Fall 2022 DSI Capstone Project
Final Report

EV Charging stations in Washington State

Dec 17, 2022

Student Team Members

Anqi Lin al4215
Clarissa Tai rt2822
Mengchen Xu mx2257
Yue Zhang yz4155
Yu-Chieh Chen yc4015

Industry Mentors

Alexander Voet, Data Scientist, Associate at KPMG Lighthouse
Arvind Sathi, Director at KPMG Lighthouse
Chengwei Wang, D&A Data Scientist, Associate at KPMG Lighthouse
Jeffrey Lee, Lighthouse D&A Modeler, Associate at KPMG
John Heath, Director at KPMG
Jordan Jalving, Sr Associate at KPMG
Laura Uguccioni, D&A Data Scientist, Manager at KPMG Lighthouse
Michelle Zee, D&A Consultant, Sr. Associate at KPMG Lighthouse
Sydney (Bolim) Son, D&A Data Modeler, Associate at KPMG Lighthouse

1. Project Background and Problem Definition

In November 2021, President Biden officially signed a $5 billion investment in state-administered grants for nationwide EV charging stations. These funds will help states develop charging networks across rural, disadvantaged, and hard-to-reach areas communities.[1] People living in rural areas are deterred from owning an EV car because of the lack of EV chargers, especially when they plan to travel far. The Chicken-and-Egg Conundrum exists here is that drivers choose not to own an EV being afraid of not enough power, while lack of enough EV owners will not support the idea of establishing an EV charging station.[2] Some states have already taken actions to plan ahead for planning, prioritization, and implementation of a statewide network of charging stations along state highways[3].

In this project, we will tackle the challenge of how to best electrify our roads and communities to encourage balanced growth. We defined our research question as: what are the most important factors to consider when choosing EV charger locations in Washington state and where are potential optimal locations. To solve the problem, we will firstly use machine learning methods to predict the number of charging stations by each census tract area, and use optimization models to choose locations from candidate gas stations and optimize the number of charging stations and chargers along major routes in WA. With the optimized locations, we will create a scoring system suggesting the recommendation level to build a station depending on features that we wanted to focus on.

2. Exploratory Data Analysis

2.1 Why Washington State?

To start from small, we choose a specific state or region to focus our efforts on. We chose WA as our target for a couple of reasons. Firstly, WA has enough data to build models and at the same time potential to put additional stations. The comparatively small number of EV stations shows the potential and the comparatively large percentage in registered EV and EV stations ensures enough data to build models as shown in Fig. 2.1.1. Secondly, WA has an EV Deployment Goal that all light-duty vehicles sold, purchased, or registered in WA must all be EVs by model year 2030[4]. This indicates a large potential market for EVs along with its supporting infrastructure. Also, in Fig. 2.1.2 and 2.1.3, we observe that EV registration density is not the same as EV charging station density, which suggests the necessity for us to make optimization on EV charging station locations.

[1] EU Support Grows for Russia Oil Ban Over Ukraine War - WSJ
[2] Biden Has a $5 Billion Plan to Eliminate America's EV Charging Deserts
[3] Washington State Plan for Electric Vehicle Infrastructure Deployment
[4] Electric Vehicle (EV) Deployment Goal

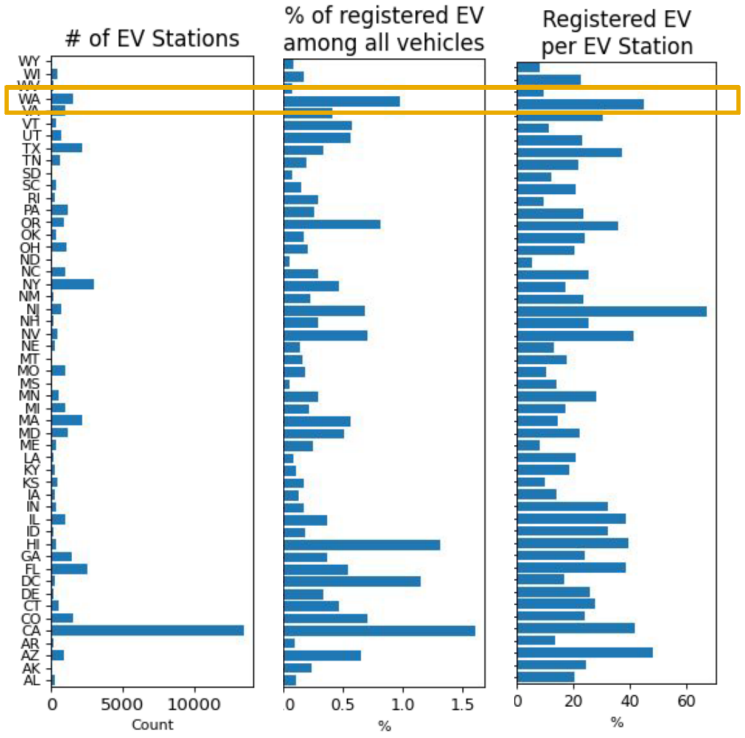


Figure 2.1.1 EV demographics

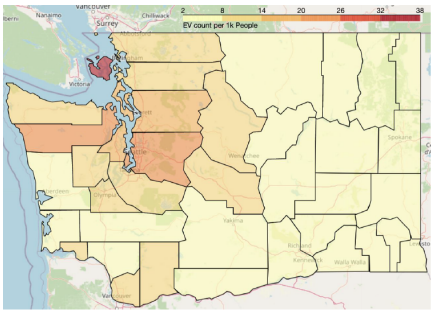


Figure 2.1.2 EV registration density (# of EVs per 1k people by county)

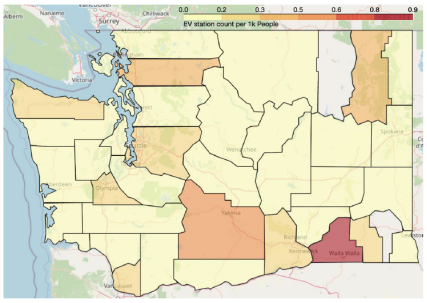


Figure 2.1.3 EV station density # of EV stations per 1k people by county)

