Forecasting Key Performance Metrics and Growth of Electric Vehicles Using Supervised Machine Learning

Mikayel Grigoryan (mikayelg@stanford.edu), Millie Pu (pml256@stanford.edu), Zhihao Zhang (Zihow Cheung) (fdpw14@stanford.edu)

Stanford University, Summer 2024
DATASCI112 Principles of Data Science
Dr. Alexander Dekhtyar (dekhtyar@calpoly.edu)

Introduction

The rise of Electric Vehicles (EVs) has had a pivotal role in the development of the automotive industry. From plug-in hybrid electric vehicles (PHEVs) all the way to purely battery electric vehicles (BEVs), these cars have revolutionized the way that we think about the automotive industry and vehicles as a whole. In fact, the number of electric vehicles in the U.S. alone accounted for 9.7% of all new EV registrations worldwide in 2022, with about 1.6 million EVs being sold in the U.S alone (Straughan, 2024). Inspired by this growth and innovation, we decided to choose a topic related to electric vehicles for our final project.

With the increasing popularity of electric vehicles worldwide, we were curious whether it would be possible to predict some key performance metrics of the electric vehicles based on a part of their parameters, such as their range, model, make. Additionally, inspired by the fast growth in EV sales, we were curious whether we would be able to accurately predict the number of cars sold in individual countries and around the world in 2030, given that we have collected the information about worldwide sales from 2011 to 2022. Finally, we were interested in whether we could accurately predict whether a given vehicle could be classified as a PHEV or a BEV based on the parameters that we provided.

To address all the questions we had, we created supervised machine learning models using Scikit-learn and Pandas and have done extensive data collection and preparation, cleaning and filtering, feature engineering, exploratory data analysis, data visualization. Additionally, we have used various techniques for performing our predictions such as linear regression and K-nearest neighbors regression and classification and based our results on comparing the mean squared errors (MSEs) and root mean square errors (RMSEs) for each of the models where appropriate. Finally, to make sure that our models are as accurate as possible, we have used two testing and training techniques including cross-validation and splitting our datasets into training and testing pairs for validating our prediction results.

Data Provenance

Our project utilizes several datasets, each contributing unique insights into electric vehicle (EV) trends and characteristics. Below is a detailed overview of the sources:

- **Washington State EV Registration Data**
  Data.gov - Electric Vehicle Population Data
  This dataset provides detailed records of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) currently registered with the Washington State Department of Licensing (DOL). It is one of the largest official EV datasets available in the United States, offering critical information on the distribution and adoption of EVs within the state.
- **EV Specifications and Pricing**
  Kaggle - Electric Vehicle Specifications and Prices
  Both of these datasets are derived from the ev-database.org website, providing comprehensive details on the specifications and pricing of various electric vehicle models. Given their similar structure and relatively small size, these datasets will be merged, with duplicate entries removed, to create a unified dataset for model training.
- **Historic EV Sales Data**
  Kaggle - Historic Sales of Electric Vehicles
  This dataset contains historical sales data for electric vehicles, providing valuable insights into market trends and consumer preferences over time.
- **Additional EV Population Data**
  Kaggle - Electric Vehicle Population
  This dataset offers another perspective on the distribution of electric vehicles, complementing the data from Washington State. It will be used in conjunction with the other datasets to enhance the robustness of our analysis.

We will primarily reuse the first dataset (Washington State EV Registration Data) for the analysis, while the remaining datasets will be integrated to enrich the overall dataset, ensuring comprehensive coverage of both vehicle specifications and market trends.

Analytical Question 1

Can we accurately predict the range of battery electric vehicles (BEVs) based on their production year and CAFV-eligibility in the state of Washington D.C?

We were interested in finding an answer to this question, because we wanted to get an idea of how technologies advance over the year and impact the range of the electric vehicles. During our research, we were able to discover a credible data source from the U.S. government which gave us precise data, including information about the **types** of the vehicles, **ranges** and other specifications, such as their **model** and **make**. The dataset can be found at https://catalog.data.gov/dataset/electric-vehicle-population-data.

Exploratory Data Analysis

To understand the data at a deeper level, we performed exploratory data analysis and were aiming to find some categorical columns as well as numerical columns, which would help us get a rough idea of how our data is structured. We discovered a few columns of interest including, ‘*Clean Alternative Fuel Vehicle (CAFV) Eligibility*’, ‘*Electric Vehicle Type*’, ‘*MSRP*’ and ‘*Year*’, with ‘MSRP’ and ‘Year’ being numerical variables. By analyzing the variable *Electric Vehicle Type* discovered that the data contained 2 types of electric vehicles, namely, **plug-in hybrid electric vehicles** (PHEVs) and **battery electric vehicles** (BEVs). This distinction between the two types was **crucial for the formulation of the analytical question** and the implementation of the final predictors.

Project Design

Before proceeding further with the EDA, we wanted to focus on data cleaning and filtering, and keeping only the columns that we were going to work with, as well as rename some of the columns and drop some of the rows which had invalid range values such as *NaN* or an empty string. After successfully accomplishing that, our final dataset had a structure similar to **figure 1.1**.

Figure 1.1

Year	Electric Vehicle Type	(CAFV) Eligibility	Range	MSRP
2000	BEV	YES	356	65000
2003	PHEV	NO	34	24000

After thoroughly filtering and cleaning our data, we ended up with ~90,000 rows from the original ~190,000 row dataset, which still was a significant number, considering the nature of our question that we were trying to give an answer to. For now, we will not be using the MSRP variable in any way, but for the upcoming question, we will be creating a subset which has valid MSRP values and will be trying to predict the MSRP of a vehicle, given its range, year and the rest of the parameters. We were able to successfully create a subset of ~90,000 records which did not contain MSRP information and ~3,000+ records which did contain such information.

Exploratory Data Analysis (Continued)

Our next step in understanding our data was to analyze the distribution of different categories of electric vehicles by their driving range. After our analysis and visualizations, we discovered that **BEV**s had higher median and mean ranges, equivalent to **215 miles** (median) and **161 miles** (mean), whereas **PHEV**s only got **30 miles** (median) and **30.86 miles** (mean). As an additional metric, we decided to analyze the distributions of the production years of these two types of electric vehicles, to further understand whether there are any trends in the increase and/or decrease of the driving range. In turn, this helped us discover that the average production year for the **BEV**s in our dataset was **2017** and for **PHEV**s the average production year was **2019**. To give a more in-depth view on these two metrics, we have made sure to provide two figures, **figure 1.2** and **figure 1.3** which we have created using the matplotlib library for plotting.

