HOW DOES MACHINE LEARNING COMPARE TO CONVENTIONAL ECONOMETRICS FOR TRANSPORT DATA SETS? A TEST OF ML VS MLE

Weijia (Vivian) Li
School of Highway, Chang'an University, South Erhuan Middle Section, Xi'an, Shaanxi 710064, China
Visiting Scholar at University of Texas at Austin
liweijiachd@126.com

Kara M. Kockelman, Ph.D., P.E.
(Corresponding Author)
Dewitt Greer Centennial Professor of Transportation Engineering Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin – 6.9 E. Cockrell Jr. Hall
Austin, TX 78712-1076
kkockelm@mail.utexas.edu
Phone: 512-471-0210 & FAX: 512-475-8744

Growth and Change, 2021 https://doi.org/10.1111/grow.12587

Word Count: 6,664 words + 3 tables = 7,414 words (+ 3 figures)

ABSTRACT

Machine learning (ML) is being used regularly in transportation and other applications to predict response variables as a function of many inputs. This paper compares traditional econometric methods of data analysis to ML methods in the prediction and understanding of four key transport planning variables: household vehicle-miles traveled (a continuous variable), household vehicle ownership (a count variable), mode choice (a categorical variable), and land use change (a categorical variable with strong spatial interactions). Here, the results of decision trees (DTs), random forests (RFs), Adaboost, gradient boosting decision trees (GBDTs), XGBoost, lightGBM (gradient boosting methods), catboost, neural networks (NNs), support vector methods (SVMs) and Naïve Bayesian networks (BN) are compared to methods of ordinary least squares (OLS), multinomial logit (MNL), negative binomial and spatial auto-regressive (SAR) MNL methods using the U.S.'s 2017 National Household Travel Survey (NHTS 2017) and land use data sets from the Dallas-Ft. Worth region of Texas. Results suggest that traditional econometric methods work pretty well on the more continuous responses (VMT and vehicle ownership), but the RF, GBDT and XGBoost methods delivered the best results, though the RF model required 30 to almost 60 times more computing time than XGBoost and GBDT methods. The RF, GBDT, XGBoost, lightGBM and catboost offer better results than other methods for the two "classification" cases (mode choice and land use change), with lightGBM being the most time-efficient. Importantly, ML methods captured the plateauing effect that modelers may expect when extrapolating covariate effects.

Key words: machine learning; artificial intelligence; econometric methods; travel behavior prediction; model estimation comparisons; ensemble methods

INTRODUCTION

Machine learning (ML) is a branch of artificial intelligence (AI) that employs a variety of statistical, probabilistic and optimization techniques. Instead of making strict assumptions about random components and equations linking predictor variables or covariates to response variables (X's to Y's), ML methods allow computers to "learn" and detect hard-to-discern patterns from large, noisy and/or complex data sets. Thanks to computing advances, ML is already being widely tested across various areas of transport data analysis, including truck-type classification (Huang et al., 2018), mode recognition (Xie et al., 2003, Wang et al., 2019), travel time prediction (Zhang et al., 2015), injury severity prediction (Delen et al. 2017, Hamad et al., 2019, Das et al., 2019), traffic flow prediction (Hosseini et al., 2019, Cui et al., 2019), trip purpose identification (Deng et al., 2010), and automated-vehicle applications (Liang et al., 2019). To date, the most commonly used ML algorithms appear to be logistic regression (LR), using gradient ascent to calibrate model parameters (Caruana et al., 2006; Chen et al., 2016; Delen et al., 2017), decision tree (DTs), the support vector method (SVM), naïve Bayes (NB), k-nearest neighbors (KNN), K-means, random forest (RF), bagging algorithms, gradient boosting decision trees (GBDT), XGBoost (extreme gradient boosting), lightGBM (light gradient boosting method), and categorical boosting algorithms (catboost). Some of these are old, traditional methods (based on basic clustering math or a single sigmoidal equation). Others do not have target values or training data and are used for smoothing behaviors with reinforcement along the way (e.g., a self-driving car avoiding near-crashes and high centrifugal forces gets a higher score).

Data mining models have more flexible structures than traditional econometric models in representing the relationship between the attributes of alternatives and choices. Thanks to such flexibility, they may be able to offer valuable insights into relationships that random-utility-based choice models cannot recognize (see, e.g., Xie et al., 2003, Xu et al., 2018, Keya et al., 2019). Tree-based ML methods are often preferred by modelers because their specifications are relatively clear (as a series of discrete branches, from variable to variable), and this method appears to be quite resistant to outliers, noise/random errors, missing data, and overfitting (using training data) (Harb et al., 2009). Compared to other ML methods, which are regarded as black boxes, the tree-based ensemble methods (i.e., RF methods) are easily interpreted and can solve complex nonlinear relationship, which enable a better understanding of Y-vs-X relationships (Zhang et al., 2015). The following will describe the ensemble methods from Bagging to Boosting.

Bagging

RF is a common bagging methods that combines Brieman's bagging idea with Ho's random subspace method (Harb et al., 2009). Cheng et al. (2019) and Seksar et al. (2014) reviewed many recent RF studies, finding consistently strong predictive accuracy across distinctive data sets. RF also can identify, interpret, and rank-order the relevance of competing covariates and their interactions. Bagging methods involves random sampling of small subsets of data, and gives all resulting weak models equal probability. So each sample and its associated model have the same weight, running in parallel, which is very different from boosting methods, since those give more weight or credit to stronger/less weak models among the samples.

Boosting

Freund (1997) proved that using "boosting" algorithms to combine many weak models can result in a single and highly accurate model - one of the most powerful ideas for prediction over the past 20 years (Hou et al., 2018). Boosting methods use the same data set to run the prediction models, with the same weight given to each data set when predicting the first weak model, then using prediction error rates to adjust sample and model weights to predict the next model combination. Thus, boosting methods must run in sequence, which slows prediction speeds but delivers better prediction accuracy, as compared to bagging methods (Caruana et al., 2006). Gradient boosting, which combines gradient descent and boosting to improve prediction speed and accuracy, is an extension of the over-boosting method (developed by Friedman [2001]). Such algorithms are typically one of the following four methods: GBM (gradient boosting method), XGBoost, lightGBM, and catboost.

GBMs generate base models sequentially and improve prediction accuracy since misclassified or incorrectly estimated samples are more likely to be selected (using larger weights) on successive estimation rounds. Zhang et al. (2015) used a gradient boosting regression tree (a kind of GBM) to predict travel times with greater accuracy than ARIMA and RF. Ding et al. (2016) devised the GBDT method (a kind of GBM) to estimate short-term subway ridership levels (by station) and showed how GBDT can reliably rank the relative influence of different covariates - like bus transfer activities and temporal features (Time of day, day, week and month were incorporated in the prediction model). It is also used compared to BP-NN, SVM and RF. For BP-neural network, and found that GBDT model receives the best model performance (higher R^2). Proposed by Fruend in 1997, Adaboost is another kind of GBM to compare to RF (Miao et al., 2010), NN (Alfaro et al., 2008), and SVM (Wang et al., 2006). All of the results showed that AdaBoost outperforms other methods and is more robust than SVM.

