Vehicle Range Prediction Final Report

Min Zhou
minzhou@bu.edu

Yuchen Wang
wangyc95@bu.edu

Xinqiao Wei
weixq95@bu.edu

1 Introduction

The electric vehicle has been a hot topic for a long time due to its environmental-friendly, efficiency and free of noise. However, the difficulty of predicting vehicle range efficiently and accurately and managing energy storage of electric vehicles become the most challenging problems for improving the long-term performance of electric vehicles.

In this project, we are working with Electra Vehicles on predicting electric vehicle range from provided system specifications. The vehicle range is defined as how far can you drive on a full charge. However, there are many sources of uncertainty that make this difficult. For example, the road condition, driving style, and the limitation of vehicle system configuration. We have to find the relationship among those main variables to predict the vehicle range.

1.1 Problem statement

Considering the specifications of the energy storage system, the vehicle modeling parameters, and the simulated vehicle operating conditions, our goal is to project the resultant simulated vehicle range without performing a detailed cycle analysis. The resulting models will be generated for each of the supplied

EPA driving cycles and will account for any possible simulated vehicle or energy storage system. As vehicle range is a nonlinear factor, as opposed to system cost, system power, and system range, such a calculation would serve to better estimate all core energy storage system specifications prior to performing a detailed electrochemical analysis.

1.2 Dataset

The raw dataset is collected from Amazon S3 Buckets in JSON format provided by Electra Vehicles. We transformed 20,000 JSON files to Pandas DataFrame and split the dataset into 13,400 training examples and 6,600 test examples to train and evaluate our model.

2 Method

There are some existing researches for predicting electric vehicle range. Oliva and others proposed a model-based approach under a stochastic framework[1]. Another model uses a feature-based linear regression framework[2] for a real-time application.

We approach the solution by using Random Forest, Decision Tree, K-Nearest Neighbors Regression, and Ordinary Least Squares algorithms to train 4 different regression models. The trained models are used to predict the vehicle range given the specific input, and we compared them using commonly used regression evaluation metrics: MAE (Mean Absolute Error), MSE (Mean Squared Error) and R-Squared Score.

2.1 Data preparation

We are using the dataset collected by Electra Inc. Before choosing the features, we transformed the JSON files to Pandas DataFrame. To feed our regression model, we need numeric values as our input features. As our first try, consider the curse of dimensionality, we selected 7 important specifications as our features, specifically, system weight, the number of HEUs, the number of HPUs, and dP threshold, etc. The output variable would be the system range. Furthermore, after discussing with our partner and comparing feature importances we added some new features calculated by existing variables. Finally, 14 significant features are used to train the models.

2.2 Random Forest Model

Random forest is an ensemble learning method by building a multitude of decision trees at training time and the class that most trees predicted would be the final prediction of the Random Forest algorithm. By randomly select features and samples, random forest algorithm correct for decision trees' habit of overfitting to their training set.

When we applied the Random Forest algorithm to the dataset, the model is more likely overfitting at the beginning. To avoid that, we optimized a tuning parameter that governs the number of features that are randomly chosen to grow each tree from the bootstrapped data.

2.3 K-Nearest Neighbors Model

K-Nearest Neighbours is a popular classification and regression algorithm in machine learning area by its accuracy and simplicity. The algorithm is calculating the distance of between each test set sample and all training set samples, then check the labels of K nearest training samples, the label that appeared for most times would be the prediction of the test sample. So, the picking of K could be really important, one safe way is trying every K values and compare the results.

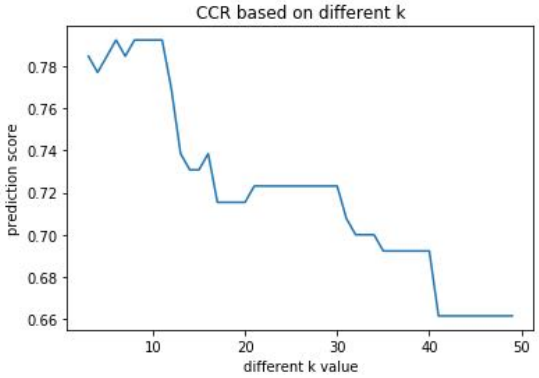

Figure 1: The CCR with different K

As we can see in Figure 1, we tried many different K values from 3 to 50 and plot the accuracy against different K, when K = 6, 8, 9, the KNN models output the best result which is 79.23%. Furthermore, we applied the model into the training set, and when K = 6, it outputs the best training accuracy which is 87.46%, as a result, we set K = 6.

2.4 Decision Tree Model

Decision tree learning uses a decision tree as a predictive model to go from observations about an item represented in the branches to conclusions about the item's target value represented in the leaves. The algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items.

