Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 337-344
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Comparative Analysis of Machine Learning Models for Predicting Electric Vehicle Range

Gregorius Airlangga
Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id

Abstract

This research presents a comprehensive analysis of various machine learning models to predict the electric range of electric vehicles (EVs). In the context of growing environmental concerns and the push for sustainable transportation, accurate prediction of EV range is crucial for consumer trust and wider adoption. We evaluated five different models: Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Gradient Boosting Regressor, using a dataset that included a diverse array of EV attributes. The primary evaluation metric was the Mean Squared Error (MSE), applied both in cross-validation and on a test set. Our findings revealed significant differences in performance between linear models and ensemble methods. Linear models, while computationally efficient and interpretable, showed modest predictive capabilities, likely limited by their inability to capture complex, non-linear relationships in the data. Notably, Lasso Regression exhibited the highest error rates, possibly due to its feature exclusion in regularization. In contrast, ensemble methods, particularly the Random Forest Regressor and Gradient Boosting Regressor, demonstrated superior performance, effectively modeling non-linear relationships and intricate feature interactions. This study underscores the importance of model selection in predictive tasks, highlighting that more complex models, such as ensemble methods, are often more suitable for datasets with multifaceted interactions and non-linearities. The results of this research contribute to the evolving field of electric vehicle technology, providing insights that can guide future developments in EV range prediction, a key factor in the advancement of sustainable transportation. This research aids in understanding the application of machine learning in EV range prediction and lays the groundwork for future exploration, potentially incorporating real-time data and external factors for enhanced accuracy.

Keywords*: Electric Vehicles, Machine Learning, Ensemble Learning, EV Range Prediction, Transportation*

1. INTRODUCTION

The evolution of transportation towards sustainability is a cornerstone in the global fight against climate change, with electric vehicles (EVs) emerging as a pivotal element in this transformation [1]–[3]. The shift to electric mobility not only promises reduced emissions but also challenges traditional concepts of vehicle performance and utility [4]–[6]. A critical aspect in this context is the accurate prediction of an electric vehicle's range. The range, or the distance an EV can travel on a single charge, is a key determinant of user adoption and market penetration [7]–[9]. The existing body of research in this domain has primarily revolved around different facets of electric vehicles, ranging from their environmental benefits to the challenges of charging infrastructure and advancements in battery technology [5], [10], [11]. However, when it comes to predicting the electric vehicle range, the complexity intensifies, given the myriad of influencing factors such as vehicle attributes, environmental conditions, and user driving patterns.

In recent years, machine learning has emerged as a powerful tool in forecasting various aspects of electric vehicles [12]–[14]. Studies have shown that techniques like linear regression, when applied to vehicle specifications, can yield significant insights into range estimations. For instance, linear models have been employed to correlate vehicle characteristics with battery performance [15]. On a more advanced front, ensemble methods such as Random Forest [16] and Gradient Boosting [17] have been leveraged to enhance prediction accuracy, harnessing their ability to handle complex, non-linear relationships within the data. These methodologies have marked a significant step forward; however, they often fall short in fully integrating the diverse array of data types, particularly when dealing with both categorical and numerical variables, which is crucial for a holistic understanding of EV range dynamics [18].

The urgency to develop and refine predictive models for EV range cannot be understated, especially in the context of the global push towards reduced carbon emissions. The role of electric vehicles in this paradigm is increasingly critical, and the ability to accurately predict their performance is essential for consumer confidence and broader adoption [19]–[21]. Despite significant strides in utilizing machine learning for range prediction, the current state of the art often lacks a comprehensive approach that encompasses the multifaceted nature of EV data [22]–[24]. The goal of this research is to address these gaps by implementing a diverse array of machine learning algorithms, ranging from straightforward linear regressions to more complex models like support vector machines and advanced ensemble methods. The intent is to not only predict the electric range of various EV models but also to understand the relative importance of different features in these predictions. This approach seeks to expand upon existing methodologies by incorporating an extensive set of both categorical and numerical features, thereby providing a more nuanced and accurate model for EV range prediction.

