Michael Palazzolo

Final Report: EV Ownership Satisfaction Trends

Problem Statement

Hybrid and electric vehicles (EV's) have seen increased popularity over the past decade, but the factors driving satisfaction are contentious and less well understood. A major auto manufacturer is looking to understand the trends driving customers that re-purchase an EV as opposed to those who discontinue ownership and switch back to gasoline powered vehicles. A survey containing thousands of respondents in the State of California was collected to aid in bringing context to this question. A full summary is contained within this report covering the Data Science approach from data wrangling, exploration to model deployment.

Prediction accuracy of 85% was achieved by on the surveyed population and key contributing features were identified that impact model behavior. The top three are one way commute distance, backup vehicle MPG, and the months the vehicle was owned.

Contents

	precision	recall	f1-score	support
continued	0.87	0.85	0.86	327
discontinued	0.84	0.86	0.85	285
accuracy			0.86	612
macro avg	0.86	0.86	0.86	612
weighted avg	0.86	0.86	0.86	612

Data Wrangling

The data for this analysis and predictive model comes from a survey conducted by Scott Hardman at U.C Davis, linked below.

Source data link: https://zenodo.org/record/4586675#.YO-tlOhKhPZ

Raw Survey Feature Description

The raw survey data contains a combination of categorical satisfaction questions and information about the charging behavior/access that the respondent had. In addition, there are some features regarding commute distances and travel behavior.

Table 1: Raw features provided in the survey data

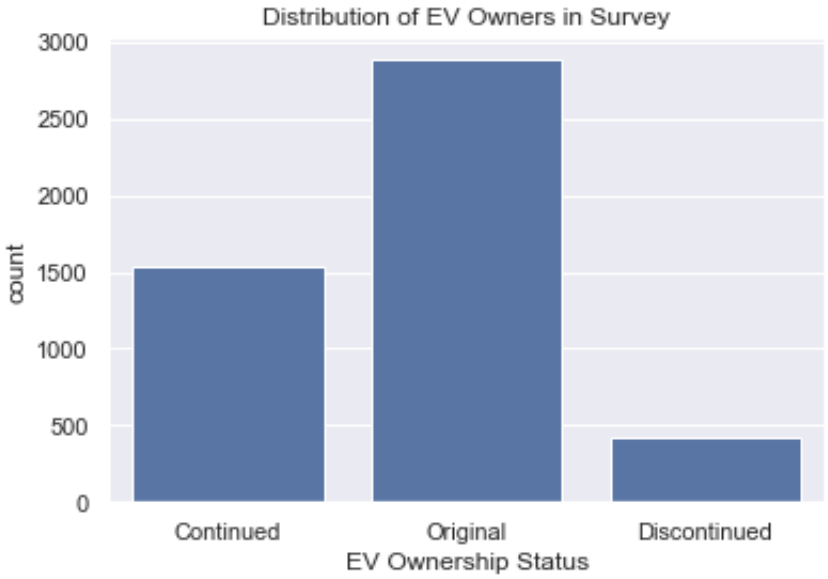
Raw Feature	**Description**
response_id	Response ID
year_submitted_survey_2	Year Survey 2, the follow up survey, was submitted
months_owned	Months PEV has been owned for, or months owned from purchase date to date ownership ceased
year_make_model	Yea Make and Model of PEV
electric_range	US EPA Electric Driving Range of PEV
ev_type	Vehicle Type

discontinuance	Whether respondents no longer own their original PEV but have another PEV (continued), or no longer own any PEVs (Discontinued), or whether they own their original PEV
surveyed_age	Age of survey taker
surveyed_gender	Gender of survey taker
dist_1	One way commute distance of main car
trips_greater_200	Number of trips over 200 miles in the past 12 months prior to taking the survey
safety_satisfaction	Satisfaction with original PEV Safety (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
comfort_satisfaction	Satisfaction with original PEV Comfort (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
refuel_recharge_satisfaction	Satisfaction with original PEV Refuelling/recharging costs (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
performance_satisfaction	Satisfaction with original PEV Environmental Impacts (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
env_impact_satisfaction	Satisfaction with original PEV Performance (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
purch_price_satisfaction	Satisfaction with original PEV Vehicle Purchase Price (including rebates, discounts, etc.) (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
reliability_satisfaction	Satisfaction with original PEV Reliability (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
range_satisfaction	Satisfaction with original PEV Electric Driving Range (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
charge_access_satisfaction	Satisfaction with original PEV Convenience of Charging (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
adas_satisfaction	Satisfaction with original PEV Driving Assistance Features (1= Very Dissatisfied, 2= Slightly Dissastisfied, 3= Indifferent, 4=Slighly Satisfied, 5= Very Satisfied)
household_income	Household Income
home_charge_type	Whether respondents had charging at home at the time of the original survey
work_charge_type	Whether respondents had charging at work at the time of the original survey
highest_charge_used	Highest level of charging used by respondents for public charging at the time of the original survey
backup_vehicle_mpg	MPG of the second vehicle in the household
finance_type	Whether the vehicles is leased or purchased
last_page	The last page survey takers go to (complete survey=40)