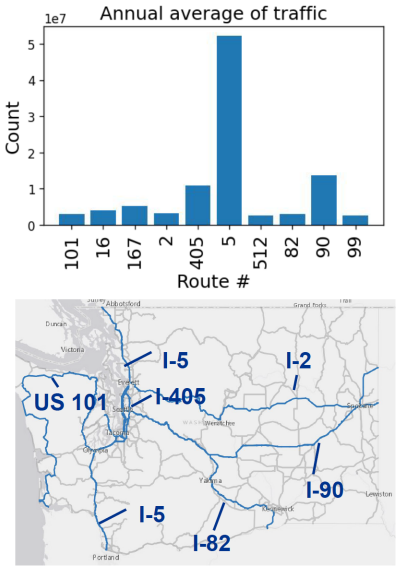


Figure 2.2.1 Annual Average Daily Traffic Distribution

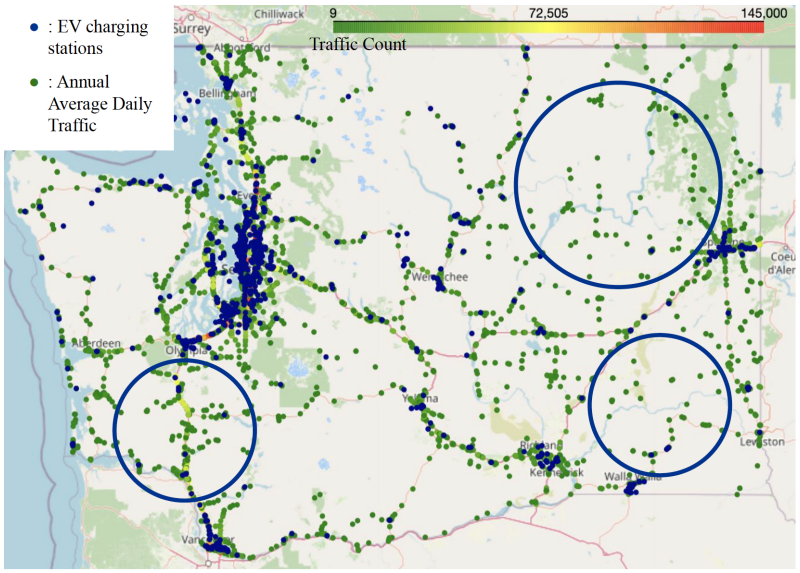


Figure 2.2.2 Distribution of traffic with EV stations

2.2 Traffic

Traffic count is an important factor to consider when determining the location to make sure the EV stations are placed around traffic accessible areas. A total of 188 highway routes exist in WA in 2021 and major traffic is happening along interstate highways and state routes in the Seattle area as shown in Fig. 2.2.1. Overlaying EV station locations with the traffic, we realized that WA

is lacking EV chargers in rural areas within comparatively large traffic access in 3 major areas (Fig. 2.2.2).

2.3 Natural Risk Index

Risk of natural disaster is an important aspect when designing EV station locations. Exposure to natural disasters will directly lead to possible electric shocks and outages which further lead to severe damage to the infrastructure.Sustainability and environmental vulnerability of the charging stations take a big part when achieving business goals. From Figure 2.3, coastal suburban areas have higher risk scores, due to lower social vulnerability and community resilience to recover from the disaster.

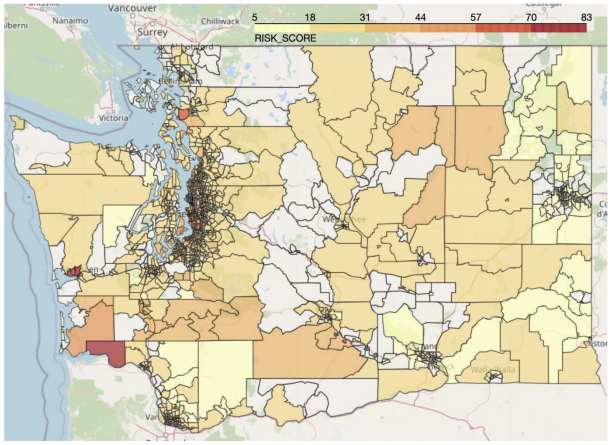


Figure 2.3 Natural Risk Index Score

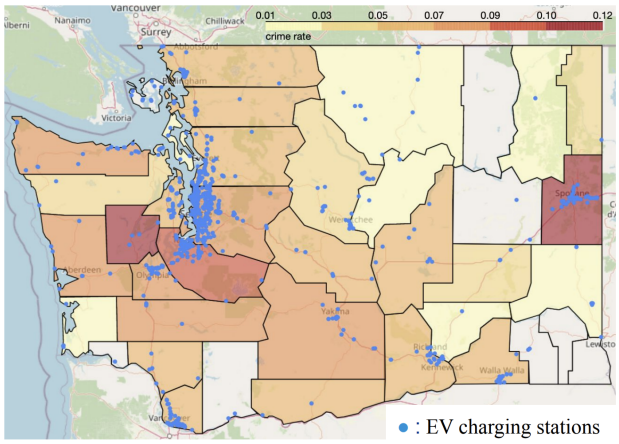


Figure 2.4 Crime Rate on County

2.4 Crime rate

Crime rate in one specific area is also an important aspect to be taken into account. Out of safety concerns, EV car owners may be unwilling to leave their car for a couple of hours to charge in dangerous areas (especially high motor theft, robbery, etc.), leaving the already built EV stations useless. Therefore, to take the crime situation into account, we collected crime count data from the FBI website and calculated the crime rate adjusted by population, which indicates the dangerousness of the neighborhood in each city. From Figure 2.4, most current EV charging stations are not located in high crime rate areas, except some crowded charging stations in Spokane. In addition, assumed ideal places for EV chargers discussed in previous slides are mostly located in low crime areas.

2.5 Tourists Attraction

Tourists account for a huge amount of EV station users. In order to determine the optimal locations for EV stations, we need to consider the places where tourists prefer to visit, and these places may provide helpful insights on potential areas that may need EV stations. From the visualization (Figure 2.5), we can see that the blue circled area includes some tourism locations without current EV charging stations along major routes (mainly I-5) with higher traffic

identified previously. Therefore, we can focus more on these areas, which may be potential baseline areas that need to build EV stations.

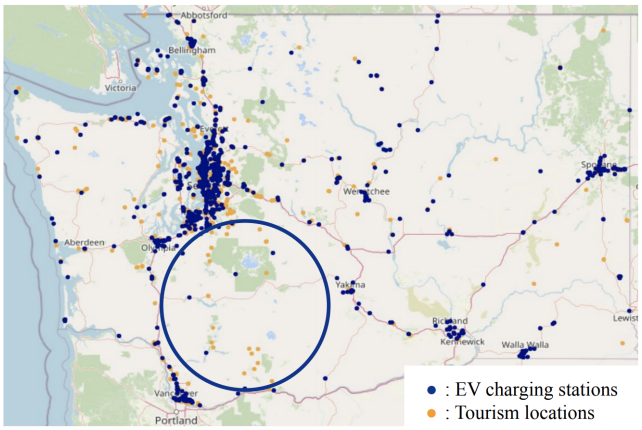


Figure 2.5 Tourist Attraction Locations in WA

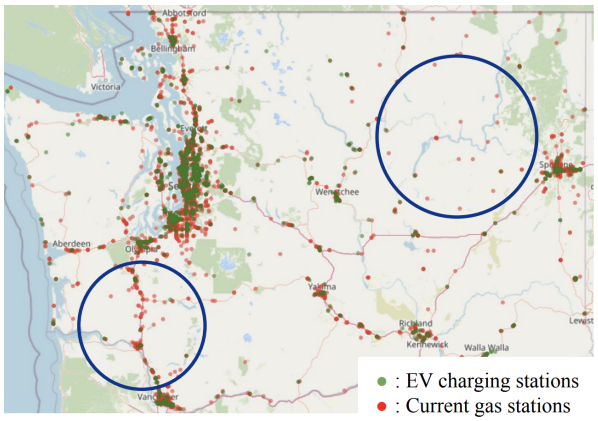


Figure 2.6 EV vs. gas stations distribution

2.6 Gas station distribution

From the above section, it shows the power plants' locations but there is no public information on where electricity is delivered from each power plant. We are unable to determine the availability of electricity. Since gas chargers and convenience stores in gas stations are powered by electricity, the locations of gas stations guarantee the electricity accessibility. Therefore, an analysis of gas stations is crucial. From the figure 2.6, we can see most current EV charging stations overlap with gas stations in cities but not in rural areas. Some rural areas still lack charging stations. When proposing new EV charging stations, we can assume that the presence of gas stations indicates the electricity can be delivered.