There are various advantages of decision trees.

- Simple to understand and interpret.
- Performs well with large datasets.
- Mirrors human decision making more closely than other approaches.

We tried both decision tree regression and decision tree classification with threshold. The test accuracy for decision tree classification is 85.38% while the R-Square score of decision tree regression is 0.997. As the property of the decision tree, the training accuracy for both classification and regression are 100%.

2.5 Ordinal Least Squares Model

Ordinary Least Squares (OLS) is a linear regression model, by applying OLS method, it is easy to obtain the unknown parameter which minimizes the sum of the squares of the differences between the ground truth and those predicted by the linear function.

We can easily construct the mathematical model for the problem using the trained OLS model coefficients and understand the relationships between each variable and the system range by visualizing the partial regression results.

3 Results

3.1 Features

We used correlation matrix to explore the correlations among 14 features we selected as shown in Figure A.1. The correlation coefficients shows the vehicle input variables are highly correlated as well as HEU power per kilogram. Besides, the number of HPU and HEU are positive correlated with their nominal energy.

By using statistical methods, we compared the importance of each feature as shown in Figure 2, we found that the top 5 important features are HEU nominal energy, HPU nominal energy, the number of HPU, the number of HEU and HEU cost per kw. We also presented the top 12 feature distributions respects to system range in Figure A.2.

3.2 Prediction Models

The performances of all models are shown in Table 1. Considering the MAE, MSE and R-Squared value, the Random Forest model is the best regressor for our problem.

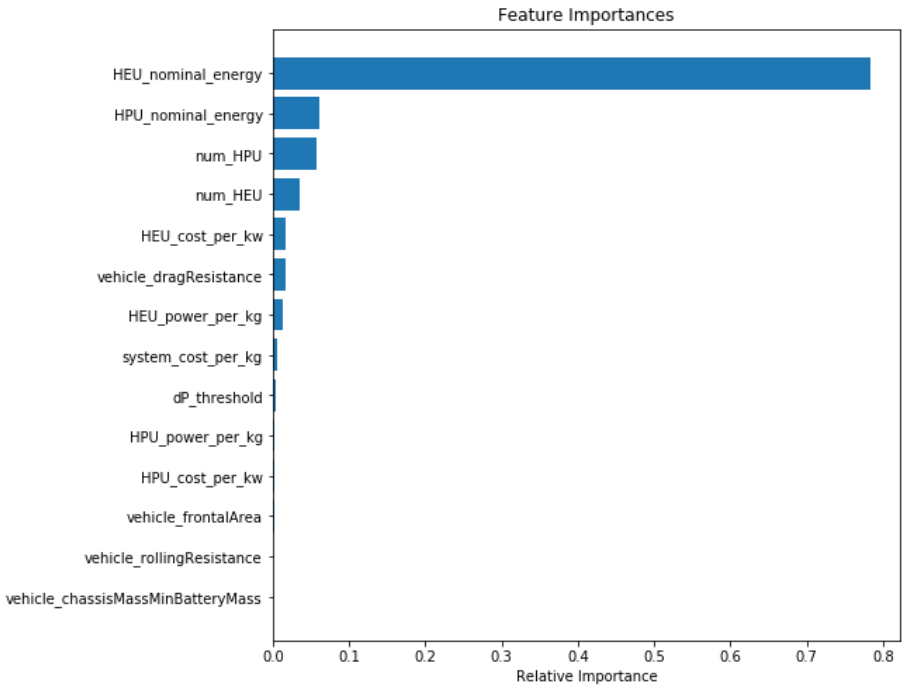

Figure 2: The feature importances

Models	MAE	MSE	R-Squared
Random Forest	0.552	1.541	0.999
K-Nearest Neighbors Regression	1.605	29.333	0.993
Decision Tree Regression	2.231	12.435	0.997
Ordinary Least Squares	3.850	89.398	0.980

Table 1: Model Performance Comparison

With the Ordinary Least Squares method, we achieved the following mathematical model for our problem:

$$SystemRange = A \cdot X$$

where A is 1x14 Linear Regression model coefficient vector and X is 14x1 input feature vector. The partial OLS regression results of each features are shown in Figure A.3

3.3 Vehicle Range Prediction API

We developed an easy-to-use vehicle range prediction API including 5 different models which can be used to predict the electrical vehicle range for Electra

Vehicles. To use the created API, it is possible to follow three simple steps:

- Input your JSON file as input.
- Select the model you would like to use.
- Get the prediction result.

4 conclusion

Our major results of this project can be summarized as follows:

- 14 Important features which including the system level, HPU, HEU specifications, and vehicle parameters are found to simulate the vehicle range.
- Best regression model reaches 0.999 R-squared value without overfitting.
- Easy-to-use API for Electra Vehicles to simulate vehicle or energy storage system or develop other products.

