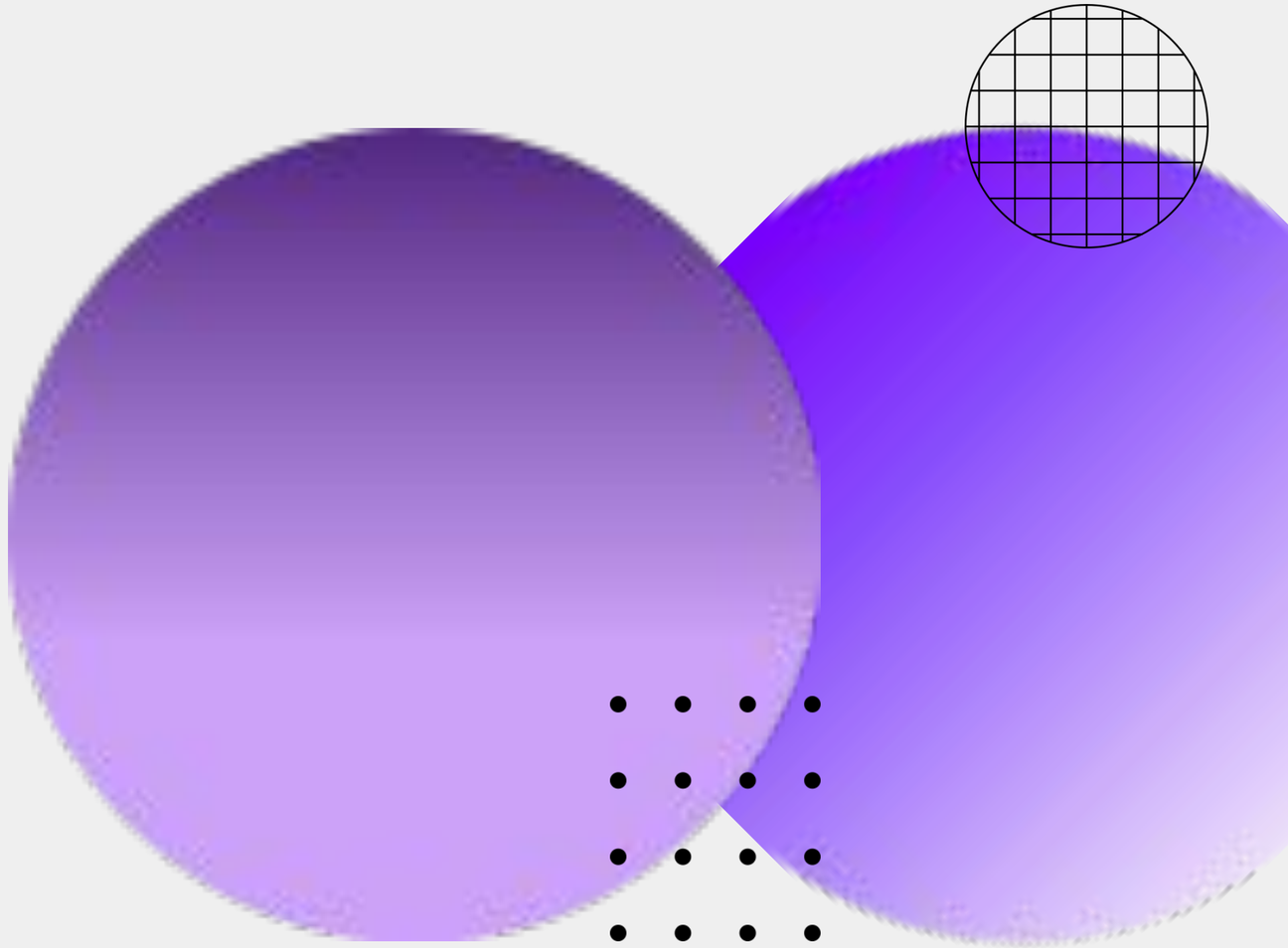
Modeling Commute Travel in the NY/NJ Metropolitan Area to Predict the Market Share of Autonomous Vehicles using Imputation Techniques with Machine Learning Algorithms

Hena Ahajjar, Ph.D. Student
C2SMART University Transportation Center | NYU Tandon School of Engineering



Introduction

Motivation

- Part of the challenge in prediction is the limited data availability on AV interest.
- Thus far, no studies have used machine learning techniques to impute data for AV interest and model the market share. However, numerous studies have used imputation to fill missing data in many fields including transportation.

Datasets

- ¡NYMTC Regional Household Survey: NY Metropolitan Area
- ¡Puget Sound Travel Survey: four-county region in Washington, which includes King, Kitsap, Pierce, and Snohomish counties.

Research Question

Can we predict the market share of AVs in the NY/NJ Metropolitan Area?

