Drivers of Electric Vehicle Prices

Omosalewa Adebooye, Joshua Linner, Jasmine Sun

Aug 12, 2024

Table of Contents

OBJECTIVES

The objective of the analysis is to determine which variables are the drivers of the cost of electric vehicles (EV). We will be using the data set "Current and Upcoming Electric Vehicles" from EV Database to explore what influences electric car prices. For potential electric vehicle buyers, it is important to be informed when making a substantial purchase. The purpose of this analysis is to help inform buyers the different features that are important to the build of an electric vehicle. Features such as battery capacity, distance per single charge, the make of a car and so on can greatly impact the overall ownership experience. Furthermore, understanding how the features affect the costs of the vehicles help inform buyers of how to best budget their spending. On a larger scale, price prediction helps to inform important business decisions in budget planning and market analysis.

Goals

Our goal is to determine the features – such as acceleration, top speed, range, efficiency, seats, charge speed, usable battery, car make – that would highest impact the cost of the electrical vehicle prices. Furthermore, we would like to use a learning model to accurately predict electric vehicle prices based on given features and evaluate the predictions from the machine learning model.

Hypothesis

We believe acceleration, max speed, range, efficiency, seats, charging speed, and battery capacity are key drivers of electric car prices. We believe that the origin and the make of the car will have little influence on the EV price.

Further Questions

At the end of the analysis we aim to address some key questions:

- What are the most important features for determining EV prices?
- What are the most expensive EV makes?
- How many unique EV makes are in the dataset?
- Which makes have the highest battery capacity?

Our Approach

To achieve this, we will be using regression analysis, machine learning, and correlation plots to analyze the data and reach a conclusion.

DATA PREPARATION

Data Collection

Sourcing

We sourced all car data from an EV database (EV Database). Originally, we planned to extract data from a kaggle database, as stated in our proposal, but settled on the original EV database because there were more cars and features available to be extracted. We believed that this would allow us to achieve a more accurate result for a more detailed analysis and to make precise predictions. The Electric Vehicle database we chose aims to provide real-world information about electric vehicles. Since the organization is based in Europe, prices are listed in euros(€) and pounds(£).

Web Scraping and Extraction

Data was extracted from the website by web scraping all 407 electric vehicle types. Web scraping was done using the popular Beautiful Soup and Requests libraries. The Requests library was imported to request the source html page from the EV database website. The html file was then parsed into a beautiful soup object for smooth extraction. After studying the html structure, we found that each car type and its features were grouped into a div element with 'list-item' as its class value. So, we extracted only the div elements with class values of 'list-item' using beautiful soup's 'find_all' method. Furthermore, we wrote a 'scrape_div' method to extract desired features from each list-item div element. Thus, scrape_div method accepts a div element as its only parameter. The method's purpose is to assign feature values to key-value pairs in a dictionary. Scrape_div returns the resulting dictionary only. This dictionary is then appended to a list of similar dictionaries. Finally, we created our main dataframe from this list.

Data Quality

One challenge we encountered during web scraping was having to extract different features from different tags which happened to have the same tag name and class name (car make and model). To get around this, we stored both tags in a list called car_title, and extracted each value based on indices. Aside from this, all electric vehicle entries on the website's main page were extracted along with the important features we required for analysis. However, we had to take several steps to clean the data. We extracted 10 features in total and all values were object data types (see appendix). This concluded the web scraping portion of our project. Using the pandas.Dataframe info() method, we got an overview of the data and saw that there were several missing values, especially in the price columns. This was particularly important to our project because price is our dependent variable. We decided that the affected rows would not be removed. Instead, we replaced the null values with computed imputations. This way, we retained all our extracted data.

Data Cleaning

Removal of unit values

Some variables such as Efficiency, Acceleration, Range and MaxSpeed were extracted along with their units. These had to be removed before we proceeded with conversions. We removed

the suffix units using the string object's replace method. Likewise, the price variables: PriceUK, PriceG and PriceN were stripped of their currency symbols (€ and £). After this process, we were able to safely convert the numerical objects to integer or float data types as needed.