This research makes several key contributions. Firstly, it introduces a detailed pre-processing pipeline that effectively combines various data types, enhancing the model's applicability to real-world scenarios. Secondly, it offers a comparative analysis of various machine learning models, examining their efficacy in range prediction. This comparison is crucial in identifying the most suitable methodologies for different types of EV data. Lastly, the study provides insights into the significance of different vehicle attributes and external factors in determining the electric range, contributing valuable knowledge to the ongoing development in the field of electric vehicle technology.

The structure of the article is designed to offer a comprehensive insight into the research process and findings. Following this introduction, the methodology section delves into the specifics of data preprocessing, feature engineering, and the selection of machine learning models. The experimental setup is then detailed, describing the dataset, its source [25], characteristics, and the design of the experiments, along with the evaluation metrics employed. The results and discussion section presents the outcomes of the experiments, shedding light on the performance of each model and discussing the implications of these findings. Recognizing the limitations of the current study, the subsequent section outlines potential areas for future research, paving the way for further advancements in this field. The article concludes by summarizing the key discoveries and their relevance to the advancement of electric vehicle technology.

2. RESEARCH METHODOLOGY

2.1. Data Acquisition and Preprocessing

The foundation of our research is a comprehensive dataset obtained from [26], which includes various attributes of electric vehicles (EVs). The dataset, titled 'Electric_Vehicle_Population_Data.csv', encompasses a wide range of features such as model year, make, base MSRP (Manufacturer's Suggested Retail Price), and electric range. Recognizing the importance of data quality, we initiated the process with meticulous data cleaning and preprocessing. We first addressed missing values, particularly in the 'Base MSRP' column, where zero values were assumed to indicate missing data and were thus replaced with NaN for subsequent imputation. Additionally, we engineered a new feature, 'Age', calculated as the difference between the current year (2023) and the vehicle's model year, to capture the potential impact of vehicle age on its electric range. Given the diverse nature of the dataset, we identified and segregated the features into numerical and categorical types. Numerical features included 'Model Year', 'Legislative District', 'Base MSRP', and the newly created 'Age'. The categorical features comprised 'County', 'State', 'Make', 'Electric Vehicle Type', and 'Clean Alternative Fuel Vehicle (CAFV) Eligibility'. To enhance the model's performance and interpretability, we removed irrelevant features such as 'VIN (1-10)', 'DOL Vehicle ID', 'Vehicle Location', and '2020 Census Tract', which were deemed non-contributory to the prediction of electric range.

2.2. Feature Engineering and Transformation

To adequately prepare the dataset for machine learning algorithms, we employed a combination of imputation and encoding techniques. The preprocessing pipeline included a ColumnTransformer to handle numerical and categorical data differently. For numerical features, we used SimpleImputer with a median strategy to fill in missing values. Categorical features were processed using OneHotEncoder to convert them into a format suitable for modeling. This approach ensured that our models could effectively learn from both types of data without any bias towards a particular data format.

2.3. Model Selection and Evaluation

Our research aimed to compare a range of machine learning models to identify the most effective approach for predicting the electric range of EVs. The selected models included Linear Regression, Ridge Regression, Lasso Regression, RandomForestRegressor, GradientBoostingRegressor, those methods are presented in the equations (1) – (5). Each model was encapsulated within a pipeline that included the preprocessor, ensuring that the same preprocessing steps were applied consistently across all models. We split our dataset into training and testing sets, allocating 80% of the data for training and the remaining 20% for testing. This split was performed to evaluate the models on unseen data, ensuring a robust assessment of their predictive capabilities. The random_state parameter was set to 42 to maintain reproducibility of the results. In addition, this split was crucial to evaluate the models on unseen data, thereby assessing their generalizability and performance in real-world scenarios. The evaluation of the models was primarily based on their mean squared error (MSE) as presented in the equation (6), both in cross-validation and on the test set. We used a 5-fold cross-validation approach to assess model performance on the training set. This method allowed us to understand the model's stability and reliability across different subsets of the data. The test set MSE provided a direct measure of how well the model could predict the electric range on new, unseen data.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{1}$$

$$\widehat{\beta^{Ridge}} = \backslash argmin_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right) \tag{2}$$

$$\widehat{\beta^{Lasso}} = \backslash argmin_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \tag{3}$$

