Prediction of Alternative Fuel Vehicle Adoption: 2017 NHTS Case

Sami SHOKER Ben ILBOUDO Senou AHOUNOU

April 2022

Abstract

The adoption of alternative fuel vehicles is seen as a major step towards the transition to sustainable energy. This topic is believed to have a wide application in government policy making and in the automotive manufacturing industry. The objective of this research is to analyze and predict the adoption of AFV using different parametric and non-parametric machine learning models. The results showed that the linear models outperformed the non-parametric models. The best performing models were Logistic Regression, Logistic General Additive Model in addition to the Naive Bayes model, but the latter model had the disadvantage of a lack of interpretability.

Keywords: AFV adoption, Imbalanced dataset, multicollinearity, parametric models, non-parametric models.

Contents

1 Introduction

The earth's greenhouse gases trap heat in the atmosphere and make the planet warmer, increasing global warming. Since the industrial revolution, greenhouse gases emitted by human activity have increased dramatically. It is estimated that between 1750 and 2011, the atmospheric concentration of methane has increased by 150%, carbon dioxide by 40% and nitrous oxide by 20% [6].

In 2019, the transportation sector accounted for 29% of total U.S. greenhouse gas emissions,making it the largest contributor to U.S. greenhouse gas emissions, followed by the electricity sector [13].

There are several ways to reduce and limit greenhouse gas emissions in the transportation sector. One of the most important ways is fuel switching, which is the adoption of a vehicle that emits less carbon dioxide (alternative fuel vehicle) than traditional vehicles that rely primarily on gasoline and diesel (conventional vehicle). However, according to a 2017 statistic, out of 17.25 million auto sales, alternative fuel vehicles accounted for only 585,930 sales, or only about 3.39% [1].

We know by far that the adoption of alternative fuel vehicles is a necessary behavior to reduce greenhouse gas emissions and preserve our planet. On the other hand, we see that this adoption is still very low and represents about 3% of the total population.

This topic is believed to have a broad application in government policy making and in the vehicle manufacturing industry. The purpose of this research is to predict and analyze the adoption of alternative fuel vehicles (AFV) by considering the individual and household characteristics in order to assess whether or not a person will adopt an AFV. Thus, it's binary classification problem.

The comparison will be between different parametric and non parametric machine learning models in order to select the optimal one. Parametric models used in this research are as follows: Logistic regression,Logistic regression with Factor Analysis of Mixed Data (FAMD), Ridge regression, Lasso regression, General Additive Model (semi-parametric). On the other hand, the non parametric models are : Random Forest, Neural Network, Support Vector Machine, K Nearest Neighbours and Naive Bayes.

The dataset used is this study is extracted from 2017 National Household Travel Survey [10] which is a large scale data for travel that takes into account more than 200 000 individual in the United State.

2 Literature Review

This topic, due of its importance with regards the preservation of our planet and its applications in vehicle manufacturing industry and government policy making, has attracted large amount research publications.

Pelsmacker and Moons [9] conducted a research to investigate the adoption of electric cars by analyzing the attitude of consumers towards the price, maintenance of the vehicle and the driving range. For Egbue et al [4], the battery range have the biggest weight for the consumer's behavior to choose an electric car while price and maintenance have lesser importance to orient the person's behavior. Diziano and Bolduc [3] found that the environmental conciseness for the consumers push them to adopt electric vehicles. Indeed, these results indicated that environmentally–conscious consumers are even willing to pay more money to get a lower emission car.

Li et al [8] summarized the factors influencing the consumer behavior to buy an electric car into three categories : Psychological, environmental an situational categories. It was also found that the individual and household factor have an impact on the person's behavior. Indeed, young and middle age good educated consumers are more likely to adopt an electric vehicle [2].

A recent paper by Jia [7] analyzed and predicted the consumer adoption to alternative fuel vehicle (electric, hybrid and other alternative fuel type vehicles). This paper was done on a large scale data using National Household Travel Survey and compared between different machine learning models in order to choose the optimal one capable of predicting if a consumer would adopt an AFV. Machine learning models were trained using person,household features. Results showed that the Random Forest algorithm has the best performance compared to other models by having the highest AUC and these results of the best model were employed to generate AFV penetration model in the state level.

3 Variables Explanation

The 2017 National Household Travel Survey was conducted in the United States from April 2016 to April 2017. It includes various questionnaires on vehicle, travel, person, and household demographics. The data were divided into four datasets: Vehicle, Household, Person and Travel data sets.

In this study, the travel dataset is not considered because the commuting

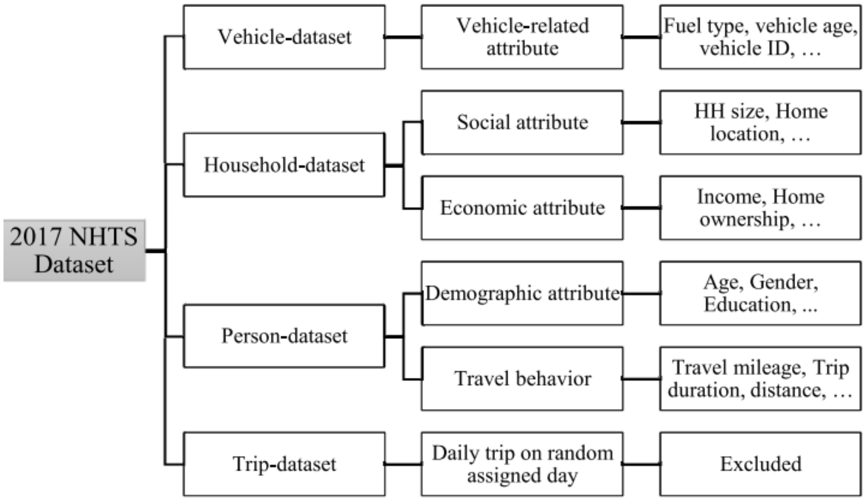
2017 NHTS Dataset
Vehicle-dataset
Vehicle-related attribute
Fuel type, vehicle age, vehicle ID, …
Household-dataset
Social attribute
HH size, Home location, …
Economic attribute
Income, Home ownership, …
Person-dataset
Demographic attribute
Age, Gender, Education, ...
Travel behavior
Travel mileage, Trip duration, distance, …
Trip-dataset
Daily trip on random assigned day
Excluded

Figure 1: 2017 NHTS Dataset Framework.

data were collected by random assignment of data for each household, including all family members. The other three data sets can be briefly explained as a total of 256,115 cars owned by 264,234 people in 159,696 households.

3.1 Vehicle Associated Variable

The main and only variable associated with the vehicle in this report is the fuel type and it's the target variable. Responses referring to gasoline and diesel fuel were converted to 0 and responses referring to electric, hybrid, and other types of alternative fuel vehicles were converted to 1 for the binary classification purpose.

Nonetheless, a person may have multiple vehicles, which leads to confusion when exploring the factors influencing vehicle type adoption. To address this issue, each individual was assumed to have only one vehicle and the older vehicle they owned, based on the vehicle age, was removed.

Moreover, each individual has been limited to adopting a vehicle over the past 3 years, as recent information better reflects market's potential[7]. Therefore, vehicle having an age older than 3 were removed from the data. Eventually, after merging vehicle, person and household data sets and cleaning them, the number of individuals that were included in this research went down to 20160 people.

3.2 Person Associated Variables

The individual associated variables express different kinds of information about the demographic attributes, travel characteristics in addition to his personal preferences with regards of transportation methods. Age, gender and race are common variables used in the research field to assess their impact on individual's behavior. In this article, "Age" is a continuous variable, "Gender" is a dummy variable where 0 refers to male and 1 refers to female and "Race" is a set of dummy variables where each variable refers to an individual's race. "White" variable was excluded from the model to be reference dummy variable with regards to other ethnic groups and to avoid strict multicollinearity. "Educational Attainment" is an ordinal variable that gives information about the person's education where the lowest value 0 refers to education less than high school and the highest value 5 indicates a level of education of graduate degree or professional degree, it was found in previous literature the significant impact of this feature with regards the adoption of an AFV vehicles [12][7]. "Relationship" is a categorical variable that indicates the respondent relationship in the household such as sibling, spouse, relative, non relative. Beside that, "Level of Physical Activity" is included in this report and it's an ordinal variable on a scale of 3 where 1 refers that the respondent never do physical activities and 3 indicates the respondent perform intense activities.

Variables that assess the information about the person's job are "More the One Job","Full Time or Part Time Worker" and they are dummy variables that describe whether the individual works a full time or part time and whether he/she has more than one job. "Job Category" is a categorical variable that is defined as: 01 = service or sale , 02= administrative support, 03= construction and manufacturing and 04=managerial, technical or professional. "Time Trip to work" is a continuous variable that describes the time in minutes that the person needs to arrive to work.

To evaluate the travel characteristics for respondents, "Count of Walk trips", "Count of Bike trips", "Estimations of Annual Miles" and "Count of Care Sharing Programs" are continuous variables included in this research. The first two variables indicates the number of walk trips and bike trips that the respondent did during their last week. The third variable is an estimation of annual vehicle mileage and the last one is the frequency of using a care sharing service during their last 30 days. Finally to evaluate the person's preferences towards the methods of transportation, multiple variables that represents questions asked to the individual where included in this research. At first, "Price of Gasoline Affects travel" and "Travel is a Financial Burden" are two categorical variable that represents questions where their values are on a scale from 1 to 5 where 1 indicates that the person strongly agrees scaling to the value 5 that indicates that the person strongly disagrees. "Public Transportation or Taxi","Passenger to Friend/Family Member or Rental Car"and " Bicycle or Walk" are used to evaluate the impact of the individual preferences towards alternative modes of transportation, these are categorical variable on a scale from 1 to 4 where 1 represents that the respondent prefers the fist alternative mode of transportation, 2 represents that he/she prefers the second mode, 3 indicates the preference to both mode and 4 indicates to preference to neither mode of transportation.