Conversion of data types

All data types were extracted as text values. Therefore, all numerical objects had to be converted to integer or float data types. We did this using pandas' to_numeric() method. Acceleration, Range, MaxSpeed, Efficiency, Seats, ChargingSpeed and Battery are the data frame columns we converted to numeric variables using pd.to_numeric().

Imputation of missing price values

Additionally, we faced a challenge with our dependent variable, price. Each of the three price columns extracted (Price in the UK, Price in Germany, and Price in the Netherlands) had null values because some price values were missing on the website during web scraping. However, after carefully studying the data, we saw that there was at least one price variable that was not null for each car type. We chose the Price in Germany to use as our dependent variable because it had the fewest null values compared to the Netherlands and UK prices. We imputed missing values using the fillna() method. We took the corresponding prices in the UK and Netherlands (whichever was available) and converted them to the price in Germany using a formula. This formula gave us a factor which we multiplied the prices with to get the price in Germany. The factors were calculated by taking the average percentage difference in prices between the locations and adding 1. This helped eliminate all missing values in the Germany Price column which would allow us to use this column as a target variable in our analysis.

Imputation of missing charging speed values

The charging speed column had one missing value which we chose to use the fillna() method to impute the mean charging speed.

Sorting the Data

The data was sorted in descending order by the Price in Germany column. This column is used as our target variable in the analysis so sorting by this column will keep the dataset more organized. This will also allow us to exclude any obvious outliers from our analysis.

Removal of outliers

After creating a histogram from the Price in Germany column we discovered the dataset to have a right-tailed distribution, as seen in Appendix Figure 2.0. Majority of the price values were between 25,000-100,000 euros and there were a few values that were much higher. We identified one major outlier which was a Rolls-Royce EV that was valued nearly 140,000 euros more than the next most expensive EV. We chose to eliminate this outlier from our analysis.

ANALYSIS

Dependent Variable

The dependent variable in the analysis is the PriceG (Price in Germany) variable. Missing values were imputed by converting UK and Netherlands prices.

Independent Variables

Independent variables Acceleration, MaxSpeed, Range, Efficiency, Seats, ChargingSpeed, and Battery are used in the machine learning model. The car make was excluded from the model because there were too many EV makers. Including car make in the model would not have added much to analysis because of the large variety.

Data Distribution of the Dependent Variable

The Price in Germany column was plotted on a histogram to determine the distribution of the values. This highlighted the right skewed distribution of the price values. Most EVs had a price clustered around the mean price, but there were some EVs with excessively high prices creating a right skewed data distribution.

Correlations

From our correlation analysis, the maximum speed of the car was shown to be the most positively correlated with PriceG at 0.75611 correlation value. The next two most positively correlated features were Battery and Range with 0.6872 and 0.5516 respectively. The efficiency of the car had the least impact with 0.3108 correlation value. Furthermore, acceleration had a negative correlation like we predicted at -0.5626. This is because the faster (in less time) a car can accelerate, the more expensive it will be.

High correlation between the variables - A positive correlation of 0.88 between Battery and Range. This can be attributed to their shared role in determining how long an EV can last on one charge. The results suggest that as the battery capacity increases, the driving range per charge also increases, which aligns with the general understanding that a longer battery life would be able to support longer driving distances in a single charge. This is supported by the plot in appendix Figure 3.1, where we can see a positive upwards trend as battery increases, so does range. We can also note from the plot that as Battery increases, there is a greater variation to how much Range increases. Later on in the report, it is noted that Battery has a high feature importance while Range does not.

Low correlation between Acceleration and Maxspeed - strong negative correlation of -0.82 between acceleration and maxspeed. While both influence the speed of the vehicle, they are important features for different uses. With a strong correlation, we can conclude that as the acceleration time (0 to 100km/h) decreases (faster acceleration), the maximum speed of the vehicle increases. This is supported by the plot in appendix Figure 3.2, where we see a negative downwards trend between the two variables.

Machine Learning Model

We used the RandomForestRegressor in our analysis to determine which variables were the most important for determining the price of an EV.