3. Data Preprocessing

3.1 Route Distance

To prepare the features needed for modeling, we consider setting gas station locations in WA as our base references to determine the optimal locations for EV stations and EV chargers. We are curious about the number of EV stations, highway exits, crime counts, traffic counts, tourist attraction places, and the risk of natural disaster around gas stations, which might be useful features to help us determine the optimal EV station locations. However, it is inexplicable to find these numbers in a circle area with gas station locations as centers and direct distance as radius.

People who need to find EV stations are drivers who can only access gas stations through existing routes. Therefore, route distance is more reasonable to be considered in this situation.

The method we come up with to figure out the route distances between locations is called MapQuest API. MapQuest API can provide route distance between two locations, and it also can perform radius search by setting center locations and route distances as radius.

3.2 Highway

Since the NEVI Formula Program (main federal funding source) requires that EV charging infrastructure projects installed with funding must be located along a designated alternative fuel corridor[5], we consider the distance to highways is an important factor. Instead of using direct distance from location to highway, we chose to use the route/driving distance from a location to its nearest highway exit. This gives us the best estimate of how much a driver needs to drive to access the EV charging station. We used MapQuest's Corridor Search API to filter the gas stations within a 5 mile driving distance from the target highway. Then we used MapQuest's Direction API to obtain the route distance from each gas station to its nearest highway exit.

3.3 Missing Value

Features related to crime and NRI contain missing values due to lack of record. To fill the missing values in NRI, we take the average since it is impossible to have no risk of natural disaster and NRI are preprocessed also using mean values. There is no crime recorded in some cities on the FBI website, so for gas stations in those cities, we have missing crime data. Considering that the missing crime data may be related to the lack of large population in some cities so small amounts of crimes weren't recorded and apparently it does not indicate no crimes happening in that city, we decided to fill in missing values in crime data with minimum count.

The full data schema used for modeling can be found in Appendix.

4. Modeling

4.1. EV Station Count Prediction Model

To tackle the problem of how many EV charging stations to put in each area, we experimented with a machine learning model to predict the number of EV stations in each census tract area by assuming that those in fully-developed areas/cities are optimal. We trained models on census tracts with non-zero number of EV stations or is labeled as cities. The model is then used to predict the number of EV stations in those without EV stations, mostly rural. The predictions will be treated as the reference number for optimization and hope to leverage further discussion. Here, we try on several regression models and select XGBoost as our final model.

[5] National Electric Vehicle Infrastructure Formula Program

4.1.1 Data Wrangling

For classification models, predictions were by each gas station. However, for this problem, we need to reform the data we have. To prepare the dataset, entries were aggregated by census tract index. New features such as average of daily traffic, maximum daily traffic, count of attractions in the area were created.

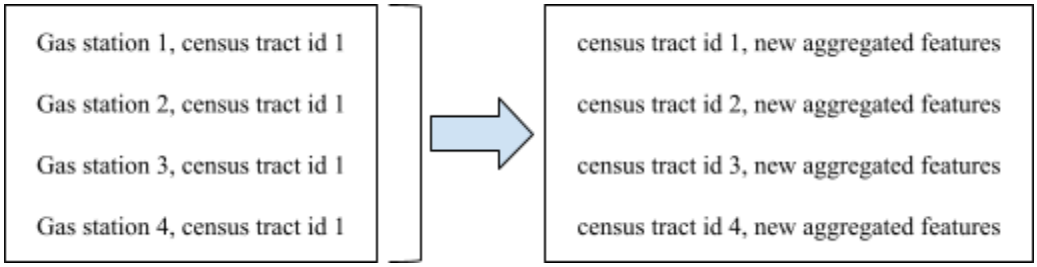


Figure 4.1.1 Generalization specific data process.

4.1.2 Feature Selection

Initially, all features are included in the dataset. However, as shown in Fig. 4.1.2, many crime-related features are highly correlated. We hence decided to only include the total crime population and also filtered other highly-correlated features. It is worth mentioning that even if some features are generated from the same data source, such as maximum daily traffic counts and average daily traffic counts, they performed low correlation, inferring that they retained different information.

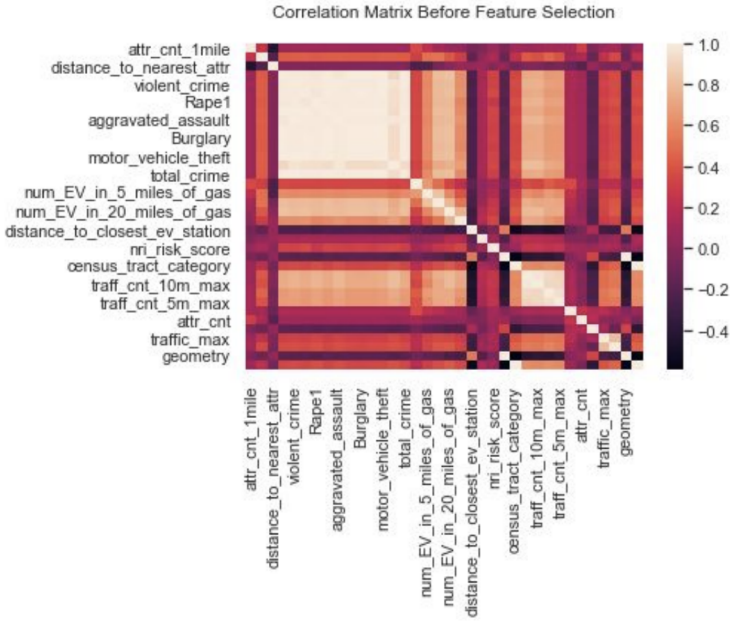


Figure 4.1.2 Correlation matrix before Feature selection

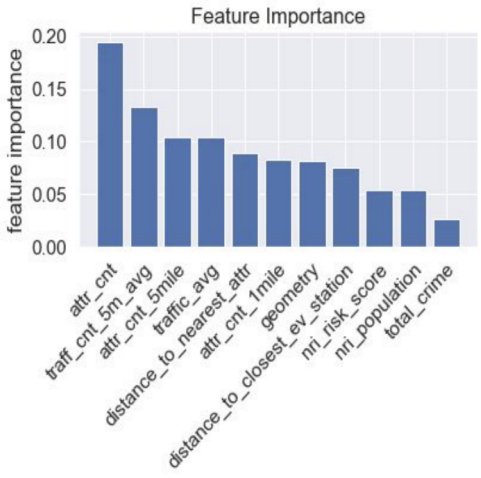


Figure 4.1.3 Feature importance of the selected model

4.1.3 Model Development & Evaluation

To predict the number of EV stations in each census tract area, we compared 4 regression models, two linear models, and two tree-based models: Linear regression, Gamma Regressor, Random Forest Regressor, and XGBoost Regressor. The Gamma Regressor was chosen because the distribution passed the Kolmogorov-Smirnov test, indicating that it follows the gamma distribution. The hyperparameters are selected through randomized grid search and 5-fold cross-validated. The results are shown in Table 2.1. Based on the results, XGBoost Regressor was selected as the final model and proceeded to predict EV station counts in each census tract.

