Predicting the Electric Vehicle Market

Utilising Machine Learning and Socioeconomic Data to Understand Features That Affect Electric Vehicle Ownership

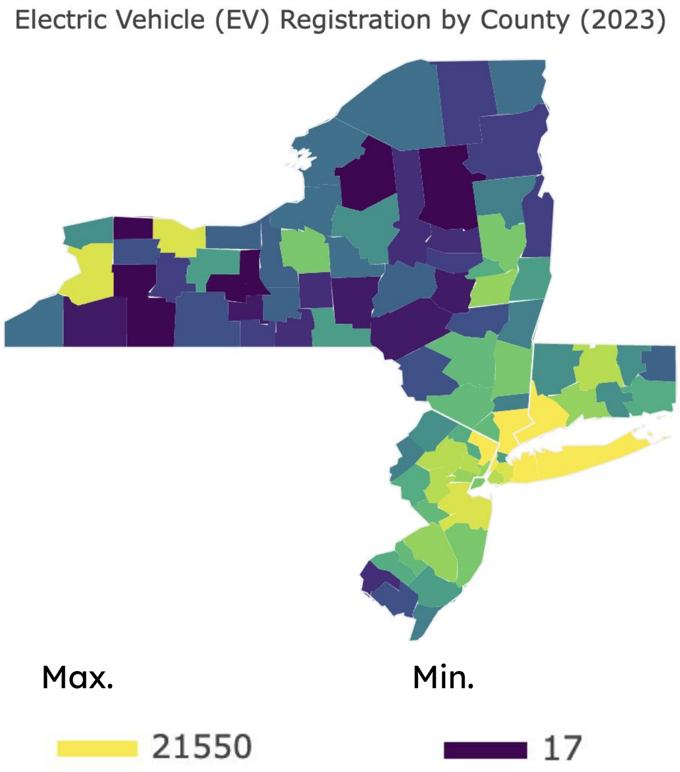
Daniel Gieseke | August 2023

Overview

The electric vehicle (EV) industry is booming. Whether it be at the automotive manufacturer level or within the EV charging space, EVs are becoming the new focus of the automotive industry. With this growth and focus, companies and public sector agencies alike are striving to understand the markets purchasing and adopting EVs within the United States (US). This project aims to understand the socioeconomic features that affect EV adoption in the US with the goal of capturing the markets currently adopting EVs, and better understanding how to pivot strategies to encourage EV adoption in slower-growing markets.

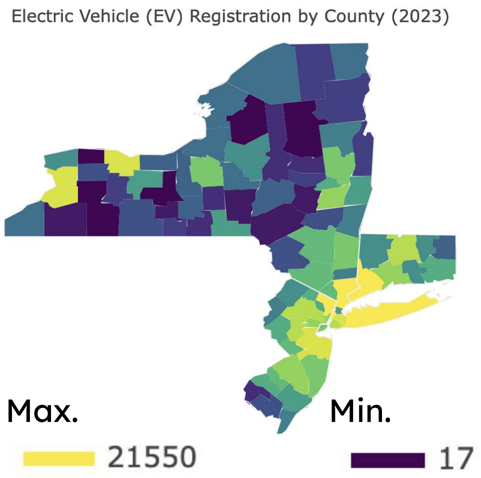
Data

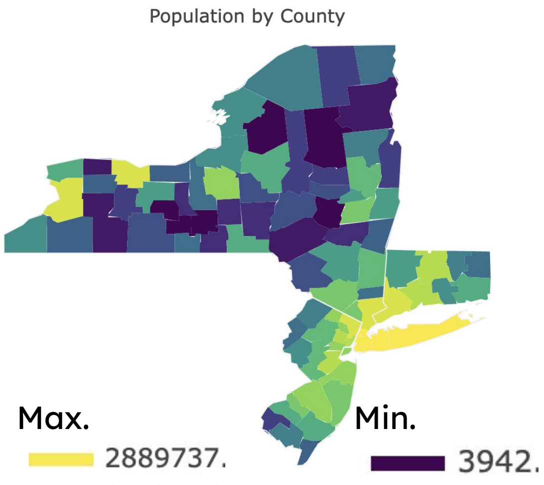
This project required the intake of data from several governmental agencies. Firstly, data on all EVs registered with the States of New York, New Jersey, and Connecticut were obtained, then grouped by ZIP code. From there, data was obtained from the US Census Bureau among other public-facing sources and joined with the EV registration data. This joining allowed for the production of several machine learning models that took in socioeconomic data by ZIP code from the US Census Bureau to predict and find correlations with the respective EVs within these locations across NY, NJ, and CT.

