Analyzing the Adoption and Usage of Electric Vehicles in Washington State

Abstract

This study analyzes the adoption and usage of electric vehicles in Washington State and develops a predictive model to estimate the electric range of an electric vehicle based on its attributes. The findings suggest that battery capacity and vehicle type are among the most important factors influencing the electric range of an electric vehicle. The dataset used in this study is well-standardized, resulting in a high accuracy of 98.5% for the classifiers on the first attempt.

However, caution is advised in interpreting these results, as the dataset only includes information on electric vehicles in Washington State and does not provide comparative data on the usage patterns of conventional gasoline-powered vehicles. Further research is needed to validate the results and examine the electric vehicle market on a more global scale. Nonetheless, this study provides valuable insights into the adoption and usage of electric vehicles and can aid in the development of more sustainable and energy-efficient transportation options.

Introduction

Electric vehicles have gained significant popularity in recent years due to increasing concerns about the environmental impact of traditional gasoline-powered vehicles and the need to reduce greenhouse gas emissions. In addition, advancements in battery technology have led to improvements in the range and performance of electric vehicles, making them a more viable option for everyday use. Consequently, several prominent automakers have declared their intentions to discontinue the production of gasoline-powered vehicles and ramp up their electric vehicle production in the foreseeable future. Nonetheless, electric vehicles encounter some obstacles, such as expensive initial investment, and the requirement for a reliable charging infrastructure. One of the biggest issues of getting an EV is range anxiety, the fear that an EV won't have enough battery capacity to successfully replace gas-powered vehicles.

The goal of this project is to create a predictive model that can accurately estimate the electric range of electric vehicles based on various attributes is the objective of this project. By achieving this, we can tackle the issue of limited electric range, which could potentially increase the adoption of electric vehicles.

Dataset

Electric Vehicle Population Data [Washington US]

The dataset pertains to the electric vehicle population in Washington, US, and provides details regarding the electric range, battery capacity, make, model, year, and other attributes of the EVs. The information was gathered by the Washington State Department of Transportation and the Department of Ecology as part of their initiatives to promote the adoption of electric vehicles and minimize greenhouse gas emissions resulting from the transportation sector.

- **Dataset Source:** Kaggle
- **Target Variable:** Electric Range

Dataset Description

VIN (1-10)	The unique identifier of the vehicle
County	The county where the vehicle is registered
City	The city where the vehicle is registered
State	The state where the vehicle is registered
Postal Code	The postal code of the registration location
Model Year	The year of the vehicle's model
Make	The manufacturer of the vehicle
Model	The model of the vehicle
Electric Vehicle Type	Type of electric vehicle (e.g. battery electric, plug-in hybrid)
CAFV Eligibility	Vehicle is eligible for Clean Alternative Fuel Vehicle (CAFV) incentives or not
Electric Range	The estimated electric range of the vehicle in miles
Base MSRP	The manufacturer's suggested retail price of the vehicle
Legislative District	The legislative district where the vehicle is registered
DOL Vehicle ID	The vehicle identification number assigned by the Department of Licensing
Vehicle Location	The location of the vehicle in longitude and latitude
Electric Utility	The utility company that supplies electricity to the vehicle owner
2020 Census Tract	The census tract where the vehicle is registered

Methodology

Based on the electric vehicle (EV) population data from Washington, USA, here is a possible analysis for each of the machine learning methods listed:

1. **Scaling/Transformation**: We used Label Encoding to convert categorical variables such as County, City, State, etc. to numerical variables. We then scaled the numerical variables to have zero mean and unit variance to improve model performance.

2. **Outlier/Anomaly Detection Method**: While scaling the data, we found that the variables Electric Range and Base MSRP were positively skewed. We used outlier removal to reduce the data skewness after detecting the outliers using the Mahalanobis distance criteria.

3. **Statistical Tests**: We used the Random Forest feature importance method to rank the features by importance. This helped us in selecting the relevant features pertaining to the target variable for building the predictive model.

4. **Splitting Data into Train-Test Sets:** We split the data into train and test sets by specifying the test dataset size to be 20% of the original dataset.

5. **Classifier**: Since the target variable is continuous, we converted it into a categorical variable ('Electric Range Category'). We then used a Logistic Regression and Decision Tree Classifier, both with k-fold cross-validation and grid search CV, to predict the "Electric Range Category" of an electric vehicle based on its attributes.

6. **Regressor**: We used Linear Regression and Random Forest Regressor, both with k-fold cross-validation and grid search CV, to predict the electric range of the Electric Vehicle based on its attributes.