Regressor	Linear	Gamma	Random Forest	XGBoost
Best hyperparameters	N/A	max_iter: 200, alpha: 0	n_estimators: 500, max_depth: 3	subsample: 0.6, min_child_weight: 1, max_depth: 5, learning_rate: 0.01, gamma: 1.5, colsample_bytree: 1.0
training MSE	9.56	11.07	5.84	1.15
validation MSE	6.49	8.72	3.31	0.94
training r-square	0.05	-0.10	0.42	0.89
validation r-square	0.20	-0.07	0.59	0.88

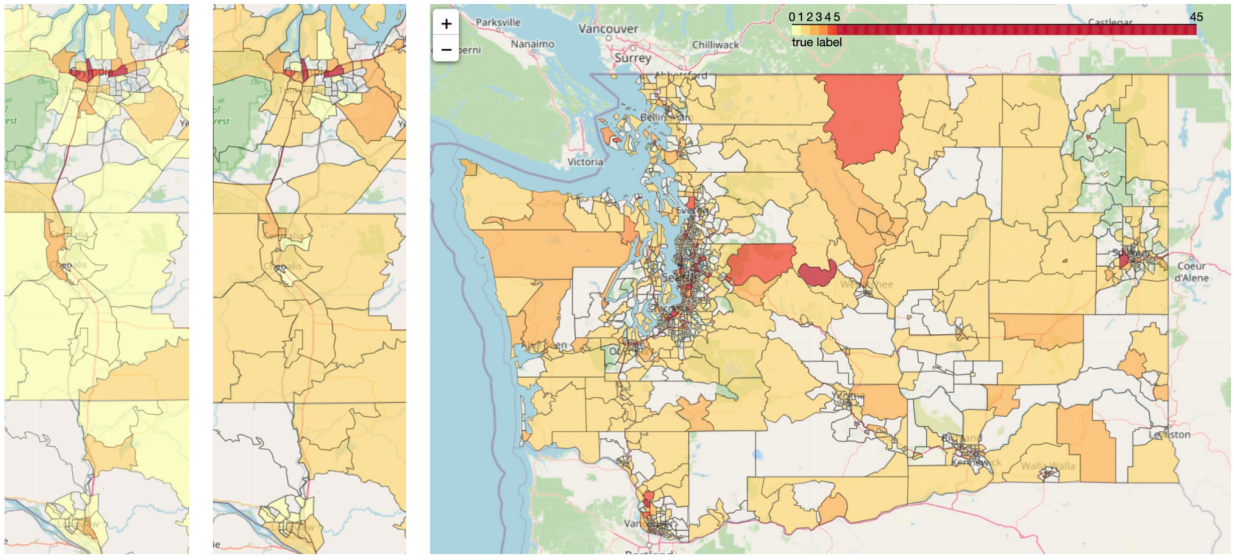
Table 4.1 Model Evaluation Results.

4.1.4 Prediction Results

The feature importance is shown in Fig. 4.1.3. Number of attractions in the specific census tract area is the most important feature, followed by the average number of daily traffic counts within a 5 mile radius.

Previously, we observed that most EV charging stations are in the cities, where the census tract area is smaller geographically and densely populated, or along highways leading to national parks and famous attractions. The imbalance between traffic demand and supply of the critical interstate highway draws our attention. For example, the current distribution of EV charging stations along the southern section of I-5 highway is shown in the left first plot of Fig. 4.1.4, and the prediction results are shown on its right side for better comparison. Advanced discussions about the findings will be proceeded in Section 3: Optimization. In this section, we will briefly show the overall patterns of the model predictions on Washington state and along southern I-5 highway. The final prediction is shown in the right first plot of Fig. 4.1.4. The figure also included the predicted results of the training subset to leverage insights. The empty census tracts

in the figures are those lacking gas stations, as we were not able to generate the features for these areas.



Current state Predicted Model prediction on whole map

Figure 4.1.4 Model Prediction, included the predictions on training data.

4.2. EV Station Location Optimization Model

After knowing the predicted number of new EV charging stations in each census tract area, we need to decide exactly where to put new EV charging stations in an optimal way minimizing the station and charger number. We decided that our constraints on this problem will be:

1. The driving distance from highway exits to each locations should be less than 10 miles (to make sure the newly suggested EV stations are near the highways)
2. Newly added EV stations can support all EV traffic demand
3. Charger count at each location should not exceed 15[6] (to make sure the charger count is reasonable at each location)

4.2.1 Model Formulation

Due to the limitation of our traffic count data which only includes all types of vehicles along highways, to calculate the EV traffic demand, we decided to multiply the total traffic count by the ratio of registered EV among all vehicles which is 1.6%[7].

To calculate the supported EV traffic by each charger, we made the assumption that, in busy Seattle city areas, the supply of EV stations is already sufficient for covering all EV traffic, so we can calculate a ratio of supported EV traffic per charger. The ratio is calculated as follows: for

[6] U.S. average EV charger per station reference:
https://www.iea.org/reports/global-ev-outlook-2022/trends-in-charging-infrastructure
[7] Reference: https://afdc.energy.gov/transatlas/#/?state=WA&view=per_capita&fuel=GA

each exit on I-5 in the Seattle city area, we count the maximum annual average daily traffic count and the number of EV stationswithin 2 miles. The ratio of supported traffic per charger is traffic count within 2 miles divided by the number of EV stations within 2 miles. Then we take the average of each ratio on exits and come to the final ratio. To approximate the EV traffic, we multiply the ratio by 1.6%. The final estimated supported EV traffic per charger is 168.

The following are the variables, objective function and constraints for our optimization model.

Variables	Description
x_i	Number of new chargers to place in each candidate gas station location
d_i	Driving distance from highway exits to each candidate gas station
y_i	Existing EV charger count within 5 miles on each candidate location
e_i	Maximum EV traffic count within 5 miles on each candidate location

The number of chargers in the area is $\sum x_i$, and we want to minimize it as the objective function.

To satisfy the constraints we mentioned previously, we firstly need to make the driving distance from highway exits to each locations less than 10 miles, which means $d_i < 10$.

According to the demand of EV traffics, our total existing number of chargers and newly proposed number of chargers should cover all the EV traffics going through the area, which gives the expression $(x_i + y_i + x_{j\in\{other\ proposed\ charger\ cnt\ within\ 5\ miles\ of\ i\}}) \times ratio \geq e_i$.

Also, the number of chargers should be an integer no more than 15 and not negative.

Combining all the results above, we formulate the mathematical model as:

$$minimize \quad \sum x_i,$$

$$subject\ to \quad d_i < 10$$

$$(x_i + y_i + x_{j\in\{other\ proposed\ charger\ cnt\ within\ 5\ miles\ of\ i\}}) \times ratio \geq e_i$$

$$x_i \in Z,\ x_i \leq 15,\ x_i \geq 0$$

4.2.2 Model Result

To start from a smaller area and generalize it to the whole map, we initially focused our model on the I-5 south area, which is a typical area with comparatively more traffic and less EV charging stations.