5 Future work

We may try some other state-of-art technique such as Neural Network or Deep Learning in the further since they did an outstanding work in most practical problems. Also, more data can be added to train a more robust model. Furthermore, we will try to scale or normalize the features and do some regularization to avoid overfitting.

6 Acknowledgement

We appreciate Electra Vehicles, Inc. for giving us this opportunity to work on this project and providing great help.

References

[1] Javier A Oliva, Christoph Weihrauch, and Torsten Bertram. A model-based approach for predicting the remaining driving range in electric vehicles. In *Annual conference of the prognostics and health management society*, volume 4, 2013.

[2] Peter Ondruska and Ingmar Posner. Probabilistic attainability maps: Efficiently predicting driver-specific electric vehicle range. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 1169–1174. IEEE, 2014.

A Appendix

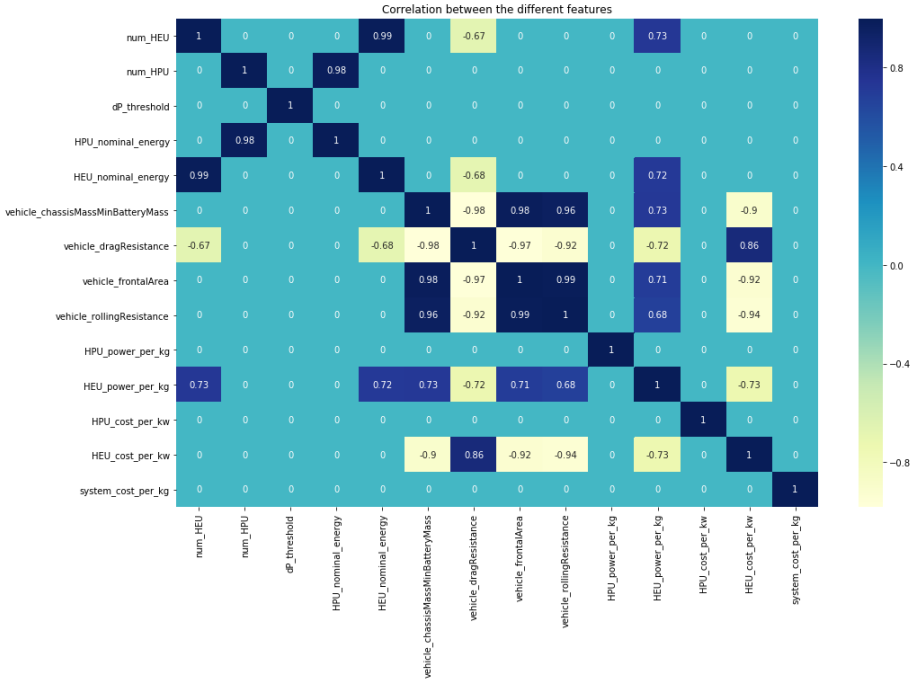

Figure A.1: The correlation between 14 features

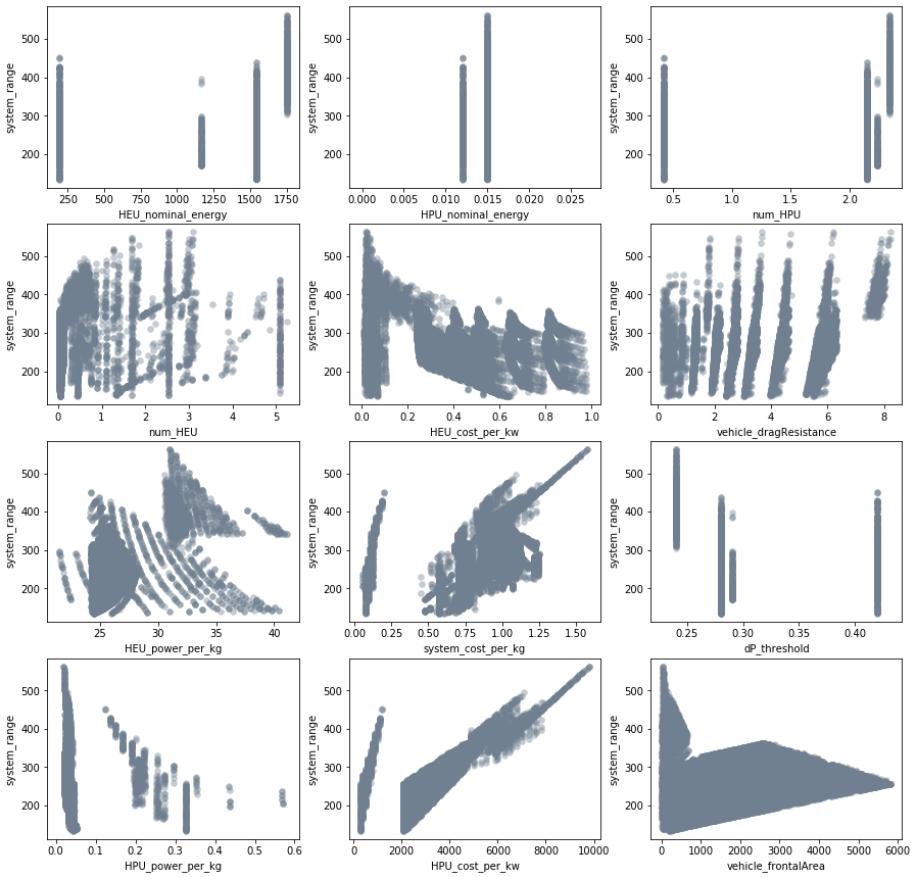

Figure A.2: The top 12 feature distributions

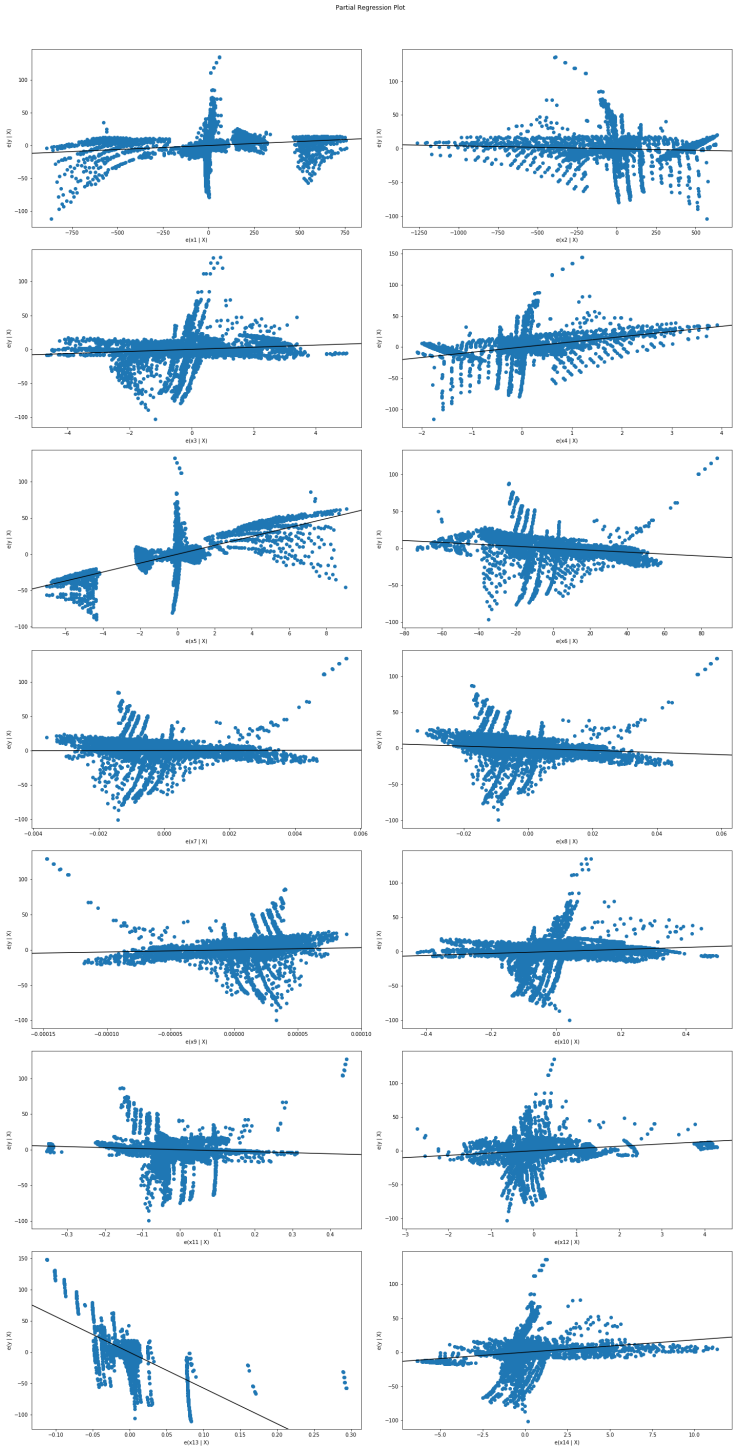

Figure A.3: The partial regression results of each features