Selecting Variables

Our X-variables (independent variables) in the analysis were Acceleration, MaxSpeed, Range, Efficiency, Seats, ChargingSpeed, and Battery. Our y-variable (target variable) was the PriceG column. In the data cleaning process, we sorted the dataset in descending order by PriceG which allowed us to exclude the last row of the dataset (expensive Rolls-Royce EV).

Training the Model

After selecting the X and y variables, we used the train_test_split function to separate the dataset into X_train, X_test, y_train, and y_test. The RandomForestRegressor() function was then used to create a regression by fitting the X_train and y_train attributes. The X_test and y_test attributes were left out of the regression to be tested later.

Predicting Prices

After the model was trained, the RandomForestRegressor() was used to predict the outputs (prices) using the X_test variables. The RandomForestRegressor() used the X_test values to predict the y_test values. A new variable, y_pred, was created in this step to evaluate the accuracy of the predictions by comparing them with the actual y_test values.

Evaluating Predictions

The predictions were evaluated using the root mean square error and calculating the R-squared score. Root mean square error was calculated using the mean_square_error function from sklearn.metrics and the numpy.sqrt() function. The y_test and y_pred values were plugged into the formula to return a root mean square error. The standard deviation of the y_test values was also calculated to compare with the root mean square error of the model. The root mean square error from the model was smaller than the standard deviation which indicated a good model. The R-squared score was calculated using the r2_score function from sklearn.metrics. The R-squared of the model is just above 0.9 which indicates a high fit of the model. Given these evaluations, we can conclude that the machine learning model does a good job at predicting prices of EVs given the independent variables.

Model Concerns

The main concern with this model is the size of the dataset. The dataset has about 400 values which is not overly large. A larger dataset with more values would allow the machine learning to train using more values and create a more accurate model. However, the dataset has a high variety of makes which gives us more confidence in the accuracy of the model. Since there is a large variety of EV makers in the dataset, the bias should not be as high as if the model had a handful of EV makers. The dataset has EV prices from many different EV makers.

Feature Importance

The most important features used in the machine learning model are Battery, MaxSpeed, and Acceleration. Efficiency, Seats, ChargingSpeed, and Range have relatively small importance for predicting price.

Analysis of Results

Battery and Range seem to be connected because of their high correlation and similar traits which explain how long an EV can last on one charge. However, Battery is the most important feature in the machine learning model, while Range is the least important (Figure 4.0). Battery capacity seems to be the most important component of an electric vehicle because it gives the vehicle its power. It makes sense for a strong battery to reflect on the quality of the EV and its price. A high range might have an impact on the efficiency of a car but this could be a result of other variables such as smaller vehicles and vehicles with fewer features requiring minimal power. Even though Battery and Range have a very strong correlation, the machine learning model relies on battery capacity significantly more for predicting prices.

Analysis of EV Make

We chose to exclude the Make variable from the machine learning model because there were too many unique EV makers. Instead, we grouped the dataset by Make and ran further analysis. In total, there are 54 makers, with about 26% of those having made only one model. The average number of models under one maker is 7.5.

Most Expensive EV Makes

Rolls-Royce, Maserati, Porsche, Lotus, and Lucid are the most expensive makes in the dataset based on the mean price of each make.

Highest Battery Capacity

Lotus, Rolls-Royce, Cadillac, Voyah, and VinFast are the EV makers with the highest battery capacities. In addition, the correlation coefficient between price and battery capacity is 0.69. As shown, Lotus and Rolls-Royce, are in the top five for most expensive and highest battery capacity. This supports the results from the machine learning model which show the importance of the Battery variable for predicting price.

Highest Range

Lucid, Lotus, Porsche, Voyah, and Rolls-Royce are the EV makers with the highest range. In addition, the correlation coefficient between price and range is 0.55. As shown, Lucid, Lotus, Porsche, and Rolls-Royce, are in the top five for most expensive and highest range. This does not align with the results from the machine learning model which shows range as a relatively insignificant variable for predicting price. There is some correlation between range and price but the machine learning model relies more on battery capacity for predicting prices.