Target Variable Insights

The target for the business problem statement is to be able to predict whether current EV owners will either continue with ownership and buy another vehicle or if they will discontinue the ownership process and go back to a gasoline vehicle.

Discontinuance Type	**Description**
Continued	The respondent had an EV and then bought another
Discontinued	The respondent had an EV and then switched to another fuel source
Original	The respondent is a first time EV owner



Observations:

- Original owners do not provide useful insights for building the classifier and are dropped from the dataset as such
- There is a **3:1 imbalance** in the *Continued and Discontinued* dataset. Oversampling of the minority class will be needed to avoid model bias.

Data Cleaning Steps

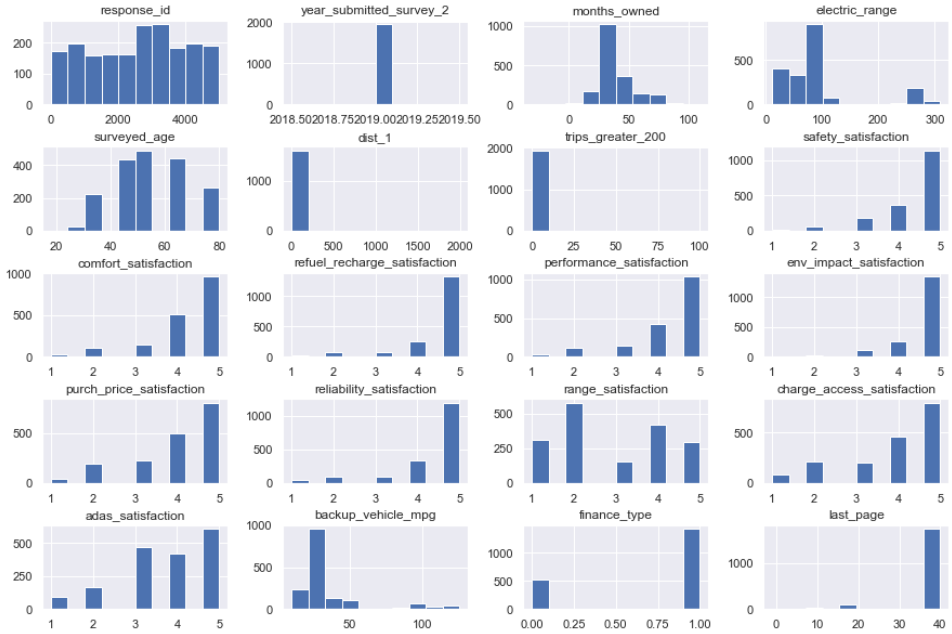
- Text Data
  - Set all column headers to lowercase
  - Set all strings to lowercase
  - Split vehicle data into three distinct columns
    - Year
    - Make
    - Model
- Target Variable (Discontinuance)
  - Drop the rows for **Original** Owners
- Filling NAN's will be handled in model preprocessing

- Negative financing and leasing terms aren't possible, make them positive numbers as this was likely a data entry error

Exploratory Data Analysis (EDA)

It's time to visualize the dataset and get familiarized with it. This will lead to a deeper understanding of the way to impute missing values and set the data up for model preprocessing.

