Machine Learning for Prediction of Mid to Long Term Habitual Transportation Mode Use

Alina Lazar, Alexandra Ballow*
Dept. of Computer Science and Information Systems
**Dept. of Physics and Astronomy*
Youngstown State University
Youngstown, OH
alazar@ysu.edu, alballow@student.ysu.edu

Ling Jin, C. Anna Spurlock
Energy Analysis and Environmental Impacts Division
Lawrence Berkeley National Laboratory
Berkeley, CA
ljin, caspurlock@lbl.gov

Alexander Sim, Kesheng Wu
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA
asim, kwu@lbl.gov

***Abstract*—Prediction of daily transportation mode use (car, public transit, or active travel) is a important task in transportation research. Unlike statistical models that impose a predetermined model structure, machine learning models are learned from the data, making them more flexible with higher prediction accuracy. However, prediction of mid- to long-term habitual modes still largely relies on traditional statistical analysis using small samples of cross-sectional data. Low interpretability of "black-box" machine learning models limits their usefulness for generating behavior insights needed for designing appropriate interventions. This paper, leveraging a set of unique longitudinal life course data, is the first use case to demonstrate machine learning methods applied for both predicting and interpreting regularly used travel modes. We combine sequence clustering and tree-based machine learning methods coupled with TreeExplainer to predict and interpret habitual travel modes using mid- to long-term predictors. Five life course clusters are derived to provide evaluation and interpretation contexts. This allows us to improve upon a recently developed TreeExplainer method to better distinguish predictor importance locally and globally; and predictor interactions across subpopulations within distinctive life history contexts. Our results demonstrate a promising step toward interpretable machine learning applications to mid- to long-term prediction of travel modes for transportation planning.**

***Keywords*-life course, life events, habitual transportation mode, sequence analysis, machine learning, gradient boosting, treeExplainer**

I. Introduction

The mode of transportation we choose can have a significant energy and environmental implications. Traditional mode choice prediction studies have been focused on trip-based analysis with trip specific attributes (such as trip cost, distance, time, terrain) being the major factors that shaped the choice preferences [1]. However, it has been suggested that regularly used modes are largely habitual and only change upon major life events, such as attending school, getting employed, getting married, and having a child [2], [3]. Methodologies for predicting these regularly used travel modes are instrumental to longer term transportation planning and creating sustainable transportation systems.

Unlike statistical models that impose a predetermined model structure, machine learning models are learned from the data, making them more flexible with higher prediction accuracy. Applications of machine learning methods to mode choice analysis have so far been limited to short-term trip data [4], while understanding mid- to long-term habitual or regularly used modes still largely relies on traditional statistical analysis with small sample cross-sectional data [3].

Predicting regularly used modes with machine learning methods needs longer term data that are difficult to collect from the life courses of individual travelers. More importantly, the models learned from the data and their performances need to be interpretable to generate travel behavior insights. This is particularly challenging for life course data because feature importance and their interactions should be understood within a dynamically changing life context rather than a simple global feature importance ranking [5].

This paper uses mid- to long term features such as life cycle stages, residence, and car ownership to predict regularly used transportation modes: driving ("used own"), public transit ("used public"), and active travel ("bike/walk"). By leveraging a unique life course data set collected by a recent transportation behavior study, this paper is the first to demonstrate machine learning methods applied to both predicting and interpreting regularly used transportation modes. We evaluate the performance of a suite of tree-based machine learning methods. To address the interpretability challenge, sequence clustering is used to derive five distinctive family and career trajectories which provide the life history contexts for TreeExplainer [5] to further probe both global and local feature importance and feature interactions. The rest of the paper is organized as follows: section II introduces the data; section III describes the machine learning methods and interpretation strategies; section IV present the results; and section V concludes.

II. Data Description and Preprocessing

We use data collected through the WholeTraveler Transportation Behavior Study [6], which is part of the U.S. Department of Energy's Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility Consortium.

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

The survey was administered in the nine core Bay Area California counties.

