Summary Report

This project was conducted to answer the problem question:

Can future rates of vehicle fuel type ownership be predicted? Is there a correlation between personal income and changes in ownership of vehicles by fuel types (year over year) that can be used to predict future vehicle ownership?

Data Wrangling

Vehicle fuel data was sourced from data.gov and published by the California Department of Motor Vehicles. The vehicle fuel data contains records of vehicles registered in each zip code from 2018 - 2022. (https://catalog.data.gov/dataset/vehicle-fuel-type-count-by-zip-code)

Income data was also sourced from data.gov and published by the California Franchise Tax Board. The income data contains anonymous tax return records by zip code till 2020. (https://data.ca.gov/dataset/personal-income-tax-statistics-by-zip-code)

However, closer look at vehicle fuel data showed that data was aggregated in October 2018 compared to 12/31 or 1/1 of subsequent years. To avoid skewed results from incomplete year, 2018 was dropped. Many features (make, duty, model year) had mostly incomplete or ambiguous data so they were dropped. A key feature – ‘ Vehicle Ratio’ – was created by calculating :
# of vehicles of a certain fuel type in a given zip code in a given year
Total # of vehicles in the given zip code in the given year

This metric is useful to compare year over year changes and compare zip codes without being skewed by zip codes with very large or small number of vehicles. Since the total number of vehicles in a zip code may change over time due to a variety of factors, the ratio can help as a standardized metric.

Because income data was only available till 2020, merging these two datasets resulted in only years 2019 and 2020 as part of the final dataset. Another important feature - ‘AGI_calc’ – was calculated by :

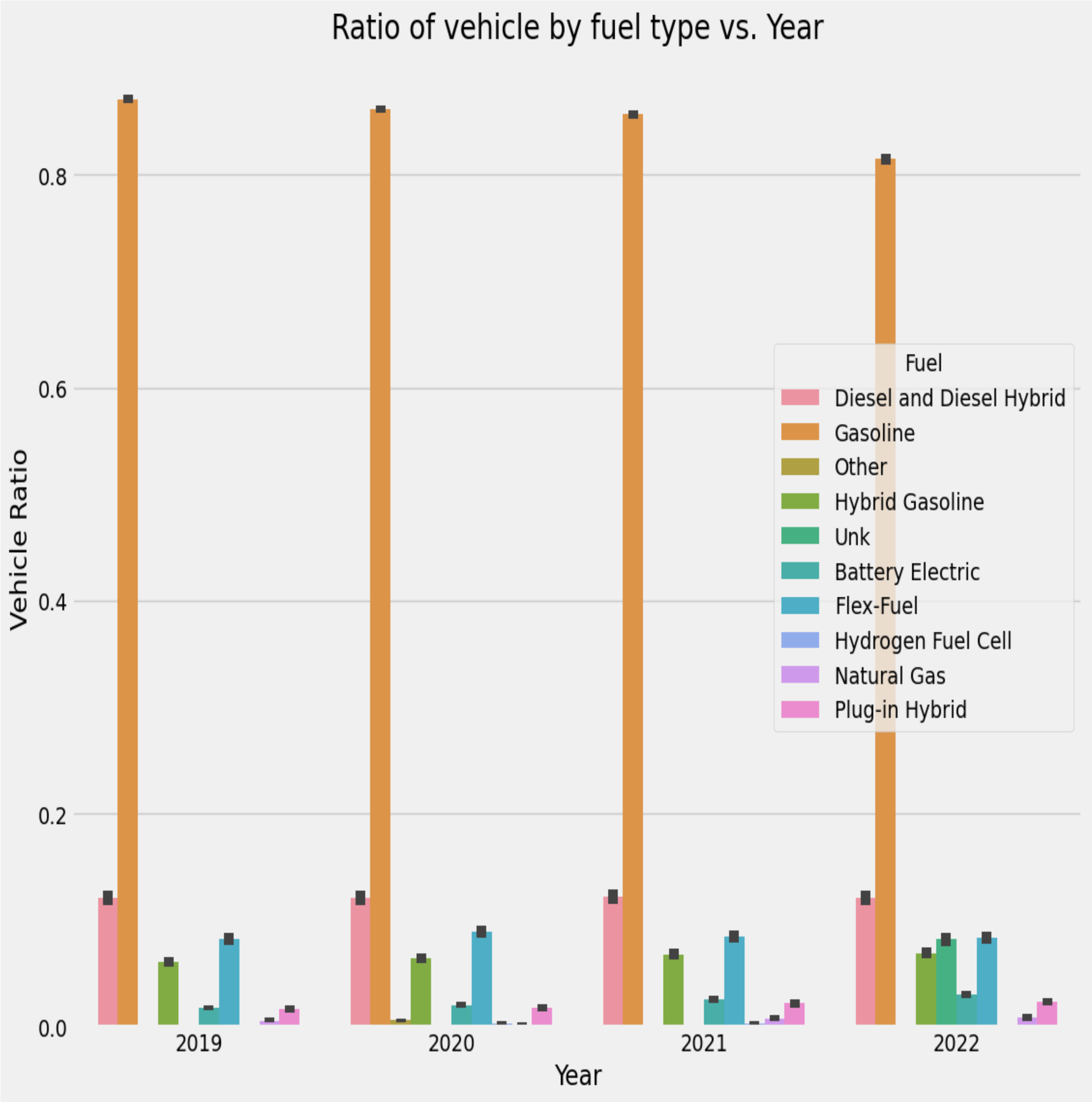
CA AGI / # of Returns to provide an average income for each zip code.