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x;\Theta_b) \tag{4}$$

$$F(x) = F_0(x) + \sum_{m=1}^{M}\gamma_m h_m(x) \tag{5}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

2.4. Experimental Setup

To conduct the experiments, we used Python as our programming language, leveraging its rich ecosystem of data analysis and machine learning libraries. The pandas library was utilized for data manipulation and analysis, while NumPy was used for numerical computations. The machine learning models and preprocessing techniques were sourced from scikit-learn, a widely used library for machine learning in Python. Additionally, we employed XGBoost, a powerful library known for its efficiency and performance in regression tasks.

2.5. Model Evaluation

To evaluate the performance of each model, we employed a function 'evaluate_model'. This function first computed the cross-validation score using a 5-

fold cross-validation on the training set, with the scoring metric being the negative mean squared error (MSE). The models were then fitted on the training set and used to predict the electric range on the test set. The MSE was calculated for these predictions, providing a measure of the model's accuracy on unseen data. Both the mean cross-validation MSE and the test MSE were reported for each model, allowing for a comprehensive assessment of their performance.

3. RESULTS AND DISCUSSION

In this study, we evaluated five different machine learning models to predict the electric range of electric vehicles (EVs). The performance of each model was assessed using the Mean Squared Error (MSE) metric, both in cross-validation and on the test set. The results as presented in the table 1. Firstly, Linear Regression exhibited a Mean Cross-Validation MSE of 692.93 and a Test MSE of 680.94. This model, while being the simplest among those tested, provided a reasonable baseline for performance. However, its relatively higher MSE indicates a modest fit to the data, possibly due to the linear nature of the model which may not capture more complex relationships within the data. Secondly, Ridge Regression model, it has Mean Cross-Validation MSE of 763.88 and a Test MSE of 746.66. The performance was slightly inferior to the linear regression model. This suggests that the addition of the L2 regularization in Ridge Regression did not significantly contribute to handling overfitting or improving the model's performance for this particular dataset.

Thirdly, Lasso Regression, resulted in a Mean Cross-Validation MSE of 879.84 and a Test MSE of 858.79. The performance was the least favorable among the models tested. This outcome could be attributed to the nature of Lasso Regression, which applies L1 regularization and can lead to the exclusion of some features entirely from the model. This might have led to the omission of relevant predictors, hence the higher error rates. Then, Random Forest Regressor, it demonstrated a Mean Cross-Validation MSE of 48.82 and a Test MSE of 47.27, marking a significant improvement over the linear models. The robust performance of the Random Forest Regressor can be attributed to its ability to model non-linear relationships and interactions between features effectively. It is also less prone to overfitting due to the ensemble learning method. Lastly, Gradient Boosting Regressor, it showed a Mean Cross-Validation MSE of 116.36 and a Test MSE of 115.53. While not as performant as the Random Forest, it still significantly outperformed the linear models. The success of the Gradient Boosting Regressor can be credited to its sequential learning of weak learners, which helps in addressing the errors made by previous models, thereby refining the predictions.

The results highlight the superiority of ensemble methods, namely Random Forest and Gradient Boosting, in predicting the electric range of EVs. These models are particularly adept at handling the complex and non-linear relationships that are typical in real-world datasets like the one used in this study. Their ability to integrate diverse feature interactions effectively explains their lower MSE scores compared to linear models. The less impressive performance of linear models (Linear, Ridge, and Lasso Regression) underscores the limitations of linear assumptions in complex predictive tasks. While these models are computationally efficient and easier to interpret, they may not capture the intricacies present in datasets with complex interactions and non-linear relationships.

The significantly higher error rates in Lasso Regression point to the potential drawbacks of overly aggressive feature selection. This raises important considerations for feature engineering and selection in predictive modeling, especially in contexts where the relationships between predictors and outcomes are not fully understood. These findings have practical implications for the development of predictive models in the EV sector. They suggest that while simpler models might provide quick and interpretable results, more sophisticated ensemble methods should be considered for higher accuracy, particularly in applications where prediction accuracy is paramount. It is also crucial to note that model performance can be highly dependent on the nature of the dataset and the specific features involved. Therefore, the choice of model should be guided by both the characteristics of the dataset and the requirements of the specific predictive task at hand.