Based on the consideration proposed by KPMG that, if we create more EV stations along the route, there will be an increased EV traffic percentage. Therefore, we ran our model on three

scenarios based on different EV% of all traffic: EV account for 1.6% of total traffic (current state), EV account for 3.0% of total traffic (projected % in 2026[8]) and EV account for 5.0% of total traffic (projected % in 2030).

Using the Mixed-Integer Linear Program Solver on Gurobi, we get the result in Fig. 4.2.1.

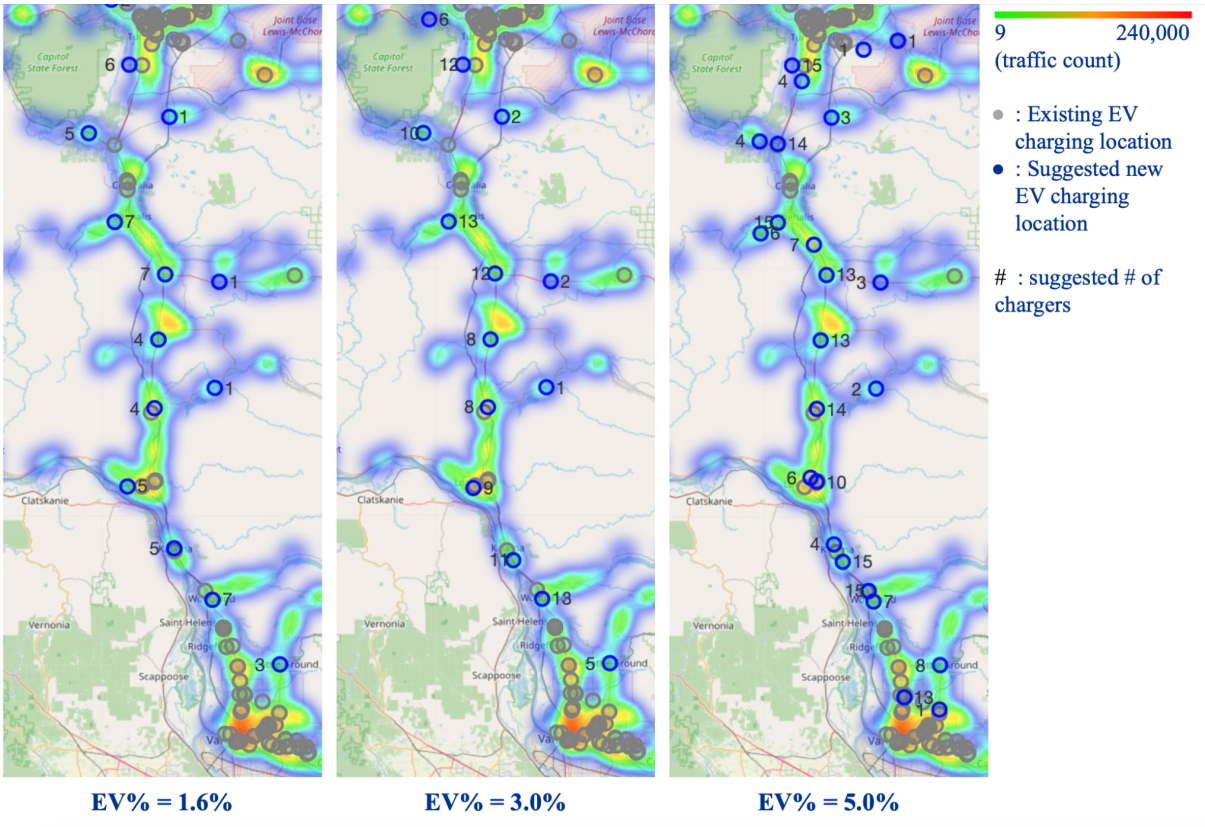


Figure 4.2.1 Optimization model result on different EV% scenarios

Model-suggested new EV charging station locations are evenly scattered around comparatively rural areas along areas where higher traffic is presented. In addition, in areas with existing EV charging locations, our model doesn't suggest a lot of newly built EV charging stations, which are performing as we expected. In addition, as the EV traffic percentage increases, the proposed number of chargers increases, and more locations appear near the city area.

Applying the current model on the entire I-5, I-90, and I-84 route on different EV% scenarios, the result is shown in Fig. 4.2.2.

[8] Quadratic-interpolated from WA registered vehicles data 2016 - 2021, includes EV and plug-in hybrid vehicles.

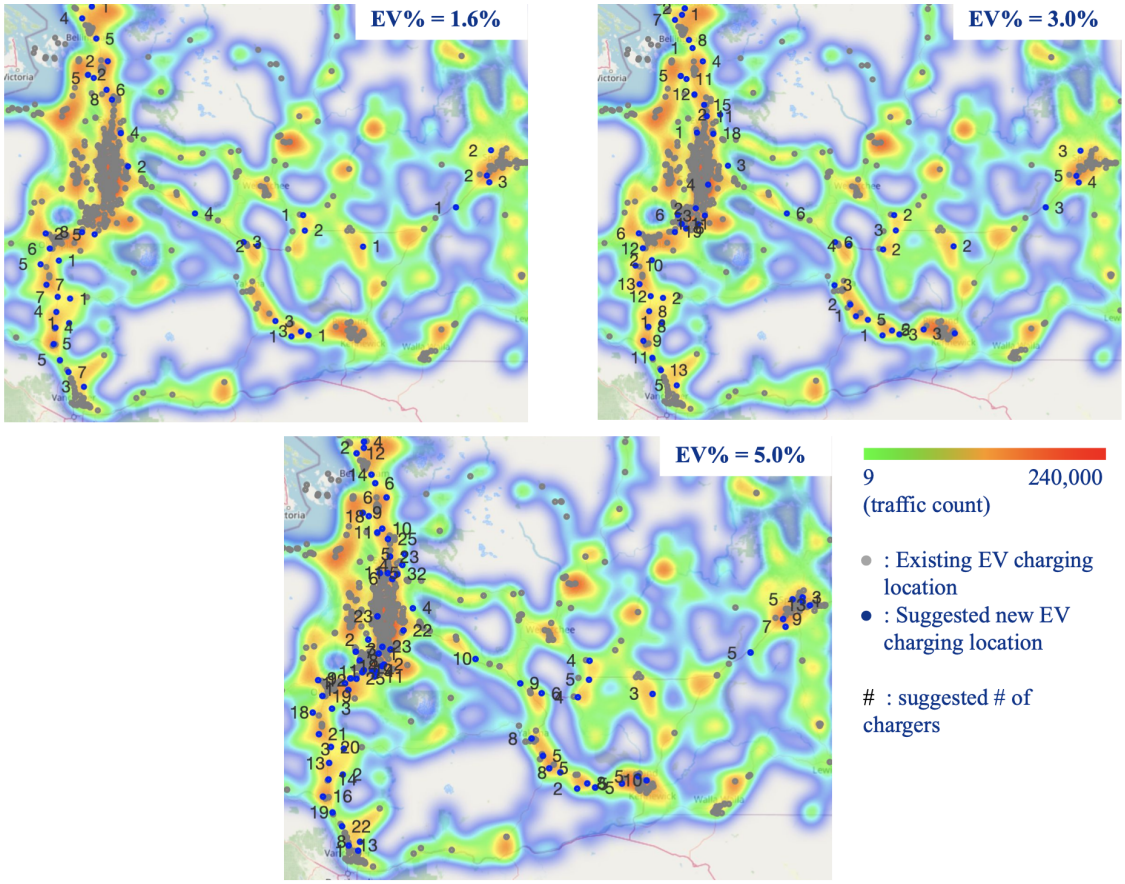


Figure 4.2.2. Model result on I-5, I-90 and I-82

4.2.3 Alternative Model

After building our model with an objective function of minimizing the number of chargers, which minimizes the cost of building new EV charging stations and chargers, we considered another way to optimize based on different objective functions that may lead to different locations with constraints remaining the same. Our alternative objective function is to minimize total distance to nearest exists from all chargers $\sum d_i * x_i$. The reason to select this objective function is that there are extra costs for each EV car driving from exits to the gas station location, which can be an extra penalty on the cost for more far-away stations.

