JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 07, ISSUE 06, 2020

Prediction of Vehicle Sales using Arima Model

Swathi Chigurlapalli, Pushpalatha Maligireddy

Department of Computer Science and Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

Abstract

The sales forecasting of vehicles plays an important role in worldwide automobile market and its gaining attractive due to the advancement in data science approaches. However, the number of efforts undertaken in this field of research is quite small to date. Methods based on statistical learning theory are powerful instruments to get insight into internal relationships within huge empirical datasets. Therefore, they can produce reliable and even highly accurate forecasts. However, data mining algorithms have become more and more complex over the last decades. In this work, the accuracy of the prediction has the same importance as the explicability of the model. Hence, only classical data mining methods are applied here.

This project presents enhanced sales forecast methodology and model for the automobile market which delivers highly accurate predictions while maintaining the ability to explain the underlying model at the same time. The representation of the economic training data is discussed, as well as its effects on the newly registered automobiles to be predicted. The methodology mainly consists of time series analysis and classical data mining algorithms, whereas the data is composed of absolute and/or relative market-specific exogenous parameters on a yearly, quarterly, or monthly base. It can be concluded that the monthly forecasts were especially improved by this enhanced methodology using absolute, normalized exogenous parameters. The main goal of this project is to consider main approaches and case studies of using machine learning for sales forecasting. The effect of machine-learning generalization has been considered. This effect can be used to make sales predictions when there is a small amount of historical data for specific sales time series in the case when a new product or store is launched. A stacking approach for building regression ensemble of single models has been studied. The results show that using stacking techniques, we can improve the performance of predictive models for sales time series forecasting.

Keywords: ARIMA model, vehicle sales, sales forecasting.

1. Introduction

Sales prediction is a complex process and a challenging task for researchers. It includes expertise in multiple disciplines. The prediction of atmospheric parameters is essential for various applications. Accurate prediction of Sales parameters is a difficult task due to the dynamic nature of atmosphere. Generally, prediction is done on Time series data. A time series is a sequence of observed values of some entity that is measured at different points in time. With the advancement of collecting data, huge amounts of data have been collected making it impossible to be processed manually. This is where the time series analysis must be automated and take advantage of modern computing mechanisms. Various techniques like linear regression, auto regression, Multi-Layer Perceptron, Radial Basis Function networks are applied to predict atmospheric parameters like temperature, wind speed, rainfall, meteorological pollution etc. It was found that the nonlinear operator equations governing the atmospheric system are the ones who can better understand the dynamics of atmosphere. In the recent past many forecast methods have been developed using Artificial Neural Networks (ANNs). Neural network techniques have the potential to handle complex, nonlinear problems in a better way when compared to traditional techniques. However, systems developed using neural network model suffer

from certain drawbacks like local minima, model over fitting etc. In this paper we are going use ARIMA model for the predicting the sales.

Intelligent vehicles are a new generation of vehicles that are equipped with advanced sensors and other devices, use new technologies such as artificial intelligence, have self-driving functions, and gradually become intelligent mobile spaces and application terminals. In recent years, a new generation of industrial technology with intelligence and the Internet as the core has been on the rise, promoting traditional industries to accelerate the "intelligent" transformation and upgrade [1]. As one of the most widespread and important industries, the automotive industry is also accelerating its progress of intelligence, and "Intelligent vehicles" are receiving more attention, especially in the areas of shared mobility, energy consumption, and vehicle safety [2]. Most intelligent vehicles are electric vehicles, and the number of electric vehicles will continue to grow in the next 40 years with an "S" trend. The Chinese government has also made intelligent vehicles a priority to the development of the automotive industry, and the development strategy released by the country has drawn strong attention to the industry. Intelligent vehicles are increasingly becoming the focus of the automotive industry, but research related to intelligent vehicles is still relatively lacking.