Recently, XGBoost has received a lot of attention. Hou et al. (2018) used single DT, RF, GBM and XGBoost to predict roadway traffic flows and found similar accuracy across methods, with XGBoost requiring the lowest computing times. Wang (2018) used the XGBoost, lightGBM and DTs methods to predict travel mode choices and found the lightGBM and XGBoost methods be more accurate than DTs, with lightGBM most preferred.

Compared to GBDT, Ke et al. (2017) verified that lightGBM reduced training times by 95% or more, while achieving nearly the same predictive accuracy (measured as AUC). The main difference between lightGBM and the XGboost algorithms is that lightGBM uses a histogram-based algorithm to speed up the training process, reduce memory consumption, and adopt a leaf-wise leaf-growth strategy with depth limitations (Guo et al., 2017). Catboost is also a cutting-edge ML technique, which is good at dealing with categorical data. (Dorogush et al. 2018).

As analysis before, ensemble methods generally deliver better predictions than other algorithms. Consider ANN, SVM and Bayesian network models are also widely used in model prediction. This paper tested the following 10 ML models: decision trees (DTs), random forests (RFs), Adaboost, gradient boosting decision trees (GBDT), XGBoost, lightGBM, catboost, ANNs, SVMs and Bayesian networks. All these methods can be used for categorical or count or continuous response variable prediction. Their results were then compared to those from traditional estimation models (OLS, negative binomial, and MNL specifications) for prediction of annual household VMT, household vehicle ownership, and mode choices. Spatial econometric techniques (MNL SAR) were also used to analyze the land use change data. Data come from Dallas-Ft. Worth's 2017 National Household Travel Survey (NHTS 2017) add-on sample and from the Dallas-Ft. Worth region more generally (via its metropolitan planning organization: the North Central Texas Council of Governments, or NCTCOG).

Interestingly, almost no ML users in the world are yet making their ML model specifications transparent. For example, Hou et al. (2018), Linero and Nethery (2015), Folden (2018), Goh (2018), Brown (2019), Pu (2019), and nearly all others only compare competing models' predictive accuracy (using, e.g., root-mean-squared errors [RMSE values]) and simply rank the explanatory variables (on the basis of dE(y)/dx slope comparisons, typically). Thus, this new paper examines not just goodness of fit in predicting hold-out sample values, but emphasizes interpretation of results for behavioral understanding, smarter planning and policy making, and better investment and management decisions.

MACHINE LEARNING (ML) METHODS

ML methods are extensive. Because ANNs, BN and SVM are three common methods which are already used before the ML and AI emerge, and many papers have describe these three algorithms, meanwhile, due to space constraints, more information on ANNs, BN and SVM methods can be found in Xiao et al. (2015), Omrani (2015), Delen et al. (2017) and Iranitalab (2017). This paper will pay more attention to ensemble methods, below are key details for specifications of Bagging and Boosting.

Bagging Trees via Random Forest

Breiman proposed the RF method in 1996. RFs combine decision tree (DT) predictors by using bootstrap method, and then using voting (for classifier) or average (for regression) the results of each DT to obtain the final prediction value. In RF, each decision node uses the best among a subset of predictors randomly chosen at that node. This method has performed very well compared to many other commonly-used classifiers and is resistant to overfitting (Breiman, 1996).

Boosting Techniques

This section describes 5 top boosting methods: Adaboost, GBDT, XGBoost, lightGBM and catboost, which are presented in the chronological order they were introduced in the literature, also from simpler to more complex.

Adaboost

Adaptive boosting (Adaboost) was introduced by Freund and Shapire in 1997; it combines a collection of weak models to provide a stronger model. With *m* training-set data points, this method uses initial weights of 1/*m* for each observation to "train" (i.e., estimate) a set of base (initial, weak) models. The predictions of these (the number of base models depend on the prediction accuracy and overfitting or not) base models are used to generate a new set of weights call *alpha* values by using prediction accuracy (*alpha = 0.5*ln (1-accuracy/total) / (accuracy/total)*), which are then used to update data-point weights. Larger weights are assigned to samples with more misclassified data points (i.e., to harder-to-predict data sets), and these updated weights are used to train the next model version. Final predictions for each data point are obtained as an *alpha*-weighted "voting" (or score) total across all upstream models (Krishnaveni et al., 2011).

Gradient Boosting Decision Tree (GBDT)

Like other boosting methods, GBDT combines several relatively weak models into a stronger model. It uses forward stage-wise additive modeling, GBDT only can uses CART (Classification And Regression Tree, which use Gini impurity to select the node classifier variables, smaller Gini impurity means better prediction accuracy) trees to be the base learners, loss of function is minimized in the direction of its steepest-descent, the aim of each iteration is to decrease the last residual. GBDT algorithms resist noise components and outlier data points by bagging and adding regularization, making GBDT more robust than other methods. For these reasons, GBDT is now widely used for classification of data points, prediction of continuous response variables (which ML users call "regression"), and ranking of outcomes or decisions (Ding et al., 2016).

Extreme Gradient Boosting Model (XGBoost)

Friedman (2001) proposed the XGBoost model as a well-developed algorithm in the boosted DT family (Hou et al., 2018). It also is trained (estimated) in a forward "stage-wise" (i.e., input by input) fashion, to minimize the "loss function" (typically the sum of squared error terms), by adjusting parameters bit by bit (to remove residual correlation between an input variable and the loss function). XGBoost uses a parallel tree boosting method (also known as GBM with DTs) that solves many data science problems in a fast and accurate way (Hou et al., 2018). It can use CART or gblinear as its base model, and add regulation to the objective function. A second-order Taylor series expansion is used to approximate the method's objective function (i.e., it's loss function, like the RMSE), based on second derivatives.

LightGBM

For very large data sets, the LightGBM technique is often helpful. It combines gradient-based one-side sampling (GOSS) and exclusive feature (covariate) bundling (EFB) to tackle the problem of model complexity (thanks to a very large set of covariates and/or training data). GOSS method keeps those instances with large gradients, and only randomly drop those instances with small gradients (leaf-wise tree growth strategy with depth limited [other algorithms grow level-wise]). LightGBM has been found to accelerate "training speed" (i.e., estimation times) while delivering accuracy compared to other ML methods (Ke et al., 2017).

Catboost

The Yandex Company proposed "Catboost" in 2017 to boost estimation for categorical data. Catboost is considered a cutting-edge ML technique, able to compete with any leading ML algorithm on the performance front (Dorogush et al. 2018). Catboost, which use complete binary tree, can be used without any explicit pre-processing of input values to convert categories into numbers, so it can handle categorical "features" (i.e., covariates) automatically. More specifically, data set is randomly permutated firstly, then for each sample, it calculates average label value for samples with the same category value prior to the given one in the permutation, which is different from one-hot encoding. It also reduces the need for extensive parameter tuning and lowers the chances of overfitting, which leads to more robust prediction (Dorogush et al., 2018).

MODEL FIT ASSESSMENT

Different model-fit statistics or loss-function values are used to evaluate and compare ML and other estimation techniques. The better score statistic to use generally depends on the response variable's characteristics (e.g., is it binary, categorical, ordered or continuous?). The following sections discuss scoring techniques, starting with a classification case.