The WholeTraveler study implemented a life history calendar survey, which asked the respondents to recall the years when certain key life events occurred and other pertinent factors on an annual basis starting at age 20 and up to age 50. Such design allows for efficiently collection of longitudinal data in a single shot of survey. The input features used in training the machine learning models include yearly life cycle status: school, employment, living with a partner, having a child, household size ("hhsize"), as well as durable mobility-related decision variables such as number of cars ("numcars") and public transit availability at the residence location ("public_avail"). Birth year and gender, fixed across time, are also included.

We restrict the analysis to the 17,777 observations from the 569 of the respondents who were age 35 or older at the survey year (2018) capturing a life period that presents the greatest heterogeneity among the population [3].

III. Methods

A. Sequence Clustering

We use sequence clustering to construct different types of life course trajectories in the family (partner and child) and career (school and employment) dimensions. Given the categorical, longitudinal characteristics of the life trajectory sequences, we use the edit-distance based dissimilarity measure called optimal matching (OM) and follow the joint sequence analysis approach by [7] to compute dissimilarities between sequences describing trajectories of multiple life dimensions. The method used is implemented in the TraMineR package version 2.0-6. We conduct hyper-parameter tuning and performance evaluation following procedures detailed in [8], [9].

B. Tree-based Machine Learning Methods

1) Random Forest: This algorithm [10] builds an ensemble of decision trees, or tree predictors, which depend on randomly and independently sampled vectors over the same distribution. The strength, correlation and monitor error are closely followed to track the growing features in response to the branches splitting.

2) XGboost, Catboost, and LightGBM: Standard gradient boosting methods managed to solve over-fitting problems, but inefficiently. In an effort to make gradient tree boosting more flexible and scalable, Chen [11] created the scalable XGBoost algorithm. XGBoost employs a new regularization technique, instead of optimizing the loss function, to minimize the over-fitting. This tactic allows XGBoost to be faster and more robust during tuning. Two other boosting methods examined here were shown to have better performance on categorical data [12]. A slightly different method, CatBoost, focuses on categorical columns using permutation techniques and target-based statistics [13]. The light gradient boosting machine (LightGBM) further improves standard gradient boosting methods. Microsoft developed LightGBM by growing the decision trees leaf-wise, allowing it to support GPU learning speed, with faster training time, better accuracy, and for larger data [14].

C. Model Interpretation

The ensemble tree methods, such as XGboost or random forest, provide robust accuracy for classification and regression tasks using large sets of shallow trees. Explaining and interpreting the prediction made by these models is very important, but not trivial. The new method TreeExplainer developed by Lundberg et al. [5] is based on game theory and provides a fast, consistent and accurate method to determine feature importance.

The SHAP values, the core of this method, represent the sequential impact on the model's output of observing each input feature's value, averaged over all possible feature orderings. In addition to just ranking the features based of their contribution to the classification, the SHAP values can be used to plot for each feature individualized explanations for every data point in the dataset and how their values affect the final prediction (Figure 2). Even more revealing are the SHAP dependence plots (Figure 3). These plots capture the impact of one feature, age for example, on the classification task. The interaction of two features can be shown by coloring the individual data points based on the values of a different feature.

IV. Results and Discussions

A. Life course clusters

Ward's linkage hierarchical clustering yields a five cluster solution based on clustering quality metrics. The five clusters shown in Figure 1 summarize the dynamic patterns across age of the percentage of sample in the family (partner and child) and career (school and employment) dimensions. The distinction between clusters is mostly driven by the timing of partner and children. Based on their observable life trajectory patterns, we refer to the five clusters as "Singles," "Couples," "Have-it-alls," "Late Bloomers," and "Family First":

1) Singles: (40% of the sample) tend to finish school and enter the workforce early and delay or eschew having a partner or children.

2) Couples: (27%) tend to finish school, work, and partner up early but delay or eschew having children.

3) Have-it-alls: (18%) finish school and start to work early in life, and partner up and have children only slightly after.

4) Late Bloomers: (8%) generally delay school, work, partnering, and children until much later in life, if at all.

5) Family First: (7%) tend to partner up and have children early and delay school and/or work.

