CS 7641 Project Proposal (Team 52)

Identifying Likely Adopters of Alternative Fuel Vehicles (AFVs): A Model Comparison Using 2022 National Household Travel Survey

Introduction

Alternative fuel vehicles (AFVs) represent a critical pathway toward sustainable transportation solutions, significantly reducing carbon dioxide emissions and improving energy efficiency. With electric motors boasting a remarkable 90% efficiency compared to the 20% efficiency of traditional internal combustion engines, AFVs offer an opportunity to transform one of the most energy-intensive sectors (Hanley, 2018). The transportation sector alone accounts for 70% of petroleum use and 30% of total U.S. energy consumption (Department of Energy, n.d.), underscoring its role as a focal point for environmental policy and innovation. Transitioning to AFVs mitigates transportation's environmental impacts and reduces dependency on finite fossil fuels, aligning with broader sustainability and energy security goals.

The rapid growth in AFV adoption highlights the importance of understanding the factors driving this transition. Between 2017 and 2023, AFVs saw a substantial increase in market share, rising from 2.21% to 16.05% (Montoya, 2024). This surge signals a significant shift in consumer preferences, technological advancements, and policy impacts, necessitating a deeper exploration of the variables influencing AFV adoption. In this background, this project seeks to examine a range of factors—from sociodemographic and economic influences to infrastructural factors—that contribute to individual’s adoption of AFVs in the U.S. Additionally, the study aims to evaluate various predictive models and methods to identify the most effective approaches for capturing these dynamics, providing insights for stakeholders seeking to accelerate the transition to sustainable transportation systems.

Problem Definition

- **Factors affecting AFV adoption**: What are the key factors that affect the adoption of AFVs? We will investigate various factors, including individual (e.g., travel patterns and household income) and regional (e.g., urban, infrastructure, and gas price) level factors.
- **Dimensionality reduction using PCA**: What factors have similar or different impacts on AFV adoption? We will employ Principal Component Analysis (PCA) to reduce the dimensionality of our dataset, as well as classify the variables into some categories. It includes a process of defining what can be explained from the clusters of factors.
- **Evaluation of model performance**: How do different models perform in predicting AFV adoption? We will compare and evaluate various predictive models, such as binary

logistic regression, Support Vector Machines (SVM), Decision Tree (DT), and Random Forest (RF).

Data

We utilized the National Household Travel Survey (https://nhts.ornl.gov/), conducted by the Federal Highway Administration. This dataset contains over 30,000 data points and 80 variables on American travel behavior, demographics, household characteristics, and vehicle information (Westin et al., 2018; Davis, 2023). The 2022 data expands on prior studies and provides a basis for applying machine-learning models to analyze AFV adoption identifiers.

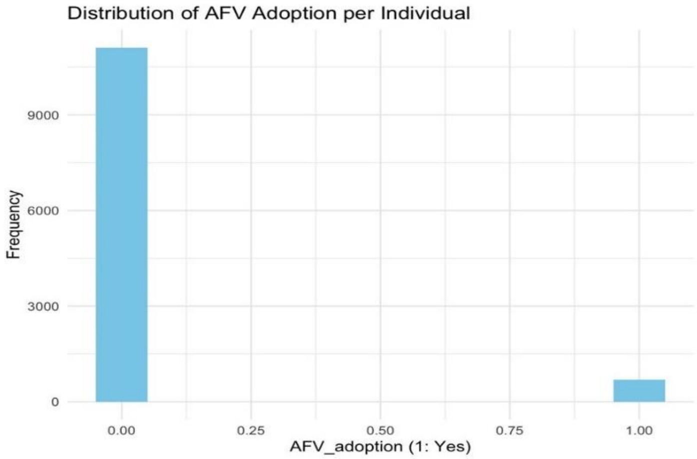


Figure 1. Distribution of AFV Adoption per individual, 2022 NHTS

Methods

1) Data Preprocessing

- **Feature Engineering:** We created a binary indicator 'AFV_adoption' representing whether individuals adopted an AFV and is the main driver of the car.
- **Handling Missing Data**: We imputed two variables ANNMILES (estimated annual travel distance), and GASPRICE. ANNMILES was input using the K-nearest Neighbors approach, HHSIZE (number of household members), HHVEHCNT (number of vehicles in household), and PPT517 (number of children in household) as predictor variables. GCDWORK was omitted due to many missing values.
- **Normalization/Standardization**: We normalized variables with skewed distributions – VEHAGE (vehicle age), HHSIZE, HHVEHCNT, NUMADLT (number of adults in household), PPT517, and WRKCOUNT (number of working household members) - using Min-Max scaling method.

