Machine Learning Approach to Travel Modeling

A Reflection with Two Case Studies

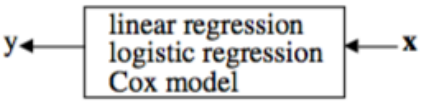
Liming Wang

Portland State University

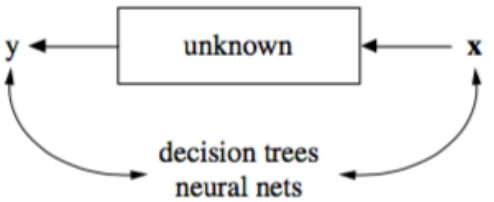
2018 TRB Innovations in Travel Modeling Atlanta, GA

Statistical Modeling vs Machine Learning

Two cultures of developing models (Breiman, 2001):



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Statistical models ("the data modeling"): Assuming a data generation model and use data and hypothesis testing framework to recover parameters of the data generation process; the focus is more on .

Machine learning ("algorithmic modeling"): With no assumption of data generation process, use computer algorithms for pattern recognition and data-driven predictions-making; the focus is

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

'Petabytes allow us to say: "Correlation is enough." (...)

We can throw the numbers into the biggest computing clusters the world ... has ever seen and let statistical algorithms find patterns where science cannot.'

Chris Anderson, WIRED.com, 2008

http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Challenges to Statistical Models

Or the case for machine learning:

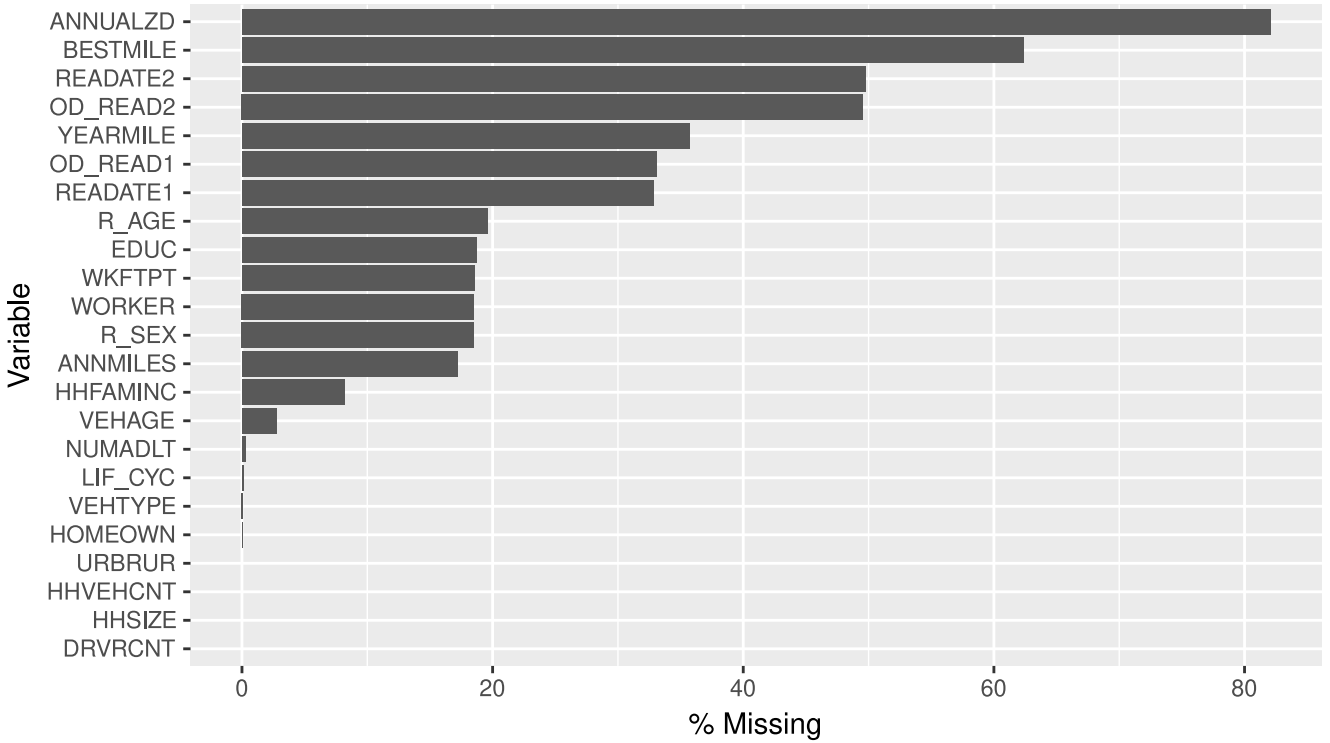
- Assumption/theory of the data generation process may be wrong
- Competing data generation models may give different pictures of the relation between the predictors and response variable;
- Changing landscape of data availability
  - Curse of dimensionality
  - Easy to detect significant correlations with large sample size
  - Increasingly models involving data of the population instead of a sample; model assumptions may not be valid
  - Missing data issue

Two Case Studies

- Imputation of missing data in travel surveys
- Models travel outcomes (VMT)

Case I: Imputation of Missing Data

Annual Vehile Miles Travelled information in the 2001 National Household Travel Survey (NHTS)



Only 12% (17037 out of 139382) observations are complete.

