CS4641_121 Project Proposal

Introduction

Literature Review

Electric vehicle (EV) charging presents a unique challenge to predicting demand and costs due to fluctuating electricity prices, usage patterns, and lack of infrastructure compared to gasoline [2]. Studies have shown that the transformer method predicted charging demands most accurately [4]. The transformer method is based on neural networks on sequential data. [1] Another potential option is random forests, which have been used in the past to predict optimal charging times based on supply and demand of electricity [3].

Dataset Description and Link

This dataset provides a comprehensive analysis of EV charging patterns and user behavior when charging at corporate settings, such as the office.

A team led by public policy professor Omar Asensio used a field experiment to collect data on 3,395 electric vehicle charging sessions between November 2014 and October 2015. The dataset “contains sessions from 85 EV drivers with repeat

usage at 105 stations across 25 sites at a workplace charging program"; it indicates the date and length of each session, total energy used, cost, and more.

https://www.kaggle.com/code/meisenbach/electric-vehicle-charging-eda

Problem Defenition

Problem

In the current electric vehicle market, there is a lack of transparency in accurate charging costs

Motivation

With rising carbon emissions/global warming, there has been a push to have car owners switch to EVs; however, due to the complexities of the charging infrastructure, it is hard for a new owner to accurately know how much it would cost them to "refill" their car, This can lead to many people avoiding the switch over to EV, as the fear of unknown cost can make people uncomfortable. Providing consumers with a tool to find the lowest prices would allow for comfortability to make the switch to EVs. This leads to less carbon emission.

Methods

Preprocessing Methods

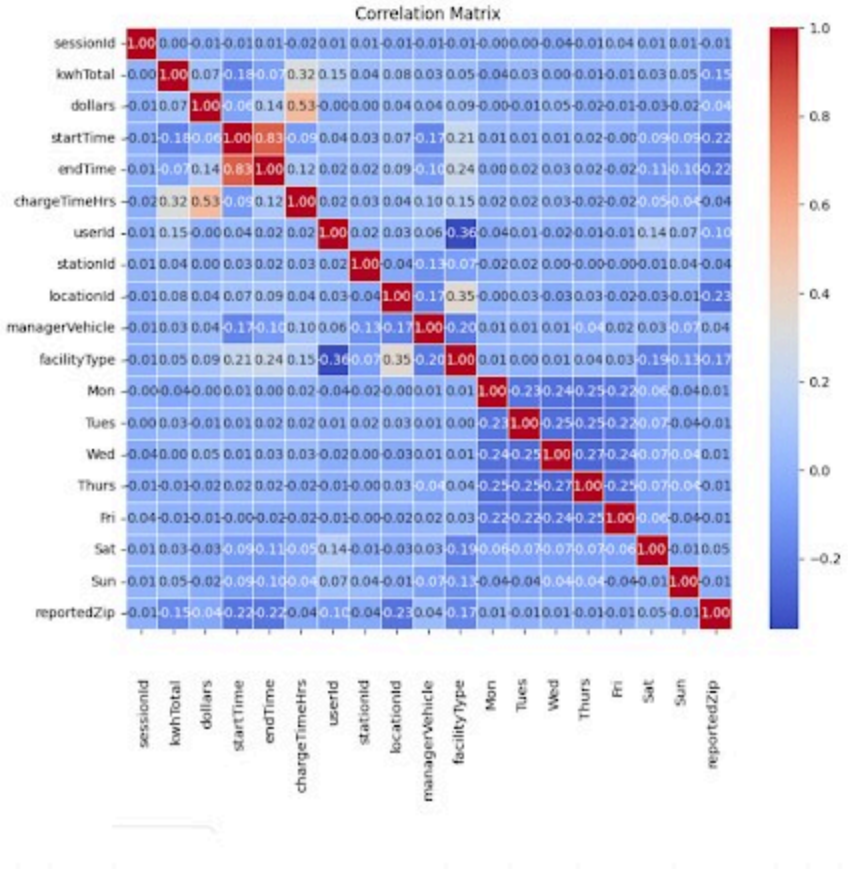
For this project, our team developed a number of preprocessing methods that were applied in order to properly prepare the dataset for modeling.