7. **Clustering**: We used the K-means clustering algorithm to group the electric vehicles into 3 clusters based on the Electric Range variable as follows - short range, medium range, and long range. The optimal number of clusters was found using the Elbow Method.

8. **Advanced Method**: We performed ensembling by using the Gradient Boosting Regressor along with Random Forest Regressor to drastically improve the model performance. This ensemble method improved the regressor’s performance by over 80%.

Results

A. Feature Engineering

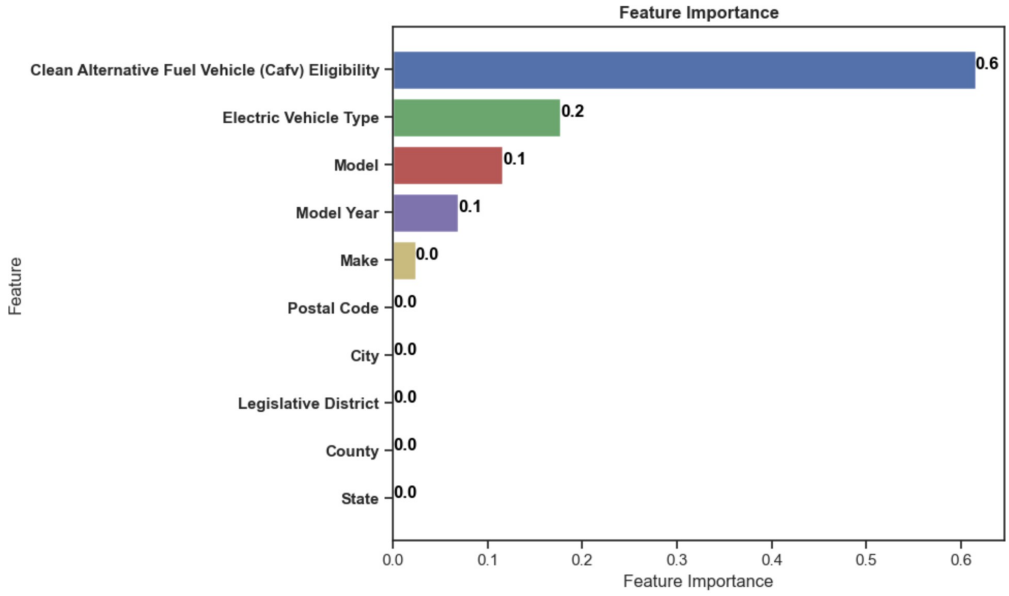

Fig. 1 Features ranked by Importance

B. Anomaly Detection

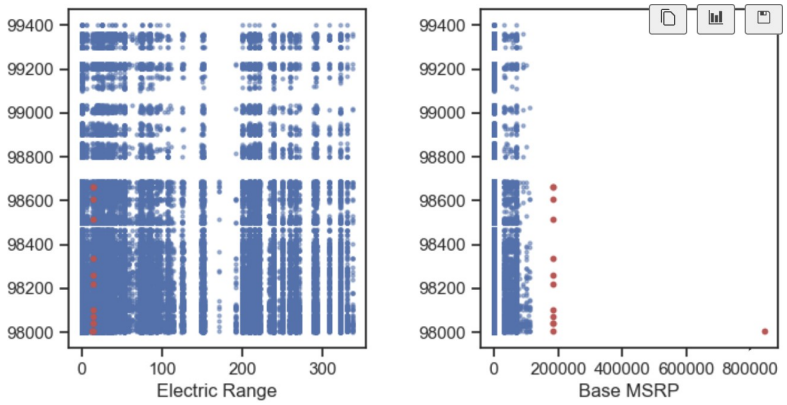

Fig.2 Outlier Detection

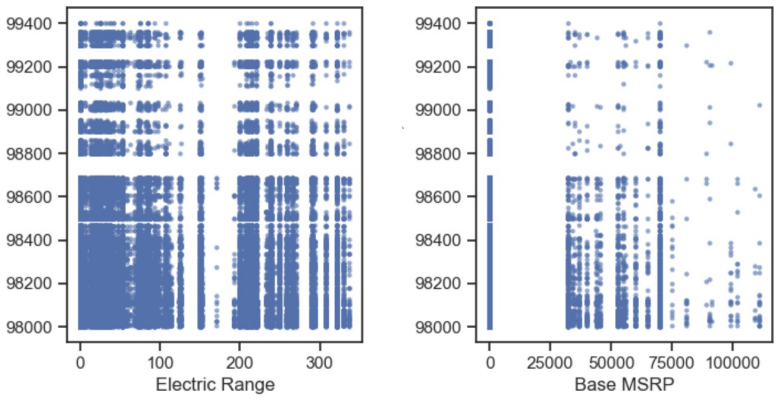

Fig. 3 Outlier Removal

C. Classifiers

	Classifiers	Accuracy Scores	Precision Scores	Recall Scores	F1 Scores	Cross Validation Scores
1.	Logistic Reg.	58.86	16.15	58.86	12.19	59.12
2.	Decision Tree	98.52	93.87	98.52	94.17	98.56

