ISyE 6740 – Spring 2022
Final Report

Team Member Names: Ikram Laaguidi, Kelly Lam, Levana Zhang

Project Title: Modeling Electric Vehicle Adoption in the United States

Problem Statement

In 2019, the Transportation sector accounts for 29% of total U.S. greenhouse gas emissions and light-duty Vehicles make up 58% of transportation emissions (EPA 2019). The adoption of electric vehicles (EVs) is a critical strategy to reduce harmful pollution. This project will demonstrate the factors that could help policymakers and utilities to target and support all income levels families to rapidly adopt EV. Early research on EV adoption focused on vehicle costs, range anxiety, charging time and economic growth. However, there is a lack of understanding consumers and adopters' decision-making patterns. This research is critical to identify the motivating factors that could optimize and allocate funds/incentives and provide support to disadvantaged communities and hesitators to adopt eco-friendly technology.

The motivation of this problem aims to shed light on identifying variables at the personal level, demographics level and socio-economic level to help develop rigorous models to classify adopters/non-adopters.

Data Source

The selected dataset for this project was collected to study the inclination of individuals towards adopting EVs and was collected through mail and online surveys in the United States. We used The National Household Travel Survey 2017 (NHTS) which provides information to assist transportation planners and comprehensive data on travel patterns in the United States. The survey collected data from 150,147 households with 115 columns including person, vehicle, daily trip, income, technology usage, costs of fuel, type of fuel, etc.

The data consisted of four different files: Household, Person, Trip, and Vehicle. The household data contained variables regarding homeownership, household income, household size, etc. and often had coinciding data with Vehicle such as the type and number of vehicles in a household. The Vehicle data had our dependent variable “HFUEL” which denotes which type of fuel a hybrid car uses; for instance, hybrid, plug-in hybrid, or purely electric. According to the 2017 NHTS, only 2.04% of the vehicles are electric vehicles. The survey had imbalanced data which generated biased classification and overfitting. One approach we used to address the imbalanced dataset is oversampling the electric vehicle category (the minority) in HFUEL using SMOTE. SMOTE (Synthetic

Minority Oversampling Technique) is a type of data augmentation by duplicating the minority data in the training set prior to fitting the model.

Person-related variables include their level of education or working status as well as occupation type. On the other hand, Trip data consisted of the "purpose of trip", mileage, and type of transportation used such as bus, walking, or car.

The majority of the data type was numerical and there was also a significant amount of categorical variables that used numerical bins. However, within the numerical category we had values ranging from -9 up to the millions. Oftentimes the values were given bins, for instance, household income would have values 1 to 11 which denoted ranges like less than $10,000 annually or between $10,000 and $14,999. The data also contained values such as -9, -8, -7, and -1 to denote "Not ascertained", "I don't know", "I prefer not to answer", and "Appropriate skip", respectively. Many of the variables were repeats of others that drilled down to become more granular. For example, we had variables which labeled the state the household was in, but we also had variables that grouped the household in census division as well as census region. This meant that we had to use our own judgment when pre-filtering features to include in our final dataset for modeling.

Methodology

Data Wrangling
The dataset provided by the National Household Travel Survey contains hundreds of variables that needed to be decoded then downsized to properly train our models. Each of the members of this team was responsible for a dataset in which we did our own exploratory data analysis. Thus, we combed through each variable and read its description in the documentation codebook supplied by the NHTS to jointly come to a conclusion of whether to include or exclude the variable to use for modeling.

We also found that there were numerous variables with the values -9, -8, -7, and -1 which accounted for more than 10% of the total number of rows. Rather than imputing, we decided to remove these variables from our model. With variables that contained less than 10% of -9, -8, -7, and -1, we removed the rows from our data. We also mapped values of yes's and no's; within the data twos were coded as 'no' and ones were coded 'yes', therefore, we replaced twos with zeros and kept the ones.