These life trajectory clusters serve as a contextual system to understand whether the XGBoost model performs equally across these sub-populations and how it uses input features to make predictions of various habitual transportation modes within specific life course contexts.

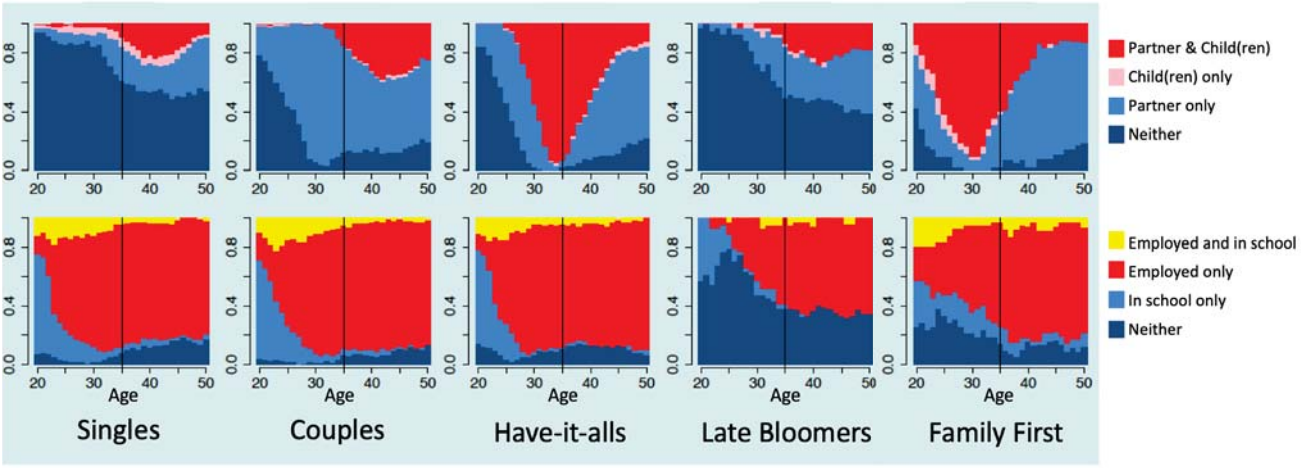

Fig. 1. Life course patterns of family and career status in five life trajectory clusters

B. Prediction Performance of tree-based learning methods

We employ a 10-fold cross validation to train the four tree-based models. As the purpose is to make temporal predictions of individuals' mode use, training and testing samples are split by the age variable. The performance metrics (accuracy, precision, recall, and F1 score) on the testing data set are shown in Table I. Model performances are consistent across the five life trajectory clusters (additional tables available upon request).

All four methods yield good performance for predicting the "used own" car mode. Random Forest and XGBoost perform similarly in predicting "use public" transit and "walk/bike", and both outperform CatBoost and LightGBM, especially in the Recall and F1 metrics. For the rest of the paper, we use XGBoost to further interpret the model predictions.

TABLE I
CLASSIFICATION PERFORMANCE ON TESTING DATA SETS

	Method	Acc.	Prec.	Recall	F1
used public	Random Forest	**0.8216**	0.6021	0.5217	**0.5590**
	XGBoost	0.8193	0.6045	0.4813	0.5359
	CatBoost	0.7761	0.4804	0.4048	0.4394
	LightGBM	0.8008	0.6274	0.1994	0.3026
walk/bike	Random Forest	0.8807	0.6066	0.3882	0.4735
	XGBoost	**0.8908**	0.6894	0.3812	**0.4909**
	CatBoost	0.8622	0.5031	0.1882	0.2740
	LightGBM	0.8700	0.8049	0.0776	0.1416
used own	Random Forest	**0.8391**	0.8495	0.9529	**0.8983**
	XGBoost	0.8258	0.8420	0.9433	0.8898
	CatBoost	0.7530	0.8479	0.8147	0.8310
	LightGBM	0.7949	0.7936	0.9795	0.8768

C. Interpretation of mode predictions

1) Feature importance: The local explanations (or SHAP values) of individual input features are computed for predicting each habitual mode (example for predicting walk/bike can be found in Figure 2 left). Positive SHAP values indicate a higher likelihood of using the mode and vice versa. The input features are ranked by their global importance (i.e., the mean absolute SHAP values) as shown in the global feature importance (Figure 2 right).