3.3 Household Associated Variables

Variables representing the characteristics of each household where included in this research to evaluate their impact on the behavior of adoption of AFV vehicles. "Household Income" is a categorical variable that represents the total annual income of the household, it's on a scale from 1 to 11 where 1 indicates an income below 10 000 $ going up to 11 that indicates an income above 200 000 $. In previous literature, family income was empirically proven to have an impact on individual's behavior to adopt an AFV vehicle [14][7][5]. "Home Own" is another income attribute for household. It's a categorical variable that is defined as : 1 = Home owned, 2 = Rented home and it was proven in previous literatures it's significant impact in the adoption of AFV vehicles [7].

"Household in Urban or Rural Area" and "Population Density", are geographical attributes for the household. The first variable is categorical where 1 means that the household exists in an urban area and 2 represents the location in rural area. The second variable is also categorical that represents the population density in a square mile, it's on a scale form 1 to 8 where 1 represents the lowest population density in the interval between 0 and 99 person and 8 represents the highest population density that is the interval between 25 000 and 999 999 person.

Lastly, "Count of household Members","Count of adult Members in household","Count of Children in Household" and "Count of Household vehicles" are additional numerical variables about the household's characteristics. The first three variables represent the total number of household members, number of adults in household and the number of children in the household. The last variable represents the number of vehicles each household has.

By total, in this research included 20160 observations having 10 numerical features and 22 categorical features. Additional variable explanation and description are included in the 2017 National Household Travel Survey documentation.

4 Data Exploration

In this section, summary statistics to the numerical and categorical variables is presented. In addition to that, variable dependency is analyzed.

4.1 Summary Statistics

Table 3 in appendix represents the summary statistics for categorical features. We observe from this table and figure 2 the existence of imbalanced data where individual adopting an AFV vehicle only represents 4 % of the total observations, this problem can affect the capability of machine learning models to predict the true positive value. Moreover, regarding categorical features, we see that the majority of the individuals own a home (81%) and lives in an urban area (78 %). Furthermore, the proportion of female respondents is slightly greater than male respondents (53 %) and 99 % of individuals have a level of education of high school degree or higher. Additionally, the majority have a job in the professional, managerial or technical field and most the respondents occupy a full-time job(86%). In addition to that, the biggest part of individuals belongs to the white ethnic group (86 %).

Table 4 in appendix represents the summary statistics for numerical features.We remark that the average age of the respondents is 47 year old and 50 % of households are composed of 2 individual or less. Additionally we observe the proportions of respondents that have children or use car share program or did bike trips during their last week are minor. On average, the time trip to work for the individuals is about 27 minutes and the average annual driving mileage is 14651.

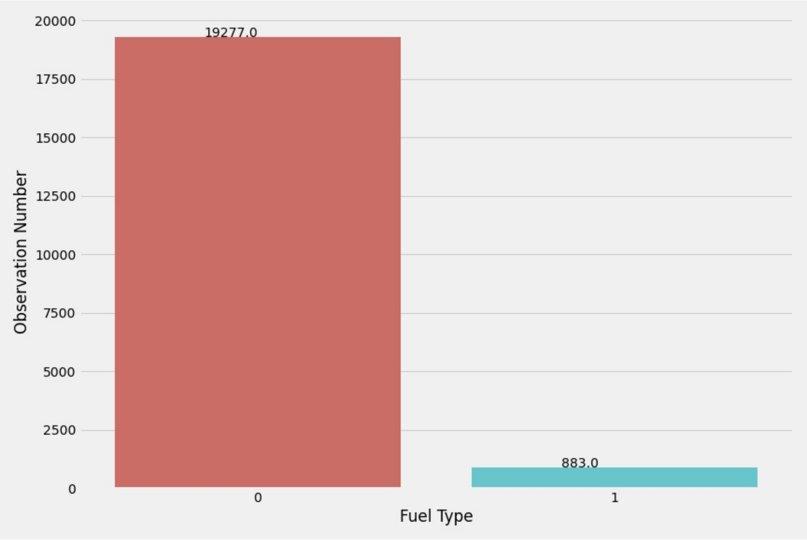

Figure 2: Count plot of the target variable

4.2 Correlation Analysis

In order to check the existence or multicollinearity problem within the features, the correlation between variables was assessed in addition to the utilization of the Variance Inflation factor.

Figure 3 represents a heat map of the Pearson correlation Coefficient between numerical features. We remark that there is low correlation between most of numerical variables. However, we notice the existence of quite strong correlation (> 0.5) in some areas between these features, which could lead to multicollinearity problem.

The analysis of correlation between categorical variables was conducted through a Chi-Square test. The objective of this test is to find if the difference between two categorical features was by chance or due to a dependent relationship between them. The null hypothesis is that the two variables are independent, whereas the alternate hypothesis is that the two variables are dependent. Figure 4 represents a heat map of the p-values for Chi-Square test. We notice that in the majority of cases the p-value is near zero, therefore we conclude the existence of high correlation between categorical variables which could indicate the presence of multicollinearity problem.

Both previous correlation tests were conducted between 2 features at each time. However, for the purpose of determination whether strong multi-

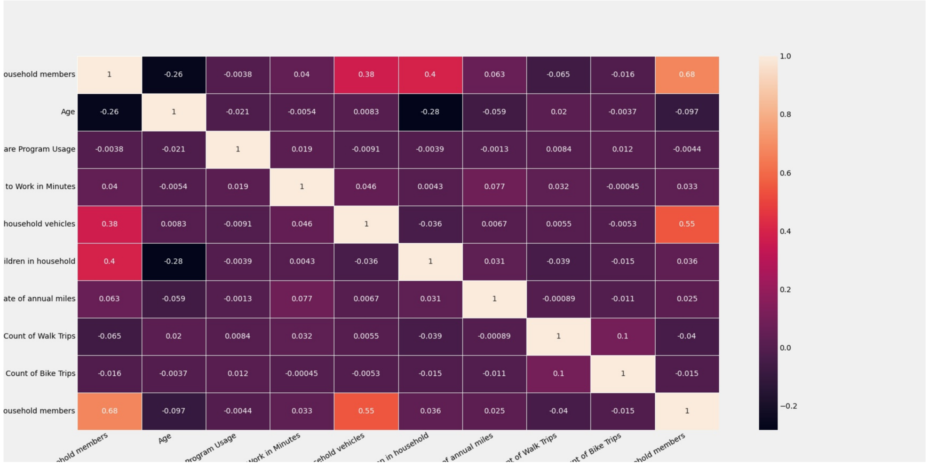

Figure 3: Pearson Correlation for Numerical Features

collinearity is present between the features, we utilized the Variance Inflation Factor which analyses the correlation between one variable X_i with the rest of variables. A VIF value above 4 indicates the possible existence of multicollinearity and a value greater the 10 indicates the significant presence of multicollinearity that needs to be corrected. VIF equation can be presented as follows:

$$VIF_i = \frac{1}{(1 - R_i^2)} \tag{1}$$

Table 5 in appendix indicates the presence of strong multicollinearity in our data such as for "Full-Time or Part-Time Worker" or "Bycicle Or Walk" variable , this problem leads to explosive variance when performing parametric models. This obstacle was addressed in this article by applying dimensionality reduction and regularization techniques.

5 Methodology

The objective of this article is based on comparison between different parametric and non-parametric machine learning models in order to be able to predict and analyse the individual's behavior towards adopting an alternative fuel vehicle. Such behavior is considered as a major step towards shifting to sustainable energy.

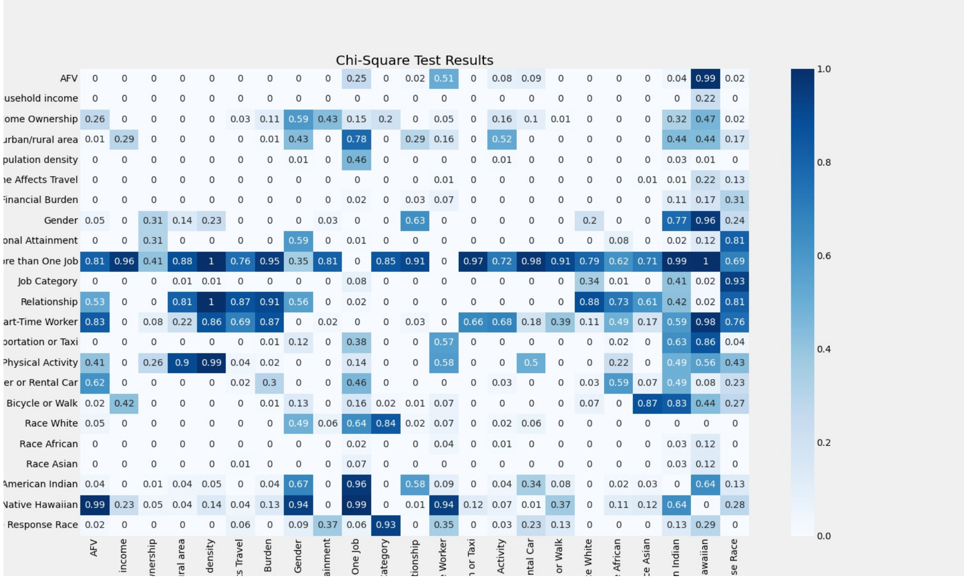

Figure 4: Correlation between categorical features

We Compared between Logistic Regression model, Regularization technique models (Lasso and Ridge), Logistic Regression with factor analysis that eliminates the multicollinearity problem, Logistic General Additive Model which is considered as semi-parametric model that is able to capture non-linear relationships while preserving in interpretability attribute that characterises the econometric models. For each parametric model we also added polynomial features with interaction to test their impact on model's performance in case non-linear relationship existed between the target variable and the features. Additionally, we compared between non parametric models that are Random Forest, Naive Bayes, Support Vector Machine, K Nearest Neighbors and Artificial Neural Network Model. Python and R programming languages were utilized for the data preparation and the data analysis in this research.

