Gap Analysis for U.S. EV Charging Infrastructure Using Supervised Machine Learning

Sanjeeb Adhikari
Master of Science in Data Science
Regis University
sadhikari006@regis.edu

Abstract

The rapid growth rate of electric cars (EVs) has brought to light serious deficiencies with the geographical distribution of public charging stations that serve EV drivers. This proposal outlines a data framework for identifying underserved areas within the EV charging network in the U.S. through the use of supervised machine learning analysis. An XGBoost regression model will analyze the factors affecting usage patterns in more established EV markets (California, Colorado, and Vermont) so that it can develop a predictive demand profile based on factors such as demographics, transportation, and built environment (population density, traffic volume, and housing type, etc.). Once the model has been successfully trained with established markets, it will be used to assess latent demand in both emerging and underserved regions of the U.S. By comparing predicted demand against current charging capacity, the study will quantify deficiencies in charging station infrastructure on a U.S. Census Tract basis, and produce a prioritized list of locations where future charging station investment would be most appropriate.

I. Introduction

The World is transitioning to sustainable transportation with the introduction of electric vehicles (EVs). While progress has been made with increased adoption of EVs, the availability of public charging infrastructure remains uneven resulting in issues such as long wait times for charging, and limited access in some regions[1]. How planners currently determine where to deploy charging stations is largely dependent on the density of EV ownership and charger availability in an area resulting in chargers being placed only where demand already exists.

This project aims to enhance the existing method of infrastructure planning for EVs by using a predictive and demand-based approach. Current infrastructure planning relies primarily on existing charging or vehicle registration as a measure of where to provide additional chargers; this project will utilize supervised machine learning to predict the expected demand for EV charging in any area, based on demographic, transportation, and land use characteristics. This approach will allow planners to identify underserved areas proactively and deploy infrastructure more efficiently and equitably.

II. Problem Statement

Charging infrastructure deployment strategies often suffer from a supply side bias. New charging stations tend to be concentrated in locations where people have already adopted electric vehicles (EVs) in large numbers [2]. As a result, many underserved areas may not have enough charging stations to accommodate potential future EV drivers. In addition to this, current methods of planning for charging stations fail to capture the overall impact of population density, traffic patterns, housing types, and socioeconomic conditions on future charging needs.

Research Question: Will an XGBoost regression model using supervised machine learning trained on established electric vehicle markets provide accurate and timely predictions of latent charging needs for electric vehicles in emerging markets while also identifying areas with insufficient charging infrastructure down to the census tract level?

III. Literature Review

As per the U.S department of Energy, there were around 4.7 million electric vehicles on U.S road as of 2025, and the number are growing more significantly over year. This rapid growth has increased the urgency of reliable and well distributed charging infrastructure, but the deployment of EV charging infrastructure has not kept the same pace as of EV adaptation. Because of this one of the most common barrier to Ev adoption is range anxiety, specially for long-distance travel along highways and interstate where charging station availability is uncertain [3].

Although the expansion of public charging stations are growing in recent years, but their geographic distribution have remains uneven [4]. As per the U.S. Department of Energy's Alternative Fuels Data Center, US currently hosts over 180,000 public charging ports, among them roughly 40,000 are DC fast chargers. Although the numbers looks large, charger density are not even across all regions, with urban areas receiving more infrastructure investment as compared to rural. This uneven distribution of infrastructure has been highlighted in multiple studies as a key challenge for large-scale EV adaptation [4].

Many studies ha shows different challenges of EV charging infrastructure planning. As per [5], there are currently three key problems with EV charging infrastructure: 1) Geographical inequity — widespread shortages, especially in rural and underserved communities; 2) Trip demand spikes, sudden and unmanageable demand spikes from conventional EV use; 3) Fragmented policy implementation, as recently happened to the NEVI Program freezing three billion dollar in state allocations, delaying its 4,000 planned charging ports. These factors make uncertainty in infrastructure planning and may also delay effective deployment of charging stations. Also, research by [2] suggest that the limited access to charging infrastructure can reduce EV adoption rates by more than 30 percent, which shows the importance of reliable and well distributed infrastructure planning.