Classification Problems for Categorical Response Data

Contents of a confusion matrix for response categorization are commonly used to compare competing models' performance, based on the metrics of recall, precision, F_1 score, and what ML modelers call "AUC" (which stands for area under the receiver operating characteristic or ROC curve).

Using a 3-category example, class or category *i*'s precision is simply the fraction of data points categorized correctly into class *i*: $p_i = n_{ii}/ (n_{ii}+n_{i2}+n_{i3})$, where n_{ij} is the number of data points with a class *i* response that are model-classified as class *j*.

The 3 positive predictive values (p_1, p_2 and p_3) and the true positive rates of the three classes ($P_1 = n_{11}/n_{i1}$, $P_2=n_{22}/n_{i2}$, and $P_3=n_{33}/n_{i3}$) are then harmonically averaged to calculate the index called F_{1score}, which is $1/3(\Sigma 2/(1/p_i + 1/P_i)$, and this statistic ranges from 0 to 1, with 1 being the best score feasible (with all class 1 data points properly categorized, and no others mistakenly placed in that category).

Regression Problem (Continuous Response Variable Prediction)

Scoring criteria commonly used to improve models and compare models with continuous response variables are the mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), root mean squared logarithmic error (RMSLE), and coefficient of determination (R^2) or correlation between actual and predict values (ρ = sqrt(R^2) in the bivariate context). The R^2 and RMSE values are defined in Eqs. 2 and 3:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum((y_i - \hat{y}_i)^2)}{\sum((y_i - \bar{y})^2)} \quad (3)$$

where y_i is the observed output value, and $\hat{y}_i$ is the model-predicted response or output.

Avoiding Overfitting

ML results can suffer from overfitting problems, since they can have hundreds, thousands or more parameters embedded in them, making them too responsive to outliers and demanding very large data sets, which many researchers will not have. ML researchers avoid this issue by holding out a testing (validation) sample of the data, apart from the training (calibration) data set used to create the model. Using random processes, if there a considerable deviation in model prediction accuracy between these two distinctive data sets, parameters in the trained model are adjusted. ML routines are typically programmed to replicate this process of partitioning the initial data set, fitting the model (on each training data set), and evaluating the success rate or scores R different times. The average of these scores is the overall success rate.

Sensitivity

Many people feel that ML methods are like black boxes, since few make explicit the parameters and equations behind them, or at least the relationships between outputs and inputs (i.e., between response values Y and each explanatory variable X_i). New sensitivity analyses have been proposed to address this long-standing weakness (Delen et al., 2017). In this paper, the average change in model-predicted output values (ΔY) with respect to a 1 standard deviation or binary (0 to 1) change in every input (for every data point) is used to appreciate the relationship between each explanatory variable and the dependent variable. This effect is calculated differently for continuous versus binary inputs (e.g., x1 = Age vs x2 = Gender or Presence of Children in the Respondent's Household), since the latter does not change smoothly/continuously in practice. The difference in the average response was divided by the response variable's standard deviation to normalize the impacts, for ready comparison across different response-variable contexts (e.g., Y1 = VMT vs. Y2 = Vehicle Ownership). As shown in Eq. 4.

$$\Delta Y\ w.r.t = \frac{y_{changed} - y_{unchanged}}{\sigma} \quad (4)$$

DATA SETS

This work relies on two data sets to do estimate four distinctive response variables. These are the Dallas-Ft. Worth or "DFW" sample in the U.S. 2017 National Household Travel Survey (NHTS 2017) data set and the land use data sets provided by DFW's metropolitan planning organization (NCTCOG). After removal of records with missing covariates, the four data sets' sample sizes are as follows: daily household-level VMT: n_1 = 8,676 households; vehicle ownership: n_2 = 8,545 households; model choice: n_3 = 13,834 person-trips; and land use change: n_4 = 99,304 grid cells (each measuring 30 m by 30 m).

The data sets contain a variety of valuable variables, including individual and household demographics, trip-level and mode-alternative attributes and neighborhood land use and access details. Table 1 provides summary statistics for all variables considered here.

TABLE 1 Summary Statistics for four Transportation-related Data Sets

HHVMT Model (n_1 = 8,676)					
Variable	Description	Mean	Std. Dev.	Min	Max
HHVMT	Vehicle-miles travelled (miles)	25,612.44	21,823.6	2	402,000
Own Home	1. Own (76.40%); 2. Rent (23.60%)	1.236	0.425	1	2
HH Size	# household members	2.229	1.202	1	10
#HH Veh	# household vehicles	1.971	1.0515	0	10

HH Income	Household income per year	$85,517	$56,975	$5,000	$202,000
#Drivers	# household drivers	1.744	0.753	0	7
#HH Trips	# person-trips on travel day	7.463	5.947	0	53
#Children	# of children in household	0.105	0.384	0	4
#Workers	# household workers	1.141	0.881	0	5
UrbRur	1. Urban (92.37%); 2. Rural (7.63%)	--	--	1	2
DistCBD	Distance to the CBD (miles)	22.67	12.34	0.119	73.82
Workerden	Workers per square mile in the census tract of the household's home location	1,892.5	1,393.7	25	5,000
Vehicle Ownership Model (n_2 = 8,545)					
Variable	Description	Mean	Std. Dev.	Min	Max
#HH Veh	Count of household vehicles	1.974	1.054	0	10
Own Home	1. Own (74.37%); 2. Rent (25.63%)	--	--	1	2
HH Size	Count of household members	2.234	1.206	1	10
HH Income	Household income	$86,245.47	$58,442.45	$5,000	$210,000
#Drivers	Count of household driver	1.746	0.754	0	7
#Workers	Count of household worker	1.144	0.881	0	5
UrbRur	1. Urban (92.29%); 2. Rural (7.71%)	1.077	0.267	1	2
Popden	Population density (persons per square mile) in the census tract of the household's home location.	3,673.54	3,189.55	5,619	41,300.6
Mode Choice Model (n_3 = 13,834)					
Variable	Description	Mean	Std. Dev.	Min	Max
Mode Choice	1. Drivealone (52.33%); 2. Carpool (39.94%); 3. Walk/Bike (3.85%); 4. Transit (3.87%)	--	--	--	--
Travel Time	Travel time (min)	53.615	63.018	1	1,160
Cost	Cost	$4.56	$18.77	$0	$654.78
Trip Purp	3 Trip purposes: 1. NHB (21.62%); 2. HBO (51.77%); 3. HBW (26.6%) = Non-home based, home-based other/non-work, and home-based work trips	--	--	1	3
Age	Age of the traveler	45	21	5	100
Driver	1. Yes (83.02%); 2. No (16.97%)	1.1697	0.3754	1	2
Sex	1. Male (47.46%); 2. Female (52.53%)	1.5253	0.4993	1	2
HH Size	Count of household members	2.8361	1.3882	1	10
#HH Veh	Household vehicle number	2.235	1.0582	0	10
HH Income	Household income	98,871	66,483	5,000	250,000
Workerden	Workers per square mile in the census tract of household's home location.	1,806.9	1,694.6	9.25	19,269.2
Land Use Change Model (n_4 = 99,304)					
Variable	Description	Mean	Std. Dev.	Min	Max
Land use change	1. Residential (6.69%); 2. Commercial/civic (2.00%); 3.Underdeveloped (60.90%); 4. Other (6.21%); 5. Ranch land (24.20%).	--	--	--	--
Builtdens SF_1mi	Pixel density per square mile within 1 mile of all single-family pixels built as of 2010.	276.757	335.7188	0	2170.5
Chgdens_MF_2mi	Pixel density per square mile within 2 miles of all multi- family pixels built between 1990 and 2010.	22.61893	37.74299	0	468.38