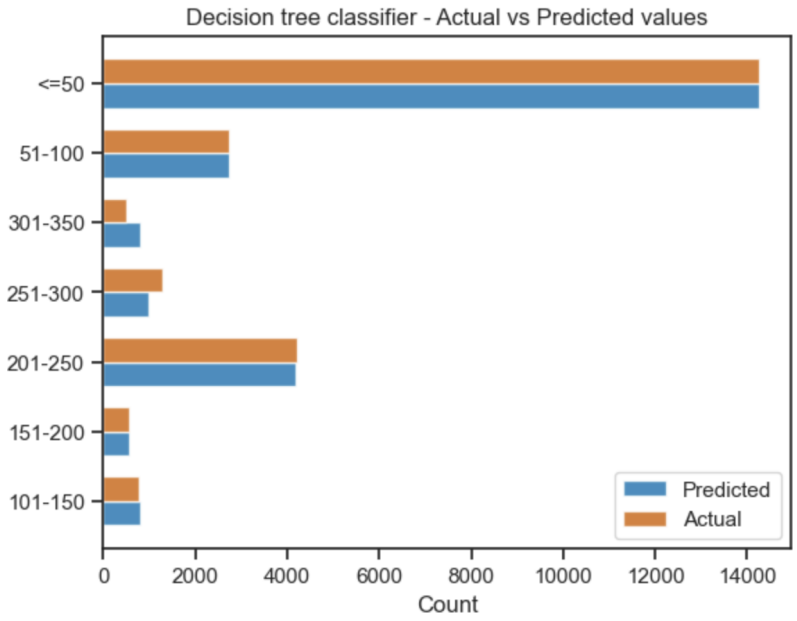

Fig. 4 Decision Tree Classifier

D. Regressors

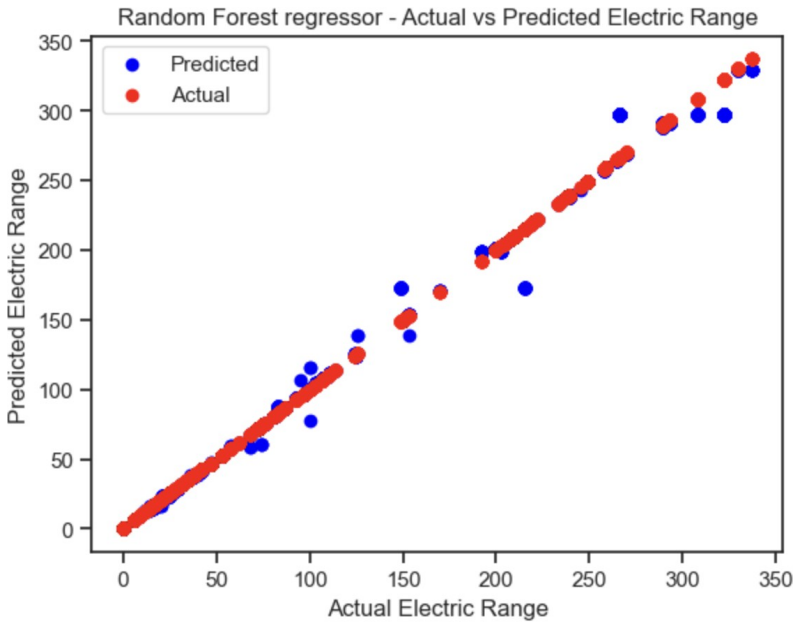

Fig. 5 Random Forest Regressor

	Regressors	Cross-validation MSE	Root Mean Squared Error	R-squared
1.	Linear Regressor	58.86	16.15	58.86
2.	Random Forest Regressor	26.051983	5.104114	0.997443