Multiple Imputation by Chained Equations

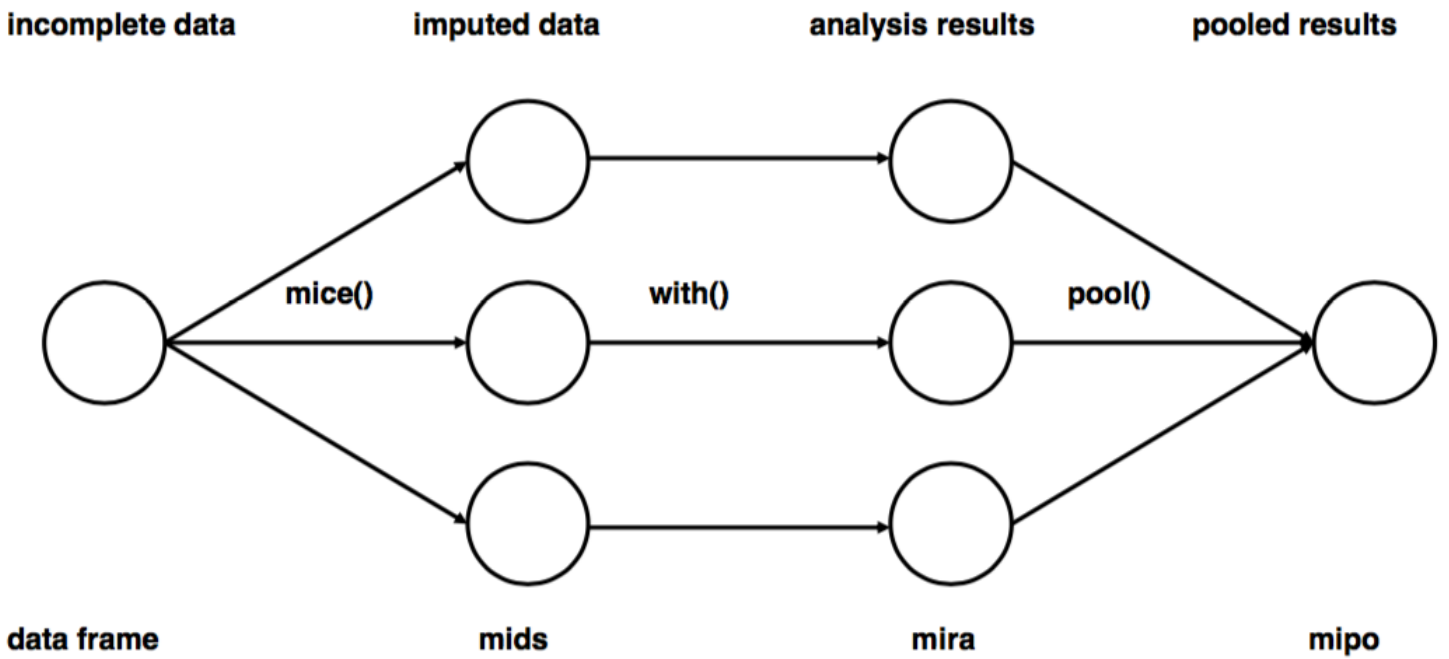


Figure 1: Main steps used in multiple imputation.

Source: van Buuren, Stef and Karin Groothuis-Oudshoorn, 2011. mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, Vol 45 (3).

Imputation Results (1)

Validation: randomly set 10% of values to missing, impute them and compare with actual values

Variable	Normalized RMSE (%)
ANNMILES (Self reported annual VMT)	31.824
ANNUALZD (VMT annualized from two Odmeter readings)	22.264
HHFAMINC (Family income)	4.750

Imputation Results (2): Comparing linear regression results (y=ANNUALZD) without and with multiple imputation

