Analyzing the Adoption of Electric Vehicles

MGT-6203 Project Final Report

Shane Wiggins (swiggins9) Neel Mehta (nmehta86) Brandon Lee (blee472)
Huy Dao (hdao3) Mitu Malek (mmalek30)

2022-11-20

Contents

Background Information

Transportation is the largest source of US greenhouse gas emissions, and the adoption of Electric Vehicles (EV) has increasingly received a push at both the federal and state level. In fact, by 2035 some states have committed to phase out the sales of new gasoline powered vehicles. Additionally, the White House has signaled further initiatives to push the adoption of EVs, including 1) targets of 50% of all new vehicles sold in 2030 to be electric 2) multi-million dollar investments into building charging station infrastructure, and 3) pushing for tax credits to make EVs more affordable.

The Greenhouse Gas Protocol (GHG Protocol) is a partnership of businesses, NGOs, governments, and other organizations, established in 1998 to address the need for a consistent and comparable framework for GHG reporting. The GHG is one of the more widely recognized standards for measuring GHG emissions today. As the largest contributors (~70%) to global GHG emissions, businesses, particularly the transport, energy, and industrial sectors, have a responsibility to track and reduce their emissions. Scope 1 emissions or GHG emissions that are emitted from resources that are owned or directly controlled by an organization, such as a company fleet of vehicles.

According to the Inventory of U.S. Greenhouse Gas Emissions and Sinks 1990–2020 (the national inventory that the U.S. prepares annually under the United Nations Framework Convention on Climate Change), transportation, including cars and trucks, accounted for the largest portion (27%) of total U.S. GHG emissions in 2020.

Extensive research has been done by the Fuels Institute into the behavior of electric vehicle consumers including answering questions about demographics about consumers and how consumers interact with charging stations[1]. However, that report was mostly assembled by compiling existing literature and publications together. Our goal aims to work directly with quantitative data to assess the demographic factors that impact EV adoption.

Problem Statement

Stringent vehicle fuel emission standards and various government incentive programs have enabled the growing adoption of electric vehicles (EV). While EV growth has been progressive, widespread adoption is still not complete. There are a variety of factors which impact a significant update of EVs and target goals for various sectors.

As shown in the figure below, sales of the adoption of EVs as a percentage of all light-duty vehicles has increased over the last few years[2]. However, total adoption still pales in comparison to gas-powered vehicles. After further evaluation and exploration of datasets available, our primary goal has shifted to identify which demographic factors impact EV adoption the greatest. A Secondary goal includes estimating EV adoption based on these factors for states that do not publicly report their EV registration.

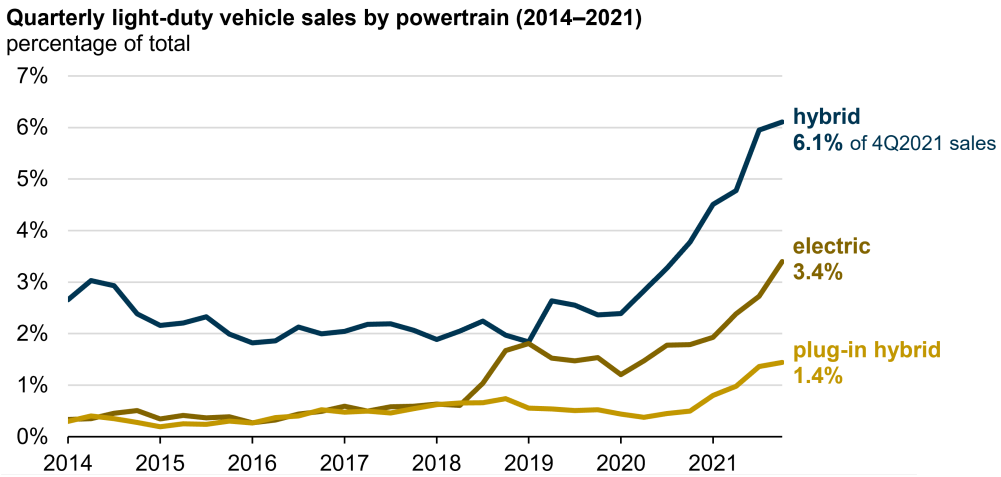

Figure 1: Light Duty Sales[2]

Business Justification

Anticipating the level of EV adoption and factors that impact adoption and emissions has an effect in multiple sectors. From a government standpoint, analysis of adoption and emissions assists in decision making of regulatory changes, additional tax incentives, or additional investment into infrastructure. Increased EV adoption also reduces dependency on foreign gasoline production. Increasing adoption (and thus demand) of EVs has a direct impact on the auto industry. Quantifying and forecasting the levels of adoption and factors impacting adoption can assist in determining to what level auto makers should prioritize the EV market over gas-powered vehicles.