We then merged four of our preprocessed files using HouseID, PersonID, and VehID. The data was already in the format of label-encoding, so what was left to do was find categorical variables to one-hot encode. We ran feature selection using Random Forest and followed up with a correlation matrix to get rid of highly correlated variables.

Feature Selection

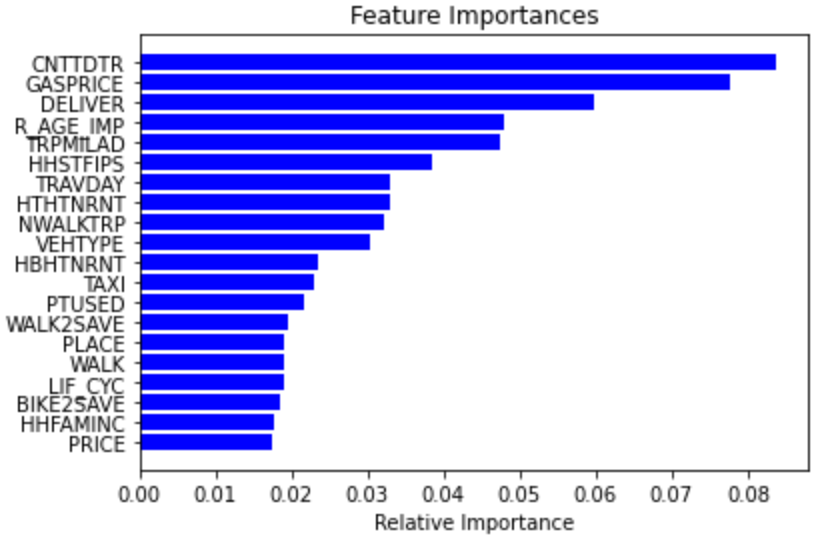
We used feature selection techniques such as manually selecting for important features, the correlation matrix and Random Forests to identify which variables at the personal, demographics, and socio-economic level contribute most significantly to an individual's

decision to adopt electric vehicles or not, as well as simplifying the model to improve modeling runtime.

