Predicting Daily Trip Frequencies of Vulnerable Households in NYS using Supervised Machine Learning Approaches

Bumjoon Bae

Ho-Ling Hwang

Shih-Miao Chin

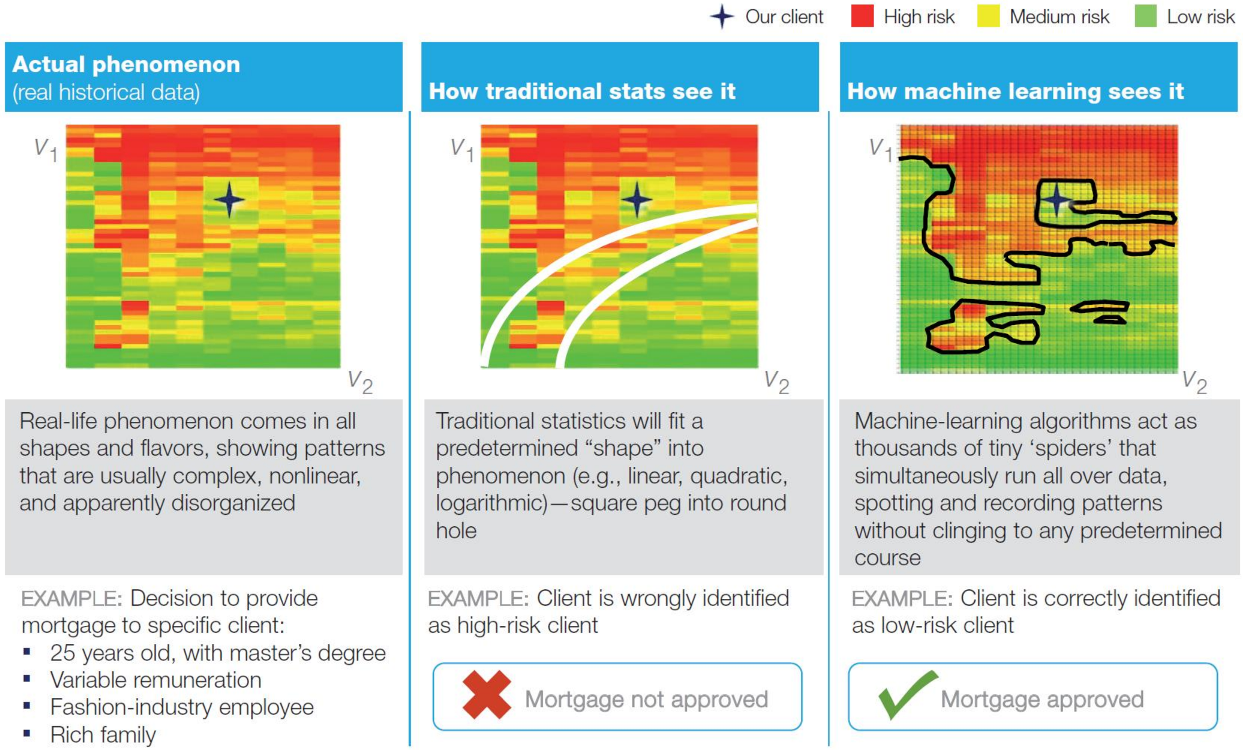
Chieh (Ross) Wang

Outline

- Introduction
- Regression Approach
- Classification Approach
- Prediction Performance Comparison
- Conclusions

Machine Learning

- "Machine learning (ML) identifies complex nonlinear patterns in large datasets, so as to make more accurate models possible." – McKinsey report (2015)



Source: https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/risk/pdfs/the_future_of_bank_risk_management.ashx

Machine Learning Algorithms

- Machine Learning
  - Supervised Learning
    - Regression
      - Linear Regression
      - Non-linear Regression
      - Decision Trees
      - Neural Networks
      - ⋮
    - Classification
      - Linear Regression
      - Non-linear Regression
      - Decision Trees
      - Nearest Neighbor
      - Neural Networks
      - Discriminant Analysis
      - Support Vector Machine
      - ⋮
  - Unsupervised Learning
    - Clustering
      - k-Means
      - Fuzzy Clustering
      - Hierarchical
      - Neural Networks
      - Gaussian Mixture
      - Hidden Markov Model
      - ⋮
  - Semi-supervised Learning

ML Applications in Transportation Studies

- Heavily applied for almost every area in transportation
  - Travel demand modeling;
  - Fuel consumption, emission estimation;
  - Real-time traffic flow & travel time prediction, congestion detection;
  - Transportation data imputation;
  - Driving behavior model calibration;
  - Object detection and path planning (CAVs);
  - Automatic vehicle classification;
  - Infrastructure condition evaluation and modeling (e.g., crack detection/classification);
  - Etc.