Data Wrangling, Cleaning and Understanding

Dataset Group One

Sixteen datasets (one for each participating state) were sourced from AtlasHub[3]. Since submission is voluntary by state, the structure of the data files vary with no set format. One of the main challenges include merging sets that were identified at the County-level versus those identified at the ZIP code-level.

Five datasets for CA, FL, MT, TN, and VA were reported at the county level. These datasets contained a total of approximately 3.2 million records, with dates ranging from 2003 to 2022. Another 11 datasets were identified for CO, CT, MI, MN, NJ, NY, OR, TX, VT, WA, WI were reported at the ZIP code level. These datasets contained a total of approximately 6.2 million records, with dates ranging from 2002 to 2022. Inconsistent data types were found throughout, such as dates being listed as strings, floats, or in datetime format. In some cases, missing or unexpected values were found that needed correction or imputation.

A sample of the raw data for each type of the datasets (California and Colorado respectively) is shown in the table below:

Dataset: CA dataset

Vehicle ID	County GEOID	Registration Valid Date	DMV ID	DMV Snapshot	Registration Expiration Date	State Abbreviation	Geography	Vehicle Name
CA-002-03597	6099	1/1/2011	2	CA Registration Data from CA (12/31/2011)	NULL	CA	County	Chevrolet Volt
CA-002-03598	6105	1/1/2011	2	CA Registration Data from CA (12/31/2011)	NULL	CA	County	Nissan Leaf
CA-002-03599	6103	1/1/2011	2	CA Registration Data from CA (12/31/2011)	NULL	CA	County	Chevrolet Volt
CA-002-03600	6099	1/1/2011	2	CA Registration Data from CA (12/31/2011)	NULL	CA	County	Tesla Roadster

Dataset: CO dataset

DMV ID	DMV Snapshot	ZIP Code	State	Registration Valid Date	Registration Expiration Date	VIN Prefix	VIN Model Year	Vehicle Name	Technology
18	CO DMV Direct (8/1/2020)	80920	CO	1/2/2020	1/31/2021	5YJ3E1EB	J	Tesla Model 3	BEV
17	CO DMV Direct (7/1/2020)	81621	CO	1/2/2020	1/31/2021	5YJ3E1EB	J	Tesla Model 3	BEV
17	CO DMV Direct (7/1/2020)	80111	CO	1/2/2020	1/31/2021	5YJ3E1EB	J	Tesla Model 3	BEV
17	CO DMV Direct (7/1/2020)	80923	CO	1/2/2020	1/31/2021	5YJ3E1EB	J	Tesla Model 3	BEV

In order to merge datasets, several steps were necessary. First, the year 2020 was selected as a primary basis for analysis as the latest US census was performed in 2020 and provided the most complete and accurate demographic data. Data types were corrected to a consistent format. The individual records were aggregated for a total count by ZIP code or County as applicable. The approximately 3.2 million County-level and 6.2 million ZIP code-level individual records were aggregated to approximately 2,200 records and 36,900 records respectively.

In order to merge the two datasets, the ZIP code-level data was mapped to County-level FIPS (Federal Information Processing System) Codes. These codes may also be referred to as County GEOID's (Geographic Identifiers) by the US Census Bureau[4]. These codes are identified by a 2-digit State code and 3-digit County code. Note that longer GEOID's may be displayed in some of the excerpts contained within this document (specifying sub-regions of the identified County). However, the first 5 digits conform to the structure previously specified.

Most of the data cleaning/wrangling work was performed in Python with some work performed in R and Microsoft Excel to finalize the datasets. Post cleaning-merging-aggregation, observations were reduced to 849 for the year 2020.

An excerpt of the aggregated ZIP-code level data, County level data showing total EV count by year and the ZIP to County FIPS mapping is shown in the table below:

Aggregated Dataset1 (zip)

STATE	ZIP_CODE	COUNTY_GEOID	COUNTY	REGISTRATION_VALID_YEAR	EV_COUNT
CA		6001		2020	7009
CA		6059		2020	51068
CA		6061		2020	6339
CA		6063		2020	52
CA		6065		2020	3523
CA		6001		2020	38776

Aggregated Dataset1 (county)

STATE	ZIP_CODE	County _GEOID	County	REGISTRATION_VALID_YEAR	EV_COUNT
CO	80002			2012	6
CO	80002			2013	15
CO	80002			2014	16
CO	80002			2015	20
CO	80002			2016	19

Aggregated Dataset1 (mapping)

	A	B	C	D	E
	ZIP	COUNTYNAME	STATE	STCOUNTYFP	CLASSF
	36003	Autauga County	AL	1001	H1
	36006	Autauga County	AL	1001	H1
	36067	Autauga County	AL	1001	H1
	36066	Autauga County	AL	1001	H1
	36703	Autauga County	AL	1001	H1
	36701	Autauga County	AL	1001	H1

Note that the excerpts of records above do not display the leading zero for County codes.