Predicting sales is a key step in making production decisions on companies and public policy for governments. Companies use product sales predictions as a basis for estimating sales revenues. They can also use product sales predictions to develop plans for marketing, sales management, production, purchasing, and logistics to improve economic efficiency and reduce losses in production planning. Intense competition, large investments, and the need for rapid model updates characterize the automotive industry, which makes predicting crucial for sale and production processes [3]. Most intelligent vehicle companies are emerging vehicle makers and have a low grasp of the market demand for their products, so it is more important to accurately predict the sales scale of intelligent vehicles for companies to set a reasonable production scale. In addition, as a strategic emerging industry supported by the Chinese government, the sales growth of intelligent vehicles is the result of the joint action between the market and the government. Therefore, the study of the intelligent vehicle market sales prediction is also important to the formulation of intelligent vehicle industry support policies and the arrangement of related supporting facilities.

Big data have become one of the important tools in the field of predicting and it has been very widely used in the traditional automotive field. Unlike traditional vehicles, intelligent vehicle brands come with Internet properties and are often also referred to as Internet vehicles, which is why they are widely followed and discussed on the Internet; with more than 6.9 million discussions about Tesla (Fremont, CA, USA), the intelligent vehicle brand, on the social platform Twitter from 2018 to 2020. In China, NIO, the emerging domestic intelligent vehicle brand, has 910,000 followers on the social platform Weibo, with more than 200,000 discussions about NIO on Weibo alone from 2018 to 2020. In this case, the brand's online public opinion and attention often affect the user's willingness to purchase and further affect the sales of the vehicle brand. In 2019, the domestic intelligent vehicle brand XPeng had a big drop in sales due to a collective online rights opinion storm, with sales down 43.9%. Previous studies have tended to focus only on traditional vehicle brands and less on the impact of online public opinion and online search index on intelligent vehicle sales. With the growing development of intelligent vehicles, it is urgent to establish a sales prediction model that applies to intelligent vehicles and considers the influence of online public opinion and attention.

2. Literature survey

Souza et al. aimed to estimate the future generation of ELVs to assist decision making and mitigate the global impact of this type of waste on the environment. For this, a hybrid forecasting model was used, based on Autoregressive Integrated Moving Average (ARIMA) methodology and on Artificial

ISSN- 2394-5125 VOL 07, ISSUE 06, 2020

Neural Networks (ANN), with a set of temporal data extracted from Brazilian sectoral platforms. Considering the scarcity of information that supports decision-making in waste management in Brazil, this study may also contribute to the proposition of alternatives that favor the proper management of automotive waste, providing a reference for the formulation and implementation of policies related to ELVs in the country.

Mehendale et al. aimed at applying Artificial Intelligence (AI) techniques to identify and predict complex sales patterns and compare the results with traditional forecasting models. Sales data related to the sample firm was used in the study for forecast modelling using both, advanced forecasting method (ARIMA) and Artificial Neural Networks (ANNs). The latter considered inputs of influential/causal factors and revisits every time to identify new trends related to those factors, thus providing a more robust forecast. This is compared with results from the ARIMA model. Though the purpose of the research is achieved, a few limitations exist because of the limited availability of data. Besides, the factors that affect the prediction are generalized and based on previous research works. A more customized approach towards the firm under study would greatly improve the accuracy.

Nigam et al. presented Box-Jenkins method used to forecast the future demand in a two-wheeler industry. An automated technique in machine learning with the help of python language has been developed and used to analyze time series data and ultimately fit the model for future demand projection. The time series data is collected for the Royal Enfield bikes' monthly sale available at the official website of Eicher motors ltd. The resulting pattern found in time series data is used to forecast the future behavior, knowledge of which will help to maintain the appropriate inventory and to reduce the risk in terms of changing customers preferences, resource availability etc. Also, the effect of covid-19 pandemic has been captured to visualize its impact.

Swami et al. analyzed the trends of production, sales, and exports in the present paper with reference to the Indian automobile sector. The sales of vehicles in countries like Europe, America, Africa, NAFTA countries, and India have been considered to forecast the future sales. Data for sales between the periods of 2005 to 2018 have been collected and analyzed using the ARIMA forecasting technique.