Data Cleaning - We began preprocessing by cleaning up our data set by removing features that showed a negligible correlation with our main variable, price. We did this because retaining this data would increase model complexity without improving its predictive power. Next we used methods to identify missing values in the data set. Missing or invalid data rows were reviewed and either corrected or removed based on their importance to the dataset.

Feature Transformation - We had many features that did not contain purely quantitative data so we had to encode them. For categorical columns like weekdays, we used one-hot encoding to create an indicator for each variable's unique category. For other categorical columns, like facilityType, we used label encoding to help the model understand the data.

Exploratory Data Analysis - We calculated the correlation matrix to understand relationships between numerical features, and we used descriptive statistics to summarize numeric columns. This helped us find outliers and inconsistencies in the data that we chose to normalize or remove.

Visualization - We used seaborn's heatmap to visualize correlations and identify potential multicollinearity issues. We also used many swarm and scatter plots to find insights into cost patterns. For example, we found that charging is free for the first 4 hours at offices / workplaces.



Data Splitting - After preprocessing, we divided the data into training (80%) and testing (20%) subsets to train and evaluate the machine learning model. This split ensured the model could generalize to unseen data while preventing overfitting.

All of the effort put into preprocessing allowed our team to have a clean dataset to develop our models on.

GLR (Generalized Linear Regression)

A linear regression model will be built to predict charging costs based on factors like charging duration, energy consumed, and time of day. This model is straightforward to interpret and can serve as a baseline for comparison with more complex models like the MLP.

Random Forests

We also chose to implement a Random forest model due to the robust nature of the model as well as its ability to showcase non-linear relationships between the features as well as its ability to show us feature importance and ability to support feature engineering. As you can see below, we are able to generate a correlation matrix for our data, allowing us to understand which features are important with regards to charging cost.

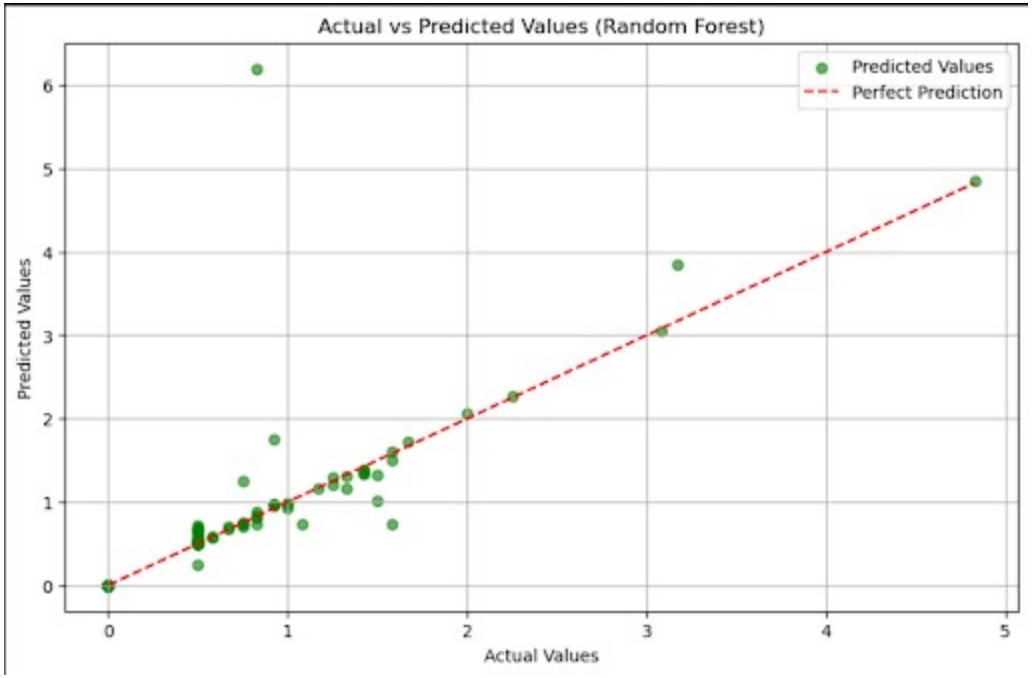
HGBR (Histogram Gradiant Boosting Regressor)

