Feb
2022

CAPSTONE TWO - FINAL PROJECT REPORT

SPRINGBOARD
KANCHANAH KANNATHASS

PROBLEM STATEMENT

Based on the recent announcement by President Biden on his goal towards the reduction of greenhouse emission by 50-52% by the year 2030, there will be a push for the government agencies at the state levels to attain these targets. One way of achieving this goal would be to promote the usage of electric vehicles across different use cases. While electric vehicles still require electricity to function, this power can be provided by cleaner alternate sources of energy. Electric Vehicles, as such, do not emit greenhouse gases in the immediate vicinity where they run as compared to an internal combustion engine (ICE) vehicle.

Based on the results of the analysis done for this project, government agencies can focus on the factors that were deemed to be important in people not adopting EV and focus more on developing policies to promote the adoption of EV based on these factors. These policies can include things such as educational policies, car sharing, and price of vehicle.

DATASET

The dataset that will be used for this project will be the 2017 National Household Travel Survey (NHTS). This survey is conducted by the federal administration (FHWA). It includes data that allows one to analyze trends in personal and household travel. It contains information on (1) daily travel linked to (2) individual personal and household characteristics, (3) socio-economic characteristics, (4) vehicle ownership and vehicle attributes.

Due to the collection on different days, the daily travel dataset was removed from the analysis. The remaining three datasets were merged to conduct the analysis.

DATA WRANGLING

The merged datasets had 256115 rows and 21 variables. Although the initial merged dataset had over 200 variables, these 21 variables were chosen based on what was most appropriate for the analysis.

VARIABLE TYPE

Since some of the variable types were incorrect, they were corrected. For example, some variables were supposed to be numeric but were provided in character format.

DROP DUPLICATE CAR INFORMATION

One respondent may have more than one vehicle which can cause the classification process to produce incorrect results. This was addressed by keeping the information for the latest vehicle and discarding the information for other older vehicles.

SPECIAL CODES

The codes provided in Table 1 were special codes used in the survey. Since there were multiple variables with these codes, the codes were converted to NA for the purpose of this project.

Variable Value	Label
-1	Appropriate Skip
-9	Not Ascertained
-7	I prefer not to answer (Selected by participant (available when no answer given)
-77	I prefer not to answer (Selected by participant (always available)
-8	I don't know (Selected by participant (available when no answer given)
-88	I don't know (Selected by participant (always available)

Table 1: Special Codes

RESPONSE VARIABLE

The response variable used for this analysis is FUELTYPE. The values for this variable are given in Table 2.

FUELTYPE Label
Gas
Hybrid, Electric OR Alternative fuel
Diesel
Some other fuel

Table 2: Value labels for Fueltype variable

Gas, Diesel and Some other fuel were recoded to non-EV. Hybrid, Electric or Alternative fuel were recoded to EV.

RECODING VARIABLES

Some of the ordinal variables were recoded to depict ordinality and the nominal variables were recoded to reduce cardinality

- HBHUR

Initial Variable Value	Recoded Variable Value
Rural	Rural
Small Town	Not Rural
Suburban	Not Rural
Second City	Not Rural
Urban	Not Rural

Table 3: Recoding for HBHUR

*The definition of variable names are given in the Appendix.

- HHSTATE

Initial Variable Value	Recoded Variable Value
WA	Pacific
OR	
CA	
AK	
HI	
NV	Mountain
ID	
MT	
WY	
UT	
CO	
AZ	
NM	
ND	West North Central
SD	
MN	
NE	
IA	
KS	
MO	
WI	East North Central
MI	
IL	
IN	
OH	

Initial Variable Value	Recoded Variable Value
NY	Middle Atlantic
PA	
NJ	
VT	New England
NH	
ME	
MA	
CT	
RI	
TX	West South Central
OK	
AR	
LA	
KY	East South Central
TN	
MS	
AL	
WV	South Atlantic
MD	
DE	
DC	
VA	
NC	
SC	
GA	
FL	

Table 4: Recoding for HHSTATE

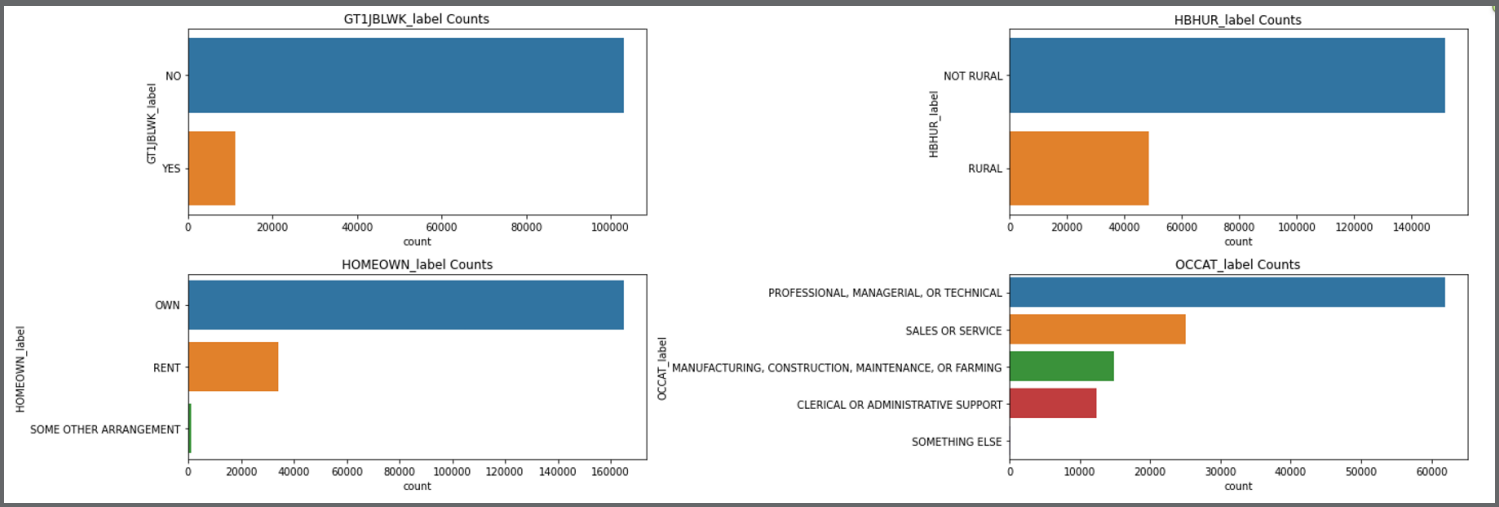
EXPLORATORY DATA ANALYSIS

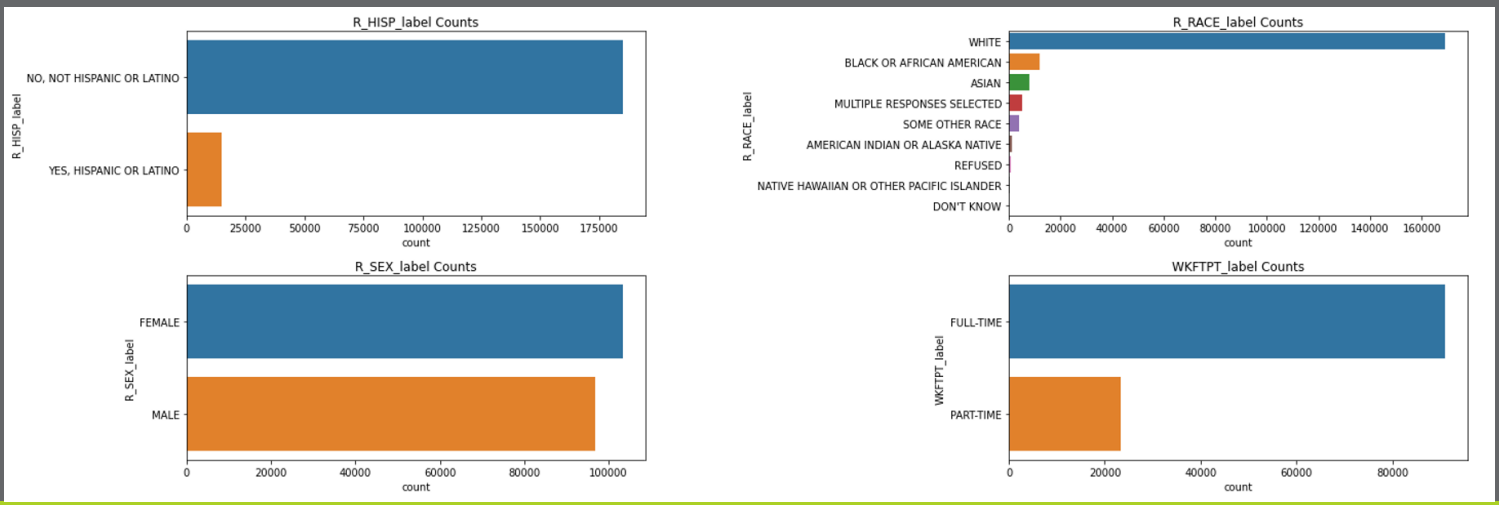
The variables could be divided into the following different types:

- Categorical
  - Nominal
  - Ordinal
- Numerical

CATEGORICAL - NOMINAL VARIABLES

The distribution of each nominal variable used for analysis is given in Figures 1.





