Price Prediction and Feature Importance Analysis of German Electric Vehicles Based on Boosted Decision Tree Model

Mingjun Sun[1, *, †], Jiahao Ye[2, †] and Peiyan Ye[3, †]

[1]Revelle College, University of California, San Diego, USA

[2]Milgard School of Business, University of Washington, Tacoma, USA

[3]College of Arts and Sciences, University of South Florida, Tampa, USA

*Corresponding author: mis005@ucsd.edu

†These authors contributed equally.

Abstract. As people realize the importance of non-renewable energy sources, more and more industries are looking for alternative sources of energy that can be used. The automotive industry is a prime example. Electric vehicles are having an increasingly dominant role in the automotive market. This paper investigates the price of electric vehicles (EVs) and the importance of consumers' characteristics for electric vehicles in the European region. In addition, feature importance is discussed with disordered data in selecting the best regression model. Finally, the Boosted decision tree model predicts future electric vehicle prices in the German region. The results show that consumers place more importance on the size of the car, the battery's durability, and acceleration. Although the penetration of electric vehicles in Germany is still tiny, government incentives for electric vehicle companies and consumers will increase the market penetration of electric vehicles. So we believe that in the future, people will be able to buy good-performing electric cars at more affordable prices.

Keywords: Electric Vehicles; Price Prediction; Feature Importance; Boosted decision tree; Renewable Energy.

1. Introduction

The development of reliable and sustainable modern energy sources has been one of the goals of the United Nations in recent years. As various environmental problems, such as global warming and air pollution, are becoming more and more serious, people are starting to pay attention to using clean energy in their daily lives. Among them, the issue of travel has received the most attention. Every year, countries worldwide are developing new environmental protection travel plans and tightening restrictions on vehicle emissions. After several years of science and publicity, many people are aware of the harmful effects of car emissions on the environment. At the same time, governments worldwide promote the development and use of electric vehicles (EVs). The increase in EVs in Germany, Denmark, and the Netherlands are statistically dependent on the increase in low-carbon power generation capacity [1]. We can find the advanced or perfected technologies that are more attractive for consumers to buy new energy vehicles. In addition, it was found that government incentives also affect the penetration of EVs in the local market. However, the magnitude of the effect is related to their vehicle-specific characteristics and the availability of local chargers [2].

This paper looks at the data's correlation based on the data visualization results. More regression models and components of analytical models are then attempted to understand which factors determine the price of EVs in Germany. Then predictions are made, and the importance of features is discussed using disordered data to improve the accuracy of the predictions.

2. Related Work

The basic starting price of an EV is relatively higher than their same-brand gas-fueled vehicle, while the future cost of owning an EV will relatively be lower because electric power is cheaper than gas power [3]. However, a recent study on Ireland's EV market shows that the key determinants of

buying a relatively higher price vehicle for people are the car's driving range, fuel cost savings, and charge time, so people are more interested in the car's performance and features while considering to buy an EV. They also calculated the marginal willingness to pay in Euros on EVs for every attribute [4]. Common sense rose that if a product has better performance, it deserves to set a higher commercial value on itself.

However, the price of EVs over the next ten years can be forecasted by analyzing the availability of chargers and consumer demand for EVs. The evolution of battery EVs over the last ten years is analyzed, and the future development of EVs in the UK is predicted. According to the study, EVs in the UK market do not have much choice for consumers with low car purchase budgets [5]. The high price of EVs is because the cost of EVs and the availability of public fast chargers limit the popularity of EVs. Although the government has pledged to invest over £1 billion in charging infrastructure, it does not fully address the shortage of chargers and the increased demand for EVs. Nowadays, home fast chargers can be purchased with the car at a discount through a government grant. However, a small group of people cannot charge at home, which makes placing public chargers and the charges for this group a concern. It has also become a concern for consumers who have not yet purchased an EV. To sum up, if the complementary goods of EVs are more affordable and convenient than fuel cars, even if the price of EV sales remains the same, EVs are still the best choice for consumers in terms of long-term spending.

3. Methodology

Our experiments is build on the previous research to and improve the accuracy of prediction by re-collecting and reprocessing the data, increasing field types, analyzing linear correlations, trying more regression models, and ultimately analyzing the importance of the features.

3.1 Data

3.1.1 Data Collection

The data we are using was collected from ev-dabase, an organization dedicated to collecting data on all types of EVs [6]. The organization believes that the combination of new energy and autonomous driving will be the future trend of the automotive industry. The data on EVs can explain many of the questions that exist today, so this database not only collect official data, but also contains a lot of real-world data. Therefore, we chose the content from this database for our analysis. Since the database is presented as an html page, we created a regular expression to get variables. 229 records are collected.

3.1.2 Data Pre-processing

In data pre-processing stage, we first converted the data into easily understandable formats. Next all variables that were not analytically meaningful were removed. For missing values, two types of processing were used. We tried to find the missing values related to car parameters through the car's official website, etc. Otherwise, the records were deleted. For Germany price, we refered to the UK and NL price because the prices in the three regions are similar. The final frame of our data set is shown in Table 1.

Table 1. Database Architecture

Features	Type	Description
brand	string	brand of the vehicle
year	int	Year of vehicle launch
month	int	Month of vehicle launch
isCurrent	int	0 AND 1: if the vehicle is already launched
battery(kWh)	float	Battery Capacity
isTow	int	0 AND 1: if tow is supported
towWeight	int	0 OR max weight supported
driveMode	string	rear OR front OR all
shape	string	cabrio OR hatchback OR liftback OR sedan OR spv OR station OR suv
size	string	A:Mini B:Small C:Medium D:large E:Executive F:Luxury N:Commercial
seat	int	seat number
acceleration(sec)	float	seconds needed to acceleration from 0 to 100km/h
topSpeed(km/h)	int	max speed
range(km)	int	how far the vehicle can run after fully charged
efficiency(Wh/km)	int	how many watt hours of electric energy is consumed per kilometer
fastCharge(km/h)	int	how far the vehicle can run after one hour of fast charge

In modeling process, encoding the categorical fields is required. Considering that one-hot encoding is a popular strategy used in some analyses nowadays and it may ignore some possible relationships [7]. In addition, our data set has only over two hundred records, so avoiding high dimension may also be important. Therefore, the data set is both target encoded and one-hot encoded, and we will try to use both datasets in the prediction and compare the results.