Figure 1.2

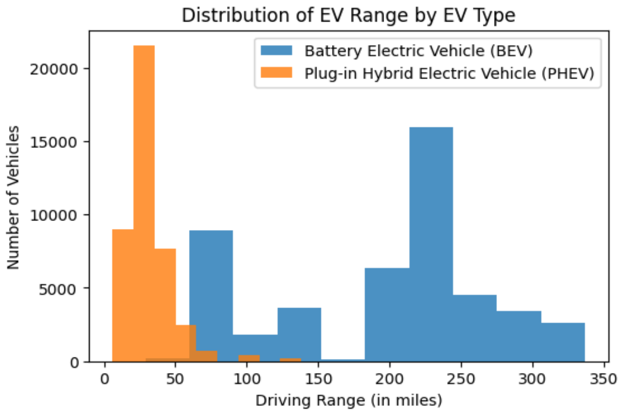
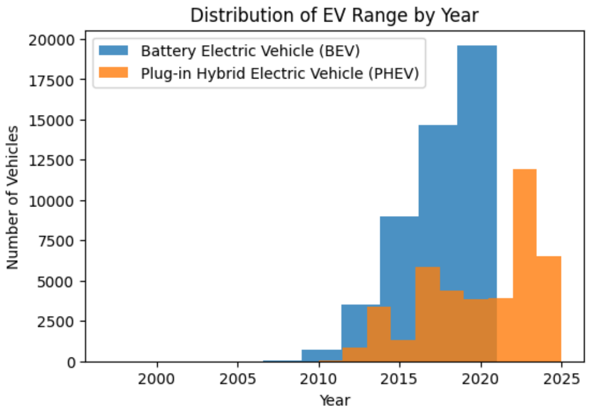


Figure 1.3



It becomes evident that the **BEV**s overall have a much bigger driving range than the **PHEV**s. To further prove this correlation, we tried analyzing the correlation coefficient between the production **years** and the **range** for both types of vehicles, using the formula for calculating the correlation coefficient as shown in **figure 1.4**.

Figure 1.4

$$\boxed{r = \frac{\mathrm{Cov}[X, Y]}{\mathrm{SD}[X]\mathrm{SD}[Y]}}$$

Figure taken from DATASCI112 lecture slides of Dr. Alexander Dekhtyar

After performing the required calculations, we discovered something interesting. It seemed like for **BEV**s there was a strong correlation coefficient and for **PHEV**s the correlation was negative and significantly weaker. We made sure to record the precise values in **figure 1.5**.

Figure 1.5

	Correlation Coefficient
BEV	0.7074844067512818
PHEV	-0.020361333597265118

Note: correlations were calculated based on the variables **'Year'** and **'Range'**

However, we have one additional column in our dataset, called '*Clean Alternative Fuel Vehicle (CAFV) Eligibility*'. Clean Alternative Fuel Vehicle (CAFV) eligibility typically refers to the qualifications or criteria that a vehicle must meet in order to be considered a clean alternative fuel vehicle. Eligibility for CAFVs can vary depending on the context, such as government incentives, tax credits, regulatory definitions, or programs aimed at promoting environmentally friendly transportation (RPubs - Electric Vehicle Data Analysis, 2023). Having this in mind, we speculated that there might be a relationship between the production years of cars marked as CAFV-eligible and their driving ranges. In **figure 1.6**, we have made sure to illustrate that this is indeed the case with vehicles which are not CAFV-eligible.

Figure 1.6

	Correlation Coefficient
Eligible	-0.05251869605597006
Ineligible	0.505698146962178

Note: correlations were calculated based on the variables **'Year'** and **'Range'**, based on CAFV eligibility (true/false)

After analyzing the coefficients, we came up with the following visualizations to illustrate the correlation of what was happening between the correlation of variables 'Electric Vehicle Type' and '*Clean Alternative Fuel Vehicle (CAFV) Eligibility*'. We rendered our results in **figure 1.7** and **figure 1.8**.

Figure 1.7

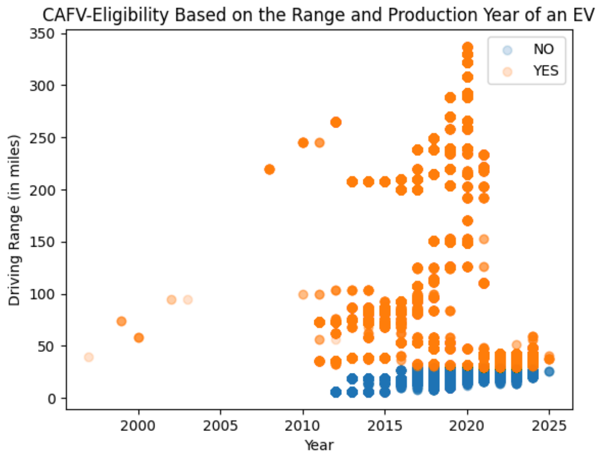
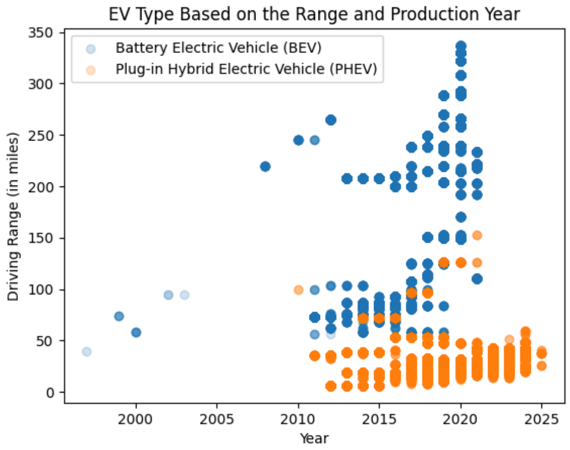
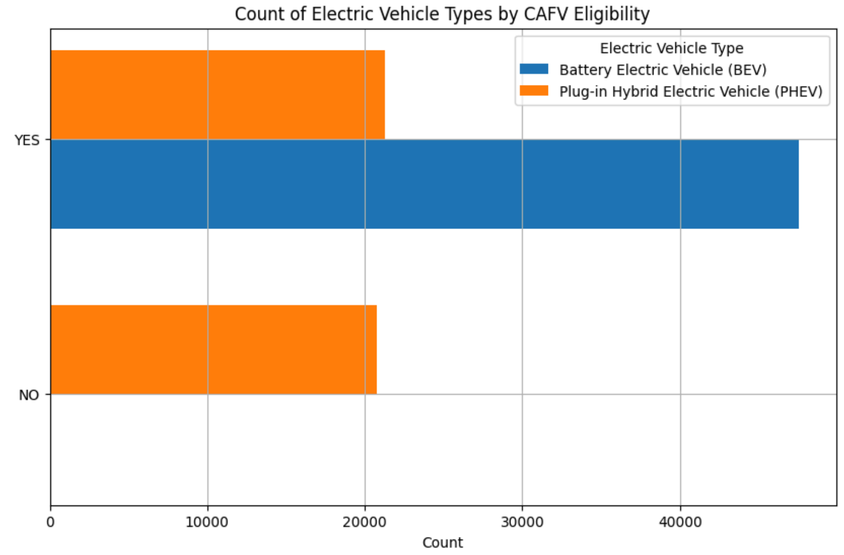