There is some uncertainty as to how the CA AGI figure is calculated and which tax returns are reported here. However, because there is a large sample and they seem to be proportional to zip codes’ populations, I continued using them as a stand in metric for average income for this project.

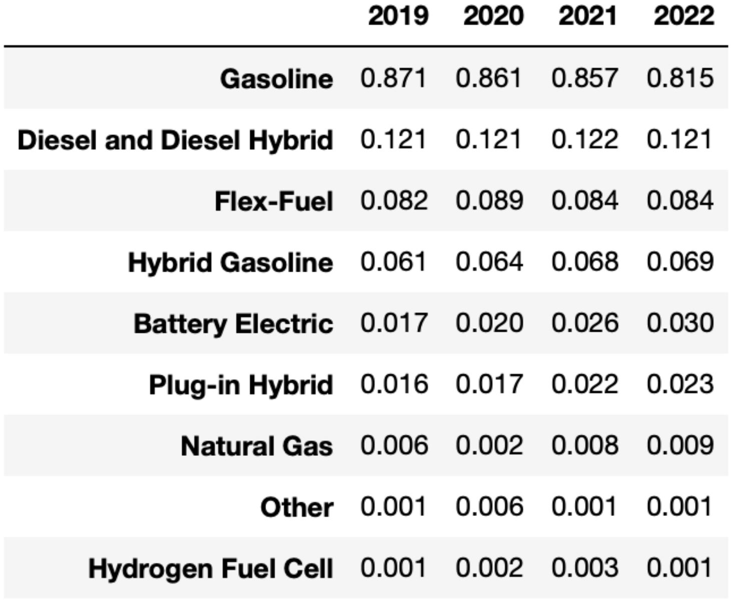
Taking a closer look at incomes showed many outliers that would skew any metrics. The incomes were standardized by removing outliers well past the inter-quartile range.