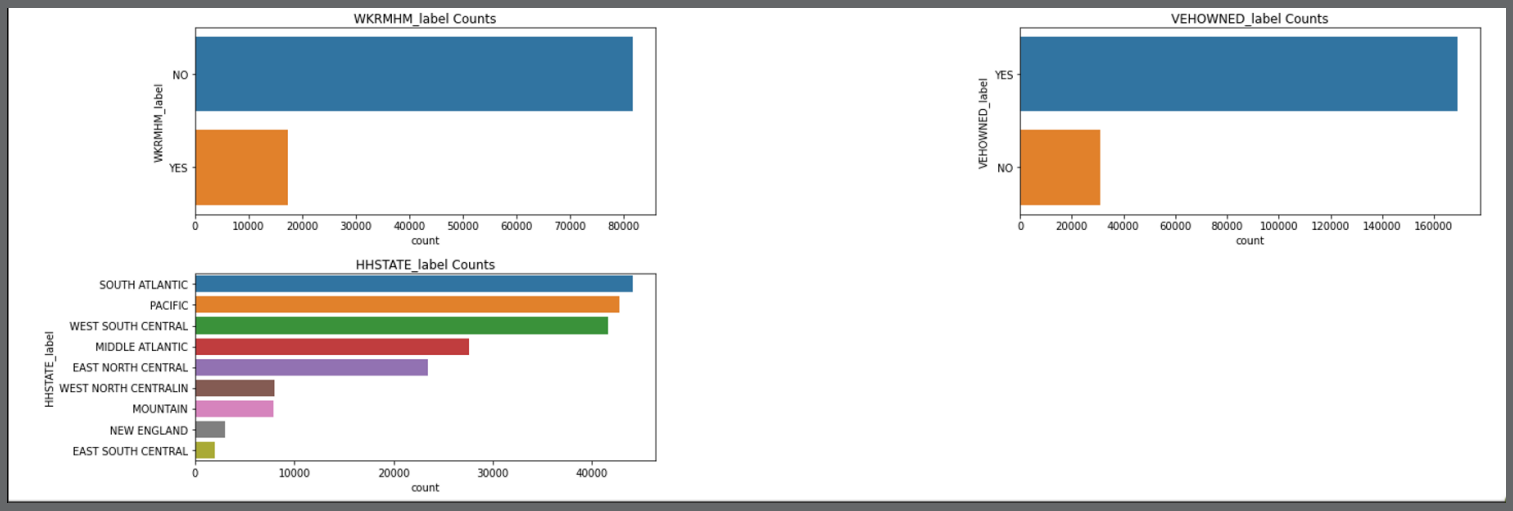
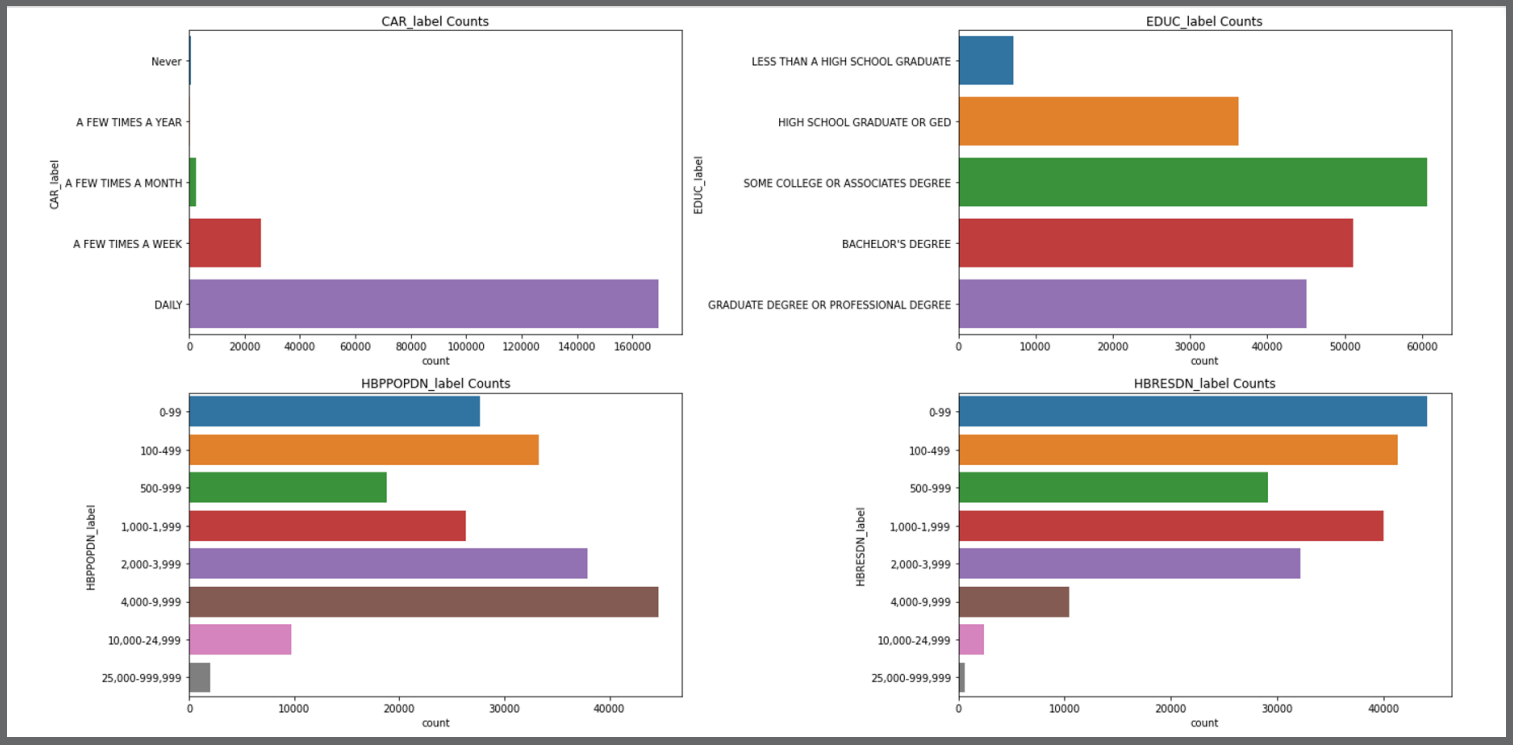


Figure 1 : Distribution of Nominal Variables

CATEGORICAL - ORDINALVARIABLES

The distribution of each ordinal variable used for analysis is given in Figures 2.