Figure 1.8



For us, it became evident that most **BEV**s fell under the category of CAFV-eligible vehicles, while for most of the **PHEV**s this was not the case. After collecting these metrics, we understood that we can use this data point as well for building our predictor models and making them as accurate as possible. Additionally, we made sure to calculate the number of both CAFV-eligible and non-eligible vehicles for both types of EVs, which is further illustrated in **figure 1.9**.

Figure 1.9



Implementation

With all this information in mind, we understood that using these 3 variables, including the CAFV-eligibility, the year, and the type of the car, to make our predictions would probably result in accurate predictions. Hence, we proceeded to the implementation part of our predictor. Our approach was to train 2 models, including a linear regression and KNN model, and determine their root mean squared errors (**RMSE**s) and mean squared errors (**MSE**s), to understand which model gave us the most accurate results.

Linear Regression

Additionally, we made sure to create a train-test split of our dataset to check the accuracies of actual predictions with ground truth. We used **40%** of our data for testing and the **60%** for training. After proper configuration, scaling, and fitting of the linear regression model, we ended up with a predictor which used linear regression and the error rates described in **figure 1.10**.

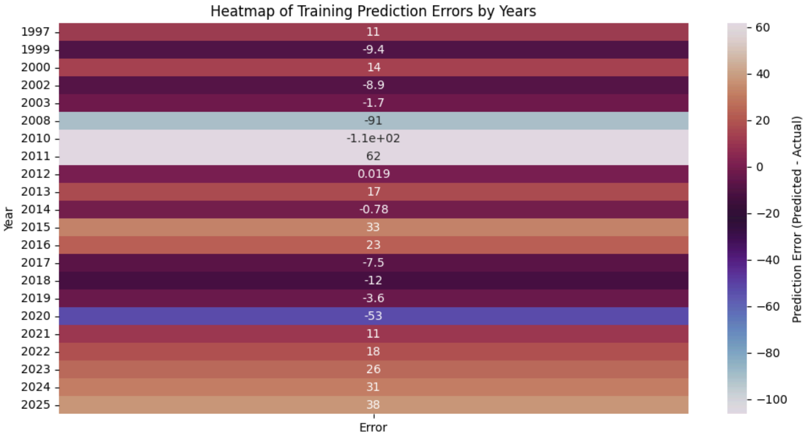
Figure 1.10

	Linear Regression *Training* Error Rate

RMSE	49.53073297484146
MSE	2453.2935090250476

As we can see from the table, our model has done relatively well. To illustrate the accuracy of our predictions, we made sure to create a heatmap visualizing the rates of errors per year, since visualizing the errors as scatterplots with a predictor line would have been hard to interpret, because of the density of our training and testing data. You can see what results we've achieved in **figure 1.11**.

Figure 1.11



It is evident that our predictor behaved well for most years, and had some difficulties with underpredicting for the years 2010 and 2008. Additionally, we made sure to check the testing error rates as well, and came up with the following error rates as shown in **figure 1.12**.

Figure 1.12

	Linear Regression *Testing* Error Rate
RMSE	49.63569331364943
MSE	2463.702050726663

The accuracy of the model is acceptable, however, we believed that we could do better, and for that reason started working on a separate predictor, which would use a KNN regression model underneath. To illustrate the testing error rate, we have created an additional **figure 1.13** in the form of a heatmap illustrated below.

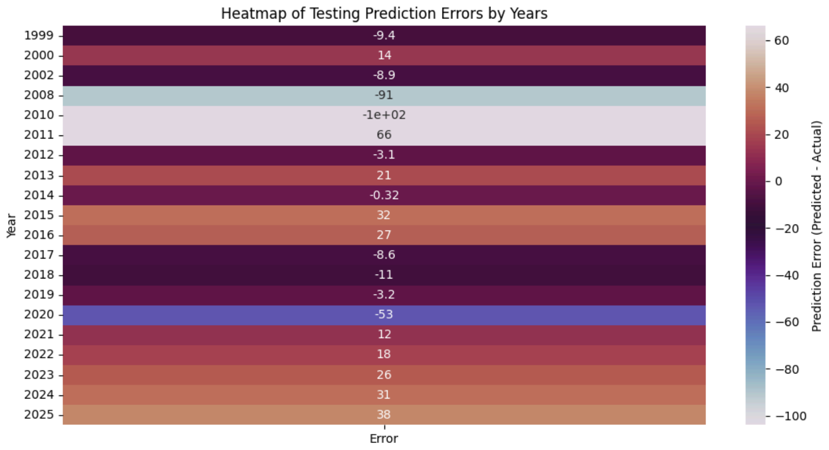


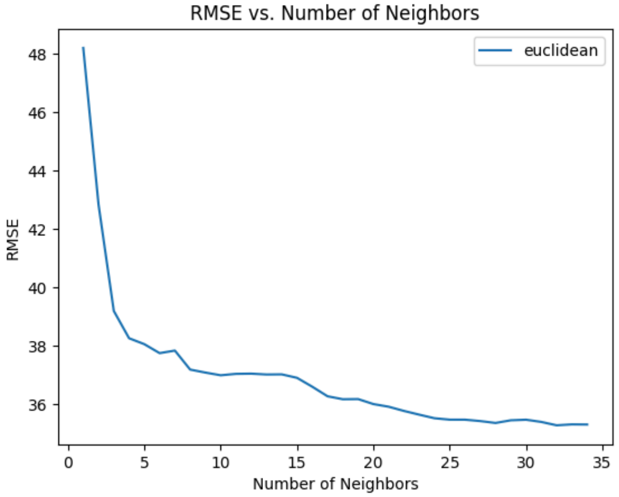
Figure 1.13

As we can see, our model outputs close to similar results when compared with our training outcomes. The model indeed had difficulties in predicting the ranges of cars manufactured in **2008** and **2010**. For the remaining years, the error margins were relatively acceptable.