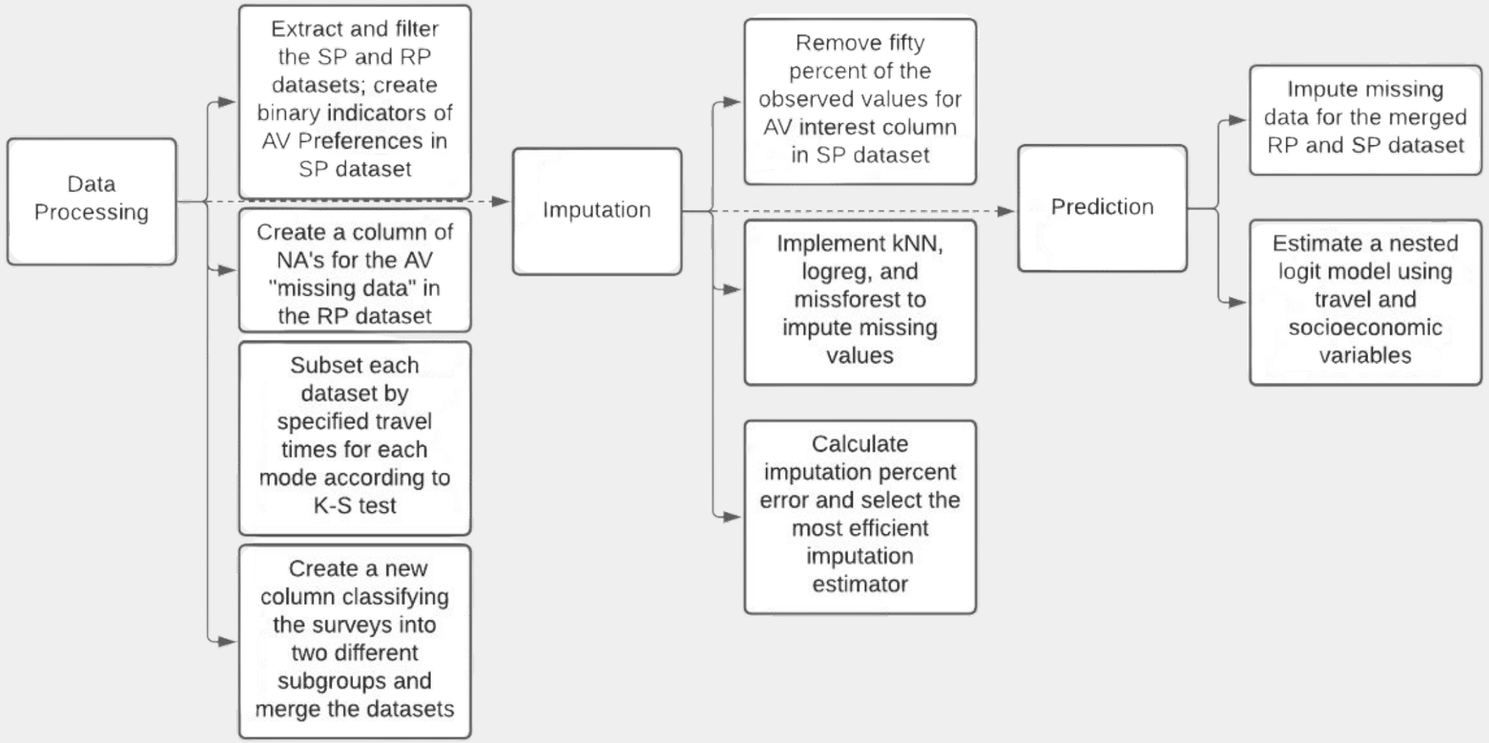
C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Methodology

A binary variable distinguishes NY/NJ metropolitan area and Washington State as two different subgroups.

The stated preference (SP) dataset, PGST, contains various questions on AV interest such as whether the respondent **will take an autonomous taxi, share a ride in an autonomous car, and own an autonomous car.**