5.1 Logistic Regression

Since the logistic regression was introduced in 1960, it had a wide application in different industries such as Economic, Medicine, Banking etc. This model is used to measure the probability of a certain event to exist or not. In addition to its big application in binary classification, its usage extended to classifying multiple categories such as multinomial logistic regression and ordinal logistic regression where there is order in the classes. In this research, logistic regression model was employed and it is considered as a benchmark to compare its output with other parametric and non parametric machine learning models. The estimated parameters of logistic regression are obtained by maximising the following log-likelihood function:

$$l(\beta) = \sum_{i=1}^{n} [y_i x_i \beta - \log(1 + e^{x_i \beta})] \tag{2}$$

- X_i : Represents explanatory variables that we want to study their influence on individual's behavior to adopt an AFV vehicle.

- β : Represents the regression coefficients.

5.2 Logistic Regression with Factor Analysis of Mixed Data

As demonstrated in the subsection 4.2, one of the obstacles faced in this research in the multicollinearity problem. One of the ways to address this problem is by performing factor analysis to the data which replace the features by factors that doesn't have correlation between them. Factor analysis of mixed data is a method of principle components that analyses mixed data containing numerical and categorical variables. By summary, FAMD works as multiple correspondence analysis (MCA) for categorical data and as principle components analysis (PCA) for numerical data. In this research, the factors chosen are the ones having an eigenvalue greater than 1 since an eigenvalue lower than 1 indicates that it explains lesser than what a singular original variable does. After that, a logistic regression is applied on these factors.

5.3 Ridge Regression

Another way to address to multicollinearity problem is by applying regularization techniques. Ridge regression,which is considered as "L2 regularization" technique, penalizes the log-likelihood function,which is needed to be maximized, by subtracting a penalty equivalent to the sum of square of the magnitudes of coefficients. The penalized log-likelihood function is presented as follows [11]:

$$l_{ridge}(\beta) = l(\beta) - \lambda \sum_{i=1}^{n} \beta_j^2 \tag{3}$$

- λ : The hyper-parameter that needs to be optimized.

5.4 Lasso Regression

Lasso regression which is also called "L1 regularization", like the ridge regression model, it penalizes the log-likelihood function. This penalization is done by subtracting a penalty equivalent to the sum of absolute value of the magnitudes of coefficients to the log-likelihood function that needs to be maximized. Apart from shrinking the coefficients of the parameters, Lasso technique performs feature selection by giving a value zero for some predictors which is equivalent of excluding this predictor from the model. The penalized log-likelihood function is presented as follows[11]:

$$l_{lasso}(\beta) = l(\beta) - \lambda \sum_{j=1}^{n} |\beta_j| \tag{4}$$

5.5 Logistic General Additive Model

Linear models might miss some nonlinear relationships between the dependent variable and the explanatory variables. For these reason we applied Logistic General Additive Model in this research which is able to capture the non-linear relationship. One of the important aspects in logistic GAM is that there is no curse of dimensionality. This model can be presented as follows:

$$\log \frac{P(Y=1)}{P(Y=1)} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{j=1}^{m} S_j(X_j) \tag{5}$$

- β_0 : The intercept term.
- β_i : The coefficient for dummy variables.
- X_i : The dummy variables.

- X_j : The continuous or discrete variable.
- S_j : The estimated density function related to the predictors.

5.6 Support Vector Machine

Previously, we presented parametric machine learning models used in this work. One of the objectives of this research is to compare the performance of those previous models with the non parametric models. We strated by Support Vector Machine model that was created by V. Vapnik and A. Chervonenkis in 1963 [15].The main idea of this algorithm can be simply explained by developing a hyperplane that is able to separate the data into classes. In our case the 2 classes are Alternative Fuel Vehicles (AFV) and Conventional Vehicles (CV).

5.7 Neural Network

In this work, we also applied Deep Learning model to test it's predictive performance. Inspired by biological neurons , a neural network is constructed by nesting many neurons. A neuron is a computational unit that takes inputs and returns outputs. In such way that the output of one neuron is the input of another. In our work, since we have a binary classification problem, the output layer in the neural network model is the "Sigmoid Activation Function" and the loss function is the "binary-crossentropy".

5.8 Random Forest

Another popular non-parametric machine learning algorithm that was utilized in this work is Random Forest. This algorithm constructs several classification trees (a forest) and slices the space, branch by branch, on one variable at a time. Each tree is built on a subset of observation and variables. For a given observation to be predicted, the prediction of a tree is done by passing this observation within the branches, based on the splitting conditions. The prediction of the Random Forest corresponds to the average of the prediction of each of its trees. Additionally, for the purpose of exploring features importance in this algorithm, we utilized the GiniIndex that can be expressed as follows[7]:

$$GiniIndex = \sum\sum_{j\neq i}(\frac{f(Y_i,TR)}{|TR|})(\frac{f(TR_i,TR)}{|TR|}) \tag{6}$$

- TR : The training set.

- $\frac{f(Y_i,TR)}{|TR|})$: The probability that select sample would be affiliated to the class Y_i.

5.9 K-Nearest Neighbors

An additional supervised learning algorithm employed in our work is K-Nearest Neighbors algorithm which is considered as a very simple technique. This algorithm simply creates fictional boundaries in order to classify the data and at each time a new data point comes in, the algorithm would try to predict that to the closest of the boundary line.

5.10 Naive Bayes

Finally, we also utilized from the Naive Bayes Classifier algorithm to predict the person's behavior towards adopting AFV vehicles. It's a popular classifier with a special characteristic which is its high computational speed on a large scale data. This algorithm is built on the Bayes theorem that have the assumption of independence among predictors and the outcome class, which is in our case is AFV or CV vehicle, would be the one having the highest posterior probability. Bayes theorem can be presented as follows:

$$p(Y_i/X) = \frac{p(X|Y_i)p(Y_I)}{p(X)} \tag{7}$$

- $p(Y_i/X)$: The posterior probability of the class Y_i given the predictor X.

- $p(X|Y_i)$:The likelihood of the predictor given the class.

- $p(Y_I)$: The prior probability of the class.

- $p(X)$: The prior probability of the predictor.

5.11 SMOTE NC

As mentioned in the section 4.1, one of the main problems faced in this study is having an imbalanced dataset where only 4 % of the vehicles are AFV. This problem leads to unreliable predictions because of the existence of minority class that is not sufficient to train the models. In order to solve this problem we applied the synthetic minority oversampling technique (SMOTE) for numerical and categorical data (NC) which picks 2 near observations in the numerical feature space an draws new observations between them. That is for the numerical side of the observation, for the categorical side, this technique generates the new sample based on the most recurring category of the closest neighbors present during the generation.

In order to evaluate the impact of SMOTE NC technique on models predictive power, this technique was only applied to training data while testing data were kept the same.

5.12 Model Evaluation Metrics

The important aspect in our analysis is to measure the predictive power for our parametric and non-parametric machine learning models. Evaluation metrics are the following : Accuracy, Recall, Precision, F1 Score and AUC. Accuracy is the ratio of the sum of correct true positive TP and true negative TN out of all the predictions that were made. Recall is the ratio of the correct true positive out of all actual true positive values. Precision is the ratio of the correct true positive out of all positive predictions that was made. F1 score represents a harmonic average of recall and precision score and finally AUC which means Area Under the Curve is by summary evaluation of how good is the model is predicting the true positive values as true positive and the true negative values as true negative. A higher AUC mean a better model and the opposite is true.

6 Results and Discussion

In order to build a model that predicts the individual's behavior to adopt an alternative fuel vehicle, data was divided into train and test data. As mentioned in subsection 5.11, SMOTE NC technique was only applied on training data, therefore the correct true positive values only accounts for 4 % of the testing data. Data was divided into 70 % train and 30 % test. In our work we compared the results before and after the synthetic minority oversampling technique. Moreover, for parametric models, we tested the impact of adding polynomial features on the model's performance by comparing the results before and after.

6.1 Results Before SMOTE NC

Table 1 shows the performance evaluation metric for machine learning models. We clearly notice the poor performance of these models where we observe an AUC score of 0.5 for all models. This points on the challenge faced when working with imbalanced dataset. Indeed, all models failed to predict correct true positive values in the testing set for exception of K-nearest neighbors that had 0.07 recall score and Neural Network model that had a perfect recall score but a very low accuracy and that was due to the weight parameter where we gave greater weight for the class of AFV than the class CV.

Table 1: Prediction Evaluation Metrics Before SMOTE NC on testing Data

Classifiers	Accuracy	Recal	F1	Precision	AUC
Logistic regression	0.96	0.00	0.00	0.00	0.50
Lasso regression	0.96	0.00	0.00	0.00	0.50
Ridge regression	0.96	0.00	0.00	0.00	0.50
Naive Bayes	0.95	0.00	0.00	0.00	0.50
K-Nearest Neighbors	0.92	0.07	0.06	0.06	0.51
Neural Network	0.05	1.00	0.08	0.04	0.51
Support Vector Machine	0.96	0.00	0.00	0.00	0.50
Support Vector Machine	0.96	0.00	0.00	0.00	0.50
Random Forest	0.96	0.00	0.00	0.00	0.50

6.2 Results After SMOTE NC

In the section 6.1, almost all machine learning models failed to predict the true positive values due to imbalanced dataset problem. In this subsection, we explored the impact of applying SMOTE NC on train data in improving the model's predictive power.

6.2.1 Logistic regression

In this part the results of logistic regression model are analyzed and compared with the case of adding polynomial features.

Without Polynomial Features

The performance of logistic regression model after applying SMOTE NC improved significantly in its power to predict the true positive values. AUC score for this model is 0.6 and its recall score is 0.47, which means that this model was able to predict correctly 47 % of correct true positive values. In addition to that, the improvement of the model's capability to predict the correct true positive values, taking into account that these values only represent 4 % of the testing set, is accompanied by an accuracy of 73 %.

