Big Data Analysis for Vertical Industries, Automotive Industry

Huahui Mo
Computer Science
University of Calgary
Calgary, Canada
huahui.mo@ucalgary.ca

Jianan Ding
Computer Science
University of Calgary
Calgary, Canada
jianan.ding1@ucalgary.ca

Zheng Liang
Computer Science
University of Calgary
Calgary, Canada
zheng.liang@ucalgary.ca

***Abstract*—The automotive industry is an example of a technological revolution. It is considered one of the fastest growing sectors in the world. Our project is using data set about the automotive industrys production and the economy to dig more into the detail of its impact on the economy. This project will develop a time series model for the total production volume of the automotive industry and the GDP for United States. We can forecast the future production volume and its relationship with the economy by using machine learning techniques.**

***Index Terms*—automotive industry, time series analysis, forecast, GDP, machine learning**

I. Introduction

The automotive industry is an example of a technological revolution, which consist of companies involved in design, development, manufacturing, marketing, and selling of vehicles. It is considered one of the fastest growing industries in the world. What does this huge industry brought to our world, and what will this industry be in the future is the question we have. However, the automotive industry is short on data in one aspect naturally because of it's existence time and limited countries or brands. Which means the analysis or the prediction is unreliable with these limited data. It has to combine with related data that from other dimensions. As a result, our team built a time series model for the automotive industrys production volume and GDP data of United States from the past 20 years to make predictions on production, as well as examine the relationship between production and GDP, which this relationship could be the validation of the predictions. This study is done by analyzing the time series and built machine learning models to predict the automotive production volume and GDP. This topic is fascinating because people get used to the vehicle in daily life but always ignore this huge industry.

By following sections, we will discuss the basic idea related to the topic with relevant citations in the project overview section, propose our solution in the discussion section, illustrate our implementation and contribution and give a future suggestion in the result section and conclude all in the conclusion section.

II. Overview

A. Dataset

Web crawler was used to get the United States' automotive production volume and GDP. Collecting out dataset uses the following steps:

1) : Come out with a list of URLs to visit. As the crawler visits these URLs, it looks for all the hyperlinks in the page and adds them to the list.

2) : Analyze the content of websites in the list, and looks for the target data. Sort the data into the specified format.

3) : Save the target data into an CVS format file.

We have collected the automotive production volume and GDP for United States between 1993 and 2017. We built our time series using all the data we have collected from the web crawler. These data will be utilized as two time period, "Forecasting the future" and "Look into the past", which will be discussed below in discussion section. For the "Forecasting the future" part, the training data set is the data from 1993 to 2017, and the data beyond 20017 are the predictions. As for the "Look into the past" part, this part is focusing on validation of the models. The training data set is the data from 1993 to 2007, and the validation data set is the data from 2007 to 2017.

B. Analytical Tools

1) Autoregressive Integrated Moving Average Model: [1] An autoregressive Integrated Moving Average Model (ARIMA) is a statistical analysis model that uses time series data to better understand the data set and predict the future trends. [1] "Autoregressive" means lags of stationarized series in the foresting equation, lags of the forecast errors called "moving average" and a time series which needs to be differenced to be made stationary is called "integrated". [2] An ARIMA model has three components as follows: [1]

a) : Autoregression(AR), indicates a model that displays a evolving variable that regresses on its own laggeg, or prior,values.

b) : Integrated(I), represents the data have been replaced with the difference between the data values and previous data values.

c) : Moving Average(MA), incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each component functions as a parameter with a standard notation. A stand notation for an ARIMA is p, d, and q. p is the number of lag observations in the model. d is the number of times that the raw observation are differenced. q is the size of the moving average. Differencing, autoregressive, and moving average come up a ARIMA model as follows:

$$y_t = c + \phi_1 y_{dt-1} + \phi_p y_{dt-p} + ... + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t \quad (1)$$

where y_d is Y differenced d times and c is constant.

2) Long Short-Term Memory Net: [3]

Long short-term memory(LSTM) is an artificial recurrent neural network. LSTM are widely used for sequence prediction problems. LSTM are able to store pass information that is important, and drop the information that is not. A common LSTM unit is composed of following units:

a) : Cell: responsible for keeping track of he dependencies between the elements in the input sequence.

b) : Input gate: controls the extent to which a new value flows in to the cell.

c) : Forget gate: controls the extent to which a value remains in the cell.

d) : Output gate: controls the extent to which the value in the cell is used to computed the output activation of the LSTM unit

We split the data into training data set and testing data set. We fit the traning data into the LSTM model and make forecast, at the end, we evaluate the model by using the RMS Error (Root mean square) [5].