3.2 Correlation Between Fields

3.2.1 Data Visualization

Before analyzing the correlation between each field (including price), visualizing some general distributions and correlations of each fields presents an intuitional perspective on their relationships. The following correlation analysis only contains fields with numerical values.

Figure 1 shows that all vehicles launched to the public are no more than two years before now or in the upcoming year to preserve the authenticity of the data and compatibility with the current vehicle market.

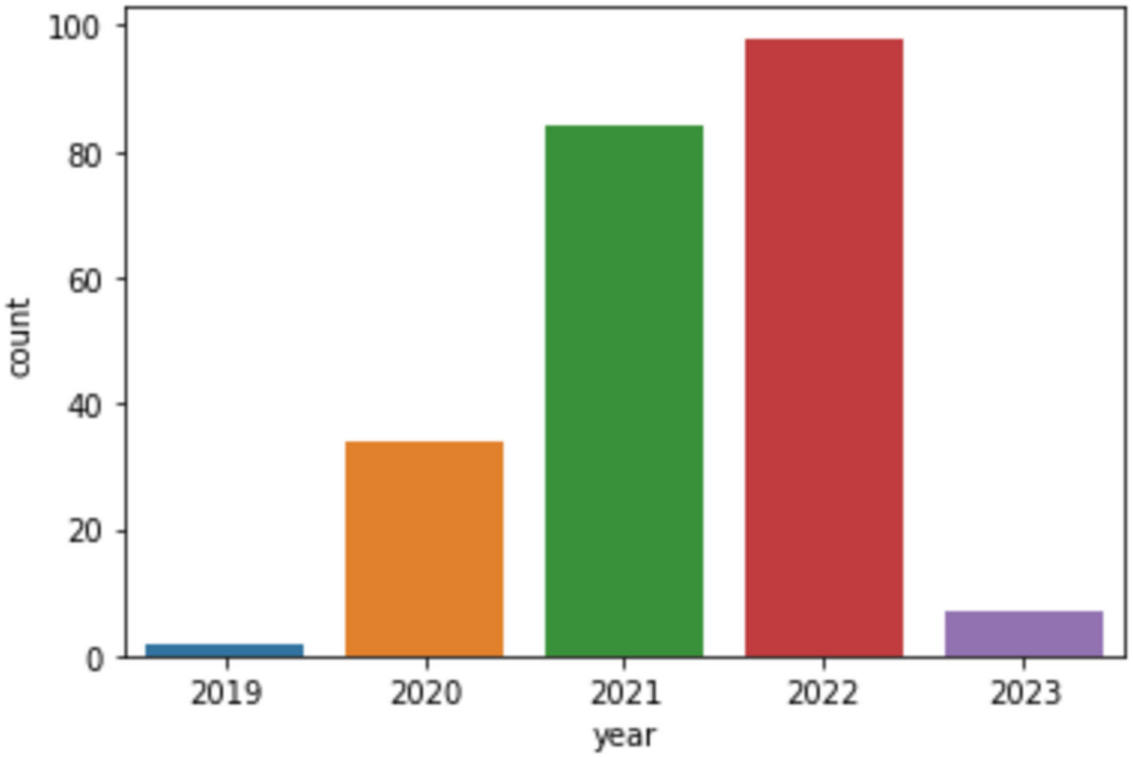

Fig. 1 Vehicle Launch Time (Year)

Besides, the Figure 2 shows that most EVs are released around November every year.