Exploratory Data Analysis

An exploratory look at the data showed some trends in fuel types over the years. Gasoline is still the dominant fuel type (over 80% of vehicles on average)



Average Ratio of Fuel Type VS. Year

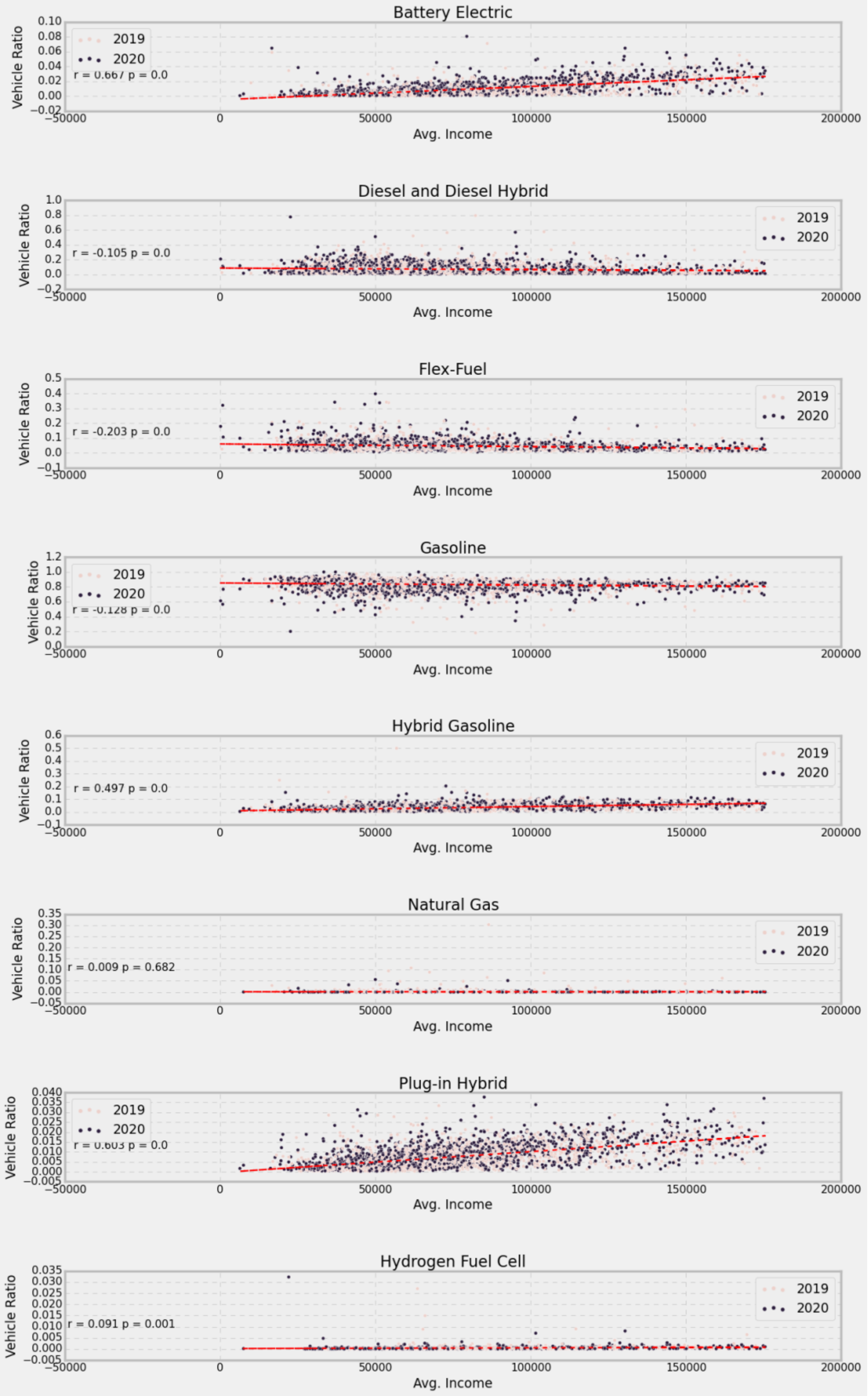


	2019	2020	2021	2022
Gasoline	0.871	0.861	0.857	0.815
Diesel and Diesel Hybrid	0.121	0.121	0.122	0.121
Flex-Fuel	0.082	0.089	0.084	0.084
Hybrid Gasoline	0.061	0.064	0.068	0.069
Battery Electric	0.017	0.020	0.026	0.030
Plug-in Hybrid	0.016	0.017	0.022	0.023
Natural Gas	0.006	0.002	0.008	0.009
Other	0.001	0.006	0.001	0.001
Hydrogen Fuel Cell	0.001	0.002	0.003	0.001

A closer look shows Diesel and Diesel Hybrid are at a distant second at about 12 % of vehicles. Notably, battery electric appears to be have grown from ~1.7% in 2019 to 3% in 2022 but is still a very small part of the market.