To address these challenges, many early research focused mainly on optimization based charging station placement models. These approaches mainly uses traffic flow patterns, travel demand, and transportation networks to determine optimal charger location. For example [6] applied maximun coverage model to determine optimal charging station placement strategies that could maximize infrastructure accessibility within urban transportation networks. Likewise, [7] used spatial model to see the influence of spatial distribution of EV charging demand, but this model did not helps in dynamic demand forecasting.

After recognizing the limitation of sole spatial optimization approaches, many recent studies have explored data-driven methods for estimating charging demand. A study by [8] proposed a demand estimation framework using urban informatics techniques. A graph theory and the PageRank algorithm were used to estimate spatial attractiveness for Ev charging demand on their study. This method was highly relied on predefined spatial indicators and regression mapping, which create some limitation to capture complex nonlinear relationship between environment variable and infrastructure demand.

In recent time, many research has also explored machine learning approaches for deman forecasting. [9] developed a data-driven framework using supervised learning amd deep learning to forecast short and medium term EV charging demand. Also, [10] proposed a machine learning model that included charging station data, infrastructure accessibility indicator, and demographics variable using models like Random Forest and XGBoost achieving a good performance. The research by [1] also highlights the importance of using infrastructure data, transportation patterns, and socioeconomic indicators for more accurate forecast of charging demand.

Based on above insights, this project tries to develop a ML based framework for predicting EV charging demand at the census tract level using demographic, built environment features, accessibility, and spatial data. This project also integrates demand prediction with infrastructure gap analysis by using well established market data. This study will try to find underserved and overserved area to make better infrastructure planning.

IV. Methodology

The current research used an observational data science methodology to evaluate and forecast the need for electric vehicle (EV) charging infrastructure at the U.S. census tract level. Data from many public sources were compiled, and then through feature engineering and machine learning modeling were used in order to compare predicted demand with actual infrastructure, and to reveal spatial gaps.

This workflow is separated into five different stages - collecting data, pre-processing and integrating data spatially, feature engineering, modeling, and conducting a gap analysis.

The four major data source types that were used included the American Community Survey (ACS) 5-year estimates, the Alternative Fuels Data Center (AFDC) charging station dataset, the EPA Smart Location Database, and ZIP Code Business Patterns (ZBP). One limitation of these datasets is the difference in spatial resolution, which is addressed through spatial processing and tract-level aggregation.

All data processing and modeling were implemented using Python. Pandas and NumPy were used for manipulating data; GeoPandas and Shapely were used for geospatial processing; and visualization was accomplished with Matplotlib and Plotly. The machine-learning algorithms XGBoost and LightGBM, which have advantages in handling structured tabular data as well as the ability to learn nonlinear relationships between variables, were used in the development of the machine-learning models[10].

For the preprocessing of data, datasets were harmonized based on census tract GEOID identifiers. The locations of charging stations were then assigned to the appropriate census tract using spatial joins based on latitude and longitude, and data that businesses reported on a ZIP code level were assigned to census tracts through area-weighted spatial allocation. For feature-engineering purposes, log transformations, percentiles, and composite indices (e.g., destination intensity) were used.

To estimate how many charging ports will be required at the census tract level, we will be using a modeling approach using XGBoost (a Gradient Boosting algorithm). The resulting model performance will be assessed using Mean Absolute Error (MAE) and R^2 metrics. In addition, SHAP will be used to interpret feature importance and to understand how the model is making predictions. Lastly, we will conduct a gap analysis to compare the estimated charging demand to the available charging infrastructure, thus highlighting data sparse areas as well as areas with excess charging infrastructure. As a result, we will have developed a reproducible pipeline that can go from raw data to useful information for the planning of EV infrastructure.

V. Data Analysis

The data analysis process consists of data collection, preparation, exploratory data analysis (EDA), visualization, and reporting to understand the spatial distribution of EV charging infrastructure and the factors that are influencing charging demand.

A. Data Collection

The data collected for this study is from different U.S governmental API and websites. Firstly, for the demographics and socioeconomic indicators, API provided by U.S Census Bureau was used. The data was from American Community Survey (ACS) 5-Year Estimates. To fetch data from API, the main issue was that each columns was represented by a specific field code, like population was indicated by field "B01003_001E", so all required field were identified using guide provided at census website. Features like population, household, income, vechicles_total, commute_time and education were obtained including commercial establishment data from ZIP Code Business Patterns (ZBP) dataset.