Dataset Group Two

The second dataset group was sourced from the US Census Bureau and USDA for all counties residing in 50 States and District of Columbia. Although this data represents many demographic factors, a single combined data source is not readily available. One of the major challenges with this data-group was identifying factors of interest and merging the necessary data for this task.

In some cases data could be streamed directly from the website for aggregation; in other cases it was necessary to locally download the data for manipulation and merging. As previously mentioned the year 2020 was focused on since this was the latest US Census year and provided the most complete demographic data.

An excerpt of compiled demographic factors including the GEOID is shown below:

Dataset2 Excerpt

GEO_ID	NAME	S0101_C01_001E	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0	S0101_C0
Geography	Geographic Area N	Estimate!!Total!!Total population	Margin	Annota	Annota	Estimat	Annota	Margin	Annota	Estimat	Annota	Margin	Annota
0500000US72011	Aï¿½asco Municipio, F	26555	*****	*****	null	927	null	*****	*****	1299	null	249	null
0500000US45001	Abbeville County, So	24582	*****	*****	null	1275	null	50	null	1186	null	211	null
0500000US22001	Acadia Parish, Louisi	62371	*****	*****	null	4348	null	10	null	4529	null	409	null
0500000US51001	Accomack County, Vi	32560	*****	*****	null	1822	null	13	null	1756	null	235	null
0500000US16001	Ada County, Idaho	469473	*****	*****	null	26884	null	165	null	29098	null	1338	null
0500000US19001	Adair County, Iowa	7048	*****	*****	null	402	null	14	null	464	null	64	null

Merged Example

FIPS	STATE	COUNTY	TOTAL_POPULATION	COLLEGE_GRAD	UNEMPLOYED_COUNT	MEDIAN_INCOME	POVERTY_RATIO	MALE_MIDAGE	PREF_PARTY	Metro_TYPE	CHARGING_PORT_COUNT	TOTAL_EV_REGISTERED_2020
53033	WA	KING COUNTY	2269675	851818	100232	102620	7.6	280890	DEMOCRAT	Metropolitan Statistical Area	2257	385497
06037	CA	LOS ANGELES COUNTY	10014009	2317682	612986	75624	13.2	1240635	DEMOCRAT	Metropolitan Statistical Area	10997	154234
36103	NY	SUFFOLK COUNTY	1525920	385916	63095	109084	6.1	212745	REPUBLICAN	Metropolitan Statistical Area	566	129414
53061	WA	SNOHOMISH COUNTY	827957	185346	38700	93589	7.1	110058	DEMOCRAT	Metropolitan Statistical Area	391	101723
36059	NY	NASSAU COUNTY	1395774	441749	56913	122730	5.7	185047	DEMOCRAT	Metropolitan Statistical Area	288	100615
08031	CO	DENVER COUNTY	715522	261452	32569	75355	12.4	78253	DEMOCRAT	Metropolitan Statistical Area	915	89052
08013	CO	BOULDER COUNTY	330758	135876	11127	88289	9.5	40934	DEMOCRAT	Metropolitan Statistical Area	769	85167
36119	NY	WESTCHESTER COUNTY	1004457	333530	38862	103340	7.6	130202	DEMOCRAT	Metropolitan Statistical Area	392	80397

All data was aggregated to the County level (Identified by the 5-digit FIPS number). The final data set contains the following variables:

Final Dataset

Two datasets were merged on common unique identifier: FIPS code. Final dataset contains a total population of 3,143 counties. Of which, 849 counties with available EV registration will be served as sampling dataset. Other 2,294 counties will be used for prediction. Variables for final dataset are shown below with data definitions:

Type	Quantitative	Name
Dependent	True	EV_COUNT: Electric vehicle registration counts by county FIPS
Independent	False	FIPS: The 5-digit County Level FIPS/GEOID code
Independent	False	STATE: Two-character State ID
Independent	False	COUNTY: County Name
Independent	True	TOTAL_POPULATION: Total population
Independent	True	COLLEGE_GRAD: Total population with a bachelor's degree or higher
Independent	True	UNEMPLOYED_COUNT: Total unemployed population
Independent	True	MEDIAN_INCOME: Median income reported for each county
Independent	True	POVERTY_RATIO: Percent of population out of total population below poverty level
Independent	True	MALE_MIDAGE: Total males between 45-years and 65-years age
Independent	True	MALE_OVER16: Total males over 16 years age
Independent	True	FEMALE_OVER16: Total females over 16 years age
Independent	True	PREF_PARTY: 2020 General election winning party
Independent	True	METRO_TYPE: Categorical value identifying if the region is Metro- or Micropolitan area (if applicable)
Independent	True	CHARGING_PORT_COUNT: Count of public EV charging stations
Independent	True	X2020_GDP: County GDP in 2020

Model Intent, Election, and Summary

Originally, our primary intent was focused on the emissions impact of EV adoption. However, as data exploration and collection progressed, it was determined available emissions data was only readily available at the state (not county) level and greatly varied by state or was unavailable for many states. Our focus has shifted into primarily identifying which factors impact EV adoption the greatest.