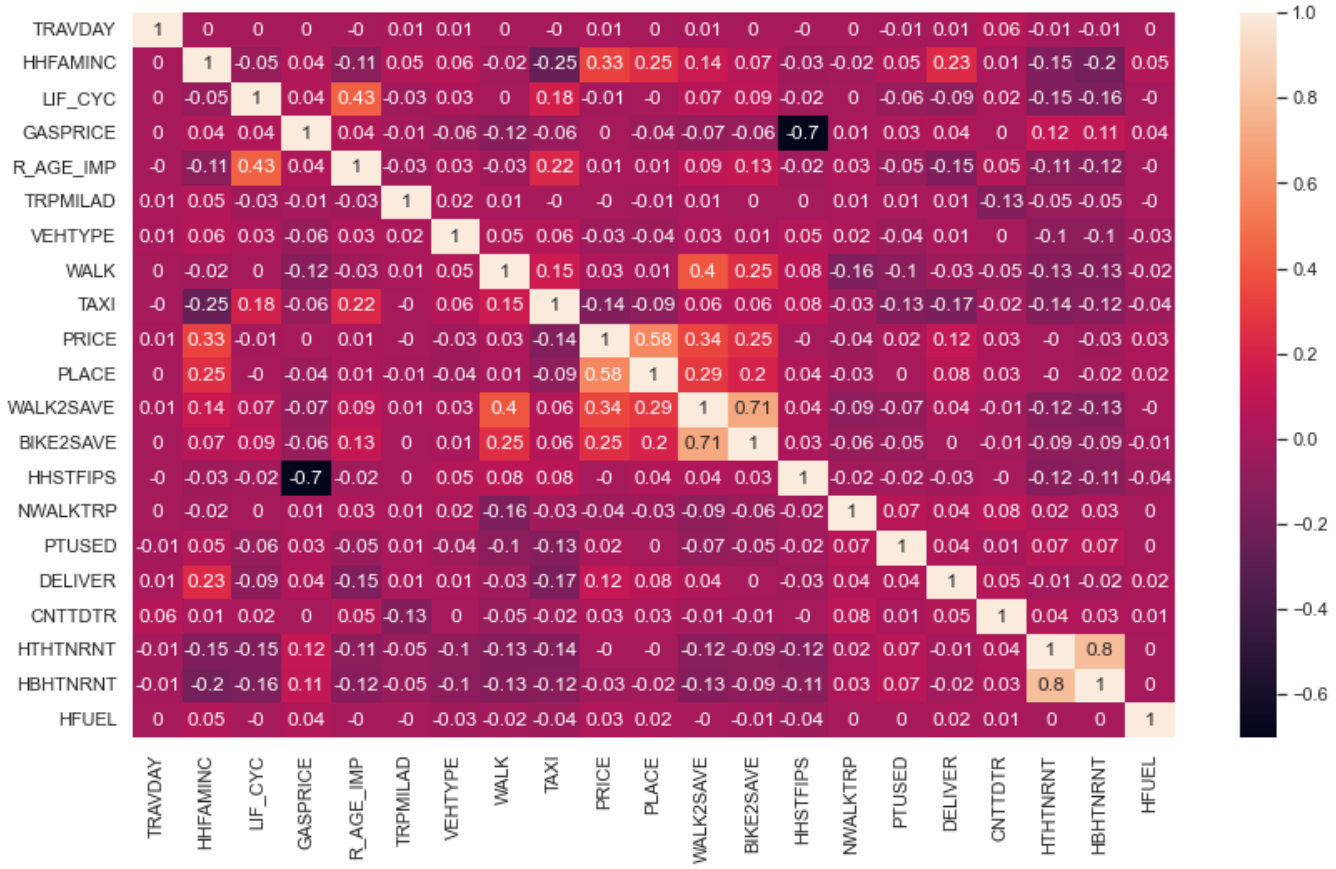
We chose to use Random Forests as our main algorithmic feature selection tool because many of our features were either categorical or label-encoded, and Random Forests can still function effectively as an algorithm for non-one-hot-encoded categorical or label-encoded features. We wanted to avoid one-hot-encoding because feature selection algorithms tend to separate one-hot-encoded variables, and LASSO required one-hot-encoding. We chose Random Forests over Lasso because we did not need to one-hot-encode our categorical variables for Random Forests, while we did for LASSO.

To optimize the Random Forests algorithm that we would use for feature selection, we conducted some hyperparameter tuning using cross-validation to find the optimal hyperparameter combination. The hyperparameters we tuned were the number of trees: 10, 50, 100, 150, 300 trees and the maximum tree depth: 3, 5, 10, 15, 20, and we used MSE as the metric to evaluate and compare our Random Forests models on. We ultimately settled on 100 trees and a maximum tree depth of 10 as our optimal hyperparameter combination in terms of model quality and speed in model fitting.

When using Random Forests for feature selection, we used scikit-learn's Random Forests package's feature_importances_ attribute to determine which feature(s) improves the model, and hopefully the model prediction quality, the most. We chose to keep the top 20 most significant features according to this metric to form the basis of our modeling dataset.



For the final phase of feature selection, we chose to use the correlation matrix to remove highly correlated features. Highly correlated features essentially give the same information to our models, thus removing one of a pair of highly correlated features will help reduce the feature space, and thus improve model quality, training time, and interpretability, while still preserving information gained by the model. We chose to remove one feature at random for pairs of features with a correlation coefficient of 0.5 or higher.



Below are the variables outputted after feature selection.