For the present charging Infrastructure data U.S. Department of Energy's Alternative Fuels Data Center's (AFDC) API was used. Detailed information was found including location represented by latitude and longitude, charger type, number of ports for each charger type.

Finally, features like walkability indices, employment density, activity density, intersection density, transit accessibility were derived from the Environmental Protection Agency's Smart Location Database.

B. Data Preprocessing and Spatial Integration

Since, the collected data were from different source and spatial unit, many processing steps were performed to integrate them all. The basic overview of data preprocessing steps is represented by below diagram.

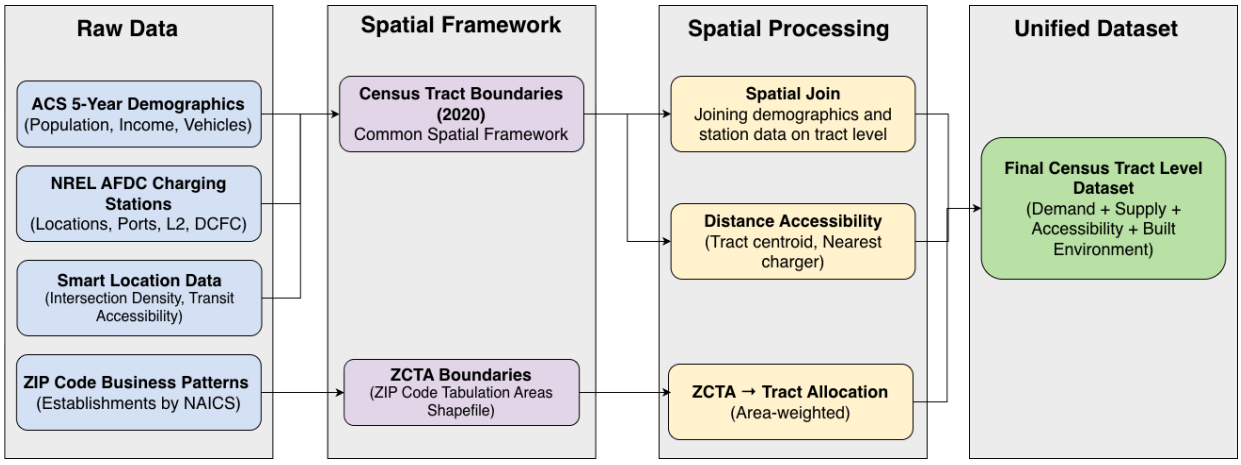

Fig. 1. Overview of data collection, spatial framework, and preprocessing pipeline used to construct the unified dataset.

The demographics data from the U.S census API and the built-environment data from Smart Location Database were already at census tract level where each tract is identified by an 11-digit GEOID, so they were merged with spatial census tract boundaries data.

However, the EV charging station data from AFDC dataset just provided a location with latitude and longitude coordinates. So, to map charging station to tract level, the coordinates were converted into spatial point geometrics and joined with census tract boundary shapefiles and by aggregating on tract level total number of level 2 and DC fast charging ports were calculated.

The ZIP code business patterns dataset provides data on ZIP code level instead of census tracts. To integrate this dataset ZIP Code Tabulation Area (ZCTA) boundary shapefile were first obtained and the ACTA polygons were overlaid with census tract polygons to identify the overlapping areas between them. An area-weighted allocation method was used to distribute business count from ZCTA to census tract level.

In addition, a distance accessibility feature was calculated to measure accessible charging stations from each census tract. For that, the centroid of census tract polygon was calculated, and the distance from centroid to the nearest charging station was computed.

After completing these steps, all the datasets were merged using census tract GEOID and a final unified dataset was obtained containing demographic variables, built-environment indicators, accessibility features, and charging infrastructure supply.

C. Exploratory Data Analysis (EDA)

To understand the distribution of features, correlation with target variables and to see the relation with different variables, EDA was conducted.