Most efficient EV makes

Hongqi, Maxus, Toyota, Jaguar, and Cadillac make the most efficient EVs based on watt-hours per kilometer. The correlation coefficient between price and car efficiency is 0.31. As shown, none of the makes with the highest efficiency are also in the top five for price. This supports the results from the machine learning model which show the insignificance of the Efficiency variable for predicting price.

CONCLUSIONS

This report aimed to identify the primary factors that drive the pricing of electric vehicles (EVs). Through comprehensive data collection, cleaning, and machine learning modeling, it was found that battery capacity, maximum speed, and acceleration are the most significant features influencing EV prices, while efficiency, seats, charging speed, and range have relatively less importance. Additionally, although the brand of the vehicle did not significantly contribute to the machine learning model, further analysis showed that brands like Rolls-Royce and Lotus tend to price their vehicles higher on average. This can be attributed to its higher battery capacity on average, which is found to be a major driver for EV prices. The study also revealed a strong correlation between battery capacity and driving range. However, the machine learning model emphasized the importance of battery capacity in determining vehicle price.

The conclusion of this analysis provides valuable insights for potential EV buyers, which may help them understand which features are most likely to impact the cost of a vehicle as well as what features to prioritize. By focusing on key attributes such as battery capacity and performance metrics, consumers can make more informed decisions when selecting an EV that fits their budget and needs.

APPENDIX

Figure 1.0: Data Fields

Names	**Description**
Make	The make of the electric vehicle.
Model	The model name of the electric vehicle.
Acceleration	The acceleration time from 0 to 100 kilometers per hour.
MaxSpeed	The maximum speed the vehicle can achieve kilometers per hour.
Range	The driving range of the vehicle on a single charge in kilometers.
Efficiency	The energy efficiency rating of the vehicle in watt-hours per kilometer (Wh/km).
Seats	The number of seats in the electric vehicle.
ChargingSpeed	The fast-charging capability of the vehicle in minutes for a certain charging percentage.
PriceG	The price of the electric vehicle in Germany (euros).
Battery	The capacity of the vehicle's battery in kilowatt-hours (kWh).