Research Background & Objective

- Limitations of the conventional trip generation model using linear regression in literature.
  - Negative trip rates likely
  - Continuous nature in trip rates
  - Lacks in a traveler's behavior mechanism (e.g., cost minimization or utility maximization)
- Nonetheless, linear regression has shown comparable or better performance, compared with alternative models (e.g., tobit, Poisson, negative binomial, truncated normal, ordered logit).
- This study is to explores supervised machine learning methods to predict trip rates of individual travelers using 2017 NHTS data.

Datasets

- 2017 NHTS data
  - Travelers living in New York state
  - Low income household (below 2017 Poverty Threshold by Census Bureau)
  - Sample size: 1,731 (70/30 splits for training/testing, 100 runs)
  - 20 predictors from the person/household data

Traveler characteristics	Household characteristics	Regional characteristics
Age, Educational attainment, Sex, Race, Medical condition, Opinion of Health, Born in U.S., Public Transit Usage, Worker status, Driver status, Home ownership,	Household size, Count of household vehicles, household income, Number of drivers, Number of workers, Household in urban/rural area, Number of children	Population density of the household's home location, Employment density of the household's home location

Linear Regression Approaches

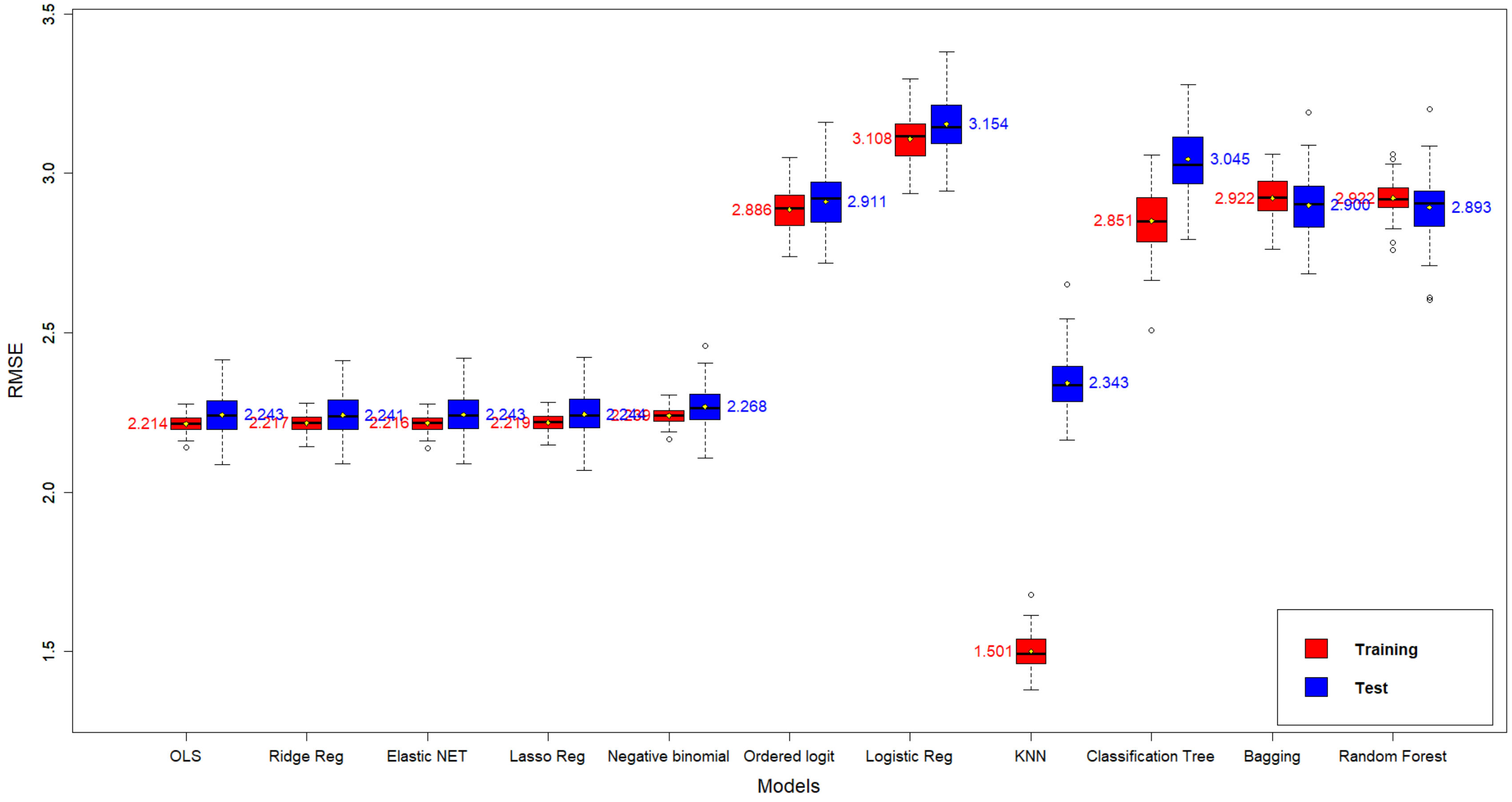
Model	Pros	Cons
Ordinary least square (base model)	Simple functional structure	Multicollinearity; Curse of dimensionality;
Ridge regression	Shrink variable estimates when multicollinearity exists	Doesn't produce a sparse model (i.e., no subset selection)
Lasso regression	Drop off variables with less effects; Can be used when the number of predictors exceeds sample size.	Doesn't address multicollinearity issue; May introduce bias.
Elastic net regression	Hybrid model of Ridge and Lasso regression; Address both multicollinearity and variable selection	Tuning parameter selection problem
Negative binomial	Count data model; Overdispersed dependent variable	
Ordered Logit	Can treat an ordinal dependent variable using a latent continuous variable and cutoff values	Bias if the ordered-response choice mechanism is not true.