E. Clustering - KMean

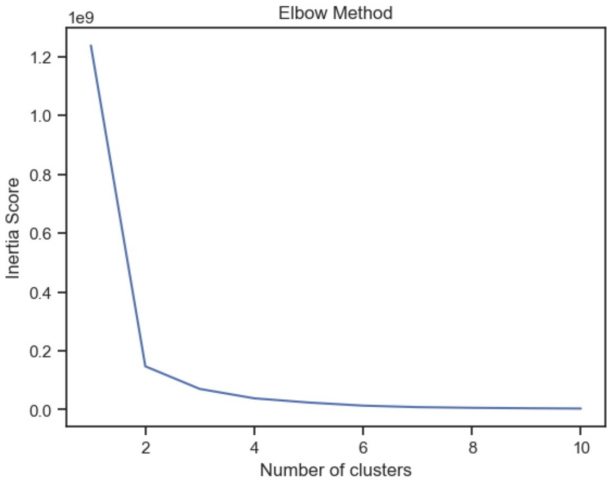

Fig. 6 Elbow Method to Find Number of Optimal Clusters

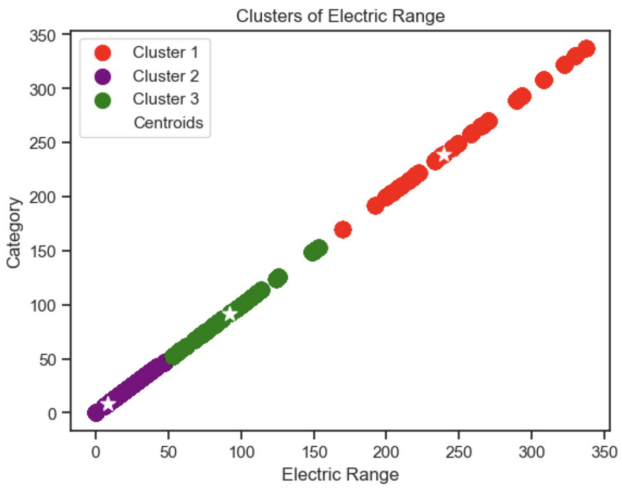

Fig. 7 Cluster Found by KMeans

F. Advanced Method

	Mean Squared Error
Random Forest Regressor	26.051983
Random Forest + Gradient Boosting	5.066162

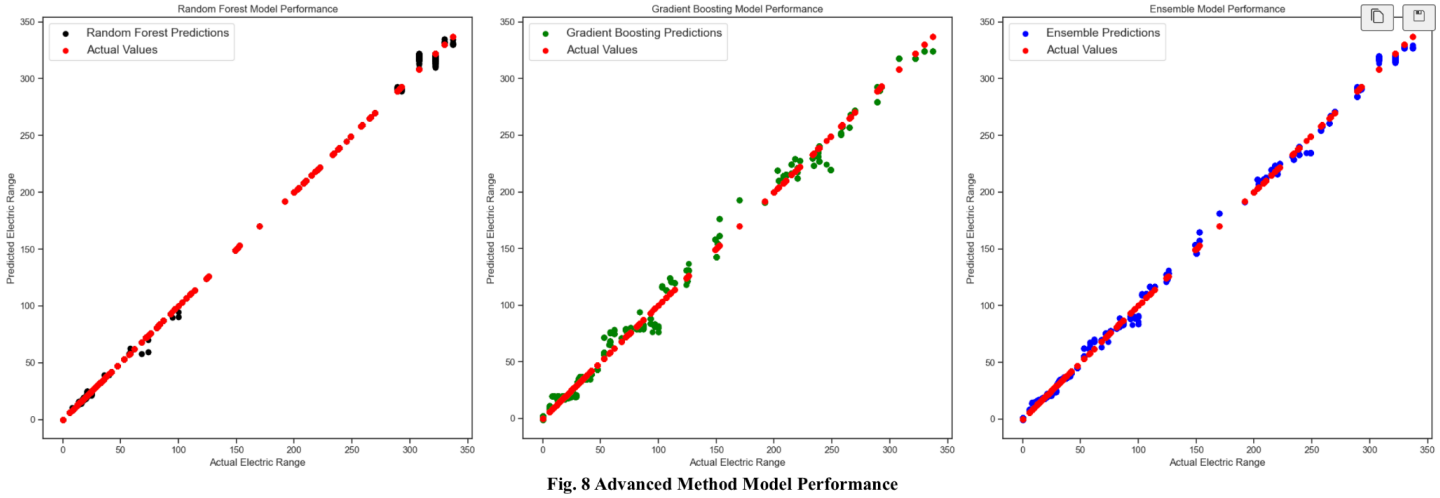

Fig. 8 Advanced Method Model Performance

Actionable Insights

Our analysis of the electric vehicle population in Washington, US from 1997 to 2022 has yielded the following findings:

EDA

- The dataset contains records ranging from the year 1997 to 2023, with the majority of the electric vehicles manufactured between 2016 and 2022 and having a range between 0 to 250 miles. (Refer Fig. 1 - Appendix A)
- Vehicles that are eligible for the CAFV program seem to have a slightly higher range, with a peak around 200 to 250 miles. The non-CAFV eligible vehicles, on the other hand, are concentrated around the 0 to 100 miles range. (Refer Fig. 2 - Appendix A)
- The majority of battery electric vehicles (BEVs) have an electric range of less than 300 miles, while plug-in hybrid electric vehicles (PHEVs) have a wider range distribution with some vehicles having an electric range of over 500 miles. (Refer Fig. 3 - Appendix A)
- EVs with longer ranges tend to be more expensive, and those with higher base MSRP are eligible for CAFV incentives. (Refer Fig. 4 - Appendix A)

Data Preprocessing

- Features VIN (1-10), DOL Vehicle ID, Vehicle Location, Electric Utility, and 2020 Census Tract have no effect on the target variable and are hence dropped.
- The target variable is continuous and is converted into a categorical variable (‘Electric Range Category’) for classification by defining range intervals and assigning a category to each interval.
- The features Electric Range and Base MSRP are positively skewed with 0.93 and 6.4 respectively.
- Outlier Detection using Mahalanobis distance criteria is used to remove all the discrepancies in the Electric Range feature.
- Features with importance scores close to zero, including 'Electric Range', 'Electric Range Category', 'City', 'State', 'County', 'Postal Code', and 'Legislative District' are removed from the list of features for ML models.

ML Methods

- **Classifier**: The Decision Tree classifier achieves a significantly higher accuracy score of 98.52% compared to the Logistic Regression classifier's accuracy score of 58.86%. The cross-validation scores suggest that the Decision Tree classifier is likely to generalize better to new data than the Logistic Regression classifier.
- **Regressor**: The Random Forest Regressor achieves a much lower cross-validation MSE of 26.051983 compared to the Linear Regressor's cross-validation MSE of 4888.003900.
- On the other hand, the R-squared score for the Linear Regressor is much lower, suggesting that the independent variables do not explain as much of the variance in the dependent variable.
- **Clustering**: K-means clustering algorithm is used to group the electric vehicles into 3 clusters based on the Electric Range variable as follows - short range, medium range, and long range. The optimal number of clusters is found using the Elbow Method.
- **Advance Method**: The ensemble model achieves a much lower MSE of 5.07 compared to the Random Forest model alone, indicating that the ensemble model is better at predicting the electric range of vehicles than the Random Forest model alone. (Refer Fig. 5 - Appendix A)

Based on these findings, we can suggest the following actionable insights:

- To incentivize the adoption of EVs, policymakers could consider increasing incentives for vehicles with longer ranges, particularly those that are not eligible for the CAFV program.
- Dealerships and manufacturers can use the clustering results to target marketing efforts and product development towards specific segments of the EV market.
- Machine learning models, particularly the Random Forest Regressor and ensemble model, can be used to predict the electric range of new EV models with reasonable accuracy, thereby aiding in the product development process.

Discussion and Critics

This study has two limitations. Firstly, it only examines electric vehicles in Washington State, and therefore the findings may not apply to other regions or countries. Secondly, the study doesn't provide any comparative data on the popularity of conventional gasoline-powered vehicles in the state.

Although the dataset used in our study was well-generalized and we didn't encounter any major issues with data quality or formatting, resulting in a high accuracy of 98.5% for our classifiers on the first attempt, this may not necessarily reflect the model's performance on real-world data that could be less standardized and more prone to errors. Despite our efforts to minimize overfitting using various methods, the accuracy remained consistently high, leading us to believe that the model was accurately predicting labels and not overfitting.

Therefore, it's important to be cautious when interpreting our results and further testing should be conducted to validate the effectiveness of our models on new and diverse datasets.

Conclusion

In conclusion, this study examined the adoption and use of electric vehicles in Washington State and developed a model to predict an EV's electric range based on its individual features. The results show that an EV's electric range is majorly influenced by the battery capacity and type of the vehicle. For manufacturers and policy-makers looking to create more efficient and ecologically friendly electric vehicles, the study provides useful data.

To verify the findings and conduct a more thorough analysis of the electric vehicle market, more study is required. Nevertheless, this study serves as a springboard for additional investigation into the subject of electric cars and can help advance the creation of more environmentally friendly and energy-efficient modes of transportation.