We achieved these goals by developing a multi-linear regression for analysis. Regression model development is performed in R. Variable selection is performed with a shrinkage-selection method such as Elastic Net Regularization to identify which factors impact EV adoption the greatest while at the same time deal with multicollinearity.

Quality of the model is evaluated via cross-validation using the 16 states of readily available data.

A summary of our goals for this project are listed below:

1. Primary: Determine which factors impact EV adoption the greatest.
    - Using the Census data and the 16 states of EV adoption data develop a validated multi-linear regression and determine which factors impact EV adoption greatest.
2. Secondary: Estimate EV adoption in the 34 remaining states based on the previously developed model.
    - Using the previously developed multi-linear regression, forecast adoption in states that do not readily provide EV adoption data.

Model Development

We introduced new interaction factors such as percentages in addition to raw population metrics. By log-transforming the various population and count data (EV_Count), we were able to reduce variance and improve our model fit. This resulted in an improvement of an initial R^2 value of approximately 0.4 to 0.6124 in the final model.

The `glmnet` package in R was used to create the model with Elastic Net Regularization. The responses and predictors are fed as a matrix of values to `glmnet`. Values were standardized using the package to ensure all features were weighted equally against each other during the tuning stage. Elastic Net Regularization requires a mixing parameter, α, ranging from 0 to 1 to define the balance between Ridge ($\alpha = 0$) and LASSO ($\alpha = 1$) regression components. Additionally, a penalty/shrinkage parameter, λ, must be defined for each model generated. The λ-value determines the extent to which fewer coefficients are favored in the model at the cost of increased error.

The following plot of MSE vs each α value depicts that the models were all relatively on par with each other:

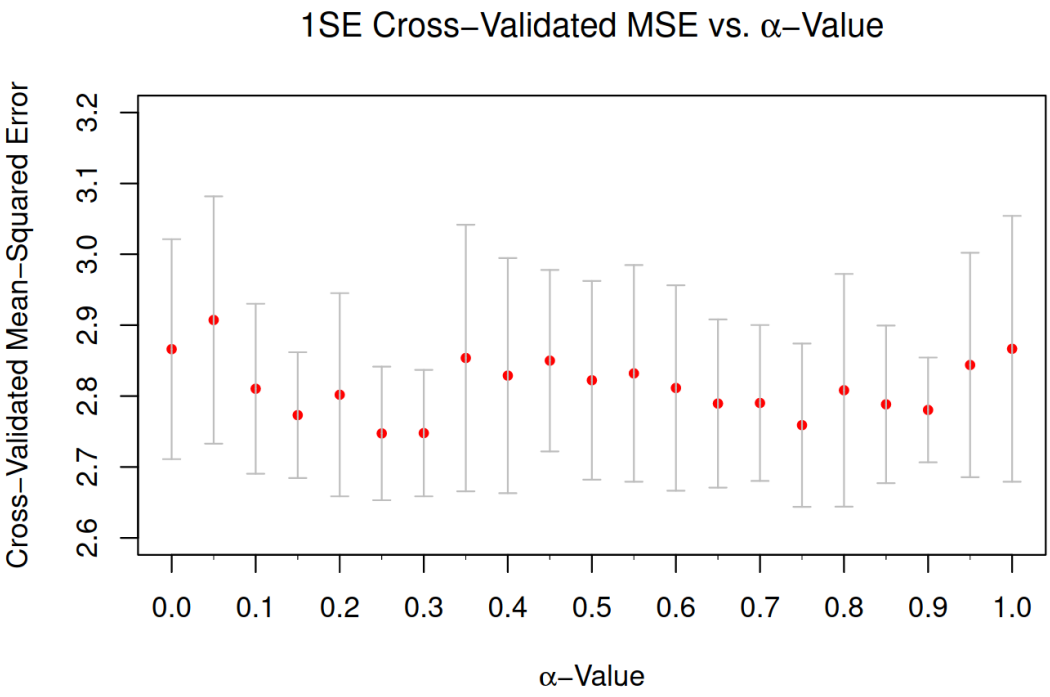

Figure 2: 1 Standard Error MSE vs. α

Using the λ at 1 Standard Error (1SE) MSE, the values are minimized at $\alpha = 0.25$, 0.3, and 0.75. We elected to use the model at an α value of 0.75 as this model favors LASSO regression which results in a more parsimonious model (i.e., fewer variables). Note that the 1SE-λ is generally preferred over the minimium-λ as it generates the model with fewer variables with only a slight increase in error.