III. Discussion

To find the answer for our project, we use pyramid.arima python library to build our ARIMA (Autoregressive integrated moving average) model. Keras python library was used to build our deep learn LSTM (Long Short-Term Memory) model, and our result contain two part, forecasting the future and look into the past.

A. Forecasting the future

In this part, we would use the automotive industry production data set and United States' GDP from 1993 to 2017 to forecast the future production for the industry and the future GDP for United States.

1) ARIMA model: In our ARIMA forecast result, the blue line is our collected data set, and the orange line is the model foecasting. Figure 1 shows that the time series model is non-stationary since the mean change over time in this model. Figure 1 forecast the Automotive industry production in United States, which predict a decreasing trend after 2019. As we can see in Figure 2, it also predicting a significant decline, but the decline will stop at 2020 and there will be a rebound for the industry GDP after 2020. And for the Figure 3, the prediction is showing a huge drop around 2020, which is like the financial crisis happened in 2008.

2) LSTM model: In our LSTM forecast result, the blue line is our collected data set, and the red line is the model forecasting. LSTM model gave the prediction in Figure 4, which is also showing and decreasing trend similar to ARIMA model's prediction in Figure 1. In Figure 2, LSTM model gave an prediction that is different to ARIAM model (Figure 2), it forecast the Automotive industry GDP in United States will continue increasing until 2020, after that there will be a significant decline. And for the forecasting for United States' GDP, LSTM model predict that it will have an continue increasing trend as before. In our LSTM forecast result, the blue line is our collected data set, and the red line is the model prediction.

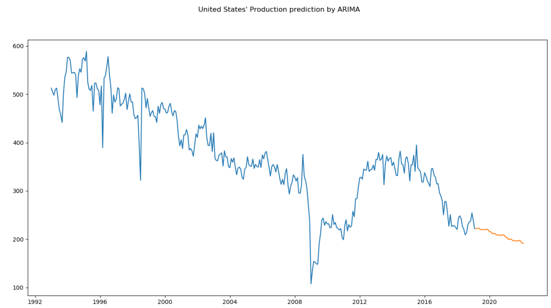

Fig. 1. United States Auto Production by ARIMA

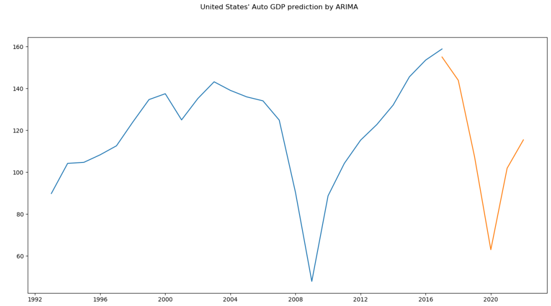

Fig. 2. United States Auto GDP by ARIMA

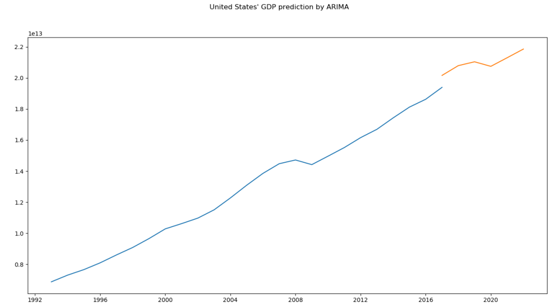

Fig. 3. United States GDP by ARIMA

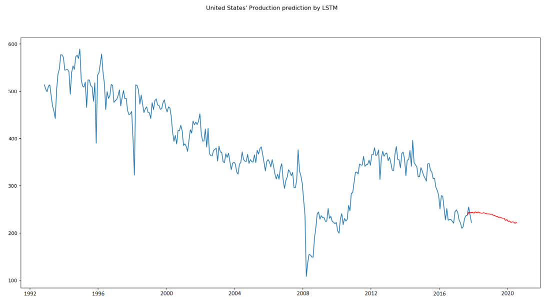

Fig. 4. United States Auto Production by LSTM

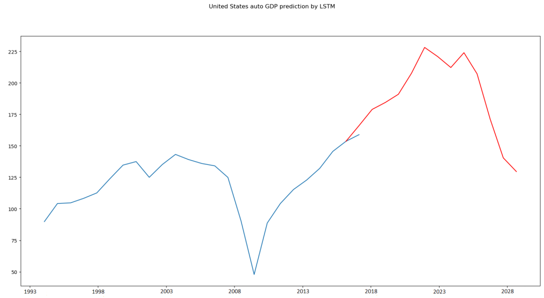

Fig. 5. United States Auto GDP by LSTM

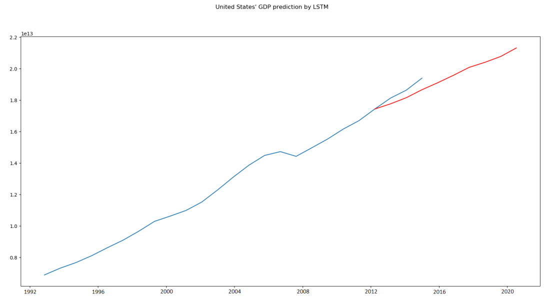

Fig. 6. United States GDP by LSTM

B. Look into the past

In this part, we would use the data set from 1993 to 2007 to forecast the next few year. As we all know that there was an financial crisis happened in 2008, we are trying to see if our ARIMA model and LSTM model would be able to predict the financial crisis may happen base on the given data set.