On the I-5 south area, based on three scenarios with different EV% of all traffic: EV account for 1.6% of total traffic (current state), EV account for 3.0% of total traffic (projected % in 2026) and EV account for 5.0% of total traffic (projected % in 2030), our alternative model result is shown in Fig. 4.2.3.

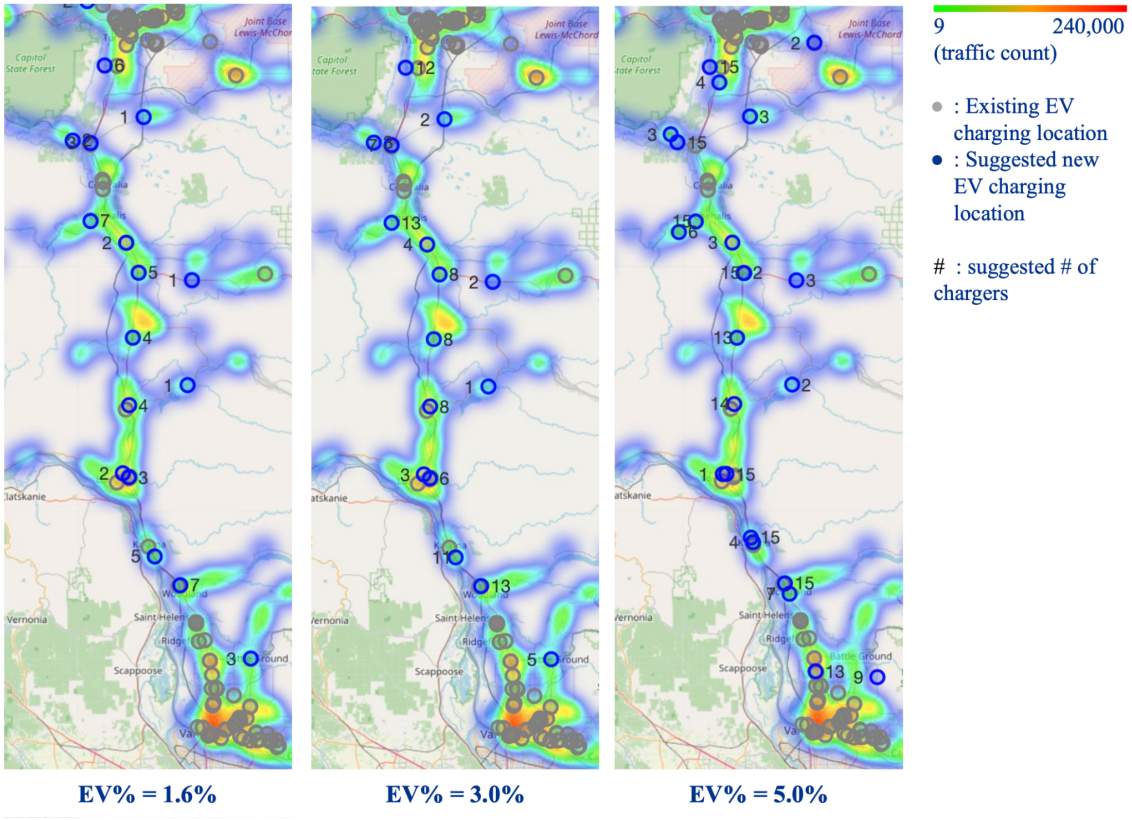


Figure 4.2.3 Alternative model result on different EV% scenarios

Compared to the original model, our alternative model minimizing total distance to nearest exits from all stations gives similar optimized locations with some small differences.

4.2.4 Comparison Result

Considering that there are extra costs on each car driving from exits to the gas station location, we decided that our evaluation metric will be the total cost of building stations, adding chargers, and the penalty of the cost of each EV car driving from exits to the stations.

After researching online, we estimate the rough cost for building an EV charging station in WA as $30,000, the cost for adding an EV charger in a station in WA as $8,000[9], and the cost for an EV per charger driving a mile as $0.15[10]. Therefore, the total cost will be calculated as

$$30000 \times \sum 1_{\{if\ x_i>0\}} + 8000 \times \sum x_i + 0.15 \times ratio \times e_i \times d_i.$$

After deciding the evaluation metrics, we calculated our cost on both models. The result for I-5 south region is as follows:

[9] Charger cost reference: https://afdc.energy.gov/files/u/publication/evse_cost_report_2015.pdf
[10] Driving cost reference: https://ecocostsavings.com/electric-car-cost-per-mile/

EV%	Model	Total suggested locations	Cost	Cost Difference
1.6%	Original: minimizing charger count	13	$785,086	~$90,000
	Alternative: minimizing distance	16	$874,278	
3.0%	Original: minimizing charger count	13	$1,288,670	~$60,000
	Alternative: minimizing distance	16	$1,348,591	
5.0%	Original: minimizing charger count	23	$2,063,685	~29,000
	Alternative: minimizing distance	24	$2,092,414	

Although the cost on alternative models is always higher than the cost on the original model, this is caused by the fact that the original model produces less charging stations, which accounts for a larger portion of the total cost. When the EV traffic count increases, the penalty of the cost of each EV car driving from exits to stations plays a bigger role, causing the cost difference to decrease. Therefore, by adjusting the cost coefficients, there will be a point where our alternative model outperforms our original model. However, under the current situation, we can conclude that the model minimizing the charger count works better regarding the cost.

5. Proposed Location Scoring Model

We decided to propose a score for all the optimized EV station locations to give insights on the different importance levels of each optimized EV station location. Based on our expectation, the higher the score, the more we recommend building an EV station at the location.

5.1 Intuitive Base Model

We used 4 features which are not used during the optimization process in this model: number of attractions, total crime, median household income and natural risk score. To calculate the final score, we firstly give each feature the same weight (weight = 1), and scale the range of each column of the data (one column is one feature) from 0 to 1 by applying Min-Max Scaler. Then we sum up all the numbers to get a final score for each EV station. Finally, applying Min-Max Scaler again on the score, we can make them range from 0 to 1 and, if multiplying it by 100, we can get a final score ranging from 0 to 100.