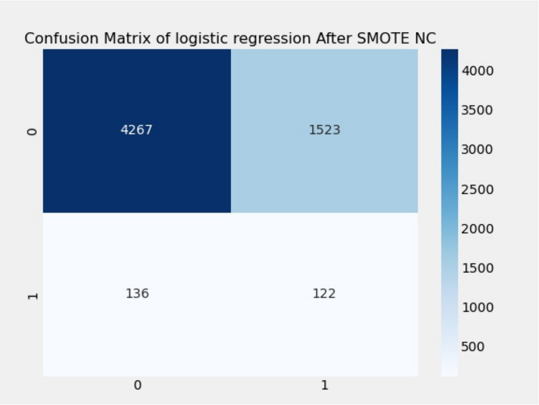

Figure 5: Confusion Matrix for Logistic Regression After SMOTE NC

Figure 6 represents the plot for the estimated coefficients of Logistic regression model after SMOTE NC. The first thing we notice in this figure is the large coefficient magnitudes. This is due to the multicollinearity issue that leads to explosive variance such as we observe the coefficient of "Home Ownership" variable, this problem also leads to decrease or remove the statistical power of other significant features. This model indicates that the most powerful feature in decreasing the log odds (which decreases the odds itself) of owning an AFV vehicle is "Home Ownership" variable which means that individual's owning a home have lower chance to adopt an AFV vehicle than those who rents one. On the other hand, the most powerful feature in increasing the odds of adopting an AFV car is educational attainment. This means individual's having higher education level would have higher chance in owning an AFV car.

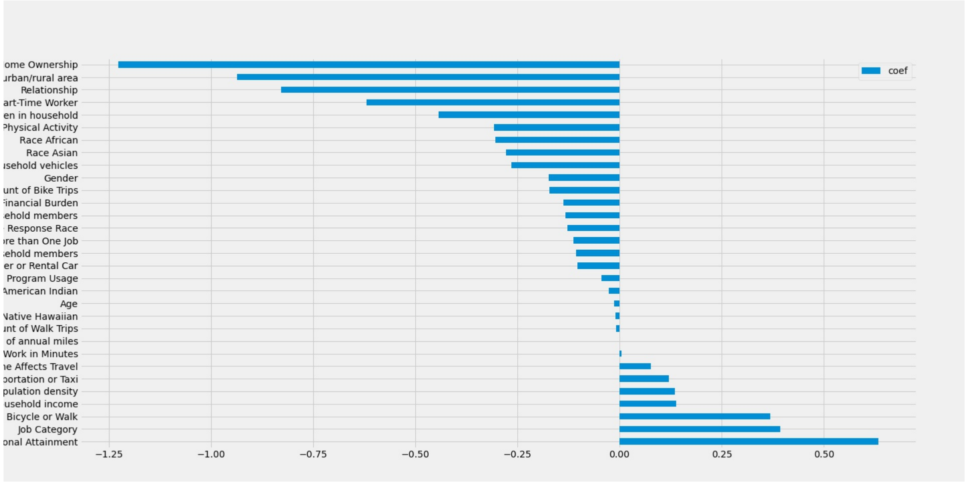

Figure 6: Feature Importance for Logistic Regression After SMOTE NC

With Polynomial Features

In order to assess the impact of the adding non linear components on the model's performance, we estimated logistic regression model with polynomial feature and interactions which increased the number of features in our model to 528 features. Figure 7 represents the confusion matrix of logistic regression model after adding polynomial features. We observe from this figure that this model fails to predict any correct true positive values. Additionally, this model have some False negative values. This big decrease in model performance can be explained by the fact that the original model already had multicollinearity issue and by adding this great amount of feature would lead to expand the model's explosive variance which eventually decrease it's performance.

6.2.2 Logistic regression with FAMD

In order to solve multicollinearity problem, factor analysis of mixed data was applied which creates components with no correlation between them as it's shown in figure 21.The number of components used was 30 components having a cumulative variance percentage of 55 %, which means that this

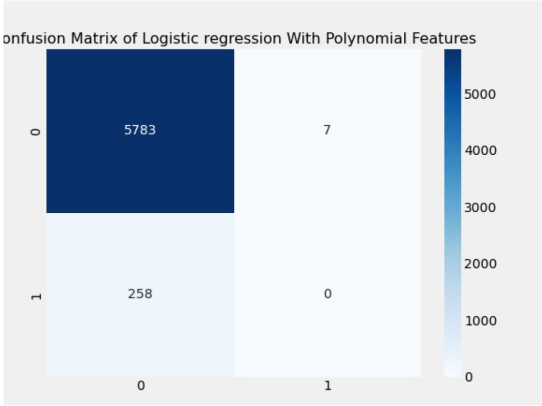

Figure 7: Confusion Matrix for Logistic Regression After SMOTE NC

model was abel to represent 55 % of the original one as shown in figure 22 in appendix.

Figure 8 represents the confusion matrix of logistic regression with FAMD. This model have an AUC score of 0.6 same as for the logistic regression after SMOTE NC. The main improvement in this model is it's predictive power to the true positive values that only represents 4% of testing data. Indeed this model was able to predict 67% of the true positive values while preserving the same AUC of the logistic regression model. However, the major drawback of this predictive model is its lack of interpretability which makes us unable to identify the feature's importance that influences the individual's behavior to own an AFV vehicle.

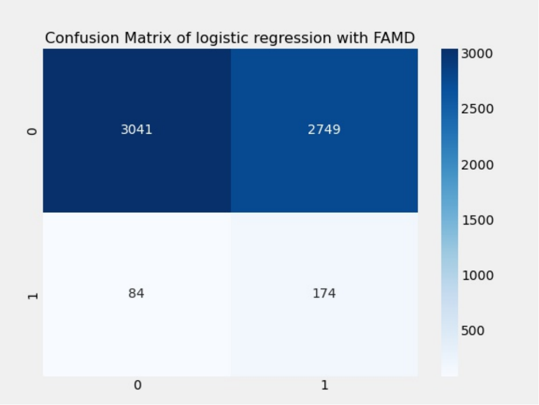

Figure 8: confusion matrix for logistic regression with FAMD

6.2.3 Ridge regression

Another way to solve multicollinearity problem is by applying regularization techniques. In this subsection we analyzed ridge regression model performance after SMOTE NC. Additionally, we tested the impact of polynomial features on model's performance. λ penalizing parameter was chosen based on grid search cross validation technique where it was based on $F1$ score.

Without Polynomial Features

After performing grid search cross validation on the train data, the optimal penalization parameter λ value was $1e-05$. Figure 9 represents the confusion matrix where we see that this model's predictive performance is similar to the logistic regression model performance after SMOTE NC but it's slightly lower by having an AUC score of 0.59.

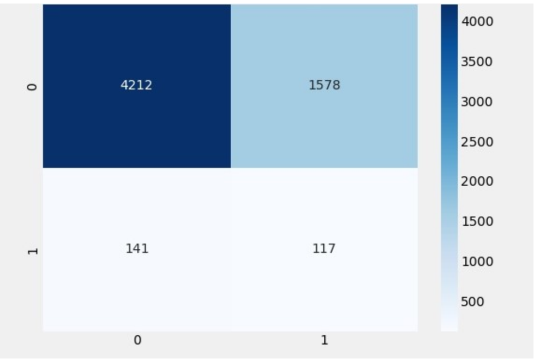

Figure 9: confusion matrix for ridge regression after SMOTE NC

Figure 10 represents the feature importance for this model. The main thing we notice is the significant decrease in coefficient's magnitude due to applying "L2 Regularization" technique. For this model, the 2 most powerful features in decreasing the log odds of owning an AFV vehicle are "Race African" and "Home Ownership" features. In other words, individual's belonging to the African ethnic group have lower chance to own an AFV vehicle then those belonging to the white ethnic group (reference dummy variable), same as for those owing a home rather than renting one.

With Polynomial Features

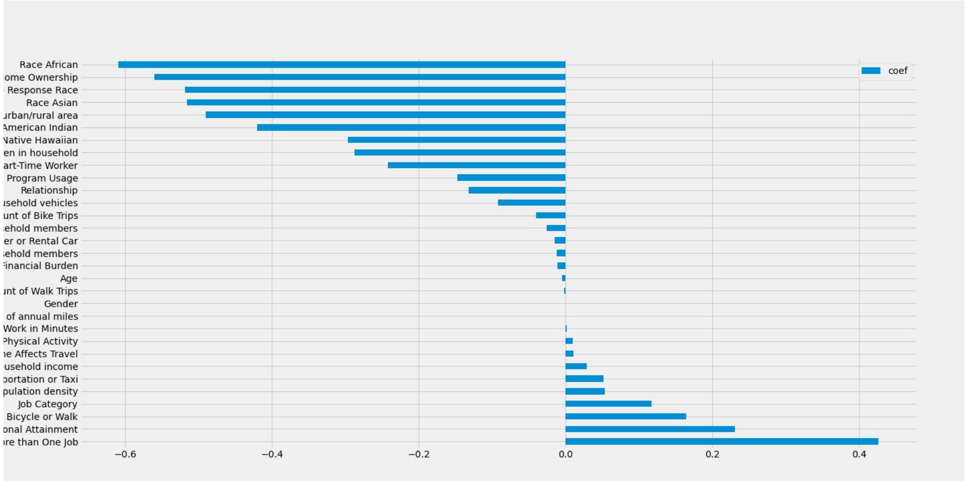

Figure 10: Feature importance in Ridge Regression after SMOTE NC

After adding polynomial features with interactions, we applied grid search cross validation and the optimal λ value was 1.5. Figure 11 represents the confusion matrix for this model where we observe an increase in the model's accuracy but a decrease in the recall score while preserving the same AUC score. Indeed, this had an AUC score of 0.59 accompanied with an improved accuracy score of 0.79 but a recall score of 0.37.

