Predictive Analysis on Mean Commute Time

Based on Census Tract Dataset

Yiyang Zhou & Haoyuan Xia

AAE 722

Nov.17 2022

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Outline

Outline

Introduction

Labor Force and Regional Development

- Labor force plays an important role in regional development (Florida et al., 2008; Shapiro, 2006; Jones, 2002; Simon, 1998; Rauch, 1993).
- Ambitious cities should try to attract labor force to boost regional development (especially the development of certain industries).
- More importantly, cities need to make labor forces stay.

Commute Time, Job Satisfaction, and Turnover Rate

- Long commute time significantly increases employees' perceived stress (Gottholmseder et al., 2009), decreases job satisfaction (Amponsah-Tawiah et al., 2016), and does harm to some employees' psychological health (Roberts et al., 2011).
- Low job satisfaction leads to high turnover rate (Porter et al., 1974).

Urban Planning

- Focus more on urban planning to avoid traffic digestion and decrease commute time.

Outline

ACS Census Tract Data

American Community Survey

- Originally from 2015 American Community Survey 5-year estimates
- Randomly choose 35,000 Census Tracts from the full dataset
- Census Tract: a small, relatively permanent statistical subdivisions of a county
- Attributes: *Demographics, Occupation, Employment*

Data Cleaning Process

Proportion of Missing Values

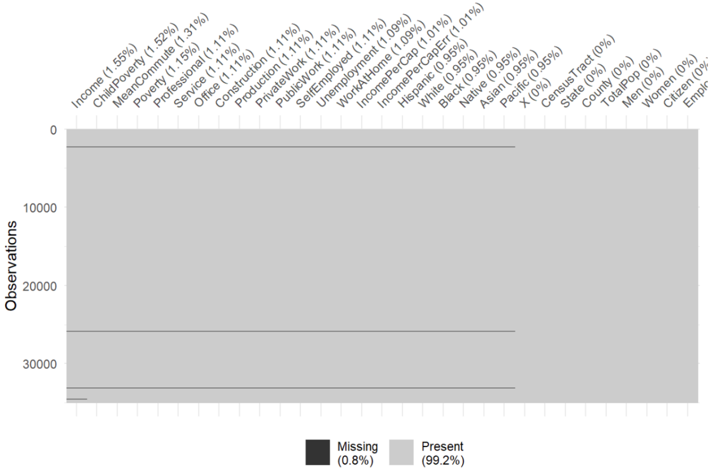


Figure 1: Proportion of Missing Values

Data Cleaning Process (Cont'd)

Drop Problematic Observations

Operation	N Dropped	N Remaining
Raw Census Tract Data	0	35,000
Drop Missing Values	623	34,377
Drop Duplicated Census Tract Number	1,158	33,219

Table 1: Data Cleaning Process

Drop Unused Variables

- IncomePerCapErr: a variable measuring margin of error when estimating Income Per Capita

Split Sample

- Train: 75%; Test: 25%

Descriptive Statistics

Variable	N	Mean	SD	Min	Max
TotalPop	33219	4401.127	30104105.335	66	39454
Men	33219	2162.097	8370440.716	30	27962
Women	33219	2239.03	7670967.316	26	18182
Hispanic	33219	16.723	1901.627	0	100
White	33219	62.51	3260.629	0	100
Black	33219	12.947	1738.043	0	100
Native	33219	0.747	363.031	0	100
Asian	33219	4.568	432.139	0	91.3
Pacific	33219	0.141	35.3	0	64
Citizen	33219	3092.963	13803746.041	53	28932
Income	33219	57655.036	4797093957.032	2611	245870
Poverty	33219	16.677	700.895	0	98.6
ChildPoverty	33219	22.205	1288.455	0	100
Professional	33219	34.909	1203.893	0	100
Service	33219	18.955	375.188	0	69
Office	33219	23.91	288.143	0	74.4
Construction	33219	9.34	171.673	0	66.8
Production	33219	12.887	255.029	0	59.6
WorkAtHome	33219	4.364	99.158	0	90.6
MeanCommute	33219	25.817	396.779	5.3	70.5
Employed	33219	2024.512	7166949.803	9	18538
PrivateWork	33219	78.944	1215.477	14.1	100
PublicWork	33219	14.63	325.335	0	85.7
SelfEmployed	33219	6.255	75.016	0	44.2
Unemployment	33219	8.966	182.995	0	68.8