Table 1. Comparison Results

Methods	Mean Cross-Validation MSE	Test MSE
Linear Regression	692.93	680.94
Ridge Regression	763.88	746.66
Lasso Regression	879.84	858.79
RandomForestRegressor	48.82	47.27
GradientBoostingRegressor	116.36	115.53

4. CONCLUSION

This research set out to explore the efficacy of various machine learning models in predicting the electric range of electric vehicles (EVs). The models assessed included Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and Gradient Boosting Regressor. The primary evaluation metric used was the Mean Squared Error (MSE), both in cross-validation and on a test set. Our findings revealed a clear distinction in performance between linear models and ensemble methods. Linear Regression, Ridge Regression, and Lasso Regression demonstrated modest predictive capabilities, with their performance constrained likely by their inherent linear nature, which may not capture the complex, non-linear relationships present in the dataset. Among these, Lasso Regression showed the highest error rates, potentially due to its feature selection approach that could have excluded relevant predictors.

In contrast, the ensemble methods, particularly the Random Forest Regressor and the Gradient Boosting Regressor, exhibited superior performance. Their ability to model non-linear relationships and interactions between a wide range of features was reflected in their significantly lower MSE scores. The success of these models underscores the value of ensemble methods in handling complex predictive tasks, such as EV range prediction, where the interplay of various factors determines the outcome. The study highlights the importance of selecting appropriate machine learning models based on the nature of the dataset and the specific characteristics of the prediction task. While simpler models like linear regressions offer ease of interpretation and computational efficiency, more complex models like ensemble methods can provide greater accuracy in predictions, which is crucial in fields like EV technology where precision is key.

Future research could expand upon this work by exploring additional models, incorporating larger and more diverse datasets, and possibly integrating real-time data to enhance prediction accuracy further. Moreover, examining the impact of external factors such as environmental conditions and driving patterns could offer a more holistic view of EV range prediction. In conclusion, this research contributes to the growing body of knowledge in the field of electric vehicles and machine learning. It provides valuable insights into the application of different machine learning models for predictive tasks and lays the groundwork for future studies aimed at enhancing the reliability and accuracy of EV range predictions, a critical factor in the advancement and adoption of electric vehicle technology.

REFERENCES

[1] L. Xin, M. Ahmad, and S. I. Khattak, “Impact of innovation in hybrid electric vehicles-related technologies on carbon dioxide emissions in the 15 most innovative countries,” *Technol. Forecast. Soc. Change*, vol. 196, p. 122859, 2023.

[2] G. Bhatti, H. Mohan, and R. R. Singh, “Towards the future of smart electric vehicles: Digital twin technology,” *Renew. Sustain. Energy Rev.*, vol. 141, p. 110801, 2021.

[3] M. Mohammadi, J. Thornburg, and J. Mohammadi, “Towards an energy future with ubiquitous electric vehicles: Barriers and opportunities,” *Energies*, vol. 16, no. 17, p. 6379, 2023.

[4] A. Ghosh, “Possibilities and challenges for the inclusion of the electric vehicle (EV) to reduce the carbon footprint in the transport sector: A review,” *Energies*, vol. 13, no. 10, p. 2602, 2020.

[5] M. Muratori *et al.*, “The rise of electric vehicles—2020 status and future expectations,” *Prog. Energy*, vol. 3, no. 2, p. 22002, 2021.

[6] J. Van Mierlo *et al.*, “Beyond the state of the art of electric vehicles: A fact-based paper of the current and prospective electric vehicle technologies,” *World Electr. Veh. J.*, vol. 12, no. 1, p. 20, 2021.

[7] S. C. Mukherjee and L. Ryan, “Factors influencing early battery electric vehicle adoption in Ireland,” *Renew. Sustain. Energy Rev.*, vol. 118, p. 109504, 2020.

[8] V. Singh, V. Singh, and S. Vaibhav, “A review and simple meta-analysis of factors influencing adoption of electric vehicles,” *Transp. Res. Part D Transp. Environ.*, vol. 86, p. 102436, 2020.