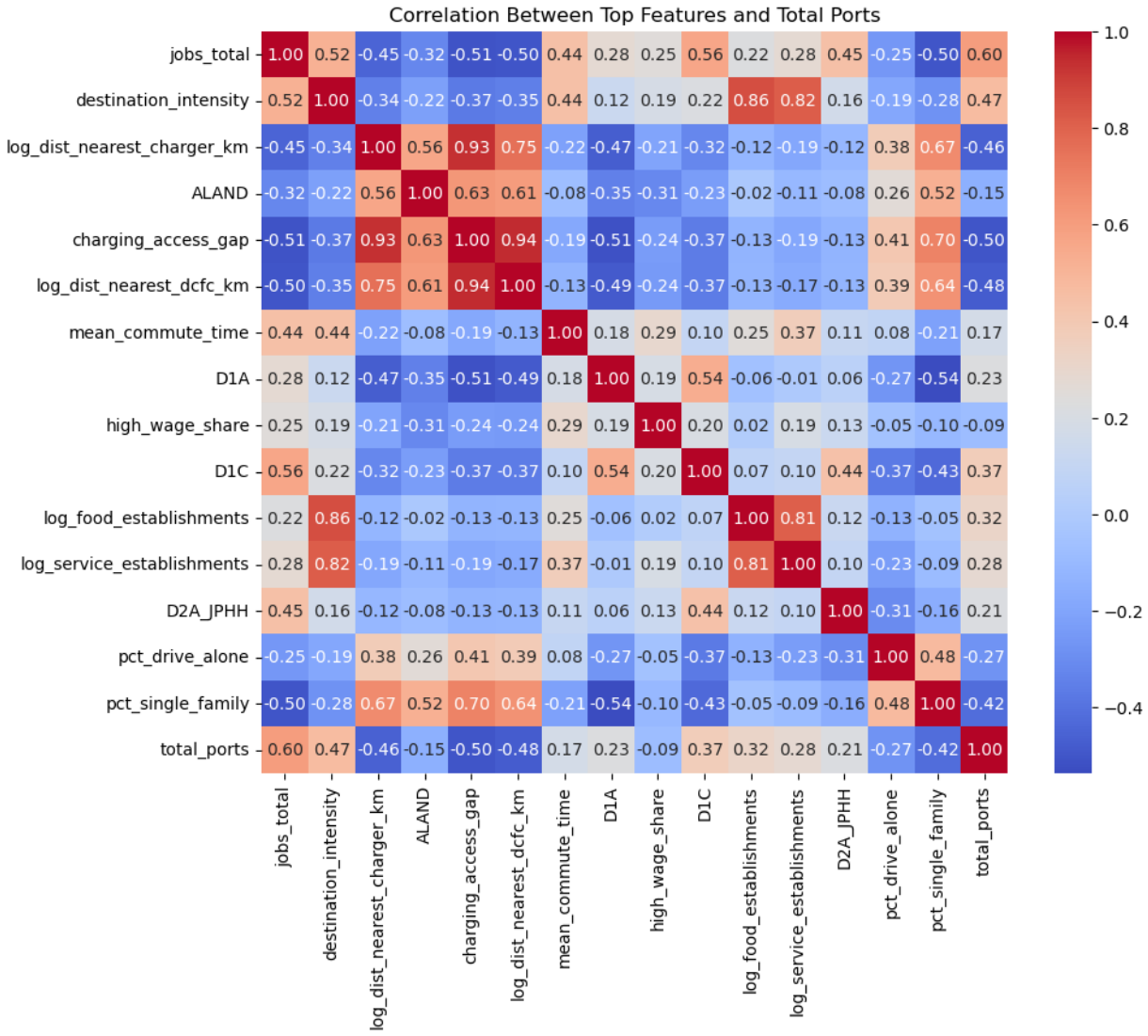

Fig. 2. Correlation Heatmap

From above Heatmat diagram, we can see that total jobs has the strongest positive correlation with total charging ports which is 0.6, which suggest that tract that have more employments tends to have more charging infrastructure. Also, feature like destination intensity index and D1C also show moderate positive relationships. Also, we can see the negative correlation liwh distance nearest charger which tries to indicate that location farther from existing chargers tend to have fewer chargiing ports. Hence, employment density, commercial activity and accessibility to existing charger are major influencing factors.

D. Model Performance and Results

This seection provides the outcomes of our model and discuss the findings of our model. The analysis focuses on model performance, feature importance, and the identification of gaps across census tracts.