We selected the Histogram Gradient Boosting Regressor (HGBR) because it is particularly efficient in handling large datasets and it is able to process different kinds of data which is crucial because our dataset has many diverse data types. HGBR uses a histogram to sort continuous features into concrete intervals which greatly reduces computational complexity while still keeping the information accurate; this is ideal for sets with many dimensions. HGBR supports quantitative and categorical data which allows it to easily handle encoded data and numerical features. The model is then able to understand cross feature relationships and can better discover more nuanced patterns in the data set. Additionally, HGBR has an iterative boosting process that works to minimize prediction errors while still allowing for regularization so that overfitting does

not occur. Lastly, HGBR also allows us to interpret which factors most strongly influence charging prices by highlighting feature importances.

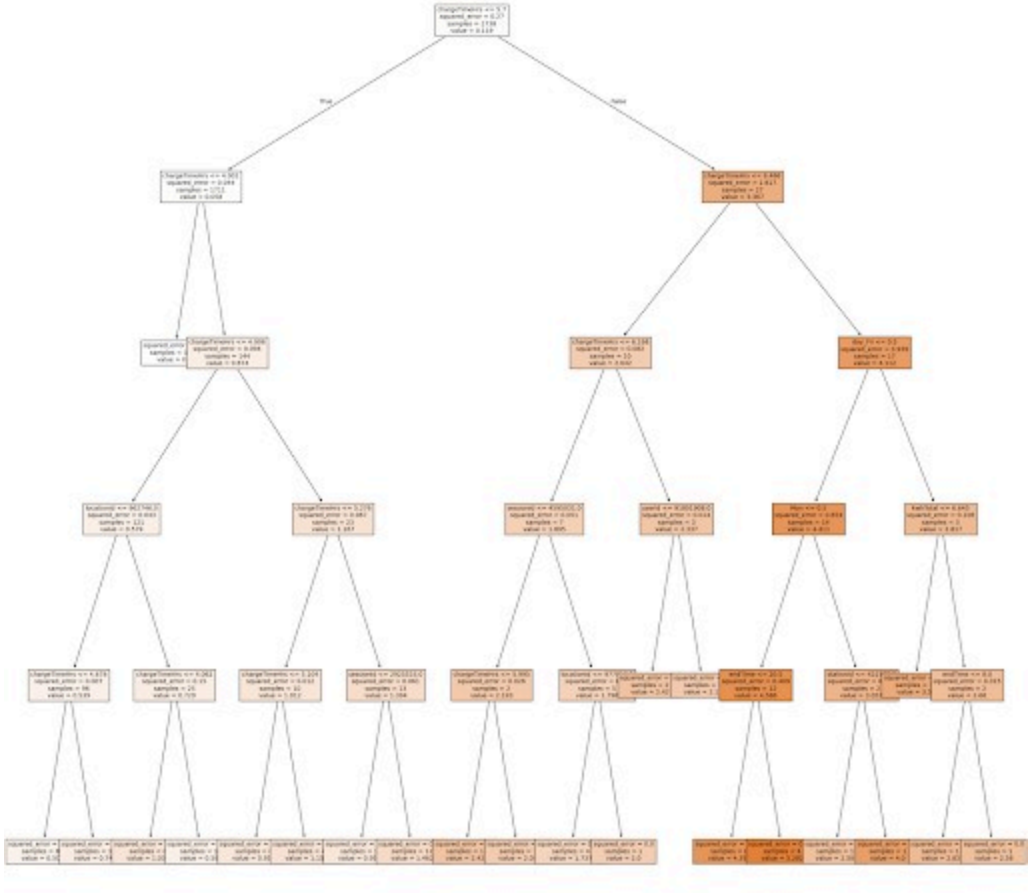
Results / Discussion

Random Forests



The Random Forest model's decision trees consistently prioritized splitting based on charge duration, effectively distinguishing between free and paid charging periods—specifically, workplace policies that often provide free charging for the first few hours. This clear division indicates that charge duration acts as a key threshold, directly correlating with whether costs are incurred. In our random forest model, We saw strong correlations between price and charging time, and our forest model had a strong accuracy of predicting our