Variable	**Label**	**Code/Range**
WALK	Frequency of Walking for Travel	-9 = Not ascertained -8 = I don't know -7 = I prefer not to answer 01 = Daily 02 = A few times a week 03 = A few times a month 04 = A few times a year 05 = Never
TAXI	Frequency of Taxi Service or Rideshare Use for Travel	Same as above
WALK2SAVE	Walk to Reduce Financial Burden of Travel	Same as above

R_AGE_IMP	Age (imputed)	Responses = 5 to 92
LIF_CYC	Life Cycle classification for the household, derived by attributes pertaining to age, relationship, and work status.	-9 = Not ascertained 01 = one adult, no children 02 = 2+ adults, no children 03 = one adult, youngest child 0-5 … and so on
TRPMILAD	Trip distance in miles, adjusted for comparability to past	-9 = Not ascertained Responses = 0 to 9,621.053
TRAVDAY	Travel day - day of week	01=Sunday 02=Monday 03=Tuesday … and so on
PTUSED	Count of Public Transit Usage	-9=Not ascertained -8=I don't know -7=I prefer not to answer Responses = 0 to 30
VEHTYPE	Vehicle Type	-8=I don't know -7=I prefer not to answer 01=Automobile/Car/Station Wagon 02=Van (Mini/Cargo/Passenger) 03=SUV (Santa Fe, Tahoe, Jeep, etc.) 04=Pickup Truck 05=Other Truck 06=RV (Recreational Vehicle) 07=Motorcycle/Motorbike 97=Something Else
PRICE	Price of Gasoline Affects Travel	-9 = Not ascertained -8 = I don’t know -7 = I prefer not to answer 01 = Strongly agree 02 = Agree 03 = Neither Agree or

		Disagree 04 = Disagree 05 = Strongly disagree
NWALKTRP	Count of walk Trips	-9 = Not ascertained -8 = I don’t know -7 = I prefer not to answer Responses = 0 to 200
HTHTNRNT	Category of the percent of renter-occupied housing in the census tract of the household's home location	-9 = Not ascertained 0 = 0-4% 05 = 5-14% … and so on
HHFAMINC	Household income	-9 = Not ascertained -8 = I don’t know -7 = I prefer not to answer 01 = Less than $10,000 02 = $10,000 to $14,999 … and so on
DELIVER	Count of Times Purchased Online for Delivery in Last 30 Days	-9 = Not ascertained -8 = I don’t know -7 = I prefer not to answer -1 = Appropriate skip Responses = 0 to 99
CNTTDTR	Count of person trips on travel day	Range 0 to 50
GASPRICE	Price of gasoline, in cents, on respondent's travel day	Responses = 201.3 to 295.9

One-Hot-Encoding
We chose to one-hot-encode our categorical variables after feature selection from Random Forests because we were concerned that Random Forests would split up a categorical variable’s one-hot-encoded sub-variables through its decision trees, as one-hot-encoded variables were considered by Random Forests to be variables of their own right. We were concerned about the loss of model informativeness resulting from Random Forests focusing on a few sub-variables instead of the categorical variable as a whole. Thus to prevent this information loss, we one-hot-encoded our categorical variables after using Random Forests for feature selection.

Scaling
The final step of our preprocessing was to scale our final independent variables using MinMaxScaler and export the final dataset to use for modeling.

K-fold Cross Validation
To avoid any bias when splitting the data into training and testing sets, we opted to use k-fold cross-validation with f1-score as the evaluation metric to compare models with different hyperparameters before settling on the optimal hyperparameters for each of the four models we ultimately trained.

Modeling

We decided to build four machine learning models and evaluate its ability to correctly classify adopters and non-adopters i.e. HFUEL.