Shakti et al. used the ARIMA model for predicting the number of sales for a Time series data. The dataset tractor sales data for a period of ten years (2003-2014) obtained from the Mahindra Tractors Company are used from which used to classify the performance by drawing various scattered plots and graphs. The result of the ARIMA results showed that which predicted better for the sales prediction of the next following 5 years.

Irhami et al. obtained the best model for forecast the number of cars and the number of motorcycles in the next 11 years. For the purpose, ARIMA method was used. Using the historical data of the number of cars and the number of motorcycles from 2001 to 2019, the best model for forecasting the number of cars and the number of motorcycles is ARIMA (1,1,0) and ARIMA (2,1,2), respectively. The models have MAPE of 7.01% and 7.24% for cars and motorcycles, respectively.

Permatasari et al. predicted the printed newspaper demand as accurately as possible to minimize the number of returns, to keep off the missed sales and to restrain the oversupply. The autoregressive integrated moving average (ARIMA) models were adopted to predict the right number of newspapers for a real case study of a newspaper company in Surakarta. The model parameters were found using maximum likelihood method. Then, the software Eviews 9 were utilized to forecasting any variables in the newspaper industry. This paper finally presented the appropriate of modeling and sales forecasting newspaper based on the output of the ARIMA models.

Kaya et al. recommended an 8-layer Deep Neural Network model for vehicle sales prediction. The inputs of the model consist of features, such as exchange rate, the gross domestic product, consumer confidence index, and the consumer price index. The vehicle sales forecast was made according to the output of the model. This work analyzed a total of 90 data monthly between the years of 2011 and 2018 was collected.

Haffner et al. devoted to the sales prediction of Svijany Slovakia, Ltd. Using modern cloud computing tool Microsoft Machine Learning Studio. In the paper there is briefly described characteristics of the company and sales, described the software system Helios Green which is used in the company. There is described the process of data processing from this system for further using. After data processing, predictive models and predictions of the sales are made. Results of the predications played a significant role in the management negotiations for the building of a new warehouse.

Zhang et al. introduced the seasonal decomposition and ARIMA model, this paper proposed a sales forecasting model for the consumer goods with holiday effects. First, a dummy variable model is constructed to test the holiday effects in consumer goods market. Second, using the seasonal decomposition, the seasonal factor is separated from the original series, and the seasonally adjusted series is then obtained. Through the ARIMA model, a trend forecast to the seasonally adjusted series is further carried out. Finally, according to the multiplicative model, refilling the trend forecast value with the seasonal factor, thus, the final sales forecast results of the consumer goods with holiday effects can be obtained. Taking the cigarettes sales in G City, Guizhou, China as an example, the feasibility and effectiveness of this new model is verified by the example analysis results.

3. Proposed system

3.1 ARIMA: (Auto Regressive Integrated Moving Average)

ARIMA model is done on time series data. Time series data is a sequence of observations collected from a process with equally spaced periods of time.

Examples: Industrial Averages, Daily data on sales, Daily Customers.

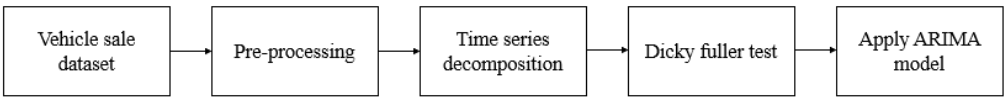

Fig. 1: Block diagram of proposed system.

Stages of Time series model process using ARIMA:

ARIMA is also known as Box-Jenkins approach. To build a time series model issuing ARIMA, we need to study the time series and identify p, d, q. Were,

p – Auto Regressive (Auto Correlation)

d - Integrated (Stationary / Trend)

q - Moving Average (Shocks / Error)

- Identification: Determine the appropriate values of p, d, & q using the ACF, PACF, and unit root tests. p is the AR order, d is the integration order, q is the MA order
- Estimation: Estimate an ARIMA model using values of p, d, & q you think are appropriate.

- Diagnostic checking: Check residuals of estimated ARIMA model(s) to see if they are white noise; pick best model with well-behaved residuals.
- Forecasting: Produce out of sample forecasts or set aside last few data points for in sample forecasting.