KNN Regression

We knew we could do better, and so, we decided to train our KNN regression model based on the same train-test split. After performing the training, we discovered that the most accurate KNN model worked with **euclidean** as its distance metric and **33** was the ideal number of neighbors where the errors were present the least. We illustrated this correlation between the number of neighbors and the RMSE in **figure 1.14**.

Figure 1.14



After our GridSearchCV completed, it was evident that the most optimal number of neighbors (in our chosen range) was 32. Initially, this model took many hours to train. However, after switching from hosted Google Colab runtimes to a local containerized Docker runtime and setting the n_jobs variable to -1, we were able to leverage multithreaded processing, to get the predictions in less than 2 minutes. Additionally, we discovered that the euclidean metric was the most accurate one, and so we resorted to using 32 as the number of neighbors and euclidean distances as our main metric.

Results

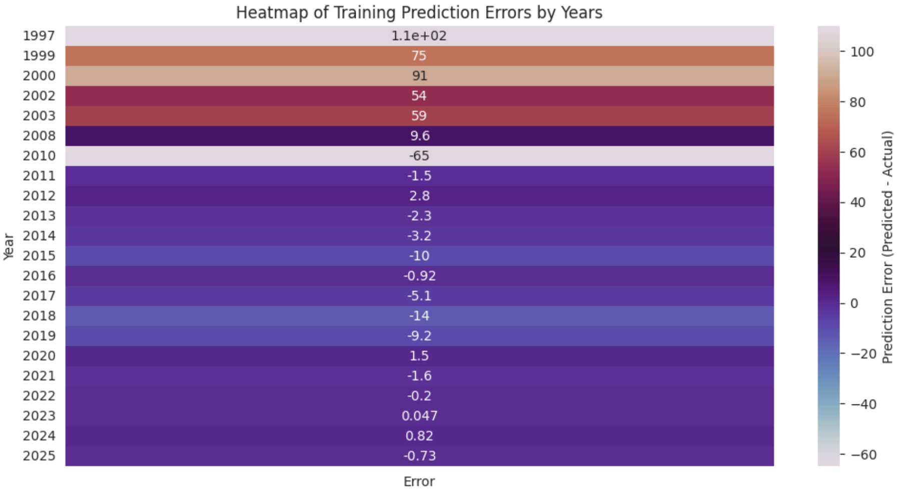
After running additional tests to verify the accuracy of our model for the training dataset, we indeed discovered that our KNN model showed a significant increase in accuracy in comparison to the linear regression model. To illustrate the differences between the training accuracies of our linear predictor and KNN predictor, we created **figure 1.15** and **figure 1.16**.

Figure 1.15

	Linear Regression	**KNN Regression**
RMSE	**49.53**073297484146	**35.52**410909683425
MSE	**2453.29**35090250476	**1261.96**2327123782

Note: Comparison of **Training** Accuracies for the Linear and KNN Regression Predictors

Figure 1.16



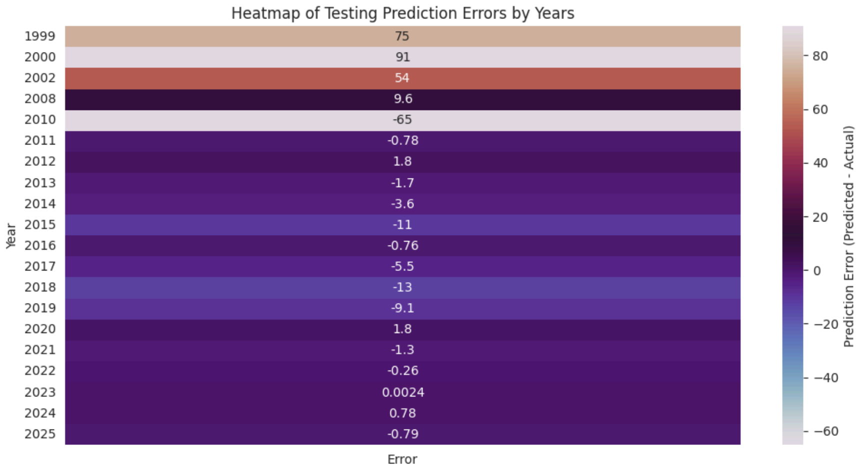
As an additional measure of verifying our results, we also made sure to compare the testing accuracies and report them accordingly in **figure 1.17**. We have used the same train-test split for this analysis as well.

Figure 1.17

	Linear Regression	**KNN Regression**
RMSE	**49.63**569331364943	**35.83**2822054413406
MSE	**2463.70**2050726663	**1283.99**1136383256

Note: Comparison of **Testing** Accuracies for the Linear and KNN Regression Predictors

Figure 1.18



It is evident that we were able to successfully create a reasonable predictor, which could predict the **range** of a potential battery electric vehicle (**BEV**), given its **year** and **CAFV-eligibility**. To further illustrate our results, we created **figure 1.18**, which indeed indicates that the model performs relatively well in most cases, except for cars of older years (which is usually less than **2010**).

Analytical Question 2

Question: Can we accurately predict an electric vehicle's efficiency (range) using parameters like acceleration, speed, and price, and would this be more accurate than the CAFV-eligibility and year-based model?

To predict the efficiency (range) of an electric vehicle based on factors acceleration, speed, and price, we decided to use the dataset with various vehicle attributes. Approaching involves Linear and KNN regression models, with hyperparameters fine-tuned using GridSearchCV and n-fold cross-validation for evaluation.

Exploratory Data Analysis

In our exploratory data analysis, we focused on identifying key variables that could influence the efficiency of electric vehicles (measured in Wh/km). We explored both categorical and numerical columns to understand the structure of our dataset better. Among the variables of interest were **‘AccelSec’, ‘TopSpeed_KmH’, ‘Range_Km’, ‘Price’, ‘FastCharge_KmH’, ‘RapidCharge’,** and **‘BodyStyle’**. By analyzing these variables, we aimed to predict the **efficiency (Efficiency_WhKm)** of electric vehicles.