Table 2: Descriptive Statistics

Distribution of Mean Commute Time

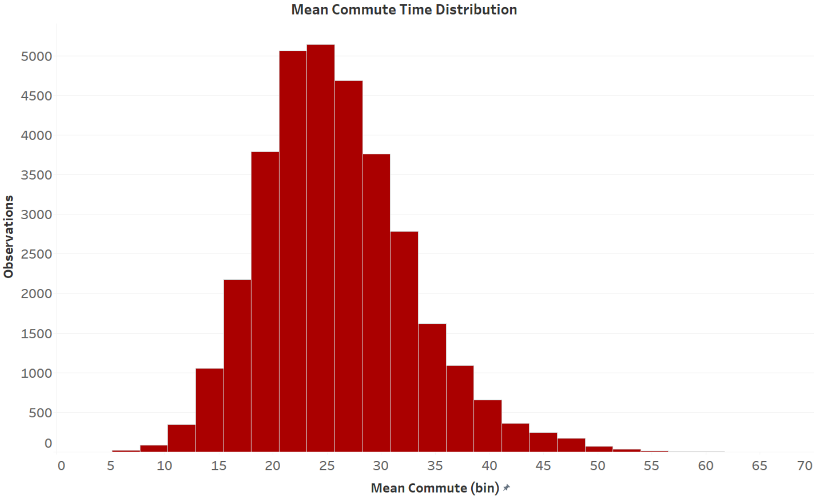


Figure 2: Distribution of Mean Commute Time

Mean Commute Time by State

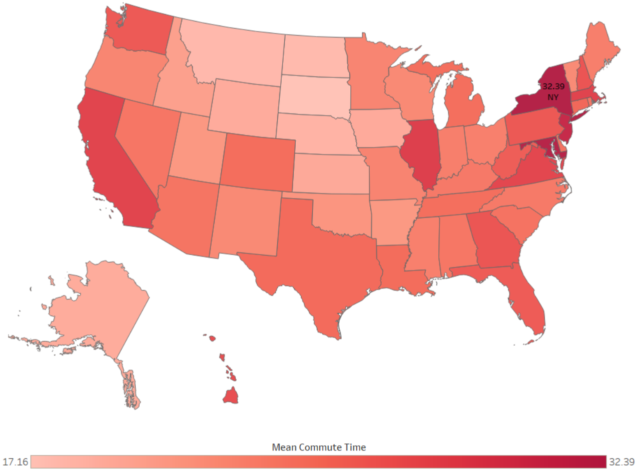


Figure 3: Mean Commute Time by State

Outline

Linear Regression Model

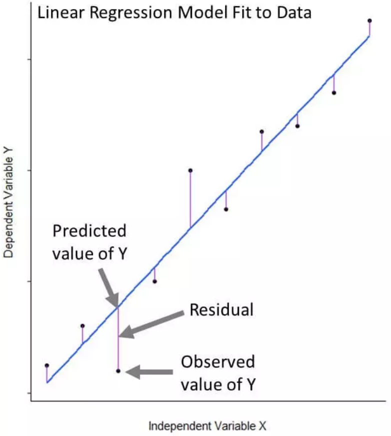


Figure 4: Linear Regression Diagram[1]

[1] Source: DataQuest

OLS with Best Subset Selection

Forced Out Variables (Avoid Perfect Multicollinearity)

- *Men, White, IncomePerCap, Professional, PublicWork*

Cross Validation

- 5×5 folds

Search Method

- Exhaustive search

Ramdom Forest

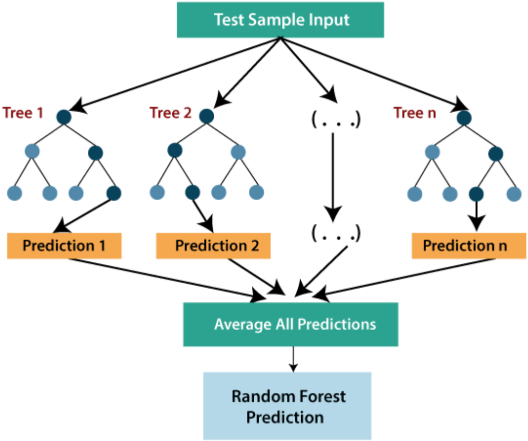


Figure 5: Random Forest Diagram[2]