Chgdens_com_3mi	Pixel density per square mile within 3 miles of all commercial pixels built between 1990 and 2010.	29.696	41.116	0	405.82
Chgdens_2mi	Pixel density per square mile within 2 miles of all pixels built between 1990 and 2010.	253.630	236.559	0	1416.3
Disemp	Distance to an "employment cluster" zip code in 2010.	27451.25	17072.42	0	66798
Empdens_3mi	Density of all jobs located within 3 miles in 2010.	266.607	791.504	0	14033
Elevation	Elevation in feet above mean sea level.	754.799	193.668	340.01	1357.8

Notes: 1 pixel = 30 m * 30 m (approx. 0.25 acres). Only 3.7% of all pixels were used, due to excessively large data set size being transferred to the research team, for this very large metroplex region. These are uniformly sampled in space, so their centroids are 150 meters apart in a grid across the region, to give wide dispersion in land use settings.

Experiments

Each of the four data sets were randomly split into two sets (70% for training and 30% for testing) ten times, with results averaged over the 10 different estimation runs for each of the ML methods deemed most competitive for the different response variables. Python codes from the scikit-learn were used in this paper to run the ML methods, lightGBM and catboost packages were installed in Python, and Stata was used to run the MLE model.

RESULTS

Tables 2 and 3 summarize model results in terms of computing time, predictive accuracy (using R^2 values), and RMSE or recall and F_1scores (depending on type of response variable being estimated). Carrion-Flores' (2009) multinomial logit model with spatial auto-regression (MNL-SAR) specification was used to predict the land use changes across much of the DFW region (11 counties) between 2010 and 2015. For the regression models, The GBDT and RF achieve a better prediction accuracy, while RF takes longer running time than the GBDT. In the classification models, the catboost, GBDT, XGboost and lightGBM achieve better prediction accuracy, while the running time of lightGBM being the most time-efficient.

TABLE 2 Regression Model Comparisons based on Computing Time, Predictive Accuracy, R^2 and RMSE (for HH VMT& HH Vehicle Ownership)

Method Used	Run Time (seconds)	Training Accuracy (R^2)	Test Data Set Accuracy (R^2)	RMSE
Annual HHVMT Prediction Results using OLS vs. ML Methods				
OLS	0.14	0.479	0.468	15754.88
Decision Tree (DT)	0.07	0.526	0.431	14534.29
Random Forest (RF)	49.49	0.594	0.497	**11433.93**
AdaBoost	17.48	0.382	0.323	17318.64
GBDT	3.56	0.642	0.542	**9589.14**
XGBoost	0.98	0.569	0.488	**11867.06**
lightGBM	4.49	0.512	0.452	**12885**
NN	9.56	-0.0123	-0.0359	21567.33
SVM	5.48	-0.06603	-0.0666	22043.15
HH Vehicle Ownership Prediction Results using Negative Binomial vs. ML Methods				
Poisson regression	0.169	0.4526	0.4319	0.779
Decision Tree (DT)	0.03	0.5372	0.5029	**0.699**
Random Forest (RF)	18.28	0.5786	0.5251	**0.639**
AdaBoost	1.82	0.5138	0.4804	0.729
GBDT	0.61	0.5855	0.5246	**0.645**

XGBoost	0.68	0.5618	0.492	**0.680**
LightGBM	0.26	0.539	0.4924	**0.683**
SVM	0.56	-0.2536	-0.1322	1.121
NN	2.52	-0.7005	-0.6547	1.131

TABLE 3 Computing Time, Training Accuracy, Precision, Recall and F_{1score} Values for Mode Choice & Land Use Change Prediction

Method Used	Time (seconds)	Training accuracy	Precision	Recall	F_1score
Mode Choice Prediction Results by MNL and ML					
MNL	0.83	0.658	0.63	0.64	0.63
Decision tree	0.09	0.976	0.98	0.98	0.98
Random Forest (RF)	13.24	0.989	0.98	0.98	0.98
AdaBoost	0.24	0.877	0.81	0.88	0.84
XGBoost	6.15	0.989	0.99	0.99	**0.99**
GBDT	7.00	0.994	0.99	0.99	**0.99**
NN	0.71	0.556	0.65	0.56	0.60
NBayes	0.06	0.856	0.85	0.86	0.84
catboost	60.63	0.993	0.99	0.99	**0.99**
lightGBM	2.20	0.997	0.99	0.99	**0.99**
Land Use Change Prediction Results by MNL SAR and ML					
MNL SAR	8.93	0.643	0.64	0.64	0.64
Decision Tree (DT)	1.04	0.714	0.67	0.70	0.68
Random Forest (RF)	30.97	0.744	0.72	0.71	**0.71**
AdaBoost	29.24	0.612	0.53	0.61	0.57
Extra trees	49.27	0.65	0.66	0.65	0.65
XGBoost	143.82	0.76	0.74	0.73	**0.73**
GBDT	208.82	0.769	0.73	0.74	**0.73**
NN	256.56	0.626	0.55	0.63	0.59
NBayes	0.32	0.348	0.47	0.34	0.39
catboost	161	0.743	0.74	0.74	**0.74**
lightGBM	19.34	0.733	0.73	0.72	**0.72**

Notes: Extra trees is a variation of RF model, it is tested and shows prediction accuracy worse than RF.

In terms of **HHVMT** prediction, the GBDT proved most effective relative to the many other models examines, achieving highest prediction accuracy in just 3.56 seconds, followed by the random forest technique, which took 49.5 sec, (almost 14 times longer), but is considered a more transparent model (with interpretable branches across all input variables used)

In terms of HH **vehicle ownership** prediction, the GBDT and RF methods outperformed the other models (with the RF requiring 30 times more computing time than the GBDT). RF's performance was followed by a simple, single DT.

In terms of **mode choice** prediction, the DT and ensemble methods beat the MNL approach, while catboost required much longer running times than XGBoost, GBDT and lightGBM.

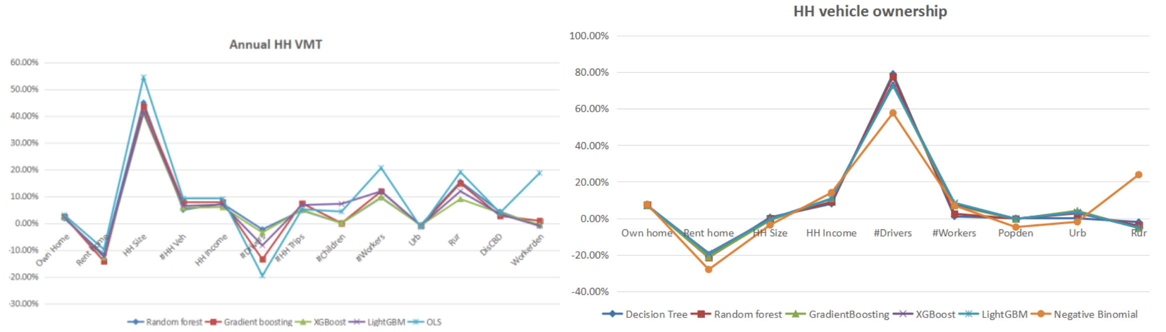
For **land use change** predictions (over the 20-year data period), the catboost method achieved the highest F_1score (of 0.74), which is only slightly higher than that of the XGBoost (0.73), GBDT (0.73), lightGBM (0.72) and RF (0.71) methods. This required only slightly (11%) more run time than XGBoost required, 8.32 times more than lightGBM, and 5.2 more than RF, but 77% less computing time than GBDT. So lightGBM may be best used in predicting land use changes, if run times are a concern (which will depend on application needs and data set size).