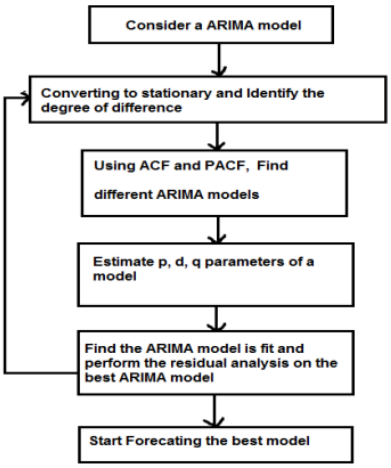

Fig. 2: Flow diagram of ARIMA model.

3.1.1 Autoregressive (AR) Process

Series of current values depend on its own previous values. In an AR(p) model the future value of a variable is assumed to be a linear combination of p past observations and a random error together with a constant term. Mathematically the AR(p) model can be expressed as:

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots\ldots\ldots + \varphi_p y_{t-p} + \varepsilon_t$$

Here y_t and ε_t are respectively the actual value and random error (random shocks) at period t, ϕ_i (i = 1, 2..., p) are model parameters and c is a constant. The integer constant p is known as the order of the model.

3.1.2 Moving Average (MA) Process

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots\ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

The current deviation from mean depends on previous deviations. An AR(p) model regress against past values of the series, an MA(q) model uses past errors as the explanatory variables. The MA(q) model is given by:

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots\ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

Here μ is the mean of the series, (j=1, 2..., q) θ_j = are the model parameters and q are the order of the model.

ISSN- 2394-5125 VOL 07, ISSUE 06, 2020

Conceptually a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations.

3.1.3 Stationary Series (Integrated)

To model a time series with the Box-Jenkins approach, the series must be stationary. In practical terms, the series is stationary if tends to wonder uniformly about some fixed level. In statistical terms, a stationary process is assumed to be in a particular state of statistical equilibrium, $i.e., p(xt)$ is the same for all t.

To achieve a series as stationary we need to do Regular differencing (RD),

$$\text{(1st order)}\ \nabla x_t = (1 - B)x_t = x_t - x_t\text{-}1$$
$$\text{(2nd order)}\ \nabla^2 x_t = (1 - B)^2 x_t = x_t - 2x_{t-1} + x_{t-2}$$

"B" is the backward shift operator. It is unlikely that more than two regular differencing would ever be needed.

3.2 Decomposition of time series

The decomposition of time series is a statistical task that deconstructs a time series into several components, each representing one of the underlying categories of patterns. There are two principal types of decomposition, which are outlined below.

4. Results

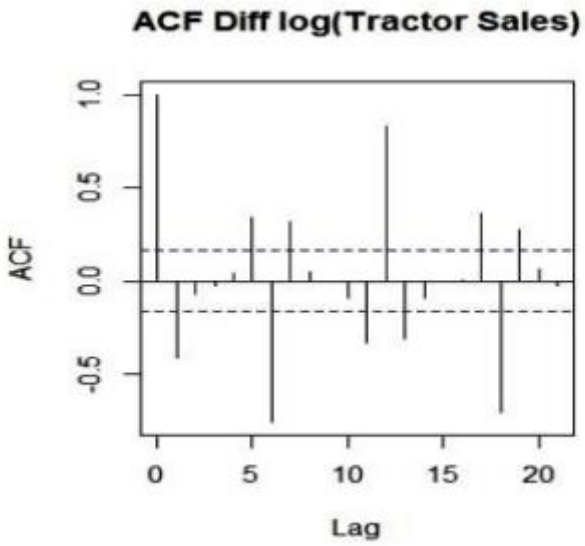

Fig. 3: ACF graph.

