Universe
Scientific Publishing

Analysis and Forecast of New Energy Vehicle Sales Volume Based on Industry Data under the Epidemic

Yuchen Ji

No.2 High School Of East China Normal University, Shanghai 200000, China.

Abstract: In daily life, the cars play an important role as a means of transportation. However, economies of many fields have been indelibly hit this year due to the pandemic. In order to help automobile manufacturers, consumers and the government take countermeasures in case of similar man-made or natural disasters in the future, this paper will predict the future automobile sales based on China's GDP data in 2020 and performance source of automobile's parameters and sales volume in the two years before and after the epidemic. Previous studies on car sales mostly focus on future sustainability development in the industry, but there is a lack of research using data and quantitative analysis. This paper will use data processing and machine learning to predict the sale volume as well as finding the key factors that will influence the sale volume.

Keywords: Statistics; Pandemic; Vehicle Sale; Finance

Introduction

We use vehicles everywhere in our daily life. It's no doubt that vehicles play an important role in our life. Furthermore, buying a new car is not as easy as taking a piece of cake. As a result, it's quite important for us to recognize what's going on in the field of cars. However, as we all know the pandemic has an inevitable negative influence on economies of many fields. Because the pandemic as well as the epidemic prevention policies, the sale volume of new energy vehicles were greatly affected.

From this point of view, using the previous data to predict what kind of situation will happen when similar man-made or natural disasters happen is very essential. These predictions can help automobile manufacturers, consumers and the government take countermeasures to reduce their lost as much as possible. For the automobile manufacturers, the predictions can help them to predict how many cars they should produce in this period of time and also suggest how many components they should import as well. If they do not make a proper plan for the production, unsalable and the great pressure on the cash flow will be destructive for some small-scale company. For the government law maker, a prediction can assist them publish a law to stimulate the economy as soon as possible to minimize the lost. It's the same for the consumer, as they know more about the situation in the field of vehicles and enable them to get a more financial price.

Now the society do need a qualified prediction of the vehicle industry. To make the final result as accurate as possible, this paper is going to focus on the new energy automobile industry aim at collecting all the data in this area from 2018 to 2021 including the sale volume and other parameter about the cars. In the following passage, this paper will use data processing and machine learning to predict.

Method

Since there is no available online data set summarizing various reference indicators of new energy vehicles from 2018 to 2021, I made the data set myself before the start of the study. In terms of automobiles, there is no website where all the results can be found directly. I crawler down some data I need especially the sales volume in the network of dashi data. (https://www.daas-auto.com/supermarket_data_De/120.html) and search for more characters of these kinds of vehicles on the internet, for instance, the length, width and height of each kind of vehicles. The original data set reach a scale of 2100 pieces of data containing 8 characters. The next step is to do data cleaning—— to change the literal words into digital characters.

Firstly, use pandas (a data analysis tool) to import the data sets into jupyter. The following step is to do data preprocessing. Due to one of the 1348 pieces of data is shown as None in the result, I use the function "dropna" to remove the empty data in order to avoid it disturbing the final result. After removing the empty data, the next step is to deal with some data includes Chinese character. To solve this problem, we could use the ordinal code to preprocess these data, turning these categorical data to several numerical features. When processing the length, width and height data of the vehicle, the original unit is millimeter, so the value of the overall data will affect the prediction result greatly. In order to keep the dimensional consistency, I do the process of normalization. All the numbers are divided by 1000. All processed data is generated using the COPY tool to create a new data set without modifying the original one. With these methods, the preprocessing of all text data is finished. When some extremely small data is found, for example, there is only one data in the "manual" shift, it can be "dropped" to reduce one feature and improve the accuracy of the model.

When some extremely small data is found, for example, there is only one data in the "manual" shift, it can be "dropped" to reduce one feature and improve the accuracy of the model. Completion of individual data processing using the tool "pd. Concat" will have to complete the processing of data is merged into one table. After all the processed data are put into a table, the data of 'manufacturer ',' brand ', 'model ',' time 'and' sales volume 'in this table are taken out separately, and then take out the data of each year from this table in turn. Because sales of a particular kind of car is not available every year, I only use all the overlapped data from 2019 to 2021 to complete the data preprocessing.