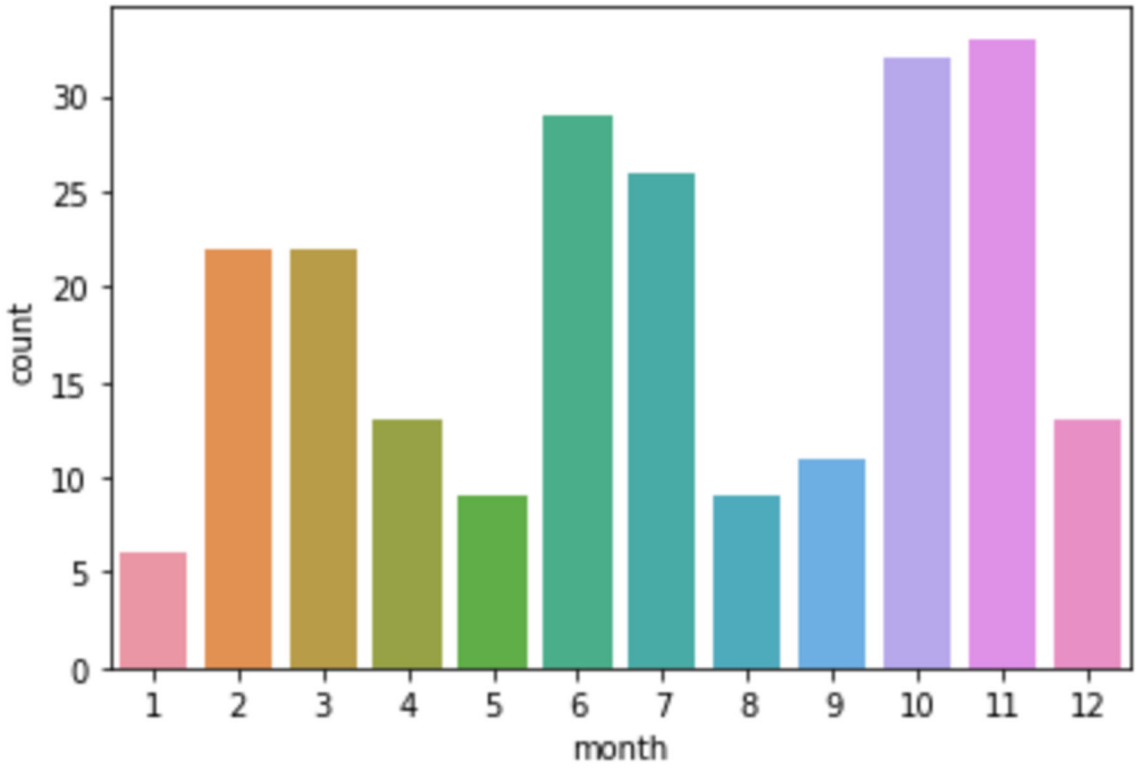

Fig. 2 Vehicle Launch Time (Month)

Figure 3 gives an insight that Mercedes, Porches, and Audi have the most amount of car in the database.

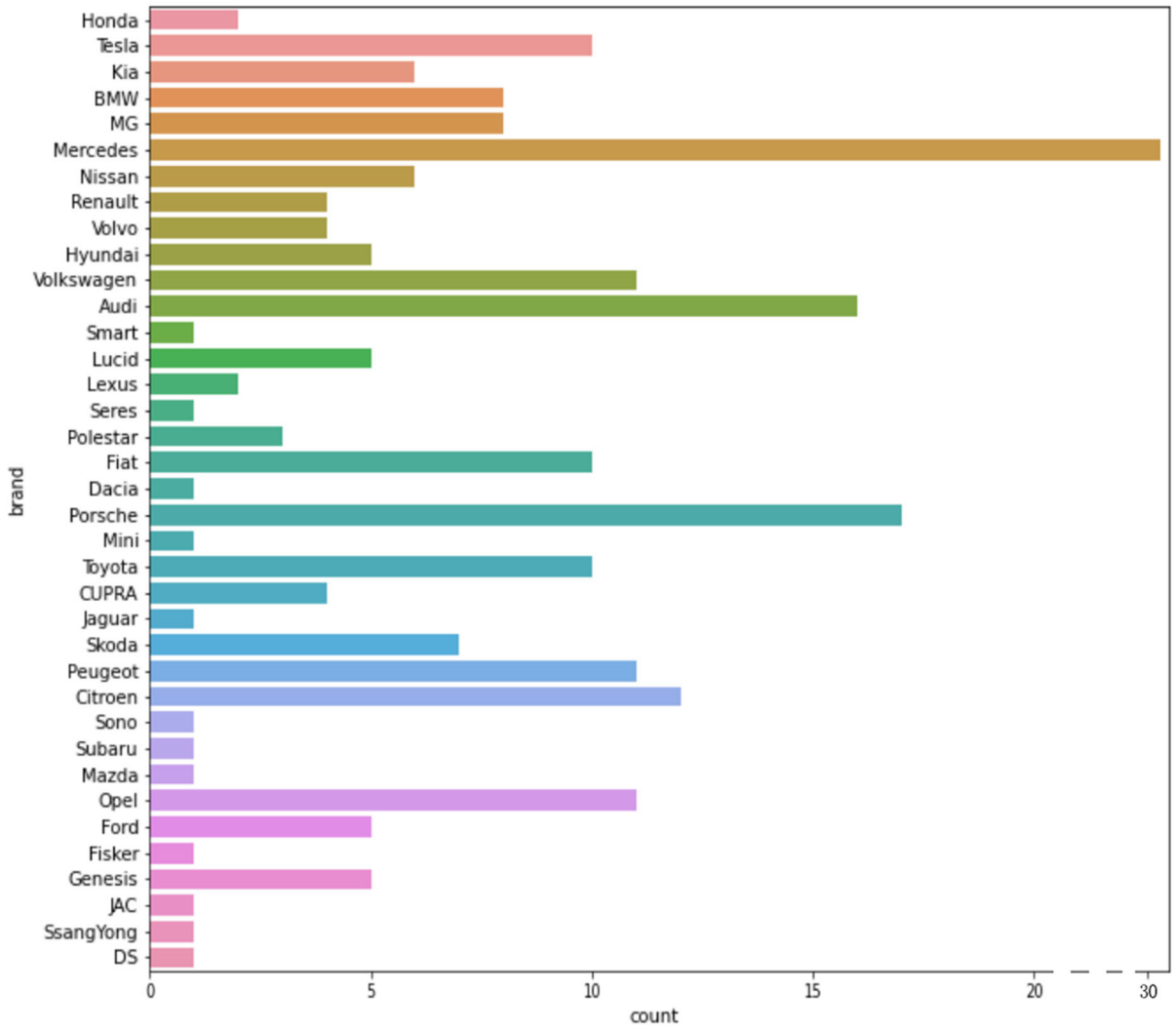

Fig. 3 Brand Distribution

Figure 4 includes fields which have large correlation with price and tries to decipher potential correlations between them.