1) Model Performance: In this project, two tree-based ensemble models XGBoost and LightGBM were tested intially. Both model provides almost same accuracy and works well with nonlinear feature. But, after hyperparameter tuning and feature selection, the performance of XGBoost was more accurate.

Feature Importance analysis shows that many features were actual reducing the accuracy instead of helping to get better result. So, the final XGBoost model trained with reduced feature set produced best results and selected for final gap analysis.

Model Performance was evaluated using standard regression metrics like Mean Absolute Error (MAE) and the coefficient of determination (R^2).

Model	Features Used	MAE	R^2
LightGBM	All features	3.9	0.40
XGBoost	All features	4.01	0.37
XGBoost	Reduced features	3.7	0.44

TABLE I
MODEL PERFORMANCE COMPARISON

2) Model Result: The final XGBoost model was then used to predict the number of charging ports for Vermont State, which predict for almost 15 percentage more to the present number of charging ports. For a 1370 existing total number of port, model predict the 1563 total predicted ports creating a gap of around 193 ports.

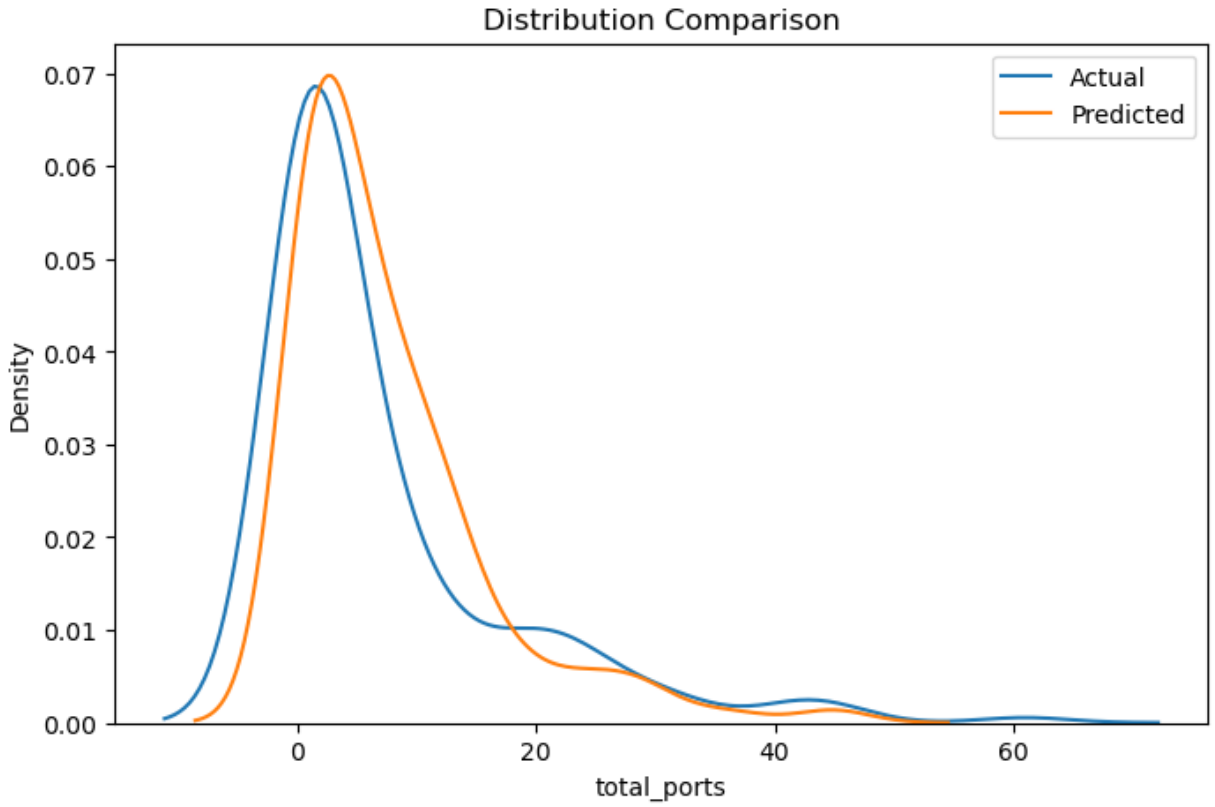

Fig. 3. Actual and Predicted ports Comparision for Vermont State

Also, to interpret the prediction of XGBoost model, Shapley Additive ecPlanations (SHAP) was used to see the contrivution of each feature to the output. The summary plot clearly illustrate the importance of features as well as the direction of their influence on the result.