The graphs below plot average income of a zip code versus the ratio of vehicles that are of each fuel type in that zip code. Few of the fuel types show stronger correlations between income and fuel type. A closer look at the correlation can help determine the strength and reliability of these correlations.

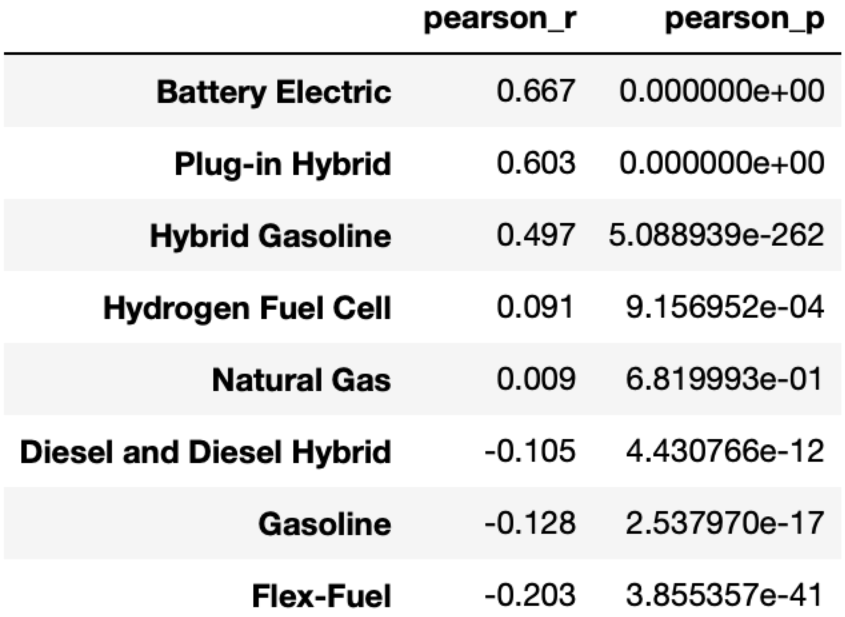
Average Income vs. Vehicle Ratio for each Fuel Type



.

Pearson_r coefficients show moderate positive correlation between average income and three fuel types (battery electric, plug-in hybrid, hybrid gasoline). This implies that the higher the average income is in a given zip code, the higher percentage of these vehicles there are in that zip code. Conversely, flex-fuel, gasoline and diesel show a small negative correlation with average income. It is encouraging that the p values for all these are extremely small if not 0; giving further confidence in these correlations.