5.2 Logistic Regression

Since our intuitive model is based on our expectation rather than the reality, another way to calculate the score is to use the Machine Learning models, specifically classification models, to learn from existing patterns and give the predicted score. Here our target is whether to label the gas station as an EV station (if an EV station appears nearby within 20 meters).

We first fit a Logistic Regression model and predict the probability of labeling each proposed location as EV station locations. Multiplying the predicted probability by 100, we get the recommendation score of that proposed EV station location.

After we have dropped all highly correlated features (>0.85) in the Logistic Regression Model, the model result is shown in Figure 5.1. The model shows a large false positive rate, which will cause the probability score to be a bit higher since the model tends to predict positive results. The feature with the highest importance is motor vehicle theft, which is counter-intuitive based on our expectation that we are more inclined to put EV stations in lower crime rate areas. We explored the possible reason behind this counter-intuitive model result and found that motor vehicle theft rates are higher in places with more EV stations based on current EV station location patterns. Logistic regression learned this pattern and embedded this pattern in the training model.

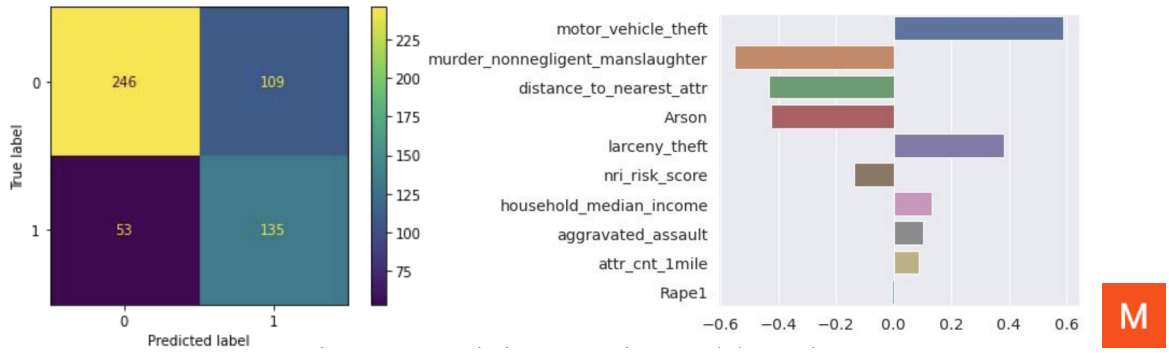


Figure 5.1 Logistic Regression Model Result

5.3 Random Forest Classification

Another model we implemented is the Random Forest Classification Model with GridSearch Cross Validation. Similarly, we predict the probability of labeling each proposed location as EV station locations and multiplying it by 100 to come with the score. The model performance is shown in Fig. 5.2.

The probability of predicting the proposed EV locations as optimized locations are not high overall in the Random Forest model. However, the relative probabilities among all the proposed locations have some overlap with the result obtained from the logistic regression model. The random forest model is more reasonable compared to the logistic regression model since the top 4 important features are reasonable to have high influences on choosing optimized EV charging locations.

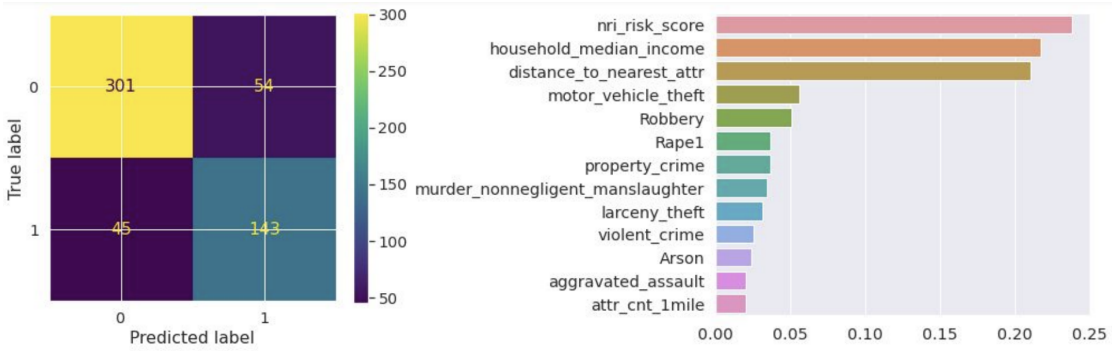
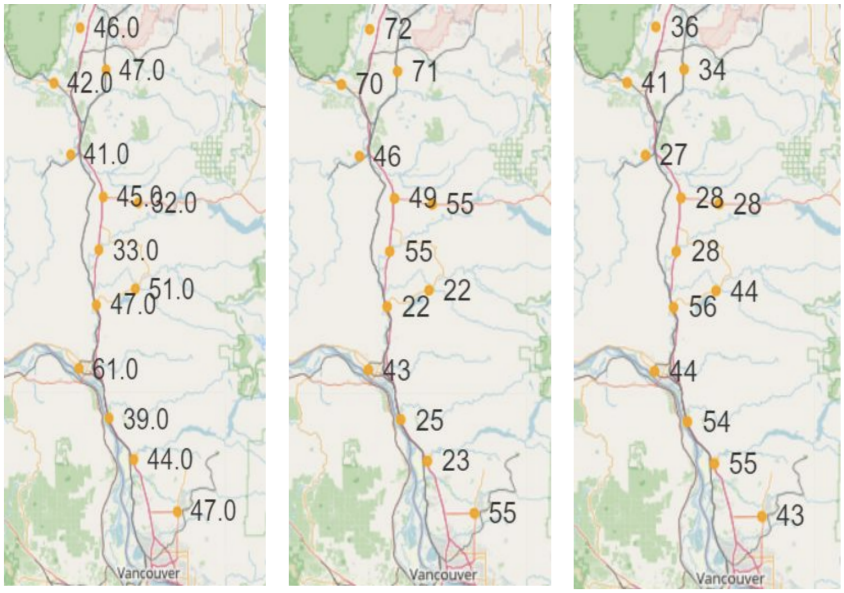


Figure 5.2 Random Forest Model Result

5.4 Model Comparison

The predicted scores of different models are shown in Figure 5.3.



Intuitive Base Model | Logistic Regression | Random Forest

Figure 5.3 Model Score Result Comparison

Comparing the scores, there are some differences between the intuitive base model with the ML models. In the intuitive base model, the scores are mostly similar along the road with only 2 or 3 locations with extreme scores, but in Logistic Regression, locations in the north are having much higher scores compared to locations in the south. Also, locations in the middle from Random Forest results are having much lower scores compared to locations in north and south. Based on the accuracy from the table below, we would prefer Random Forest over Logistic Regression because of the better performance.

Model	Features Used In Model [11]	Best Estimators	Precision	Recall	F1 Score	Accuracy
1. Intuitive Base Model	N4 N17 N18 N19	N/A	N/A	N/A	N/A	N/A
2. Logistic Regression	N4 N6 N8 N9 N11 N14 N15 N16 N18 N19	N/A	0.55	0.72	0.63	0.70
3. Random Forest	N4 N6 N7 N8 N9 N10 N11 N12 N14 N15 N16 N18 N19	'max_depth'= 50, 'max_features'= 'log2', 'n_estimators'= 150	0.73	0.76	0.74	0.82

The intuitive base model and the Random Forest model give two perspectives to choose the score. Depending on the necessity from the user, they can choose to refer to one of the model results. If the user wants to know the ideal or expected score, he/she can refer to the intuitive base model score, and if the user wants to know the score that is based on the existing pattern, he/she can refer to the Random Forest model score.

