A Machine-Learning Decision-Support Tool for Travel-Demand Modeling

C. Scott Brown, Venu Garikapati, and Yi Hou

Transportation Research Board Annual Meeting
January 13–17, 2019

NREL/PO-5400-72993

MOTIVATION

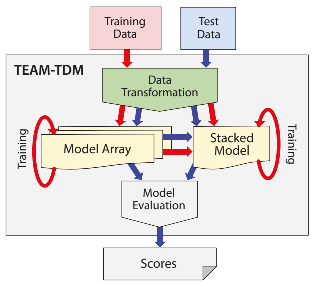
- Logit-based choice models have long been the golden standard for classification modeling in transportation
- This is partly motivated by the simplicity of interpretation of logit models and the fact that they are so deeply ingrained in the current infrastructure of transportation modeling
- Machine-learning (ML) models are being adopted in various domains and have been shown to be more accurate than traditional models at many tasks
- We propose a modeling pipeline to provide practitioners with a simple yet effective means of gauging the predictive abilities of utility maximization ML algorithms for a given modeling context

OBJECTIVE

- Where do we focus our efforts in introducing new model families?
  - Is there a simple heuristic to determine if alternative model families may have superior performance to linear models?
  - If so, can we automate the process?
- In other words, can we make a tool that allows us to quickly evaluate the pros and cons of using different model families for a given problem?

TOWARD A SOLUTION: A MACHINE-LEARNING EVALUATION ASSISTANT

- TEAM-TDM: A Tool for Evaluating Applications of Machine Learning in Travel-Demand Modeling
- Dummy variables/data scaling, (some) model tuning, model training, and model evaluation are automated



FEATURE TRANSFORMATION AND SELECTION

- For a given dataset, categorical variables are identified a priori and a list of which variables are categorical is given as a parameter to TEAM-TDM
- TEAM-TDM transforms these variables via encoding, which is equivalent to the introduction of a dummy variable for each possible class of the variable
- In order to streamline the selection of features, a random-forest classifier is fit to the data, and the mean decrease impurity is computed for each feature

MODEL EVALUATION

$$precision = \frac{n_{true\ positives}}{n_{true\ positives} + n_{false\ positives}}$$

$$recall = \frac{n_{true\ positives}}{n_{true\ positives} + n_{false\ negatives}}$$

$$MAMSE_{macro} = \left\| \frac{\sum_i C_{ij} - \left(\sum_j C_{ij}\right)^T}{\sum_i C_{ij}} \right\|_{L_1}$$

$$MAMSE_{weighted} = \frac{\left\| \sum_i C_{ij} - \left(\sum_j C_{ij}\right)^T \right\|_{L_1}}{\sum_{i,j} C_{ij}}$$

Where C is the confusion matrix for the model

EXPERIMENTS

- Given a modeling pipeline that requires little or no adjustment, we a priori pick a handful of model families and determine which hyperparameters need tuning
- Data from 2017 National Household Travel Survey (NHTS) is used for analysis
- Given this configuration, we train on two different problems:
  1. Vehicle ownership (number of vehicles owned by a household)
  2. Work schedule (start and end times)