Correlation coefficients for Average Income vs. Vehicle Ratio for each Fuel Type

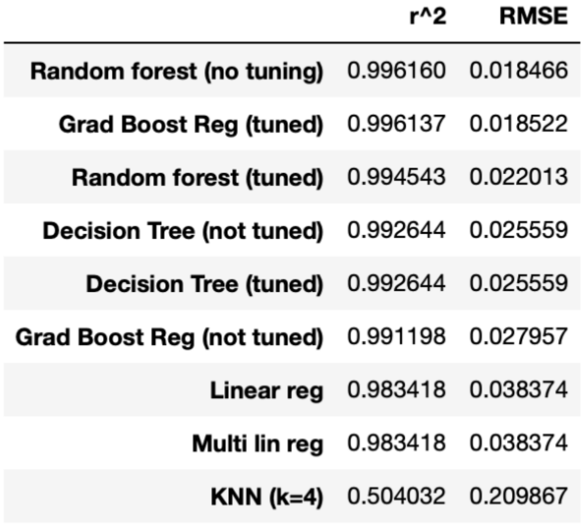


	pearson_r	pearson_p
Battery Electric	0.667	0.000000e+00
Plug-in Hybrid	0.603	0.000000e+00
Hybrid Gasoline	0.497	5.088939e-262
Hydrogen Fuel Cell	0.091	9.156952e-04
Natural Gas	0.009	6.819993e-01
Diesel and Diesel Hybrid	-0.105	4.430766e-12
Gasoline	-0.128	2.537970e-17
Flex-Fuel	-0.203	3.855357e-41

Modeling

A part of this project was also to create a model that was able to predict popularity of vehicle fuel types (vehicle ratios). To do so, multiple model types were tried and evaluated (root-mean square error). The charts below show predicted vs. actual values of the test data using different models. While linear regression showed good performance metrics (r^2 and RMSE), the plot clearly showed that the model was not fitting to large portions of the data well. Alternatively, tree-based models (random forest, decision tree, and Gradient Boosting Regression) performed the best in terms of metrics and plotting.

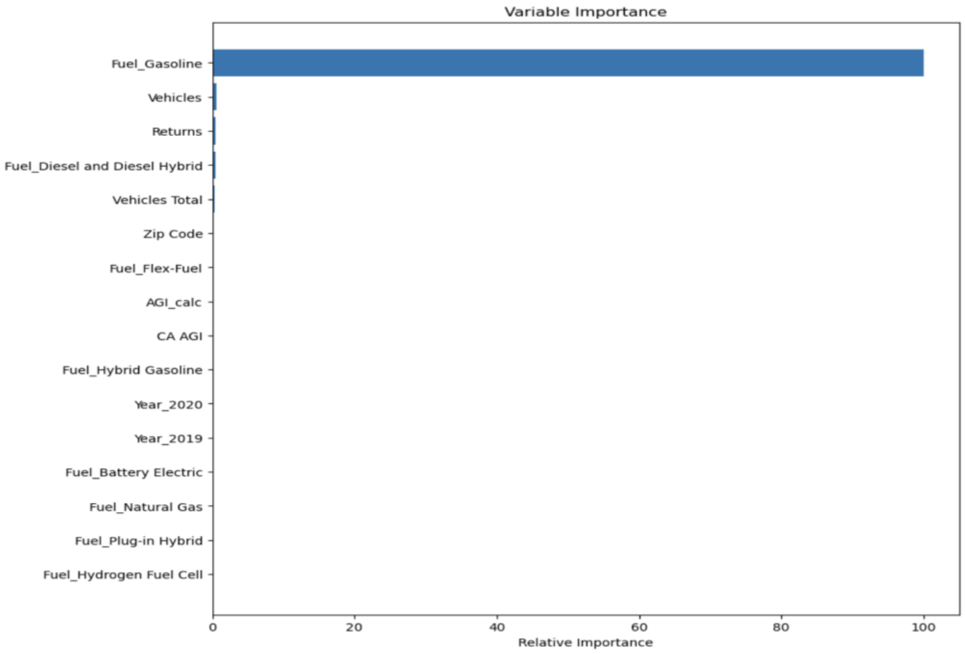
Table showing r^2 coefficient and root-mean square error (RMSE) for each model