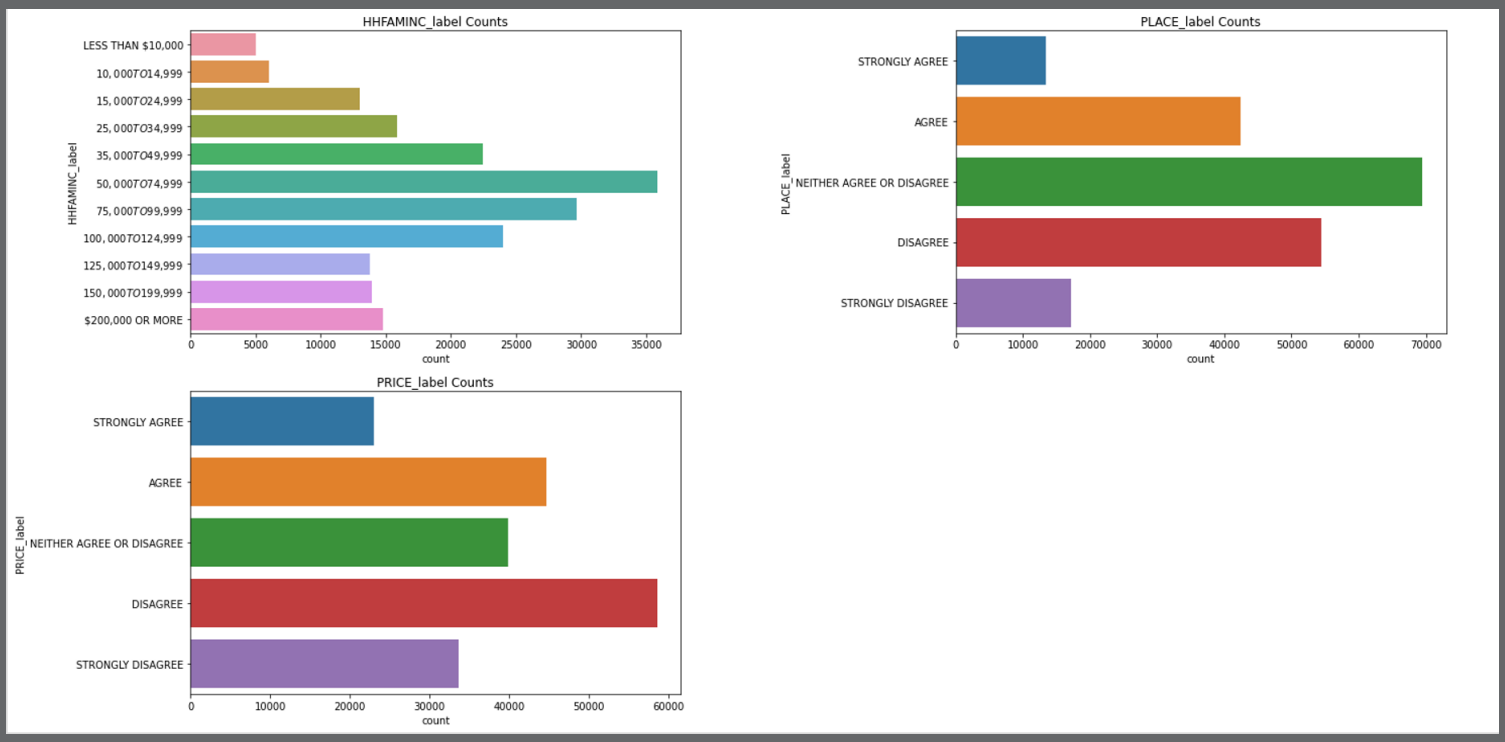
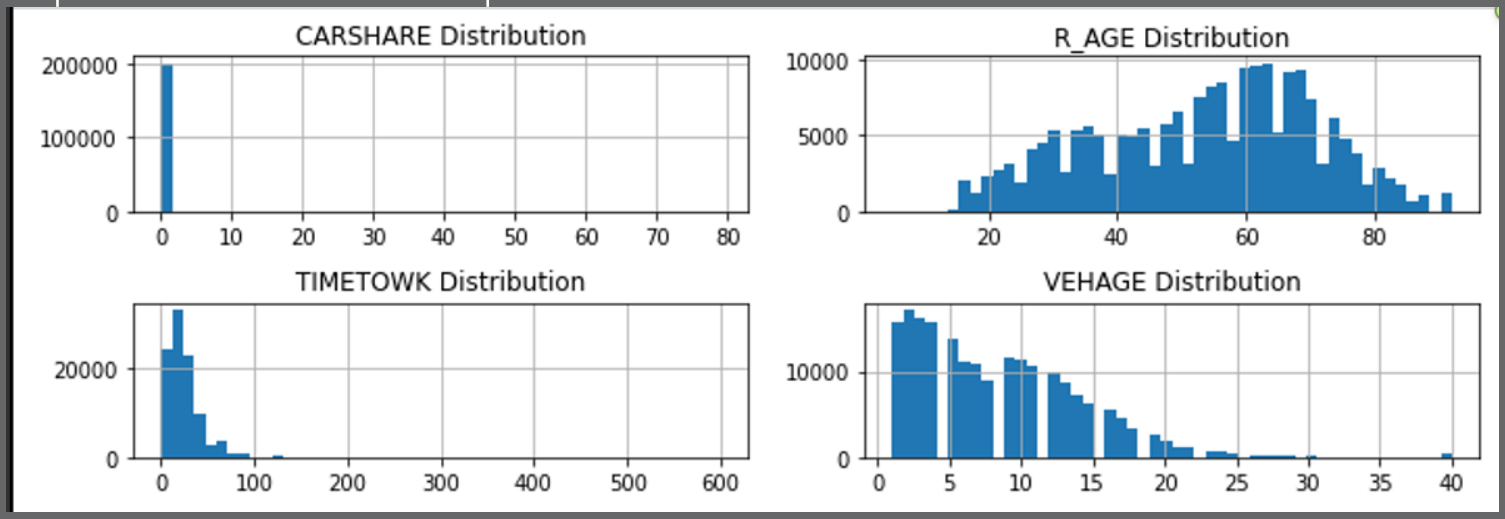


Figure 2: Distribution of Ordinal Variables

NUMERICAL VARIABLES

Apart from the age and the months of vehicle ownership variables, the other variables were skewed to the right indicating more values for each variable were populated at lower values.

Variables such as CARSHARE and YOUNGCHILD were zero-inflated, indicating most respondents chosen 0 as an option.



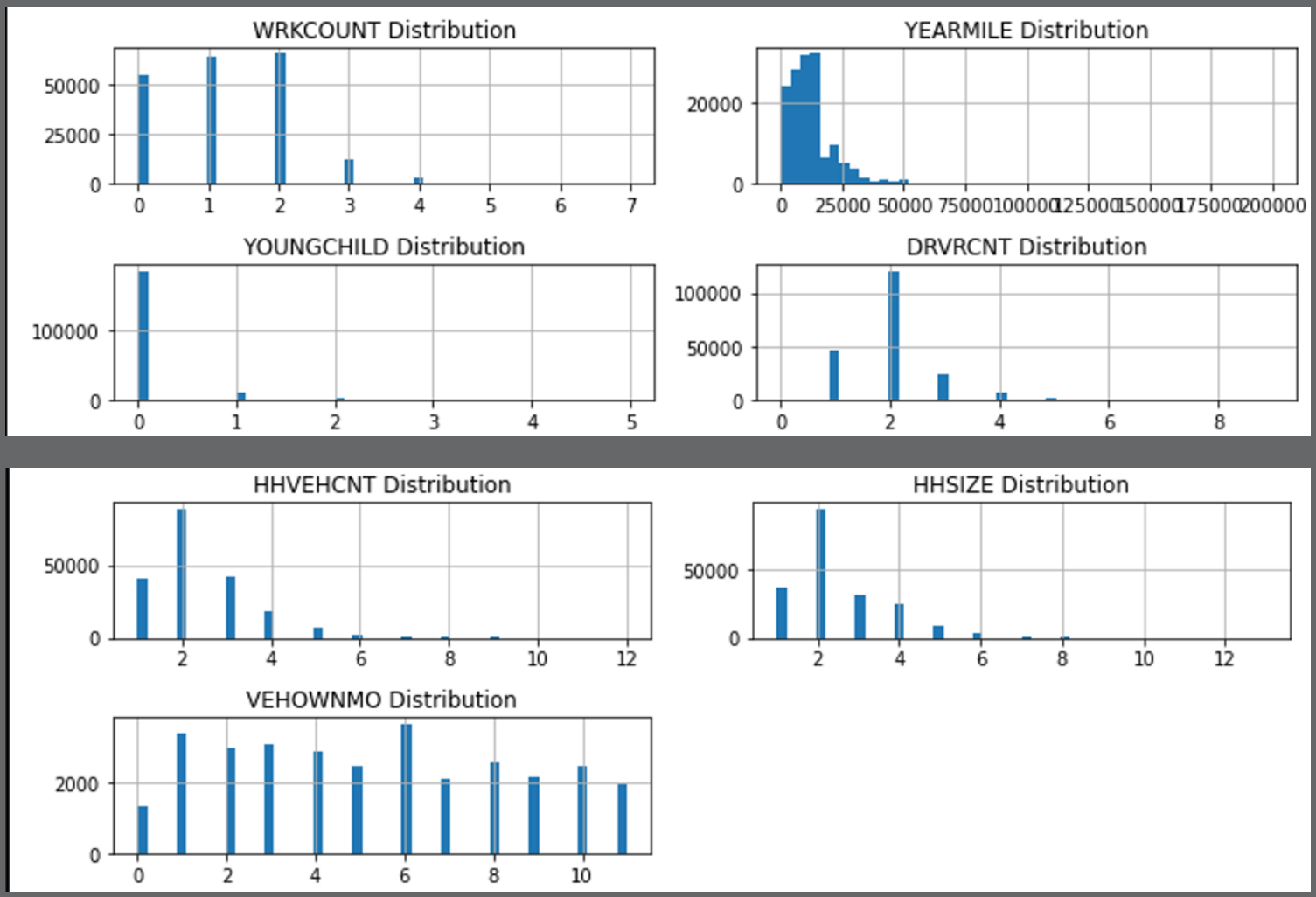


Figure 3: Distribution of Numerical Variables

RESPONSE VARIABLE

About 97% own non-EV in the sample, indicating that based on the sampled population most respondents still prefer to use non-EV vehicles as their mode of travel.