References

Dataset: Kaggle.com
Python: https://docs.python.org/3/
Scikit-learn: https://scikit-learn.org/stable/
Matplotlib: https://matplotlib.org/stable/index.html
Seaborn: https://seaborn.pydata.org/
DSCI 633: Lecture notes, labs and assignments

APPENDIX A

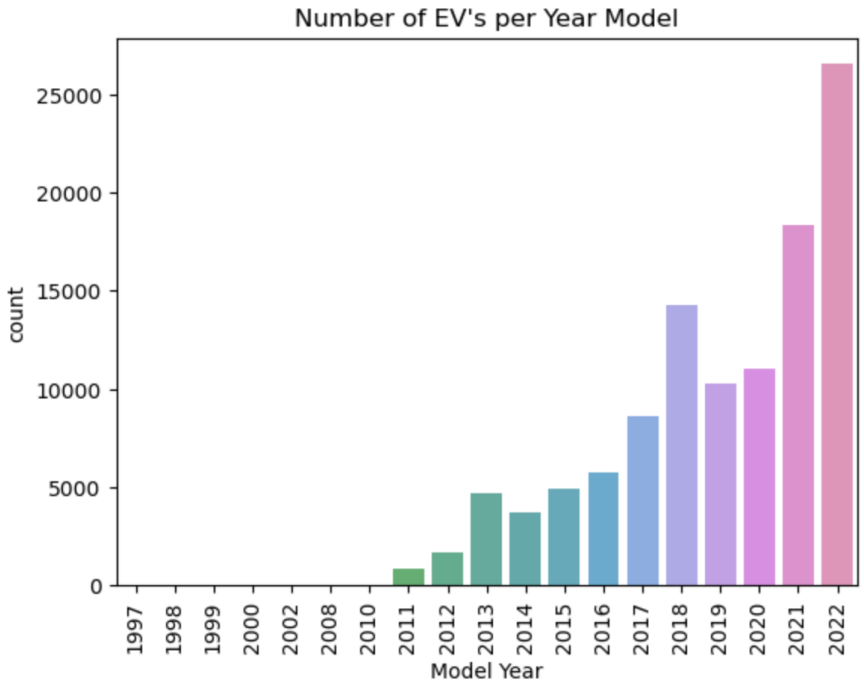

Fig. 1

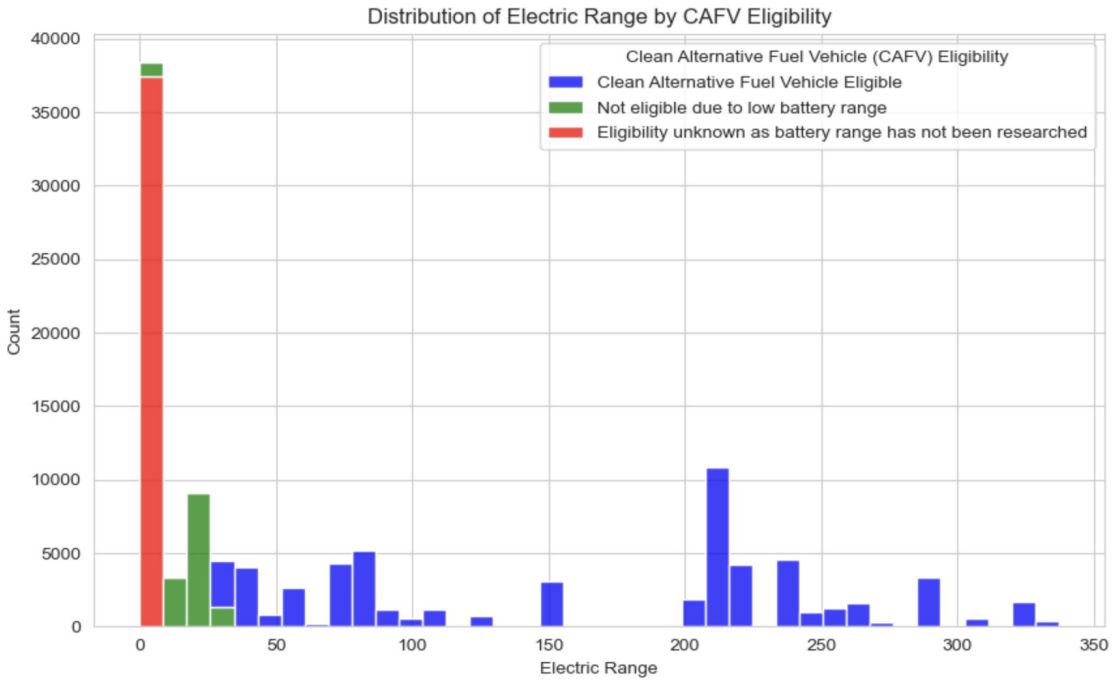

Fig. 4

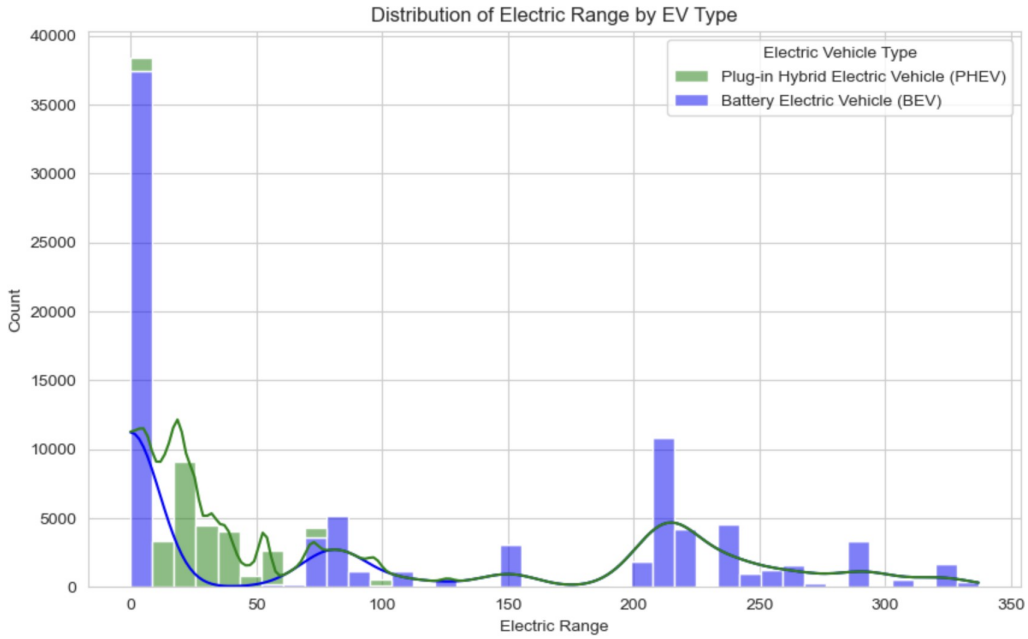