- **Synthetic Minority Over-sampling Technique, SMOTE** (Imbalanced-learn): Since the adoption rate in our data is as low as 1.2%, it is concerned our models might contain biases towards the majority (non-adopters), which can result in skewed decision boundaries (Chawla et al., 2002). To deal with this problem, a new sample will be created using SMOTE to reduce overfitting of the classifiers.

2) Principal Component Analysis (PCA) - *Unsupervised Learning*

Principal Component Analysis (PCA) is employed to address the challenge of redundancy in datasets, where variables may exhibit significant correlation or overlapping information. By transforming the original data into a smaller set of uncorrelated components, PCA effectively captures the essential patterns within the data while filtering out noise. This redundancy reduction enables a more streamlined analysis, improving computational efficiency and enhancing the interpretability of results.

In addition to reducing redundancy, PCA ensures that most of the dataset's variability is retained by choosing the number of components that explain over a certain threshold of the total variance. This approach minimizes information loss while significantly reducing dimensionality, allowing for a more efficient representation of the data.

The effectiveness of this approach was validated through a cumulative explained variance plot (Figure 2), which demonstrated that 18 principal components were sufficient to capture the dataset's variability (over 90%), ensuring a balance between data simplification and fidelity. The plot shows that the first 10 principal components capture most of the variance, with 18 components explaining around 90% of the variance. This confirms effective dimensionality reduction without excessive information loss.

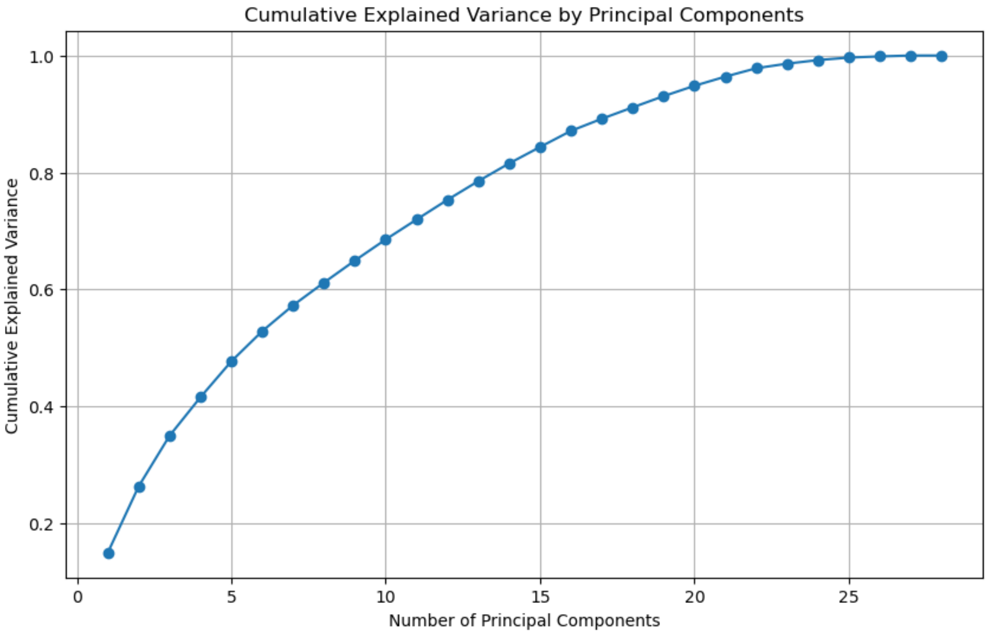


Figure 2. Cumulative Explained Variance by Number of Principal Components

Each principal component was examined for its top contributing features, allowing us to label components with themes like “Household Vehicle Usage” and “Regional Demographics.” This interpretive approach gives us insights into the primary drivers behind AFV adoption trends while maintaining an efficient feature set for modeling.