ISSN- 2394-5125 VOL 07, ISSUE 06, 2020

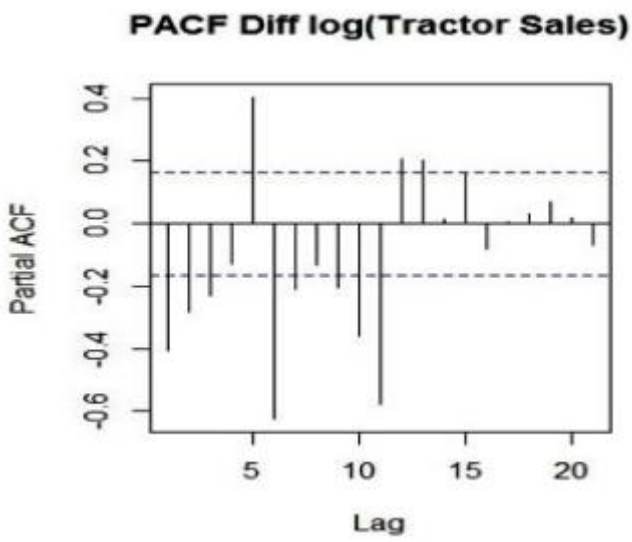

Fig. 4: PACF graph.

```
Series: log10(data)
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
          ma1     sma1
      -0.4047  -0.5529
s.e.   0.0885   0.0734

sigma^2 estimated as 0.0002571:  log likelihood=354.4
AIC=-702.79   AICc=-702.6   BIC=-694.17

Training set error measures:
                      ME       RMSE        MAE          MPE      MAPE      MASE
Training set 0.0002410698 0.01517695 0.01135312 0.008335713 0.4462212 0.2158968
                    ACF1
Training set 0.01062604
```

Fig. 5: Choosing best ARIMA fit.

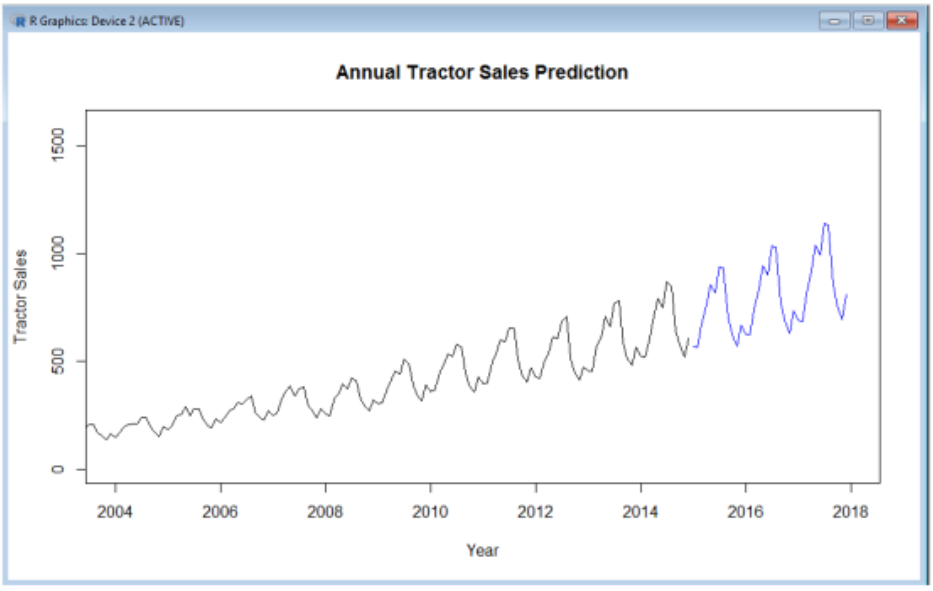

Fig. 6: ARIMA model prediction up to 2018.