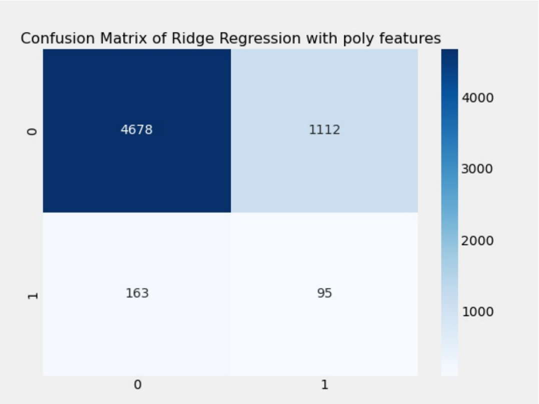

Figure 11: confusion matrix for ridge regression with polynomial features

6.2.4 Lasso regression

For Lasso regression models, the hyper-parameter λ was chosen based on the model's information criterion that can be defined as follows:

$$AIC = 2K - 2\ln(\hat{L}) \tag{8}$$

$$BIC = K\ln(n) - 2\ln(\hat{L}) \tag{9}$$

- AIC: Akaike information criterion.
- BIC : Bayesian information criterion.
- K : Number of estimated parameters.
- n : Number of observations.
- $\hat{L}$: Maximum values of the Likelihood function of the model.

Without Polynomial Features

Figure 23 in appendix represents the selection procedure of the hyper-parameter λ based on the information criterion. After optimal λ was chosen, we applied lasso regression. Figure 12 represents the confusion matrix of lasso regression model after SMOTE NC. The model's performance is similar to Logistic regression and Ridge regression models. Indeed, this model got an AUC score of 0.59 with a precision of 0.72 and a recall of 0.42.

Figure 6 shows the feature importance for this model. We see that the features impacting the odds of adopting alternative fuel vehicle either in a positive or negative way are the same as for the ridge regression model. However, the difference we notice is that the coefficient magnitude is lower than ridge and logistic models.

With Polynomial Features

Same as for logistic and ridge regression models, we added polynomial features with interactions to test their impact on model's performance. Figure 24 in appendix shows the selection procedure of the penalization term value based on the information criterion. We notice a significant increase in model's capability to predict the true positive values by having a Recall score of 0.81. However, this increase was on the expense to a great decrease

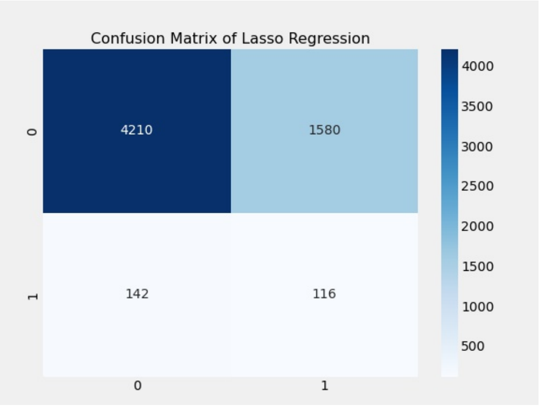

Figure 12: confusion matrix for lasso regression

in model's accuracy that became 0.21 and the AUC score that became 0.49 (figure 14).

6.2.5 Logistic GAM

Logistic GAM model is able to capture the non-linear relationships between the features and the log-odds of adopting an AFV vehicle. Figure 25 in appendix represents the plots of non-linear components that were produced by this model, we observe that in most features there is non strong non linear relationship with the target variable. Figure 15 shows that the GAM model had a better performance in predicting the true positive values while to that was on the cost of a big decrease in the model's accuracy. Indeed, logistic GAM model got an AUC score of 0.61 along with a recall score of 0.54 and an accuracy of 0.67.

6.2.6 Random Forest

Despite the creation of new artificial observations in the training set, random forest model had poor performance in predicting the true positive values in the testing set as seen in the figure 16, this substandard performance manifests by having a recall score of 0.12. With regards to feature importance in this model, we see that the 3 most valuable features impacting the choice of vehicle type are the annual mileage driven, the educational attainment and the job category as presented in figure 27.

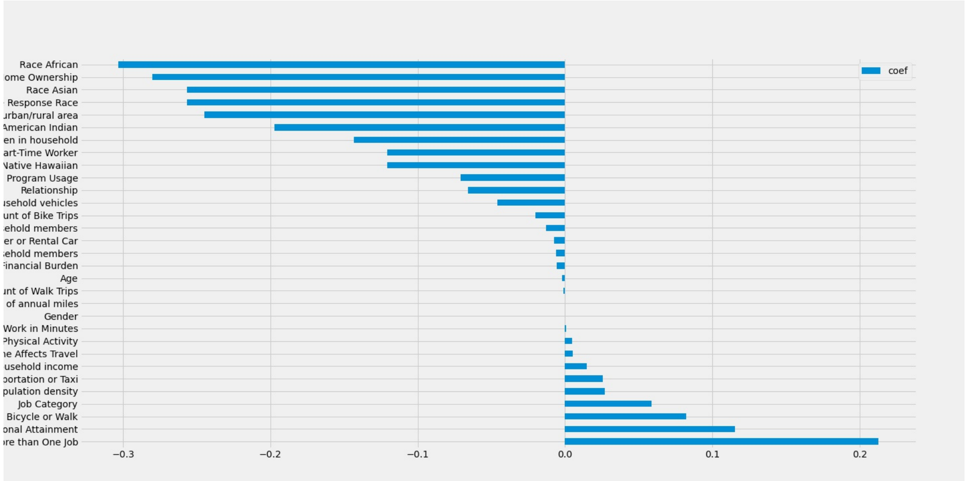

Figure 13: Features Importance For Lasso Regression Model

6.2.7 Naive Bayes

The confusion matrix of Naive Bayes model presented in figure 17 shows a good performance for this model in predicting the true positive values by comparison with the previous models. However, we observe a poor performance with regards to the true negative value. In fact, this model got an AUC score of 0.6 along with a Recall score of 0.68 and an Accuracy of 0.52. Despite that, this model got a good recall score while preserving the AUC score, however one of its major drawback is its lack of interpretability.

6.2.8 Support Vector Machine

SVM model had a high recall score of 0.71 in the testing set but it was in exchange for low precision score as presented in figure 18. This model had an AUC score of 0.49 which is a bad indicator of it's predictive performance.

6.2.9 K Nearest Neighbors

Figure 19 show the poor performance of K-nearest neighbors model. We notice that this model's performance slightly increased after applying SMOTE NC, however still had a low AUC score of 0.52 and a recall score of 0.29.

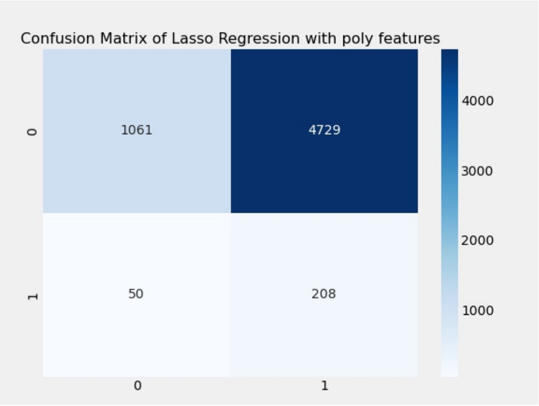

Figure 14: confusion matrix for polynomial lasso regression model

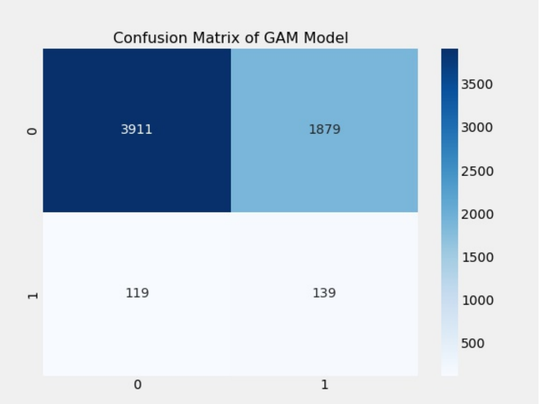

Figure 15: confusion matrix for logistic GAM Model

6.2.10 Neural Network

In the neural network model, we have utilized the weight parameter that gives greater weight to a class during the model training process. Even though we observe a big decrease in the loss function for both train and validation set as shown in figure 26, however this performance wasn't the case in the testing set as we observe in figure 20 where we observe that this model was able to predict all true positive cases, however, it failed in most true negative ones.

