Real-time Decision Making for Train Carriage Load Prediction via Multi-Stream Learning

Hang Yu, Jie Lu, Anjin Liu, Bin Wang, Ruimin Li and Guangquan Zhang

University of Technology Sydney, Faculty of Engineering and Information Technology (FEIT),

Australia Artificial Intelligence Institute (AAII)

Abstract

In this study, we present a machine learning application that forecasts, in real time, the passenger loads of each carriage in a train when it departs the platform. Developed in collaboration with Sydney Trains using publicly-available data from Transport for NSW's Open Data Hub, the framework advances public transport and carriage load modeling in several key ways: 1) Pre-processing the data with fuzzy metrics helps to reduce the impact of noisy data; 2) Predictions are made with the LightGBM model but with an incremental learning scheme that allows for real-time forecasting; 3) Moreover, this scheme, called multi-stream learning, pioneers a new strategy of merging data streams with similar concept drift patterns to increase the amount of training data while reducing generalization errors. Experiments conducted on a real-world dataset over the period Nov to Dec 2019 comprehensively demonstrate our solutions. We hope researchers and industry analysts facing similar problems will benefit from our findings.

Contributions

1. A discussion on the factors that influence model selection and feature engineering, which provides valuable insights on predicting train carriages loads;
2. An incremental LightGBM model designed to predict train carriage loads based on streaming data, taking noise, drift, and disruption into account;
3. A selection of fuzzy learning methods designed to improve prediction accuracy with noisy samples; and
4. A discussion of our findings from this exploration of multi-stream learning strategies.

Data Examples

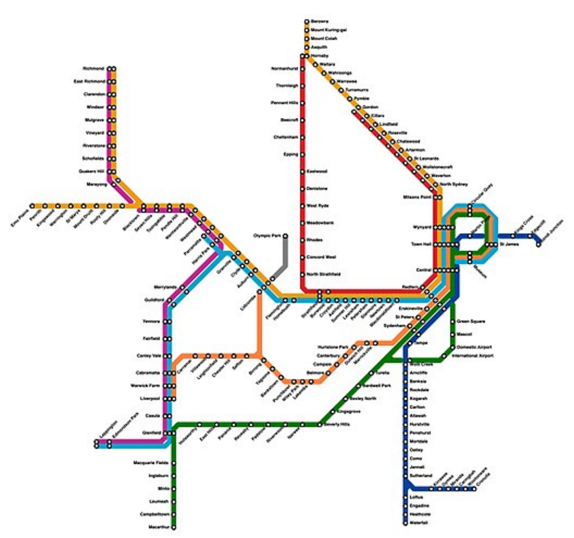

Fig. 1 Sydney Trains network runs Waratah and non-Waratah trains. Waratah trains are equipped with an occupancy weighing system, whereas non-Waratah trains are not. This means the prediction model is built only from data of Waratah trains, but need also address forecasting for non-Waratah trains.

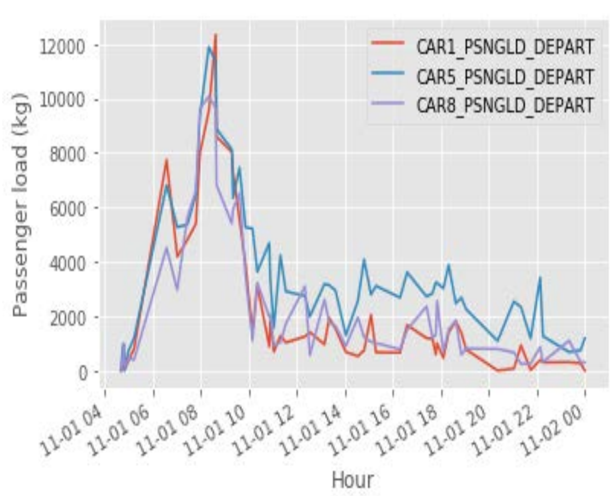

Fig. 2 Examples of loading data
Date: 01/11/2019 to 03/11/2019.
Stations: Central to Schofields;
Current station: Redfern;
Platform: RD05;
Carriage ID: CAR1, CAR5, CAR8.