Birth-year, age, and number of cars owned are consistently among the top five globally important features for all modes and the remaining two features are mid- to long-term mobility or life-event-related inputs, which are different across the modes as shown in Figure 3. While the importance of life-event-related features are mostly similar to previous literature, especially for school and employment status, difference is seen in household size on predicting bike/walk across travelers with different family forming histories. Despite its overall low importance for the bike/walk mode, household size replaces school in the top five features (Figure 2 left insert) for the "Family-first" travelers who partner up and have children early and delay school and/or work. Current literature usually bases their variable selection on global feature importance ranking. Our result here suggests that such a practice may miss important features for certain sub-population and therefore bias predictions.

2) Life event interactions with age: Interpretation of input features is straightforward when their effects (as represented by the summary of SHAP values Fig 2 left) are monotonic with mode use. For example, owning more cars decreases walk/bike use, while attending school increases it. However, the interpretation is more challenging for features (such as age and familial events) whose effects are conditioned on multiple other variables. The generic variable interactions from TreeExplainer handle two features at a time. We use life history clusters to provide further conditional context and illustrate it with child and age interactions in Fig 4. Aggregating over the full sample, having children appears to significantly increase driving when survey respondents are in their 30s. However, such a pattern is not representative of the two clusters that had children relatively early in life. "Have-it-alls" increase their car usage much earlier than the average population. However, those categorized as "Family-first", who

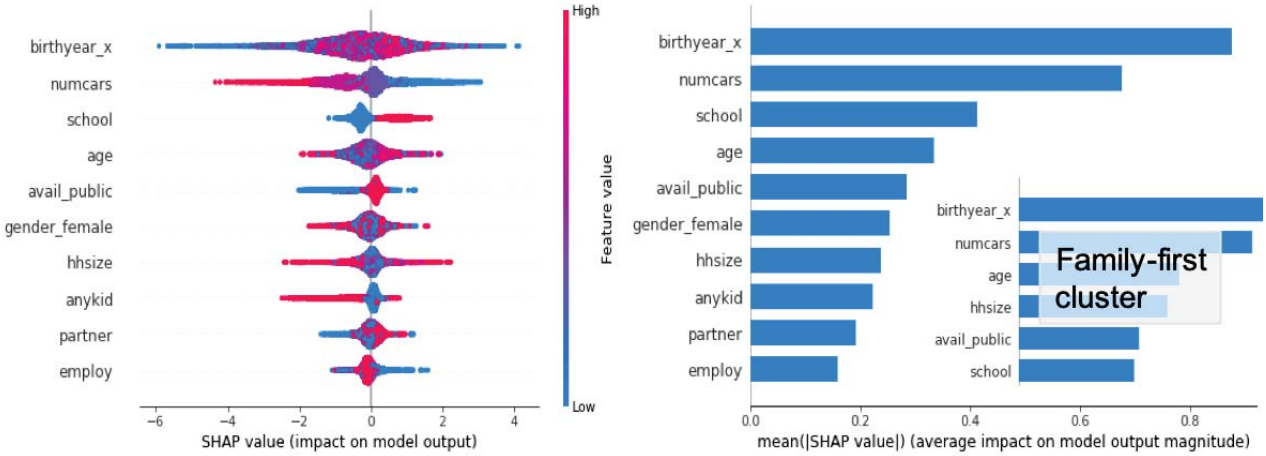

Fig. 2. Local explanation summary (left) and global feature importance (right) for predicting use of own car.