1) ARIMA model: In our ARIMA forecast result, the blue and orange line is our actual data set, and the green line is the model forecasting. In Figure 7, ARIMA model gave an forecast that show the huge decreasing trend from 2008 to 2019, as we can see that there is a big difference between the prediction and our actual data. Predicting an decreasing trend has also been seen in Figure 8, the prediction and actual data set go to an opposite side. In Figure 9, the prediction gave an smoother increase than the actual data.

2) LSTM model: In our LSTM forecast result, the blue and orange line is our collected data set, and the red green is the model prediction. LSTM gave an more precise prediction than ARIMA model in Figure 10, but they both showing a decreasing trend in automotive production. As Figure 11 show, it gave an sightly increase trend after 2008 to 2019, which is a huge difference to the actual data we had. The prediction in Figure 12 gave a smoother increasing start form 2008, and skip the financial crisis that happened in 2008.

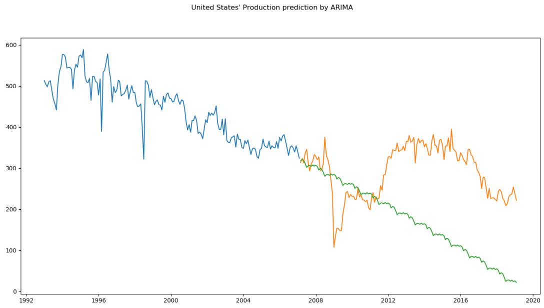

Fig. 7. United States 2008 Auto Production by ARIMA

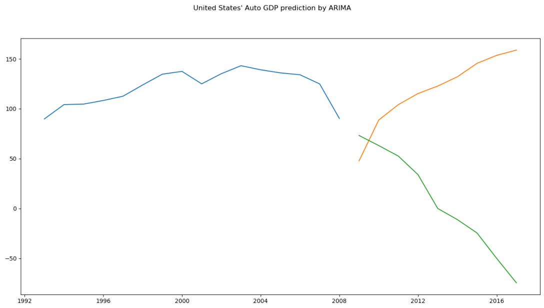

Fig. 8. United States 2008 Auto GDP by ARIMA

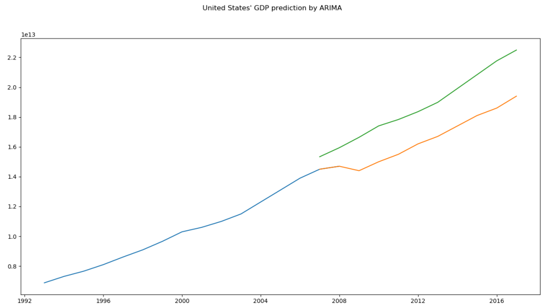

Fig. 9. United States 2008 GDP by ARIMA

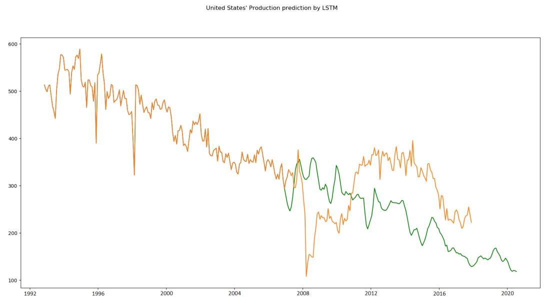

Fig. 10. United States 2008 Auto Production by LSTM

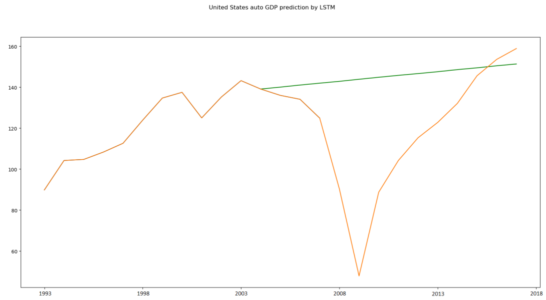

Fig. 11. United States 2008 Auto GDP by LSTM

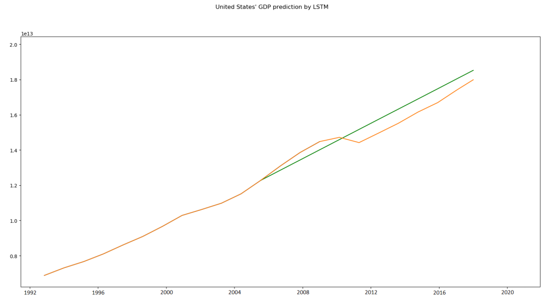

Fig. 12. United States 2008 GDP by LSTM

C. Know the problem

Since ARIMA model and LSTM model can not precisely predict the 2008 financial crisis, we can see that the relationship between the automotive industry production and GDP are actually two-sided, they influence each other and depended on each other. Comparing between these two forecasting model, overall, we can see that the LSTM model give an more accurate prediction than the ARIMA model.

IV. Result

With the models built in Fig.1 - Fig.12. Obviously there is pattern between GDP and automotive industry. As decreasing of the GDP, the production was decreasing as well, and vice versa. Definitely this is the connection between GDP and production and examined by models.

As models of production of the future are calculated with data, there is a way to validate the production prediction's correctness. Which is by applying GDP prediction with the production prediction the model made. With the connection found above, it is easy to point out the errors in production prediction.

The implement above could be concluded in such:

1) Collect the related data
2) Plug data into the model and make prediction
3) Validate the production prediction with the GDP prediction

This implementation could also be concluded into a method for analyzing this industry. Compare to only focusing on the production data, a better way could be combining the production data with related data that from other dimension, like GDP in this case. However, with only a few of dimensions is not enough to make a perfect prediction. On the other hand, when making a prediction for production, GDP is not involved in this implementation, a Non-linear Prediction of Time Series Based on Dual-forecasting Model [4] would be perfect for this. Since for now this paper is focusing on the validation of this implementation. It has some improvements to be done since it is the first step. For the future work, it could be collecting more related data, to form a bigger data set in order to get a more precise prediction. Another could be applying the Non-linear Prediction of Time Series Based on Dual-forecasting Model [4].