Distribution of Electric Range by EV Type
Electric Vehicle Type
Plug-in Hybrid Electric Vehicle (PHEV)
Battery Electric Vehicle (BEV)
Count
Electric Range
0
5000
10000
15000
20000
25000
30000
35000
40000
0
50
100
150
200
250
300
350

Fig. 3

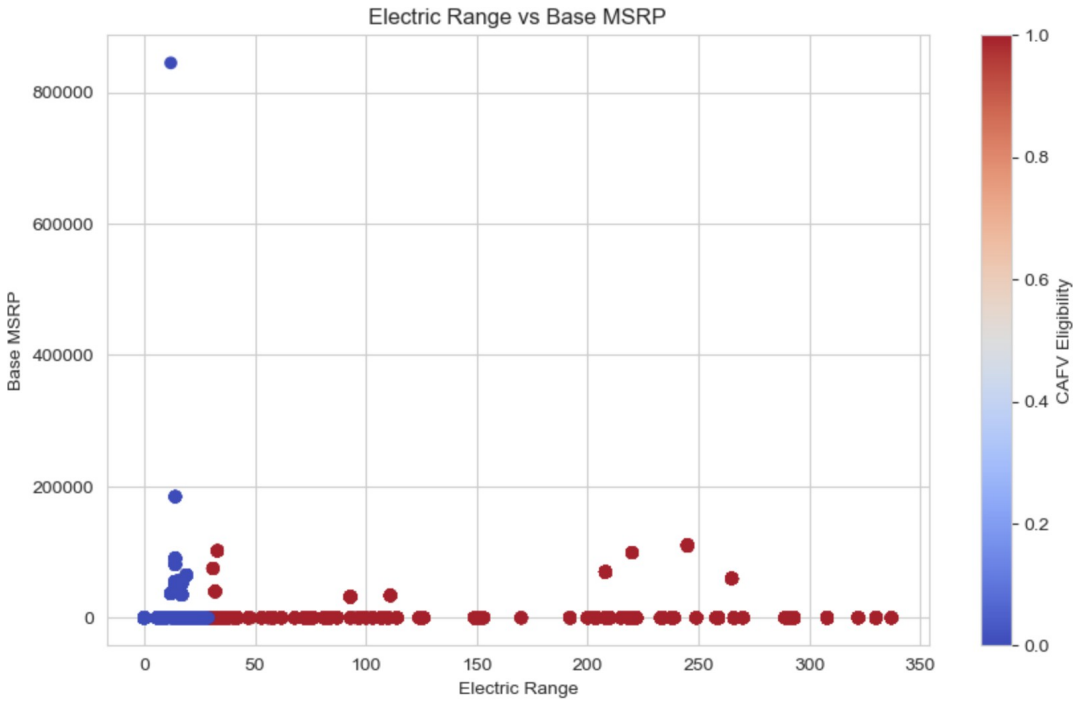
Electric Range vs Base MSRP
Base MSRP
Electric Range
CAFV Eligibility
0
200000
400000
600000
800000
0
50
100
150
200
250
300
350
0.0
0.2
0.4
0.6
0.8
1.0