Below is the plot of MSE versus log(λ) at the chosen $\alpha = 0.75$, with the lower x-axis displaying the log(λ), the y-axis showcasing MSE, and the upper x-axis displaying the number of variables (factors) remaining if a particular λ is chosen. Additionally, the left vertical dashed line represents the λ which results in a minimum MSE and the right vertical dashed line represents the λ which results in the 1SE MSE.

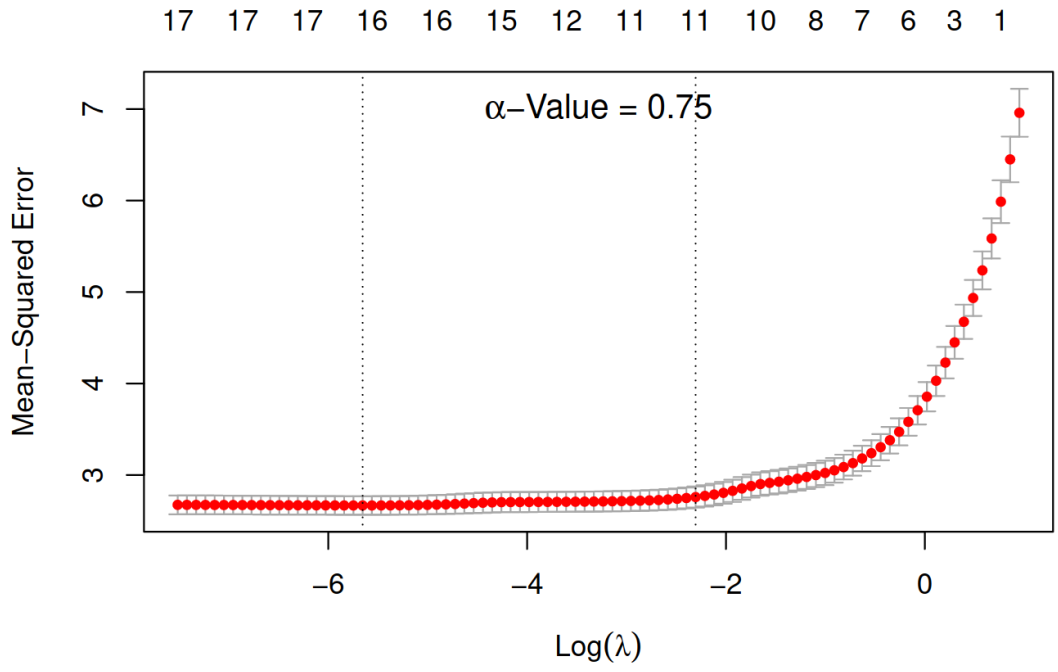

Figure 3: MSE vs. λ at $\alpha = 0.75$

We chose the 1SE λ to increase penalty favoring less variables, resulting in 11 variables in the final model.

Model Results

The following plot highlights the fitted values vs. the actual values of the final model:

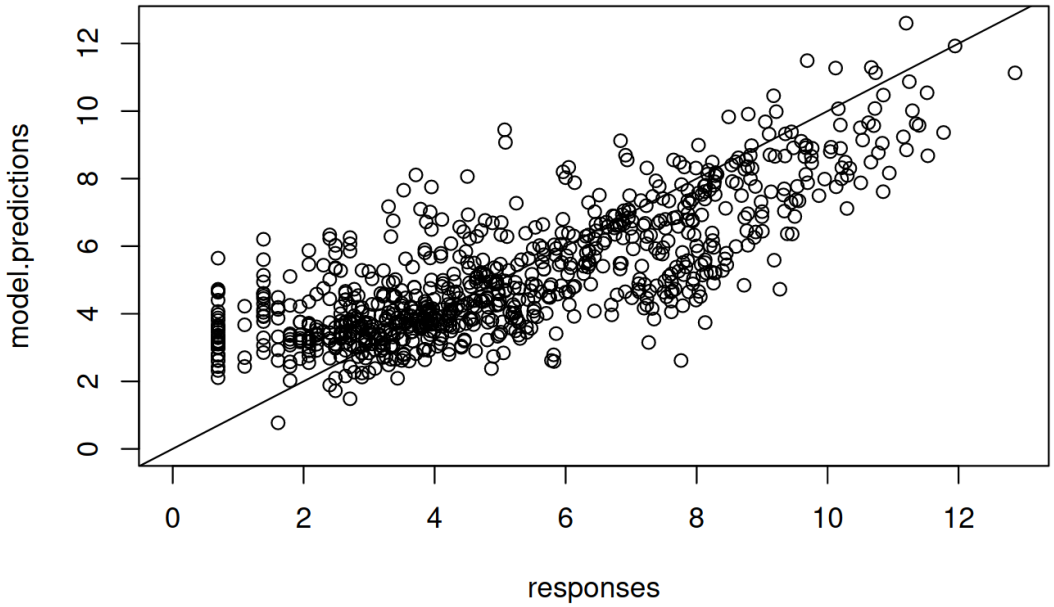

Figure 4: Fitted vs. Response

As observed in the plot above, the model's predictions under-estimate the groundtruth at the top end by up to a factor of approximately e^2 and over-estimate the groundtruth at the bottom by a factor up to roughly e^2.

The R^2 value of the final model (with log-transformation) is 0.6124.

The table below shows the resulting 11 selected coefficients from the original 18.

Coefficient	Value
(Intercept)	-3.662554e+00
PREF_PARTY_democrat	4.028185e-01
METRO_TYPE_metro	5.551182e-01
METRO_TYPE_micro	.
TOTAL_POPULATION	5.574941e-02
COLLEGE_GRAD	4.751138e-01
UNEMPLOYED_COUNT	.
MEDIAN_INCOME	2.994747e-05
POVERTY_RATIO	5.489383e-02
MALE_MIDAGE	.
MALE_OVER16	1.382409e-01
FEMALE_OVER16	.
CHARGING_PORT_COUNT	1.746109e-04
COLLEGE_GRAD_pct	5.407996e+00
UNEMPLOYED_COUNT_pct	-1.240783e+01
MALE_MIDAGE_pct	-5.913667e+00
MALE_OVER16_pct	.
FEMALE_OVER16_pct	.
GDP_2020	.

Some interesting insights can be made about the coefficient values. On average, and holding all other variables constant, the following is implied by the model:

1. Counties with a Democratic Party preference (PREF_PARTY_democrat) tend to have ~50% more EVs (vs. the baseline with a Republican Party preference).

- Note this binary predictor does not differentiate the level of preference for each party in each county.

2. Metropolitan Areas (METRO_TYPE_metro) have ~74% more EVs on average,
   - Micropolitan Areas (METRO_TYPE_micro) seem to have no impact on the number of EVs
3. A 1% increase in Total Population of a County (TOTAL_POPULATION) results in an increase of 0.056% in EVs.
4. A 1% increase in the absolute number of College Graduates (COLLEGE_GRAD) of a county results in a 0.475% increase in EVs.
   - A +1% increase in the percentage of College Graduates to Total Population (COLLEGE_GRAD_pct) results in 0.054% increase in EVs.
5. A $1000 increase in the Median Income of a county (MEDIAN_INCOME) results in a 3.0% increase in EVs.
6. A 1% increase in the absolute number of Males Over 16 (MALE_OVER16) results in a 0.14% increase in EVs.
   - Interestingly, a +1% increase in the percentage of middle-aged men to Total Population in a county (MALE_MIDAGE_pct) results in a 0.06% decrease in EVs.
7. A +1% increase in the Unemployment Rate (UNEMPLOYED_COUNT_pct) results in a 0.124% decrease in EVs.

The following heatmaps compare the EV count plotted on a US map between the ground-truth and the predictions from our model for the 16 states that have data:

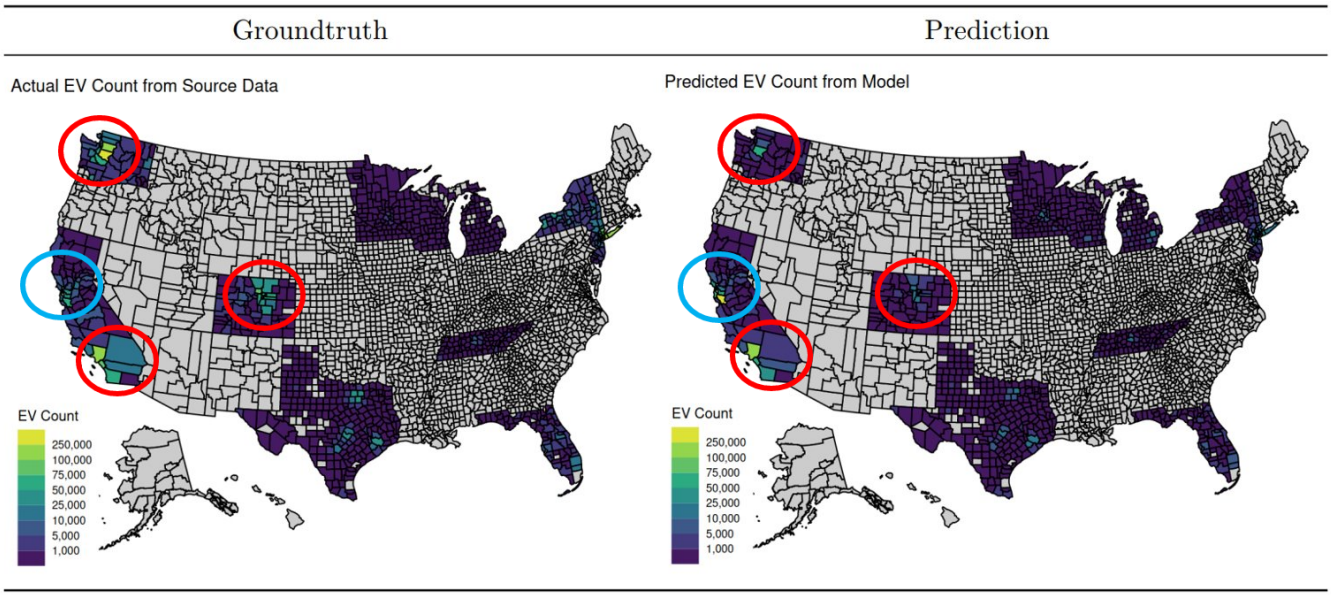

Figure 5: Heatmaps of 16 states, Groundtruth vs. Predictions

As previously described and from the above heatmap, it can be seen that the model tends to underestimate at the top-end and overestimate at the low-end. However, the model still recognizes areas where EV demand is greater than the average county which can be beneficial for targeting for specific business use cases.

Potential Model Improvements

Some potential improvements can be made to improve the model's predictive power and descriptive power.

For prediction improvement, we could explore the data further and apply more appropriate transformations to the features. Additional interaction variables may exist that would help predictions. Regarding the political preference, the data we have is currently binary for each county (Republican or Democratic); having a percentage or ratio data at county level will likely improve predictions.

Additional demographic features may help both descriptive and predictive powers. Finally with additional time, we could perform a detailed outlier analysis which would help both prediction accuracies and model explainability.

Model Conclusions in Regards to Stated Goals

Factors Impacting EV Adoption

We found the following factors impactful to EV adoption: Percent of College Graduates, Percent of Unemployed, Percent of Middle Age Males, Democratic party preference, Metropolitan demographics, Population count, Median Income and Poverty Ratio, Males Over 16 Years Old, and Charging Port counts.

These factors were eliminated by the model and likely have no significant effect on EV adoptions: Micropolitan demographics, Females Over 16, Absolute Unemployed Population, and County GDP.

Estimate EV Adoption of Different Counties for States in Absent of Data

Below is the model predictions for the remainder of the US counties for which data are unavailable. The heatmap highlights hotspots where the number of EVs is significantly greater than the status quo.