ISSN- 2394-5125 VOL 07, ISSUE 06, 2020

```
$pred
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2015 2.754168 2.753182 2.826608 2.880192 2.932447 2.912372 2.972538 2.970585
2016 2.796051 2.795065 2.868491 2.922075 2.974330 2.954255 3.014421 3.012468
2017 2.837934 2.836948 2.910374 2.963958 3.016213 2.996138 3.056304 3.054351
          Sep      Oct      Nov      Dec
2015 2.847264 2.797259 2.757395 2.825125
2016 2.889147 2.839142 2.799278 2.867008
2017 2.931030 2.881025 2.841161 2.908891

$se
            Jan         Feb         Mar         Apr         May         Jun         Jul
2015 0.01603508 0.01866159 0.02096153 0.02303295 0.02493287 0.02669792 0.02835330
2016 0.03923008 0.04159145 0.04382576 0.04595157 0.04798329 0.04993241 0.05180825
2017 0.06386474 0.06637555 0.06879478 0.07113179 0.07339441 0.07558934 0.07772231
            Aug         Sep         Oct         Nov         Dec
2015 0.02991723 0.03140337 0.03282229 0.03418236 0.03549035
2016 0.05361850 0.05536960 0.05706700 0.05871534 0.06031866
2017 0.07979828 0.08182160 0.08379608 0.08572510 0.08761165
```

Fig. 7: Prediction of 2017 and 2018 in month wise.

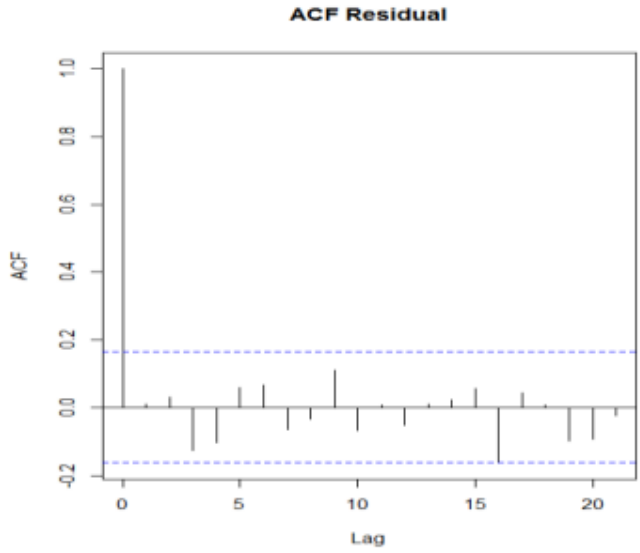

Fig. 8: ACF residual model for the AR model.

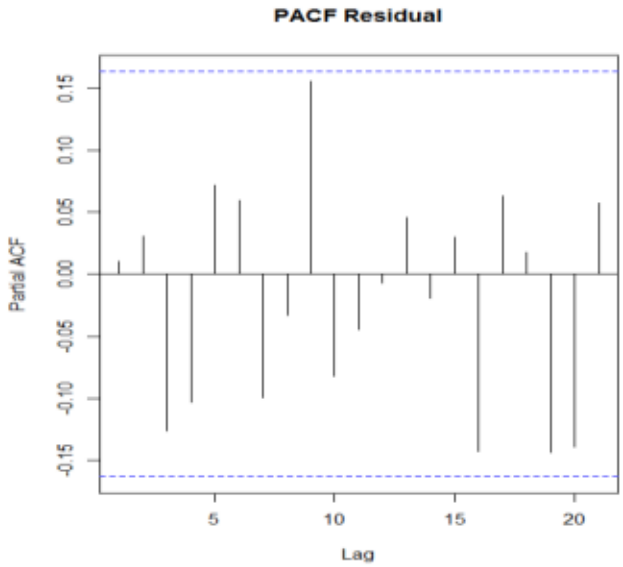

Fig. 9: PACF graph for MA model.

Since there are no spikes outside the insignificant zone for both ACF and PACF plots we can conclude that residuals are random with no information in them.

5. Conclusion

In this paper the performance of ARIMA is done by showing different visualization graphs. Results obtained shows that ARIMA performs better for the next following years. Number of sales parameters observed that it has significant effect for ARIMA model, that is it predicts has a wide deviation from the data taken from the previous years. That is ARIMA model predicted values has a large difference from the expected values. know there no model which is predicted perfectly when it comes to sales forecasting. But as of now ARIMA model is more suitable for sales forecast for the static time series data.