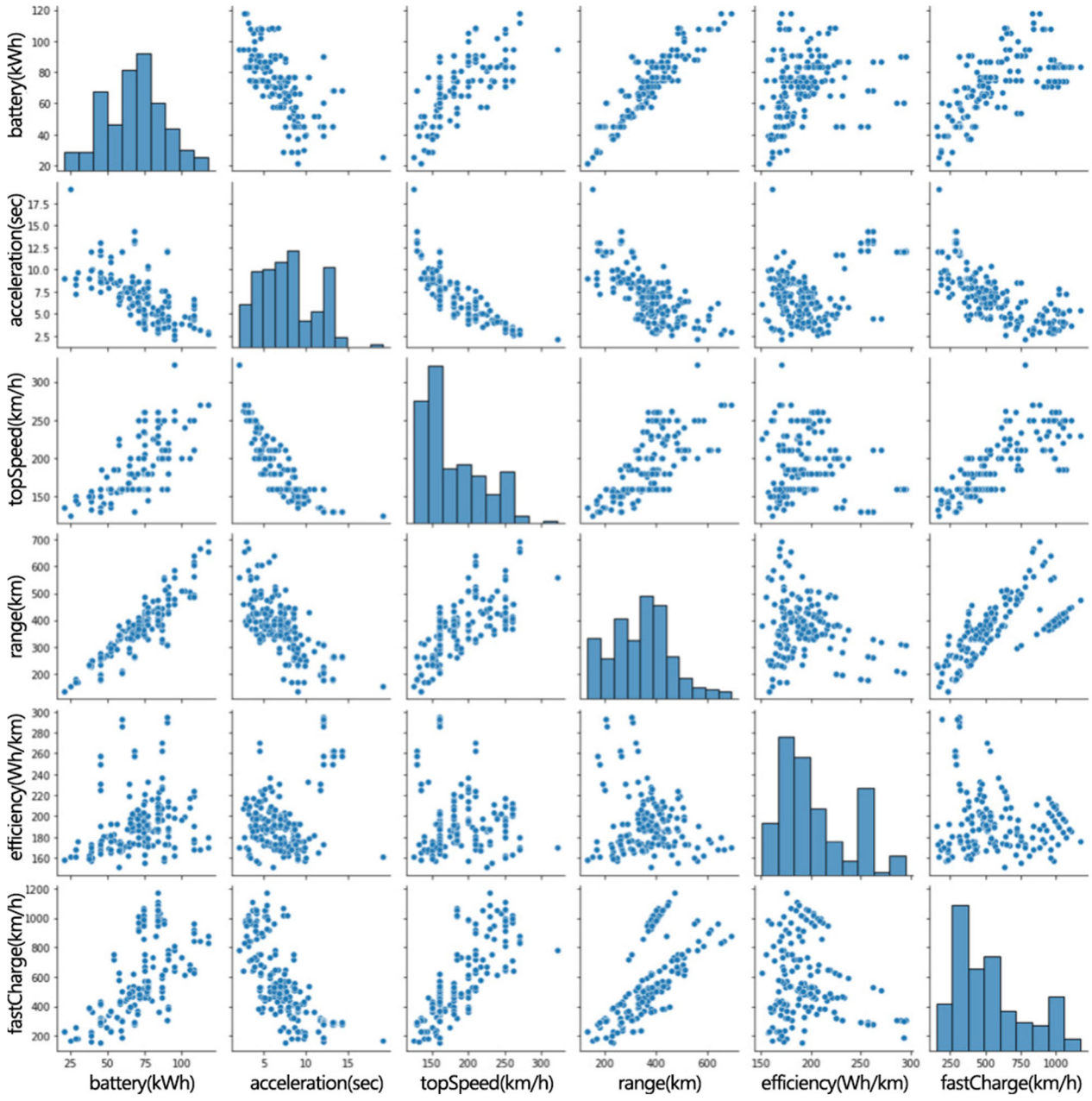

Fig. 4 Fields Pairplot

It surprisingly shows that battery(kWh) has positive linear correlation with range(km), which corresponds with the great magnitute(0.89), and it was positively correlated with most fields. For acceleration(sec), the pairplot shows observable negative correlation with fastCharge(km/h), range(km), topSpeed(km/h), and battery(kWh), while it was negatively correlated with most fields. Topspeed(km/h), range(km), and fastCharge(km/h), on the other hand, has macroscopic similar positive correlation with fields other than acceleration(sec). Topspeed(km/h) and range(km) has distinct positive correlation with fastcharge(km/h) and battery(kWh). Plots on efficiency(Wh/km) is messier that it does not show any distinct correlations with other fields, which implies an independence on efficiency(Wh/km).

3.2.2 Possible Correlation Between Numerical Fields

Numerical correlation values clearly show every possible correlation of each field in the dataset. Basic features of an EV do not show large correlation with any other performance features because theoretically each of the individual basic features are specifically designed to individual vehicles, which would not directly affect the price setting. Besides, according to the EV industry, basic features are easy to be duplicated while the technological features, which determines the key performances

are firms' core competitiveness. Since all performances are related with the core technology while producing, their firm correlations are truly presented on the heatmap in Figure 5.