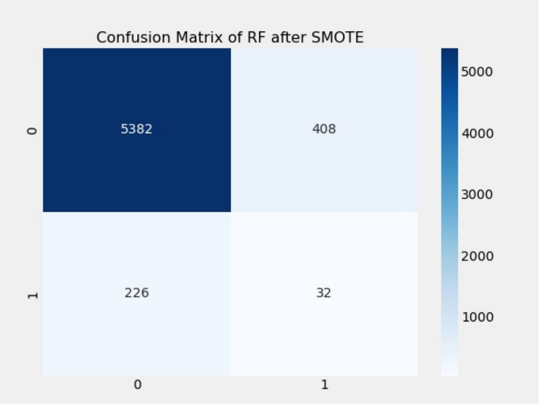

Figure 16: confusion matrix for Random Forest Model

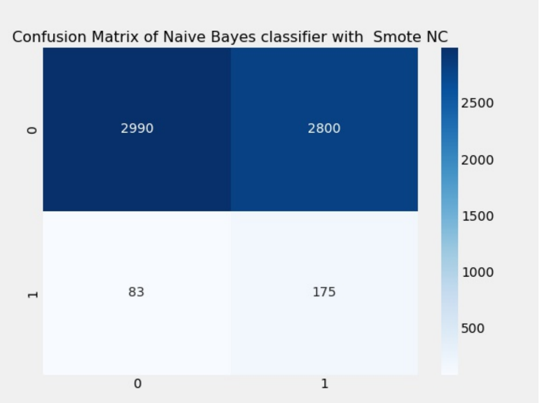

Figure 17: confusion matrix for Naive Bayes Model

6.2.11 Model performance comparisons after SMOTE NC

Table 2 is a summary of all model's performance after applying SMOTE NC. We conclude that the linear econometric models outperformed all non-parametric machine learning models by having the highest AUC score for exception of Naive Bayes model that has a disadvantage in the lack of interpretability. The best performing model was the Logistic GAM model with an AUC of 0.61, followed by both Logistic Regression and Logistic Regression with FAMD having an AUC of 0.6. Linear models utilizing the information reduction techniques such as Logistic regression with FAMD, Ridge and Lasso also had a similar performance. However, polynomial features that adds non-

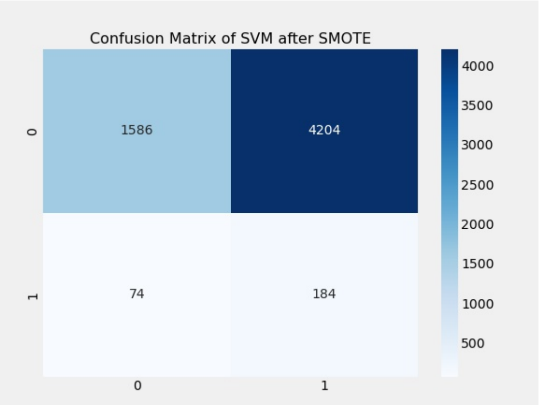

Figure 18: confusion matrix for SVM Model

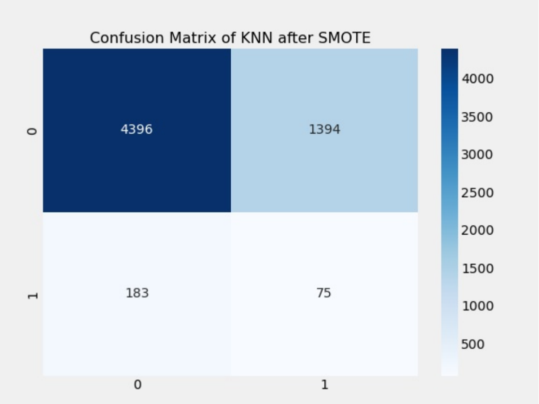

Figure 19: confusion matrix for KNN Model

linearity to these models didn't help in increasing their performance. We can conclude by that and by the out performance of the parametric models over non-parametric models that the relationship between the features and the target variable is more close to a linear relationship which was affirmed by the output plots of Logistic GAM model (25).

7 Conclusion

In this research we explored and analysed the behavior of adopting an alternative fuel vehicle by utilizing parametric and non-parametric machine

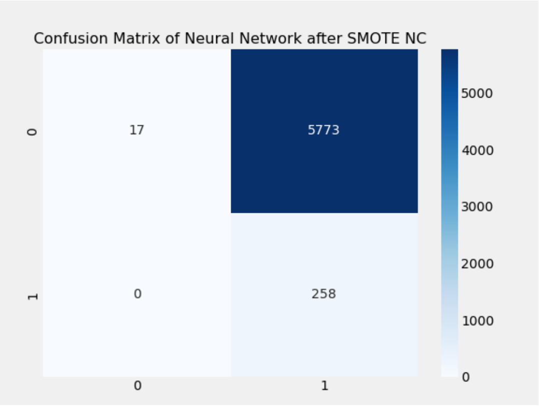

Figure 20: confusion matrix for Neural Network Model

learning models where the feature set is composed of individual and household characteristics. The main challenge in this work was having an imbalanced data where the individuals adopting AFV represent only 4% of the total observations. All models had an AUC score close to 0.5 in the testing set before creating artificial observations. After applying SMOTE NC technique to the training set only, the predictive power of models in the testing set increased. Econometric parametric models surpassed all the non-parametric models for exception to Naive Bayes. Another challenge faced in this work is multicollinearity problem where we observed irregular feature coefficients magnitude in the estimation of Logistic Regression model. However, this problem was solved by applying regularization techniques models such as Ridge and Lasso that gave a realistic feature coefficient magnitude that would serve well in the interpretability aspect, but it didn't help in increasing the predictive performance. Additionally, we added polynomial features to the parametric models but this non-linearity added didn't improve the model's performance. We derived a conclusion that the out-performance of the parametric models over the non-parametric models is explained by the fact that the relationship between features and target variable doesn't contain strong non-linearity, and that manifested in the plots presented by the GAM model (25). In the end, we also conclude that the model's predictive performance is still low by having an AUC score of 0.61 in the best case scenario and that can be explained by the fact that individual and household characteristics aren't sufficient features to predict the individual's behavior to adopt an AFV. Indeed, such behavior would need additional features such

Table 2: Prediction Evaluation Metrics After SMOTE NC for testing Data

Classifiers	Accuracy	Recal	F1	Precision	AUC
Logistic regression	0.73	0.47	0.13	0.07	0.60
Polynomial Logistic Regression	0.96	0.00	0.00	0.00	0.50
Logistic regression with FAMD	0.53	0.67	0.11	0.06	0.60
Lasso regression	0.72	0.45	0.12	0.07	0.59
Polynomial Lasso Regression	0.78	0.36	0.12	0.07	0.58
Ridge regression	0.72	0.45	0.12	0.07	0.59
Polynomial Ridge Regression	0.79	0.37	0.13	0.08	0.59
GAM Model	0.67	0.54	0.12	0.07	0.61
Naive Bayes Model	0.52	0.68	0.11	0.06	0.60
K-Nearest Neighbors	0.74	0.29	0.09	0.05	0.52
Neural Network	0.06	0.97	0.08	0.04	0.49
Support Vector Machine	0.29	0.71	0.08	0.04	0.49
Random Forest	0.90	0.12	0.09	0.07	0.53

as information about price range, maintenance of the vehicle, battery quality, driving range etc. With that being said, it would be interesting in the future to create a model combining features from all these aspects in order to be able to predict accurately this behavior that is considered essential towards shifting to cleaner and more sustainable energy.

References

[1] "Alternative Fuels Data Center: Maps and Data". In: (). URL: https://afdc.energy.gov/data/.

[2] Sanya Carley et al. "Intent to purchase a plug-in electric vehicle: A survey of early impressions in large US cites". In: *Transportation Research Part D: Transport and Environment* 18 (2013), pp. 39–45.

[3] Ricardo A Daziano and Denis Bolduc. "Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model". In: *Transportmetrica A: Transport Science* 9.1 (2013), pp. 74–106.

[4] Ona Egbue and Suzanna Long. "Barriers to widespread adoption of electric vehicles: An analysis of consumer attitudes and perceptions". In: *Energy policy* 48 (2012), pp. 717–729.

[5] Kelly Sims Gallagher and Erich Muehlegger. "Giving green to get green? Incentives and consumer adoption of hybrid vehicle technology". In: *Journal of Environmental Economics and management* 61.1 (2011), pp. 1–15.

[6] "Greenhouse Effect 101 | NRDC". In: (). URL: https://www.nrdc.org/stories/greenhouse-effect-%20101.

[7] Jianmin Jia. "Analysis of alternative fuel vehicle (AFV) adoption utilizing different machine learning methods: a case study of 2017 NHTS". In: *IEEE Access* 7 (2019), pp. 112726–112735.

[8] Wenbo Li et al. "A review of factors influencing consumer intentions to adopt battery electric vehicles". In: *Renewable and Sustainable Energy Reviews* 78 (2017), pp. 318–328.

[9] Ingrid Moons and Patrick De Pelsmacker. "Emotions as determinants of electric car usage intention". In: *Journal of Marketing Management* 28.3-4 (2012), pp. 195–237.

[10] "National Household Travel Survey". In: (). URL: https://nhts.ornl.gov/.

[11] Jose Manuel Pereira, Mario Basto, and Amelia Ferreira da Silva. "The logistic lasso and ridge regression in predicting corporate failure". In: *Procedia Economics and Finance* 39 (2016), pp. 634–641.

[12] William Sierzchula et al. “The influence of financial incentives and other socio-economic factors on electric vehicle adoption”. In: *Energy policy* 68 (2014), pp. 183–194.

[13] “Sources of Greenhouse Gas Emissions | US EPA.””. In: (). URL: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions.

[14] Gil Tal and Michael A Nicholas. “Studying the PEV market in California: Comparing the PEV, PHEV and hybrid markets”. In: *2013 World Electric Vehicle Symposium and Exhibition (EVS27)*. IEEE. 2013, pp. 1–10.