After preprocessing the next step is to do data prediction. In the prediction "x", the characters collected to do prediction, is the whole table without the sale volume of 2020; "y", the data needed to be predicted is the sale volume of 2020. In the prediction I divided all of the data into 10 groups by the tool "cv" and let all of them to be mostly two parts ——the training group and the test group. The "param_dict" is a function including standard deviation which are used to correct the result of prediction. In order to do the prediction, I decided to use 6 different models——AdaBoost, Extra Tree, Random forest, MLPR, Decision Tree, and Bagging —— these six models to predict the sale volume separately. Because this prediction is in the area of regression, the six models are all famous models being widely cited in regression problems.

However, in the first training by the decision tree model does not preforms well, the R2 is -1.37, which suggests that the prediction is not accurate. One of the reasons for that is the number of pieces is not large enough in the database, which may be the mean reason for the inaccurate. The second reason is that there are many noises in the dataset. As a result, these companies' sales volume has an enormous change and cause the prediction result being hard to fitting. In order to solve this inaccuracy problem, I use two ways to optimize the model. The first way is to turn the sale volume into its logarithm. Let's take the sale volume prediction of 2020 as an example. It's easy to observe that in 2020 the greatest sale volume of new energy vehicles is about 30000, however, the least one is only 1. This enormous gap will cause the model to be quite inaccurate and easy to overfitting. Using the logarithm could reduce the difference and range of sale volume efficiently. This is the most important step of optimization. Furthermore, make some character that are both very high and very low as a pair to consider. Take the type of displacement and the value of displacement as an example. Even though these two characters are completely two different features, they are closely related. The same kind of displacement usually have the same kind of value. Which is also shown in the "pearson correlation coefficient" graph. To make a balance, I drop the features whose pearson correlation coefficient's absolute value is under 0.002.

Evaluation

This final selling of 2021 is made by the "Random forest Tree" model, containing 80% of training and 20% of data test. It's clear to find that even though not all the lines are coincided, the tendency of prediction is approximately the same with the true one. It can also be found that the range of value of prediction is smaller than the real one, attribute to the predict model is easier than the reality in order to avoid overfitting. The reason why at first the model seems to perform well but does not continuously fit the true data is that in the dataset the order is made by the volume of sale. For the large companies, who seem to be fewer influenced by the pandemic still have an increasing sales volume. But those small or little-known companies will have a larger change in sales volume and hard to predict and train.

We could find that the accuracy of 2021 is higher than the one of 2020, probably due to one more year's sale volume. And in 2020 China undergo the pandemic, which cause the whole society's economic system to change, so the sale volume of cars in 2021 will greatly relay on the one of 2020, which is quite different from previous years.

I also make another dichotomy model to predict whether the sale volume will increase or not in the second year. It's obvious that the accuracy is quite high. The optimistic of data play an important role in this success. The accuracy is higher than the sale volume may also due to the prediction of sale volume is numerical while the prediction of tendency only has two chooses, which does not require so many pieces of data as sale volume prediction.

Discussion

After the whole experiment, I am also wondering that what factors contributing to the vehicles' sale volume the most. In order to find this answer, I first search it on the Internet. The mean answer could be summarized as the society's economic situation and the parameter of the car. To verify the result, I use the data processing to find out the top 10 important features of predicting. It's self-evidence that the sale volume of 2019 is the most essential feature when predicting the sale volume of 2020. The second important feature is the time the kind of new energy has being sold. This character is closely related to the car brand because the longer the car being sold, the possibility that more people will know it will be bigger. The time the car being on sold could reflects the reliability and knowability in some extent. Then the following three features are the length, width, and height of the car, which are the parameter of the car. The size of the car is directly relative to the need of energy. In other words, the bigger the car is, the more energy it will cost, the more money you will have to spend on it. Furthermore, from previous analysis, the sale volume of SUV is affected greatly during the pandemic. This phenomenon is probably a result from the price of the SUV. As SUV is bigger relative to other types of cars, it will probably be more expensive, has a wider limitation on car parking and actually needs more energy to maintain driving. As a result, the parameter of the car could partly reflect the consumption level of the customers, which is also a very important factor that will influence the sale volume. During the pandemic, the middle class receive the greatest influence on living standard, who are the mean consumers in buying SUV. This fact is also matched with the result getting from the experiment.

References

[1] Cai B, Rui MJ. Research on Automobile Sales Forecasting in China based on improved differential Evolution Algorithm and Grey Model [J]. Shanghai Management Science, 2015, 37(01):14-20.

[2] Zhu LR, Ge DD, Pan LD. Response analysis of new energy vehicle industry chain based on the COVID-19 epidemic [J]. Southern Agricultural Machinery, 2019, 51(23): 44+58.

[3] Schoenfeld David. (1982). Partial residuals for the proportional hazards regression model. Biometrika.