Estimates of Practical Significance across Competing Models

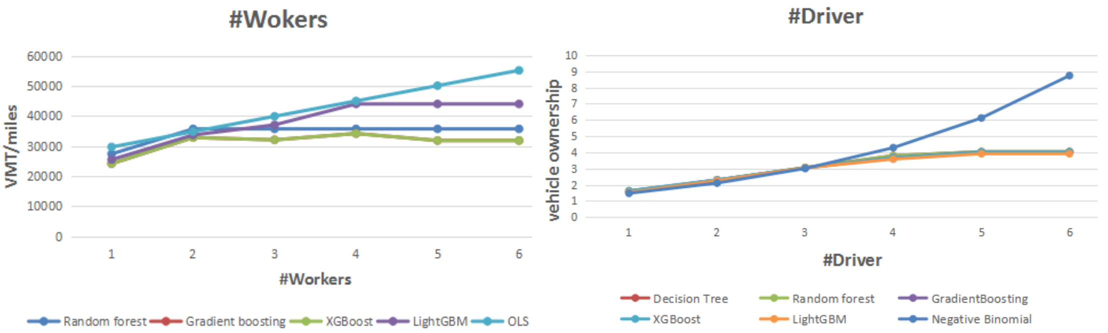
Using Eq. 4, estimates of the practical significance of input variables for each of the four output variables can be calculated. Here, only the top ML models' estimates are compared to those from the traditional econometric methods used, and just for a few of the input variables, to illustrate differences, while respecting word-count constraints.

Practical significance is imputed by increasing each input value for each data point in each data set. For example, inputs of HHsize and #Workers were varied from 1 to 6 to appreciate each model's estimate of how this would impact a household's expected VMT.

The prediction results comparison for the four models and practical significance analysis results for some important variables are shown in Figure 1 through 3.



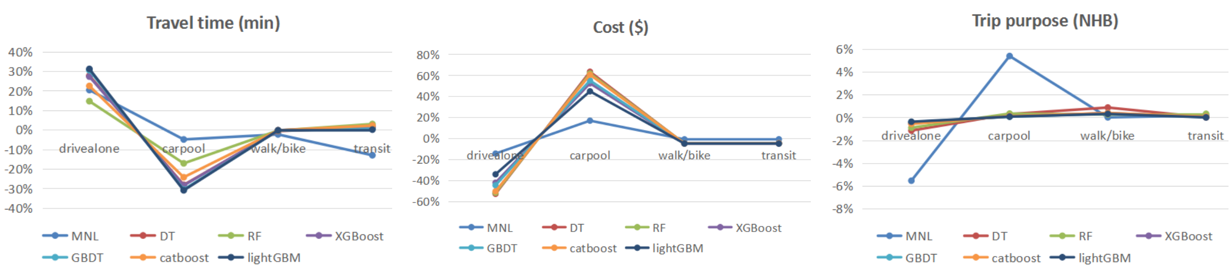
(a) Influence of Variables on VMT (b) Influence of Variables on HH Vehicle Ownership