Fig. 4

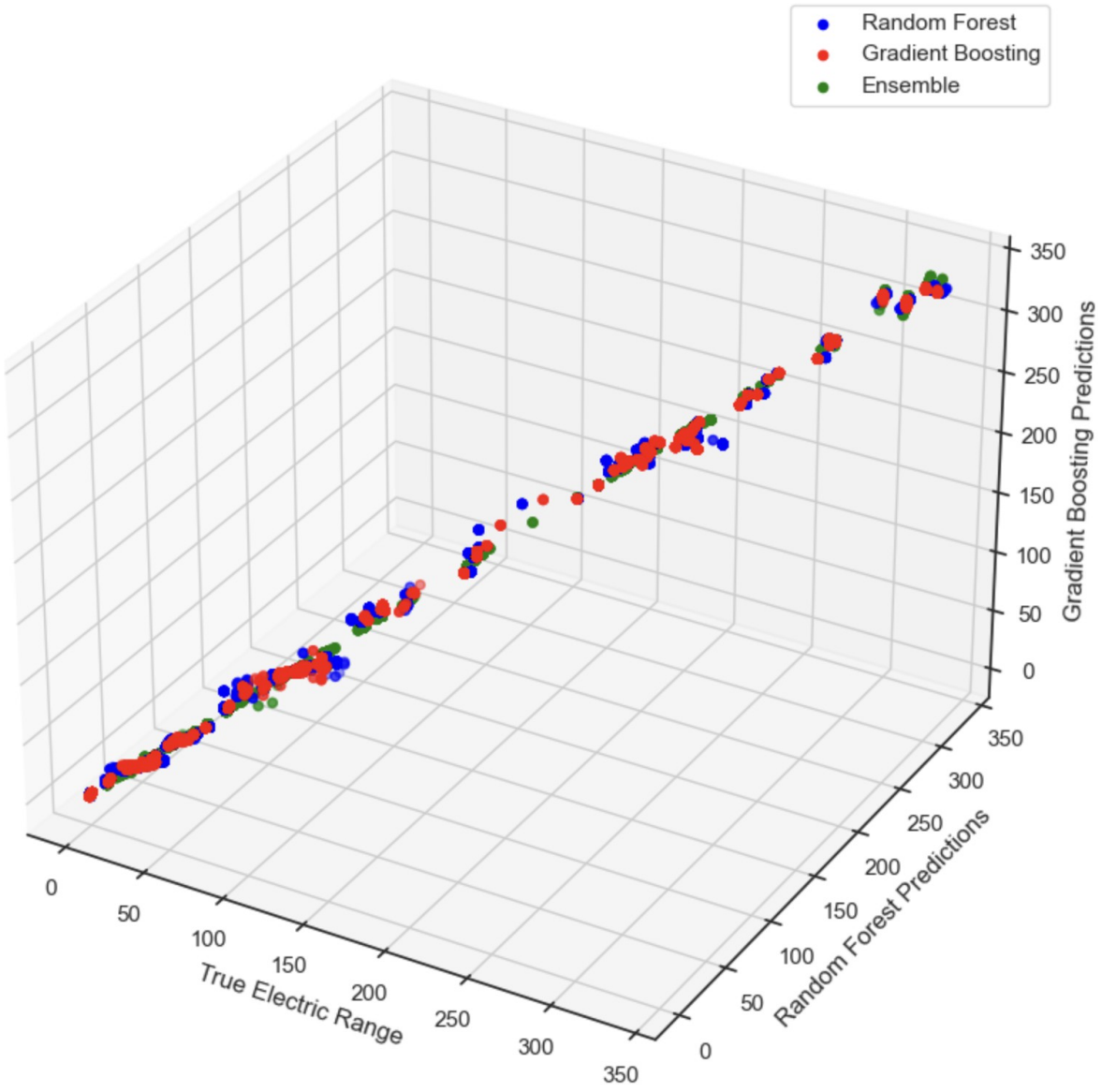

Fig. 5 Advanced Method Model Performance