Teaching Freight Mode Choice Models New Tricks Using Interpretable Machine Learning Methods

Xiaodan Xu[1], Hung-Chia Yang[1], Kyungsoo Jeong[2], William Bui[3], Srinath Ravulaparthy[4], Haitam Laarabi[1], Zachary Needell[1], C. Anna Spurlock[1]

[1] Lawrence Berkeley National Laboratory, 1 [2] National Renewable Energy Laboratory, [3] University of California, Berkeley, [4] University of California Santa Barbara

Research Question and Objectives

Why combine machine-learning (ML) and discrete choice models?

- Conventional discrete choice models:
  - Theory-driven and provide clear subject-matter interpretations.
  - Widely used in understanding the travel behavior of passenger and freight and support policy making.
  - Lacking efficient and systematic way to identify non-linear and interactive effects.
- Machine-learning (ML) methods:
  - Often provide better out-of-sample accuracy, but hard to extrapolate.
  - Often capture complex and non-linear relationship among the data.
  - Recently become interpretable and transparent applying SHapley Additive exPlanations (SHAP).

Research goal:

- Develop a multinomial logit (MNL) model for freight mode choice using the insights from ML models.
- Showcase how interpretable ML methods help enhance the performance of MNL models and deepen our understanding of freight mode choice.

Proposed Workflow

Develop freight mode choice models for Austin

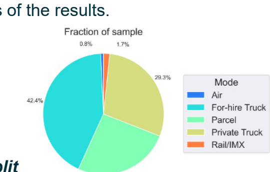
- Using 2017 Commodity Flow Survey (CFS) data (sample size = 247,073).
- For-hire truck (base), private truck, air, parcel, and rail + intermodal truck/rail (rail/IMX).

Compare the performance of two approaches:

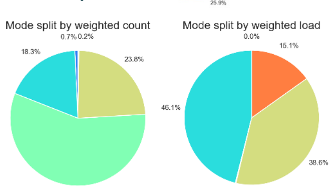
- A conventional logit model approach
  - Baseline MNL models (**'bMNL'**) with mostly linear specifications.
- A machine-learning (**ML**) guided approach
  - Advances MNL model ('**aMNL**') using ML and SHAP interpretations.

Investigate the results in two aspects:

- Accuracy measures of predicted mode choice.
- Interpretations of the results.



Austin mode split



Interpretable Machine-learning Results

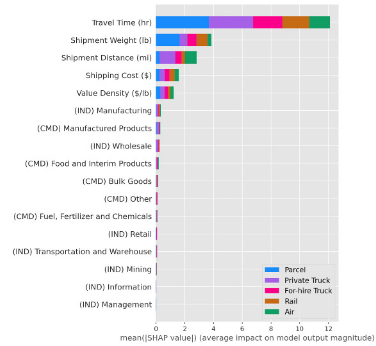
Machine-learning models overview:

- Select ML methods that are
  - Suitable for resolving nonlinear relationships in mode choice models.
  - Seamless connection with SHAP TreeExplainer.

Selected methods:

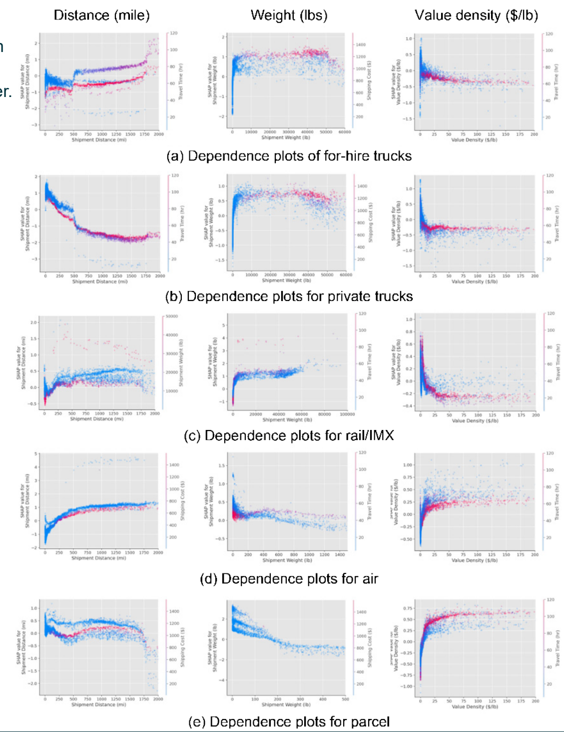
- **Random forest (RF):** builds a large collection of de-correlated trees and then averages them.
- **Boosting Trees:** combines the outputs of many "weak" classifiers to produce a powerful "committee".
  - **XGBoost:** a scalable ML system for tree boosting.
  - **CatBoost:** specialize in categorical data.

SHAP feature importance (CatBoost)



**CMD – Commodity; IND - Industry*

SHAP dependence plots (CatBoost)



(a) Dependence plots of for-hire trucks

(b) Dependence plots for private trucks

(c) Dependence plots for rail/IMX

(d) Dependence plots for air

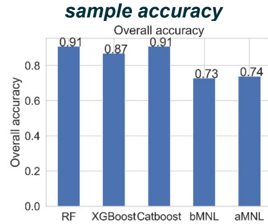
(e) Dependence plots for parcel

Performance Comparison

Out-of-sample accuracy of mode choice models:

- RF and CatBoost have the highest accuracy, followed by XGBoost.
- Tree-based MLs outperform the MNL models.
- aMNL model has higher accuracy than bMNL.

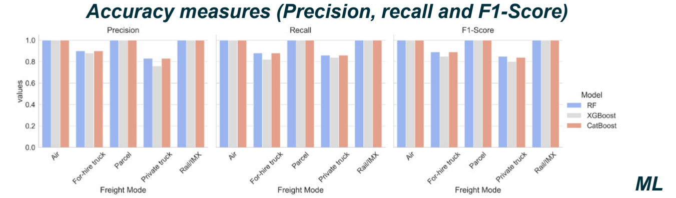
Overall out-of-sample accuracy

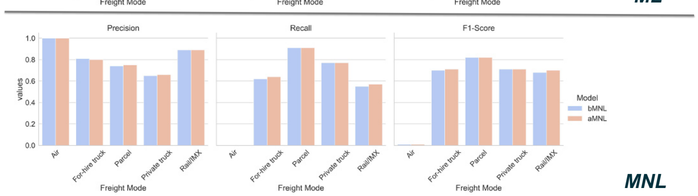


Performance measures (Precision, Recall, F-1 Scores) by mode:

- ML generate accurate predictions for all modes, while the accuracy of the two truck modes are slightly lower.
- MNL models have larger errors for air and rail/IMX, potentially due to low sample size.

Accuracy measures (Precision, recall and F1-Score)





Baseline Multinomial Logit Model

Highlights of bMNL model estimation:

- Results from MNL do not capture the intricate relationship demonstrated in SHAP.
- Low-impact factors (e.g., some industries) may absorb the effects from more influential factors.

Variables	Mode (for-hire truck as the base) Air	Parcel	Private Truck	Rail/IMX
Constant	-5.05***	0.472***	1.395***	-5.49***
Distance (mile)	0.002***	0.001***	-0.005***	-6.1e-5
Value density ($/lb.)	1.4e-5***		-0.001***	
Weight between 150 and 1,500 lbs.	-3.211***			1.798***
Weight between 1,500 and 30,000 lbs.	-3.784***		0.044**	3.352***
Weight between 30,000 and 45,000 lbs.			-0.67***	2.627***
Weight greater than 45,000 lbs.			-1.207***	4.296***
Commodity is bulk		-1.107***	-0.78***	-2.094***
Commodity is fuel, fertilizer or other chemical		-0.77***	-0.338***	-1.296***
Commodity is interim product or food	-0.98***	-1.312***		-3.077***
Commodity is manufactured goods	0.882***		-0.912***	-1.458***
Information industry		0.126**	-0.968***	
Manufacturing industry		0.327***	-0.319***	-0.558***
Management industry		0.325***	0.202*	-2.185***
Retail industry	0.558***	2.459***	1.08***	
Transport and Warehouse industry				-2.311***
Wholesale industry			0.493***	-1.403***
Shipping Costs	-0.001***			
Shipping Time	-0.003***			
Number of parameters	47			
Number of observations	247,073			
Log-likelihood	-157,515			
Adjusted ρ^2	0.567			