3) Machine Learning Algorithms for Predicting AFV Adoption:

Using the principal components derived from the dataset and the AFV adoption labels, several machine-learning algorithms were implemented to predict AFV adoption based on sociodemographic characteristics, travel patterns, and regional factors.

Binomial logistic regression, implemented through Scikit-learn, was used to estimate the coefficients of the principal components, providing insights into the relative influence of each factor on AFV adoption.

Additionally, **Naive Bayes, Decision Tree, Support Vector Machine (SVM), and Random Forest classifiers** were employed to classify individuals into AFV adopters and non-adopters based on the transformed features. These models were chosen for their diverse approaches to classification, ranging from probabilistic reasoning (Naive Bayes) to ensemble learning (Random Forest), allowing for a comprehensive comparison of their predictive performance. The analysis aims to identify the most effective model for understanding and predicting AFV adoption patterns.

Results and Discussion

1) Balancing Data using SMOTE

The SMOTE (Synthetic Minority Over-sampling Technique) output demonstrates its effectiveness in addressing the class imbalance in the AFV adoption dataset, where the initial distribution was heavily skewed (7778 instances of class 0 compared to only 463 instances of class 1). After applying SMOTE, the class distribution was balanced, with both classes containing 7778 instances (Table 1). We also visualized the distribution of principal components in the minority class (AFV_adoption == 0) to check if the distribution is retained after SMOTE and found that the distribution is reasonably generated (Figure 3).

Consequently, the training data size increased from 8241 observations with 18 features to 15,556 observations, maintaining the original feature dimensionality while enhancing the dataset's suitability for training predictive models (Table 1).

Table 1. Class Distribution and Training Data Size after SMOTE

		Before SMOTE	After SMOTE
Class Distribution	0 (No AFV)	7,778	7,778
	1 (Has AFV)	463	7,778
Training Data Size		(8241, 18)	(15,556, 18)

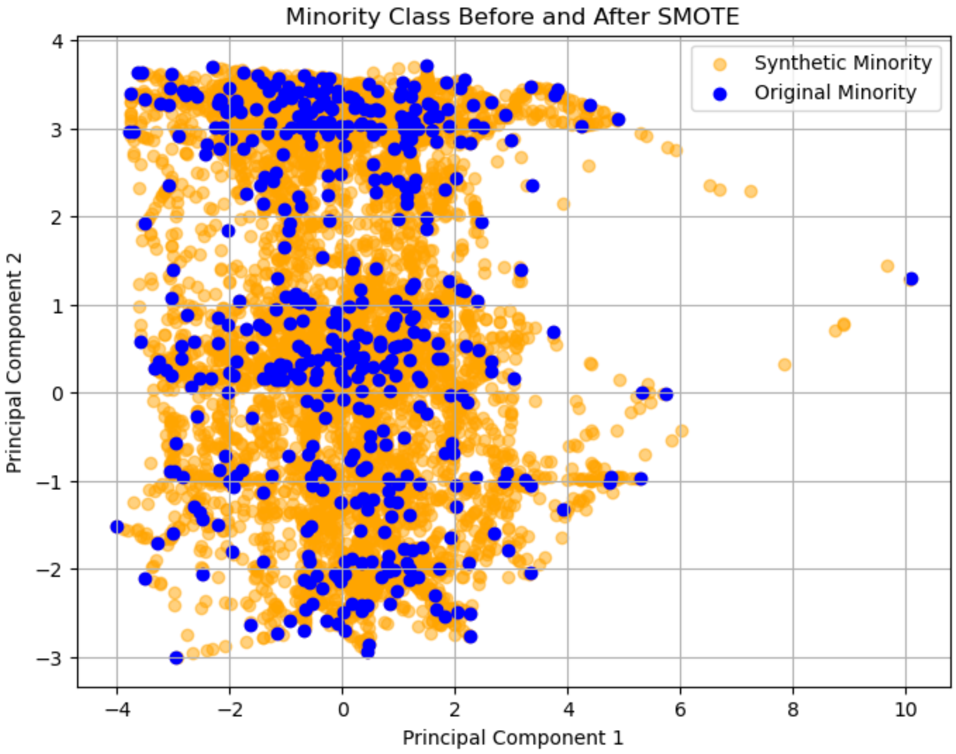


Figure 3. Visualization of PC1 and PC2 after SMOTE

2) Dimension Reduction using PCA

We set the cumulative variance threshold at 0.9 and defined 18 principal components for further analysis. We created labels for each principal component based on its top 5 features with the highest absolute loadings, as shown in Table 2.