	No Imputation	w/ Imputation
(Intercept)	8904.27***	8722.34***
	(306.50)	(185.89)
Workers	2497.73***	1912.89***
	(173.60)	(120.73)
Urban	-1321.43***	-908.94***
	(159.87)	(126.99)
Income ($ 30k – 60k)	837.84***	769.23***
	(167.85)	(103.22)
Income ($ 60k+)	1713.78***	1515.67***
	(173.96)	(81.99)
Parents with young children	2003.62***	1356.74***
	(272.99)	(148.76)
Couples w/o children	771.46**	259.57
	(274.32)	(141.73)
Empty nesters	-1195.52***	-1400.96***
	(283.42)	(151.84)
# Drivers	544.21***	508.79***
	(96.55)	(57.33)
Pop. Density (bg)	-0.05*	-0.017
	(0.03)	(0.014)
Emp. Density (tract)	-0.33***	-0.124***
	(0.08)	(0.037)
R^2	0.08	
Adj. R^2	0.08	
Num. obs.	23647	

***p < 0.001, **p < 0.01, *p < 0.05

Case II: Travel Behavior Modeling

$$\text{VMT}_h \leftarrow (\text{SES}_h, \text{regional characteristics}, \text{built environment})$$

Data Sources:

- 2009 NHTS for household's SES, travel outcome (VMT);
- EPA's Smart Location Database (for blockgroup level 5D built environment measures);
- Highway Performance Measure System for regionwide roadway information;
- National Transit Database for regionwide transit supply.

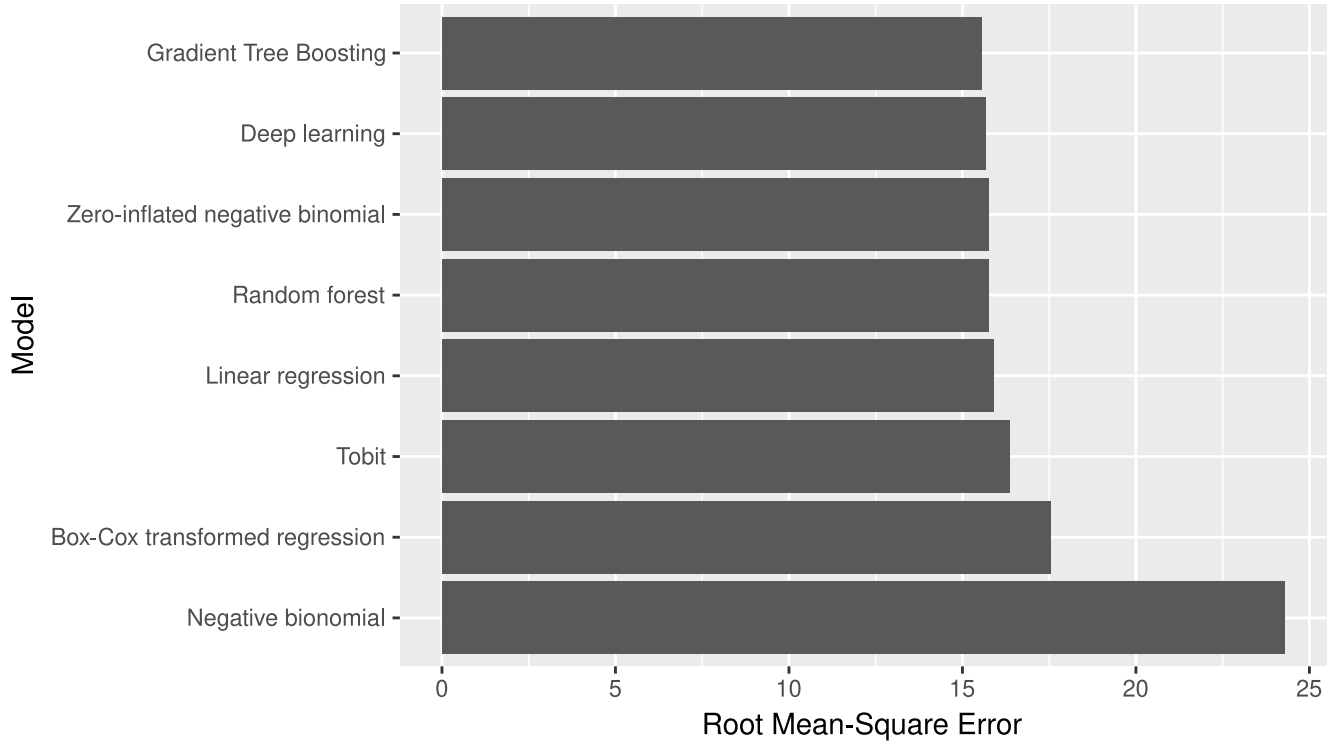
150,000 households with more than 180 independent variables (before considering non-linear transformation or interaction between variables)

VMT models

- Statistical Models
  - linear regression
  - non-linear regression (transformed dependent variable)
  - tobit model
  - zero-inflated negative binomial model
- Machine learning algorithms
  - Random Forest
  - Gradient Tree Boosting
  - Deep Neural Network

Cross Validation Results

- Dependent variable is household VMT on the day of survey
- Data are randomly partitioned into 5 parts for a 5-fold cross-validation