Project Design

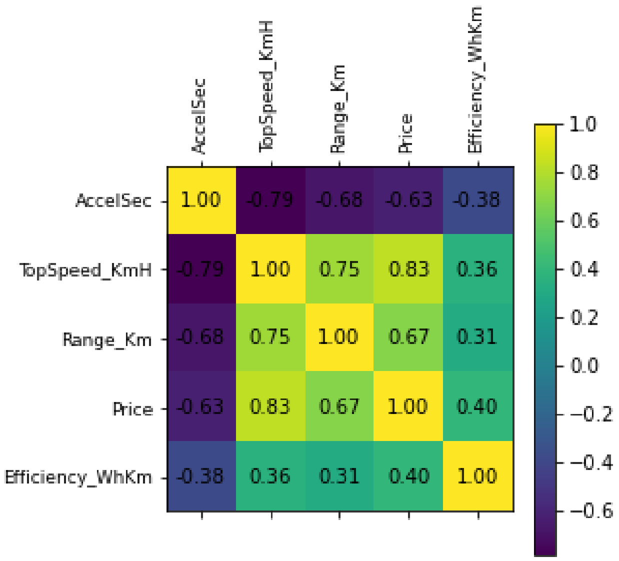
This project aims to explore and predict specific outcomes related to electric vehicles by leveraging a cleaned and structured dataset. The primary focus is on understanding the factors that influence vehicle efficiency and using these insights to develop predictive models. We filter out all the prices and after the cleaning process, the final dataset was structured with the following key columns in **Figure 2.1**

Figure 2.1

AccelSec	TopSpeed	Range	Price	FastCharge	RapidCharge	BodyStyle
4.6	233	450	55480	940	Yes	Sedan
10.0	160	270	30000	250	Yes	Hatchback
4.7	210	400	56440	620	Yes	Liftback

After cleaning the dataset and selecting key variables, we calculated the correlation matrix to understand the relationships between **AccelSec, TopSpeed, Range, Price, FastCharge, RapidCharge, BodyStyle,** and **Efficiency.** This matrix helped identify the strength and direction of the correlations between these features. Most of the indicators contain the correlation value around 0.3-0.4 but the AccelSec is -0.38.

Figure 2.2



Since the dataset that we were going to work with contained both, categorical and numerical variables with different units, we used the Standard Scaler and OneHotEncoder from the Scikit-learn package to process the data, the Standard Scaler is same as used in **Figure 1.3**

Implementation

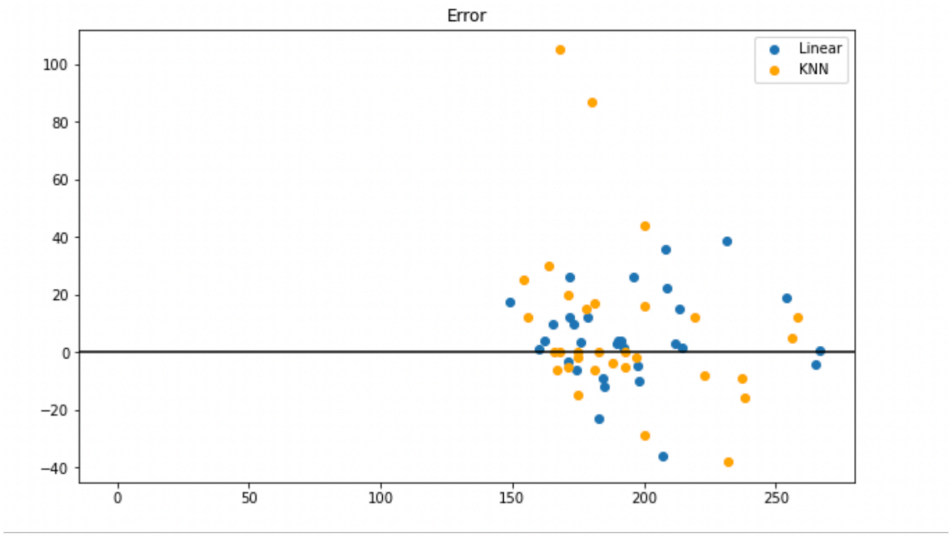
With all this information in mind, we understood that using these variables, including the AccelSec, TopSpeed, Range, Price, FastCharge, RapidCharge,BodyStyle, to make our predictions would probably result in accurate predictions. Hence, we proceeded to the implementation part of our predictor.Our approach was to train 2 models, including a linear regression and KNN model. We used Mean Absolute Error(MAE) and Root Mean Square Error(RMSE) to figure out which model is more accurate. We use 30% and 70% on split test trains. Figure 2.8 is the MAE and RMSE in **figure 2.8**.

Figure 2.8

	MAE	**RMSE**
KNN regression	17.580645161290324	29.36917409626587
Linear Regression	12.22849533442446	16.440309671728016

Residuals represent the difference between the observed values and the values predicted by a model. In the context of a graph, residuals are often plotted to assess the model's performance. And the Figure 2.3 is the residuals of KNN regression and linear regression. The Y-axis means the errors, the X-axis represent the original predict. The KNN regression has two main outliers in the **figure 2.3**.

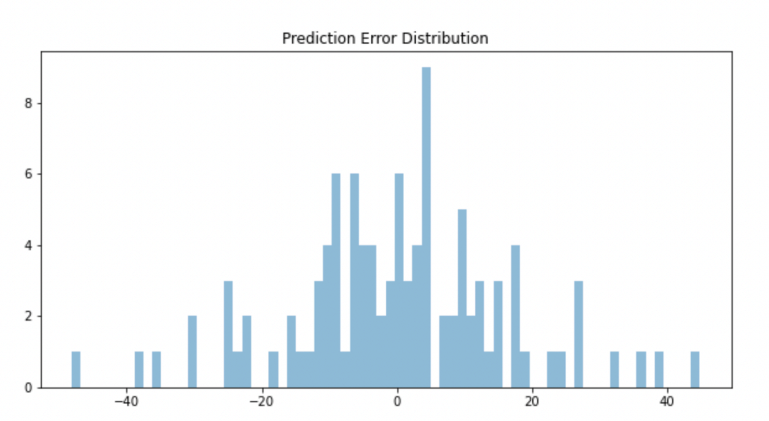
Figure 2.3



Linear Regression

An important aspect of assessing our linear regression model is analyzing the distribution of residuals, which represent the differences between observed values and those predicted by the model.The X-axis represents the residuals and y-axis is the sum of the error in different values. The error most focus on around the zero and some small number farther.

Figure 2.4



Also, the r^2 score is 0.759201393379708, indicating that 75.9% of the variability in efficiency is explained by the independent variables included in the model. This suggests that the model has a fairly good fit, capturing a substantial portion of the variation in efficiency. However, the remaining 24% of the variability is not accounted for, which could be due to factors not included in the model or inherent randomness in the data.

KNN Regression

As we can see, our model outputs close to similar results when compared with our training outcomes. And we tried to use KNN regression to predict with the same train-test split. After performing the training, we discovered that the most accurate KNN model worked with the Manhattan distance metric and the first neighbor was the best number. Please refer to **figure 2.5** and **figure 2.6**.