Methodology

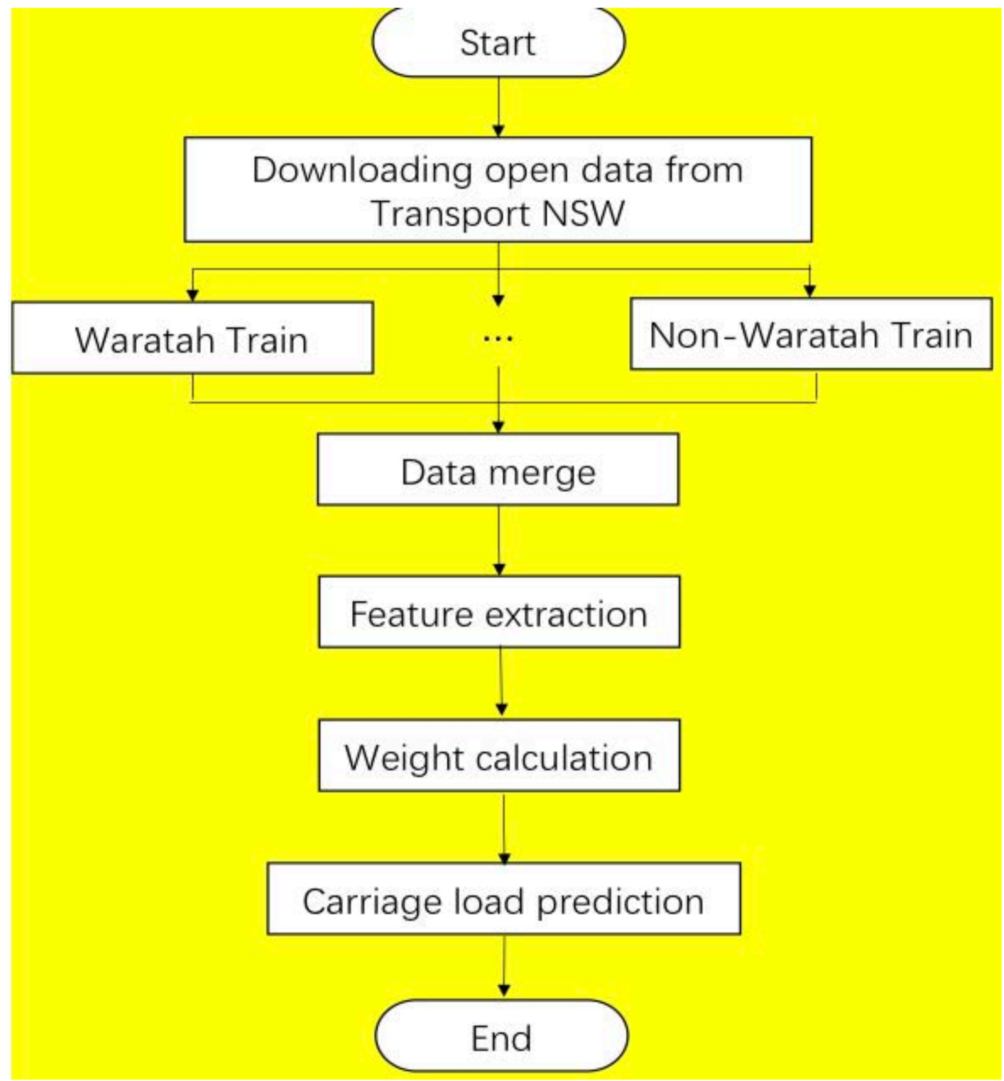

Fig. 3 A flow chart of the prediction model

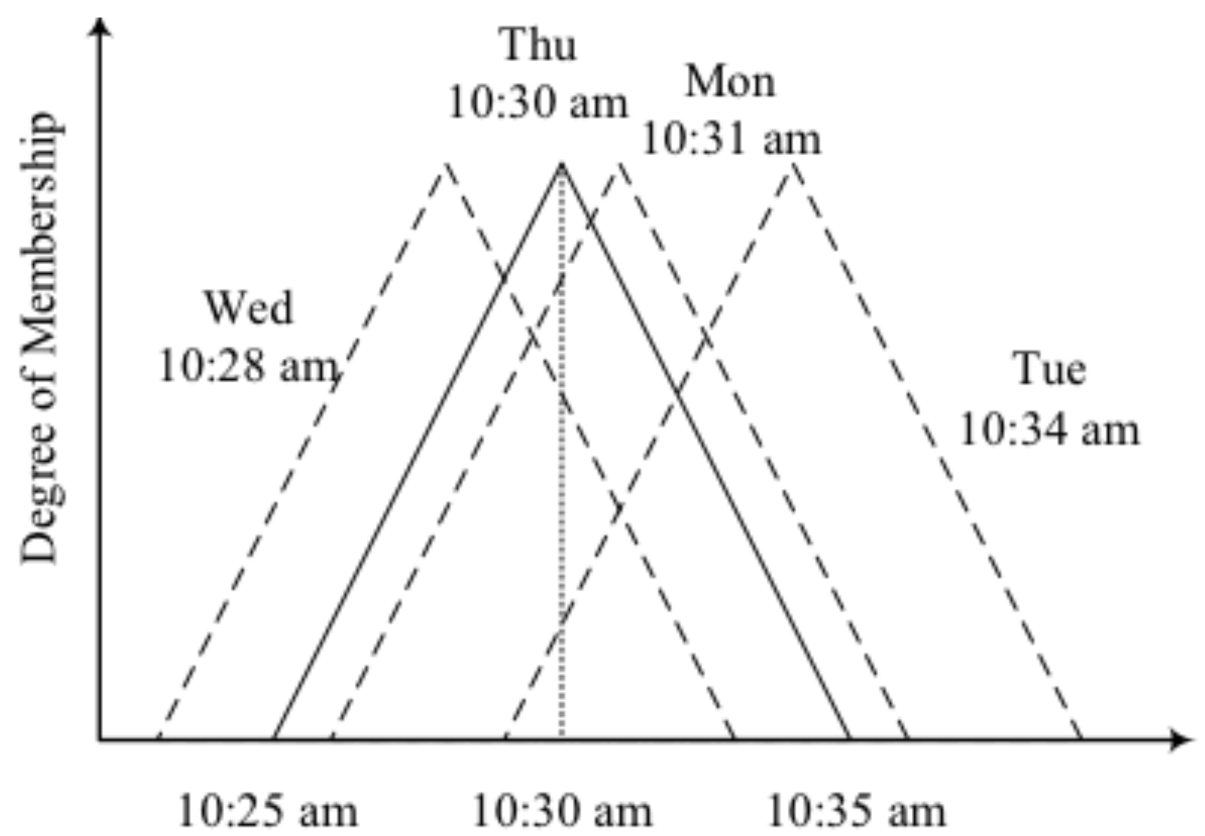

Fig. 4 A demonstration of fuzzy time interval matching for calculating the contextual moving average

