Electric Vehicle Ownership

Gillian Tatreau

Department of Data Science, Bellevue University

DSC550: Data Mining

Brett Warner

3 June 2023

Introduction

Electric vehicle sales are rapidly increasing and have been for the past several years. The International Energy Agency has found that electric vehicles' shares in overall car sales have increased from 4% in 2020 to 14% in 2022, with a projected increase to 18% of overall car sales by the end of this year; therefore, it is becomingly increasingly important for car manufacturers to understand where potential future consumers are coming from (IEA 2023).

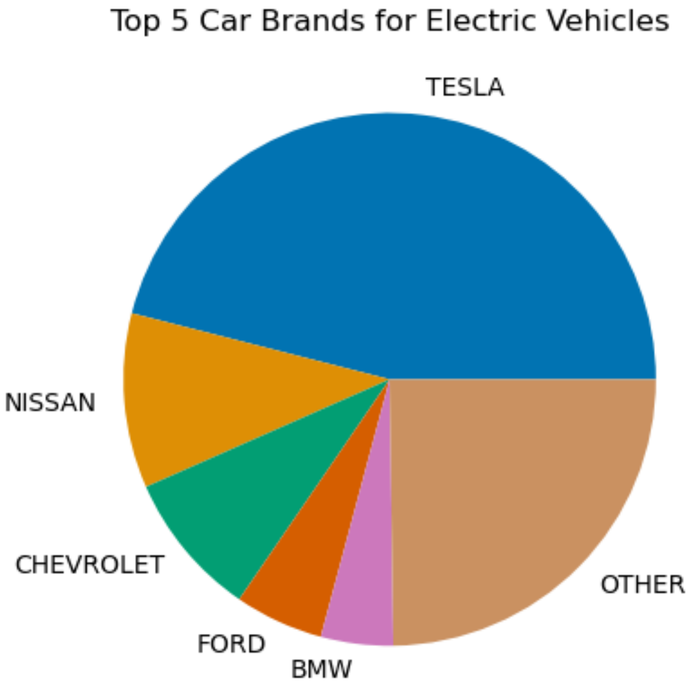
If it were possible to predict the brand of an electric vehicle purchased, given geographical information about the consumer, it would be very beneficial to these manufacturers at various levels of their corporation. For example, if the brand of electric vehicle were predicted given a potential client's geographical area, a local dealership could predict the probability of closing the deal for a particular brand. This could give manufacturers information about where to build more electric vehicle support infrastructure, such as charging stations, and licensing mechanics that specialize in their vehicles. Being able to predict a car's manufacturer based on geographical demographic information can have a lot of importance to that manufacturer: it gives insight into places to focus their advertisements to increase sales.

Using Electric Vehicle Population Data on electric vehicles currently registered through Washington State Department of Licensing, the model will attempt to predict the make of the car given geographical demographic information. The original dataset data includes VIN, city, county, state, postal code, legislative district, location, electric utility, 2020 census tract, Department of Licensing vehicle ID for each vehicle as well as vehicle specific information (make, model, year, type, electric range, and base MSRP). The target for the model will be the car's make. Since we are trying to predict a vehicle's make, we cannot include any data that would not be known if we do not know yet what car was purchased, which would be the model, type of electric vehicle, base MSRP, year, electric range, VIN, or Department of Licensing vehicle ID. Due to the quantity of car manufacturers listed in the make column, an alternate target is manufactured which includes the top 5 most common car manufacturers and the others recoded to “Other” that represents a slightly more balanced target with fewer possibilities to predict.