6. Conclusion, Future Works and Ethical considerations

Our model can flexibly propose optimized EV charging station locations based on different objectives, different EV traffic level scenarios, and give realistic scores for proposed locations based on multiple factors, and it is fairly generic and can be applied to other regions if respective data is available.

For the optimization model, we can make future improvements such as choosing candidate locations from arbitrary latitude and longitude (not necessarily gas stations), adding more constraints (electricity capacity, budget, geographics etc), or considering extra construction cost and land cost which may vary from place to place. For the scoring system, future improvements can be systematically choosing weights, using adjacency to police stations to evaluate safety level (except crime rate), or including electricity capacity if we have access to e-grid datasets. In addition, there are no ethical considerations to be paid attention to in this project.

[11] The feature names refers to the Data Features Form in Appendix

7. Contributions

- Anqi Lin: Team captain. Gathered, cleaned and preprocessed traffic and crime data, and responsible for exploratory data analysis on those two dataset. Main contributor for creating and improving the optimization model, and forming the business case question: how to score each optimized EV station.
- Clarissa Tai: Gathered, cleaned, preprocessed, and visualized natural risk index data. Responsible for exploratory data analysis on census tract level data, and aggregate feature and dataset for generalization. Main contributor for creating and improving EV Station Count Prediction Model; researching supplementary documents and datasets to support optimization model assumptions.
- Mengchen Xu: Gathered, cleaned and preprocessed commute and attraction data. Responsible for exploratory data analysis on commute and attraction data. Main contributor on MapQuest API for route distance. Improved the optimization model, and formed the business case question. Main contributor for scoring model and comparison.
- Yue Zhang: Gathered, cleaned and preprocessed government and exit data. Responsible for exploratory data analysis on government funding data. Main contributor on MapQuest search API, creating and improving the optimization model, and forming the business case question: how to score each optimized EV station.
- Yu-Chieh Chen: Gathered, cleaned and preprocessed gas station and e-grid data. Responsible for exploratory data analysis on gas station and e-grid data, as well as creating and maintaining merged main dataset. Main contributor for highway exit distance calculations. Researched supplementary documents and datasets to support optimization model assumptions.

8. Appendix - Data Source and Schema

8.1 Data Sources

- Washington State Plan for Electric Vehicle Infrastructure Deployment, July 2022 https://wsdot.wa.gov/construction-planning/statewide-plans/washington-state-plan-electric-vehicle-infrastructure-deployment
- U.S. Energy Information Administration. "Electricity." https://www.eia.gov/electricity/. Accessed 5 October 2022.
- Energy Efficiency & Renewable Energy. "Alternative Fuel Data Center." https://afdc.energy.gov/fuels/electricity_locations.html#/analyze?fuel=ELEC. Accessed 5 October 2022.
- National Risk Index https://hazards.fema.gov/nri/data-resources. Accessed 5 October 2022.
- Annual Average Traffic Count https://gisdata-wsdot.opendata.arcgis.com/datasets/WSDOT::wsdot-traffic-counts-aadt-1/explore?location=47.271387%2C-119.745108%2C6.90 Accessed 11 October 2022.
- Crime https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/tables/table-8/table-8-state-cuts/washington.xls Accessed 5 October 2022.
- MAPQUEST https://developer.mapquest.com/documentation Accessed 15 October 2022.
- Washington Geospatial Open Data Portal https://geo.wa.gov/datasets/WSDOT::wsdot-interstate-exit-numbers-1/api Accessed 12 October 2022.
- Tourist Attraction https://mygeodata.cloud/data/download/osm/tourist-attractions/united-states-of-america--washington Accessed 5 October 2022.

8.2 Dataset Schema

Column Name	Data Type	Description
gas_key	Numerical	Gas station key
gas_name	Categorical	Gas station name
gas_lat	Numerical	Gas station location latitude
gas_long	Numerical	Gas station location longitude
attr_cnt_1mile	Numerical	Number of tourism attractions within 1 mile distance
attr_cnt_5mile	Numerical	Number of tourism attractions within 5 mile distance

distance_to_nearest_attr	Numerical	Distance to the nearest tourists attraction
crime_coord	Numerical	Crime data location coordinate: (longitude, latitude)
crime_county	Categorical	Crime happened county label
total_crime	Numerical	Number of summed crimes (theft, burglary, murder, etc.)
highway	Categorical	Name of Highway, for example “I5”
distance_to_nearest_exit	Numerical	Distance to the nearest highway exit
num_EV_in_2_miles_of_gas	Numerical	Number of EV stations within 2 mile distance
num_EV_in_20_miles_of_gas	Numerical	Number of EV charging stations within 20 mile distance
distance_to_closest_ev_station	Numerical	Distance to the nearest EV charging station
nri_geoid	Categorical	Geographic identifiers that uniquely identify all administrative geographic area
nri_county	Categorical	County names that corresponds to the Geographic identifiers in the previous column, for example “King” as “King County”
nri_risk_score	Numerical	Relative risk of natural hazards at a location, higher value means higher risk, a number in range [0, 100]
nri_risk_rating	Categorical	Relative rating of communities at the same level, categories include “Very High”, “Relatively High”, “Relatively Moderate”,“Very Low” and “Relatively Low”
traff_cnt_5m_max	Numerical	Maximum traffic count within 5 miles
traff_cnt_10m_max	Numerical	Maximum traffic count within 10 miles

7.3 Data Features for Modeling

Column Name	**ID**	**Data Type**	**Description**
gas_key	N1	Numerical	Gas station key
gas_lat	N2	Numerical	Gas station location latitude
gas_long	N3	Numerical	Gas station location longitude
attr_cnt_1mile	N4	Numerical	Number of tourism attractions within 1 mile distance
attr_cnt_5mile	N5	Numerical	Number of tourism attractions within 5 mile distance
distance_to_nearest_attr	N6	Numerical	Distance to the nearest tourists attraction
violent_crime	N7	Numerical	Number of violent crime cases

murder_nonnegligent_manslaughter	N8	Numerical	Number of muder/non-negligent manslaughter cases
Rape1	N9	Numerical	Number of rape cases
Robbery	N10	Numerical	Number of robbery cases
aggravated_assault	N11	Numerical	Number of aggravated assault cases
property_crime	N12	Numerical	Number of property crime cases
Burglary	N13	Numerical	Number of burglary cases
larceny_theft	N14	Numerical	Number of larceny-theft cases
motor_vehicle_theft	N15	Numerical	Number of vehicles stolen cases
Arson	N16	Numerical	Number of arson offenses
total_crime	N17	Numerical	Number of summed crimes (including theft, burglary, murder, etc.)
nri_risk_score	N18	Numerical	Relative risk of natural hazards at a location, higher value means higher risk, a number in range [0,100]
household_median_income	N19	Numerical	Median household income in Zip Code Magnitude
matched	B1	Boolean	True if an EV station is located within 1 mile of the gas station; false otherwise
census_tract_city	B2	Boolean	True if the gas station location is in city; false otherwise
highway	C1	Categorical	Name of highway, for example “I5”