Table 2. PCA Results

	Theme / Title	Top Features
PC1	Household Vehicle Usage	DRVRCNT, NUMADLT, HHSIZE, WRKCOUNT, HHVEHCNT
PC2	Regional Demographics	CENSUS_D, CDIVMSAR, CENSUS_R, GASPRICE, URBAN
PC3	Urban vs. Rural Composition	URBAN, WORKER, MSACAT, EDUC, HHFAMINC_IMP
PC4	Employment & Worker Characteristics	LIF_CYC, HOMETYPE, WORKER, HOMEOWN, HHFAMINC_IMP
PC5	Urban vs. Rural Composition	URBAN, MSACAT, URBANSIZE, WORKER, HOMEOWN
PC6	Urban vs. Rural Composition	URBANSIZE, MSASIZE, MSACAT, URBAN, CNTTDTR
PC7	Household Composition	PPT517, HHSIZE, MSASIZE, HHVEHCNT, WRKCOUNT
PC8	Gender & Regional Details	R_SEX_IMP, VEHAGE, VEHTYPE, CNTTDTR, WHOMAIN
PC9	Vehicle Type Preferences	VEHTYPE, VEHFUEL, R_RACE_IMP, VEHAGE, HOMEOWN
PC10	Home & Ownership Attributes	R_RACE_IMP, VEHFUEL, HOMEOWN, CNTTDTR, HOMETYPE
PC11	Vehicle Type Preferences	ANNMILES, VEHTYPE, R_RACE_IMP, VEHFUEL, PPT517
PC12	Travel & Commute Distance	ANNMILES, R_RACE_IMP, CNTTDTR, VEHTYPE, GASPRICE
PC13	Vehicle Type Preferences	CNTTDTR, VEHFUEL, R_SEX_IMP, VEHTYPE, ANNMILES
PC14	Gender & Regional Details	VEHAGE, VEHFUEL, GASPRICE, EDUC, R_SEX_IMP
PC15	Economic Indicators	GASPRICE, CNTTDTR, VEHFUEL, MSASIZE, CENSUS_R
PC16	Gender & Regional Details	EDUC, VEHAGE, R_SEX_IMP, CNTTDTR, VEHTYPE
PC17	Miscellaneous Factors	WHOMAIN, HHFAMINC_IMP, HOMETYPE, CNTTDTR, WORKER
PC18	Home & Ownership Attributes	HOMETYPE, HOMEOWN, LIF_CYC, HHVEHCNT, NUMADLT

3) Prediction of AFV adoption using multiple ML algorithms

The evaluation metrics provide insights into the performance of various machine learning models, including Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine (SVM), and Random Forest, in predicting alternative fuel vehicle (AFV) adoption. These metrics are derived from the original dataset and a SMOTE-balanced dataset to address class imbalance.

Table 3. Model Summary Table

Model	Dataset	TP	FP	FN	TN	TP Rate	TN Rate	Accuracy
LR	Original	225	0	0	3308	1.000	1.000	1.000
Naïve Bayes		219	14	6	3294	0.973	0.996	0.994
Decision Tree		213	6	12	3302	0.947	0.998	0.995
SVM		222	0	3	3308	0.987	1.000	0.999
Random Forest		216	1	9	3307	0.960	0.999	0.997
LR	SMOTE	225	0	0	3308	1.000	1.000	1.000
Naïve Bayes		224	13	1	3295	0.996	0.996	0.996
Decision Tree		211	14	14	3294	0.938	0.996	0.992
SVM		224	1	1	3308	0.996	1.000	0.999
Random Forest		225	0	0	3306	1.000	0.999	0.999

