Jack Hu, Boyuan Lu, Haoxiang Yi, Jaida Adamson

General Motors: Strategy for EV and EV Charging Stations

Background

General Motors (GM) is aware that Electric Vehicles (EVs) are becoming more crucial in the car industry, and they aim to grow their EV range. They asked our team to help them research this industry, specifically focusing on investigating the factors affecting EV usage and charger location. The main goal of this report is to give GM suggestions and understandings for shaping their EV strategy by answering these important questions:

1. What are the factors that contribute to EV adoption, and what role does charging infrastructure play?
2. What areas (zip codes) should see the most growth if new EV chargers are installed?
3. Where, specifically (venues), should EV chargers be installed in the highest priority areas?

Approach

To respond to the main questions, our group set up two target variables: "EV Adoption Rate" and "Charging Stations Count." We forecasted the EV adoption rate using all useful variables, not including those that caused data leakage or could be utilized to calculate the original target variable. By finding the most important variables for predicting adoption, we also understood how GM should promote EVs and improve its market share.

Likewise, we anticipated the Charging Stations Count by using all useful variables and discovered the most vital factors that affect this number. Our examination assisted in identifying the places for charger setup which hold more influence. We emphasized on top 100 zip codes having considerable contrast between forecasted and factual charging station counts. From the study of zip codes, we gave suggestions on where to put 100 charging stations for maximum effect in increasing the EV adoption rate.

Starting from predicting the EV Adoption Rate, we utilized every related variable. We left out variables that resulted in data leakage or allowed us to calculate the main target variable. Selecting important variables that influenced EV adoption helped us find out how GM could best promote EVs, using these influential factors as guidelines.

After that, we concentrated on forecasting the Charging Stations Count. We utilized every pertinent variable to estimate how many charging stations there should be and identified which variables had the most impact on this count. This analysis helped us pinpoint the utmost important places for charger placement according to those recognized variables.

Ultimately, we looked into the 100 zip codes with the biggest gaps in charging station count forecasts versus actual counts. This allowed us to concentrate on these areas and comprehend what elements are adding to the variations, along with formulating suggestions for distributing 100 charging stations to get the highest marginal rise in EV adoption percentage.

In this way, we presented GM with practical knowledge and suggestions to steer their EV adoption and charger location tactics. Our study recognized the main elements that affect EV acceptance and suggested the best spots for setting up additional charging points for maximum influence on EV adoption percentages.

Data Collection and Transformation

For our analysis, we had to gather data from different places. We gathered information about weather, like average precipitation and temperature for each zip code, to understand how climate might affect EV adoption and where charging stations should be placed. Data connected with tourists, such as the number of tourists and hotels per zip code or county, were collected so as to evaluate how tourism can influence EV infrastructure requirements. We got data about traffic like how many main roads and transit places there are in every zip code or county. We use this information to assess the connection between transport networks and the adoption of EVs. Also, we gathered details on where public and private charging stations are located, along with their counts for understanding the existing condition of EV infrastructure. Venue data was collected too; it tells us how many retail establishments, offices, parking facilities as well as medical/educational institutions are present per zip code or county. This helps identify possible spots for new charging stations.

The data our team gathered was cleaned and preprocessed. We grouped it at the county or zip code level, as per the granularity of available information. The main data from GM was combined with extra influential data we collected to create a full dataset. For handling missing values, we grouped by county and calculated averages to fill in the variables that were missing. Categorical variables were transformed into dummy variables to facilitate analysis.