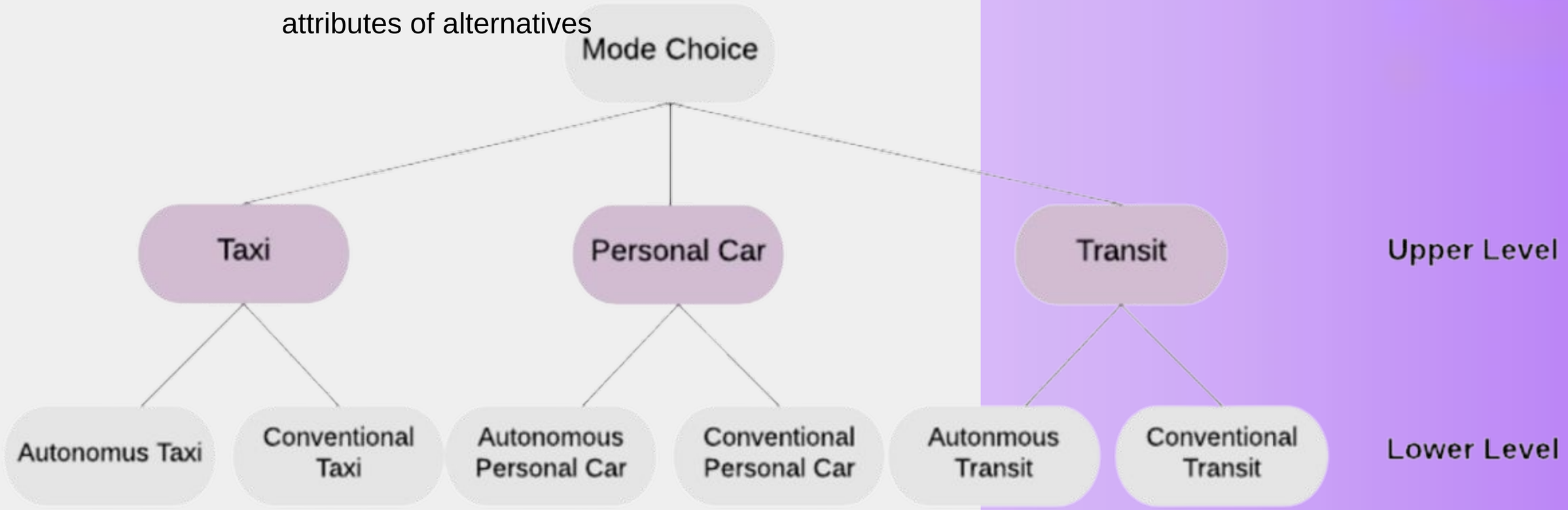
AV preference questions are not asked in the revealed preference (RP) survey which will be imputed using **logistic regression,**



C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Methodology

- A nested logit model is estimated to calculate the predicted market shares
- Method is based on **maximizing the utility function**, $V_{(n,i)}=\beta_1 X_{(n,i)}+\ldots+\beta_M X_{(n,i,M)}$, for each individual i and alternative n derived from the attributes of alternatives



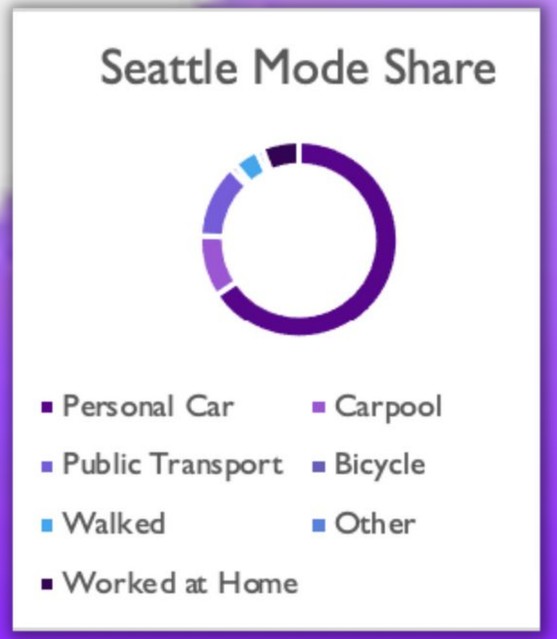
$$P_{n,\mathrm{B_k}} = \frac{\exp(Z'_{nk}\alpha+\lambda_k IV_{nk})}{\sum_{j\epsilon B_k}\exp(Z'_{nl}\alpha+\lambda_l IV_{nl})}$$

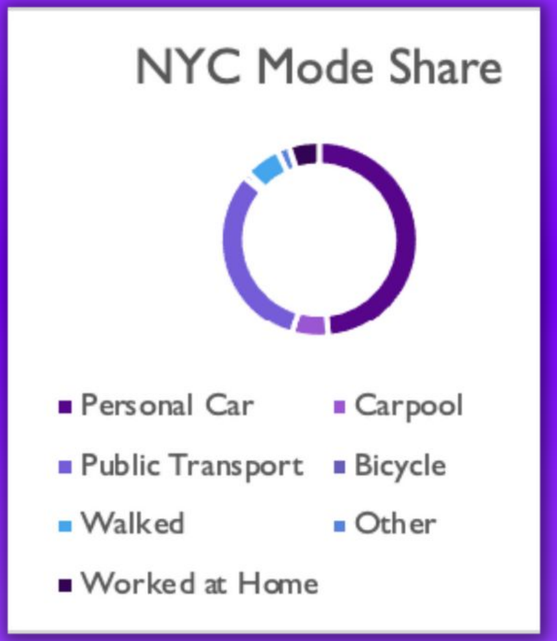
$$P_{ni|\mathrm{B_k}} = \frac{\exp(V_{ni}/\lambda_k)}{\sum_{j\epsilon B_k}\exp(V_{nj}/\lambda_k)}$$