Original Dataset

- **Logistic Regression** achieved perfect performance across all metrics, with precision, recall, and F1-scores of 1.00 for both classes. The model classified all samples correctly, as evidenced by its 100% accuracy, true positive rate (TPR), and true negative rate (TNR).
- **Naïve Bayes** performed slightly below perfect, with a precision of 0.95 and a recall of 1.00 for class 1, resulting in an F1-score of 0.97. The overall accuracy was 99.43%, showing a small number of false positives and false negatives.
- **Decision Tree** showed strong performance, with an F1-score of 0.94 for class 1, but slightly lower recall (0.94) compared to Logistic Regression. It achieved an accuracy of 99.49%.
- **SVM** delivered nearly perfect results, with precision, recall, and F1-scores of 1.00 for both classes. The accuracy was 99.92%, indicating minimal misclassifications.
- **Random Forest** exhibited high performance with a recall of 0.99 and a precision of 1.00 for class 1, leading to an F1-score of 1.00 and an accuracy of 99.72%.

SMOTE-balanced Dataset

- **Logistic Regression** maintained perfect performance across all metrics, achieving 100% accuracy, precision, recall, and F1 scores.
- **Naïve Bayes** improved slightly, achieving a near-perfect recall (0.996) and an accuracy of 99.60%. However, its precision for class 1 dropped marginally to 0.95.
- **Decision Tree** saw a minor decrease in performance, with a recall of 0.937 and an accuracy of 99.21%. This indicates some difficulty in correctly classifying class 1 samples in the balanced dataset.

- **SVM** continued to perform excellently, with precision and recall close to 1.00 for both classes and an accuracy of 99.97%.
- **Random Forest** demonstrated perfect recall and almost perfect precision, resulting in an accuracy of 99.94%.

Logistic Regression and SVM consistently delivered the highest accuracy and reliability in classifying AFV adopters. Random Forest and Naïve Bayes also performed well but showed slight variations in recall and precision for the minority class. The SMOTE-balanced dataset helped improve recall for the minority class across all models, ensuring a more equitable performance and addressing the initial class imbalance.