As per the plot below ALAND (tract land area) is the most influential feature followed by distance to nearest charger and total jobs on that tract. ALso, Commercial activity indicators like food and service establishments also show positive contribution to model output.

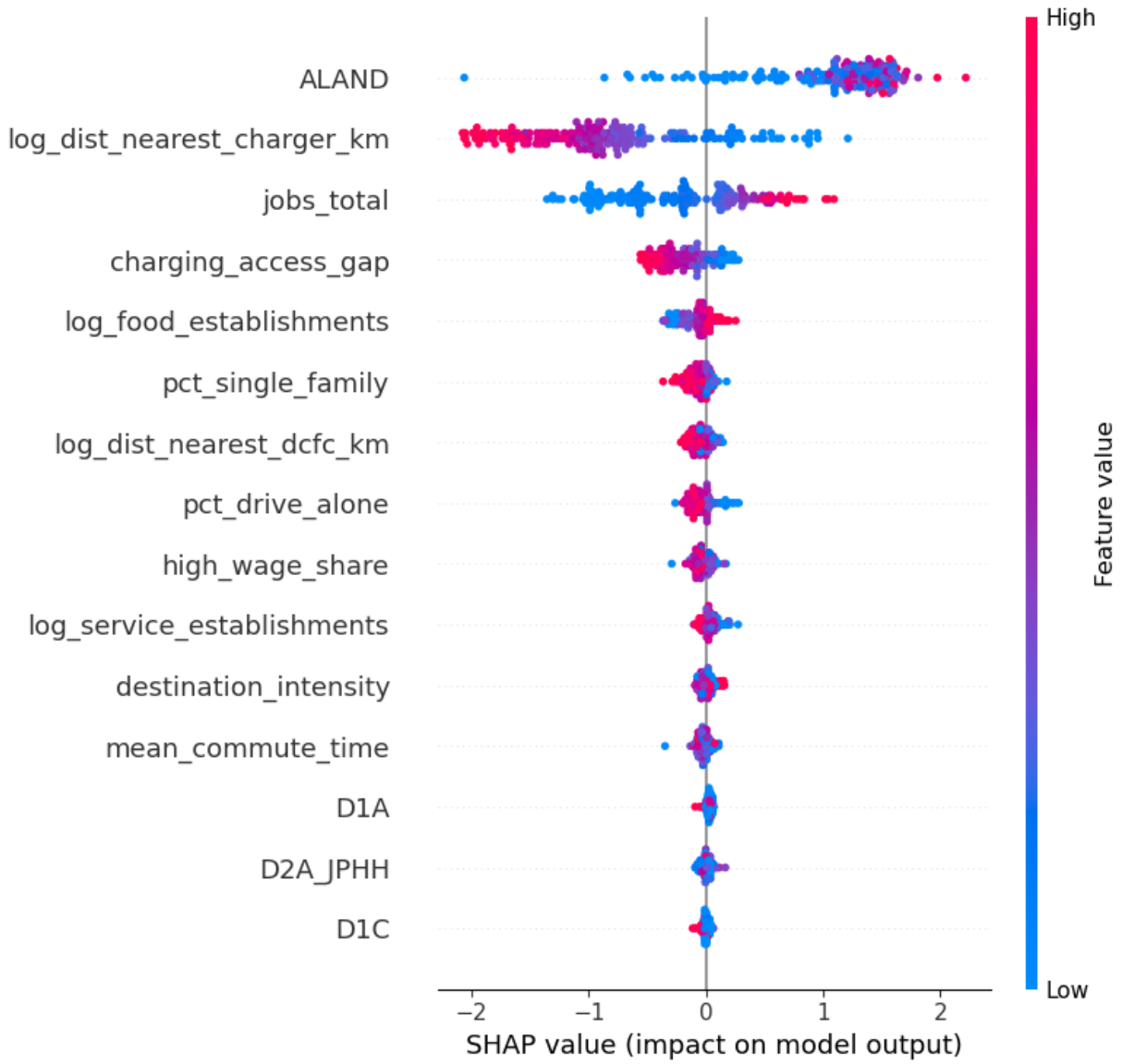

Fig. 4. SHAP Analysis for Vermont State

VI. Key Findings

The result of this projects shows that demographics, spatial, and built-environment variables can effectively estimate charging infrastructure demand at census tract level. And, among different model tested, XGBoost regression model gives best performance after feature selection outperforming the LightGBM model in terms of prediction