NYC & Seattle Profiles

Summary Statistics (Census)*

Measure	New York City	Seattle
Average Yearly Congestion per Person (hours)[1 2]	133	138
Average Yearly Cost to traffic congestion per Person (dollars)[1 3]	1859	1932
Average One Way Commute Time (minutes)[4 5]	36	33
2015 Households without Vehicles[6]	54.5%	16.6%
Population[7]	8.175133 million	761,100
Job growth of 2018[8 9]	2.2%	2.4%
Average hourly wage[10 11]	$34.4	$27
Average daily VMT per capita [12]	15.4	25.8





*See slide 14 for references

C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Kolmogorov-Smirnov Test

1. Prior to combining the datasets, the Kolmogorov Smirnov test is used to prove whether the two datasets are statistically similar.
2. By sub-setting the full datasets into only commute trips, a common distribution is found between the two areas.
3. The null hypothesis states that there is no difference between the distribution functions of these two populations and can be seen as one population. If there is statistical significance, the null hypothesis is rejected and alternative hypothesis that there is a difference between the distribution for at least one x is accepted.

$$H_0: F(x) = G(x)$$
$$H_1: F(x) \neq G(x)$$

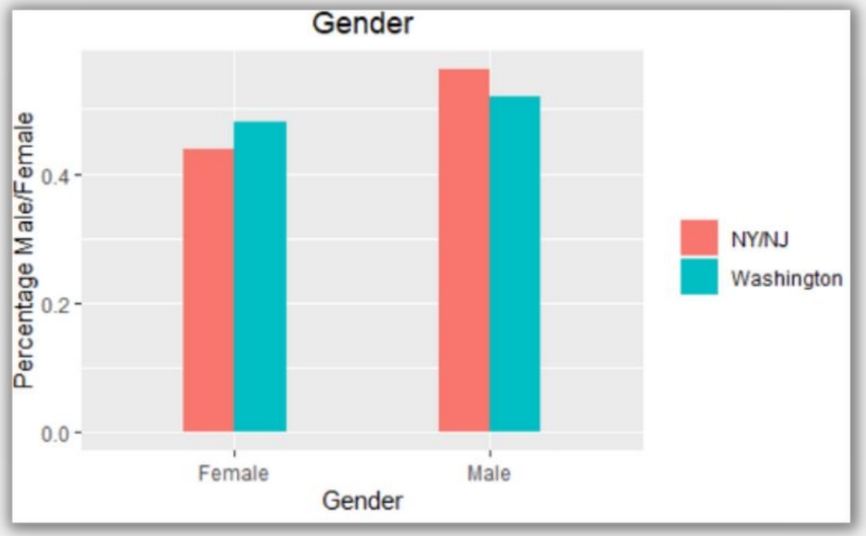
TABLE 1: Kolmogorov Smirnov Test for Long Commute Trips

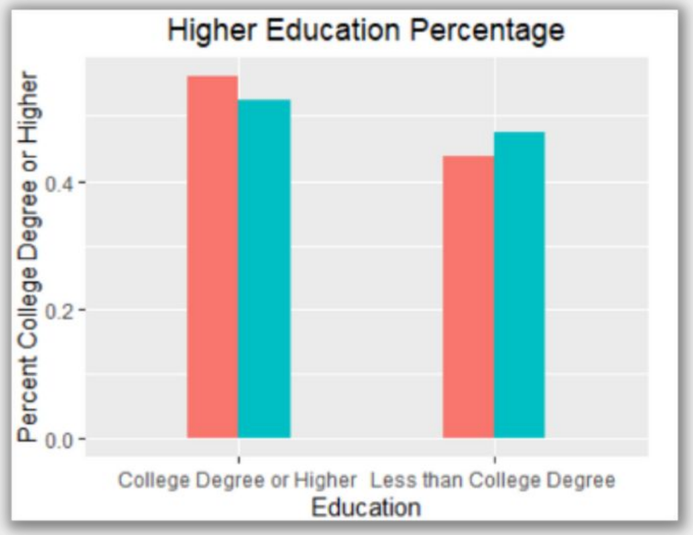
	Travel Time Period (Minutes)	D (Difference)	P Value
Car Trips	55 - 260	0.0733	.01
Taxi Trips	31 -120	0.0185	.019
Transit Trips	40 – 120	0.119	.119

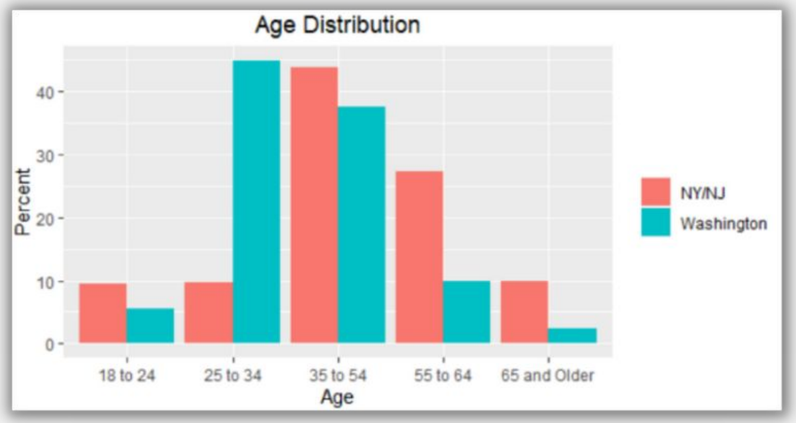
C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Long Commute Trips