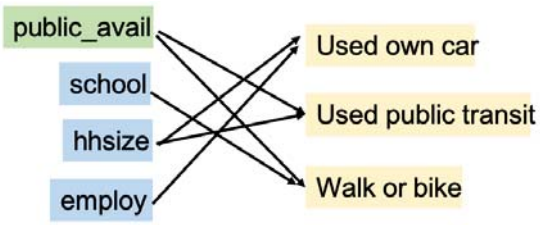

Fig. 3. Important input features vary by mode

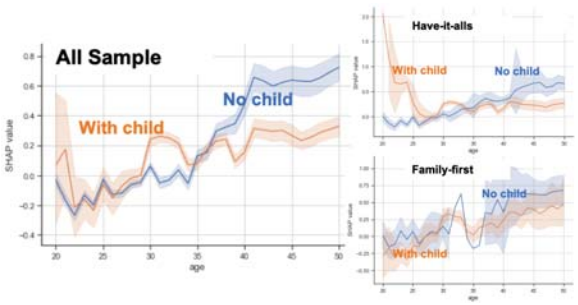

Fig. 4. Interaction of child with age on car use prediction derived from all samples (left) and within different life history contexts (right)

delay their education and work, do not increase car usage as much upon having a child, likely due to more limited resources and/or need for driving. These results suggest that the model adequately captures the complex interactions within different life histories and confirms the importance of timing in multiple life event dimensions.

V. Conclusion

This paper, leveraging a unique life course data set, is the first use case to demonstrate machine learning methods applied to both predicting and interpreting regularly used transportation modes. We design an innovative analysis framework, combining sequence clustering and tree-based machine learning methods coupled with TreeExplainer, to predict and interpret habitual modes using mid- to long-term predictors. Five life course clusters are derived to provide evaluation and interpretation contexts.

We find that the commonly used global feature importance is not representative of all the sub-populations with different life history contexts. Variable selection using the global feature importance ranking may miss important features for certain sub-populations and therefore produce biased predictions.

The TreeExplainer-derived local explanation can be straight forward for input features that have a monotonic association with the outcome variables. For input features whose contribution are conditioned on others, this paper has shown that such conditionality is better interpreted by presenting feature interactions within different life history contexts.

The analysis framework implemented here demonstrates a promising step toward interpretable machine learning applications to mid- to long-term prediction of travel modes for transportation planning.

Acknowledgment

This work was supported by the Office of Workforce Development for Teachers and Scientists (WDTS) under the Visiting Faculty Program (VFP), Office of Advanced Scientific Computing Research and Vehicle Technologies Office of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] F. Wang and C. L. Ross, "Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transp. Res. Rec.*, vol. 2672, no. 47, pp. 35–45, Dec. 2018.

[2] B. Clark, K. Chatterjee, and S. Melia, "Changes to commute mode: The role of life events, spatial context and environmental attitude," *Transp. Res. Part A: Policy Pract.*, vol. 89, pp. 89–105, Jul. 2016.

[3] S. Beige and K. W. Axhausen, "Interdependencies between turning points in life and long-term mobility decisions," *Transportation*, vol. 39, no. 4, pp. 857–872, Jul. 2012.

[4] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Modeling stated preference for Mobility-on-Demand transit: A comparison of machine learning and logit models," Nov. 2018.

[5] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "Explainable AI for trees: From local explanations to global understanding," May 2019.
[6] C. A. Spurlock, J. Sears, G. Wong-Parodi, V. Walker, L. Jin, M. Taylor, A. Duvall, A. Gopal, and A. Todd, "Describing the users: Understanding adoption of and interest in shared, electrified, and automated transportation in the san francisco bay area," *Transp. Res. Part D: Trans. Environ.*, vol. 71, pp. 283–301, Jun. 2019.
[7] G. Pollock, "Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 1, pp. 167–183, 2007. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2006.00450.x/full
[8] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, and others, "Data quality challenges with missing values and mixed types in joint sequence analysis," *2017 IEEE International*, 2017.
[9] A. Lazar, L. Jin, C. A. Spurlock, K. Wu, A. Sim, and A. Todd, "Evaluating the effects of missing values and mixed data types on social sequence clustering using t-SNE visualization," *J. Data and Information Quality*, vol. 11, no. 2, pp. 7:1–7:22, Mar. 2019.
[10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.
[12] E. Al Daoud, "Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset," *pdfs.semanticscholar.org*.
[13] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.
[14] E. A. Minastireanu and G. Mesnita, "Light GBM machine learning algorithm to online click fraud detection," *J. Inform. Assur. Cybersecur*, 2019.