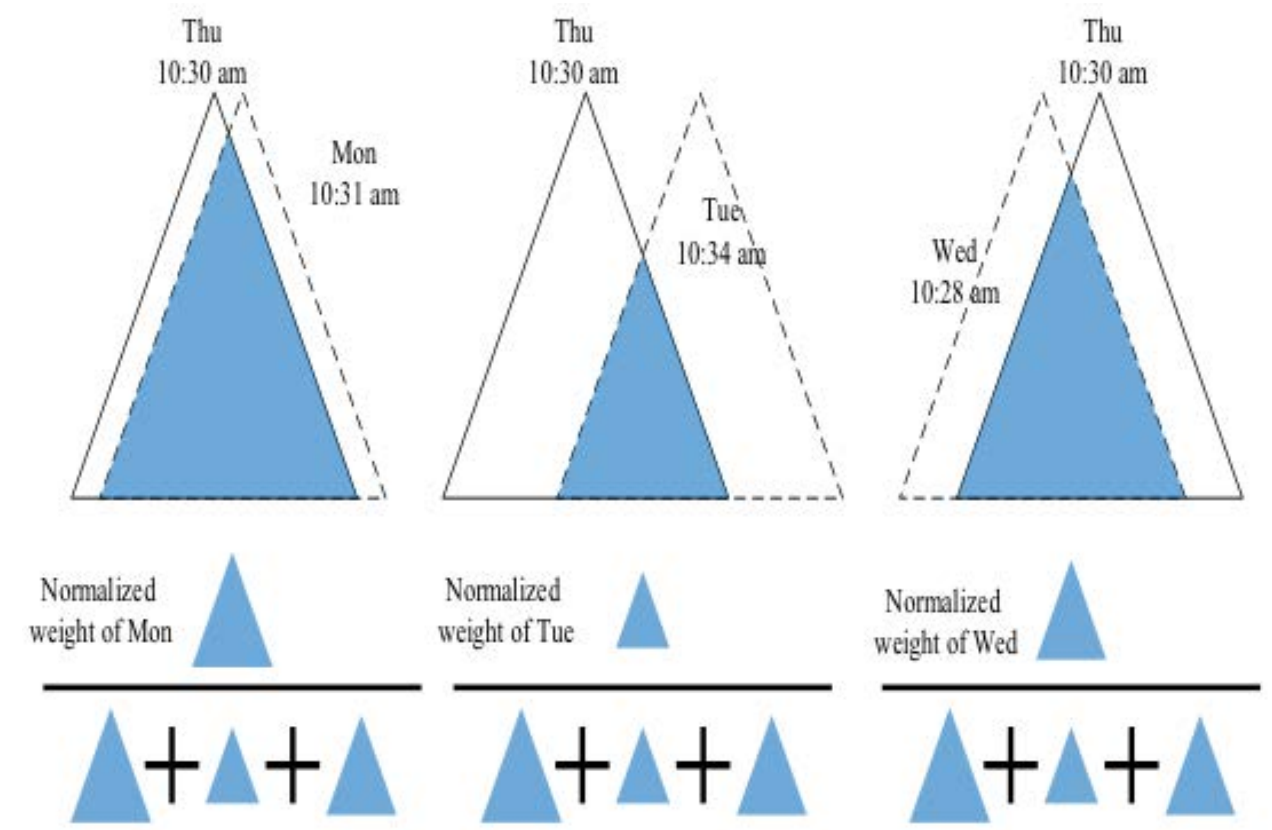

Fig. 5 Calculating the normalized weight

Experimental Results

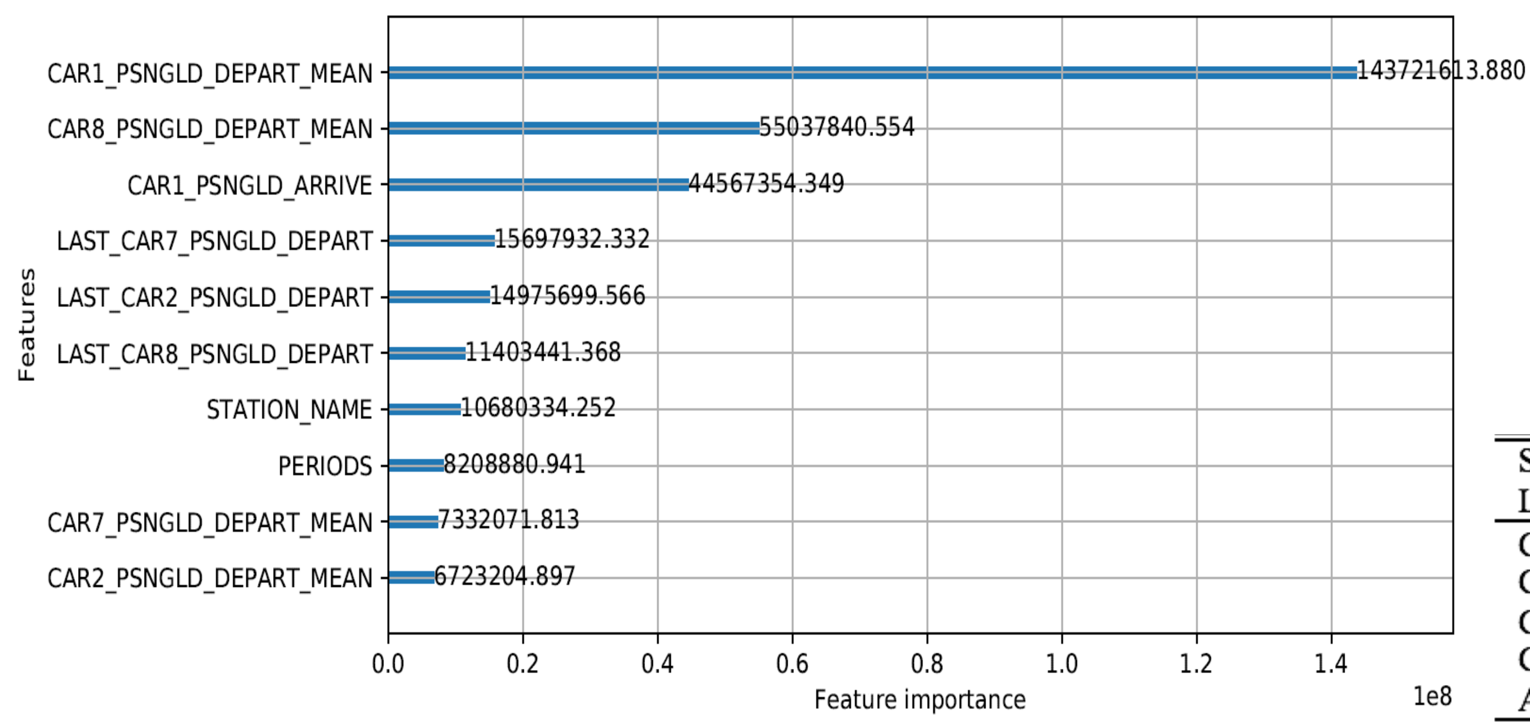

Fig. 6 LightGBM predictions of feature importance for carriage 1.

	WaraNorm	WaraAbno	NonwNorm	NonwAbno	Average
Accurate Matching	299.68	489.49	550.33	882.10	475.37
Fuzzy Matching	279.22	469.34	541.08	847.16	465.01
Fuzzy Matching + Fuzzy Weighting	**255.75**	**420.51**	**526.12**	**806.70**	**445.48**

Fig. 7 MAE for the LightGBM with different fuzzy learning schemes

Sample Size	25000		20000		15000		10000	
Learning Method	MulOutput	MulStream	MulOutput	MulStream	MulOutput	MulStream	MulOutput	MulStream
Car 1-8	386.90	382.14	392.52	386.15	395.32	386.98	397.46	391.44
Car 2-7	492.63	486.58	498.25	489.57	501.72	493.28	510.84	502.23
Car 3-6	513.15	505.01	518.86	510.44	520.38	512.51	531.32	518.40
Car 4-5	487.22	470.69	492.00	471.34	496.92	472.85	507.69	481.03
Average	469.98	461.10	475.41	464.37	478.58	466.40	486.83	473.28
Improvement	8.87		11.03		12.18		13.55	

