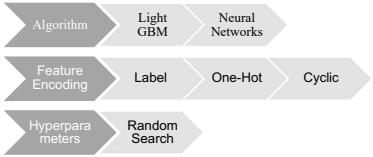
An Ablation Case Study of Flight Delay Predictions: What Modelling Decisions Most Impact Prediction Errors

Dimka Dellenbag (*11403373*), Michelle Dijkstra (*11270926*), Jonathan A. Harris MSc (*13711180*), Sang Pham (*13064878*)
Applied Machine Learning, Team 6

UNIVERSITY OF AMSTERDAM

Purpose

An ablation study was conducted to investigate the effect of decisions related to algorithm selection, encoding techniques, and hyperparameter tuning on flight delay predictions.

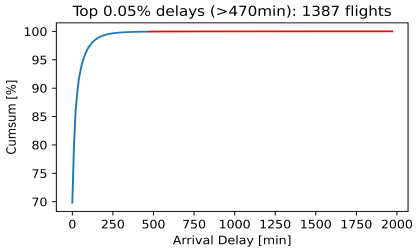


Datasets

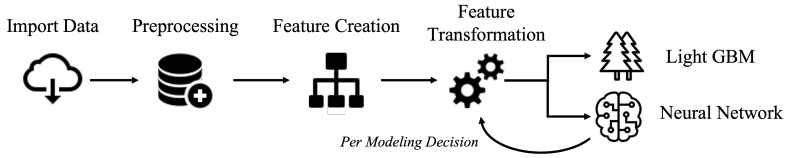
Train: 2.8M flights (Jan 1st, 2015 – June 30th, 2015)
Test: 515K flights (July 1st, 2015 – July 31st, 2015)

Time features:
Day, Month
Day_of_week
Scheduled departure
Scheduled arrival

Categorical Features:
airline code,
origin_airport_code
destination_airport_code

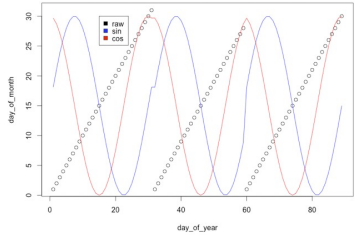


Process



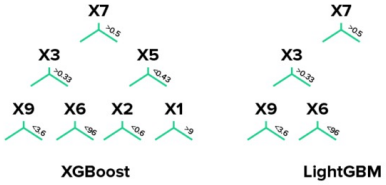
Feature Creation:
is_summer (May-Aug)
is_winter (Jan-Feb)
Time-Zone Difference
Departure Delay
Minutes Between Flights
Scheduled Departures During Same Hour
Scheduled Arrivals During Same Hour

Feature Transformation:
Label-Encoding – categorical, time features
One-Hot-Encoding – categorical, time features
Cyclic Encoding (ex. below)[1] – time features

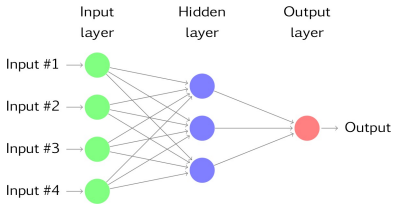


Models

LightGBM grows vertically (faster!)[2]



Neural Network for Linear Regression[3]



Results

LightGBM Models	MSE		
	Train	Test	Kaggle
Cat. (OHE)	1388.613	1413.354	1459.168
Cat. (LE)	1394.447	1418.027	1440.671
Cat. (LE) + Time (LE)	1392.820	1416.634	1449.406
Cat. (LE) + Time (CE)	1404.323	1427.223	1439.164
Cat. (LE) + Time (OHE)*	1409.049	1432.834	1434.121
Encoding* + Feat. Creation	110.557	116.837	90.482
Encoding* + Feat. Creation + Tuning	40.681	77.803	84.453
Encoding* + Feat. Creation + NN	141.463	109.448	NaN

OHE = One-hot encoding; LE = Label encoding;
CE = Cyclic encoding; NN = Neural network model
** Selected encoding transformation workup protocol*

Final Kaggle Submission: 84.453 (8th Place)

Discussion

- Model Mean Squared Error (MSE) results from the aggregated effect of many non-trivial decisions, not just algorithm selection.
- Underlying algorithmic math may explain differences in MSE results between transformation options.
- Creatively extracting additional features may dramatically improve MSE and is worth the time investment.
- Experimental Neural Network produced Train and Test MSE values, despite using a truncated subset of the data (300.000 samples).

References:
[1] Bescond, Pierre-Louis. Cyclic Feature Encoding, it's about time!. 2020. **[2]** Nahon, Aviv. XGBoost, LightGBM or CatBoost – Which algorithm should I use?. 2019 **[3]**Hyndman, RJ, Athanasopoulos G. Forecasting: Principles & Practice. 2018