Classification Approaches

Model	Pros	Cons
K-nearest neighbor	Simple nonparametric method; Computational efficiency; Can handle missing values; Robust to outliers; Predictive power	Prone to overfit (depending on k); Low interpretability; Sensitive to distance function selection
Multinomial logistic regression	Easy interpretation (probability scores for observation) Computational efficiency (linear model)	Low predictive power with large number of categorical variables; IIA assumption
Classification trees	Can handle missing values; Robust to outliers; Easy interpretation;	Instability with high variance (highly rely on training data); Computations become prohibitive with a large number of multi-class categorical predictors
Bagging trees	Reduces squared error by decreasing variance compared to classification tree	Limited variance reduction due to high correlation between trees by using all predictors
Random Forest	Reduces squared error by decreasing variance to classification tree	

Performance comparison

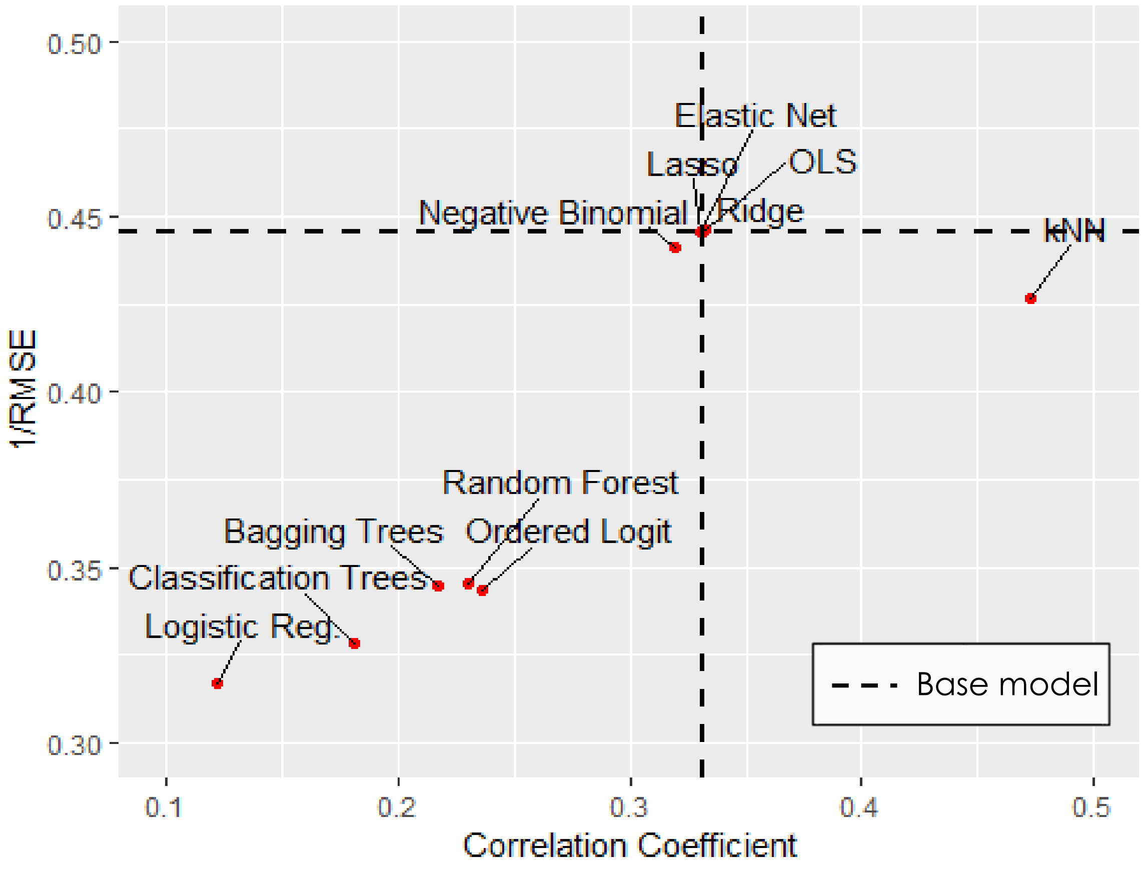
- Root mean squared error (RMSE)