Data Description

Variable	Description	Mean	Median	Standard Deviation	Importance	
DRVRCNT	Number of drivers in household	1.677	2.0	0.767	0.075	NHTS Vehicle Count
RESP_CNT	Count of responding persons	2.129	2.0	1.167	0.033	
$HHRELATD_0$ (dummy)	No household members are related	0.664	1.0	0.473	0.030	
CNTTDHH	Count of household trips on travel day	7.121	6.0	5.810	0.025	
WRKCOUNT	Number of household workers	0.989	1.0	0.899	0.022	
NUMADLT	Count of household member > 18 y.o.	1.781	2.0	0.712	0.020	
CAR_0 (dummy)	Respondent never uses personal vehicle	0.026	0.0	0.160	0.012	
HHSIZE	Count of household members	2.129	2.0	1.167	0.012	
LIF_CYC_0 (dummy)	Household has one adult, no children	0.212	0.0	0.409	0.011	
$HHRELATD_1$ (dummy)	At least 2 household members are related	0.336	0.0	0.473	0.009	
$HOMEOWN_0$ (dummy)	Respondent owns home	0.759	1.0	0.428	0.009	
CAR_1 (dummy)	Respondent uses personal vehicle daily	0.776	1.0	0.417	0.008	
DWELTIME	Time at destination	473.055	512.000	161.722	0.009	NHTS Work Schedule
GCDWORK	Geodesic distance to work	12.473	6.380	67.014	0.005	
R_AGE	Age of respondent	45.130	46.000	14.734	0.005	
$TRPMILES_0$	Trip distance to work	13.534	8.595	47.846	0.005	
DISTTOWK17	Road network distance to work	16.107	8.830	75.840	0.005	
$TRVLCMIN_0$	Trip duration to work	26.389	20.000	24.509	0.005	
$TRPMILES_1$	Trip distance from work	13.355	7.998	52.757	0.005	
VMT_MILE_0	Personal vehicle trip miles to work	11.776	7.507	28.206	0.005	
TIMETOWORK	Reported average trip time to work	24.674	20.000	25.151	0.005	
$TRVLCMIN_1$	Trip duration from work	28.356	20.000	28.432	0.005	
VMT_MILE_1	Personal vehicle trip miles from work	11.358	6.926	26.682	0.005	
CNTTDHH	Count of household trips on travel day	8.862	8.000	5.978	0.005	

Results

Evaluation Criteria	RF	MNL	MLP	NB	Dummy	OP	NL	Stacked	Best Model	
accuracy	0.630	0.611	0.643	0.614	0.255	0.640	0.650	0.655	NL	NHTS Vehicle Count
weighted precision	0.611	0.572	0.583	0.599	0.258	0.620	0.623	0.631	NL	
weighted recall	0.630	0.611	0.643	0.614	0.255	0.640	0.650	0.655	NL	
macro precision	0.245	0.219	0.201	0.248	0.078	0.263	0.248	0.262	OP	
macro recall	0.199	0.199	0.211	0.222	0.078	0.219	0.223	0.229	NL	
mean log loss	1.062	1.062	1.105	1.947	25.349	1.061	1.040	1.038	NL	
macro MAMSE	10.301	33.761	9.365	20.631	7.678	11.448	8.716	19.389	NL	
weighted MAMSE	0.401	0.320	0.224	0.190	0.051	0.306	0.220	0.210	NB	
training time (s)	268.247	190.623	6701.169	0.650	0.068	144.772	11.390	4795.457	NB	
accuracy	0.212	0.572	0.626	0.260	0.032			0.593	MLP	NHTS Work Schedule
weighted precision	0.214	0.571	0.585	0.438	0.029			0.587	MLP	
weighted recall	0.212	0.572	0.626	0.260	0.032			0.593	MLP	
macro precision	0.064	0.334	0.242	0.179	0.004			0.339	MNL	
macro recall	0.025	0.292	0.286	0.142	0.004			0.292	MNL	
mean log loss	3.656	3.942	1.968	20.363	33.417			12.839	MNL	
macro MAMSE	211.669	152.784	279.702	1116.607	250.511			15.454	MLP	
weighted MAMSE	0.146	0.235	0.348	0.886	0.304			0.249	MNL	
training time (s)	1383.772	2214.177	181095.857	10.359	1.171			194215.423	NB	

Acronyms: random forest (RF), multinomial logit (MNL), multi-layer perceptrons (MLP), naïve bayes (NB), ordered probit (OP), nested logit (NL), and mean absolute market share error (MAMSE)

RESULTS

- For vehicle ownership prediction, the nested logit model seems to perform best, although the MLP and OP models are not far behind
- For work schedule prediction, the MLP model is somewhat better than the MNL model in certain aspects but performs worse on minority classes and market share
- Results are comparable to other experiments in the literature using similar model families
- The resulting metrics can be used as post hoc heuristics for deciding which model families will provide the most value with the least effort

CONCLUSIONS

- Accommodate handling of messy data, unbalanced data, and outliers
- Extend the tool to handle regression, clustering, and mixed discrete continuous models
- Inclusion of more model families and extension to applications beyond travel-demand modeling

FUTURE DIRECTIONS

- Adapt the tool to handle regression and classification
- Add additional model families and perform more experiments
- The tool, currently tuned to travel-demand modeling problems, could be adapted to other problem areas in transportation

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Vehicle Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.