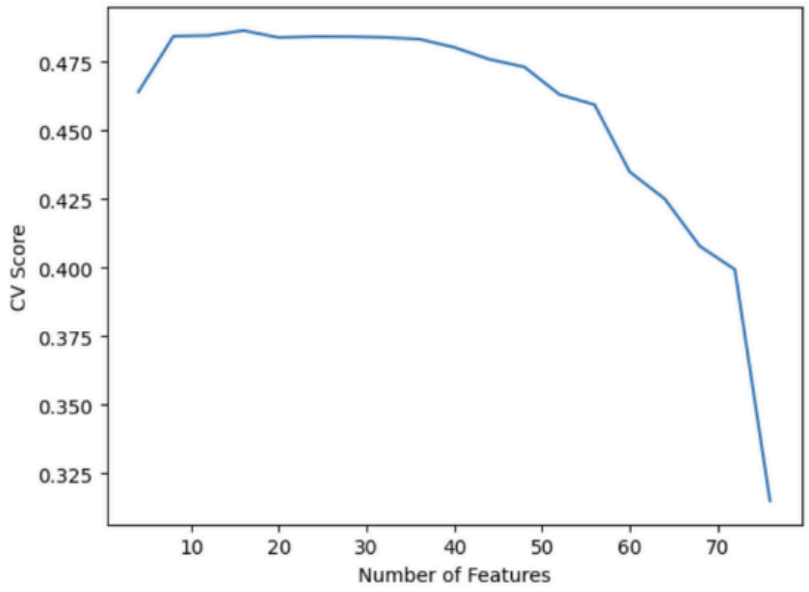
In data integration, we filtered the dataset to Texas state only, as asked in the project. When merging datasets from various sources, we maintained uniformity by meticulously examining data structures and resolving any differences. Using feature engineering, we created new columns to enhance the explanation of target variables. Let's say we made a column that shows the proportion of charging stations for EVs in every zip code or county. This gives us an understanding of how easily accessible the charging infrastructure is.

The dataset is comprehensive and ready for analysis. It was gathered from many sources, transformed into a usable format, and includes all the variables we need to answer questions

about Texas EV strategy. The variables in this dataset consist of weather data (temperature, humidity), tourism details (number of visitors in Texas each year), traffic information like daily vehicle miles traveled or DVMTs on average by an individual Texan per day as well as charging station numbers along with their locations' current status such as available or being used at present time etc., and venue details (number of restaurants, hotels, shopping centers). All these variables are aggregated at proper geographic levels to match the needs of our analysis. We have a complete set of data for making recommendations on how GM can best implement their EV strategy in Texas, thanks to the robustness and integration provided by this dataset.

EV Adoption Rate Modeling

First, data exploration was performed on the target variable, and a highly skewed distribution was obtained, clustering a majority of data points towards the lower end of the EV adoption rate spectrum, close to zero. Few cases of higher EV adoption are available, hence the long tail seen on the graph pointing towards the right. One can conclude, thus, from the visualization that adoption of EVs is generally low in the dataset, with few areas or groups displaying higher value adoptions. Applying a logarithmic scale to the data could be seen to provide the distribution in a more normalized form, hence preferable for some type of statistical modeling. We applied log transformation to the target variable. We started by introducing each feature into the model one by one and see how they performed based on the CV score. Looking at the graph below, as The CV score starts at a relatively high value when the number of features is small (near 4), we can infer that a few features already provide much predictive power.



The major drop is after about 50 features, and here the CV score actually increases majorly, further telling us that including more features at this point is detrimental to model performance. Looking at the graph, we can consider the best number of features used in this linear regression model to lie somewhere around 20 since that's where the model seems to give a balance of complexity enough to capture underlying patterns but yet simple enough not to overfit. We then fit the testing set to the data, where it yields an R-squared value of 0.732 and an MSE of 1.59e-05. We will then apply the same approach but use a random forest model and see how the model performs. The best number of features for the random forest is 40. Comparing the two models, both models have an R-squared close to 0.732, indicating that 73.2% of the variance is explained for either model in the adoption of EV. This is a pretty decent level of explanatory power for such complex social phenomena. The Mean Squared Errors are also quite close, wherein the Linear Regression model has a slightly lower MSE (1.5907e-05) than the Random Forest model (1.5910e-05). The difference is not significant, saying they would have almost equal accuracy when making predictions over test data. But, in general, a Linear Regression model is simpler and trains faster than a Random Forest model, and it has the advantage of interpretability. Each of the coefficients of the linear model relates to one feature and gives direct relations between the feature and the target variable. These are the advantages that could go a long way in the considerations for deployment—definitely simplicity and speed in achieving performance. Linear models are simple to implement and computationally inexpensive, hence very suitable for environments where these are limiting factors. Analyzing the intersection of features between the models offers several insights:

- Policy Implications
    - RUCC_2013: This suggests that the level of urbanization or rurality of an area plays a role in EV adoption. Policymakers might focus on tailoring incentives for different areas based on their urbanization level.
    - Transport_car_rate: The importance of car usage rates indicates that regions with higher reliance on cars might be more receptive to adopting EVs. Infrastructure investments like charging stations could be strategically placed in these high-usage areas.
    - WFH_rate: The work-from-home rate's significance could reflect a lifestyle or socio-economic status that correlates with EV adoption. It could also hint at potential EV adopters having more flexible schedules to manage charging times, suggesting that promoting home charging solutions might be effective.
- Economic Considerations
    - Individual_income_percapita: Higher individual income levels being a predictor point to the affordability of EVs. As such, financial incentives or models that lower the effective cost of EVs could be crucial in adoption rates.
    - LPG_stations: The presence of liquefied petroleum gas (LPG) stations as a predictor could imply that areas with alternative fuel options are more open to

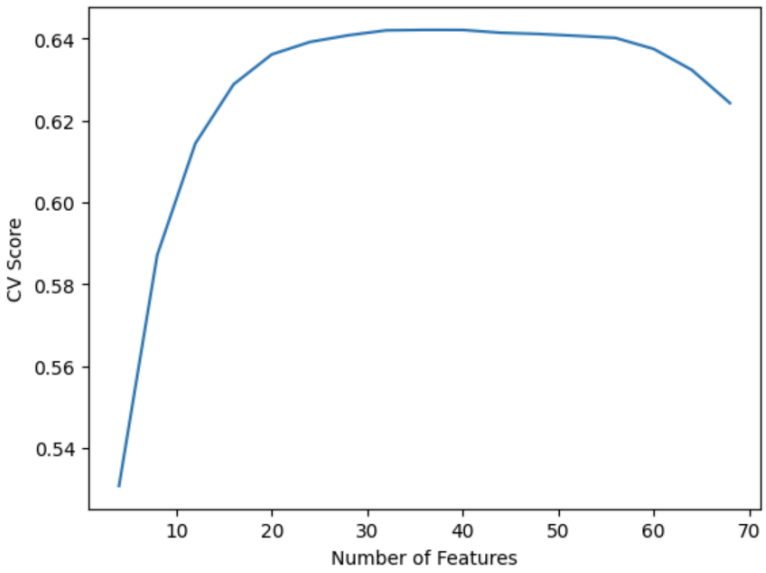
adopting different types of vehicles, including EVs. This could guide the expansion of EV infrastructure in regions already accustomed to alternative fuels.

- Tourism and Hospitality Factor:
    - Total_Hotel_Airbnb: The significance of accommodations like hotels and Airbnb might suggest that areas with higher tourism have greater visibility of EVs and potentially more charging infrastructure. This can inform marketing efforts, indicating that exposure to EVs through travel experiences could influence adoption rates.

Charging Stations Modeling

The second target variable in our project is the Number of Charging Station. This target variable is used to show how many charging stations each zip code should have based on all independent variables. First, data exploration was performed on the target variable, and a similar result from the previous section was obtained as the data is highly skewed. A few zip codes have a large number of charging stations while the majority don't have any stations. Therefore, applying a logarithmic scale to the data could be seen to provide the distribution in a more normalized form, hence preferable for some type of statistical modeling. Similar to the EV adoption target variable, we decided to construct both a linear regression model and a random forest model for this variable.