*p<0.1, ** p<0.01, *** p<0.001

Advanced Multinomial Logit Model

Highlights of new findings in aMNL model:

- SHAP results help remove nine low-impact factors.
- Binned specifications of distance and value density help reveal nonlinear relationships of mode preferences.

Variables	Mode (for-hire truck as the base) Red cell highlights removed variables in aMNL Air	Parcel	Private Truck	Rail/IMX
Constant	-5.258***	0.237***	1.405***	-6.366***
Distance*(Distance <= 500 miles)	0.004***	0.004***	-0.005***	0.001***
Distance*(Distance > 500 miles)	0.002***	0.001***		
(Distance > 500 miles)		2.220***		0.321*
Value density*(Value density <= $5/lb.)	-0.114*	0.012	0.009	
(Value density > $5/lb.)			-0.301***	
Value density*($5/lb.<Value density<= $25/lb.)	0.039***	0.025***		
(Value density>$25/lb.)	1.557***	0.372***		
Value density*(Value density <= $1/lb.)				-0.223*
Value density*($1/lb.<Value density<= $10/lb.)				0.124***
Weight*(Weight <= 150 lbs.)	-46.389***	-33.591***	2.815***	
Weight between 150 and 1,500 lbs.	-3.329***			
Weight between 1,500 and 30,000 lbs.	-3.619***			2.151***
Weight between 30,000 and 45,000 lbs.			-0.749***	1.606***
Weight greater than 45,000 lbs.			-1.281***	3.232***
Commodity is bulk			-0.732***	-1.273***
Commodity is fuel, fertilizer or other chemical		-0.329***	-0.248***	-0.843***
Commodity is interim product or food	-0.642**	-0.790***	0.144***	-2.681***
Commodity is manufactured goods	0.354***	0.089**	-0.847***	-1.049***
Information industry			-1.111***	
Manufacturing industry		0.155***	-0.375***	0.519***
Management industry				
Retail industry	-1.625***			
Transport and Warehouse industry				
Wholesale industry			0.469***	
Shipping Costs	-0.001***			
Shipping Time	-0.003***			
Number of parameters	51			
Number of observations	247,073			
Log-likelihood	-145,857			
Adjusted ρ^2	0.576			

Findings and Recommendations

- Using insights from SHAP, aMNL's accuracy surpass that of bMNL.
- The estimated aMNL reveals significant and complex relationships that are hidden in bMNL.
- The directions of impacts from aMNL and CatBoost are often aligned.
- Interpretable ML can be a useful tool to enhance the practice of freight behavior analysis and modeling.

The work described was sponsored by the U.S. Department of Energy (DOE) Vehicle Technologies Office (VTO) under the Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility Laboratory Consortium, an initiative of the Energy Efficient Mobility Systems (EEMS) Program. Lawrence Berkeley National Laboratory operates under DOE Contract No. DE-AC02-05CH11231. The National Renewable Energy Laboratory is operated for DOE by the Alliance for Sustainable Energy, LLC, under Contract No. DE-AC36-08GO28308. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. 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.

103rd Transportation Research Board (TRB) Annual Meeting
Washington DC
January 7-11, 2024
NREL/PO-5400-88324