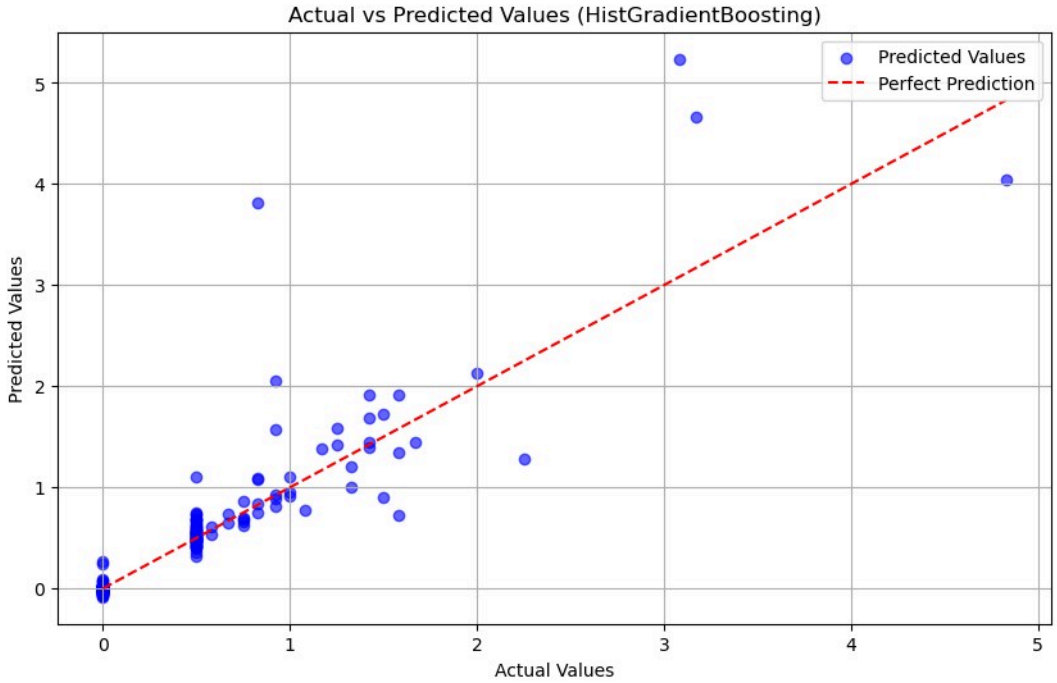
testing data (MSE of 0.046 and R^2 of .69). Due to having such an important feature in the form of charge time however, the model tended to prioritize it in tree generation and feature importance.



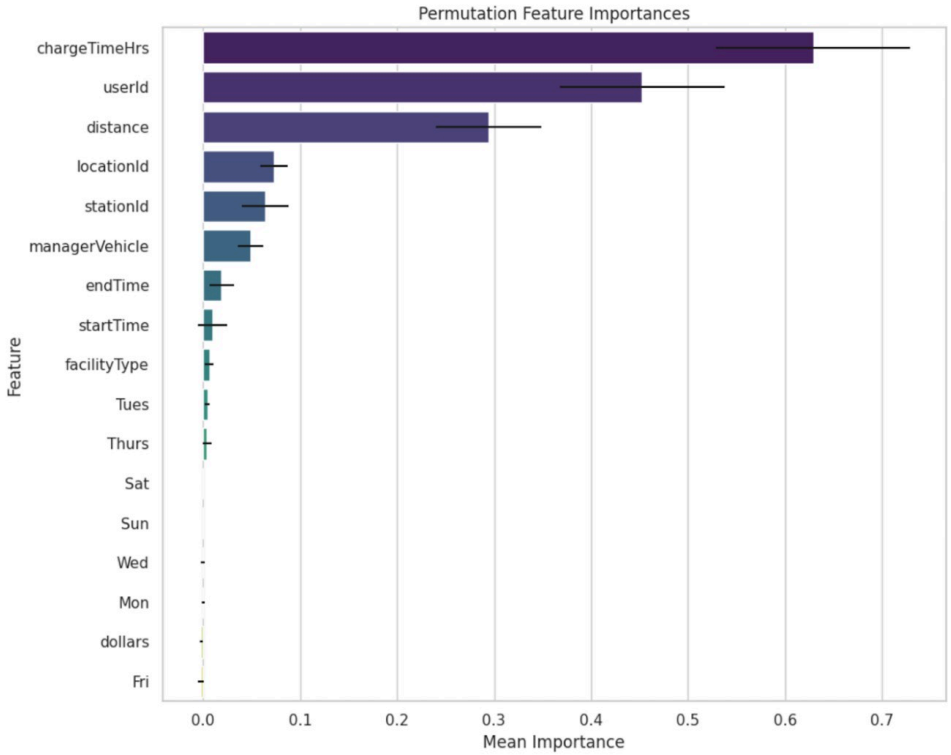
This sample tree shows at most steps, the model would focus on the charge time to solve for charging price, finding a clean divide of the “free charges” (charge times ≤ 4 hours) then dividing the rest subsequently. Overall, this model shows us how while it is possible to accurately predict the cost of your charge,