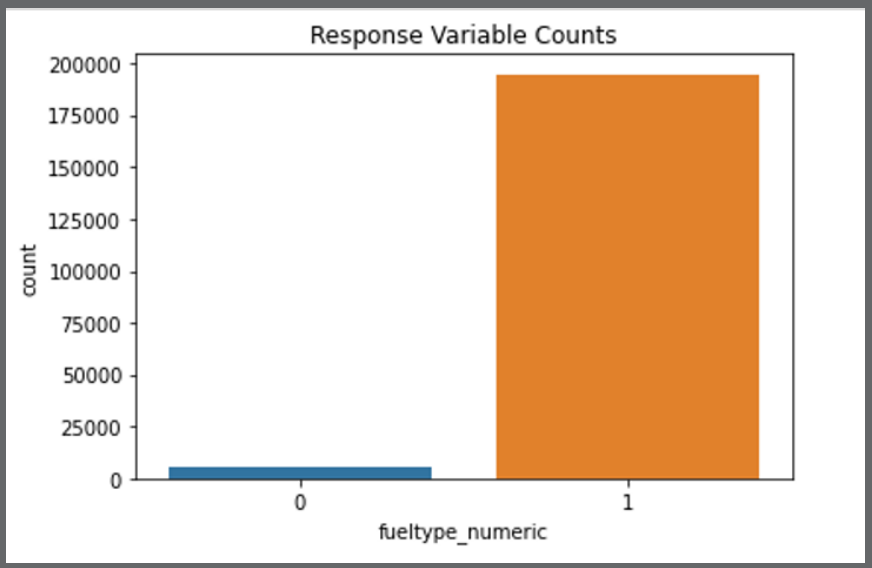


Figure 4: Distribution of response variable

HEATMAP

A heatmap was done to understand the relationship between ordinal and numerical variables.

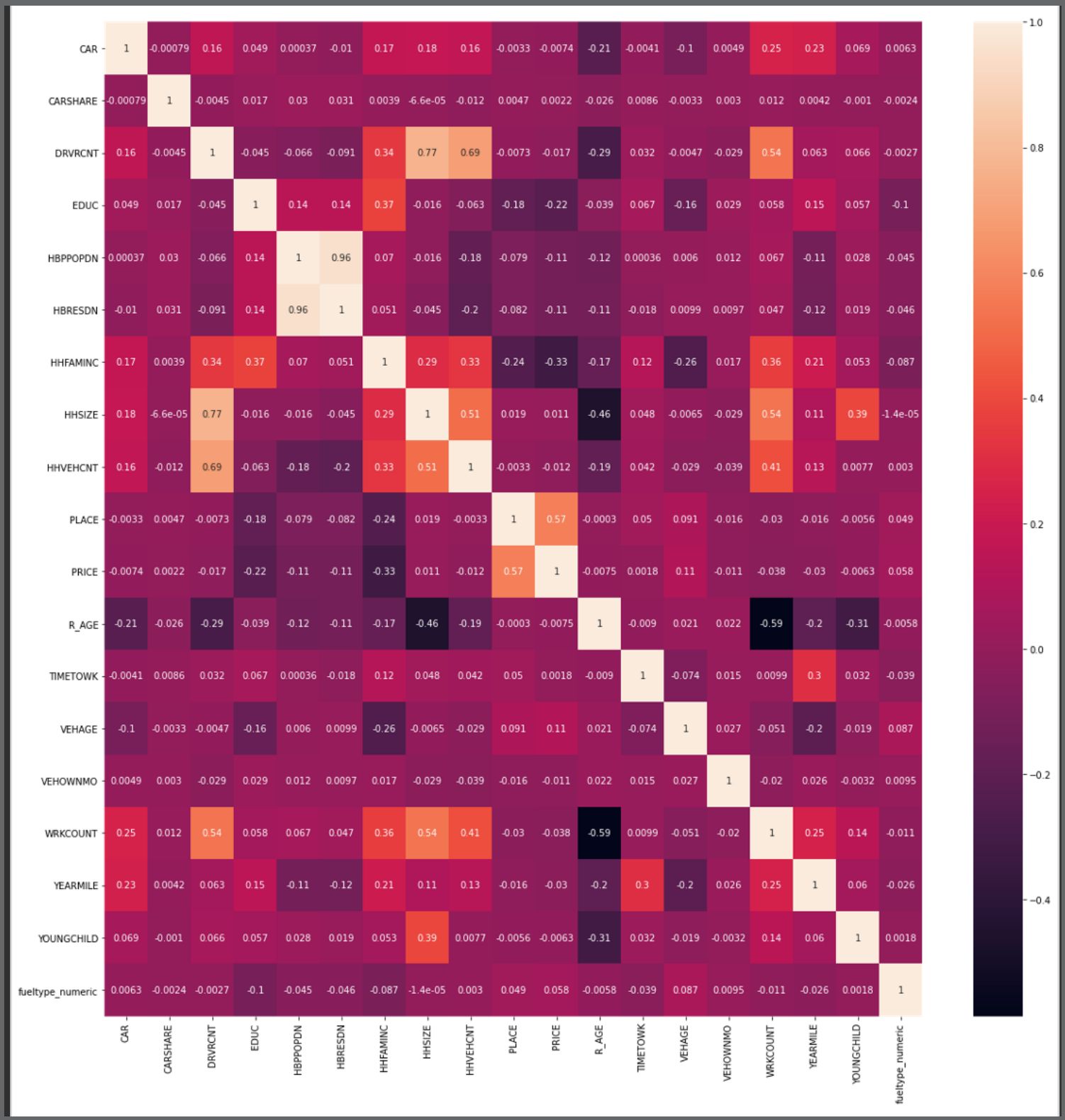


Figure 5: Heatmap

Based on the heatmap, many of the variables seem to be relatively correlated with each other.

- **HBPPOPDN and HBRESDN**

Category of population density and category of housing units are positively correlated. This makes sense as one increases, we would expect the other to increase as well.

- **HHSIZE and DRVRCNT**

As the count of household members increase, number of drivers could increase as well.

- **WRKCOUNT and DRVRCNT**

As the number of workers in the house increases, we would expect drivers to increase as well.

- **R_AGE and YOUNGCHILD**

These are negatively correlated. As the number of young children increases, we could expect the age groups of people to decrease. This could of course be negatively correlated the other way round as well.

- **WRKCOUNT and R_AGE**

The number of workers and ages of people in a household are negatively correlated. We would expect the working class to be younger so this is possible.

- **R_AGE and HHSIZE**

The number of people in a their household could have been negatively correlated with age.

VISUALIZATION OF EXPLANATORY AND RESPONSE VARIABLE

GT1JBLWK

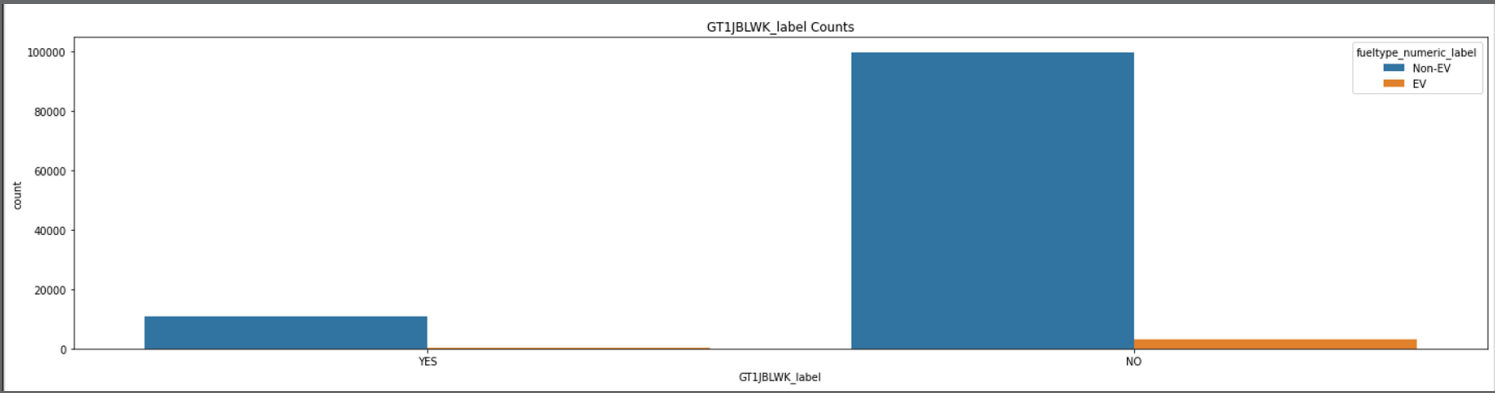


Figure 6: Bar graph of GT1JBLWK and Response

Based on the graph, the number of people who own non-EV cars is much higher among the people who have lesser than or equal to one job than people with more than one job.

