Team 5: U.S. Electronic Vehicle Sales and Technical Paper Citations Analysis

UH SPE Machine Learning Bootcamp First Project: Non-Linear Modelling

Dung Nguyen, Xu Yang, Charles White, Jose Benavides

1. Project Introduction:

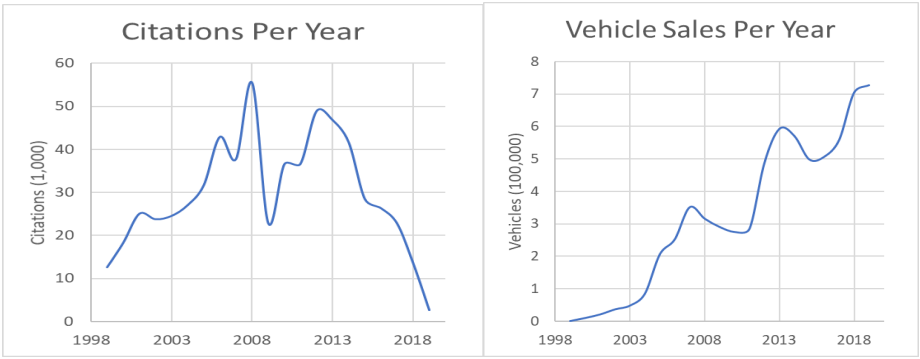
Sales of electric, hybrid electric, and plugin hybrid vehicles have increased dramatically in the last decade. This is due to multiple advances in lithium ion battery technologies. These technological advances have their roots in technical papers that are written for multiple industry journals. These papers represent the origin of the technologies that will later appear in the vehicles that are being sold. By establishing a metric for understanding the impact that a technical paper has, a prediction can be made as to the number of electronic vehicles that will be sold in the coming years after the paper is released. Understanding future vehicle sales will allow vehicle manufacturers to know the number of vehicles that will need to be produced in a given year.

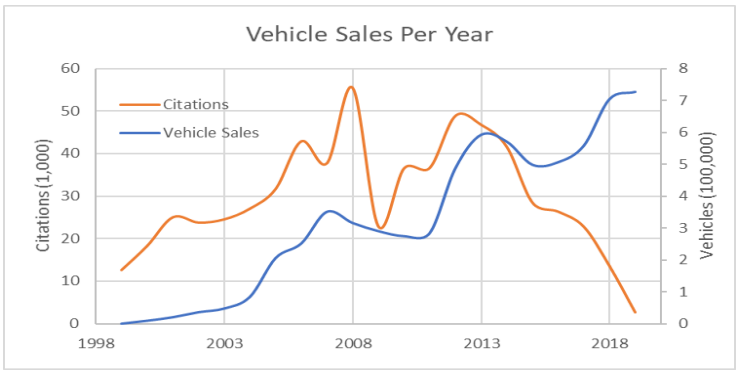
2. Dataset

The dataset for our independent variable will be the time series of technical publications that are related to lithium ion batteries technologies. This will be done by scraping down to a local machine by using Publish or Perish 7 software. The dataset for our dependent variable will be the time series of include electric, hybrid electric, and plugin hybrid vehicles which will be gathered from AFDC.Gov.