accuracy. Also, the SHAP analysis have revealed that economic activity and spatial data like land area, distance to nearest charger, employment and commercial density were most influential features used in model. Finally, after the analysis of predict outcomes we can see that there are more balance and overserved tracts in urban area, and many suburban area are underserved.

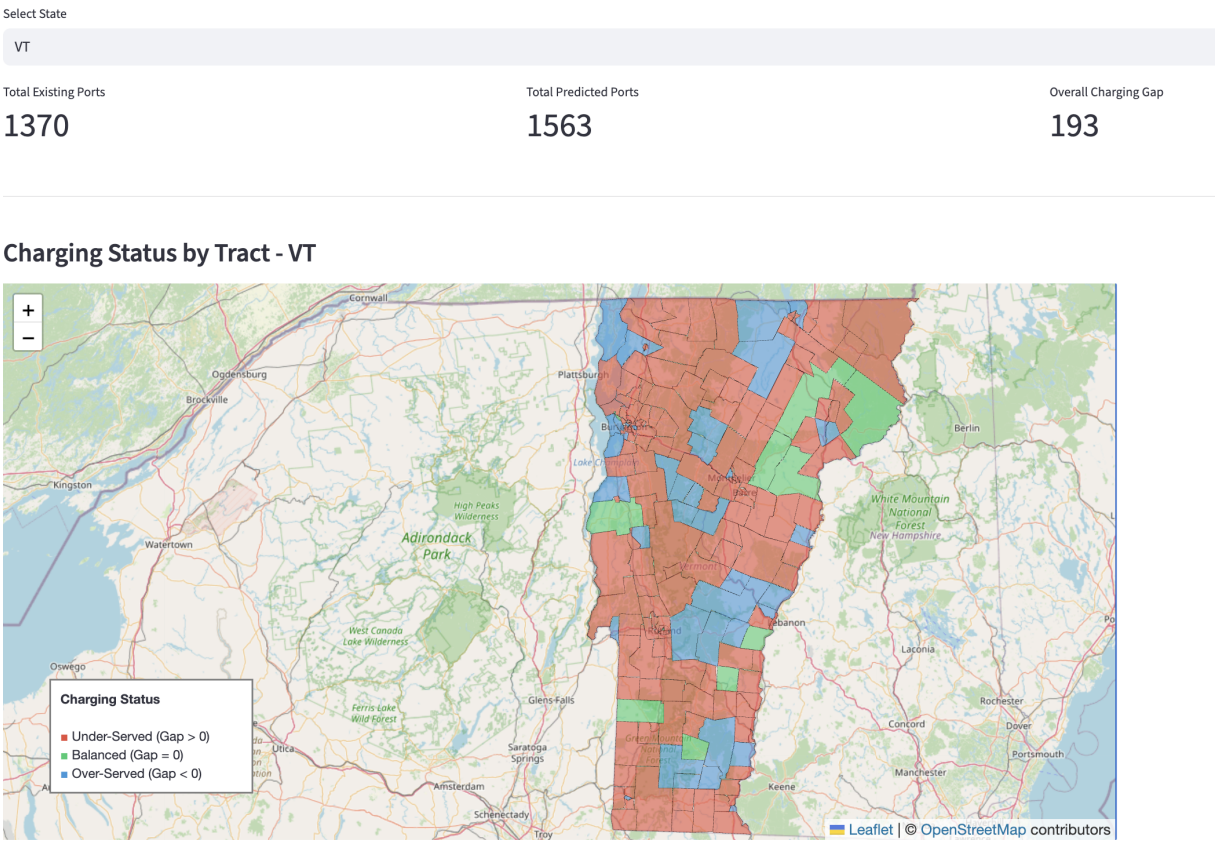

Fig. 5. EV charging infrastructure gap analysis showing underserved, balanced, and overserved census tracts, Vermont State

VII. Project Timeline

The time frame for completing this project is 8 weeks according to the MSDS schedule of practicum activities shown in Table II below.