[2]Source: Broadinfinity

Ramdom Forest

Tuning Parameters

- Minimal Node Size: From 4 to 10
- Number of Predictors for Each Split: From 4 to 15

Split Rule

- Variance

Number of Trees

- 1000

Neural Network

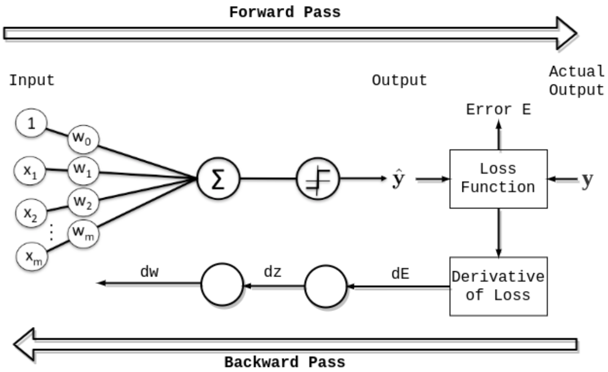


Figure 6: Neural Network Diagram[3]

[3]Source: Baeldung

Neural Network

Tuning Parameter

- Number of Hidden Layers: From 1 to 5

Other Parameters

# of Hidden Layers	# of Neurons	Dropout Rate	Activation Function
1	32	0.05	ReLU
2	32, 16	0.05	ReLU
3	32, 16, 8	0.05	ReLU
4	32, 16, 8, 4	0.05	ReLU
5	32, 16, 8, 4, 2	0.05	ReLU

Table 3: Other Parameters in the Neural Network

Data is properly scaled for Neural Network

Outline

OLS: Best Subset

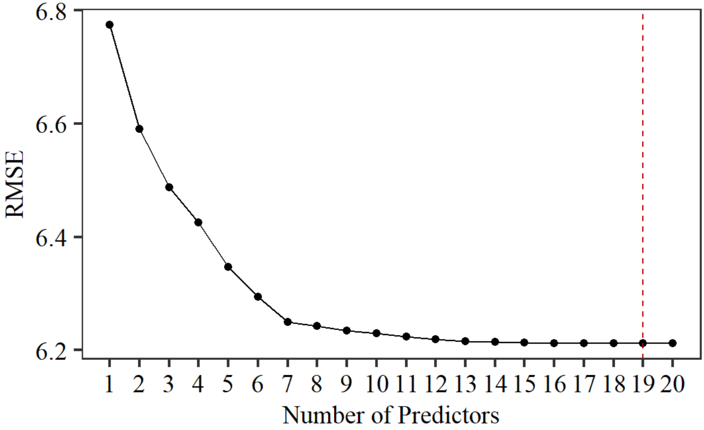


Figure 7: Best Subset of Predictors

OLS: Final Model

$$
\begin{aligned}
MeanCommuteTime = \beta_0 &+ \beta_1 TotalPop + \beta_2 Women + \beta_3 Hispanic \\
&+ \beta_4 Black + \beta_5 Native + \beta_6 Asian + \beta_7 Pacific \\
&+ \beta_8 Income + \beta_9 Poverty + \beta_{10} ChildPoverty \\
&+ \beta_{11} Service + \beta_{12} Office + \beta_{13} Construction \\
&+ \beta_{14} Production + \beta_{15} WorkAtHome \\
&+ \beta_{16} Employed + \beta_{17} PrivateWork \\
&+ \beta_{18} SelfEmployed + \beta_{19} Unemployment \\
&+ \epsilon
\end{aligned}
$$