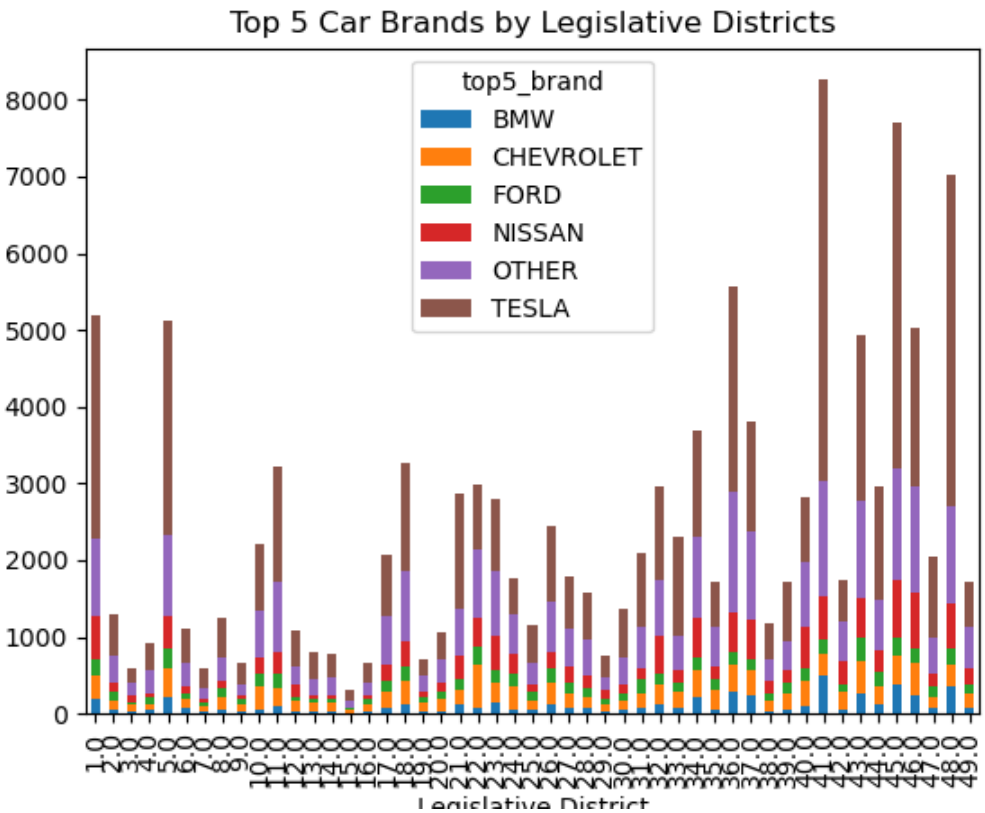
Summary of Milestones 1-3

EDA

The data that is available for the model building process is entirely categorical, and thus the exploratory data analysis looked at the class imbalances as they were to see any potential relationships between the variables.



In the above pie chart, the classes are deeply unbalanced, but the imbalances are much less severe than if we were to populate a pie chart with all the brands listed in the Make column. Also, the prevalence of one company over another in the population of electric vehicles might be explained by the companies' philosophy as well. Tesla only sells electric vehicles, and have very few model options on the market, while the other car brands have many more models available on the market and have a variety of electric vehicles, hybrids, and gas cars. This might suggest that other brands that do not put as much emphasis on electric vehicle models may not be able to sell as many electric vehicles because their electric vehicles only represent a portion of the cars they manufacture.



In many of the legislative districts, we see a very similar distribution of the car brands, with some exceptions such as in the 3rd, 4th, 7th, 12th, 16th, 19th, 20th, 24th, 29th, 35th, and

42nd districts that appear to have a much lower proportion of Tesla vehicles than the average; in contrast, the 1st, 5th, 41st, 45th, and 48th districts have a much higher proportion of Tesla vehicles than the average. We typically see fewer Tesla vehicles and more "Other" vehicles in districts that have fewer total electric vehicles and more Tesla vehicles in districts with more total electric vehicles.

We see some potentially interesting trends in geographic distribution of brands, with could have sociopolitical weight- for example, the distribution of brands could be related to socioeconomic standing, and which could be correlated to geographical boundaries, such as legislative districts. There are areas that have significantly smaller and larger populations of total electric vehicles. The imbalanced class categories will present a challenge in model accuracy, but by focusing on the top 5 brands and putting all others into an "Other" category, the imbalances are slightly less severe and can be accommodated better through careful model selection. Because the distributions of the brands remain even across geopolitical boundaries, that feature might not have a lot of predictive power, which further complicates the model building process.

Data Preparation

Data preparation began by graphically evaluating the presence of outliers in the numerical columns, which were not used in the model. Missing data was dealt with in the following two ways: columns that were missing more than 50% of their data were removed (which removed only the Base MSRP column) and then missing data was imputed by column with either the column’s mode or median. Upon further reflection, missing data was not imputed and instead any rows with missing data were removed to decrease the dimensionality of the dataset.