Fig. 8 Multi-stream learning with different sample sizes

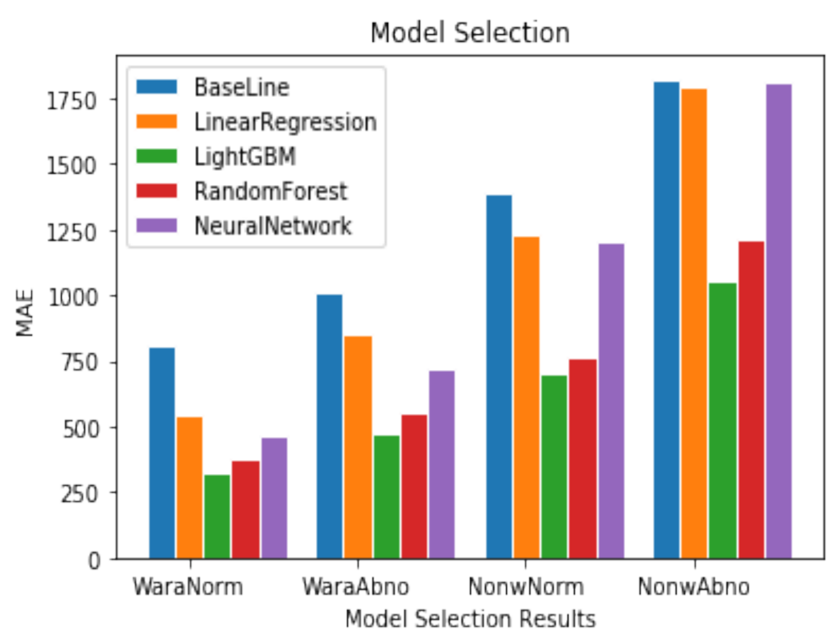

Fig. 9 Model selection results for four tasks.