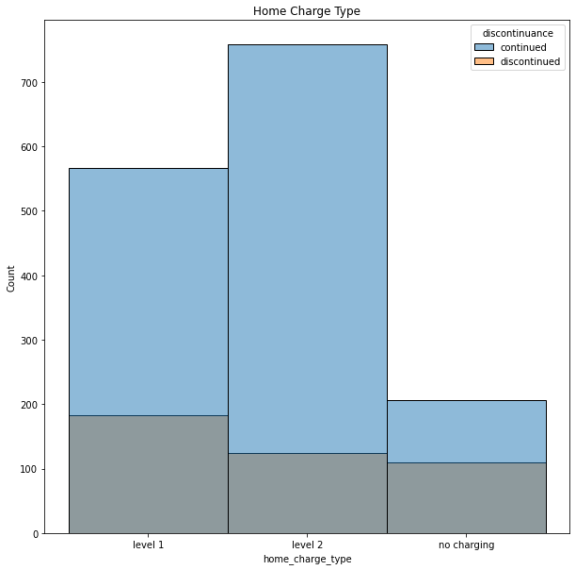
Raw Data Distributions

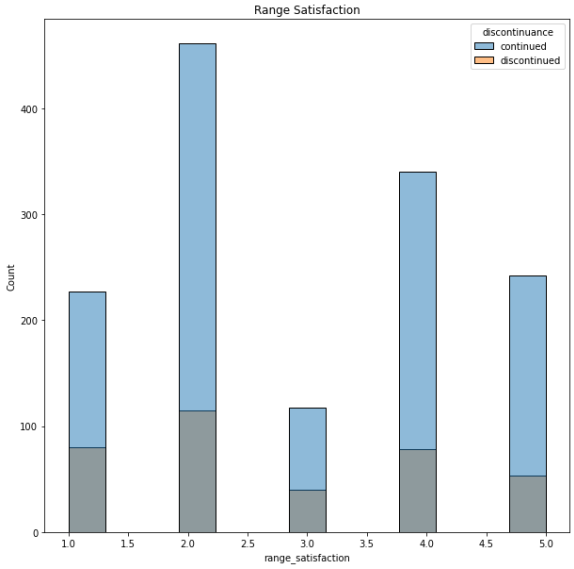


Observations:

- The one way commute distance **Dist_1** has a long right tail distribution due to 2 respondents with >1500 mile travels noted.
- The months of ownership appears approximately normally distributed.
- About 300% more of the respondents lease rather than finance the EV.
- The surveyed population shows a bimodal distribution regarding **range_satisfaction** of the EV.
-

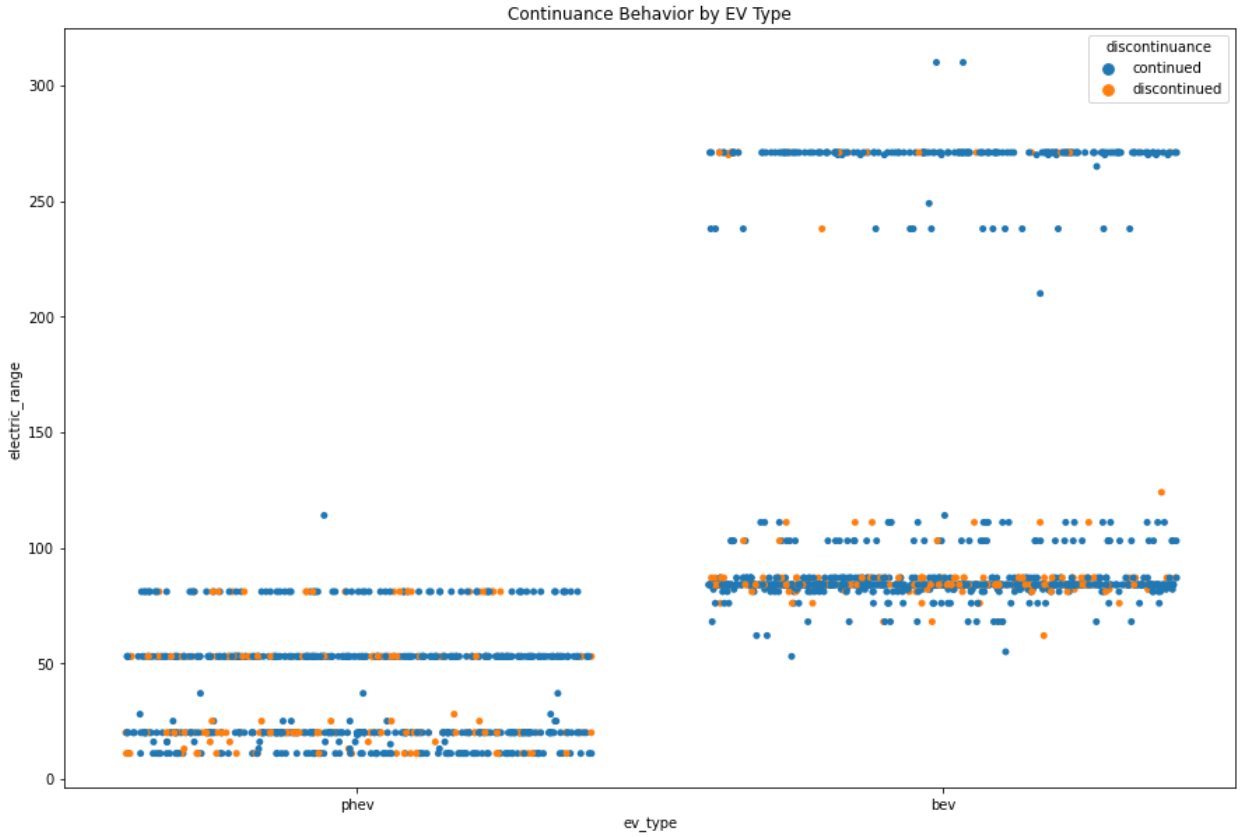
What can we learn about the Contininuance behavior?