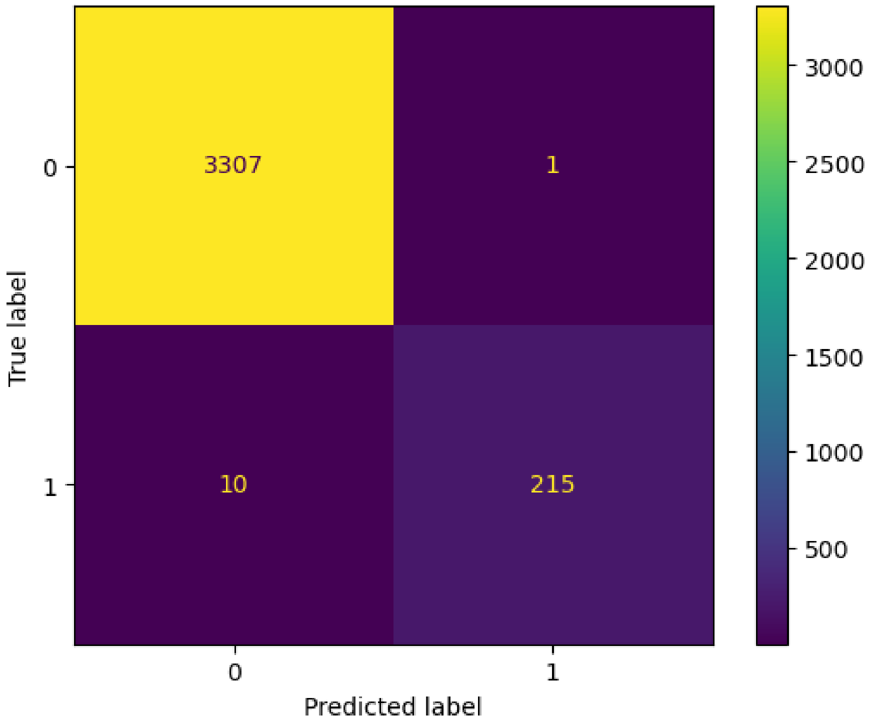


Figure 4. Confusion Matrix

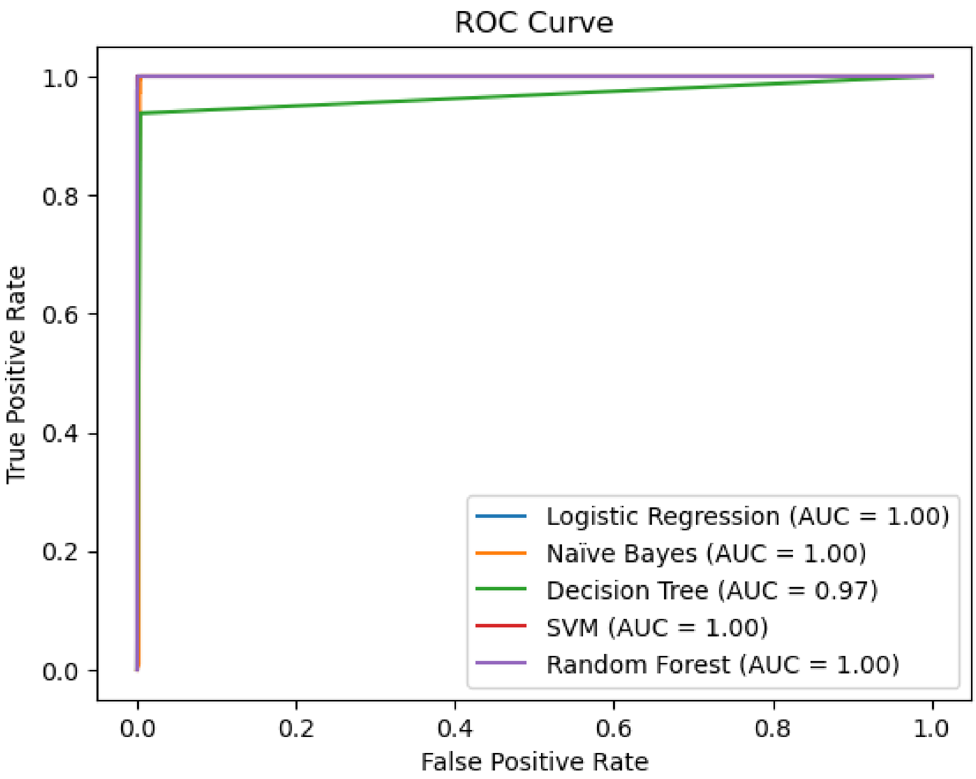
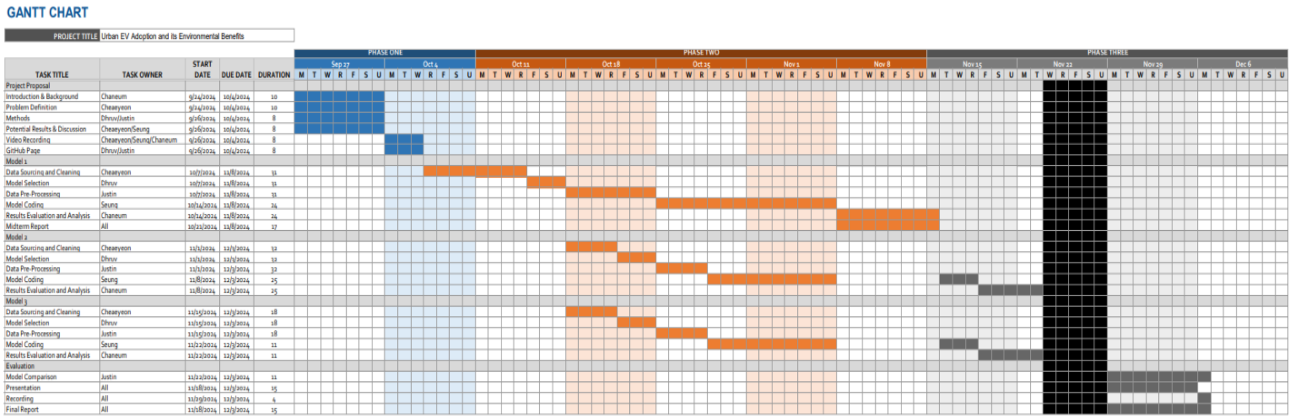