there are a lot of factors that are not impactful in settings with corporate chargers, with the key one being how long you leave your car on charge.

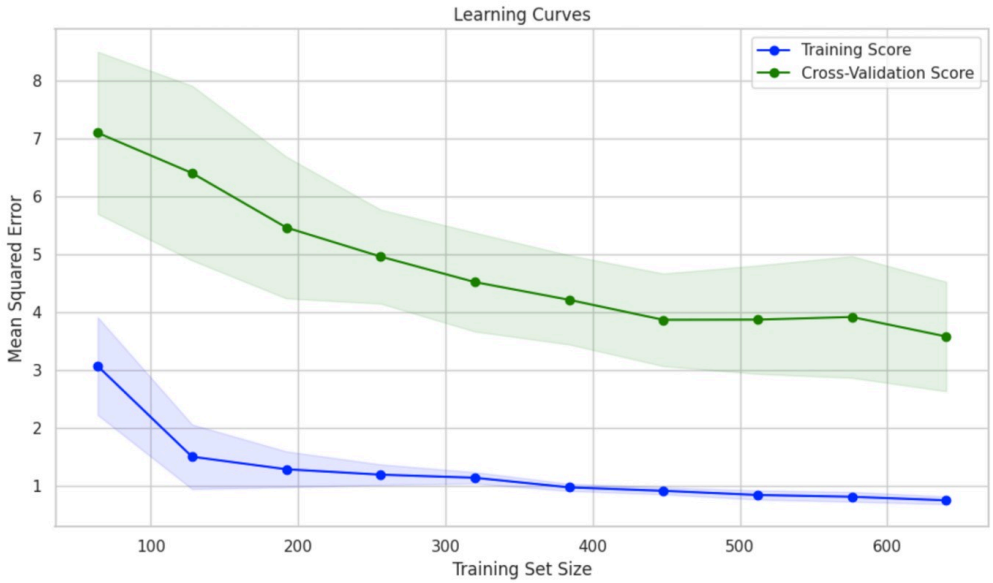
HGBR Results



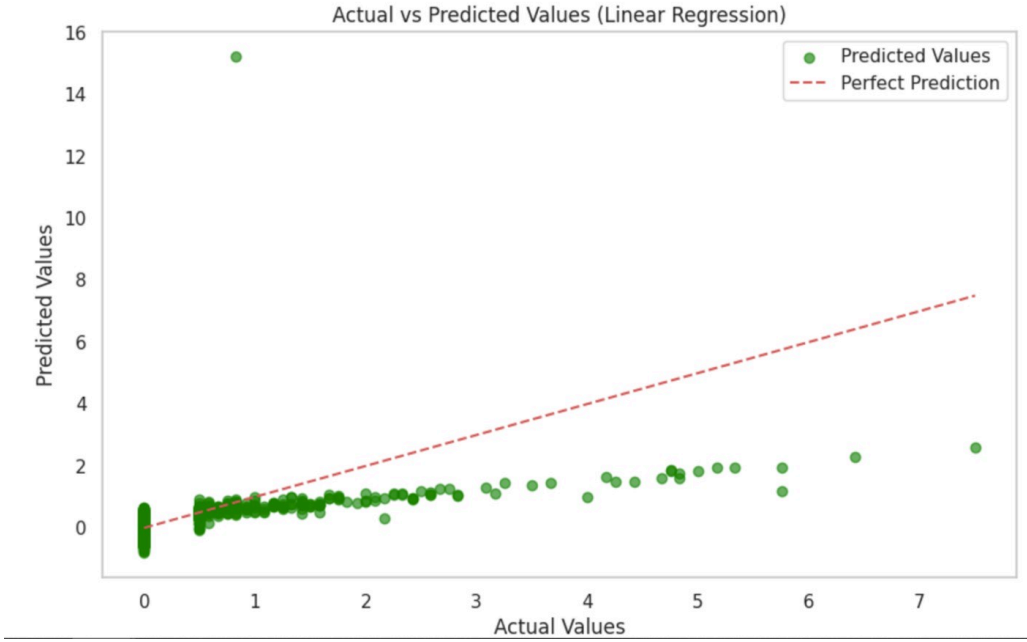
In our HGBR, we saw another relatively accurate prediction model from our samples, resulting in a MSE of .03 and R^2 of .78. Along with this, we also saw a similar