Conclusion and Discussion

Conclusions:

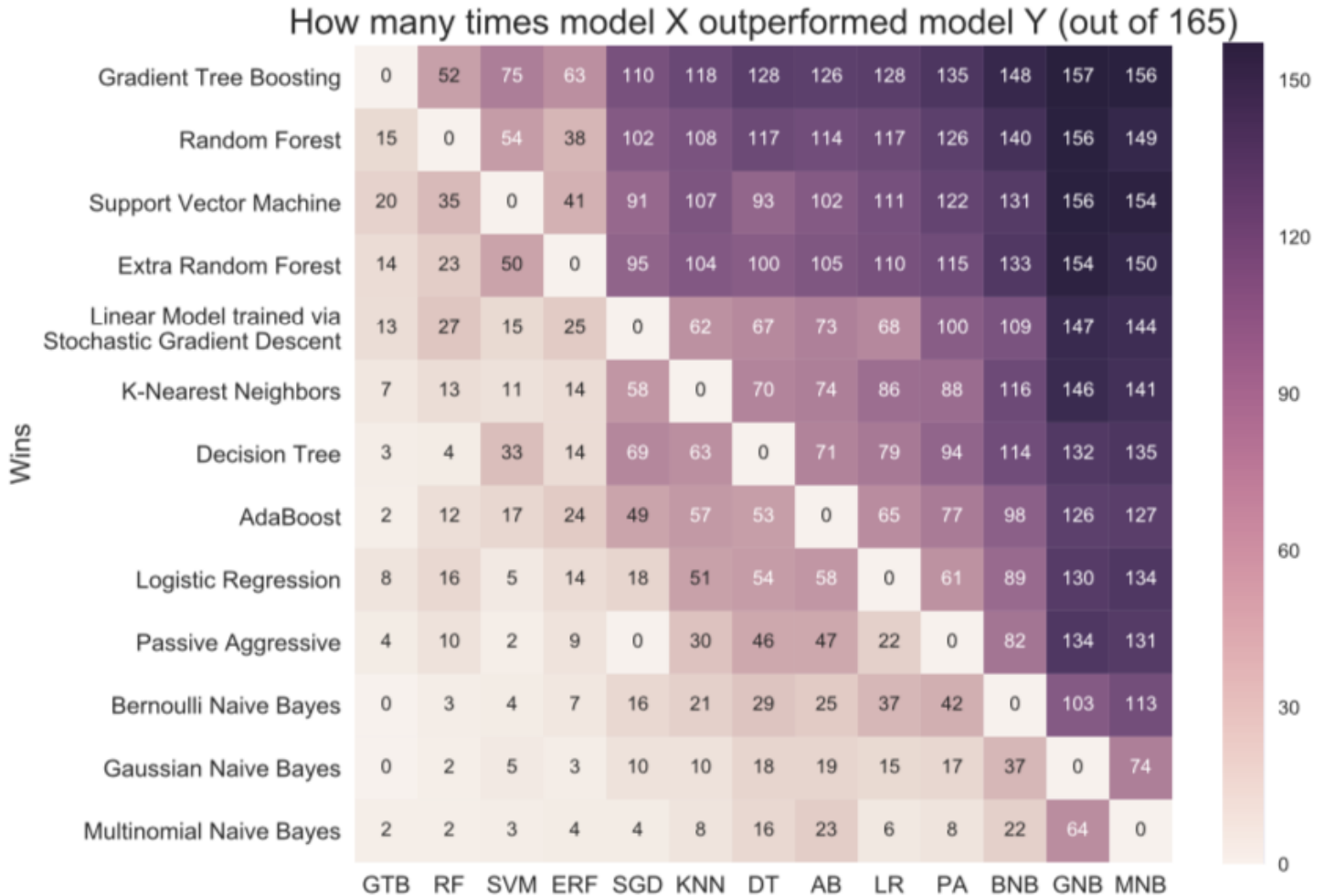
- Some tasks, such as multivariate data imputation, are hard or impossible to do with statistical models but possible with machine learning,
- Growing modeling complexity adds challenges to statistical models, machine learning has an advantage
- If you're developing models for prediction, there are few reasons not to look into machine learning algorithms

Challenges

- Combining machine learning skills with the domain knowledge;
- Train students with machine learning skills;
- Computation intensity & access to computer resources

Acknowledgements

- Oregon Department of Transportation (SPR 788)
- National Institute for Transportation and Communities (NITC-881)
- Portland Institute for Computational Science (PICS) and its resources acquired using NSF Grant #DMS 1624776 and ARO Grant #W911NF-16-1-0307



Benchmarking Machine Learning Algorithms

Source: Randal S. Olson and William La Cava et al., 2018.