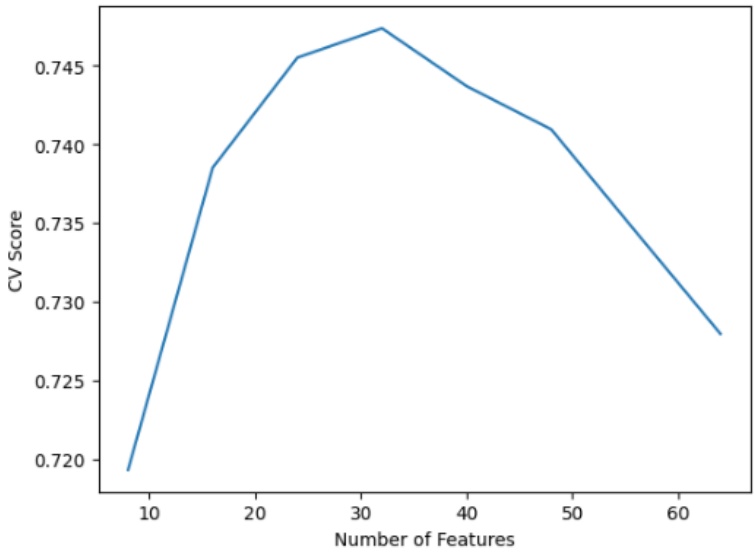
For the linear regression model, we started with the feature selection process. Similar to the approach in the previous section, we introduced each feature into the model and saw how they performed based on the CV score. Looking at the graph below indicates the optimal number of features is 36:



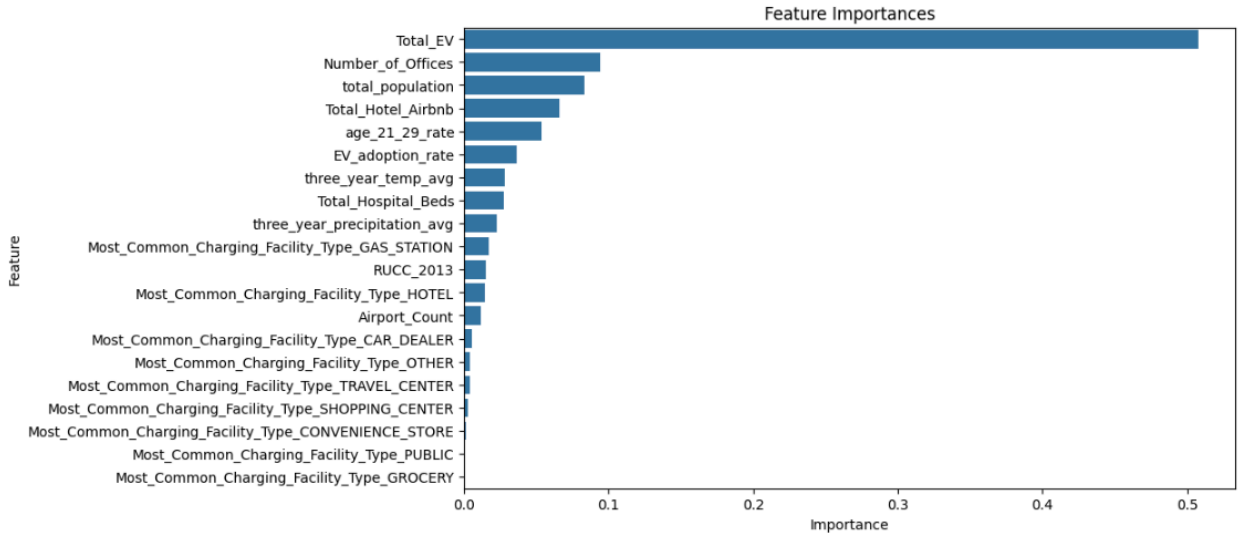
The forward selection method is used to select the number of features needed for this model. The linear regression model shows that there are 28 statistically significant variables, which proved hard to analyze given the amount and different nature of these variables. The complete result is shown below:

RUCC_2013	-0.0879	0.009	-10.145	0.000	-0.105	-0.071
total_population	6.179e-06	1.42e-06	4.342	0.000	3.39e-06	8.97e-06
age_0_20_rate	0.7304	0.183	3.982	0.000	0.371	1.090
age_21_29_rate	1.8680	0.268	6.974	0.000	1.343	2.393
age_30_39_rate	1.3135	0.336	3.913	0.000	0.655	1.972
age_50_64_rate	0.2491	0.244	1.019	0.308	-0.230	0.729
age_65_more_rate	0.4669	0.212	2.197	0.028	0.050	0.884
ethnicity_wnh_rate	-0.3099	0.092	-3.362	0.001	-0.491	-0.129
transport_public_rate	11.8130	2.656	4.447	0.000	6.603	17.023
average_traveltime_work	0.0054	0.003	1.765	0.078	-0.001	0.011
individual_income_percapita	1.135e-05	1.47e-06	7.708	0.000	8.46e-06	1.42e-05
Total_EV	0.0004	7.72e-05	5.307	0.000	0.000	0.001
Traffic_Count	-5.931e-09	2.24e-09	-2.652	0.008	-1.03e-08	-1.54e-09
Total_Hotel_Airbnb	0.0002	3.82e-05	5.402	0.000	0.000	0.000
three_year_precipitation_avg	0.0481	0.018	2.659	0.008	0.013	0.084
three_year_temp_avg	-0.0010	0.001	-0.861	0.389	-0.003	0.001
Most_Common_Charging_Facility_Type_AIRPORT	1.8061	0.517	3.494	0.000	0.792	2.820
Most_Common_Charging_Facility_Type_BREWERY_DISTILLERY_WINERY	0.2257	0.295	0.766	0.444	-0.352	0.804
Most_Common_Charging_Facility_Type_B_AND_B	0.3069	0.658	0.467	0.641	-0.983	1.597
Most_Common_Charging_Facility_Type_CAR_DEALER	0.7933	0.097	8.200	0.000	0.604	0.983
Most_Common_Charging_Facility_Type_FED_GOV	0.7633	0.210	3.630	0.000	0.351	1.176
Most_Common_Charging_Facility_Type_GAS_STATION	1.5009	0.139	10.777	0.000	1.228	1.774
Most_Common_Charging_Facility_Type_GROCERY	1.6798	0.466	3.603	0.000	0.765	2.594
Most_Common_Charging_Facility_Type_HOSPITAL	0.8161	0.382	2.136	0.033	0.067	1.566
Most_Common_Charging_Facility_Type_HOTEL	0.7272	0.065	11.251	0.000	0.600	0.854
Most_Common_Charging_Facility_Type_MUNI_GOV	0.6565	0.330	1.990	0.047	0.009	1.304
Most_Common_Charging_Facility_Type_OFFICE_BLDG	0.5472	0.211	2.594	0.010	0.133	0.961
Most_Common_Charging_Facility_Type_OTHER	0.8589	0.151	5.704	0.000	0.564	1.154
Most_Common_Charging_Facility_Type_PARKING_GARAGE	0.6502	0.334	1.949	0.051	-0.004	1.305
Most_Common_Charging_Facility_Type_PARKING_LOT	0.4398	0.220	1.995	0.046	0.007	0.872
Most_Common_Charging_Facility_Type_PAY_GARAGE	1.8369	0.658	2.790	0.005	0.546	3.128
Most_Common_Charging_Facility_Type_PUBLIC	1.8274	0.465	3.928	0.000	0.915	2.740
Most_Common_Charging_Facility_Type_REC_SPORTS_FACILITY	1.1601	0.657	1.766	0.078	-0.128	2.449
Most_Common_Charging_Facility_Type_SHOPPING_CENTER	0.7725	0.153	5.036	0.000	0.472	1.073
Most_Common_Charging_Facility_Type_TRAVEL_CENTER	1.9499	0.329	5.921	0.000	1.304	2.596
Most_Common_Charging_Facility_Type_TRUCK_STOP	0.7468	0.383	1.948	0.052	-0.005	1.499

Next, we utilized the same approach to the random forest model. We applied the logarithmic transformation for the target variable as well and performed feature selection. According to the CV score, the optimal number of features is 32 as shown below:



Similar to the linear regression model, forward feature selection is used to select the optimal features. Once the model is built, we extracted the importance of each feature and formulated the following graph:



In the random forest model, Total EV, number of officers, and total population have the highest feature importance among all features. Both models are compared using R-Squared value and Mean Standard Error (MSE), and the results are shown below:

	R-Squared	MSE
Linear Regression	0.718	0.621
Random Forest	0.805	0.267

This table clearly shows that the Random Forest model has a better performance for this dataset, hence it was used for predictive purposes in the next step. We also identified 14 variables that are significant for both models, and they would be good indicators of having a large number of charging stations in the future.