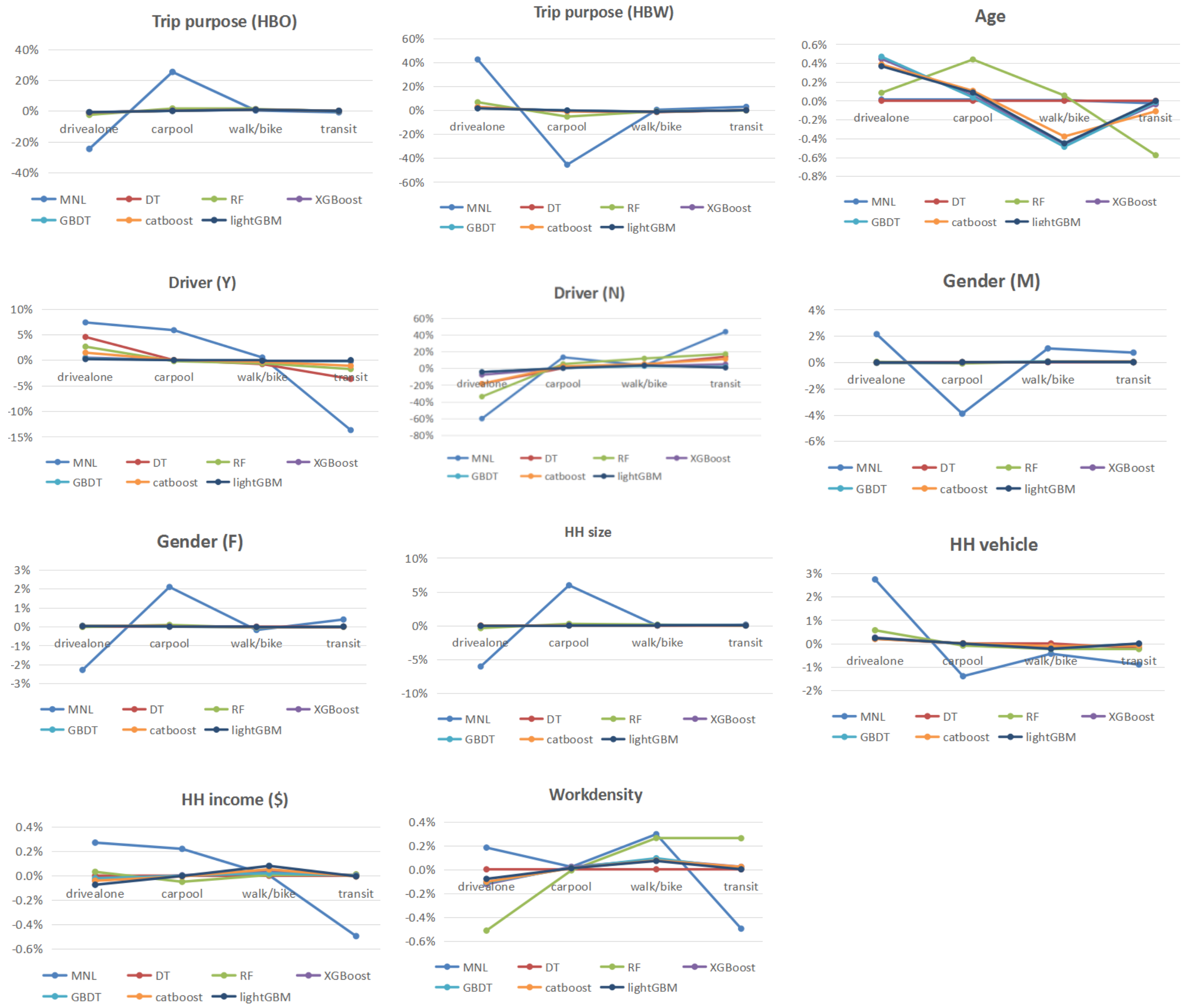


(c) Influence of Number of Workers Increase on VMT

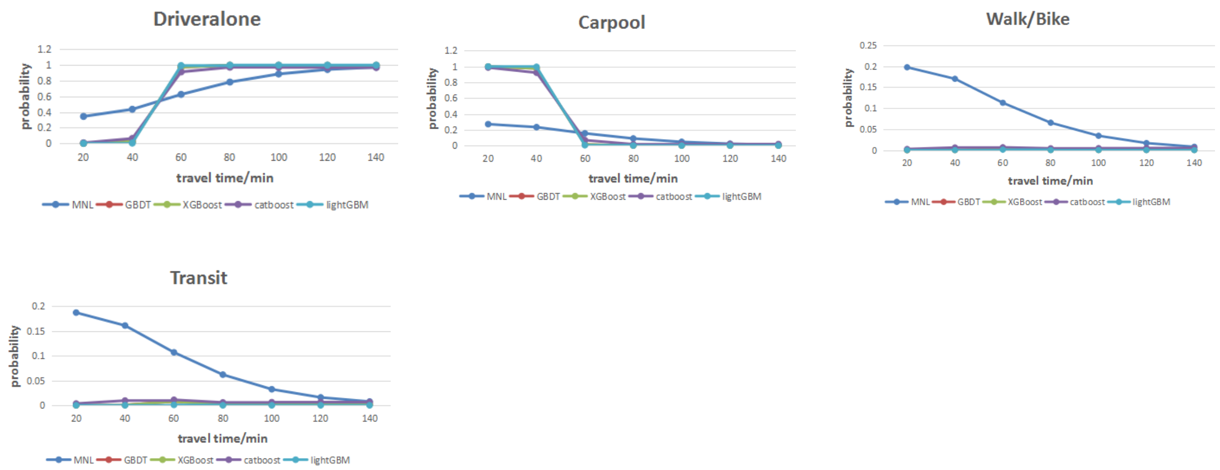
(d) Influence of Number of Drivers Increase on HH Vehicle Ownership

Figure 1 VMT & HH Vehicle Ownership Changes: Comparisons of Practical Significance Estimates



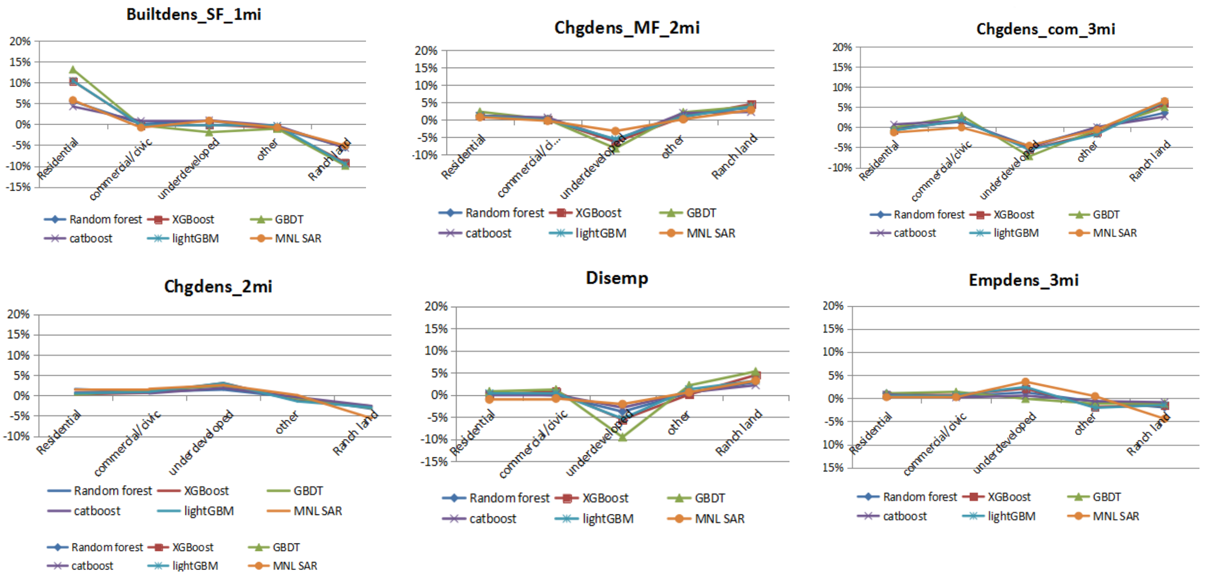


(a) Models Comparison of Influence of Variables on Chosen Probabilities of Each Mode

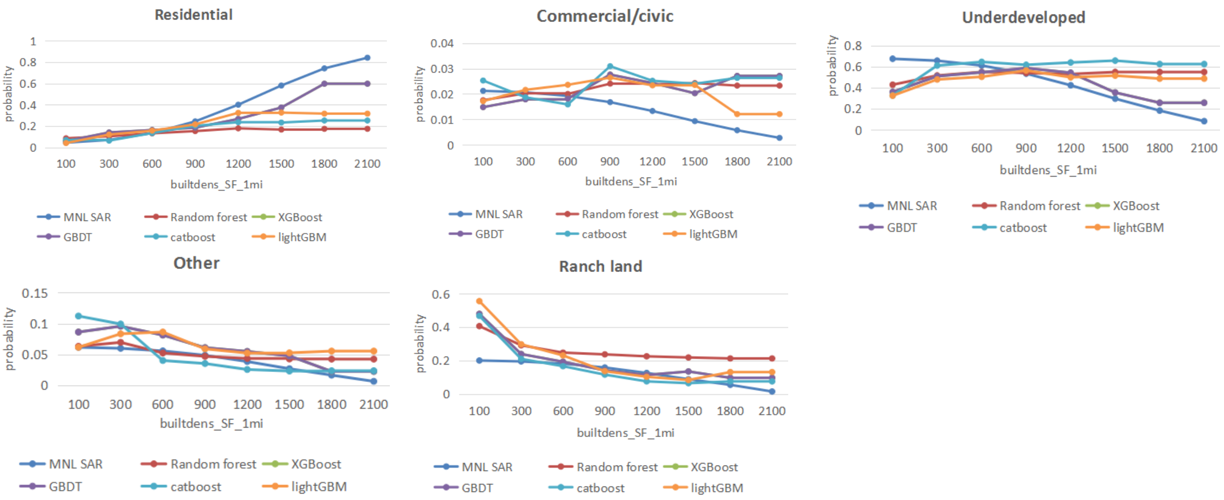


(b) Influence of Travel Time Increase on Chosen Probabilities of Each Mode

Figure 2 Mode Choice Changes: Comparisons of Practical Significance Estimates



(a) Models Comparison of Influence of Variables on Chosen Probability of Each Land Use Type



(b) Influence of Builtdensity of Single Family within 1mi on Land Use Change Probabilities

Figure 3 Land Use Change: Comparisons of Practical Significance Estimates

Annual HHVMT

As seen in Figure 1 (a), HHSize has a significant impact (around 40%) on HHVMT. The results predicted by the OLS is 1.11 times more than ML models, the practical significance of the number of HH workers is around 2 times more than results of ML methods, and the results of worker density show a huge difference between the OLS and ML methods, while the ML methods share very similar results.