HBHUR

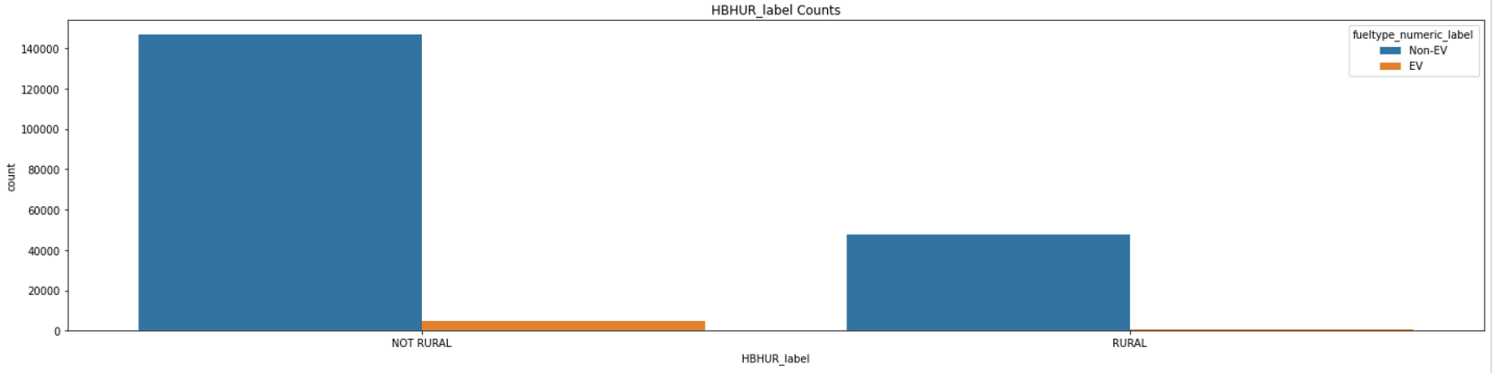


Figure 7: Bar graph of HBHUR and Response

Based on the Figure 7, the number of people who own non-EV cars is much higher among the people are in non rural areas than people in rural areas.

HOMEOWN

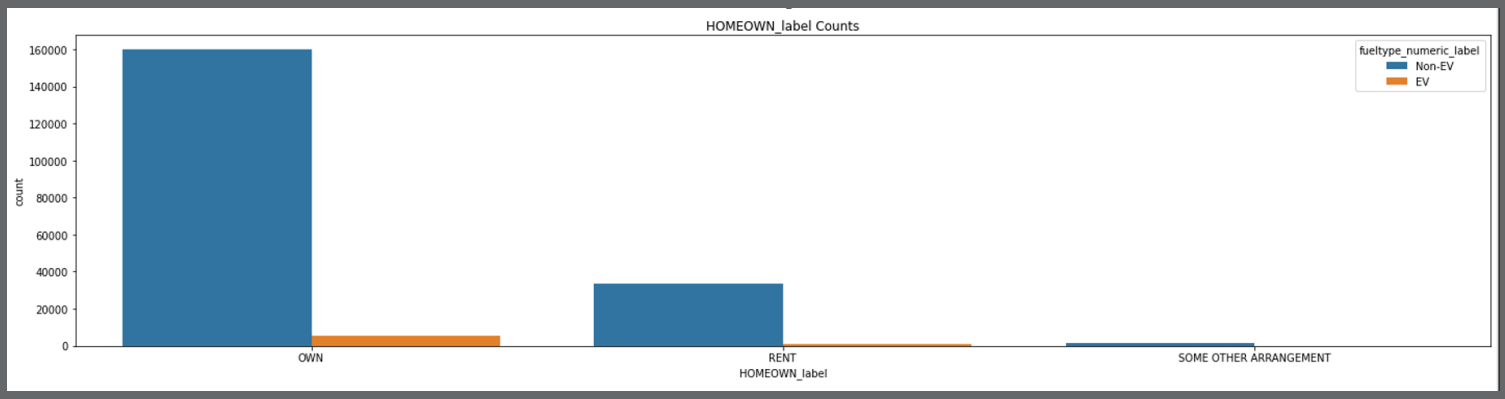


Figure 8: Bar graph of HOMEOWN and Response

Based on the Figure 8, the number of people who own non-EV cars is much higher among the people who own their own home than all other groups such as people who rent or have some other arrangement for their home.

OCCAT

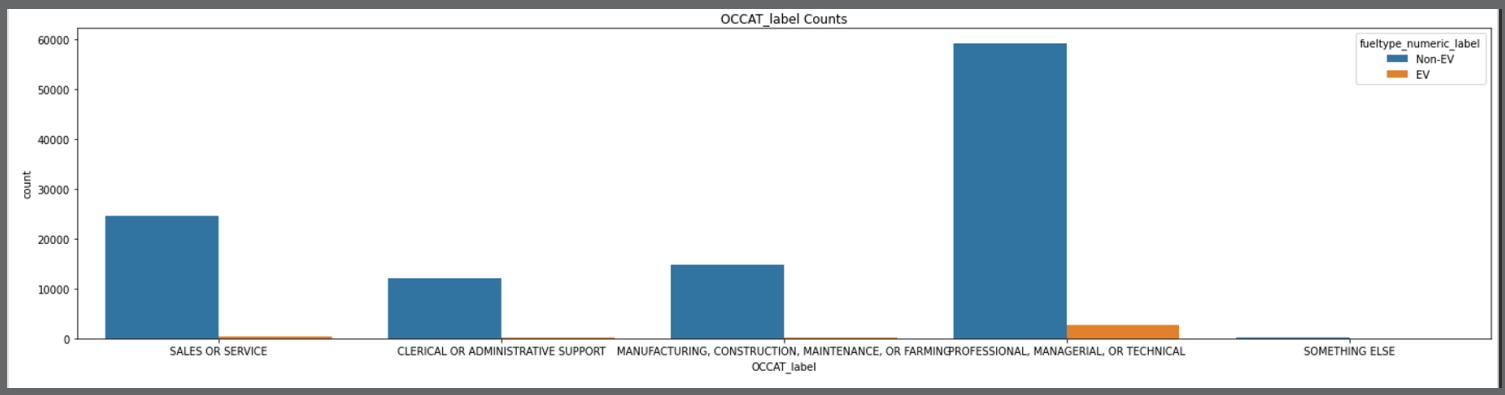


Figure 9: Bar graph of OCCAT and Response

Based on the Figure 9, the number of people who own non-EV cars is much higher among the people have professional, managerial or technical jobs than all other groups.

R_HISP

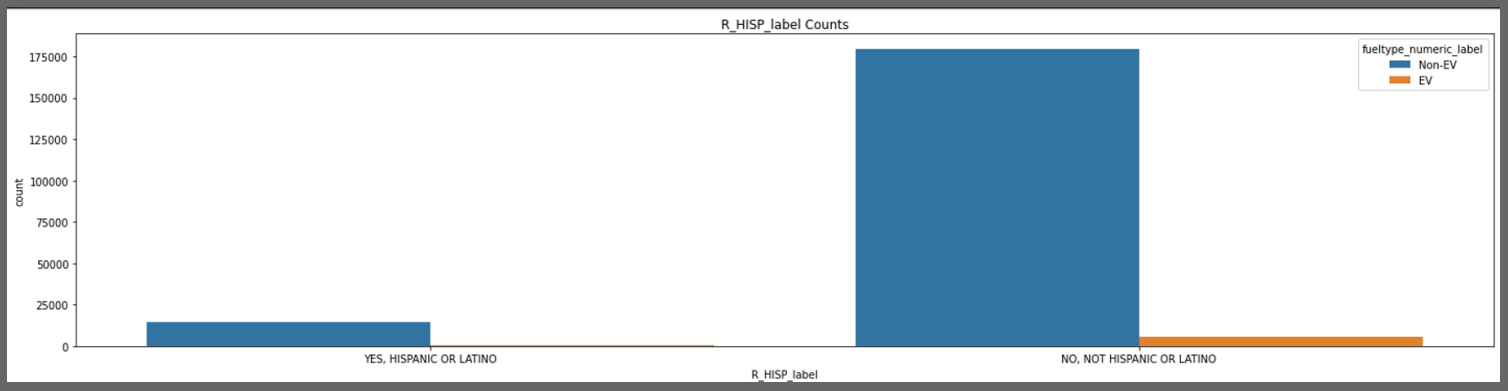


Figure 10: Bar graph of R_HISP and Response

Based on the Figure 10, the number of people who own non-EV cars is much higher among the people who are not Hispanic or Latino than among people who are Hispanic or Latino.

R_RACE

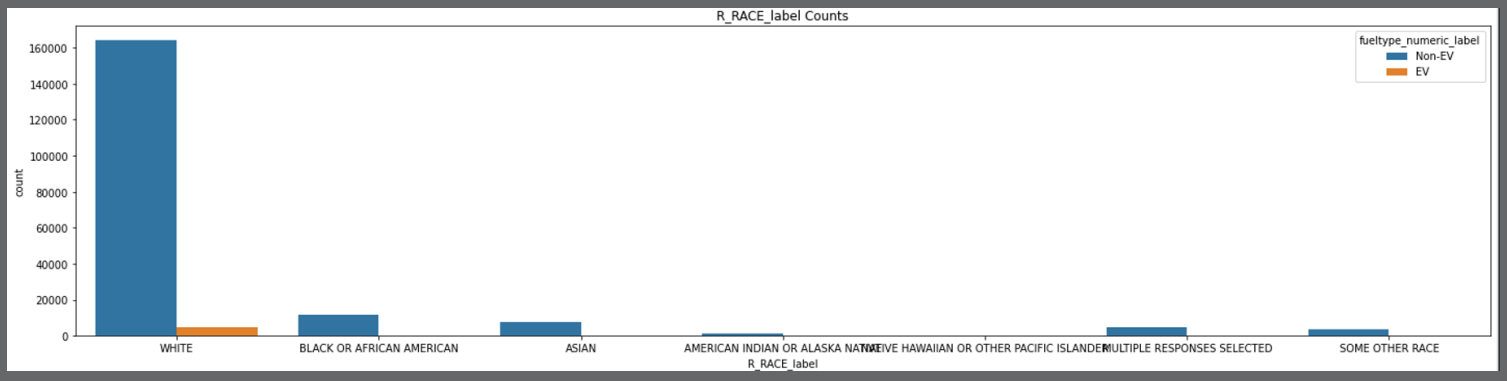


Figure 11: Bar graph of R_RACE and Response

Based on the Figure 11, the number of people who own non-EV cars is much higher among the people who are White than people who belong to other races.