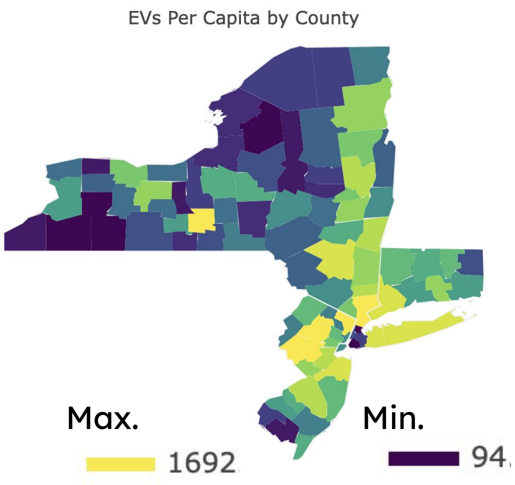


Data Spread

As highlighted in the visualization on the left, EV ownership and registrations are highly variable across New York, New Jersey, and Connecticut. EV ownership ranges from almost 22,000 in the county with the highest EV adoption rates, to 17 in the county with the lowest. This variance of EV ownership can leave gaps in EV infrastructure such as charging stations, highlighting the importance of researching and understanding these variations.







Data Spread

Even more curious, we may expect the number of EVs in an area to be directly correlated with population as are traditional internal combustion engine (ICE) vehicles, however, our data demonstrates a different story. Above we see that EV registration by county (left image) seems similar to population by county (center image). However, when looking at the EVs per 100,000 residents in our image on the right, we can see how EV adoption per capita ranges across geographies in our States of interest. This variance lends us an opportunity to understand the underlying socioeconomic features that are affecting EV adoption rates throughout this region.

Features Most Correlated with EV Ownership

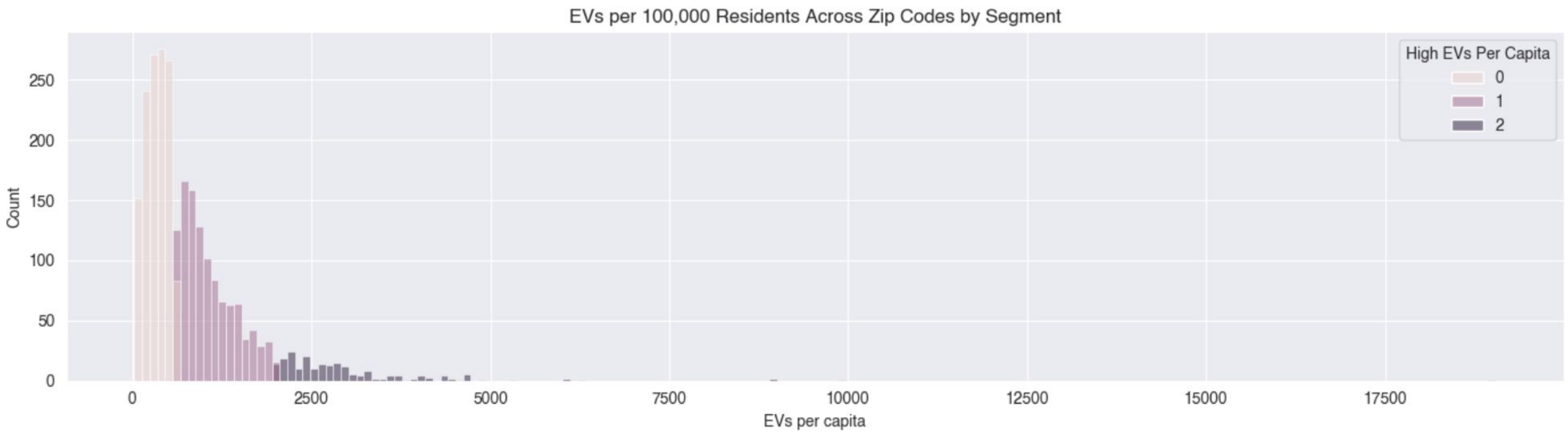
Within our data collected, there are three factors that stand out the most when predicting EVs registered within a ZIP code: Median Household Income, Percent of Population with Bachelor's Degree, and Population. The regression plots above illustrate the correlation these factors hold with EVs registered by ZIP code. Other notable, correlated features include the number EV charging stations within a ZIP code as well as well several geographic features (e.g. whether a ZIP code is within a New York City suburb). From these features collected, we are able to create our first model.