3. Features and Pre-Processing:

a. Features





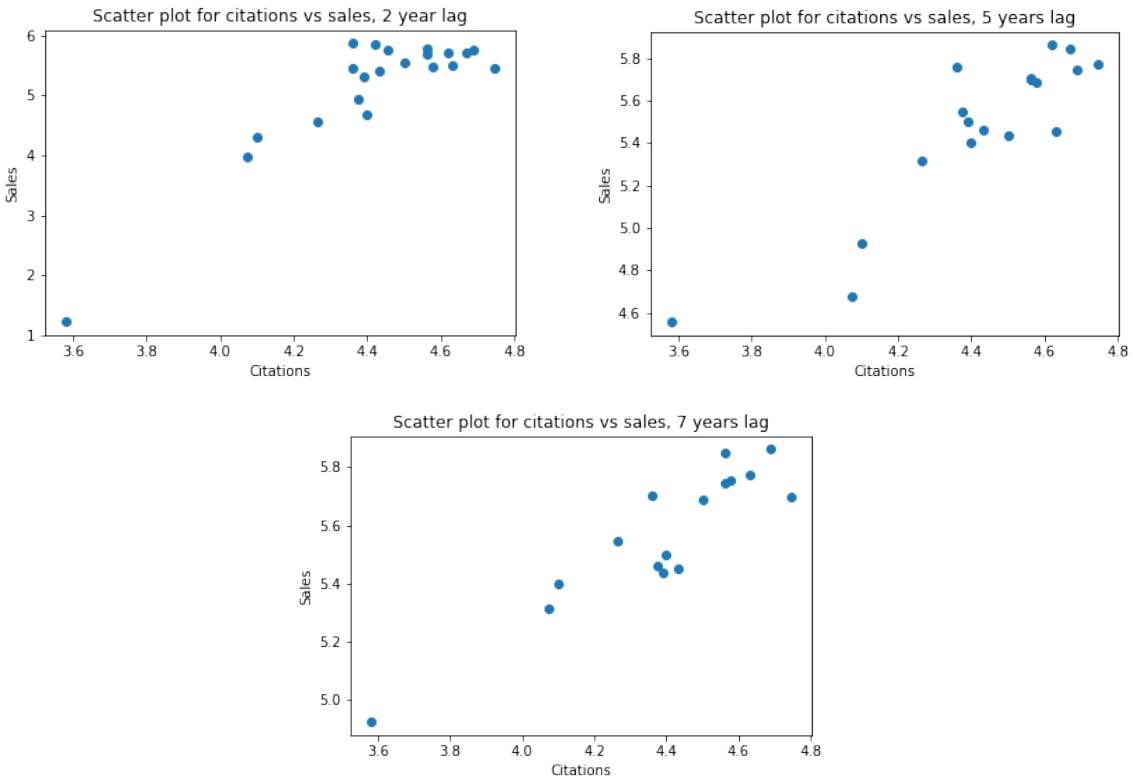
b. Pre-Processing

Once all data is gathered it will be evaluated and all unnecessary data columns will be dropped according to domain knowledge. The organized columns will contain keywords containing Booleans to ask as separate feature vectors. Because 2020 would represent an incomplete year without a full accounting of citations and vehicle sales, it is not used in our data. This provides 21 years of data to draw conclusions from. The titles of the citated paper will be searched for specific key words and all combined into one dataframe. Finally, a log of the entire dataframe will be taken to put the numbers into a scale, and that processed data will be used in further analysis.

4. Model and Techniques:

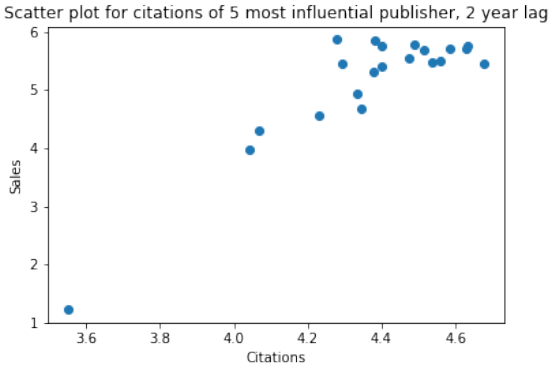
a. Data visualization

Because the publishing of technical papers will have a delayed effect on the sales of vehicles, multiple models are created to determine the best lag time to predict vehicle sales.

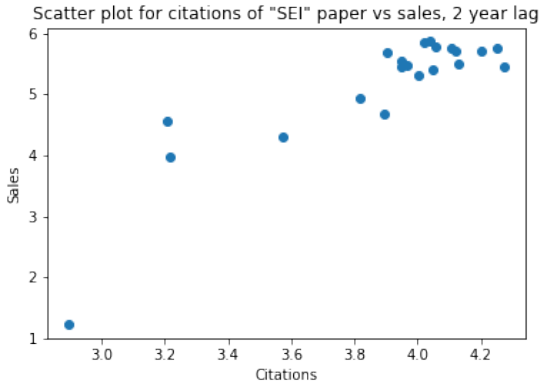


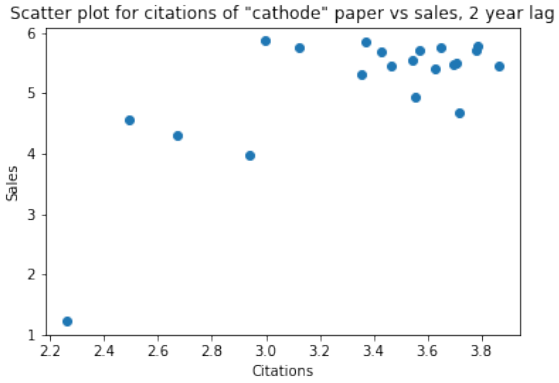
These models show that there is a seven-year lag between the publishing of technical papers and electric vehicle sales. This allows for an accurate prediction of the sales of electric vehicle so that manufacturers of these vehicles can plan production accordingly.