5.1 Future scope

Future scope of the work is to optimize the accuracy of sales forecasts and is a guide to setting a reasonable production scale. The limitations of this study are mainly reflected in the following aspects: (1) This paper only considers the effect of Weibo content text on sales prediction. As an open and interactive social platform, Weibo's comments, likes, and reposts data can further reflect the public's sentiment towards specific content. In the next step, Weibo comments, likes, and reposts will be considered in the prediction model. (2) This paper only considers the sentiment of the Weibo textual content. The emoji expressions and pictures deleted in the pre-processing stage are more likely to express the publisher's sentiment, and the next step will be to consider including the above content in the scope of sentiment analysis. (3) The factors that affect intelligent vehicle sales are complex, such as policies, gas and electricity prices, and loan rates. In this paper, only the effects of online public opinion and online search index are considered, and the next step is to fully consider various factors and further improve the prediction model. (4) The ensemble learning method can improve the accuracy of model prediction and will be applied in future studies.

References

[1] Miao, X. Smart Factory and the Transformation & Upgrading of Equipment Manufacturing Industry. Process Autom. Instrum. 2014, 35, 1–6.

[2] TuroŃ, K.; Czech, P.; TÓTh, J. Safety and Security Aspects in Shared Mobility Systems. Sci. J. Sil. Univ. Technol. 2019, 104, 169–175.

[3] Geva, T.; Oestreicher-Singer, G.; Efron, N.; Shimshoni, Y. Using Forum and Search Data for Sales Prediction of High-Involvement Products. MIS Quart. 2017, 41, 65–82.

[4] J. A. F. D. Souza, M. M. Silva, S. G. Rodrigues, S. M. Santos, "A forecasting model based on ARIMA and artificial neural networks for end–OF–life vehicles", Journal of Environmental Management, Volume 318, 2022, 115616, ISSN 0301-4797, https://doi.org/10.1016/j.jenvman.2022.115616.

[5] Mehendale, Abhang, and Nadheera Sherin HR. "APPLICATION OF ARTIFICIAL INTELLIGENCE (AI) FOR EFFECTIVE AND ADAPTIVE SALES FORECASTING." Journal of Contemporary Management Research 12.2 (2018).

[6] Nigam, Bhanuj, and A. C. Shukla. "Sales forecasting using Box Jenkins method based Arima model considering effect of covid-19 pandemic situation." International Journal of Engineering Applied Sciences and Technology (2021).

[7] Swami, Mandeep Nim1 Kunal. "Sales Forecasting of Global Automobiles Trends using ARIMA Model."

[8] Shakti, Sana & Hassan, Mohan & Zhenning, Yang & Caytiles, Ronnie & Iyenger, N Ch Sriman Narayana. (2017). Annual Automobile Sales Prediction Using ARIMA Model. International Journal of Hybrid Information Technology. 10. 13-22. 10.14257/ijhit.2017.10.6.02.

[9] Irhami, E. Arif Farizal, F, Forecasting the Number of Vehicles in Indonesia Using Auto Regressive Integrative Moving Average (ARIMA) Method, IOP Publishing, vol. 1845, pages 1742-6596, 2021, DOI 10.1088/1742-6596/1845/1/012024.

[10] Permatasari, Carina Intan, Wahyudi Sutopo, and Muh Hisjam. "Sales forecasting newspaper with ARIMA: A case study." AIP Conference Proceedings. Vol. 1931. No. 1. AIP Publishing LLC, 2018.

[11] KAYA, Sema KAYAPINAR, and Özal YILDIRIM. "A PREDICTION MODEL FOR AUTOMOBILE SALES IN TURKEY USING DEEP NEURAL NETWORKS." Endüstri Mühendisliği 31.1 (2020): 57-74.

[12] O. Haffner, E. Kučera and M. Moravčík, "Sales Prediction of Svijany Slovakia, Ltd. Using Microsoft Azure Machine Learning and ARIMA," 2020 Cybernetics & Informatics (K&I), 2020, pp. 1-9, doi: 10.1109/KI48306.2020.9039875.

[13] Zhang, Mu, Xiaonan Huang, and Changbing Yang. "A sales forecasting model for the consumer goods with holiday effects." Journal of Risk Analysis and Crisis Response 10.2 (2020).