Model 1

The first machine learning model created in conjunction with this project is a ridge, linear regression predicting the total number of EVs within a ZIP code. Our model took in 18 different geographic and socioeconomic features that were able to predict the number of EVs registered in a zip code with a mean absolute error of 31 vehicles. This model yields an R-Squared score of .82, meaning 82% of the variance in our target (EVs in a ZIP code) is predicted by our model. As flagged previously, population, income, and education features served to be among the most significant to this model.

Mean Absolute Error:
31 Vehicles

R-Squared Score:
.82



Model 2: Segmenting the Data

With model 2, we aim to predict not just the total number of EVs in a ZIP code, but the EVs per capita. In analyzing the distribution of EVs per 100,000 residents across zip codes in our data, though not exceptionally distinct, there seems to be three segments within the data. After ascertaining the median among other statistical measures in the distribution of this feature, a categorical feature was created for this second machine learning model. This target, “High EVs Per Capita”, is broken into three groups: zip codes with below average EV adoption rates per capita, zip codes with above average adoption rates per capita, and zip codes with exceptionally high adoption rates per capita.

Model Used:
Random Forest Classifier

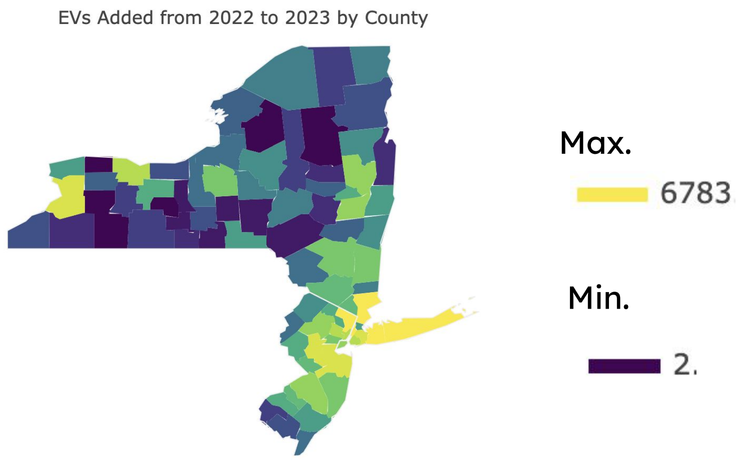
Accuracy Score:
82%

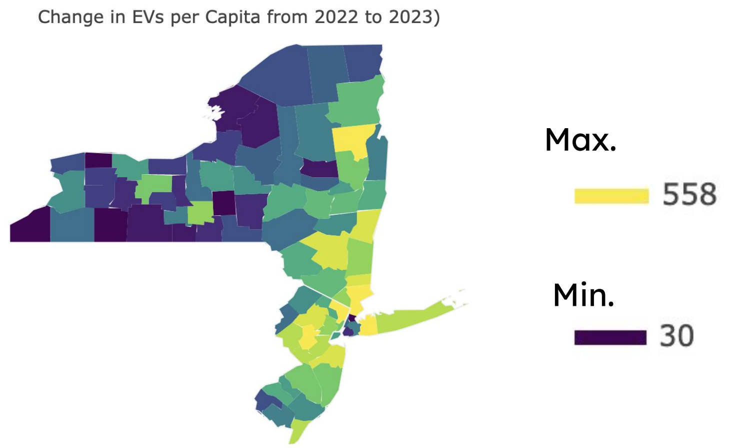
Model 2

Following the segmentation of this target feature, a random forest classifier model was successfully created yielding an accuracy score of 82%. Similar to the first model created with this project, income and education stood out in terms of feature importance in this model. Other features with relatively high importance to this model include population, distance from New York City, as well as several unique, feature engineered variables that are more complex to interpret.