[15] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

A Appendix

A.1 Supplement Data Exploration

Table 3: Summary Statistics for categorical features

Variable	Outcome	Count	Percent
AFV	0	19277	95.62
	1	883	4.38
Household income	6	3388	16.81
	7	3369	16.71
	8	3276	16.25
	11	2505	12.43
	10	2292	11.37
	9	2090	10.37
	5	1687	8.37
	4	813	4.03
	3	438	2.17
	2	152	0.75
	1	150	0.74
Home Ownership	1	16353	81.12
	2	3807	18.88
Household in urban/rural area	1	15855	78.65
	2	4305	21.35
population density	6	4618	22.91
	5	4025	19.97
	2	3256	16.15
	4	2797	13.87
	1	2223	11.03
	3	1843	9.14
	7	1124	5.58
	8	274	1.36
Price of Gasoline Affects Travel	2	5572	27.64
	4	5178	25.68
	3	3708	18.39
	5	2897	14.37
	1	2805	13.91

Travel is a Financial Burden	3	7087	35.15
	2	5219	25.89
	4	4884	24.23
	5	1597	7.92
	1	1373	6.81
Gender	1	10632	52.74
	0	9528	47.26
Educational Attainment	4	6315	31.32
	5	5839	28.96
	3	5592	27.74
	2	2288	11.35
	1	126	0.62
More than One Job	2	18396	91.25
	1	1764	8.75
Job Category	4	12630	62.65
	1	3399	16.86
	2	2363	11.72
	3	1768	8.77
Relationship	1	13914	69.02
	2	5174	25.66
	3	733	3.64
	7	139	0.69
	6	73	0.36
	4	66	0.33
	5	61	0.30
Full-Time or Part-Time Worker	1	17450	86.56
	2	2710	13.44
Public Transportation or Taxi	4	9659	47.91
	2	4435	22.00
	3	3882	19.26
	1	2184	10.83
Level of Physical Activity	2	12467	61.84
	3	5778	28.66
	1	1915	9.50
Passenger to Friend/Family Member or Rental Car	1	7424	36.83
	4	6062	30.07
	3	5399	26.78
	2	1275	6.32

Bicycle or Walk	4	13626	67.59
	3	2952	14.64
	2	2845	14.11
	1	737	3.66
Race White	1	17423	86.42
	0	2737	13.58
Race African	0	19086	94.67
	1	1074	5.33
Race Asian	0	19128	94.88
	1	1032	5.12
Race American Indian	0	20067	99.54
	1	93	0.46
Race Native Hawaiian	0	20114	99.77
	1	46	0.23
Multiple Response Race	0	19668	97.56
	1	492	2.44

Table 4: Summary Statistic For numerical Features

	mean	std	min	max	50%
Count of household members	2.52	1.22	1.0	12.0	2.00
Age	47.50	13.30	18.0	92.0	49.00
Count of Car Share Program Usage	0.01	0.17	0.0	8.0	0.00
Trip Time to Work in Minutes	27.70	27.81	1.0	600.0	20.00
Count of household vehicles	2.46	1.24	1.0	12.0	2.00
count of children in household	0.13	0.41	0.0	3.0	0.00
Best estimate of annual miles	14651.56	15697.54	0.0	200000.0	12082.69
Count of Walk Trips	5.29	7.53	0.0	89.0	3.00
Count of Bike Trips	0.22	1.10	0.0	70.0	0.00
Count of adult household members	2.00	0.75	1.0	7.0	2.00

Table 5: Variance Inflation Factor Analysis

	Feature	VIF
0	Household income	19.592299
1	Home Ownership	11.415045
2	Count of household members	13.772372
3	Household in urban/rural area	16.251976
4	population density	10.410594
5	Price of Gasoline Affects Travel	9.937182
6	Travel is a Financial Burden	13.482565
7	Age	16.624895
8	Gender	2.237904
9	Educational Attainment	19.784738
10	More than One Job	37.911517
11	Job Category	10.140491
12	Count of Car Share Program Usage	1.006758
13	Trip Time to Work in Minutes	2.064529
14	Best estimate of annual miles	1.901082
15	Count of household vehicles	8.014692
16	count of children in household	1.536414
17	Relationship	4.919301
18	Full-Time or Part-Time Worker	12.773288
19	Count of Walk Trips	1.566344
20	Count of Bike Trips	1.073221
21	Level of Physical Activity	14.967480
22	Count of adult household members	22.234177
23	Public Transportation or Taxi	9.708427
24	Passenger to Friend/Family Member or Rental Car	5.011663
25	Bicycle or Walk	17.142553
26	Race African	1.110708
27	Race Asian	1.129838
28	Race American Indian	1.009115
29	Race Native Hawaiian	1.007079
30	Multiple Response Race	1.037149

A.2 Supplement Machine Learning Models

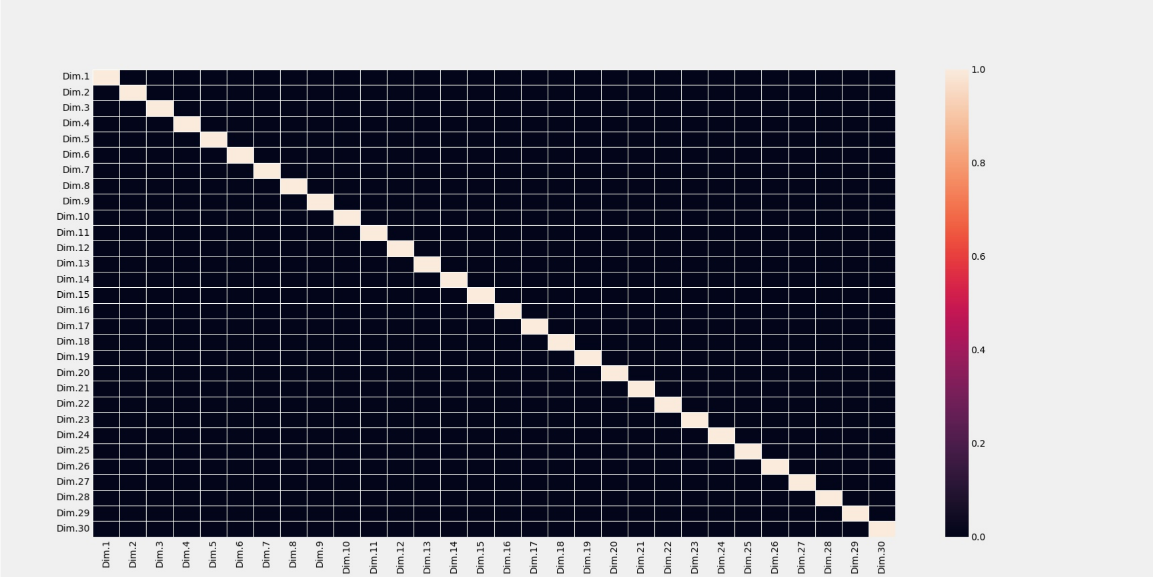

Figure 21: Correlation matrix for FAMD components

Table 6: Cumulative explained variance percentage

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.95	4.28	4.28
Dim.2	2.39	3.47	7.75
Dim.3	1.91	2.77	10.52
Dim.4	1.60	2.33	12.84
Dim.5	1.48	2.15	14.99
Dim.6	1.44	2.09	17.08
Dim.7	1.38	2.00	19.08
Dim.8	1.31	1.90	20.98
Dim.9	1.29	1.87	22.85
Dim.10	1.23	1.79	24.63
Dim.11	1.20	1.73	26.37
Dim.12	1.19	1.72	28.09

Dim.13	1.17	1.69	29.78
Dim.14	1.15	1.66	31.45
Dim.15	1.14	1.65	33.09
Dim.16	1.13	1.63	34.73
Dim.17	1.10	1.59	36.32
Dim.18	1.09	1.58	37.90
Dim.19	1.08	1.57	39.47
Dim.20	1.06	1.54	41.01
Dim.21	1.06	1.54	42.55
Dim.22	1.05	1.53	44.08
Dim.23	1.05	1.52	45.60
Dim.24	1.04	1.50	47.10
Dim.25	1.03	1.50	48.60
Dim.26	1.03	1.49	50.09
Dim.27	1.02	1.48	51.57
Dim.28	1.01	1.46	53.03
Dim.29	1.01	1.46	54.49
Dim.30	1.00	1.45	55.94
Dim.31	0.99	1.44	57.38
Dim.32	0.99	1.43	58.81
Dim.33	0.99	1.43	60.24
Dim.34	0.98	1.42	61.66
Dim.35	0.97	1.40	63.06
Dim.36	0.96	1.39	64.45
Dim.37	0.96	1.39	65.84
Dim.38	0.95	1.37	67.21
Dim.39	0.94	1.36	68.58
Dim.40	0.93	1.35	69.93
Dim.41	0.93	1.34	71.27
Dim.42	0.92	1.34	72.61
Dim.43	0.91	1.32	73.93
Dim.44	0.91	1.32	75.24
Dim.45	0.90	1.31	76.55
Dim.46	0.89	1.29	77.84
Dim.47	0.89	1.29	79.13
Dim.48	0.88	1.27	80.40
Dim.49	0.84	1.22	81.62
Dim.50	0.84	1.21	82.83

Dim.51	0.82	1.19	84.02
Dim.52	0.82	1.18	85.20
Dim.53	0.81	1.17	86.37
Dim.54	0.80	1.16	87.53
Dim.55	0.78	1.14	88.67
Dim.56	0.78	1.13	89.80
Dim.57	0.73	1.06	90.86
Dim.58	0.72	1.04	91.91
Dim.59	0.70	1.01	92.92
Dim.60	0.67	0.97	93.89
Dim.61	0.64	0.92	94.81
Dim.62	0.59	0.85	95.66
Dim.63	0.54	0.78	96.45
Dim.64	0.51	0.74	97.18
Dim.65	0.48	0.70	97.88
Dim.66	0.47	0.68	98.56
Dim.67	0.46	0.66	99.22
Dim.68	0.31	0.45	99.67
Dim.69	0.23	0.33	100.00