[9] C. Chen, G. Z. de Rubens, L. Noel, J. Kester, and B. K. Sovacool, “Assessing the socio-demographic, technical, economic and behavioral factors of Nordic electric vehicle adoption and the influence of vehicle-to-grid preferences,” *Renew. Sustain. Energy Rev.*, vol. 121, p. 109692, 2020.

[10] M. Kumar, K. P. Panda, R. T. Naayagi, R. Thakur, and G. Panda, “Comprehensive Review of Electric Vehicle Technology and Its Impacts: Detailed Investigation of Charging Infrastructure, Power Management, and Control Techniques,” *Appl. Sci.*, vol. 13, no. 15, p. 8919, 2023.

[11] P. Franzese *et al.*, “Fast DC Charging Infrastructures for Electric Vehicles: Overview

of Technologies, Standards, and Challenges," *IEEE Trans. Transp. Electrif.*, 2023.

[12] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal, "A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability," *Energy \& Environ.*, vol. 33, no. 8, pp. 1583–1612, 2022.

[13] V. Chandran, C. K. Patil, A. Karthick, D. Ganeshaperumal, R. Rahim, and A. Ghosh, "State of charge estimation of lithium-ion battery for electric vehicles using machine learning algorithms," *World Electr. Veh. J.*, vol. 12, no. 1, p. 38, 2021.

[14] M. S. Ibrahim, W. Dong, and Q. Yang, "Machine learning driven smart electric power systems: Current trends and new perspectives," *Appl. Energy*, vol. 272, p. 115237, 2020.

[15] S. B. Vilsen and D.-I. Stroe, "Battery state-of-health modelling by multiple linear regression," *J. Clean. Prod.*, vol. 290, p. 125700, 2021.

[16] K. Liu, X. Hu, H. Zhou, L. Tong, W. D. Widanage, and J. Marco, "Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 6, pp. 2944–2955, 2021.

[17] A. Manoharan, K. M. Begam, V. R. Aparow, and D. Sooriamoorthy, "Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review," *J. Energy Storage*, vol. 55, p. 105384, 2022.

[18] A. Almahdi *et al.*, "Boosting Ensemble Learning for Freeway Crash Classification under Varying Traffic Conditions: A Hyperparameter Optimization Approach," *Sustainability*, vol. 15, no. 22, p. 15896, 2023.

[19] W. Zhang, S. Wang, L. Wan, Z. Zhang, and D. Zhao, "Information perspective for understanding consumers' perceptions of electric vehicles and adoption intentions," *Transp. Res. Part D Transp. Environ.*, vol. 102, p. 103157, 2022.

[20] S. E. Bibri, J. Krogstie, A. Kaboli, and A. Alahi, "Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review," *Environ. Sci. Ecotechnology*, vol. 19, p. 100330, 2024.

[21] N. V Martyushev, B. V Malozyomov, S. N. Sorokova, E. A. Efremenkov, D. V Valuev, and M. Qi, "Review models and methods for determining and predicting the reliability of technical systems and transport," *Mathematics*, vol. 11, no. 15, p. 3317, 2023.

[22] M.-F. Ng, J. Zhao, Q. Yan, G. J. Conduit, and Z. W. Seh, "Predicting the state of charge and health of batteries using data-driven machine learning," *Nat. Mach. Intell.*, vol. 2, no. 3, pp. 161–170, 2020.

[23] J. Zhao *et al.*, "Battery safety: Machine learning-based prognostics," *Prog. Energy Combust. Sci.*, vol. 102, p. 101142, 2024.

[24] M. Sharif and H. Seker, "Smart EV Charging with Context-Awareness: Enhancing Resource Utilization via Deep Reinforcement Learning," *IEEE Access*, 2024.

[25] D. A. Ramadhan, S. Rochimah, and U. L. Yuhana, "Classification of non-functional requirements using Semantic-FSKNN based ISO/IEC 9126," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 13, no. 4, pp. 1456–1465, 2015.

[26] Y. Singhal, "Electric Vehicle Population Dataset." 2022.