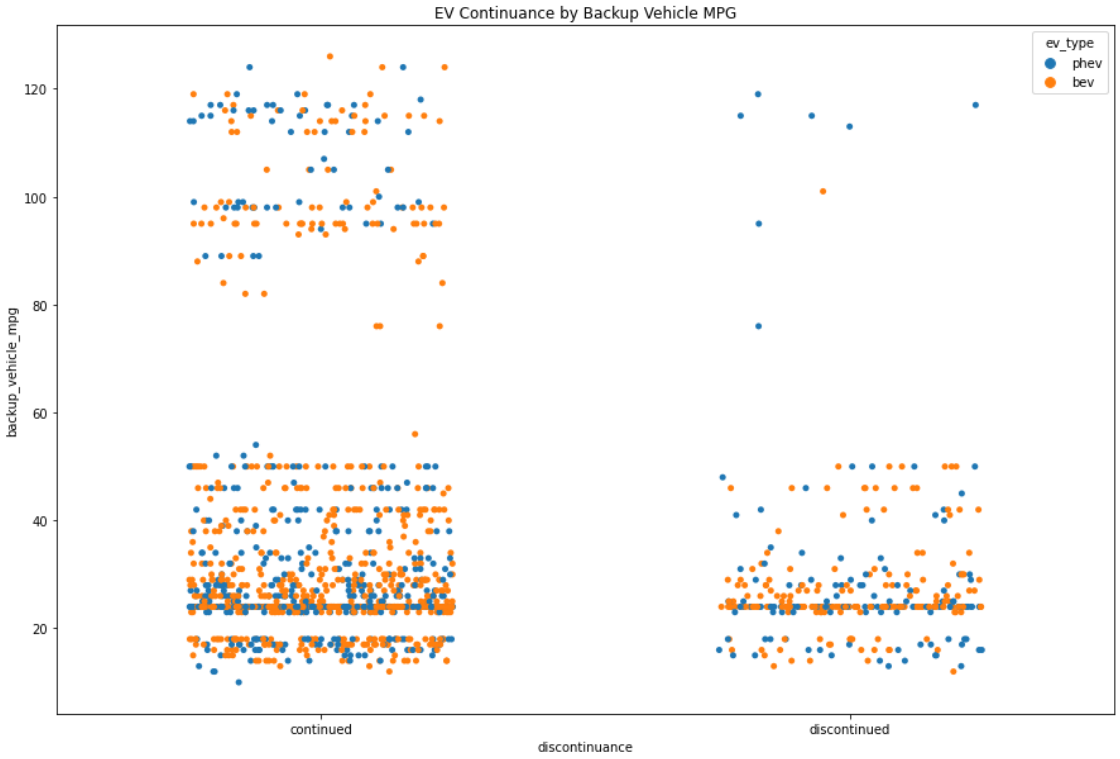
Observations:

- Range Satisfaction
  - **Range_Satisfaction** appears bimodally distributed with relatively larger discontinuance behavior at the Level 1 response.
  - Discontinuance frequency tends to drop off as the **Range_Satisfaction** increases.
- Home Charging
  - The majority of EV owner have capacity to charge the vehicle at home.
  - When the EV owner **doesn't have a charging** method at the house, they discontinue ownership at a proportionately higher level than when charging.
  - The survey respondents who have a **Level 2** charger (the highest residential version) discontinue EV ownership at the proportionally lowest rate.