Figure 2.5

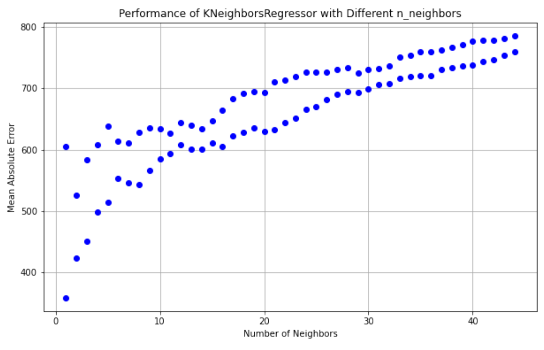


Figure 2.6

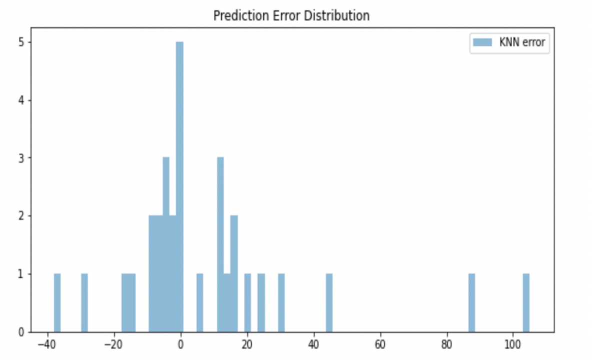
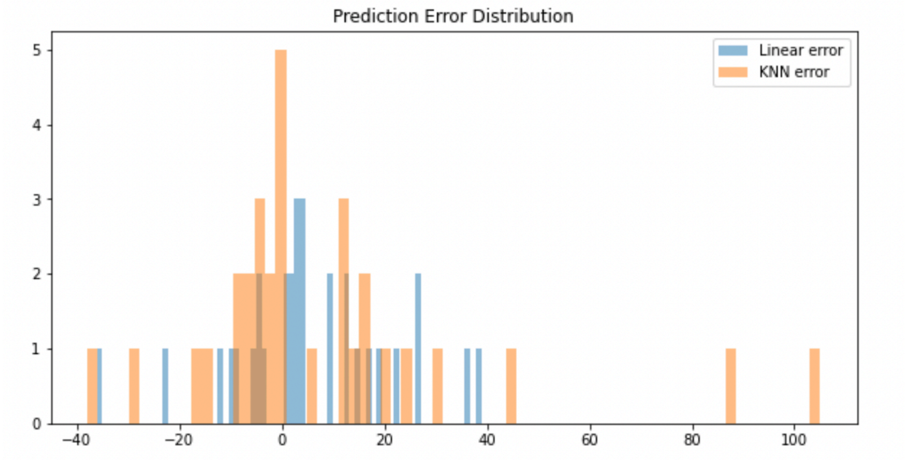


Figure 2.7

Result

In summary, we predict the efficiency of different types of cars by using its accelerations, top speed, range, price, fastcharge, and rapid charge and body style. The linear regression represents a smaller error. And to predict the value, it get better score and other.



Analytical Question 3

Can we accurately predict the number of EV sales in the US in 2020 based on the previous historic worldwide sales data of EVs?

This dataset contains all the information about the sales history from 2011 to 2022 in different countries. This question would like to only explore the changes in U.S. electric vehicles from 2011 to 2022 and predict its sales in 2020. Linear regression would be a rational approach to solve this project question, particularly for identifying and quantifying the trend over time. It can be applied to calculate the average annual growth rate in sales, providing a straightforward method to project future sales based on historical information. The year is the independent variable, while the sales of EVs are the dependent variable. The linear regression for the USA is y=31836.82x-64056272.76 , which suggests that one more increase in a year would be associated with an increase of 31836.82 in the sales of EVs in the USA (Non-causal statement).

Project Design, Implementation, and Results – Linear Regression

The scatter diagram shows that the sales of EVs in the USA has a strong positive correlation with 2017. The growth rate of EVs sales has grown slowly before 2016, the sales of Battery Electric Vehicles and Plug-in Hybrid Electric Vehicle sales are roughly the same. While the sales of EVs increased significantly after 2017, particularly the sales of Battery Electric Vehicles have substantially exceeded the sales of Plug-in Hybrid Electric Vehicles. According to Gessaroli (2024), the Federal Government offers a tax credit of up to $7,500 to consumers who purchase battery electric vehicles, which reduces the cost of purchasing a vehicle. The tax credits are not high when they purchase Plug-in Hybrid Electric Vehicles, thus PHEVs become less attractive over time. It is important to mention that an improvement in battery allows a longer range of BEVs. Modern lithium-ion batteries not only increase the longer driving distances but also improve the speeds of charging, which releases the anxiety of the range of EVs (Ambrose and O'Dea, 2021). Whereas PHEVs usually have smaller capacities of battery, which limits the range. The technological breakthrough in quick-acting charging has improved the efficiency of charging, which reduced the waiting time. The expansion of facilities of quick-acting charging is enhancing the convenience of BEVs. See **figure 3.1** and **figure 3.2** below.

Figure 3.1

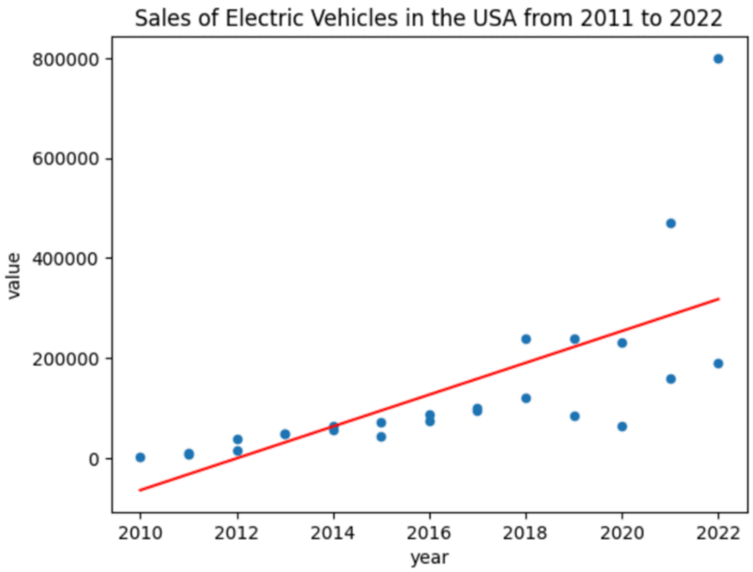
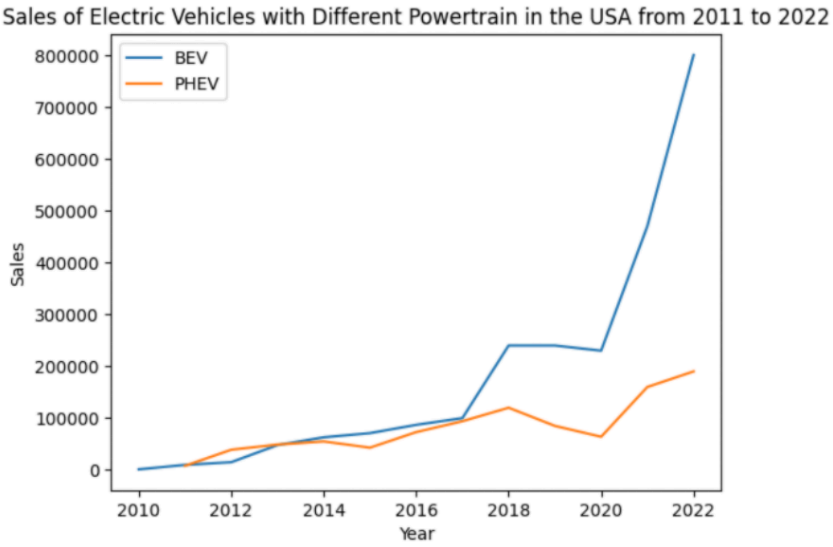


