayed]	Validation methods	Unavailable	Avg Adult Effective Dose (mSv)	Unavailable
Number of Labels	745919	Processing methods	Unavailable	Operational voltage range	Unavailable
Average labels per instance	2	Annotation methods	Unavailable		
Missing Labels	No missing labels.	Additional Notes:			
Additional Notes:					

DATA PRE-PROCESSING	DEMOGRAPHIC GROUPS	EVALUATION DATA	
<i>Describe any augmentation methods used during pre-processing to attain the requisite format. Are there any criteria that data points must satisfy to be included in the training set?</i>	<i>Does the data contain any labeled** groups, or attributes that suggest demographic group membership? Describe any demographic groups considered when assessing distributions in the data.</i>	<i>Describe any notable factors about your final test set, including your train/test/dev split, any notable differences between the collection protocols for training & test data.</i>	
Some column names in the weather data were renamed and merged with the flight data based on unique key. OriginAirportID and DestAirportID were used as categorical variables for decision trees.	No	Random split data was performed (80% for training, 20% for testing). The training and test data used to report model performance do not contain any data augmentation.	
		Use /Application	
		Training/Eval	610378/152675
		Use /Application	
		Training/Eval	[provide details]
<i>**If there are groups that may be present, but are not labeled in the training data, please note this in the Ethical Considerations section below.</i>			

Evaluation Results	
Aggregate Evaluation Results	
<i>Document your aggregate or overall model performance evaluation.</i>	
EVALUATION PROCESS	EVALUATION RESULTS
<i>Describe any notable factors in your process for evaluating your model's overall performance.</i>	<i>Summarize and link to evaluation results for this analysis.</i>
<p>Metrics: Model performance is measured using Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC).</p> <p>Evaluation Set: A random split of the data was performed for the evaluation set (20%).</p> <p>Process: A decision tree model was built on the training data set. Next, the best value of the complexity parameter was calculated for pruning the tree. Finally, prediction was performed on the test dataset.</p>	<p>Model results compares models trained with logistic regression and decision tree. Decision tree learning approach provides an AUC of 0.71 compared to 0.68 with the logistic regression method.</p>



Document your disaggregated (e.g. fairness) evaluation. Duplicate this section (subgroup, evaluation process and data, evaluation results) for each subgroup evaluated.


SUBGROUP EVALUATED	EVALUATION PROCESS & DATA	EVALUATION RESULTS
Which subgroup was evaluated?	Describe any notable factors in your process for disaggregated or sliced evaluation of model performance. Please include any assumptions made when disaggregating the data.	Are there any known and preventable failures about this model?
There were no subgroup for which evaluation was performed.	Not applicable.	Not applicable.

Fairness Evaluation Results		
FAIRNESS CRITERIA	FAIRNESS METRICS & BASELINE	FAIRNESS RESULTS
<i>How did you define fairness? Describe the target fairness criteria you hoped to satisfy or optimize for before launch.</i>	<i>Describe the metrics and the baseline for fairness against which you present your fairness results and how they are calculated.</i>	<i>Describe the results of your fairness analysis. Include any specific callouts or points that you would want to highlight for readers.</i>
Flight delay/on-time predictions: Fairness criteria are defined as predictions being equally accurate for data from any other geographical region. [There is no clear fairness criteria available from the source of the model.]	The model needs to compare sub-group performance (runway length, airline operators, taxi time) to overall model performance with data augmentation techniques.	Overall sensitivity and specificity can be seen from Evaluation Results section.

Model Usage & Limitations		
SENSITIVE USE	LIMITATIONS	ETHICAL CONSIDERATIONS & RISKS
<i>Are there any use cases where deployment of this model would be considered sensitive?</i>	<i>What factors might limit the performance of the model? What conditions must be satisfied to use the model?</i>	<i>What ethical factors did the model developers consider? Were any risks identified? What mitigations or remedies were undertaken? Where possible, link to additional documents.</i>
<p>Application: This model is not a substitute for Air Traffic Controller (ATC). Using this model's outcomes to confirm with the ATC results can be considered sensitive given its limited parameters.</p> <p>Pre-requisite training: This model requires Airline companies to provide on-boarding for their staff to use it wisely and in accordance with the air traffic safety norms.</p>	<p>Input conditions: This model may not perform well not for the connections between cities where there are less flights on daily basis. In addition, delays in loading luggage and lack of staff can further cause delay in the boarding, these factors are not taking into account when deploying this model.</p> <p>Output Caveats: This model doesn't provide its confidence score and can only predict the delay greater than 15 minutes between the cities which are already present in the dataset.</p>	<p>Research & Development: Considering that almost all datasets are focused on delays in the USA, we strongly suggest to investigate flight arrival delay in other countries to find out how well the models and results generalize.</p> <p>Deployment: Automation bias and human-error in reporting data are additional factors to consider when deploying this model in real setting.</p>


Terms of Art

Concepts and Definitions referenced in this Model Card

 Use this space to include the expansions and definitions of any acronyms, concepts, or terms of art used across the Model Card. Use standard definitions where possible (e.g. [MLCC Glossary](#)). Include the source of the definition where indicated. If you are using an interpretation, adaptation, or modification of the standard definition for the purposes of your Model Card or model, include your interpretation as well.

Runway	Bias	Accuracy
<p>Definition: a paved strip of ground on a landing field for the landing and takeoff of aircraft.</p> <p>Source: Wikipedia</p> <p>Interpretation: -</p>	<p>Definition: Bias is the conflict in trying to simultaneously minimize two sources of error that prevent supervised learning algorithms from generalizing beyond their training.</p> <p>Source: Wikipedia</p> <p>Interpretation: It is a phenomenon that skews the result of an algorithm in favor or against an idea.</p>	<p>Definition: Accuracy is the most intuitive performance measure and is simply a ratio of the correctly predicted classifications (both True Positives+True Negatives) to the total Test Dataset.</p> <p>Source: Wikipedia</p> <p>Interpretation: It is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.</p>
Specificity	Sensitivity	ROC-AUC
<p>Definition: Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).</p> <p>Source: Wikipedia</p>	<p>Definition: Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category.</p> <p>Source: Wikipedia</p> <p>Interpretation: It is the metric that evaluates a model's ability to predict true negatives of each available category.</p>	<p>Definition: ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.</p> <p>Source: Wikipedia</p> <p>Interpretation: Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1</p>

Interpretation: Specificity is defined as the proportion of actual negatives, which got predicted as the negative		

Reflections on the Model	
 Use this space to include any additional information about the model that has not been captured by the Model Card.	
Usage of the model for stakeholders other than Airline companies	As a next step it will be interesting to see how model information is interpreted and utilized by different stakeholders such as catering agencies and advertisement industry.
Passenger Dissatisfaction w.r.t correct prediction	A possible human-centered evaluation of this model will be to display the prediction values to the passengers and understand their satisfaction & possibility to take the same airline again.
xAI & Trust	This model makes no attempt to provide any explanation to its end user as in how prediction took place. Additionally, to built appropriate trust in the model, the model can a 'goodness' indicator about the flight Based on the past history, a single line explanation of the possible delay reason, predicting whether flight is likely to be delayed or not, and an mouse hover icon that can open the complete explanation which is used to compute the prediction.