case of strong correlations between key features such as charge time and distance since last drive (shown below)



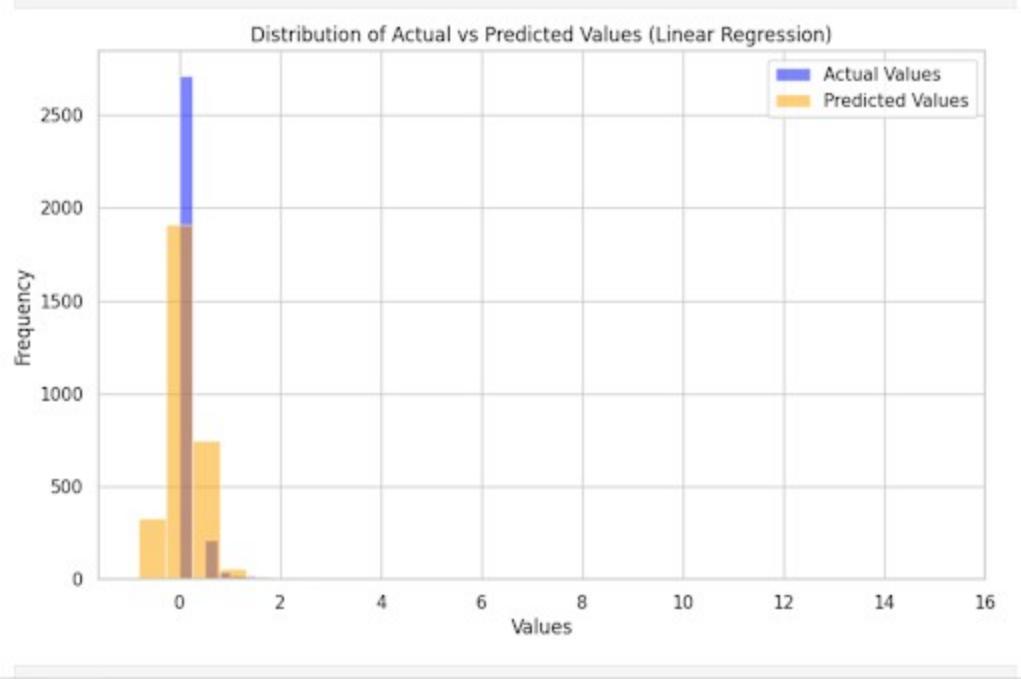
We also saw the model had a strong learning curve, showing it was able to find key features and achieve dramatically smaller MSE starting at a training set size of 200, then seeing diminishing returns as more and more data was added to the training set.