Figure 3.1



The predicted sales in the US in 2020 is 253906 based on the linear regression model, in comparison of the actual sales of 294000. It would be better to measure the error between actual sales and predicted sales, which is to ensure the accuracy of prediction. This includes some basic metrics, such as R-squared, mean squared error, and mean absolute error. The R-squared is 0.46, which indicates that 46% of the variance in the sales of EVs in the USA can be explained by the year in the linear regression model. *year* has a large effect on the sales of EVs, while it is not the only regressor. However, this would be sufficient to identify the upward trend in the sales of EVs, due to the simplicity of the linear model. *year*, as the only predictor in current linear regression, would not capture all the complexities. Many factors, such as policies, and technological development, can affect the relationship between the sales of EVs and year, while these factors are difficult to be quantified. Therefore, it suggests that the linear model is useful, but it also needs to be improved by adding more factors in the further analysis. The mean squared error is 15496672561.10. The data of sales of EVs is large, especially the measurement is united in total value. Therefore, the value of MSE would be large, it is reasonable in the current dataset.

KNN Regression

The KNN model predicts that the sales of EVs in the US in 2020 is 201800. It is noticeable that the predicted sales of EVs based on KNN model is different from those based on linear regression (253906). The KNN model uses the information from the nearest neighboring points to make predictions, which means that the prediction for a given point would be affected by the values of nearby points. The KNN model predicts the value based on the similarities between points, it does not assume any linear relationships and could capture nonlinear relationships by observing the neighborhood of each data point. The MSE of the KNN model is based on the average of the squared differences between the actual value and the prediction of the model.

A lower MSE indicates that the model's prediction could be more accurate. The below table shows that the MSE of KNN is significantly lower than the MSE of the Dummy model, which means that the prediction of the KNN model is better than the simple baseline model. Therefore, KNN models would be more effective in capturing the complex relationships in the dataset. The mean squared error for k=2, which is 47579058000.0, is lower than for other k values. This indicates that k=2 provides the most accurate prediction of sales in 2020, since it minimizes the difference between the predicted sales and actual sales. When k = 2, the predicted sales would be less fluctuating than other predictions under different k values, which ensures the accuracy of the prediction. Refer to **figure 3.3** for more information.

Target Country	United States of America
Linear Regression	y=31836.82x-64056272.76
Actual sales in 2020	294000
Linear Predicted sales for 2020	253906
KNN predicted sales for 2020	201800
Sum of Squared Errors	387416814027.43
Mean Absolute Error	77894.19
R-squared	0.46
Mean Squared Error (MSE)	15496672561.10
KNN model MSE	79635834960.0
Dummy Model MSE	101962796900.0

Results

In summary, the analysis acquired two different predicted sales of EVs in the USA in 2020. Linear regression predicts that the sales in 2020 is 253906, while KNN model predicts that the sales in 2020 is 201800. Both linear regression and KNN model are close to the actual value in 2020. For future analysis, it would be better to add more regressors so that the analysis could get more accurate predicted sales in 2020.

Analytical Question 4

Based on the **range**, **base MSRP**, **top speed** and **year** of an electric vehicle, can we predict the **price**?

For this task, we wanted to understand whether we could predict the type of the vehicle based on its parameters. Since our dataset provided very limited information about our actual car, we decided to scrape the information from the links that it provided to the car datasets and persisted them into our original dataframe, which we then later saved into a separate file, to save precious time. For scraping, we used the BeautifulSoup package alongside the requests library and set a random sleeping interval ranging from 2-8 seconds, so that our scraper was able to go undetected. We were scraping the information from the footers of https://ev-database.org/, where we had the information about the types, which corresponded to one of the values Battery Electric Vehicle (BEV) or Plug-In electric Vehicle. After scraping the data and constantly getting rate limited by the servers, we understood that the data contained only battery electric vehicles, and so, we decided to shift the focus of the question a bit from predicting the type, to predicting the price of a given **BEV**.

Exploratory Data Analysis

We initially tried to understand whether there is any correlation between the acceleration, top speed, range, efficiency, and the year to the price of the vehicle. To understand this, we computed the correlation coefficients and indeed found out some interesting things, which are illustrated in **figure 4.1**.

Figure 4.1

Battery	0.702036
Efficiency	0.174505
Fast_charge	0.618951

Range	0.589158
Top_speed	0.760579
Acceleration (0-100)	-0.532272

Note, the data is in regards to the actual **price** of the vehicle.

We were actually surprised to find out that there is a weak to no correlation of the efficiency and the acceleration to the actual price of the electric vehicle, however, this information gave us a path to constructing our predictors as accurately as possible. After reading this table, we decided to omit the Acceleration and Efficiency columns, and only keep the ones which had a correlation coefficient of more than or equal to 50%, to make our predictor as reliable as possible.

Project Design

After properly cleaning our data and applying some feature engineering, we ended up with a dataset with the following structure for creating our Linear and KNN Regression predictors. Refer to **figure 4.2**. Additionally, it is worth noting that the price is indicated in Euros and not USD.

Figure 4.2

Battery	Efficiency	Fast Charge	Range	Top Speed	Acceleration	Price
75.0	172	670.0	435	217	5.0	59017.0

Implementation

For this task, we once again decided to use KNN regression coupled with Linear Regression, since we couldn’t pass through the rate limitations of the remote servers which kept the data about the types of EVs for classification.