OAK RIDGE
National Laboratory

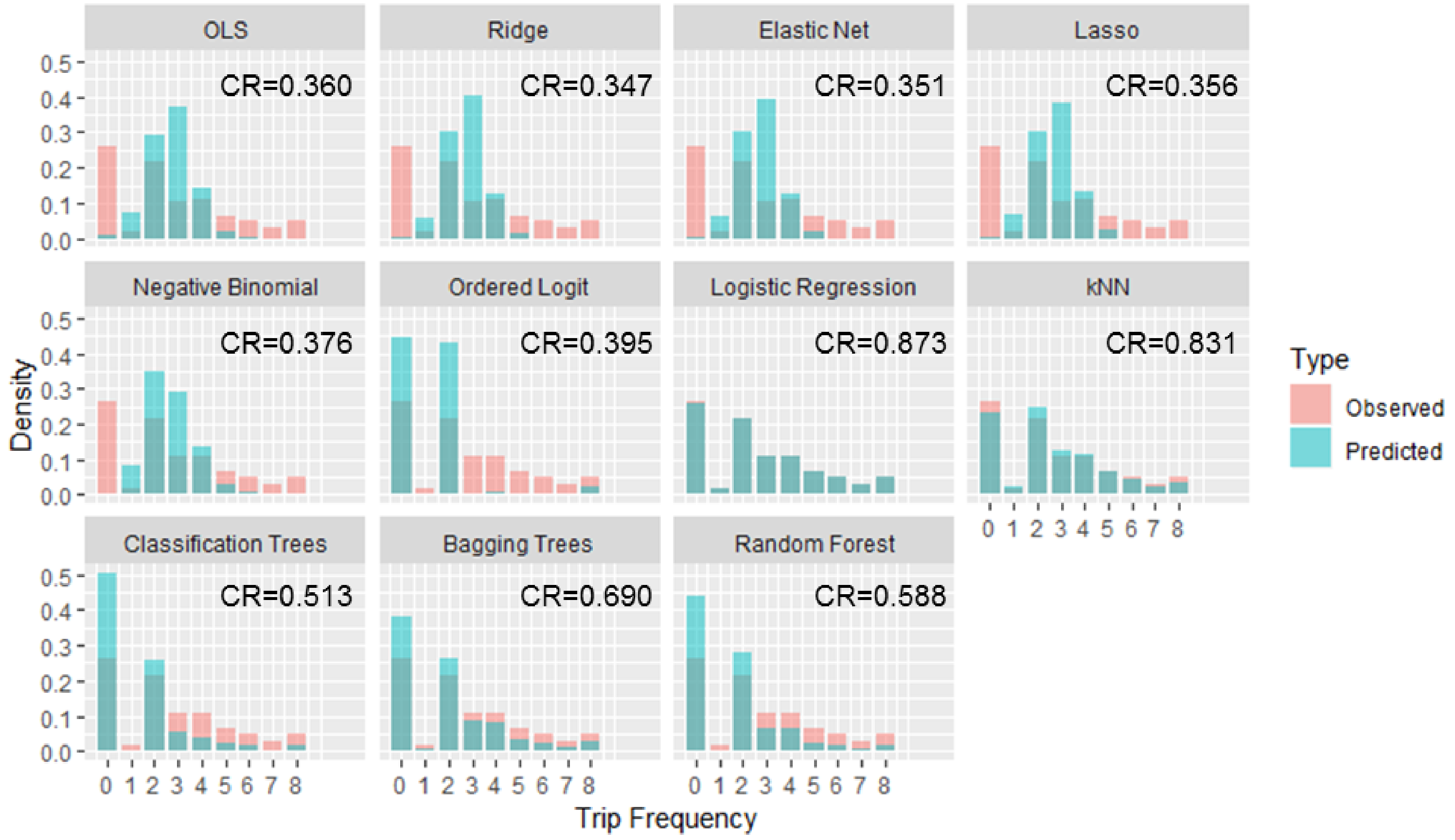
Performance comparison

- RMSE & correlation coefficient



Performance comparison

- Coincidence Ratio

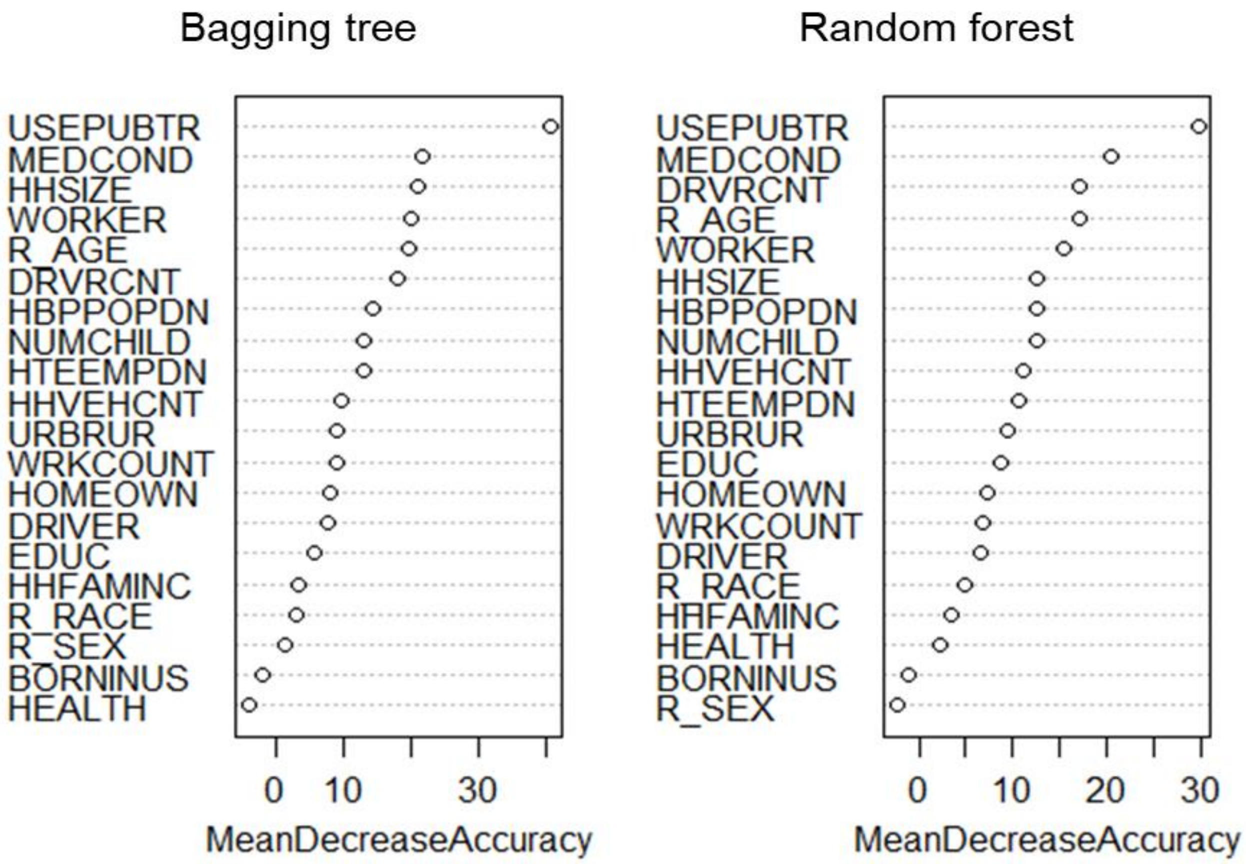


Estimated regression models

Name	OLS	Ridge	Elastic net	Lasso	Negative binomial	Ordered logit
R_AGE	-0.002	0.0004	0.003		0.003**	-0.003
EDUC	0.134	1.440	0.144	0.114	0.070	0.081***
R_SEX	0.078	-0.058	-0.049		0.074	0.052***
R_RACE	0.360**	0.118	0.208	0.165	0.181***	0.284***
MEDCOND	-0.730***	-0.499	-8.363	-0.862	-0.305***	-0.659***
HEALTH	-0.077	-0.122	-0.670	-0.044	-0.186	-0.366***
BORNINUS	0.399*	0.192	4.461	0.369	0.392***	0.278***
USEPUBTR	1.672***	0.440	1.187	1.266	0.647***	1.265***
WORKER	0.095	0.274	3.309	0.325	0.064	0.165***
DRIVER	0.614***	0.382	7.585	0.809	0.315***	0.442***
HOMEOWN	-0.360**	-0.157	-2.690	-0.226	-0.149**	-0.248***