Figure 5. Comparison of ROC Curves

Gantt Chart



PROJECT TITLE Urban EV Adoption and its Environmental Benefits

TASK TITLE	TASK OWNER	START DATE	DUE DATE	DURATION
Project Proposal				
Introduction & Background	Chaneum	9/24/2024	10/4/2024	10
Problem Definition	Cheaeyeon	9/24/2024	10/4/2024	10
Methods	Dhruv/Justin	9/26/2024	10/4/2024	8
Potential Results & Discussion	Cheaeyeon/Seung	9/26/2024	10/4/2024	8
Video Recording	Cheaeyeon/Seunq/Chaneum	9/26/2024	10/4/2024	8
GitHub Page	Dhruv/Justin	9/26/2024	10/4/2024	8
Model 1				
Data Sourcing and Cleaning	Cheaeyeon	10/7/2024	11/8/2024	31
Model Selection	Dhruv	10/7/2024	11/8/2024	31
Data Pre-Processing	Justin	10/7/2024	11/8/2024	31
Model Coding	Seung	10/14/2024	11/8/2024	24
Results Evaluation and Analysis	Chaneum	10/14/2024	11/8/2024	24
Midterm Report	All	10/21/2024	11/8/2024	17
Model 2				
Data Sourcing and Cleaning	Cheaeyeon	11/1/2024	12/3/2024	32
Model Selection	Dhruv	11/1/2024	12/3/2024	32
Data Pre-Processing	Justin	11/1/2024	12/3/2024	32
Model Coding	Seung	11/8/2024	12/3/2024	25
Results Evaluation and Analysis	Chaneum	11/8/2024	12/3/2024	25
Model 3				
Data Sourcing and Cleaning	Cheaeyeon	11/15/2024	12/3/2024	18
Model Selection	Dhruv	11/15/2024	12/3/2024	18
Data Pre-Processing	Justin	11/15/2024	12/3/2024	18
Model Coding	Seung	11/22/2024	12/3/2024	11
Results Evaluation and Analysis	Chaneum	11/22/2024	12/3/2024	11
Evaluation				
Model Comparison	Justin	11/22/2024	12/3/2024	11
Presentation	All	11/18/2024	12/3/2024	15
Recording	All	11/29/2024	12/3/2024	4
Final Report	All	11/18/2024	12/3/2024	15

Contribution Table

Name	Midpoint Contributions
Chaeyeon Han	Final Report writing, Binomial Logistic Regression

Chaneum Park	Binomial Logistic Regression, NB, DT, SVM, Random Forest
Dhruv Modi	PCA, Linear Regression, Random Forest, Presentation
Justin Siegel	PCA, Linear Regression, Random Forest, Presentation
Seung Jae Lieu	Data processing, SMOTE analysis

Reference

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Davis, L. W. (2023). Electric vehicles in multi-vehicle households. Applied Economics Letters, 30(14), 1909-1912. https://www.tandfonline.com/doi/full/10.1080/13504851.2022.2083563

Hanley, S. (2018, March 10). Electric car myth buster — efficiency. CleanTechnica. https://cleantechnica.com/2018/03/10/electric-car-myth-buster-efficiency/?utm_source=chatgpt.com

Jia, J. (2019). Analysis of alternative fuel vehicle (AFV) adoption utilizing different machine learning methods: a case study of 2017 NHTS. IEEE Access, 7, 112726-112735. https://ieeexplore.ieee.org/abstract/document/8794814

Jia, J., Shi, B., Che, F., & Zhang, H. (2020). Predicting the regional adoption of electric vehicle (EV) with comprehensive models. IEEE Access, 8, 147275-147285. https://ieeexplore.ieee.org/abstract/document/9162026

Ronald Montoya. (2024). “How many Electric Cars Are There in the U.S.?”. https://www.edmunds.com/electric-car/articles/how-many-electric-cars-in-us.html

U.S. Department of Energy. “Electric Vehicle Benefits and Considerations”. https://afdc.energy.gov/fuels/electricity-benefits

Westin, K., Jansson, J., & Nordlund, A. (2018). The importance of socio-demographic characteristics, geographic setting, and attitudes for adoption of electric vehicles in Sweden. Travel Behaviour and Society, 13, 118-127. https://www.sciencedirect.com/science/article/pii/S2214367X17300169