Linear Regression

To assert the accuracy of our predictions we resorted to using **RMSE** and **MSE** as the measures of performance for our models. We first of all created and trained the linear regression model and calculated the RMSE and MSE for both training and testing. For training, since our dataset is very scarce, we used cross-validation training, with 20 folds and for testing we selected 100 random samples and tried to make predictions for them. The errors are reported in **figure 4.3**.

Figure 4.3 (Linear Regression Errors)

	RMSE	**MSE**
Training Error Rate	**19746.19**9128725617	**389912380.03**128433
Testing Error Rate	**16588.56**2337235835	**275180400.41**63592

It is evident that our predictor is not as efficient as it could've been. This could be due to various factors, but we are guessing that this is because of the scarcity of our dataset, which is only about 300 rows in size. Additionally, we have discovered that our linear regression predictor underpredicts most of the times. We have illustrated this in **figure 4.4**. Another reason for this might be the presence of outliers, for example, the point which has a price of 200,000 Euros.

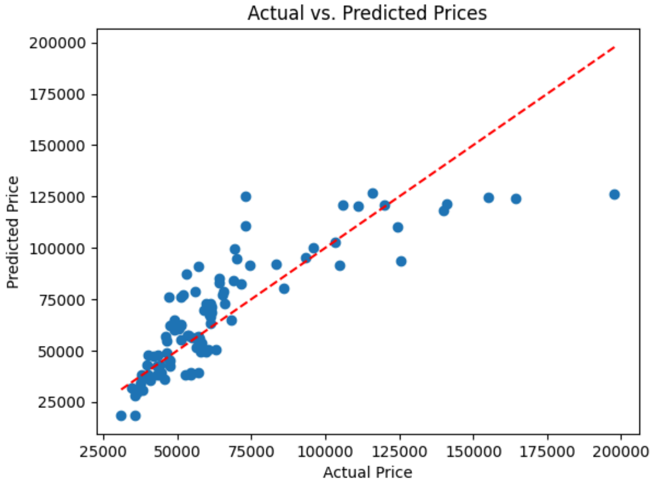


Figure 4.4

KNN Regression

For KNN regression, we once again used GridSearchCV with folds of 20, to evaluate our model on our scarce data. We discovered that our KNN regression model performed significantly better than our linear regression model and had overall smaller RMSEs and MSEs. For reference, you can compare the error rates of linear regression with **figure 4.5**.

Figure 4.5 (KNN Regression Errors)

	RMSE	**MSE**
Training Error Rate	**15499.34**8578706516	**240229806.36**42517
Testing Error Rate	**13785.91**8646054204	**190051552.91**5625

As a note, the actual training errors when running our code locally might vary, because in our implementation we have decided to sample on 200 random rows from our dataset. So it is expected for the model to perform even better and get RMSEs ranging from ~8,000 to 15,000.

This is the usual range that our accuracies were based in. After our analysis, we discovered that the optimal KNN regression model worked with Manhattan as its distance metric and 4 as the number of neighbors. We have made sure to illustrate the improvements in performance of our KNN model in **figure 4.6**.

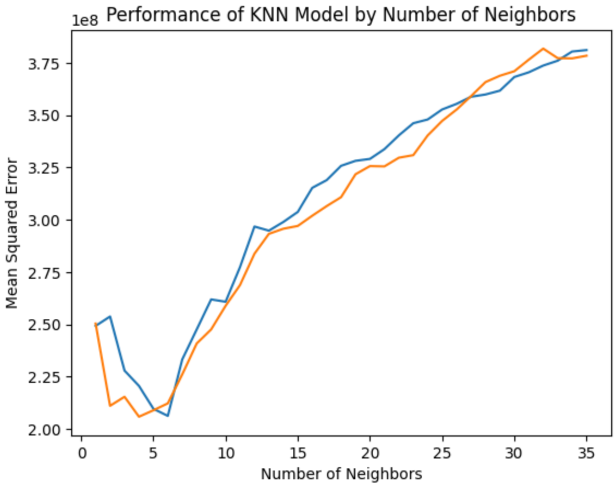


Figure 4.6 (Performance of the KNN Regression Model)

Results

We believe that with our very scarce data, we were able to train a model with moderate performance given our data constraints. The model does not perform too well, and usually underpredicts, but considering the fact that we are dealing with prices and that the error range is usually from 13K-15K, we believe that this model should be acceptable when applying it in non-production/non-critical contexts. If one was to couple it with some additional finetuning of returning the predicted prince - 12K, they might have ended up with an actually accurate price for the vehicle.

Conclusion

To conclude, we believe that we were able to successfully create accurate predictors for a few purposes. Firstly, we created an accurate predictor, which would help us to predict the **range** of a potential battery EV (**BEV**) given its **CAFV-eligibility** and the **year of production**. Secondly, we were able to create an accurate predictor for helping us predict the **range** of an electric vehicle given parameters like its **acceleration**, **top speed**, **drivetrain type**, etc. Thirdly, we were able to develop a strong predictor which would predict the number of sales of EVs in the U.S. based on historical data. Finally, we developed a classifier which would help us predict the price of an electric vehicle in Euros. After conducting extensive analysis, we believe that our results (except for the last model) will be able to be used in real-time scenarios as well and that the accuracies of most of our predictors are within an acceptable range.

As a final note, we would like to say thank you from the name of our team to our professor, Dr. Alexander Dekhtyar for providing us the necessary knowledge, resources, support and tooling for understanding the world of Data Science, and all of our TAs, who supported our journey during this quarter at Stanford. It was our pleasure to work and learn with you!

Best regards,
Mikayel Grigoryan, Millie Pu, Zhihao Zhang.

Reference lists

RPubs - Electric Vehicle Data Analysis. (2023, December 5). Rpubs.com. https://rpubs.com/Anoop-S-Hari/1124971#:~:text=Eligibility%20for%20CAFVs%20can%20vary,at%20promoting%20environmentally%20friendly%20transportation.

Straughan, D. (2024, June 7). Electric Vehicle Statistics 2024. https://www.marketwatch.com/guides/insurance-services/electric-vehicle-statistics-2024/#:~:text=accommodate%20market%20demands.-,How%20Many%20Electric%20Vehicles%20Are%20Sold%20in%20the%20U.S.%3F,EV%20 registrations%20 worldwide%20in%202022.

GESSAROLI, J. (2024). Electric vehicles are promising but still evolving. In A BUMPY ROAD AHEAD: A Critical assessment of Canada's Electric Vehicle Availability Standard (pp. 24–29). Macdonald-Laurier Institute. http://www.jstor.org/stable/resrep59865.9

Ambrose, H., & O'Dea, J. (2021). Electric Vehicle Batteries: Addressing Questions about Critical Materials and Recycling. Union of Concerned Scientists. http://www.jstor.org/stable/resrep29545