Then all features that were not useful to the model were removed, which included: VIN (1-10), DOL Vehicle ID, Model, Model Year, Clean Alternative Fuel Vehicle (CAFV) Eligibility, and Electric Range. Dummy variables were created for all the features, increasing the number of features from 9 to 3343. To reduce the dimensionality, feature selection is applied such that the top 20% of features with the highest chi-squared values are selected and the rest are discarded, which reduced the number of features down to 669. The data splitting proportion was an 80%/20% train/test split.

Model Building and Evaluation

For each of the targets (the original Make column and the alternate target with only the top 5 car manufacturers), a baseline dummy classifier that randomly made predictions based off the proportion of the classes in the target variables was created. Originally RandomizedSeachCV() and GridSearchCV() were performed only on Random Forest Classifiers for both targets, which resulted in an increase in accuracy from their baseline classifiers. Later, the same process was repeated with Decision Tree classifiers and Bernoulli Naïve Bayes.

Random Forest: Original Target

- The hyperparameters tuned are n_estimators, max_depth, bootstrap, max_features, min_samples_split, and min_samples_leaf.
- Original target best model hyperparameters: n_estimators: 100, min_samples_split: 6, min_samples_leaf: 3, max_features: sqrt, max_depth: 30, 'bootstrap': True.
- The accuracy increased from 17.1% for the baseline to 44.6%.

Random Forest: Alternate Target

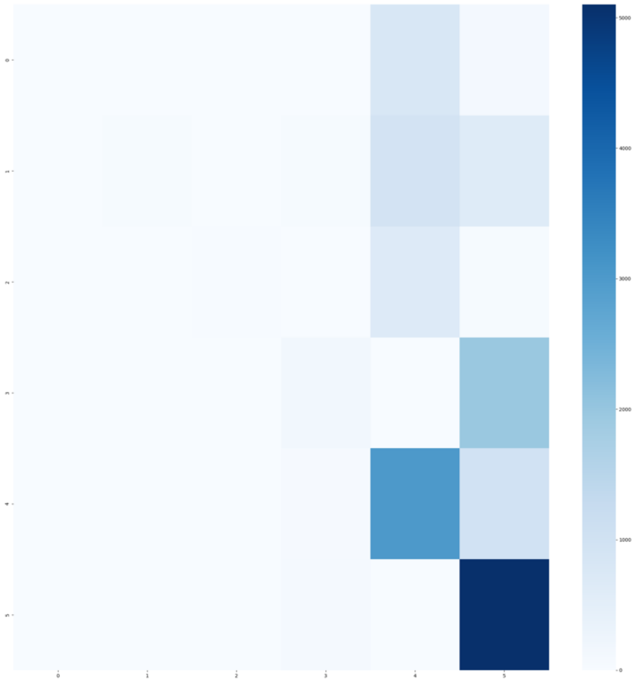
- The hyperparameters tuned are n_estimators, max_depth, bootstrap, max_features, min_samples_split, and min_samples_leaf.

- Alternate target best model hyperparameters: n_estimators: 1250, min_samples_split: 6, min_samples_leaf: 4, max_features: log2, max_depth: 40, 'bootstrap': False.
- The accuracy increased from 24.5% for the baseline to 56.4%.
- Classification report results:

```
              precision    recall  f1-score   support

         BMW       0.44      0.01      0.02       949
   CHEVROLET       0.66      0.02      0.05      1699
        FORD       0.87      0.05      0.10       769
      NISSAN       0.43      0.08      0.13      2131
       OTHER       0.55      0.74      0.63      4096
       TESLA       0.58      0.98      0.73      5200

    accuracy                           0.56     14844
   macro avg       0.59      0.31      0.28     14844
weighted avg       0.56      0.56      0.46     14844
```



The decision tree classifier and the Bernoulli Naïve Bayes classifiers were trained only on the alternate target because it had already demonstrated that it would outperform the original target.

Decision Tree Classifier

- The hyperparameters tuned are criterion and max_depth.
- Best model hyperparameters: max_depth: 3, criterion: gini.
- The accuracy is 55.8%.