Observations:

- Hybrid Electric (PHEV) show a reduced mean electric range and proportionately higher discontinuance rates. This may be attributed to the fact that they just straddle the line between fully electric vehicles and gasoline powered ones.
- The mean electric range for Battery Electric Vehicles (BEV) is higher than the PHEV population and shows a lower frequency of discontinuance.
- Once the electric range breaches the 200 mile range, the studied population tends to stick with electric vehicles.



Observations:

- The surveyed population that discontinues tends to backup vehicles at the home with lower MPG than those of the continuing population.

Model Selection

Four machine learning models were built and tuned so that a performance comparison could be made. The models selected for this project were:

1. XGBoost
2. Random Forest Classifier
3. Gradient Boosting Classifier
4. Logistic Regression

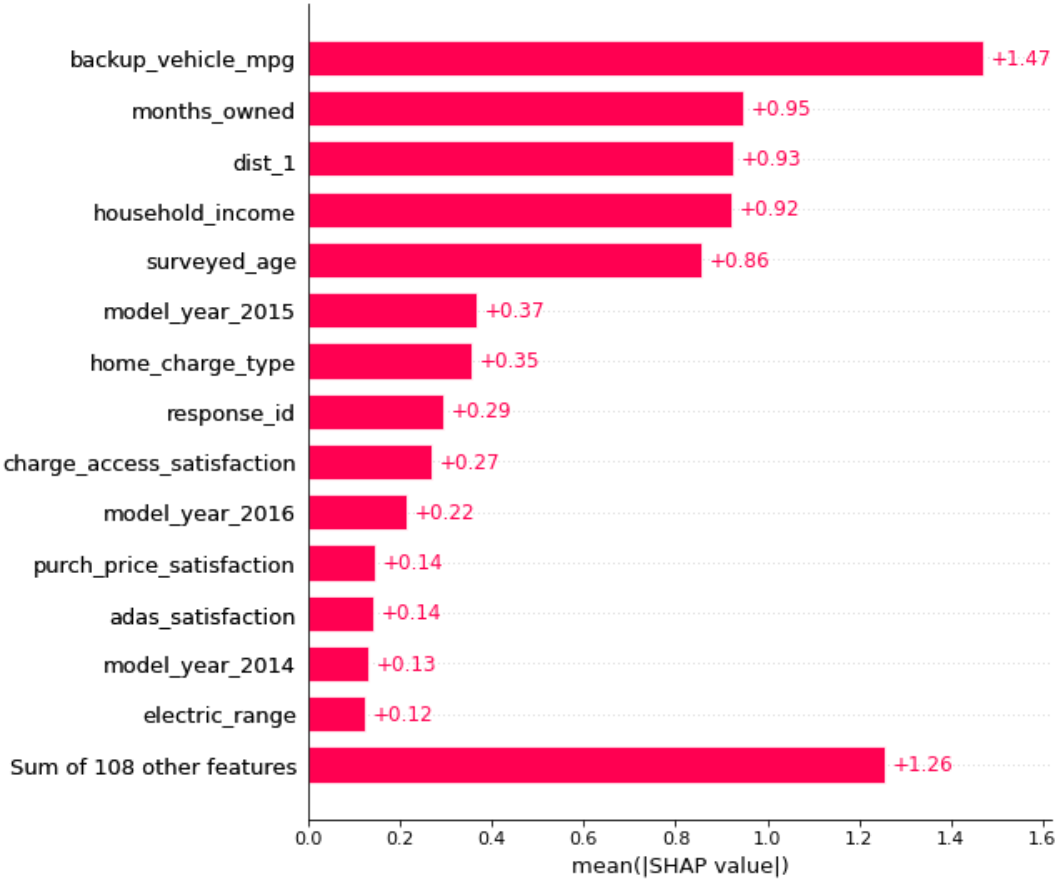
Best Performing Model – XGBoost

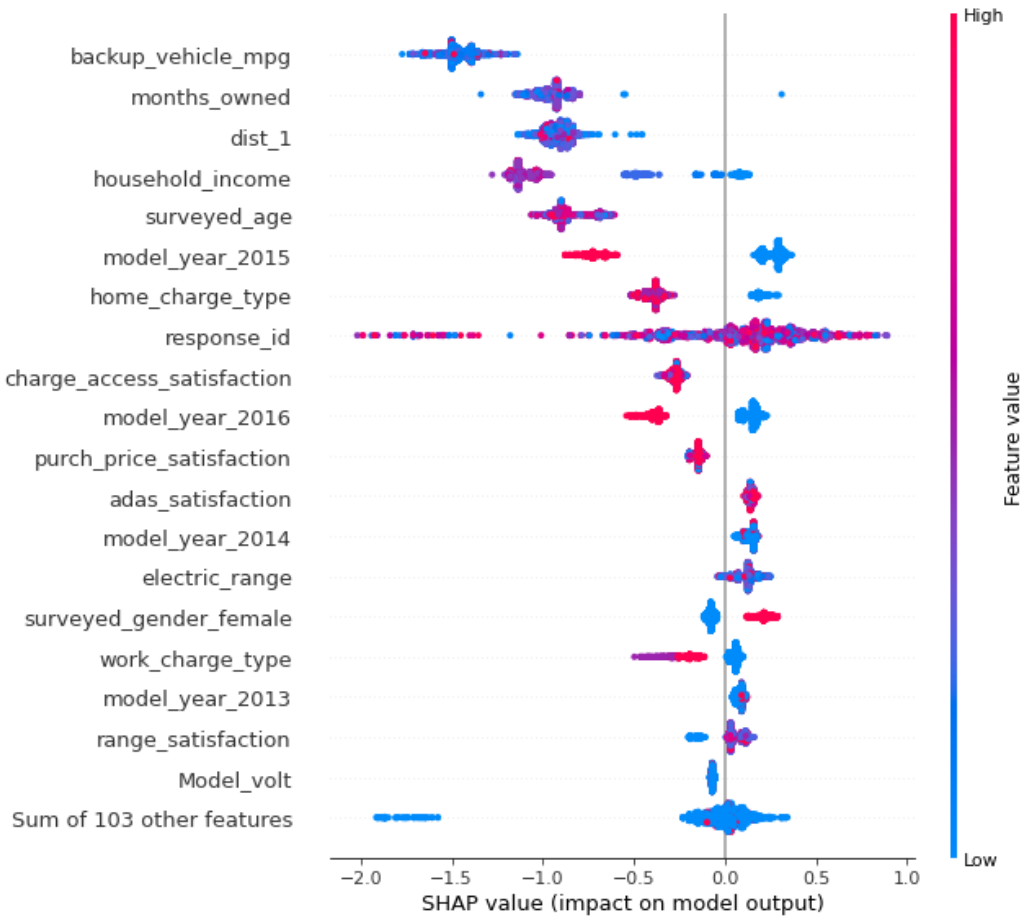
XGBoost turned out to be the most performant model that was benchmarked for this dataset. Hyperparameter selection was done using Random Search with cross validation. The best hyperparameters for the model are listed below. An f1-score of 0.85 was observed showing that the model is performant on the training set.

The top 3 features that impacted discontinuance are the one way commute distance **Dist_1,** the homes backup vehicle fuel economy **backup_vehicle_mpg**, and months of ownership.

	precision	recall	f1-score	support
continued	0.87	0.85	0.86	327
discontinued	0.84	0.86	0.85	285
accuracy			0.86	612
macro avg	0.86	0.86	0.86	612
weighted avg	0.86	0.86	0.86	612

XGBoost Feature Importance Map – SHAP output





Appendix: Model Preprocessing

Ordinal Encoding

The following features were ordinally encoded to preserve the hierarchy that exists:

- Household income
- Home charge type
- Work charge type

One Hot Encoding

The following features were one hot encoded to preserve the hierarchy that exists:

- Model year
- Vehicle make
- Vehicle model
- EV type

- Highest charge type used
- Surveyed gender

Minority Class rebalancing

The survey dataset was highly imbalanced with the minority class underrepresented by a factor of 3. Because the dataset was a mixture of nominal and continuous variables, the Synthetic Minority Over-Sampling (SMOTE-NC) algorithm was used to oversample the minority class. This approach generates synthetic samples for both the continuous and nominal features in the dataset.