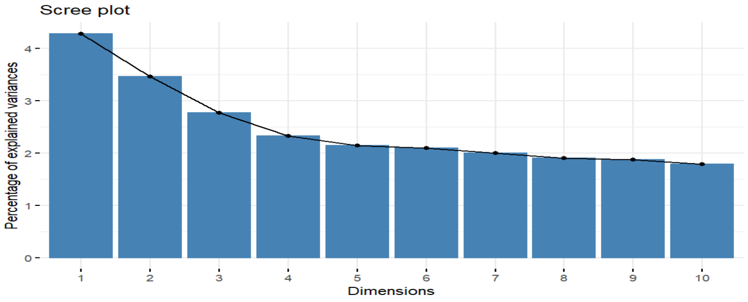

Figure 22: Histogram of Cumulative explained variance

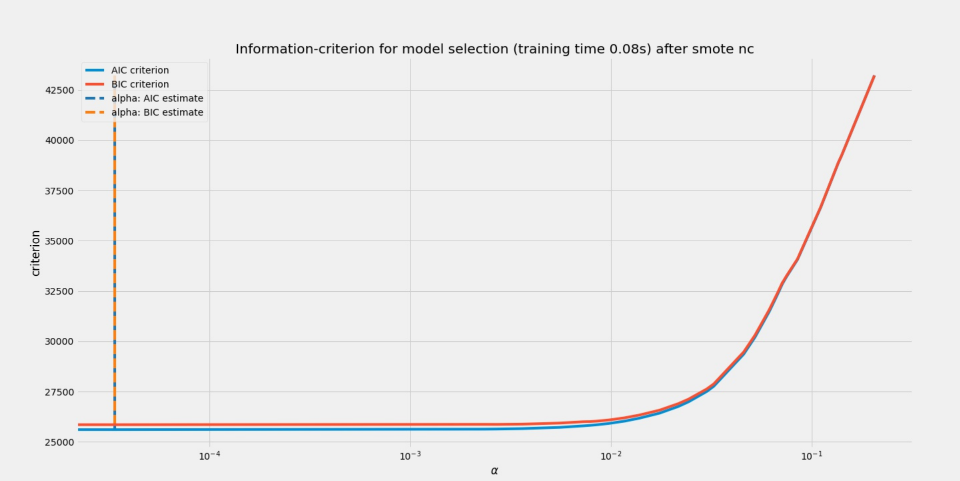

Figure 23: Selecting λ based on information criterion for lasso regression model

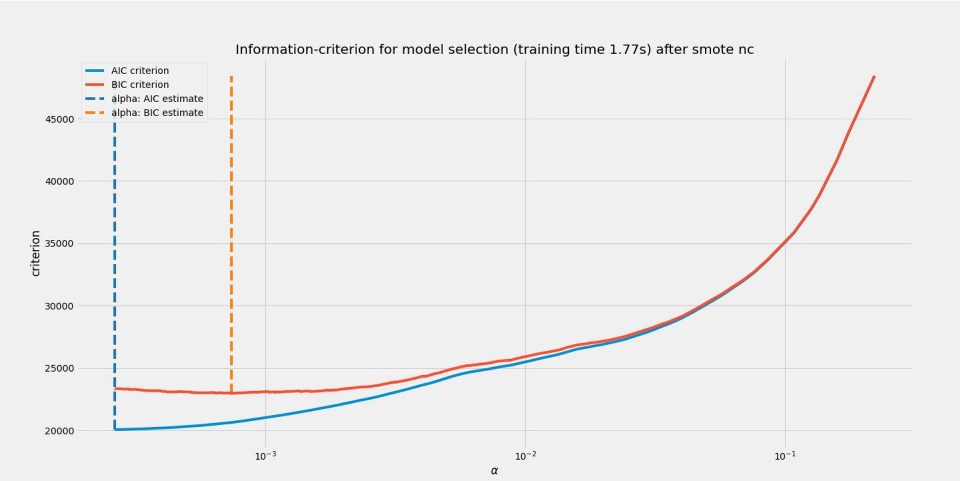

Figure 24: Selecting λ based on information criterion for polynomial lasso regression

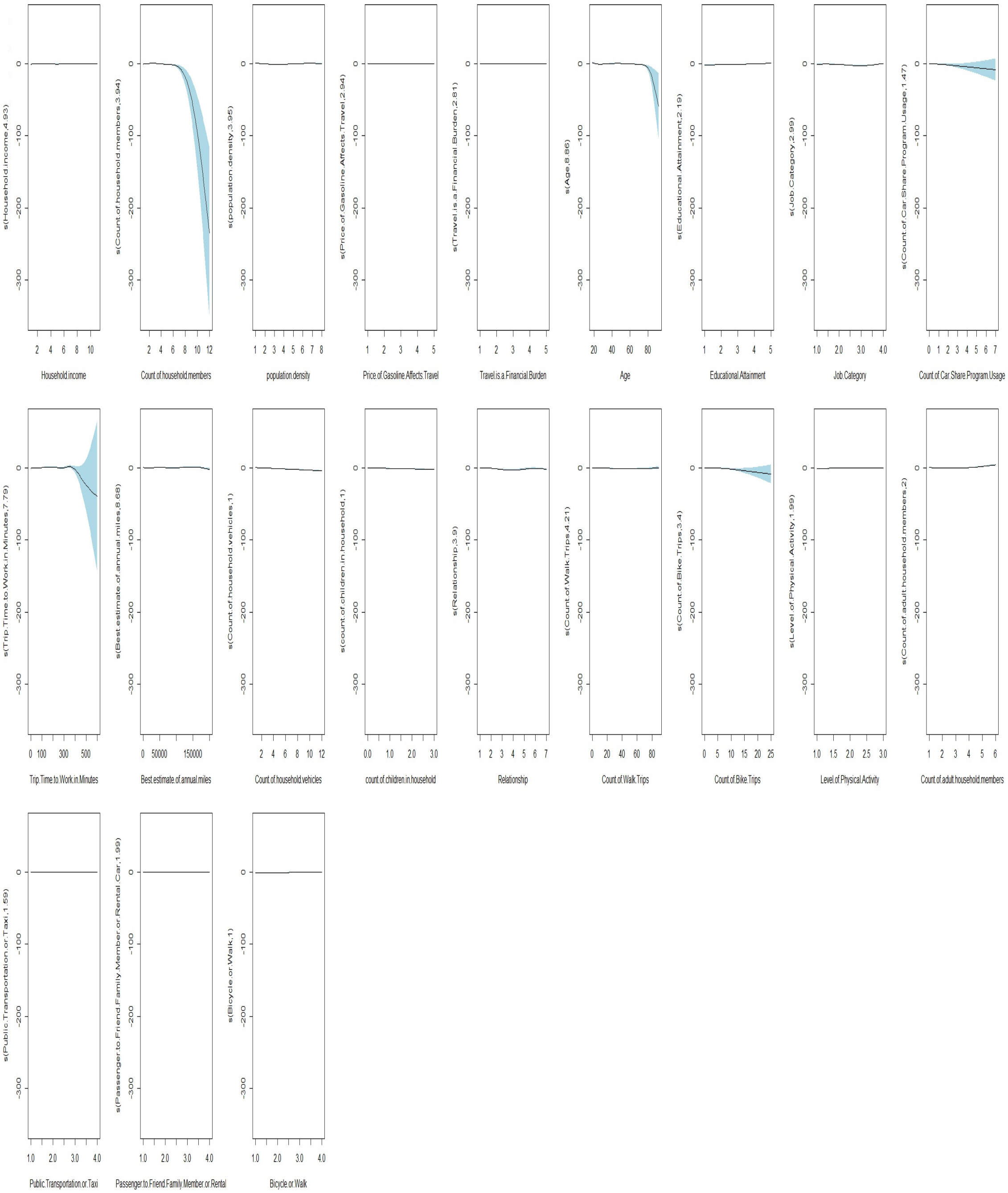

Figure 25: Non-linear components plots

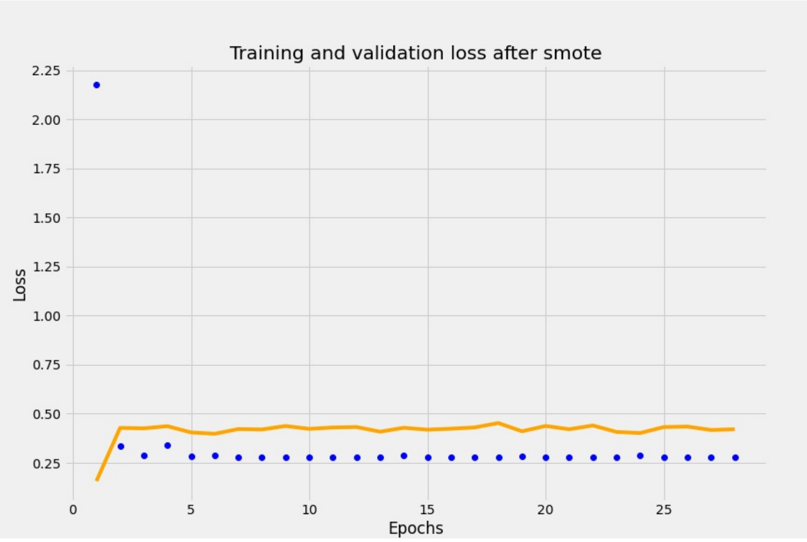

Figure 26: Train and validation loss score for neural network after smote

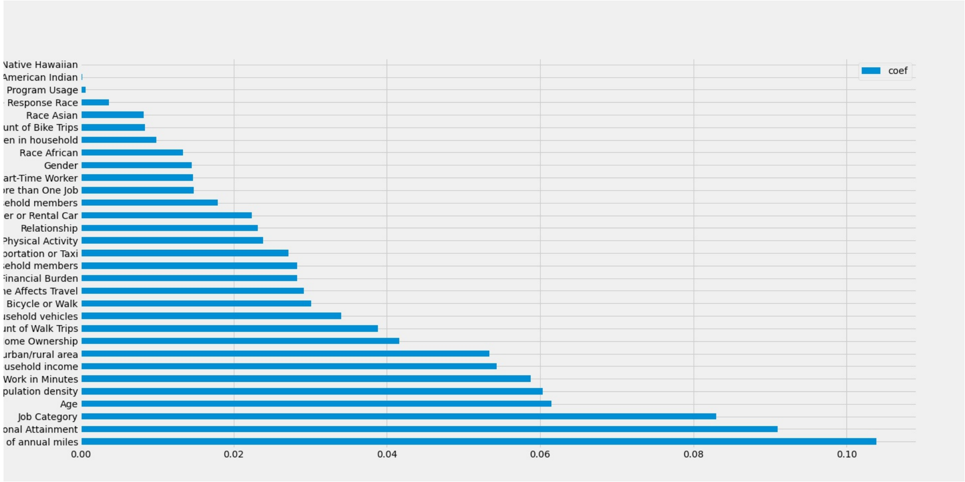

Figure 27: Feature Importance Random Forest