Socio economic and Demographic Characteristics



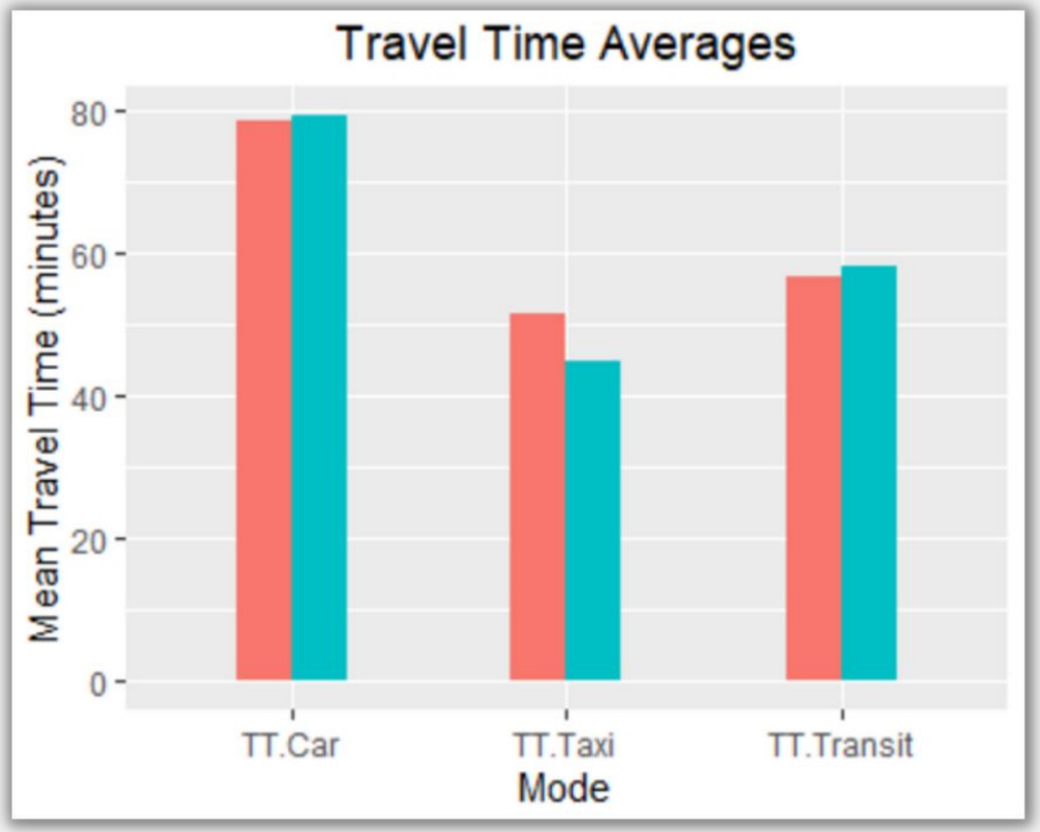


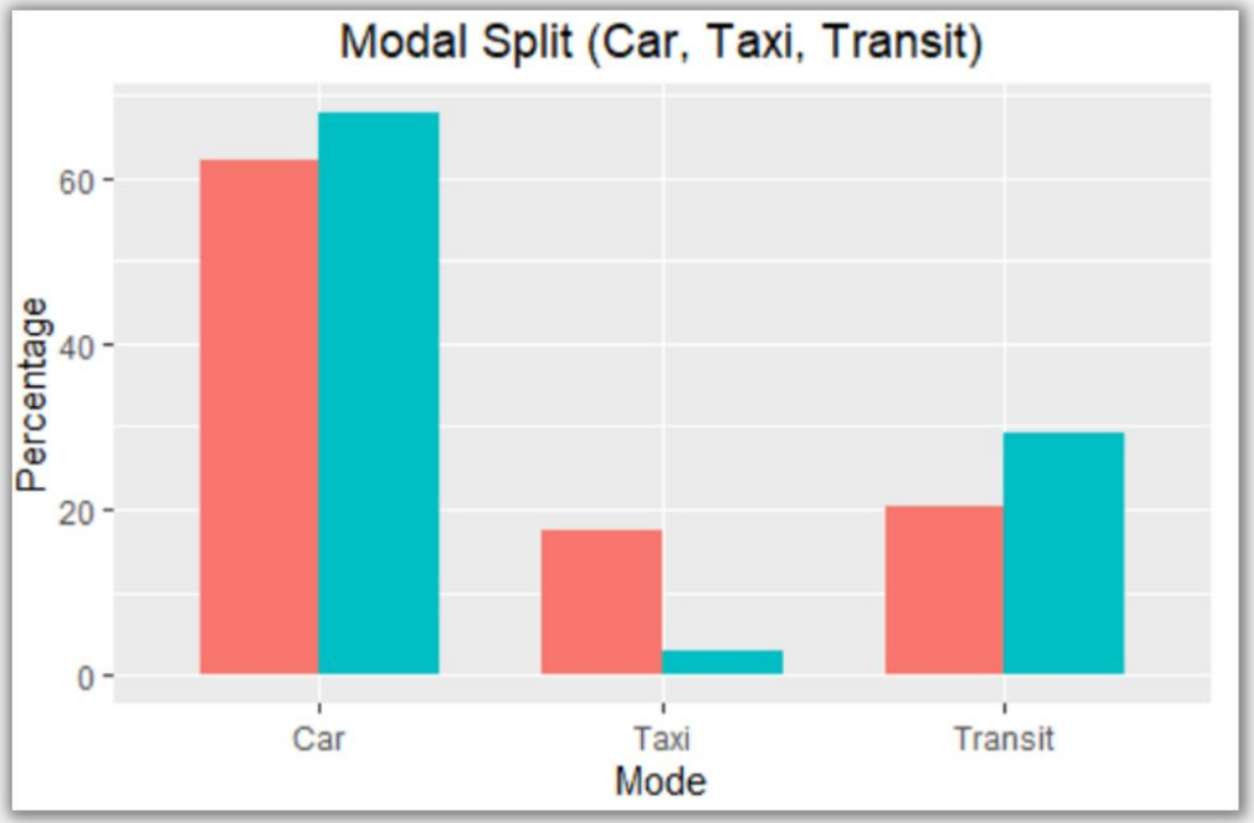


C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Travel Time Averages and Modal Split

Similar distributions are observed for travel time averages and modal split of the two regions.