OLS: Prediction on Test Set

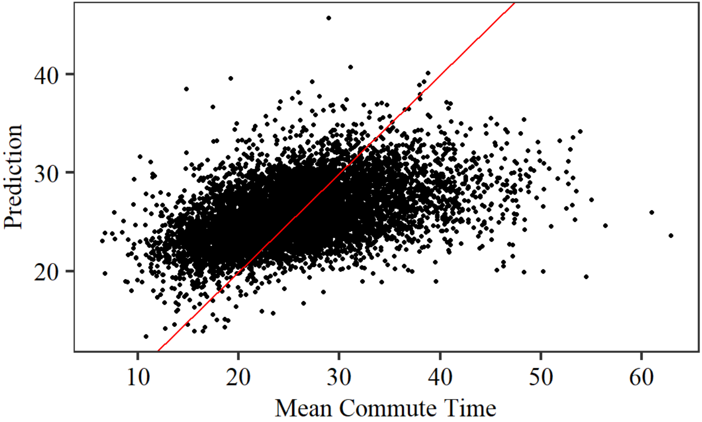


Figure 8: OLS Prediction on Test Set (RMSE = 6.168)

Random Forest: Optimal Parameters

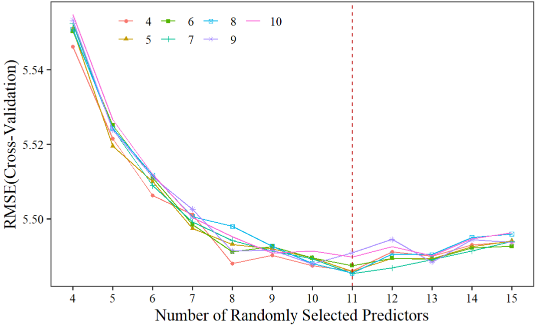


Figure 9: Tuning Parameters of the Random Forest Model

Parameter	Randomly Selected Predictors	Minimal Node Size	Split Rule
Best Tune	11	7	Variance