HHSIZE	-0.337***	-0.280	-2.173	-0.170	-0.103***	-0.296***
HHVEHCNT	0.147	0.112	1.332	0.107	0.059	0.130***
HHFAMINC	0.00001	0.0002	0.00002		0.00001	0.00001
DRVRCNT	-0.174	-0.093	-2.071	-0.237	-0.073	-0.128***
WRKCOUNT	0.372***	0.173	1.743	0.167	0.160***	0.284***
NUMCHILD	0.301**	0.232	1.986	0.171	0.110**	0.255***
URBRUR	0.154	0.144	2.556	0.340	0.164**	0.142***
HBPPOPDN	0.00003	0.00002	1.581		0.00001	0.00002
HTEEMPDN	-0.0001	-0.00002			-0.00003	-0.0001
Sample size	1,212	1,212	1,212	1,212	1,212	1,212
F-statistic	10.440***	10.183***	10.728***	13.635***		
Likelihood ratio statistic					148.638***	205.81***
R^2	0.149	0.147	0.149	0.146		
ρ^2					0.029	0.045

* 10%, ** 5%, *** 1% significance level, respectively.

Variable importance in tree methods



Concluding Remarks

- Advanced methods do not necessarily provide significant improvement in trip generation.
- Regularized regression models (Ridge, Lasso, ENET) improve prediction performance slightly.
- On average, classification methods give higher prediction errors for individual travelers; but higher accuracy in trip frequency distribution for overall sampled population.
- kNN performs best among the classification models; but prone to overfit data substantially. (It depends on k value)

Future Research

- Enhance performance of classification methods by
  - Tuning parameter settings
  - Different number of classes for trip frequency
  - Categorical variables with binary or multi classes
  - Etc.
- Test other models e.g., deep learning methods
- Compare transferability of each method using other validation sets (e.g., different region, year, etc.)
- Non-low-income population
- Person and household weights