As seen in Figure 1 (c), the plot of OLS shows a continuously increase trend, while the ML methods deliver a flattening of response with increase in input values, which is more realistic. Of course, one can add such behaviors to an OLS model, but guessing at additional input variables, but the ML method takes care of this automatically, which is valuable.

HH Vehicle Ownership

As shown in Figure 1 (b), the number of drivers is the most practically significant input used here, followed by home ownership type (rent vs. own). ML and traditional methods are found to deliver very similar estimates of impact for each covariate, apart from the indicator variable "rural", which is estimated to be 6 times more impactful when using the NegBin model than when using any of the ML methods. As shown in Figure 1 (d), with the increase of numbers of driver in a household, the plot about household vehicle ownership of NegBin model shows a continuously increase trend, while the ML methods deliver a flattening of response with increase in input values, which is more realistic.

Travel Modes

As shown in Figure 2 (a), travel cost and travel time have very sizable impacts on mode choice, in contrast to the effects of trip purpose, traveler gender and HH size. Since the accuracy of the DF, RF, XGBoost, GBDT, and lightGBM models are notably higher than those of the MNL models, the practical significance of the MNL results differ significantly from those generated using ML methods. For example, the Age and Worker density variables show that the ML models prediction results are similar with each other, while the results of the MNL is extremely different from the ML prediction results.

The MNL model suggests rather linear relationships between travel time increases and mode choice probabilities. In contrast, the ML methods' predictions of drive alone and carpool models delivered a sharp decrease or increase, followed by a flattening or plateauing, with little effect on the other two modes' probability predictions (which stayed around 0%). It is useful to have models that allow for non-monotonic effects, which the ML methods are delivering here, thanks to their behavioral flexibility – without as much mis-prediction as a traditional MNL model can produce. While the ML methods do not offer a random-utility-maximization (RUM) basis for behavior, such functional flexibility is generally realistic and useful to have.

Land Use Changes

The sensitivity tests show how local (1-mile) density of single-family land development has a strong and positive impact on land use change probabilities toward residential land use (as expected), and a negative effect on a shift toward ranchland use. A 2-mile (radius) density of multi-family use built between 1990 and 2010 has a strong negative influence on land use change toward "underdeveloped" status (as per the NCTCOG's definition of uses). Commercial land use density within 3 miles has a positive influence on land use change toward commercial activities. And distance to an "employment cluster" zip code (in the year 2010 data) is found to have a positive influence on ranch land use change.

As figure 3 (b) shows, with the increase of build density of single family within 1 mi, the plot of MNL SAR shows an increasing trend, the probability of the pixels' land use change to residual will be 1 and 0 for other land use type, while the ML methods deliver a flattened or saturated response, as this input rises by 1 SD at a time, the probability of the pixel' land use change to residual will be about 0.3. Considering the land use aggregation effect, if around of the 30m by 30m pixel are all residential, this pixels' land use type are more likely to change to residential, so, in the land use change type prediction, the MNL SAR prediction results are more realistic. For the ML, the results of GBDT are more similar to the MNL SAR results, so, the GBDT method not only has a better prediction accuracy but also predict land use change more realistically.

Compared to the traditional methods (OLS & MLE) which have own function to show practical significant among X variables and response variable, and add random error terms into their prediction models to consider the uncertainty of Y, learning from the results and analysis above, we can see that some ML methods can always achieve better accuracy no matter for classification or regression problems, and can show the practical significant among X variables and response variable in a more realistic way. By using the ML methods, the prediction model between the X variables and response variable can be obtained after training the data sets, then, new X variables (average of the X variables or increase 1 SD for some variables) can be simulated to seek the quantitative relationship between the X variables and response variable. While it is not wise to only give one predicted Y not consider the uncertainty of the predicted Y, one way to solve this problem is constructing many prediction models, input same X variables into these model and then get the interval of predicted Y.

CONCLUSION

This work explains and applies top ML methods for estimating distinct transportation variables of major interest to planners, policymakers, and the public at large. For example, household vehicle ownership, mode choice and VMT have important consequences for traffic congestion, crash counts, air pollution and energy consumption. Accurate prediction of such variables and understanding of their dependencies on household demographics, land use patterns, transport supply, and policy variables is very valuable. ML makes available a new suite of tools that transportation data analysts can exploit, but we must first understand that weaknesses and strengths, and "open up" the black boxes of the past to appreciate how our choice and attributes impact our final response values. This work demonstrates such applications for a variety of DFW data sets.

Based on testing conducted here, using the HHVMT continuous response and vehicle ownership count variables, the GBDT and RF models always performed better than the other ML models, with the RF requiring 13.9 more computing than the GBDT in HHVMT prediction and 30 times more than GBDT in vehicle ownership prediction. The XGBoost, GBDT, catboost and lightGBM all achieved better classification results – for mode choices and land use changes, with the catboost method required the most time for mode choice prediction and lightGBM requiring the least. In land-use change prediction, GBDT required more time. Overall, GBDT was found here to be the best model for the continuous and count response values, while lightGBM is preferred for categorical response prediction.

From the practical significance analysis, one finds that methods with similar predictive accuracy deliver similar estimates of each input's practical significance. For the continuous and count-based response variables (HH VMT and vehicle ownership), output accuracy and input significance are very similar between the traditional and ML methods. In the case of categorical mode choice, the MNL model's F_1score was not so competitive, so its estimates of input impact differed from the best-performing methods. In the case of land use change prediction, the ML and SAR-MNL models has similar F_1score values, and so delivered similar estimates of practical significance across inputs, while the results of SAR-MNL and GBDT are more realistic. But there is less expectation of monotonicity of effect in the ML methods, so an input's impact can change course or at least flatten, which is valuable in many example cases, as demonstrated in predicting VMT, mode choice and land use change, based on certain input variables.

Since ML methods generally outperform traditional statistically-based prediction methods and transportation projects and policies regularly have multi-million-dollar impacts, it is important to at least test such methods to ensure decisions are consistent across modeling assumptions. Planners and modelers have a duty to use limited public resources optimally. Related to this, most policymakers are not concerned about the interpretability of the tools used to reach such conclusions, but many of the best ones are. So it is valuable to try and unpack ML equations, or at least document how outputs vary with respect to each input, for the average observational unit or a sample of such units. Finally, ML methods also remain lacking in terms of conveying uncertainty in predicted output values. For example, is the ML-estimated VMT per year for the average Dallas household 25,320 miles +/- 5200 miles or +/- 8400 miles? And what is the probability that this household's oldest adult drives to work rather than bikes to work tomorrow? What software does the data analyst have access to? What programming languages is he/she comfortable with? And how important is immediate transparency in results for application of the model? Or having a behavioral foundation, like random utility maximization for mode choice? These distinctions matter, so important investigative and application work remains for ML users.