Table 4: Best Tuning Parameters

Random Forest: Number of Trees

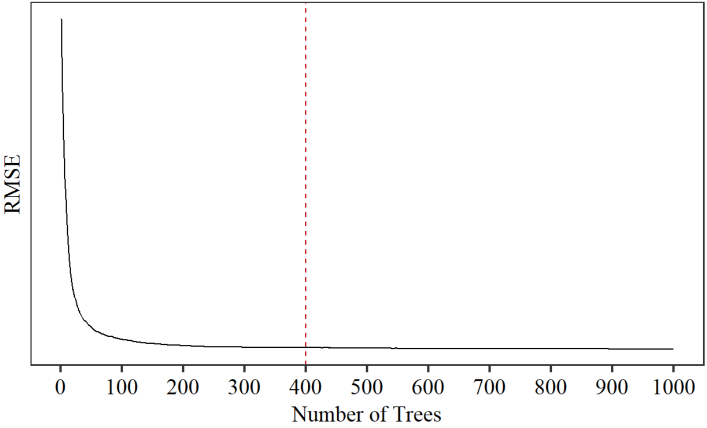


Figure 10: Number of Trees and RMSE

Random Forest: Importance of Predictors

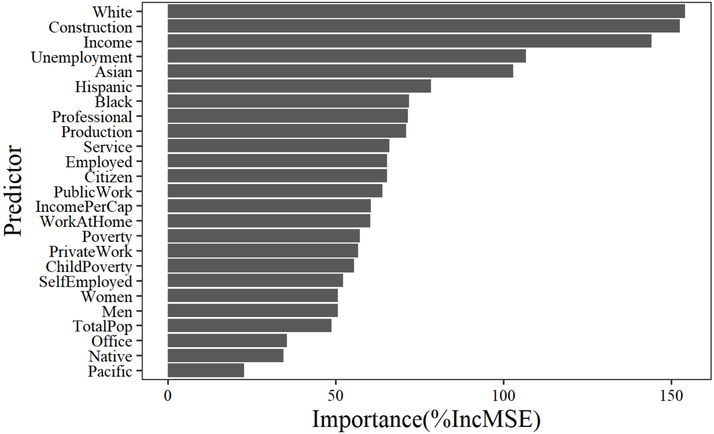


Figure 11: Importance of Predictors

Random Forest: Prediction on Test Set

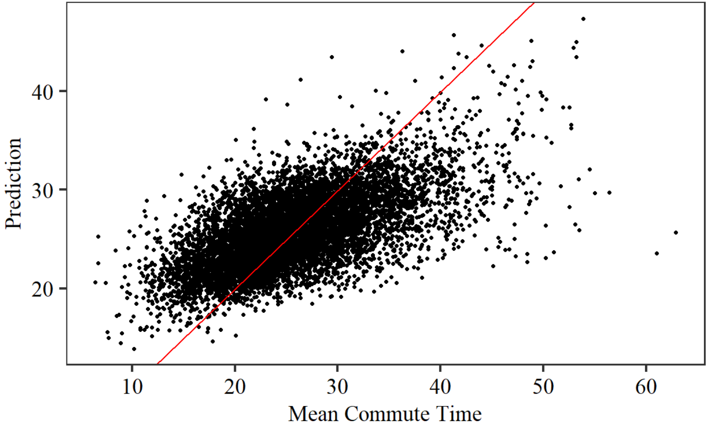


Figure 12: Random Forest Prediction on Test Set (RMSE = 5.456)

Neural Network Model with Optimal Layer: Selection

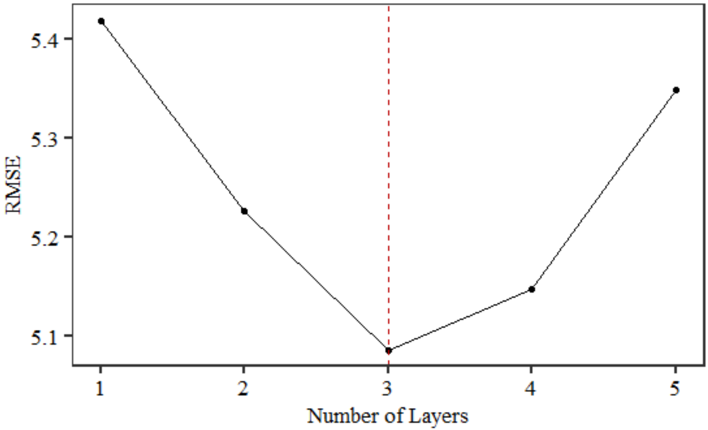


Figure 13: RMSE of the Neural Network Model on Training Set with Each Layer

Neural Network Model with Optimal Layer: Structure

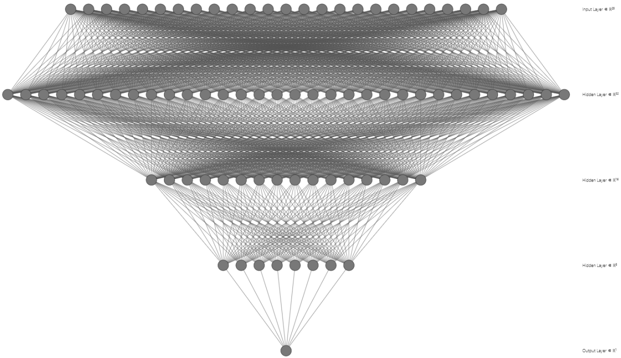
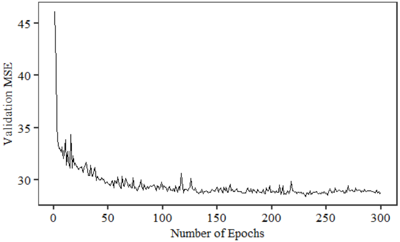
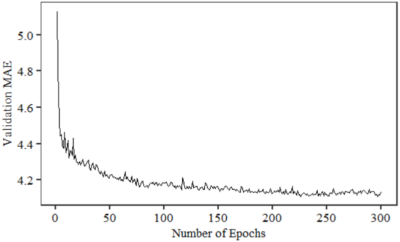