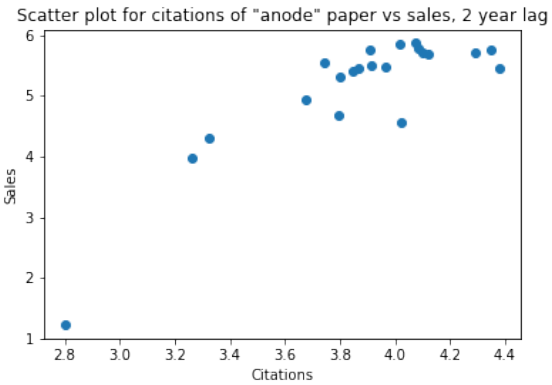
The data can also be evaluated by looking at only the most influential publishers. There are 112 individual publishers in the dataset but many of these have very few citations. These less influential publishers can be removed so that only the publishers with the greater impact on electric vehicle sales are accounted for. In this graph only the 5 most cited publishers are used.

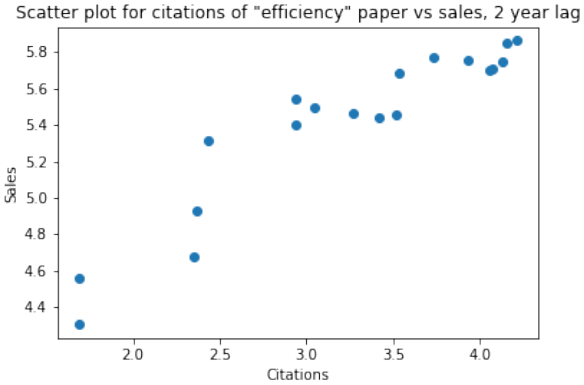


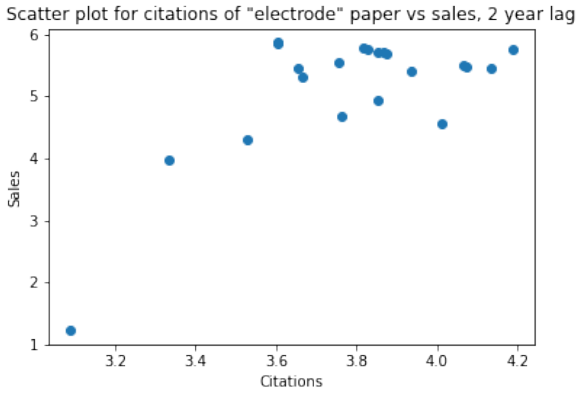
In addition, the titles of the technical papers were searched for key words that may give useful information about which technological advances have the greatest impact on electric vehicle sales. These key words were SEI, Cathode, Anode, Efficiency, and Electrode. This was done with a lag of two years to increases the number of data points.











From these graphs we can see that the keyword “Efficiency” has the best correlation with the sales of electric vehicles. Furthermore, they all seem to share a semi-linear relationship with the EV Sales.

b. Model and Techniques:

Many models were attempted. The model that brought the smallest errors utilized multivariable linear regression. The variables that are extracted from the dataset and used to create the linear regressing equations are key words in the papers. These keywords include SEI, cathode, anode, efficiency, electrode and total citations most influencing papers. 75% of the data will be fitted according to this equation using scikit-learn linear regression package:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$$

With ε as the error and $\beta_0, \beta_1, \cdots \beta_r$ will be iterated to minimize the error following this equation:

$$SSR = \Sigma_i (y_i - f(\mathbf{x}_i))^2$$

In this case, r = 6. After preprocessing, the following coefficients are obtained:

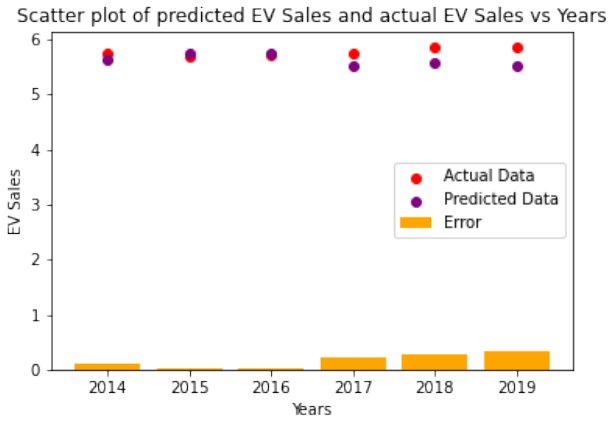
$[[\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6]]$ = [[-5.82907703, 0.02657156, -0.20898543, -0.23585407, 0.65336476, -0.15497522, 7.13711686]]

Intercept:

[0.01644511]

5. Results and Discussion:

a. Results testing: Using the above equations, the 25% left of the data (EV Sales from 2014 to 2019) are predicted.




```
Average Error:  0.1736415199065379
Percentage of Average Error to the Average Value of the Data:  3.009770814
4404987
```

b. Discussion:

The error is small, but it must be taken into account that this error is based on a log scale. Another disadvantage of this model is that very few data is collected thus making the predictions off.