ACKNOWLEDGEMENTS

The authors are grateful for provision of Dallas-Ft Worth data sets by Kathy Yu and Mark Folden at the North Central Texas Council of Governments, Albert Coleman's editing support, and financial support of the China Scholarship Council, which funded the lead author's one-year stay in the U.S.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Li, W. and Kockelman, K.; data collection: Li, W. and Kockelman, K.; analysis and interpretation of results: Li, W. and Kockelman, K.; draft manuscript preparation: Li, W. and Kockelman, K. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

Alfaro, E., García, N., Gámez, M., & Elizondo, D. Bankruptcy Forecasting: An Empirical Comparison of AdaBoost and Neural Networks. *Decision Support Systems*, 2008. 45(1): 110-122.

Breiman, L. Bagging Predictors. *Machine Learning*, 1996. 24(2): 123-140.

Brown, Garikapati, Hou. Machine Learning Decision Support Tool for Travel Demand Modeling. *Presented at the 98th Annual Meeting of the Transportation Research Board, Washington, D.C.*, 2019. (19-04806).

Carrión-Flores, C. E., Flores-Lagunes, A., & Guci, L. Land Use Change: a Spatial Multinomial Choice Analysis. *AAEA & ACCI Joint Annual Meeting*, Milwaukee, Wisconsin., 2009. (No. 319-2016-9951).

Caruana, R., Niculescu-Mizil, A. An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learnin*g, 2006. 161-168.

Chen C, Zhang G, Liu X C, et al. Driver Injury Severity Outcome Analysis in Rural Interstate Highway Crashes: A Two-level Bayesian Logistic Regression Interpretation. *Accident Analysis & Prevention*, 2016. 97: 69-78.

Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. Applying a Random Forest Method Approach to Model Travel Mode Choice Behavior. *Travel Behaviour and Society*, 2019. 14: 1-10.

Cui, Z.Y, Henrickson K., Ke R.M. High-Order Graph Convolutional Recurrent Neural Network: A Deep Learning Framework for Network-Scale Traffic Learning and Forecasting. *Presented at the 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019. (No. 19-05236).

Das S, Dutta A, Datta S. Injury Severity Analysis from Crash Narratives: A Case Study of Interpretable Machine Learning using Tree and Utility Pole related Traffic Crashes.

Presented at the 98th Annual Meeting of the Transportation Research Board, Washington, D.C., 2019. (No. 19-03241).

Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. Investigating Injury Severity Risk Factors in Automobile Crashes with Predictive Analytics and Sensitivity Analysis Methods. *Journal of Transport & Health*, 2017. 4: 118-131.

Deng, Z., & Ji, M. Deriving Rules For Trip Purpose Identification From GPS Travel Survey Data and Land Use Data: A machine Learning Approach. *Traffic and Transportation Studies*, 2010, 768-777.

Ding, C., Wang, D., Ma, X., & Li, H. Predicting Short-term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. *Sustainability*, 2016. 8(11): 1100.

Dorogush A V, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363, 2018.

Folden M., High Resolution Demographic Forecasting: The Convergence of Socioeconomic and Remote Sensing Data for Small Area Forecasting. *Applying Census Data for Transportation: 50 Years of Transportation Planning Data Progress*, 2018.

Freund, Y., & Schapire, R. E. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997. 55(1): 119-139.

Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 2001. 29 (5): 1189-1232.

Goh, Y. M., Ubeynarayana, C. U., Wong, K. L. X., & Guo, B. H. Factors Influencing Unsafe Behaviors: A Supervised Learning Approach. *Accident Analysis & Prevention*, 2018. 118: 77-85.

Hagenauer, J., & Helbich, M. A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice. *Expert Systems with Applications*, 2017. 78: 273-282.

Hamad K, Khalil M.A, Alozi A.R. Developing Machine Learning Model Learning Models to Predcit Incident Duration. *Presented at the 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019. (No. 19-02367).

Harb, R., Yan, X., Radwan, E., & Su, X. Exploring Precrash Maneuvers Using Classification Trees and Random Forests. *Accident Analysis & Prevention*, 2009. 41(1): 98-107.

Hosseini M.K, Talebpour A. Traffic Prediction Using Time-Space Diagram: A Convolutional Neural Network Approach. *Presented at the 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019. (No.19-05980).

Hou, Y., Young, S. E., Dimri, A., & Cohn, N. Network Scale Ubiquitous Volume Estimation Using Tree-Based Ensemble Learning Methods. *Transportation Research Board 97th Annual Meeting, Washington, D.C., 2018.*

Huang, Y., & Kockelman, K. M. What will autonomous trucking do to US trade flows? Application of the random-utility-based multi - regional input – output model. *Transportation*, 2018. 1-28.

Iranitalab, A., & Khattak, A. Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction. *Accident Analysis & Prevention*, 2017. 108: 27-36.

Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 2017. 3146-3154.

Keya, N., Anowar, S., & Eluru, N. Joint Model of Freight Mode Choice and Shipment Size: A Copula-Based Random Regret Minimization Framework. *Transportation Research Part E: Logistics and Transportation Review*, 2019. 125: 97-115.

Krishnaveni, S., & Hemalatha, M. A Perspective Analysis of Traffic Accident Using Data Mining Techniques. *International Journal of Computer Applications*, 2011. 23(7): 40-48.

Liang Y.Z., Wang Y.H. Data-Driven Road Side Unit location for Connected - Autonomous Vehicle -based Intersection Control. *Presented at the 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019. (No.19-01496).

Miao, X., & Heaton, J. S. A Comparison of Random Forest and Adaboost Tree in Ecosystem Classification in East Mojave Desert. *2010 18th International Conference on Geoinformatics,* 2010. 1-6.

Omrani, H. Predicting Travel Mode of Individuals by Machine Learning. *Transportation Research Procedia*, 2015. 10: 840-849.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems*, 2018. 6639-6649.

Pu, Z., Liu, C., Wang, Y., Shi, X., & Zhang, C. Road Surface Condition Prediction using Long Short-Term Memory Neural Network based on Historical Data. *Presented at the 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019. (No. 19-03118).

Wang, B., Wang, Y., Qin, K. et al. Detecting Transportation Modes Based on LightGBM Classifier from GPS Trajectory Data. *2018 26th International Conference on Geoinformatics*, 2018. 1-7.

Wang, S.H., Zhao, J.H. An Empirical Study of Using Deep Neural Network to Analyze Travel Mode Choice with Interpretable Economic Information. *Presented at the 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019. (No.19-04955).

Wang, Y., Han, P., Lu, X., Wu, R., & Huang, J. The Performance Comparison of Adaboost and SVM Applied to SAR ATR. *2006 CIE International Conference on Radar*, 2006. 1-4.

Xiao, G., Juan, Z., & Zhang, C. Travel Mode Detection Based on GPS Track Data and Bayesian Networks. Computers, *Environment and Urban Systems*, 2015. 54: 14-22.

Xie, C., Lu, J., & Parkany, E. Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2003. (1854): 50-61.

Xu, Z., Kang, J. E., & Chen, R. A Random Utility Based Estimation Framework For the Household Activity Pattern Problem. *Transportation Research Part A: Policy and Practice*, 2018. 114: 321-337.

Zhang, Y., & Haghani, A. A Gradient Boosting Method to Improve Travel Time Prediction. *Transportation Research Part C: Emerging Technologies*, 2015. 58: 308-324.