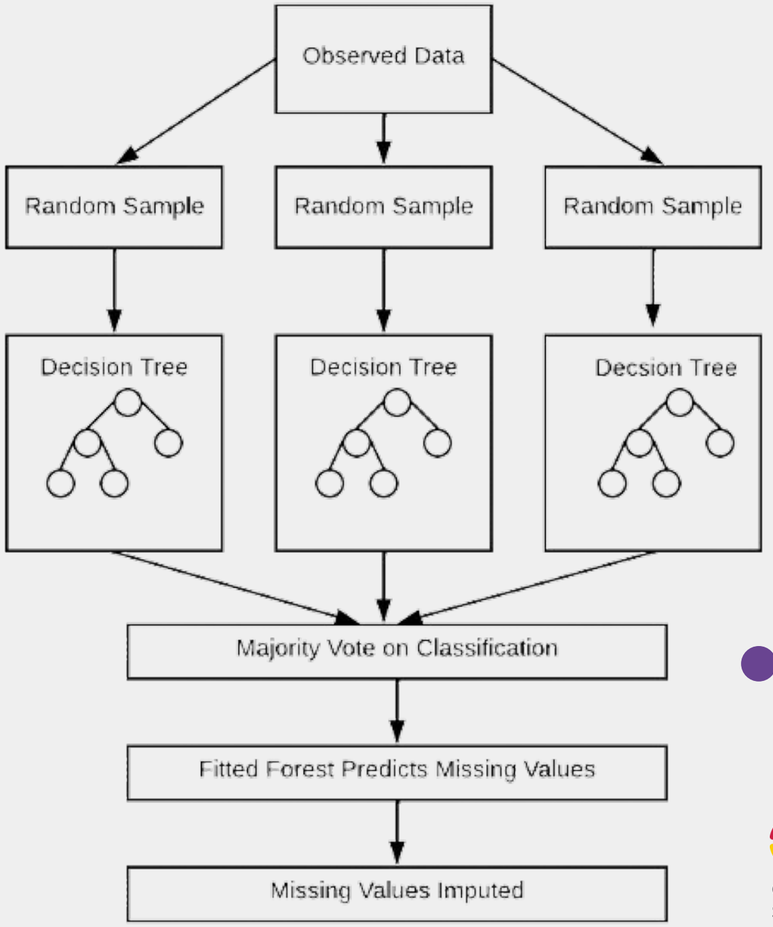
NYC

Washington

C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Imputation Methods

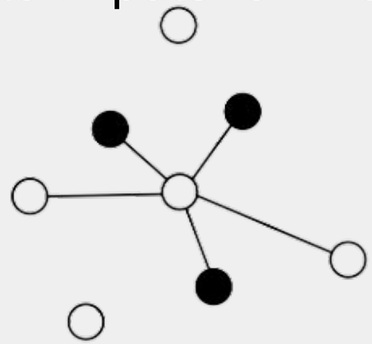
A random forest model (a forest of decision trees) for each variable is built using observed data, then missing values are predicted using the model