Figure 14: Structure of the Neural Network Model with Optimal Layer

Neural Network: Training Process



(a) Validation MSE



(b) Validation MAE

Figure 15: Training Process and Prediction Error of the Neural Network

Neural Network: Prediction on Test Set

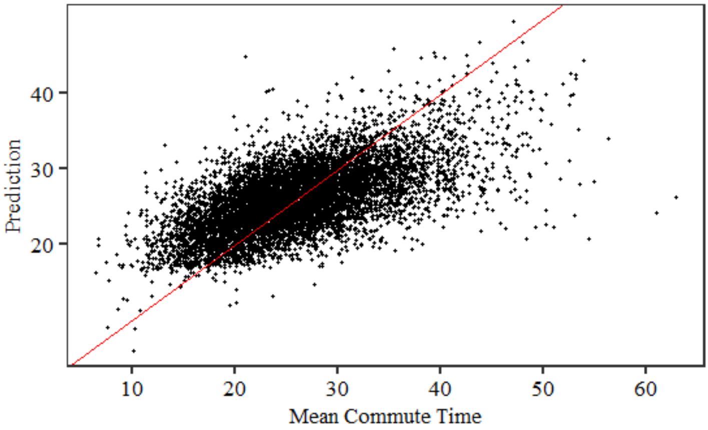
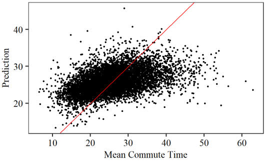


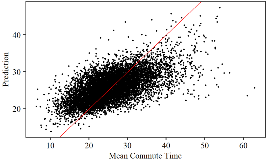
Figure 16: Neural Network Prediction on Test Set (RMSE = 5.421)

Outline

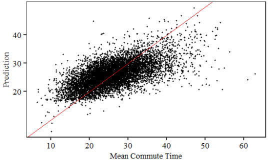
Comparison Between Three Models



(a) OLS



(b) Random Forest



(c) Neural Network

Figure 17: Prediction of Three Models on Test Set

Model	Prediction Error (RMSE)
Best Subset OLS	6.168
Ramdom Forest	5.456
Neural Network	5.421

Table 5: Comparison Between Three Models

Prediction v.s. Causal Inference

Estimate the Effect of Occupation Structure on Commute Time

- Machine Learning Results Above Imply No Causal Relationship
  - ML methods focus on prediction but not the parameters
  - Omitted Variables: e.g. *Infrastructure*
  - Reverse Causality: commute time could affect occupation structure
- Possible Ways of Causal Inference
  - Panel Data & More Variables
  - Exogenous Shock: Talent Policy (DID)
  - Alternative: Implement a Sorting Model (Kuminoff et al., 2013)

Outline

Conclusion

Variable Importance

- *Occupation Structure*, *Race* and *Income* are the three most important attributes.

Best Model

- Random Forest and Neural Network performed similarly in predicting mean commute time. They both improved RMSE by nearly 1 compared to the baseline model.

Machine Learning Methods Imply No Causal Inference

- Machine Learning focuses on prediction but not estimates of parameters. More variables and information are needed to identify the causal relationship.

References I

Amponsah-Tawiah, K., Annor, F., and Arthur, B. G. (2016). Linking commuting stress to job satisfaction and turnover intention: The mediating role of burnout. *Journal of Workplace Behavioral Health*, 31(2):104–123.

Florida, R., Mellander, C., and Stolarick, K. (2008). Inside the black box of regional development—human capital, the creative class and tolerance. *Journal of Economic Geography*, 8(5):615–649.

Gottholmseder, G., Nowotny, K., Pruckner, G. J., and Theurl, E. (2009). Stress perception and commuting. *Health Economics*, 18(5):559–576.

Jones, C. I. (2002). Sources of u.s. economic growth in a world of ideas. *American Economic Review*, 92(1):220–239.

Kuminoff, N. V., Smith, V. K., and Timmins, C. (2013). The new economics of equilibrium sorting and policy evaluation using housing markets. *Journal of Economic Literature*, 51(4):1007–62.

Porter, L. W., Steers, R. M., Mowday, R. T., and Boulian, P. V. (1974). Organizational commitment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology*, 59(5):603.

References II

Rauch, J. E. (1993). Productivity gains from geographic concentration of human capital: Evidence from the cities. *Journal of Urban Economics*, 34(3):380–400.

Roberts, J., Hodgson, R., and Dolan, P. (2011). "it's driving her mad": Gender differences in the effects of commuting on psychological health. *Journal of Health Economics*, 30(5):1064–1076.

Shapiro, J. M. (2006). Smart Cities: Quality of Life, Productivity, and the Growth Effects of Human Capital. *The Review of Economics and Statistics*, 88(2):324–335.

Simon, C. J. (1998). Human capital and metropolitan employment growth. *Journal of Urban Economics*, 43(2):223–243.