Learning Curves
Training Score
Cross-Validation Score
Mean Squared Error
Training Set Size
1
2
3
4
5
6
7
8
100
200
300
400
500
600

GLR Results


Actual vs Predicted Values (Linear Regression)
Predicted Values
Perfect Prediction
Predicted Values
Actual Values
0
2
4
6
8
10
12
14
16
0
1
2
3
4
5
6
7

With our current GLR model, it was shown to be the weakest out of our 3 chosen models. It had a low R^2 score of 0.15, indicating that a lot of the variability in the targets is still very weak. The MSE shows that the individual predictions are relatively close to the actual values which is still a good sign. The overall impact of the model is still very weak. The graph shows that the model is underfitting and the reason this could be is that it is not finding a proper correlation between charging times and price. What is most likely happening is that the model can't properly correlate that there are 2 variables that are used, "start time" and "end time" and it is not properly understanding the time in between and instead is focused more on when the charging starts and when the charging ends instead. This in turn has been predicted for almost different reasons. To improve the model's performance, we could revisit the data preprocessing and change it to address the correlation in time charged more.



Conclusions

In conclusion, this project highlighted the potential of machine learning models in predicting electric vehicle (EV) charging costs with improved transparency and accuracy. Among the approaches explored, the Histogram Gradient Boosting Regressor (HGBR) achieved the best results, with an R^2 of 0.78 and a low mean squared error (MSE) of 0.03, indicating that approximately 78% of the variance in electric vehicle charging costs can be explained by the features used in the model. These results demonstrate a strong predictive capability, slightly outperforming the Random Forest model (R^2 Score of 0.6915). The insights gained from this model can help stakeholders identify significant factors influencing charging costs, such as charging duration and distance driven. By utilizing these findings, consumers can make more informed decisions about their charging habits, enhancing transparency in pricing. This increased clarity not only empowers users to optimize their charging strategies but also contributes to the broader goal of promoting electric vehicle adoption by making the cost structure more understandable and accessible.

Visualization of feature importance further supports this finding, with charge duration accounting for the largest share of variance in cost prediction. Additionally, scatter plots of charging cost versus duration demonstrate a steep cost increase once the free period ends, solidifying the relationship. The dataset itself amplifies this pattern: in corporate settings, charging stations often employ pricing schemes tied to time, making duration a predictable and dominant factor.

Ultimately, this project serves as a step towards creating tools that can alleviate consumer hesitations regarding EV adoption by demystifying charging expenses. Future work could expand on these results by integrating real-time

pricing data or exploring additional machine learning architectures to better capture the nuances of EV charging behavior.

References

[1] R. Merritt, “What is a transformer model?,” NVIDIA Blog, https://blogs.nvidia.com/blog/what-is-a-transformer-model/ (accessed Oct. 4, 2024).

[2] W. Lee, R. Schober, and V. W. Wong, “An analysis of price competition in Heterogeneous Electric Vehicle Charging stations,” IEEE Transactions on Smart Grid, vol. 10, no. 4, pp. 3990–4002, Jul. 2019. doi:10.1109/tsg.2018.2847414

[3] Y. Mohammed, A. Manoharan, S. Kappagantula, and H. Manoharan, “Optimizing Electric Vehicle Charging costs using machine learning,” Ingénierie des systèmes d information, vol. 29, no. 3, pp. 1085–1095, Jun. 2024. doi:10.18280/isi.290326

[4] S. Koohfar, W. Woldemariam, and A. Kumar, “Performance comparison of deep learning approaches in predicting EV charging demand,” Sustainability, vol. 15, no. 5, p. 4258, Feb. 2023. doi:10.3390/su15054258

Report

Gantt Chart

https://gtvault.sharepoint.com/❌/s/Team121MachineLearning/EdIX0VWGFLNLhz6CMy1lxgEBJixHnTw4H_vmL

_SpxsBlRg?e=cY9PNd

Contribution Table

https://gtvault.sharepoint.com/❌/s/Team121MachineLearning/Ecm5regOD1BBhO6YGlpupwYBjoA48WnT8n6ZeQJSTAJzzA?e=6CRnIq

Video Link

https://youtu.be/mVZMb5IsKD4