Figure 2.0 - Distribution of PriceG

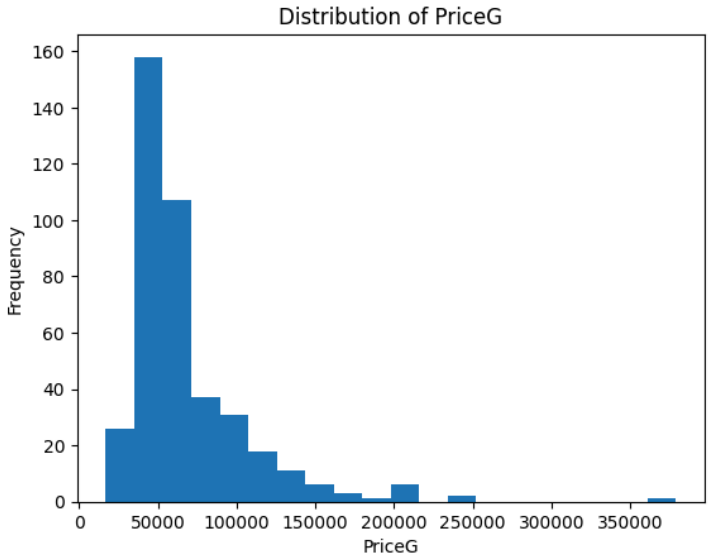


Figure 3.1 - Correlation Scatterplot between Battery and Range

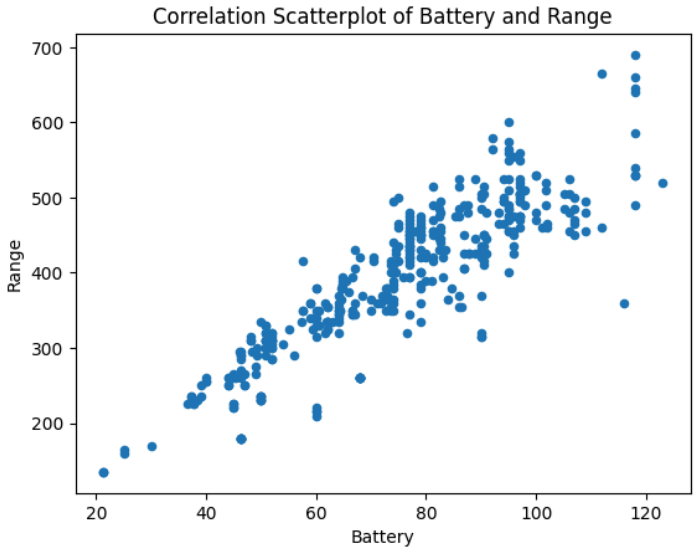


Figure 3.2 - Correlation Scatterplot between Acceleration and Maxspeed

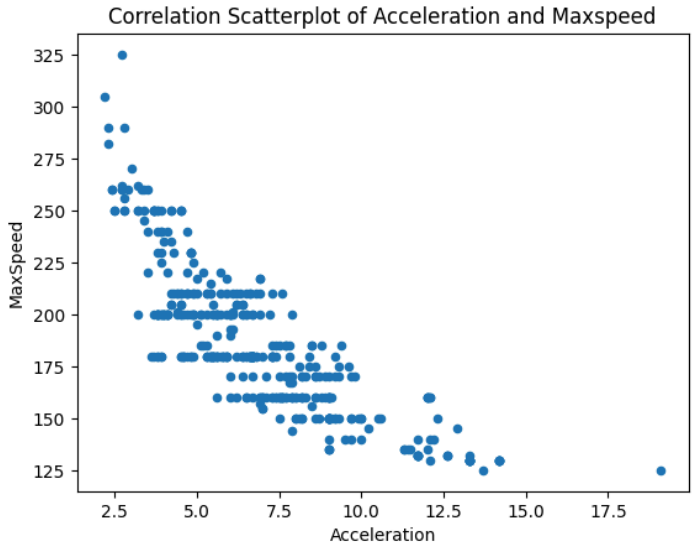
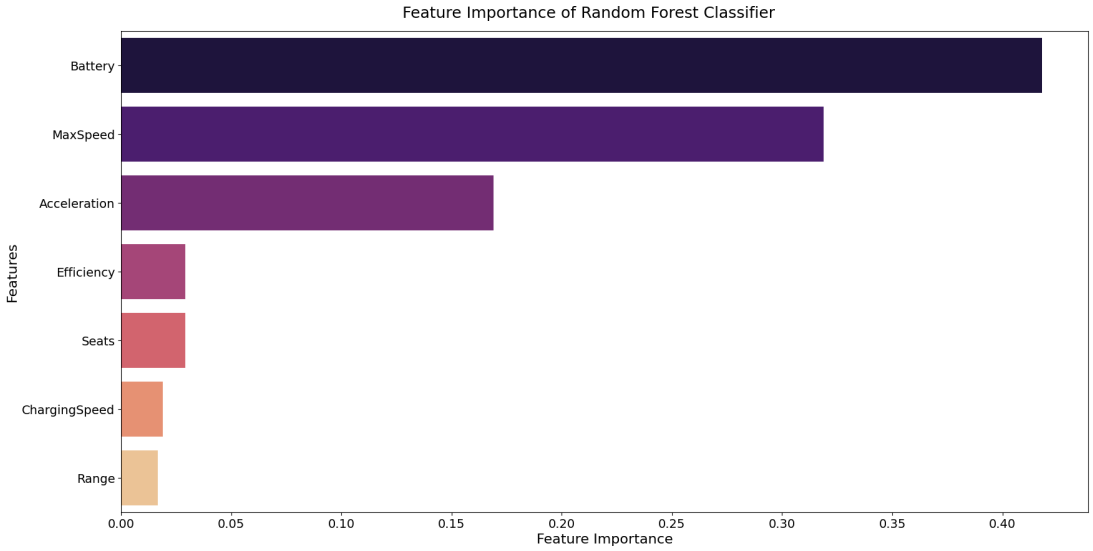


Figure 4.0 - Feature importance



CITATIONS

"Current and Upcoming Electric Vehicles." EVDB 4, 5, 2024, 41. Teleport Towers, https://ev-database.org/. Accessed 5 August 2024.