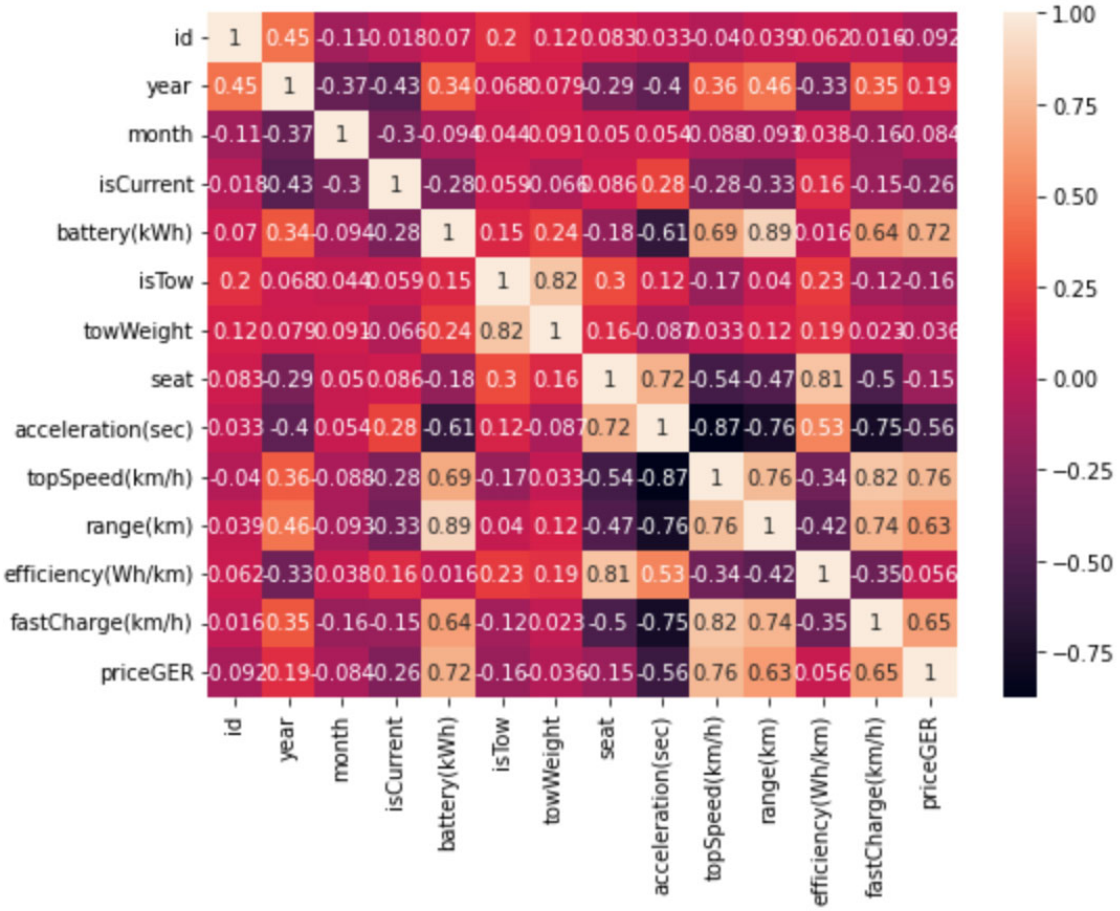

Fig. 5 Fields Numerical Correlations

89% proportion of range of the vehicle is correlated with battery capacity. Since battery creates power for the vehicles to run, the more battery a car can carry without stop to charge, the longer range the car is able to reach. Also, the low frequency a car is needed to charge also lowers the cost of refueling after purchasing it home, which undeniably has direct effect on the price of the car.

Battery is a key determination for most of the performance features that it not only shows similar pattern on the fields pairplot.

Acceleration(sec) and the efficiency(Wh/km) of a car both have high correlation with seat number, while they both have non-positive relationship with other technological feature fields. Figure 6 shows that the lower the seat number, the lowers the acceleration and the more efficient the car will be.