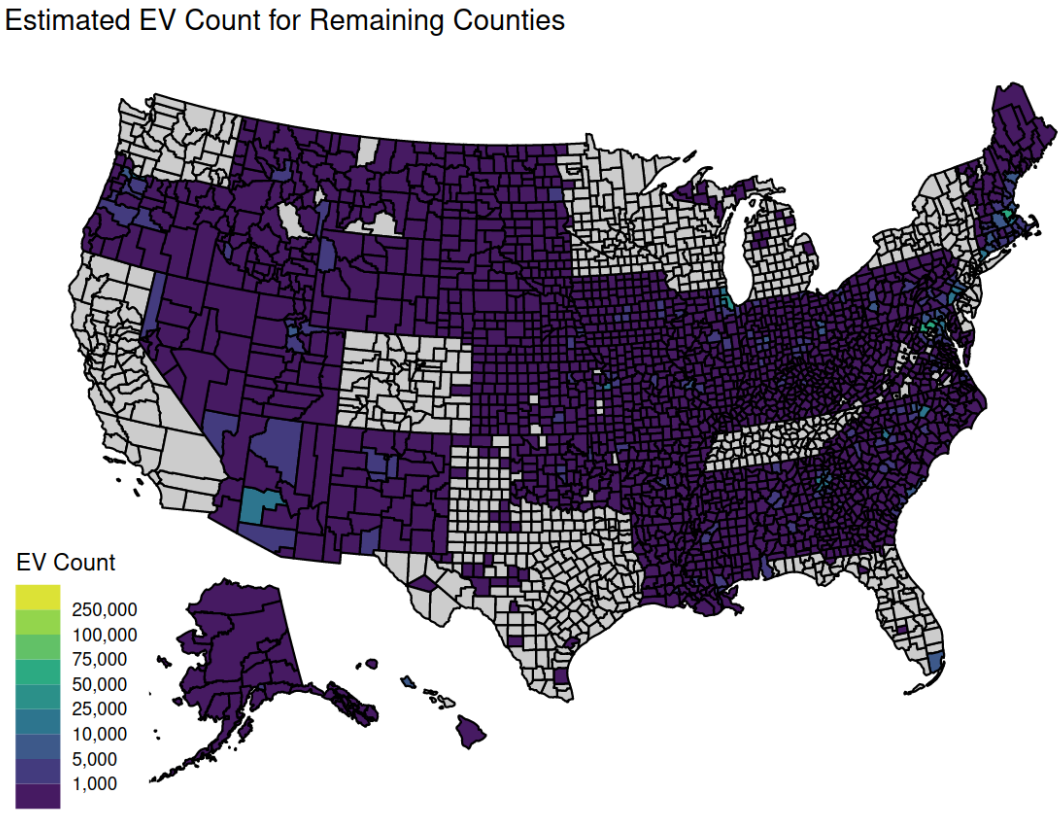

Figure 6: Model Prediction for US counties which do not have data

Model Discussion

We do have to keep in mind that correlation does not imply causation and this is an observational study based on data gathered from multiple sources. As discussed in the Data Section, the EV count data are voluntary submissions, with unexpected errors and missing values throughout. A multitude of features, interactions, and additional transformations are likely unaccounted for in this model. This likely impacts the descriptive and predictive powers of the model, therefore cautions should be taken to perform more experimental studies before suggesting any policy changes.

Team Collaboration and Project Timeline

The team formally met every Tuesday to discuss plans and milestones over Microsoft Teams along with ongoing conversation/discussion over Slack. Each team member was open to ideation and welcomed broad perspectives as a team. We collectively discussed the methodologies, algorithms and challenges related to analyzing the data which were stemmed from our course learnings.

We operated in an agile method continuing to refine the project plan as we gathered more clarity into the data and its potential. We made decisions as a team to pivot on the broader objectives to narrow the scope of our delivery in order to meet the requirements of the project.

Conclusion

Our modeling enabled us to plot the actual EV Count values compared to the predicted EV Count values. We found the definition of EVs in general is quite fragmented across jurisdictions which enables a promising opportunity to standardize across states and counties to further validate future adoption. Furthermore, this is just the beginning of transforming data into insights for making better business and environmental decisions.

References

1. Eichberger, J., Appelbaum, A., Hove, J. & Woods, D. *EV Consumer Behavior Report. Fuels Institute Electric Vehicle Council* https://www.fuelsinstitute.org/Research/Reports/EV-Consumer-Behavior/EV-Consumer-Behavior-Report.pdf (2021).
2. Dwyer, M. Electric Vehicles and Hybrids Surpass 10% of U.S. Light-Duty Vehicles Sales. *U.S. Energy Information Administration* https://www.eia.gov/todayinenergy/detail.php?id=51218 (2020).
3. Ruder, A. State EV Registration Data. *Atlas EV Hub* https://www.atlasevhub.com/materials/state-ev-registration-data/#data-format (2020).
4. Understanding Geographic Identifiers (GEOIDs). *United States Census Bureau* https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html.
5. Regional Economic Accounts: County GDP Download. *Bureau of Economic Analysis - U.S. Department of Commerce* https://apps.bea.gov/regional/downloadzip.cfm (2022).
6. Roth, J. Core based statistical areas (CBSAs) and combined statistical areas (CSAs). *NBER - National Bureau of Economic Research* https://data.nber.org/cbsa-csa-fips-county-crosswalk/cbsa2fipsxw.csv (2016).
7. Colley, N. US zipcode to county state to FIPS look up - dataset by Niccolley. *data.world* https://data.world/niccolley/us-zipcode-to-county-state (2020).
8. Sanders, A. County-level Data Sets. *U.S. Department of Agriculture - Economic Research Service* https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/ (2022).
9. Fast Facts on Transportation Greenhouse Gas Emissions. *United States Environmental Protection Agency* https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions (2018).
10. Shepardson, D. California Board Votes to Phase out Gasoline-Only Cars in State by 2035. *Reuters* https://www.reuters.com/business/sustainable-business/california-board-votes-phase-out-gasoline-only-cars-state-by-2035-2022-08-25/ (2022).
11. Cage, F. & Granados, S. The Long Road to Electric Cars in the U.S. *Reuters* https://graphics.reuters.com/AUTOS-ELECTRIC/USA/mopanyqxwva/ (2022).
12. Tanmoy, P., Bari, A. B. M. M. & Karmaker, C. L. An Integrated Principal Component Analysis and Interpretive Structural Modeling Approach for Electric Vehicle Adoption Decisions in Sustainable Transportation Systems. *Decision Analytics Journal* **4**, (2022).