By studying this industry, there are some thoughts that could be formed as suggestions for this industry. The automotive industry could use this implementation, as a base, expand it with other dimensions of data make a more accurate production prediction for the future. This could lead to a better cost control for the production, which increase the efficiency of the production. On the other hand, the automotive industry could use the idea of this implementation, predict others like consumers' preference, or consumers' need instead of production number. Then the automotive company could know better about their consumer, so they could be more efficient in design and production, which will make them more powerful in competition in automotive market.

V. Conclusion

Automotive industry, as an example of a technological revolution, is considered as one of the fastest growing industries in the world, which is also a great target for Big Data Analysis for vertical industries. To begin with, web crawlers were used to collect the related data needed, which is the vehicle production data and the GDP data. There are two models, ARIMA and LSTM, were used to analyzing the data, since two models are more reliable than only one. With the data collected and the model built, a clear connection is found between the GDP and the automotive production, which is an obvious positive relationship. This relationship not only prove the hypothesis, but also point out a way that could be used to validate the correctness of the production prediction. Compare to only focusing on the production data, combining with data the from other dimension is a better way to make the prediction more accurate and reliable. More related data from more dimensions and a Non-linear Prediction of Time Series Based on Dual-forecasting Model [4] could be the future work of this first step. The idea of this implementation is the contribution of this paper to the automotive industry. Furthermore, there are some suggestions for this industry. the industry could use this implementation but with more dimensions of data to make a better production prediction in order to lower the production cost. Or the industry could use this implementation but on predicting others like consumers' need. Which could make the design and production more efficient, and be more competitive in the market.

VI. Reference

References

[1] J. Chen, Autoregressive Integrated Moving Average (ARIMA), Investopedia, 13-Apr-2019. [Online]. Available: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp. [Accessed: 14-Apr-2019].

[2] R. Nau, Introduction to ARIMA: nonseasonal models, Introduction to ARIMA models. [Online]. Available: https://people.duke.edu/ rnau/411arim.htm. [Accessed: 14-Apr-2019].

[3] C. Olah, Keras LSTM tutorial - How to easily build a powerful deep learning language model, Adventures in Machine Learning, 03-Feb-2018. [Online]. Available: https://adventuresinmachinelearning.com/keras-lstm-tutorial/. [Accessed: 14-Apr-2019].

[4] Y. Fang and Q. Liu, Non-linear Prediction of Time Series Based on Dual-forecasting Model, System Simulation Technology, vol. 7, no. 2, pp. 117125, Apr-2011.

[5] S. Holmes , RMS Error, RMS Error, 2000. [Online]. Available: http://statweb.stanford.edu/ susan/courses/s60/split/node60.html. [Accessed: 14-Apr-2019].