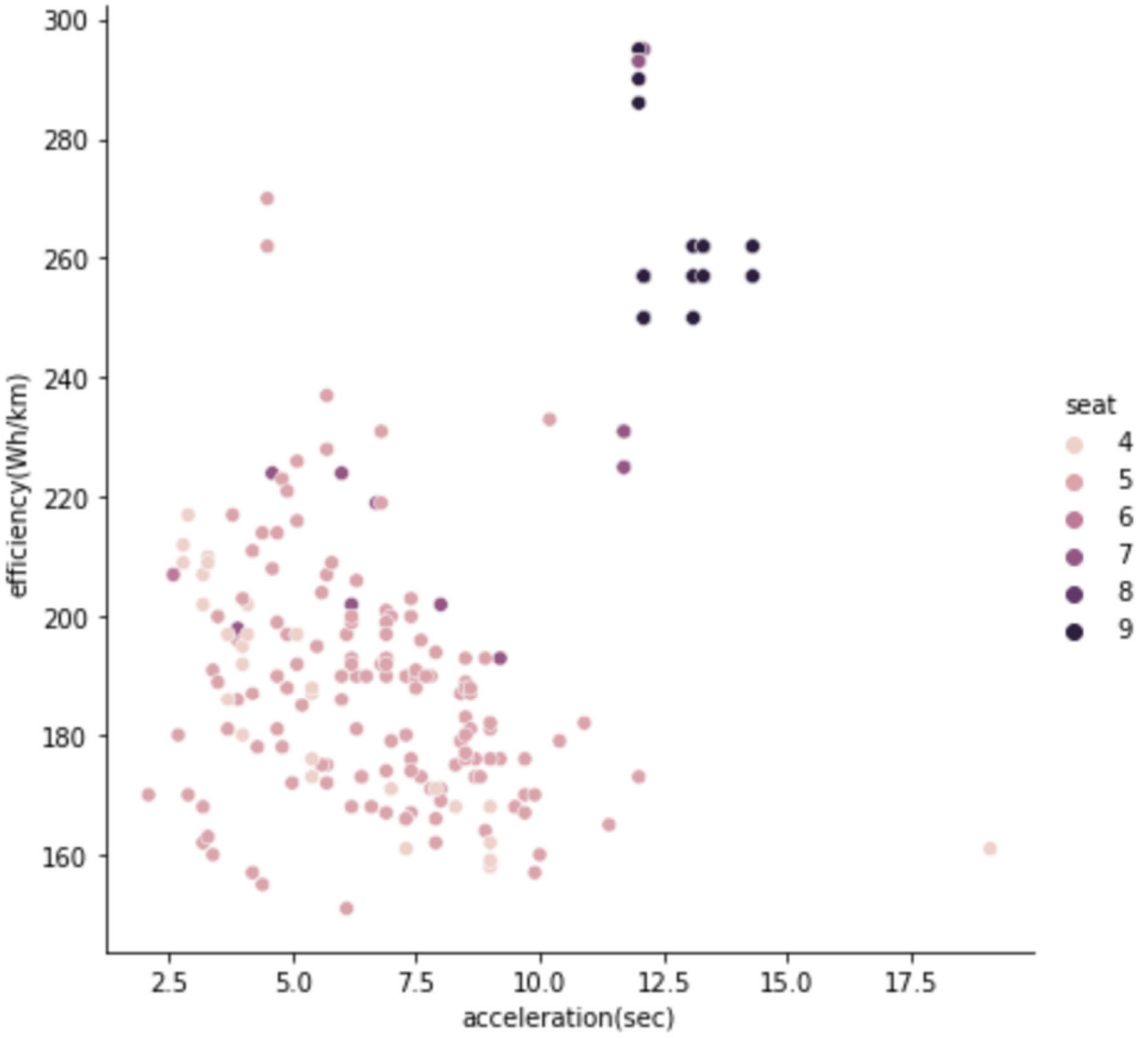

Fig. 6 Fields Numerical Correlations

Seat number determines the weight of a vehicle. Since the lower seat number decreases the overall weight of the car, a car can accelerate in a shorter time period. Efficiency is the measurement of how much electricity is converted to power rather than lost to heat. Najman implies efficiency of a car can be determined by the power of its engine, so if a car can perform better on acceleration, either benefited by its engine or weight, the less efficient it is [8]. Thus, once a car become much heavier, it can accelerate better and perform less efficiency.

Correlation does not show causal relationship between two variables, and it only gives a 2-d dimension layout of variables. Nevertheless, since the main research is about predicting price, simple correlation analysis would be more accessible for future work [9].

3.3 Regression Models

We use six types of regression models to analyze what factors determine the price of EVs in Germany, including Bayesian Linear, Boosted Decision Tree, Random Forest, Linear Regression, Neural Network, and Poisson Regression. In statistics, linear regression is a linear method that establishes a relationship between a dependent variable (Y) and one or more independent variables (X) by using a best-fit regression line. In addition, poisson regression models are best suited for modeling events where the outcome is a count. Moreover, Bayesian models probability distributions rather than point estimate to formulate linear regression. Moreover, a decision tree consists of a decision diagram and possible outcomes. At the same time, Boosted Decision Tree constructs a series of trees in a stepwise manner and then uses an arbitrarily differentiable loss function to select the optimal tree. Furthermore, a random forest is an algorithm consisting of many decision trees.

Likewise, neural networks can handle linear and nonlinear dependencies. Each model is classified into two datasets, including target encode and one hot encode. Finally, we will calculate and evaluate the MAE(Mean Absolute Error), RMSE(Root Mean Squared Error), RAE(Relative Absolute Error), RSE(Relative Squared Error), and R-Squared metrics to determine the optimal model.

3.4 Feature Importance Measure

The most accurate prediction model has been determined, and it is useful to analyze the components of the model to know what factors determine the price of EVs in Germany. This process also helps to improve the interpretability of the model. Permutation importance is an improved algorithm for evaluating feature importance. It disrupts a column order of the model, then uses the disrupted data to predict, and finally calculates the correct rate. If the correct rate decreases, it means that this variable is important. In addition, in this algorithm, to eliminate the effect of randomness on the results, it will permute the response vector several times and then find the mean and variance [10].

After using Permutation Importance to determine which factors are most important according to the scores of the variables in the results, we will also try to interpret the results by some social and economic background of German region.

4. Results

4.1 Correlation between fields and label

According to the correlation heatmap in Figure 7, identical or technological features, which are not easy to be duplicated by other firms, have a high correlation to price. They might not directly determine the final sale price of an EV, but they at least contribute to the composition of the sale price in the German market and play a critical rule on the predicting the price of EVs.

The number shows that size and brand have the highest correlation with price, which was not expected in previous general correlation analysis. The correlation value was unexpectedly up to 0.87% and 0.78%. Without the model, correlations can only be determined between existing numerical values drawn from the original dataset and could ignore significant variable relations. Despite from size and brand, topSpeed(km/h), battery(kWh), shape, fastCharge(km/h) and range(km) all exceeds 50% of correlation with price.