Random Forests

kNN

A distance metric is used to measure similarity between observed and the nearest neighbor (closest observation) is used to impute for missing values



Logistic regression

Fitting a logistic regression and then using the predicted values (to impute for missing data) based on other observed variables in the dataset

$$z_k = logit(p_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots + \beta_p x_p$$

C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Imputation Results

The results show that most travelers that adopt the autonomous vehicle technology will be using personal autonomous vehicles

The next highest percentage is transit, essentially autonomous buses and/or subway systems. These results are expected since most inner-city travelers depend on the transit system. In addition, millions of people commute everyday to NYC with personal vehicles making autonomous cars a convenient way of travel.

	kNN	Logistic Regression	Random Forests	kNN Commute	Logistic Regression - Commute	Random Forest - Commute
Personal Car	40.87	41.5	37.51	61.8	60	62
Taxi	4.5	4.6	5.13	17.2	2.8	17.6
Transit	8.69	9.4	9.57	.04	-	-
Probability of AV Adoption	54.06	55.5	52.21	79.04	62.8	79.6

Table 2: Imputation of AV Preferences for NY/NJ Metropolitan Area

Nested Logit Model Estimated Using kNN Imputation

Variable	Coefficient	P Value	Coefficient	P Value
	NY/NJ Commute Trips		NY/NJ Unsegmented Trips	
(Intercept):AVshare	-316.84	0.716	-1747.45	0.003
(Intercept):AVtaxi	24.53	$2.2*10^{-16}$	696.54	$2.2*10^{-16}$
(Intercept):Car	29.17	0.034	197.39	$2.2*10^{-16}$
(Intercept):Taxi	-28.17	0.204	692.68	$2.2*10^{-16}$
(Intercept):Transit	20.87	0.282	-7941.18	0.0001
Education	-35.47	0.012	-7.82	$2.2*10^{-16}$
Gender: Male	-55.31	0.0006	-0.88	$3.03*10^{-8}$
Hispanic	-1.09	0.598	2.27	$2.2*10^{-16}$
African	-15.33	0.0436	2.37	$2.2*10^{-16}$
White	-7.30	0.001	-1.14	$1.17*10^{-5}$
Asian	10.72	$3.94*10^{-6}$	-2.66	$2.86*10^{-12}$
Age				
16 to 24	3.00	0.440	0.52	0.095
25 to 34	-37.11	0.008	-0.12	0.705
35 to 54	-9.98	0.003	-0.49	0.052
55 to 64	7.46	0.0009	0.26	0.313
65 and older	-	-	-	-
Travel Time: Autonomous Personal Car	0.80	$8.78*10^{-11}$	0.77	$2.2*10^{-16}$
Travel Time: Autonomous Transit	1.26	0.243	-3.27	$2.2*10^{-16}$
Travel Time: Autonomous Taxi	0.19	0.044	-1.85	$2.2*10^{-16}$
Travel Time: Conventional Personal Car	-3.42	0.0008	-3.90	$2.2*10^{-16}$
Travel Time: Conventional Taxi	0.45	0.039	-1.39	$2.2*10^{-16}$
Travel Time: Conventional Transit	0.43	$1.24*10^{-6}$	2.60	$2.2*10^{-16}$

C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Market Penetration Results

C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Table 3: Nested Logit Modal Split Predictions

Mode	kNN Prediction for NY/NJ	RF Prediction for NY/NJ	kNN Prediction for NY/NJ (Commute)	Washington Area
Conventional Personal Car	43	42.6	0.72	28.72
Autonomous Personal Car	36.8	37.6	61.3	60.21
Conventional Taxi	3.4	1.5	2.7	.25
Autonomous Taxi	3.3	5.1	14.9	1.20
Conventional Transit	4	3.7	20.3	2.38
Autonomous Transit	9.5	9.6	.1	7.19

Table 4: Imputation Percent Error

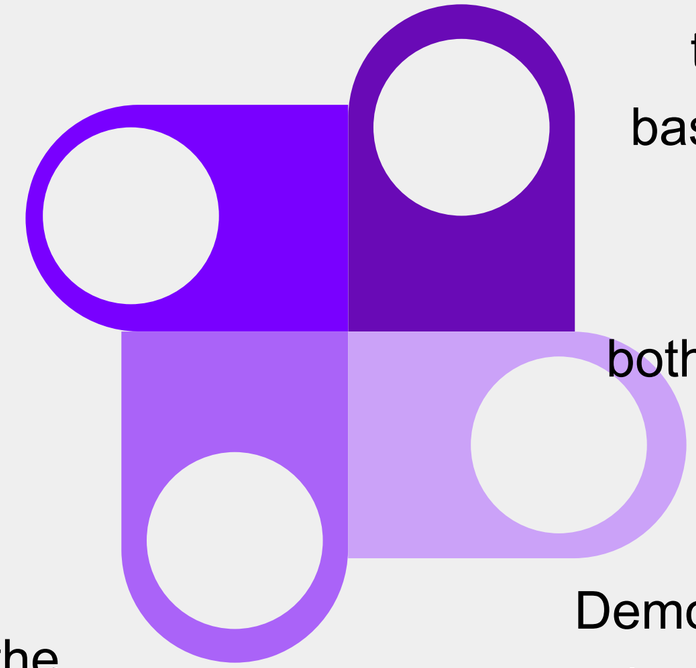
	kNN	Random Forests	kNN Commute	Random Forest - Commute
Imputation Percent Error	29	0	12	3.3