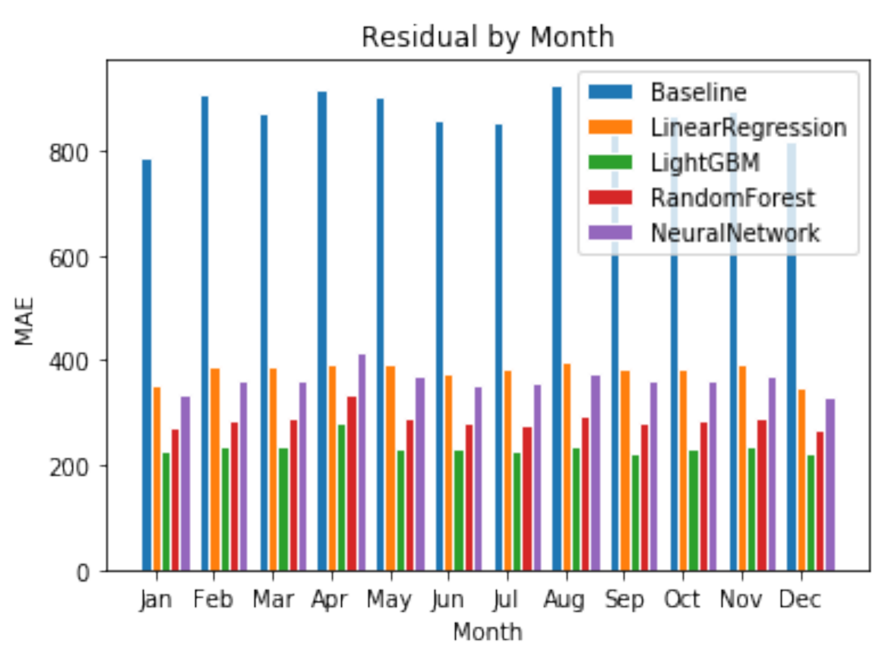

Fig. 10 MAE by month

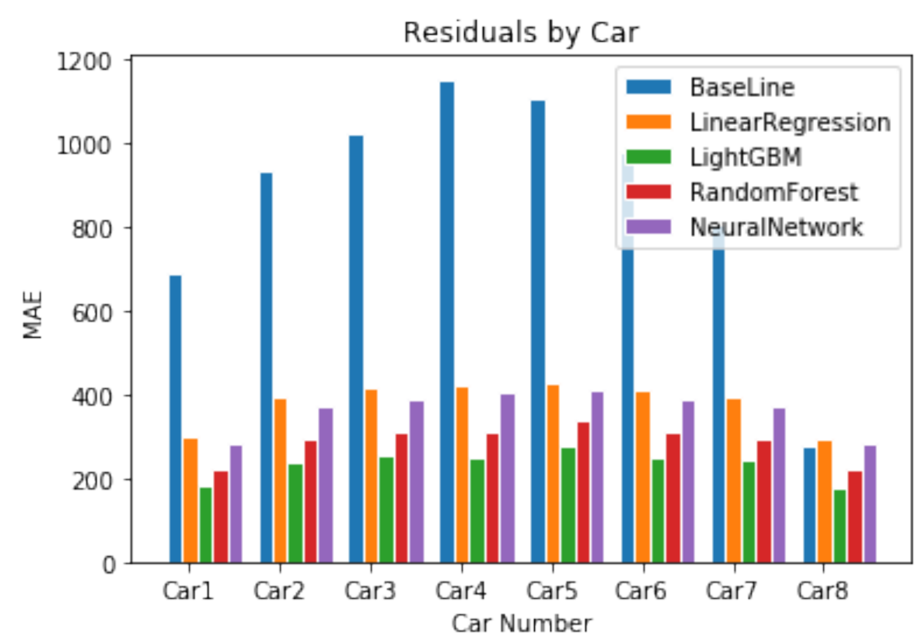

Fig. 11 MAE for each carriage

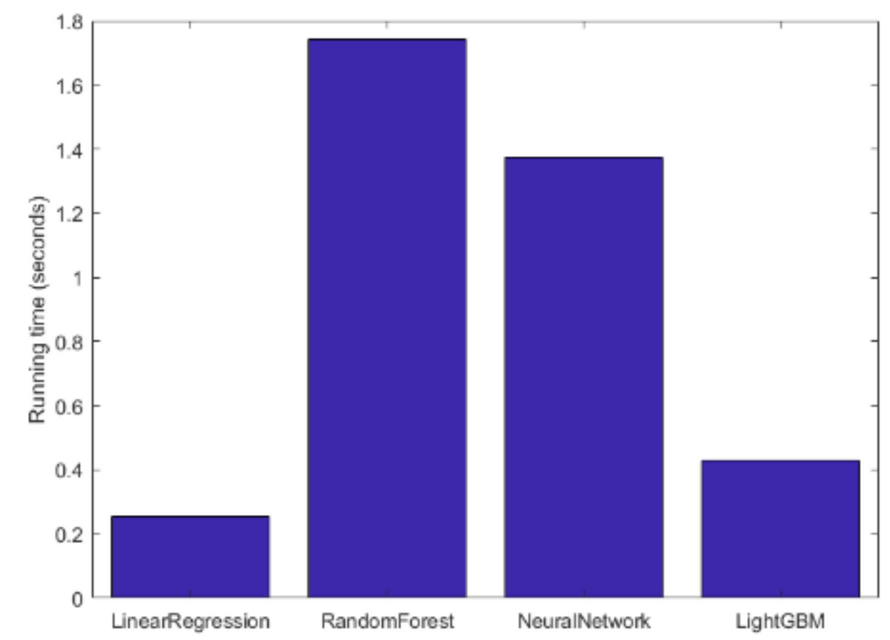

Fig. 12 Running times