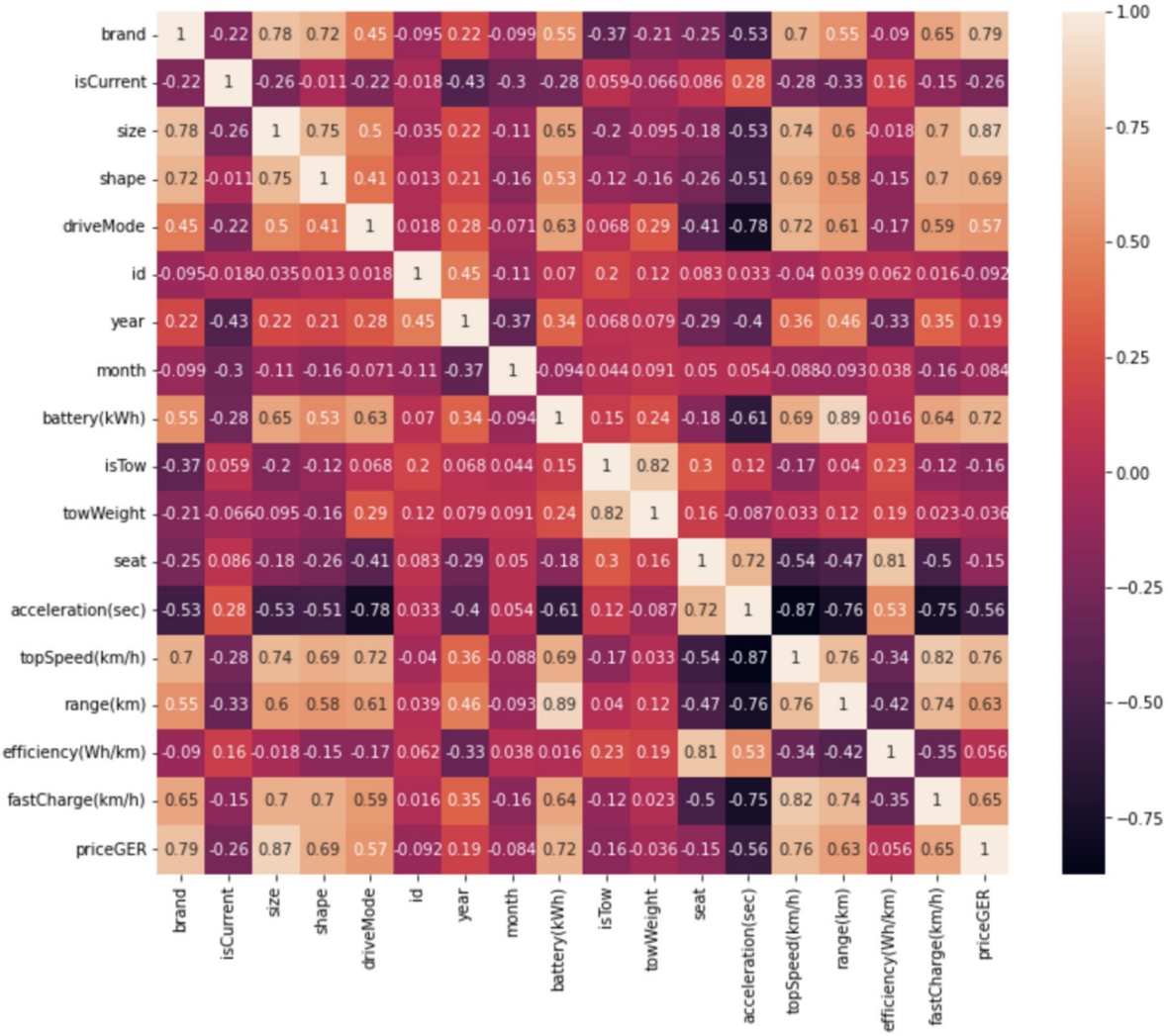

Fig. 7 Fields Correlation Heatmap After Target Encoding

Thus, the research will be taking the most correlated features as independent variable(X) to predict for price, which is the dependent variable(Y), after the above analysis.

4.2 Model Performances

This research chooses a target and one-hot encoding method to first modified the dataset. Target encoding is the easiest way of finding linear relationships, but there is the possibility of overfitting. One-hot encoding is straightforward, but it increases dimensionality and creates inconvenience. Thus, each regression model applied both encoders to find the most matching model.

The decision tree model is a long-used model which fits the surface of the data in binary classification, but then it improves into two models, that are boosted decision tree and random forest. Boosted decision tree is composited of many weak models and focuses on models which contain the minimum error, while a random forest is formed by many strong trees and averages or votes to get the final result. These two models are also the most debatable models among all models [11].

Thus, each regression model applied MAE, RMSE, RAE, RSE, and R-Squared to determine the accuracy and feasibility of each model, thereby finding the best model for price prediction. The model performance results are listed in Table 2.

Table 2. Model Performances

Regression Model	Dataset	MAE	RMSE	RAE	RSE	R-Squared
Bayesian Linear	target	9190.9743	13462.2919	0.4551	0.2548	0.7452
Bayesian Linear	one-hot	9283.0985	12675.1867	0.4596	0.2259	0.7741
Boosted Decision Tree	target	4722.8549	7509.9184	0.2338	0.0793	0.9207
Boosted Decision Tree	one-hot	4943.9819	7453.0311	0.2448	0.0781	0.9219
Random Forest	target	5930.3798	9252.8427	0.2936	0.1204	0.8796
Random Forest	one-hot	7555.2762	12551.5692	0.3741	0.2215	0.7785
Linear Regression	target	9182.9355	13459.4273	0.4547	0.2547	0.7453
Linear Regression	one-hot	9686.0375	12577.5997	0.4796	0.2224	0.7776
Neural Network	target	20181.3479	29540.0555	0.9993	1.2268	-0.2268
Neural Network	one-hot	23581.5214	34069.1559	1.1676	1.6318	-0.6318
Poisson Regression	target	10145.4532	14106.0866	0.5023	0.2797	0.7203
Poisson Regression	one-hot	10468.6697	15620.4181	0.5183	0.3430	0.6570

According to the data, the boosted decision tree has the lowest mean squared error, which means the difference between the predicted value and the actual value is the smallest. The boosted decision tree model also has the largest R-Squared value, which represents the proportion of the variance of the price that can be explained by the given features.

Figure 8 clearly presents the r-squared value of each regression model.