RUCC_2013	Most_Common_Charging_Facility_Type_CAR_DEALER
total_population	Most_Common_Charging_Facility_Type_GAS_STATION
age_21_29_rate	Most_Common_Charging_Facility_Type_GROCERY
Total_EV	Most_Common_Charging_Facility_Type_HOTEL
Total_Hotel_Airbnb	Most_Common_Charging_Facility_Type_PUBLIC
three_year_precipitation_avg	Most_Common_Charging_Facility_Type_TRAVEL_CENTER
three_year_temp_avg	Most_Common_Charging_Facility_Type_SHOPPING_CENTER

For the next step, the random forest model was used to predict how many EV charging stations each zip code has based on the selected variables. Then the predicted value is subtracted from the actual number of EV chargers to get the difference. A positive difference would indicate this zip code should have more chargers whereas a negative difference indicates the opposite. The top 10 zip codes with the largest difference is shown below:

Zip Code	Difference
78206	8
78731	7
78751	7
78256	6
77006	6
77003	5
75025	5
75251	5
77407	4
75039	4

Our model shows that these zip codes would require the most chargers to be placed. We also selected 100 zip codes with the largest difference and the list is used for the marginal analysis portion.

Marginal Analysis

Marginal analysis in this context refers to the evaluation of the incremental benefits gained by adding additional EV charging stations in specific locations. This analysis is crucial for determining where to place new chargers to achieve the maximum increase in Electric Vehicle (EV) adoption rates.

The marginal analysis utilized iterative modeling techniques, specifically employing random forest models to predict the impact of adding charging stations one at a time in various zip codes. This method helps understand the sensitivity of EV adoption rates to changes in the number of available charging stations.

The analysis focused on zip codes with the greatest discrepancy between the predicted and actual number of charging stations. Here are some of the key findings:

- Zip Code 77046 (Harris County) had the highest potential increase in EV adoption rate, projected at 0.388479862, with the addition of 67 new charging stations leading to an estimated 487 new EVs.
- Zip Code 78751 (Travis County) showed a moderate increase in EV adoption rate of 0.049241123 with an addition of 10 charging stations, potentially adding approximately 804 new EVs.
- Zip Code 75206 (Dallas County) could see an adoption rate increase of 0.043053624 by adding 9 new stations, translating to about 1655 new EVs.
- Zip Code 77407 (Fort Bend County) could achieve an adoption rate increase of 0.063810303 with 7 new stations, contributing to a significant 4888 new EVs.

Additional insights from the analysis included:

- The total number of new EVs potentially added across the analyzed zip codes amounted to 9,193.
- The economic impact was significant, with potential revenue from these new EV sales estimated at approximately $491 million, based on the average vehicle price from the 2023 Kelly Blue Book.

The marginal analysis provides a detailed roadmap for strategically allocating resources to enhance EV adoption effectively. By focusing on zip codes where additional charging stations

yield the highest increase in adoption rates, General Motors and policymakers can optimize investments and accelerate the transition to electric vehicles.

Conclusion

In this report, we conducted an in-depth analysis to help inform General Motors' electric vehicle strategy. We looked at what makes people use EVs and the best places for deploying charging stations. In terms of factors affecting EV adoption rate, the degree of urbanization, car usage rates, lifestyle factors such as work-from-home rate, income levels, and the presence of alternative fuel stations like LPG are the most significant variables correlating with higher EV adoption. This implies GM should customize incentives based on these elements.

The model we used pointed out certain zip codes in Texas. These areas would experience the biggest growth in EV use if more charging stations were set up there. The main areas are 77046, 78751, 75206, 77407, 75022,76049 and also the region around ZIP code 75056. Placing chargers carefully in these spots is expected to have a big effect.

In terms of placing EV chargers in priority zip codes, our examination shows that car dealerships, gas stations, grocery stores, hotels, public parking areas and travel centers along with office buildings are the most suitable places.

If we direct our infrastructure investments towards these places and venues with high potential, we predict that GM can encourage the use of more than 9,000 extra EVs. This could result in almost $500M as fresh revenue.

In summary, by using our data-based method, we have made recommendations that can be put into action on the areas and methods GM should use to grow charging accessibility in Texas. This will help speed up EV usage in the state's market while also establishing GM as an important figure in this electric car revolution.