R_SEX

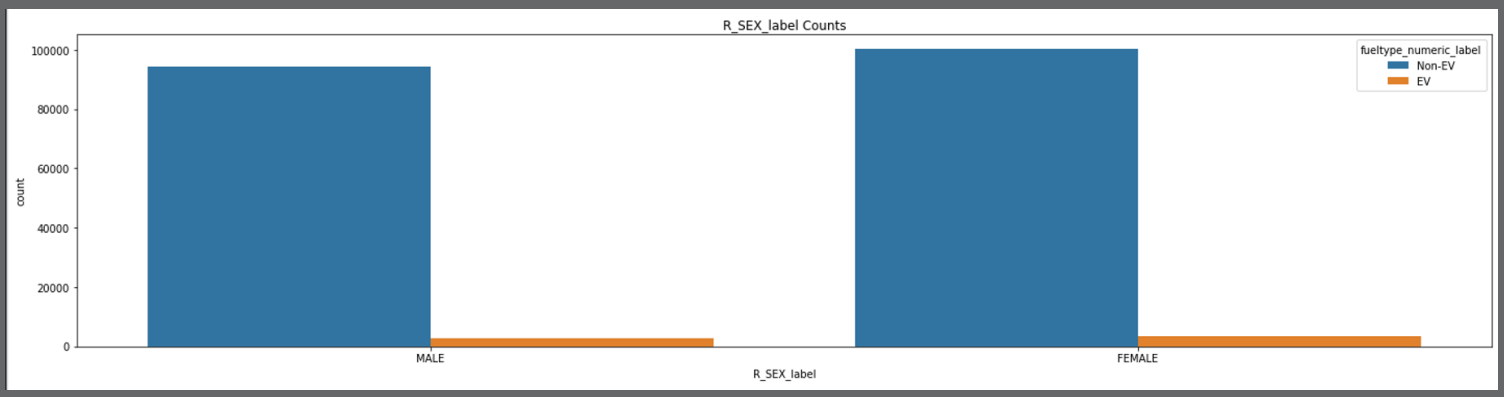


Figure 12: Bar graph of R_SEX and Response

Based on the Figure 12, the number of people who own non-EV cars not very different between males and females.

WKFTPT

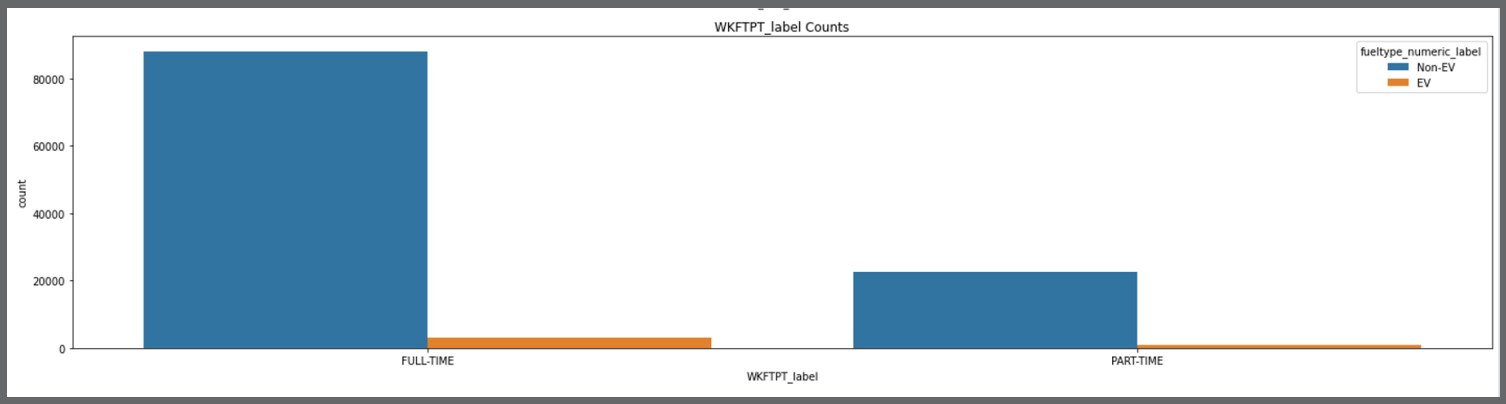


Figure 13: Bar graph of WKFTPT and Response

Based on the Figure 13, the number of people who own non-EV cars is much higher among people working full-time than people working part-time.

PREPROCESSING

Data Split

The dataset was split 70% into the train dataset and 30% into test dataset.

Imputation for Missing Variables

Numerical variables were imputed using the median of each variable and the categorical variables were imputed by the most frequent value.

Scaling of Variables

The numerical variables were scaled using the min max scaler.

Dummies

Dummies were created for the categorical variables for further analysis.

MODEL BUILDING

Different models were built to assess which model best predicts the variables that results in the a person purchasing a non-EV vehicle.

The metric that was used to assess these models were the F1 score. This metric was chosen because class imbalance was seen in the response variable.

METRIC FOR ASSESSMENT

Although F-score, recall and precision results were provided, F-score will be used as the assessment to summarize model performance. F-score provides a way to combine both precision and recall into a single measure that captures both properties.

A model can have excellent precision with terrible recall, or alternately, terrible precision with excellent recall. As a combination of both measures, F-score provides a way to express both concerns with a single score. It is most commonly used for imbalanced classification problems and will be used for this project's assessment since the response variable is imbalanced.

NAÏVE MODEL

A naïve model was used as a baseline. The purpose of this model was to assess how a random model (with random data) perform using the output.

The random data was modelled using a Logistic Regression model. The number of observations that were used for the training and test datasets input (X) for this model were similar to the original training and test dataset. However, the input data that was used were just 1's. The output that was used was the original dataset's output.

Based on the F1 score, the naïve model, based on sample data, was performing classification well. Results are shown in Table 5 under naïve model.

MODEL BUILDING

SUPERVISED LEARNING - INITIAL MODELS FOR ASSESSMENT

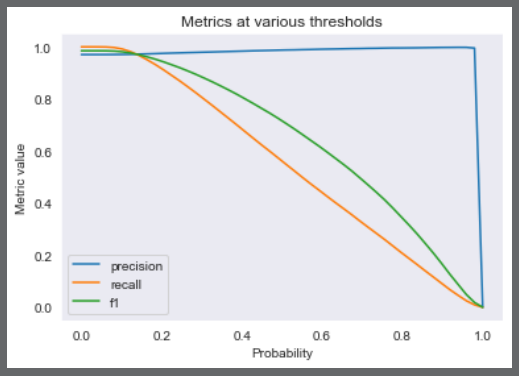
An initial logistic regression model was used to classify the data. Since the cardinality of the dataset was high, only numeric variables (including ordinal variables) were kept for further analysis.

Using the numeric variables, the Logistic Regression, Poisson Regression and Random Forest models were used for classification. Note that these results were derived by assessing the maximum threshold derived using the train datasets. Confusion matrices for all 3 models are provided below as well.

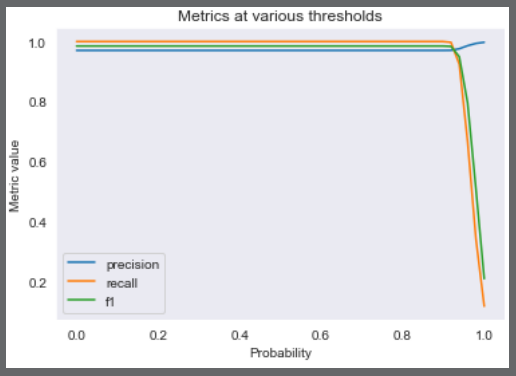
	F1 score for Train	F1 sore for Test	Precision for Train	Precision for Test	Recall for Train	Recall for Test
Naïve Model	0.956	0.956				
Logistic Regression	0.985	0.985	0.998	0.999	1	1
Poission Regression	0.985	0.985	0.997	0.998	1	1
Random Forest	0.985	0.985	0.995	0.995	1	1