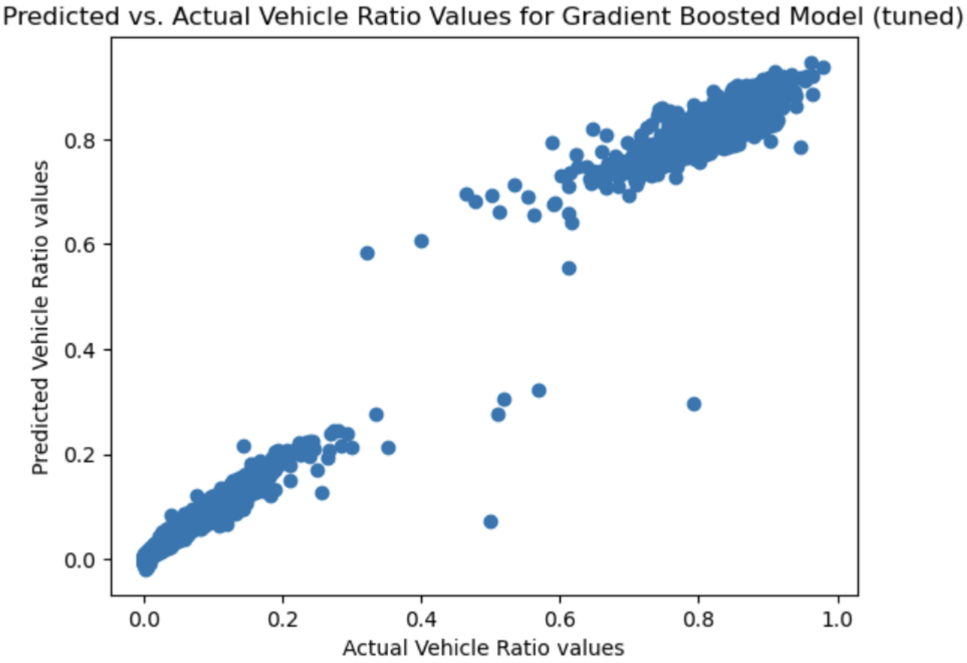
	r^2	RMSE
Random forest (no tuning)	0.996160	0.018466
Grad Boost Reg (tuned)	0.996137	0.018522
Random forest (tuned)	0.994543	0.022013
Decision Tree (not tuned)	0.992644	0.025559
Decision Tree (tuned)	0.992644	0.025559
Grad Boost Reg (not tuned)	0.991198	0.027957
Linear reg	0.983418	0.038374
Multi lin reg	0.983418	0.038374
KNN (k=4)	0.504032	0.209867

However, a closer look at feature importance showed that most of the models were predominantly fitting to 'gasoline' as a feature because the dataset is imbalanced (80% + gasoline vehicles). As a result, XGBoost model was ruled out because it does not perform well with imbalanced datasets.

Graph of feature importance in Random Forest model



While the table above showed an untuned random forest model to have the best score, subsequent cross-validation showed the score to be lower (~ 0.022 rmse). Thus, the tuned Gradient Boosting model is the best model evaluated based on metrics. The graph below shows that the model did a fairly good job in finding patterns and creating accurate predictions although there is definitely room for improvement. Especially encouraging is that the points in the lower left (lower vehicle ratios for non-gasoline fuel types) seem to be predicted accurately despite the imbalance in data.



Conclusion

There does in fact appear to be correlations between vehicle fuel type ownership and personal income. More specifically, there appears to be positive correlation between battery electric, plug-in hybrid and hybrid gasoline to average income. There appears to be slightly negative correlation between gasoline, diesel and average income.

The tuned Gradient Boosting Regressor model performed best to predict vehicle fuel type ratios. If accuracy was the main goal of the model, this would be a suitable choice for the model. However, this model is likely to be computationally expensive. Alternatively, a simple decision tree model performed almost as well in terms of accuracy and would likely be much cheaper computationally as a cheaper alternative to the Gradient Boosting Regressor model.

Future Work/Additional Info/Considerations

This analysis was only done in California so more data would be required before it can be generalized elsewhere. Also, this analysis was done using total income and not specific mappings of a person’s income and the fuel type of the vehicle they own. The latter would provide a much more specific relationship that could provide more precise results.

Future work can include further feature engineering and including other relevant facts about zip codes (climate, geography, gas prices, household size etc.). Additionally, further model tuning can be done by creating a more balanced dataset possibly by undersampling gasoline vehicles.