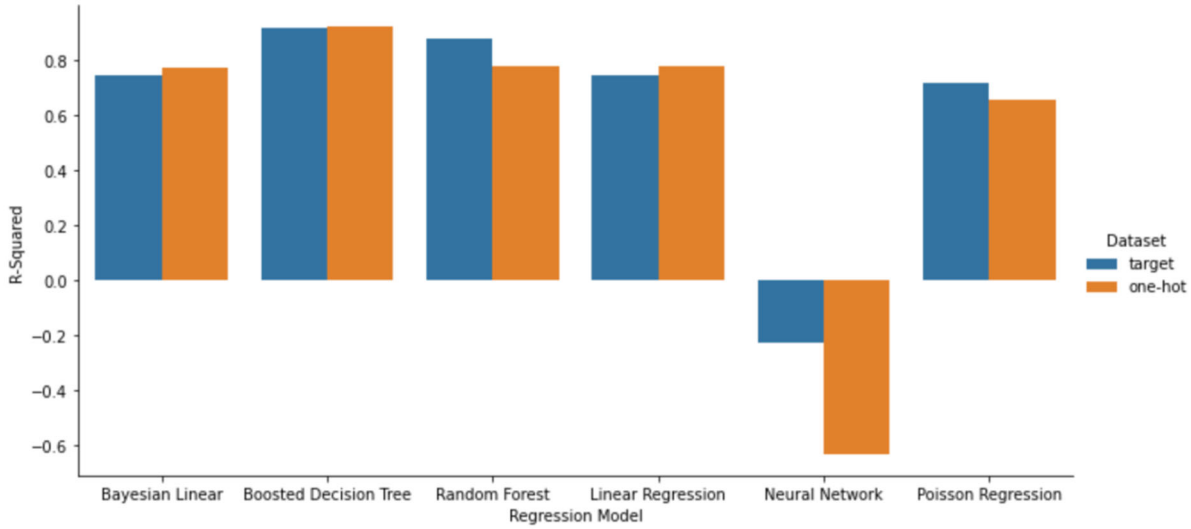

Fig. 8 Coefficient of Determination

Most model reaches a high r-squared value but boosted tree model still fits the test data the most since it has the highest r-squared. The figure also does not show much difference in using target or on-hot encoding, so choosing a regression model usually is more critical in creating a nice prediction model rather than spending too much time on figuring out the way of feature engineering.

4.3 Feature Importance

The models' training results are similar for both datasets, and the results are also similar when performing permutation feature importance. As shown in Table 3, a score was calculated for each feature, and a higher value means that the feature is more important according to this model.

Table 3. Feature Importance

Feature	score
size	0.322035
battery(kWh)	0.144694
acceleration(sec)	0.090508
efficiency(Wh/km)	0.058695
brand	0.030622
isCurrent	0.016118
seat	0.006093
fastCharge(km/h)	0.002624
topSpeed(km/h)	0.002486
isTow	0.00117
month	0.000293
year	-0.000425
towWeight	-0.000689
driveMode	-0.002099
shape	-0.00405
range(km)	-0.005159

In this experiment, size has the greatest impact on price. It is undoubtedly that the larger the model is, the higher the wheelbase and the higher the cost it requires. Next, the battery is also an important consideration. The biggest difference between electric cars and fuel or gas cars is the way the energy is provided, so it is understandable that the battery will have a greater impact on the price of the car. The following features are the car's 100 meter acceleration time, power consumption and brand.

5. Conclusion

This report uses publicly available databases to analyze the vehicle characteristics of European consumers purchasing EVs. The study concludes that consumers value the size of the car, the battery's durability, and acceleration. So we can believe that consumers are prioritizing car performance when it comes to purchasing cars. Although the popularity of EVs is still tiny compared to traditional fuel vehicles, Europe has been gaining momentum every year. Germany is a country with a high level of automotive development and a government that supports the development of EVs and consumer subsidies. We can believe that in the future, people can buy an electric car with good performance at a more affordable price.

Although we have re-collected more data and increased the number of fields, it is still not enough. First of all, electrical car is an emerging category of vehicle. There are far fewer types of EVs than fuel vehicles. Thus, it is difficult to get a large amount of records. Secondly, in addition to the parameters included in this data set, there are a number of factors that may also affect the price of the vehicle, such as horsepower and torque of the motor, battery type, etc. Therefore, more comprehensive experiments may be needed in future studies.

References

[1] Arababadi, A., Leyer, S., Hansen, J., & Arababadi, R. (2021). Characterizing the theory of spreading EVs in Luxembourg. Sustainability, 13(16), 9068.

[2] Luchs, S. , & Kasperk, G. . (2017). The influence of government incentives on electric vehicle adoption: cross-national comparison. MAGKS Papers on Economics.

[3] Ana Lucia Murillo. 2021. The Push for EVs Could Affect How Much Your Next Car Costs. https://money.com/electric-car-vs-gas-car-costs-biden/.

[4] CC Jel. Does purchase price affect preferences for electric vehicle characteristics and environmental benefits?.

[5] Gocompare. 2021. The future of electric vehicles: our predictions. https://www.gocompare.com/motoring/electric-cars/electric-car-adoption-prediction/.

[6] Ev-database Organization. 2022. Compare EVs - EV Database (Version: 4.3). https://ev-database.org/.

[7] Rodriguez, P. , Bautista, M. A. , Gonzalez, J. , & Escalera, S. . (2018). Beyond one-hot encoding: lower dimensional target embedding. Image and Vision Computing, 75(JUL.), 21-31.

[8] Liz Najman. 2021. Measuring Efficiency In Electric Cars. https://blog.wa.aaa.com/green-technology/measuring-efficiency-in-electric-cars/.

[9] Z. Zhang, K. T. McDonnell, E. Zadok and K. Mueller, "Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map," in IEEE Transactions on Visualization and Computer Graphics, vol. 21, no. 2, pp. 289-303, 1 Feb. 2015, doi: 10.1109/TVCG.2014.2350494.

[10] A André, Laura, T. , Oliver, S. , & Thomas, L. . Permutation importance: a corrected feature importance measure. Bioinformatics(10), 1340.

[11] Nabavi, & Jafari. (2013). Providing a customer churn prediction model using random forest technique. -, -(-), 202-207.