Table 5: Precision, Recall, F1 for Train and Test Datasets

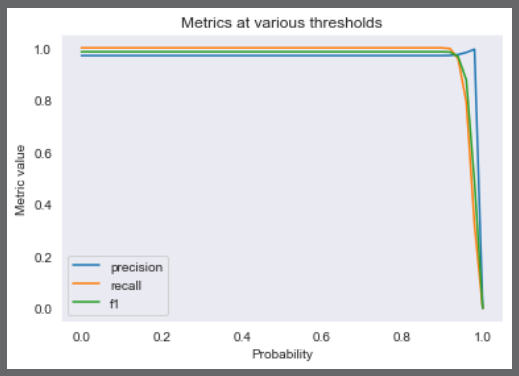
Metrics at Various Thresholds



Logistic Regression



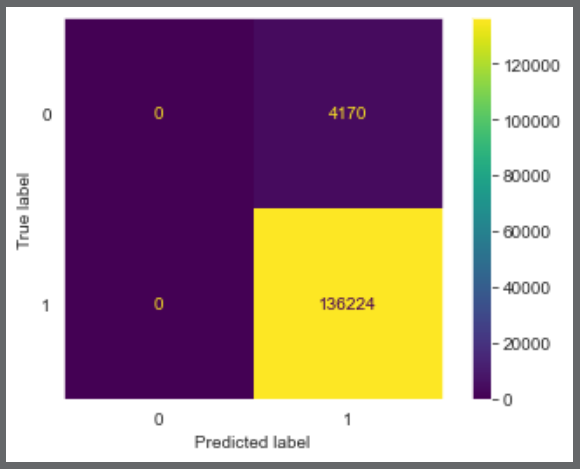
Poisson Regression



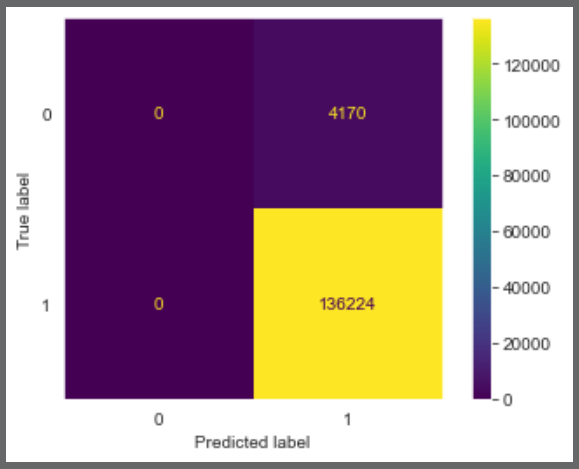
Random Forest

Figure 14: Metrics for Logistic Regression, Poisson Regression and Random Forest

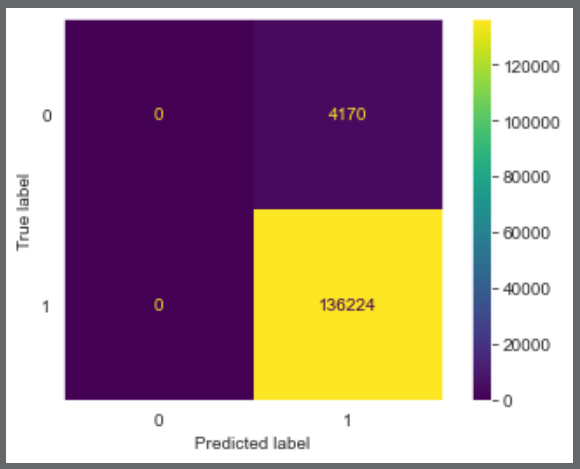
Confusion Matrix



Logistic Regression



Poisson Regression

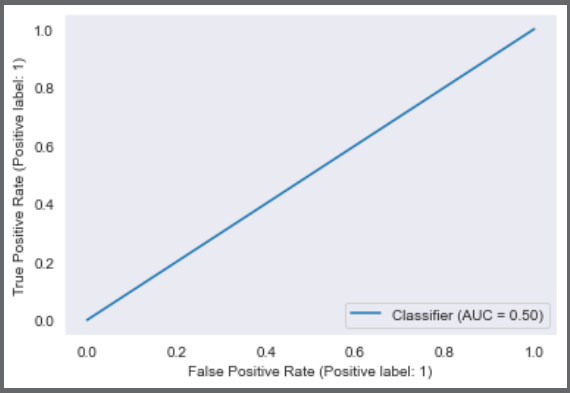


Random Forest

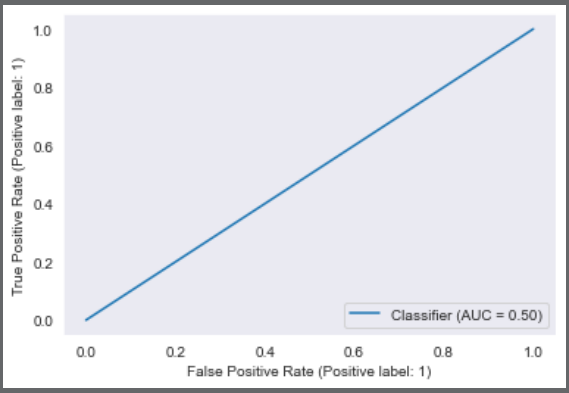
Figure 15: Metrics for Logistic Regression, Poisson Regression and Random Forest

Although the F1-score were high for all 3 models, the confusion matrices were similar for all 3 models and they showed that all the observations were predicted as non-EV. Further assessment was done using the ROC. Results are shown in the figures below.

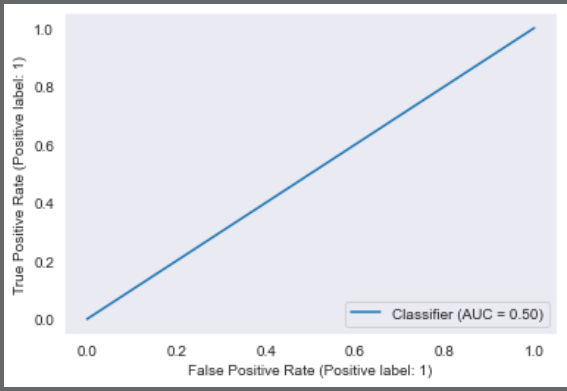
Results from ROC



Random Forest



Poisson Regression



Logistic Regression

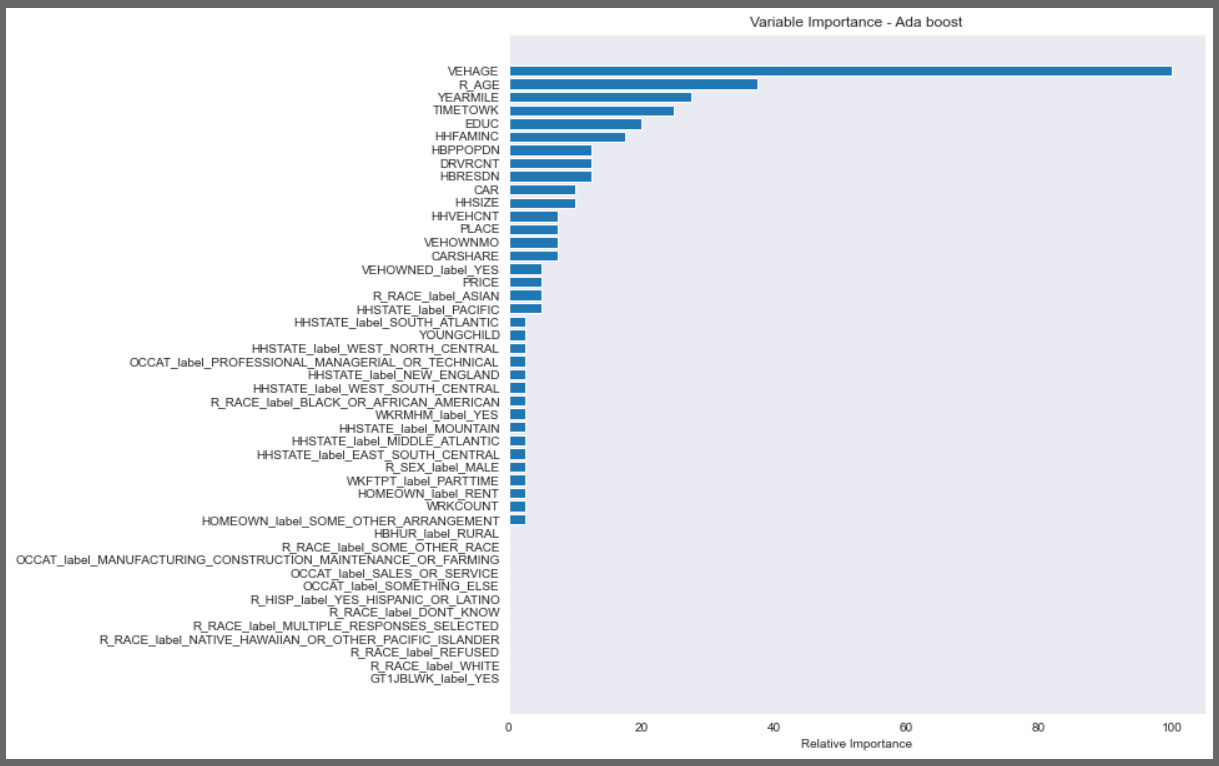
Figure 16: ROC for Logistic Regression, Poisson Regression and Random Forest

The diagonal plot shown in all 3 graphs indicated that the model was guessing how to classify the data. There was no pattern seen in the classification.