- Interestingly, similar percentages (for all NY/NJ trips) for the two algorithms are observed, 36.8 and 37.6 percent adoption autonomous personal car
- These results can be compared to a nested logit model's prediction for only the Washington in which no imputation was done.
- As expected, the results in this study show that Personal Car would make most of the trips, approximately 61.3% for commute trips.
- Although RF resulted in lower error, kNN demonstrated percent adoption for

Conclusions and Future Work

Overall, the results for both algorithms suggest high statistical significance based on the high t-values according to a 95% confidence interval. The McFadden R squared for the kNN and RF approaches are .28 and .19 respectively

Assuming the individual performances of RF and kNN imputation are both efficient, the percent adoption of AVs is about 50-52% when AVs are introduced into the mainstream marketplace (for



A nested logit model for the Puget Sound region would have predicted 68% of travelers to take an autonomous mode based on the dataset. This is significantly different than the estimated percent adoption predicted for NY/NJ area for both models (all trips – 50% and commute trips – 76.3%)

Demonstrate the validity of this study by conducting and distributing a survey study on commute trips in and out of NYC

Read more at c2smart.engineering.nyu.edu

References

- Manskar, Noah. “Only 3. U.S. Cities Have Worse Traffic Than NYC.” New York City, NY Patch, Patch, 12 Feb. 2019, patch.com/new-york/new-york-city/only-3-u-s-cities-have-worse-traffic-nyc.
- Did you lose 138 hours to Seattle-area traffic last year? *Seattle Times.* [Online] February 11, 2019. https://www.seattletimes.com/seattle-news/transportation/seattle-area-traffic-remains-sixth-most-congested-among-major-u-s-cities/.
- Seattle ranks 2nd worst commute time in US, report finds. *Komo News.* [Online] December 1, 2019. https://komonews.com/news/local/spend-a-lot-of-time-commuting-seattle-ranks-2nd-worst-in-us-report-finds.
- Hess, Abigail. “The 10 Cities with the Worst Commutes, According to U.S. News.” CNBC, CNBC, 9 Apr. 2019, www.cnbc.com/2019/04/09/the-10-cities-with-the-worst-commutes-according-to-us-news.html.
- Wynkoop, Gena. The average Seattle commute time actually seems. reasonable. [Online] May 10, 2019. http://seattlerefined.com/lifestyle/this-is-the-average-commute-time-for-seattleites.
- Governing. “Vehicle Ownership in U.S. Cities Data and Map.” Governing, www.governing.com/gov-data/car-ownership-numbers-of-vehicles-by-city-map.html.
- “U.S. Census Bureau QuickFacts: New York City, New York; Boston City, Massachusetts.” Census Bureau QuickFacts, www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bostoncitymassachusetts/PST045218.
- Thomas P. DiNapoli, Kenneth B. Bleiwas. *New York City Employment Trends.* NYC : Office of the New York State Comptroller, 2019.
- 10 useful facts and stats: the current state of the Seattle job market. [Online] [Cited: 11 13, 2020.] https://www.amaxra.com/10-useful-facts-and-stats-the-current-state-of-the-seattle-job-market#:~:text=ERFC%20economists%20predict%20the%20state's,of%20the%20state's%20net%20increase.
- Nbc. “How Much Do New Yorkers Have to Earn an Hour to Afford Rent?” NBC New York, NBC New York, 19 June 2018, www.nbcnewyork.com/news/local/This-is-How-Much-New-Yorkers-Have-to-Earn-Per-Hour-to-Rent-an-Apartment-Housing-Low-Income-Expensive-Affordable-Money-Cost-New-York-City-NY-NYC-485862001.html.
- Study: $27 hourly wage needed to afford 2-bedroom apartment in Wash. state. *KOMO Staff.* [Online] June 13, 2018. https://komonews.com/news/local/study-27-hourly-wage-needed-to-afford-2-bedroom-apartment-in-wash.
- State & Urbanized Area Statistics. *FHA.* [Online] [Cited: 11 15, 2020.] https://www.fhwa.dot.gov/ohim/onh00/onh2p11.htm.

C2SMART

CONNECTED CITIES WITH SMART TRANSPORTATION

c2smart.engineering.nyu.edu

c2smart@nyu.edu

Hella Alnajjar, PhD Candidate

C2SMART Center

New York University

Tandon School of Engineering

6 MetroTech Center, Brooklyn, NY 11201

kaan.ozbay@nyu.edu

Thank you!