TABLE II
EIGHT-WEEK PROJECT TIMELINE

Week	Phase	Key Activities and Deliverables
Weeks 1–2	Scoping	Finalizing of the project topic; Identifying project objectives, and confirming datasets and tools to be used in the project.
Week 3	Data Acquisition	Collecting EV charging station data, census demographic data, and traffic volume data using APIs.
Week 4	Data Preprocessing	Data cleaning, spatial joining of Census Tracts to traffic volume and EV charging station datasets, and feature engineering based on existing relationships between the three data sets.
Week 5	Model Development	Training of an XGBoost regression model with benchmark states' data.
Week 6	Model Optimization	Hyperparameter tuning and model explainability analysis using SHAP.
Week 7	Gap Analysis	Apply the trained model to emerging regions and compute EV charging demand gaps.
Week 8	Final Reporting	It will include the development of the Final Report, visualizations, and presentation materials for the presentation of the project results.

VIII. CONCLUSION

As electric vehicle adoption continues to grow, it will be increasingly difficult to understand how electric vehicle charging infrastructure needs to be planned. To aid future planning, this practicum project created a machine-learning-based framework to predict demand at the census tract level using demographic, spatial, and built-environment data about where people have lived and worked and where electric vehicles are most likely to be charged. The project also identified potential infrastructure gaps for the predicted demand. Ultimately the practicum work contributes to the areas of data science and electric vehicle infrastructure planning, by integrating multiple publicly available datasets, predictive modeling, and model interpretability via SHAP analysis to examine what may contribute to EV charging demand. However, the knowledge gained through this practicum project is not limited to technical results; it will help me to continued work in the field.

REFERENCES

[1] N. Al-Dahabreh, M. A. Sayed, K. Sarieddine, M. Elhattab, M. J. Khabbaz, R. F. Atallah, and C. Assi, "A data-driven framework for improving public ev charging infrastructure: Modeling and forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5935–5948, 2023.

[2] M. Hossain, M. M. K. Rabbi, N. Akter, N. N. Rimi, M. H. H. Amjad, M. H. Ridoy, and M. S. S. Shovon, "Predicting the adoption of clean energy vehicles: A machine learning-based market analysis," *Journal of Ecohumanism*, vol. 4, no. 4, pp. 404–426, 2025.

[3] X. H. Ren, W. Wu, and Z. Chen, "Integrating dynamic demand forecasting and static factor analysis for urban ev charging infrastructure: A two-stage spatio-temporal deep learning approach," Chongqing Jiaotong University and The Ohio State University, Tech. Rep., 2025, working paper, 55 pages.

[4] H. A. U. Khan, S. Price, C. Avraam, and Y. Dvorkin, "Inequitable access to electric vehicle charging infrastructure," 2021. [Online]. Available: https://arxiv.org/abs/2111.05437

[5] J. Antoun, M. E. Kabir, R. F. Atallah, and C. Assi, "A data-driven performance analysis approach for enhancing the quality of service of public charging stations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11 116–11 125, 2021.

[6] J. Dong, C. Liu, and Z. Lin, "Charging infrastructure planning for promoting battery electric vehicles: An activity-based approach," *Transportation Research Part C: Emerging Technologies*, vol. 38, pp. 44–55, 2014.

[7] J. Kang, H. Kong, Z. Lin, and A. Dang, "Mapping the dynamics of electric vehicle charging demand within beijing's spatial structure," *Sustainable Cities and Society*, vol. 76, p. 103507, 2022.

[8] Z. Yi *et al.*, "Electric vehicle demand estimation and charging station allocation using urban informatics," *Transportation Research Part D: Transport and Environment*, vol. 106, 2022.

[9] A. Orzechowski *et al.*, "A data-driven framework for medium-term electric vehicle charging demand forecasting," *Energy and AI*, vol. 14, 2023.

[10] F. R. Anonna, B. R. Chowdhury, and M. H. Ridoy, "Machine learning enabled analysis of on-the-road ev charging infrastructure: Predicting accessibility and optimizing deployment," *Journal of Computer Science and Technology Studies*, 2025.