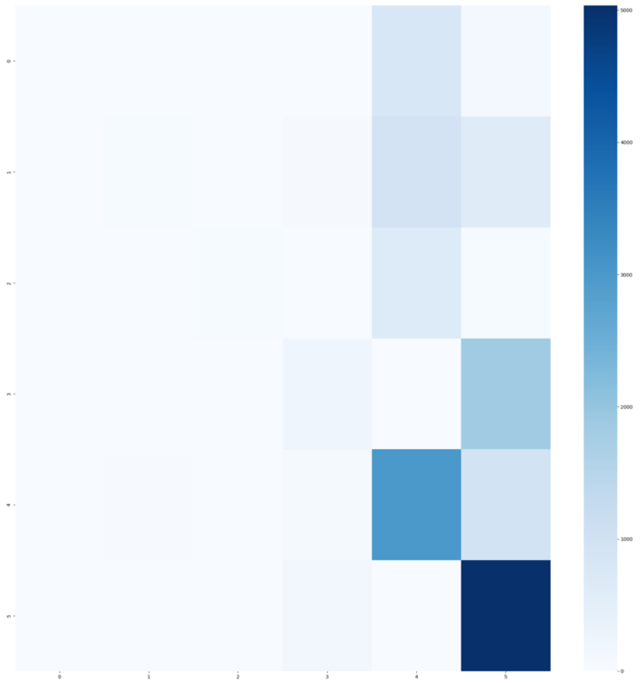
Bernoulli Naïve Bayes Classifier

- The accuracy is 47%.

A final effort to increase the accuracy found by the model was attempted by reselecting more features, such that the top 45% of features with the highest chi-squared values were used, which resulted in 1504 features.

Random Forest: Top 45% features on Alternate Target

- Hyperparameters tuned via RandomizedSearchCV(): The hyperparameters tuned are n_estimators, max_depth, bootstrap, max_features, min_samples_split, and min_samples_leaf.
- Best model hyperparameters: n_estimators: 1000, min_samples_split: 6, min_samples_leaf: 4, max_features: log2, max_depth: 110, 'bootstrap': False.
- The reported accuracy is 56.5%, which is only a 0.09% increase from the random forest classifier performed using only 20% of the features.



Conclusion

From the various models built, the best is the random forest classifier model that resulted from the grid search for the alternate target. The random forest classifier built using the top 45%

of features by chi-squared value did not result in a significant increase in accuracy and increased the computation time greatly, and the confusion matrices are nearly identical between these two models; therefore, the best model built is the random forest classifier with the following hyperparameters: n_estimators: 1250, min_samples_split: 6, min_samples_leaf: 4, max_features: log2, max_depth: 40, 'bootstrap': False. The reported accuracy of 56.4% indicates that this model is not particularly high performing.

The model at this stage is not ready to be deployed in the original context: as a predictor for where manufacturers should invest more money in advertisement or infrastructure; however, this model has the potential to be used as some form of intermediary predictor.

This model should not be used as the only metric to determine where large investments should be made, but if used in conjecture with other information, models, or as a basis for where to collect more data, it could be used to build future models with greater accuracy. If the model is used as a baseline for where to focus more data collection, manufacturers could focus on developing and paying special attention to dealerships that already exist in cities or counties that predict vehicles to exist in with somewhat high probability. They could inquire as to the existence of municipal or county data available for counties and cities that predicted high population of their vehicles, instead of relying on statewide data. None of these steps require especially large investments for manufacturers, as compared to building new dealerships or launching advertising campaigns based in certain locations.

Some additional steps that could be employed to potentially increase the accuracy of this model include resampling the data to be balanced. Additionally, some variables could potentially be removed before creating dummy variables; for example, the state column only contains WA and very few other options, which does not contain a lot of information if most of the entries have the same value, and the vehicle location column contains the approximate latitude and longitude for each vehicle which is almost entirely unique values. Also, if we were able to focus on any predictive power contained in the electric utility column, this could prove a pathway for manufacturers to possibly pursue some kind of partnership which a particular electric utility if owners of their cars use one utility provider over any other. Alternatively, this dataset contains a plethora of information about electric vehicles and their owners, and instead of focusing on the predictive power of some of the features, we performed clustering techniques to see what inherent groups exist within data, the results might yield some interesting conclusions.

References

IEA. (2023, April 1). *Demand for electric cars is booming, with sales expected to leap 35% this year after a record-breaking 2022 - news*. IEA. https://www.iea.org/news/demand-for-electric-cars-is-booming-with-sales-expected-to-leap-35-this-year-after-a-record-breaking-2022