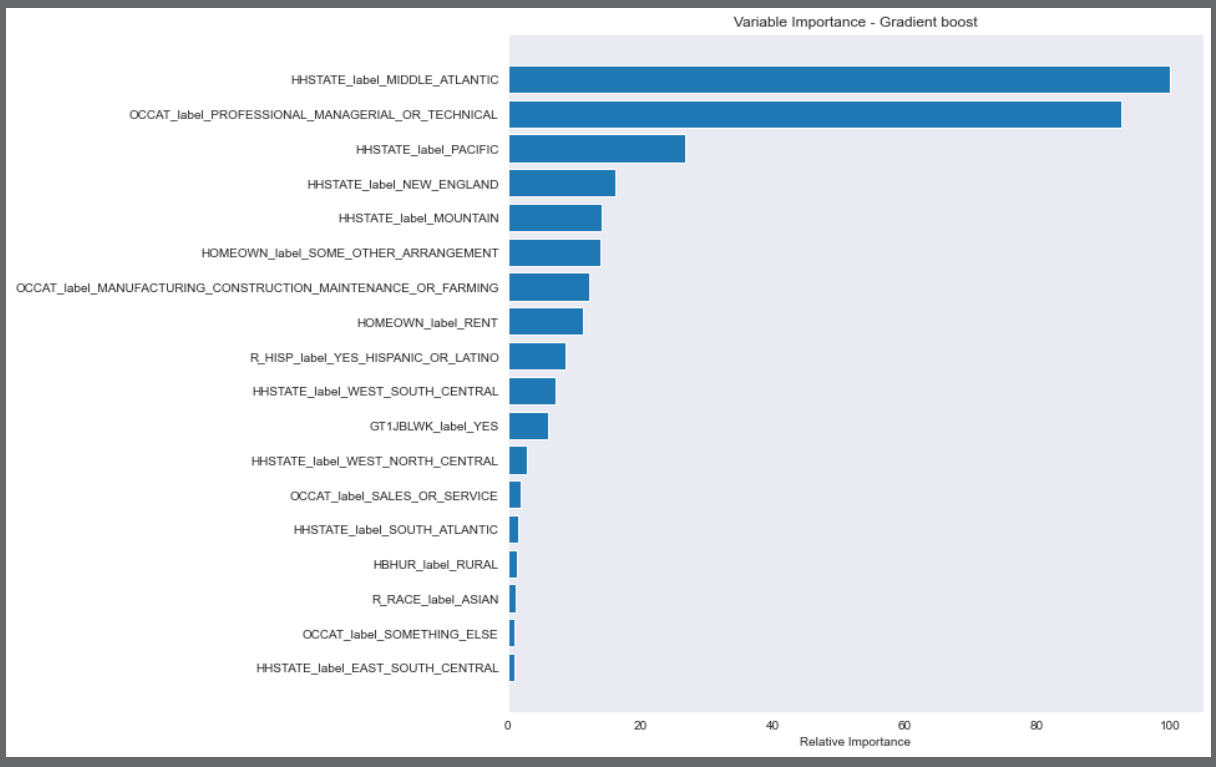
SUPERVISED LEARNING - FURTHER MODELING

Since all 3 models did not perform well in the previous models, variable importance was performed on the entire dataset using AdaBoost and Gradient Boosting.

Results from both are provided below:



AdaBoost

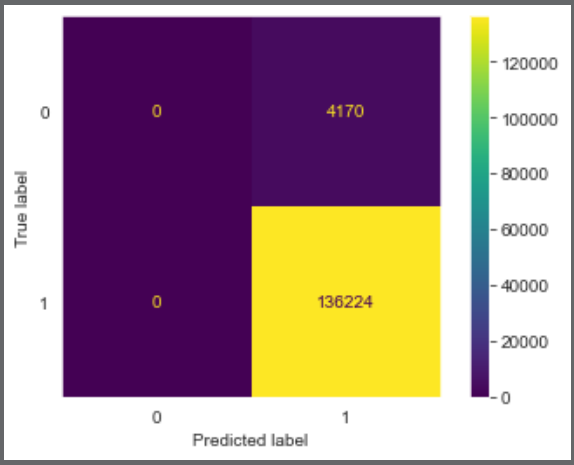


Gradient Boosting

Figure 17: Variable Importance for Adaboost and Gradient Boosting

Based on these, HHSTATE_label_PACIFIC, EDUC, HHFAMINC, R_AGE, PLACE and PRICE were chosen. Following that, a linear model with stochastic gradient descend (SGD) was implemented. However, as seen from the confusion matrix in Figure 18, the results from this model did not differ from the previous 3 supervised models. Hence, another method to model the data needs to be considered.

Confusion Matrix



Linear Model with SGD

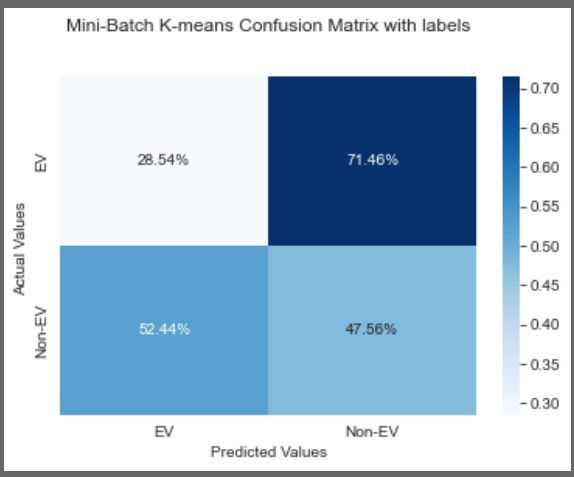
Figure 18: Confusion Matric for SGD

UNSUPERVISED LEARNING

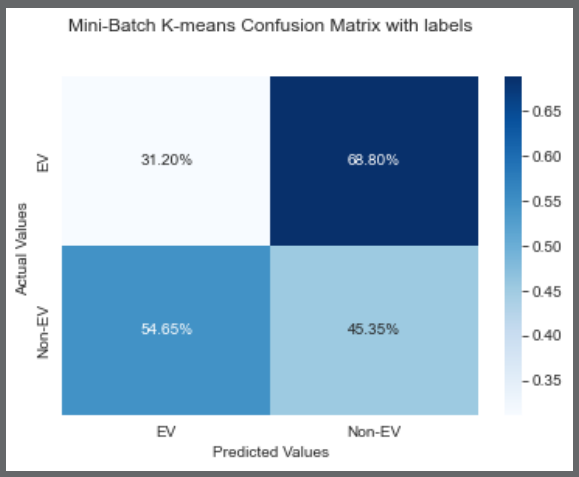
Unsupervised learning was done using K-means and Mini-batch K-means. The confusion matrices from both models are given below.

Since supervised learning models did not manage to classify the dataset well, we moved on to assess the model with unsupervised learning methods.

Confusion Matrix



K-means



Mini-batch K-means

Figure 19: Confusion Matric for K-means and Mini-batch K-means

The results in Figure 19 seem much better than the results from supervised learning as not all observations were predicted as non-EV. In order to further assess the models, the performance for each variable was visualized. Since many graphs had to be further produced, a dashboard was build using dash in python.

The interactive dashboard can be found here: https://projectev.herokuapp.com.

CONCLUSION

Overall, for this analysis, an unsupervised learning model worked better than a supervised model to classify the dataset. The dataset was particularly challenging as the response was imbalanced and many of the supervised models could not classify the adequately.

Other models that may have worked better are non-linear models such as a neural network and convolutional neural network and these models can be utilized to model the data as future work.

APPENDIX

Original Variable Names and Labels

Variable Name	Variable Label
GT1JBLWK	More than One Job
HBHUR	Urban / Rural indicator - Block group
HHSTATE	Household state
HOMEOWN	Home Ownership
OCCAT	Job Category
R_HISP	Hispanic or Latino Origin
R_RACE	Race
R_SEX	Gender
VEHOWNED	Owned Vehicle Longer than a Year
WKFTPT	Full-Time or Part-Time Worker
WKRMHM	Option of Working from Home
CAR	Frequency of Personal Vehicle Use for Travel
CARSHARE	Count of Car Share Program Usage
DRVRCNT	Number of drivers in household
EDUC	Educational Attainment
HBPPOPDN	Category of population density (persons per square mile) in the census block group of the household's home location
HBRESDN	Category of housing units per square mile in the census block group of the household's home location
HHFAMINC	Household income
HHSIZE	Count of household members
HHVEHCNT	Count of household vehicles
PLACE	Travel is a Financial Burden
PRICE	Price of Gasoline Affects Travel
TIMETOWK	Trip Time to Work in Minutes
VEHAGE	Age of vehicle, based on model year
VEHOWNMO	Months of Vehicle Ownership
WRKCOUNT	Number of workers in household
YEARMILE	Miles Personally Driven in all Vehicles
YOUNGCHILD	Count of persons with an age between 0 and 4 in household