Vehicles Added in 2022

The last feature investigated in conjunction with this project was vehicles added in 2022 compared to all previous years. Looking specifically into this feature lends the opportunity to understand if the historical features that predicted EV adoption and ownership in the region stay true to the present. As the two previous models illustrate that income and education reigned as the top two predictors, this third model may tell us that other socioeconomic features may better predict EV adoption throughout this region in the contemporary.





Model Used:
Linear Regression

R-Squared Score:
.58

Model 3

Despite the third model utilising the same same features and algorithm as our first model, our R-Squared score dropped significantly and the significance of individual features in this model changed dramatically as well. With these changes, it may be possible to suggest that the features that historically have had the largest effect on EV ownership and adoption in the past could be changing.

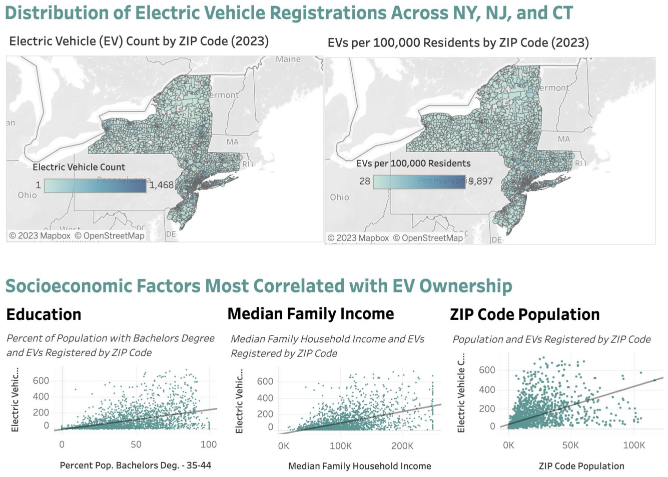
More Data!
Gather data at national level

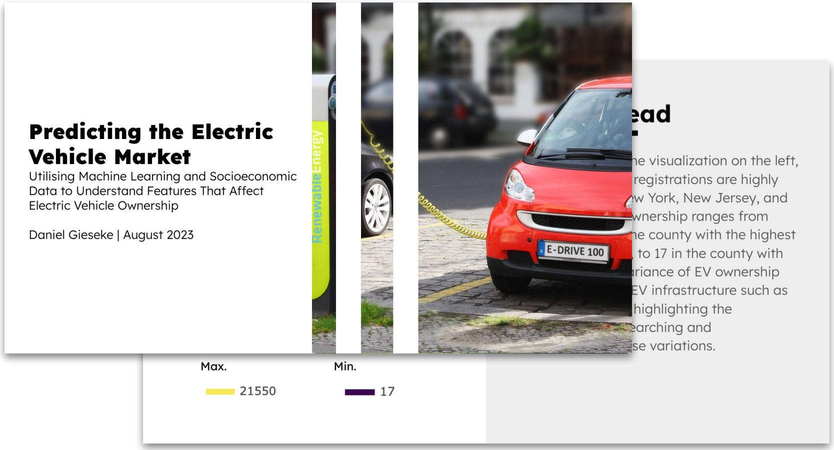
Encourage Incentives!
Provide more tax breaks and/or cash incentives based on income groups

Future Work and Recommendations

The first recommendation for future work on this project is to expand geographies analysed for this project from NY, NJ, and CT, to other states.

The second recommendation ascertained from this project, is to encourage more incentives for EV ownership. The data analyzed illustrates that EV ownership largely remains contingent on income and education; providing more incentives may allow the adoption of EVs by other income brackets.





Deployment

In addition to the data preparation, exploratory data analysis, and modelling prepared for this project, two products have been deployed for this project. Firstly, a dashboard was created and posted on Tableau Public to analyze the geographic spread of this data as well as visualizations of the relationship between EV ownership and select socioeconomic features. Secondly, this report has been created and posted on GitHub.

Questions?

Thank You!

Contact: Daniel T. Gieseke | daniel.tyler.gieseke@gmail.com
Connect: LinkedIn
Portfolio: GitHub | Tableau