Logistic regression

Confusion Matrix		
	0	1
0	73745	20681
1	68	166

Classification Report - Confusion Matrix			
	Precision	Recall	f1-score
0	1.00	0.78	0.88
1	0.01	0.71	0.02
Classification Report - Accuracies			
Accuracy			0.78
Macro Average	0.50	0.75	0.45
Weighted Average	1.00	0.78	0.87

AUC-ROC Score	0.75

Support Vector Machine: Linear Kernel

Confusion Matrix		
	0	1
0	94426	0
1	234	0

Classification Report - Confusion Matrix			
	Precision	Recall	f1-score
0	1.00	1.00	1.00
1	0.00	0.00	0.00
Classification Report - Accuracies			
Accuracy			1.00
Macro Average	0.50	0.50	0.50
Weighted Average	1.00	1.00	1.00

AUC-ROC Score	0.5

Support Vector Machine: Radial Basis Function Kernel

Confusion Matrix		
	0	1
0	46969	47457
1	65	169

Classification Report - Confusion Matrix			
	Precision	Recall	f1-score
0	1.00	0.50	0.66
1	0.00	0.72	0.01
Classification Report - Accuracies			
Accuracy			0.50
Macro Average	0.50	0.61	0.34
Weighted Average	1.00	0.50	0.66

AUC-ROC Score	0.61

Decision Tree

Confusion Matrix		
	0	1
0	93650	776
1	31	203

Classification Report - Confusion Matrix			
	Precision	Recall	f1-score
0	1.00	1.00	1.00
1	0.21	0.87	0.33
Classification Report - Accuracies			
Accuracy			0.99
Macro Average	0.60	0.93	0.67
Weighted Average	1.00	0.99	0.99

AUC-ROC Score	0.930

K-Nearest Neighbors

Confusion Matrix		
	0	1
0	94403	23
1	15	219

Classification Report			
	Precision	Recall	f1-score
0	1.00	1.00	1.00
1	0.90	0.94	0.92
Classification Report - Accuracies			
Accuracy			1.00
Macro Average	0.95	0.97	0.96
Weighted Average	1.00	1.00	1.00

AUC-ROC Score	0.967

Evaluation and Final Results

Model Evaluation

We considered several methods to evaluate our models such as accuracy, confusion matrix, F-1 score, and a quick running time to train. Through deeper analysis of our data, we found that the data was heavily imbalanced and comprised more true values of non-adapters than adapters. Our models reported high accuracies due to the imbalance of the true values. With a closer look at precision, we can see how accurate our model is at predicting and with recall we are able to have a sense of how often our models are able to

correctly identify non-adapters and adapters within the given dataset. On the other hand, the ROC is a function of classification thresholds and the AUC gives us a performance measure across possible thresholds. Our models produce moderate to good AUC-ROC scores, but low F-1 scores. We can interpret these results to mean that our classifiers are currently underperforming and can benefit from more fine-tuning to find the optimal threshold that should provide a better F-1 score.

Conclusion

The 2017 NHTS data was utilized to explore features that could impact electric vehicles adoption. With extremely imbalanced data, SMOTE was chosen to adjust the original data after feature selection was implemented. In this project, logistic regression, support vector machine, decision tree, and K-nearest neighbors were performed to build a prediction model. The model performance indicated that the K-nearest neighbors is the best model for our categorical variables, as its ROC-AUC score is 0.967. From socio-economic and demographic level; house or car ownership related to high income families which demonstrated the ability to purchase electric vehicles and can be adopters, whereas the low income are using public transportation, walking, or car sharing due to the high cost of fuels (gas or electricity). The middle income family represents 0.1% of total shared electric vehicle adopters in the NHTS survey. Due to the lack of data, the research is still needed to impact policy makers on supporting and advancing adoption of electric vehicles to reduce emissions.

Collaboration Breakdown

- Ikram Laaguidi - Data Collection, Project Proposal, Data wrangling (Vehicle and Household data), SVM (RBF and Linear), Final Report
- Kelly Lam - Project Proposal, Data wrangling (Trip data), Data Preprocessing, Logistic Regression, Model evaluation, Final Report
- Levana Zhang - Project Proposal, Data wrangling (Person data), Algorithmic Feature Selection, KNN and Decision Tree, Final Report

Reference:

https://nhts.ornl.gov/documentation
https://nhts.ornl.gov
https://nhts.ornl.gov/2009/pub/UsersGuideV2.pdf