Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241348

Analysis of the implementation for big data analysis in sales prediction: LSTM, ANN and DNN

Zhuofu Chen

Liberal arts college, Saint Micheal's College, Colchester, the United States

Zchen3@mail.smcvt.edu

Abstract. As a matter of fact, big data analysis is a choice made by the sales forecasting industry in response to the trend of the times, especially with the rapid development of computation ability and machine learning models. In general, it consists of two parts, i.e., big data and machine learning. To be specific, machine learning has received widespread attention in this field after being supported by big data. On this basis, this study will select the implementation of three machine learning models, LSTM, ANN and DNN. The prediction accuracy of these three models all meets practical requirements, and their performance in complex data is better than traditional models, but their costs are still not affordable for most companies. The purpose of this article is to help readers understand the development of big data analysis in the sales forecast industry, its current advantages and disadvantages, as well as possible future development directions.

Keywords: Sales forecasting, Machine learning, Big data analysis, LSTM, DNN.

1. Introduction

Sales forecasting is an important part of sales operations, which can help managers better allocate resources, formulate plans, and help the company gain stronger competitiveness. Therefore, this part has always received attention, and its development and progress have never stopped from the past to the present. As early as the early 1990s, major companies have realized that sales forecasting will play an important role in business activities, and related positions and work have been set up and carried out long ago.

The main forecasting methods at that time were mainly divided into two categories, known as qualitative (also called "judgmental") methods, which is a method of subjective assessment based on the speculator's experience and intuition, often as a "last resort" when no relevant data is available. The other is called a quantitative method, which is an analytical estimation method using past sales history or related variable relationships. It is roughly divided into two parts: time series and causality. The models built based on these two methods were indeed able to achieve predictions and helped the company at that time achieve positive success in the short and medium term forecasts [1].

However, the problems of the model at that time were equally obvious. Model training required a large amount of data and took a long time, and there were large errors in the prediction results of long-term or complex data [1]. For forecasters at the time, the primary task was simply to provide a company's senior management with accurate point estimates based on past sales history [2]. This fixed thinking made the industry at that time more rigid, and the functions of the models also became single. Past interviews and surveys of relevant practitioners have shown that many of them hope that the software

DOI: 10.54254/2755-2721/50/20241348

tools they use can be adjusted to their own needs, which can help them to achieve more effective data collation and analysis to help achieve more accurate sales forecasts [2, 3]. Therefore, in response to market demand, various types of algorithm models and more flexible or targeted software tools have begun to be developed and tried to be applied in the field of sales forecasting.

In 1980, the technology of machine learning (ML) has been tried to be applied to sales foresting, such as the Artificial Neural Network(ANN). The result of these methods showed good potential. However, because computer technology at that time was insufficient to support machine learning to obtain enough suitable data for model training, this method was not recognized in the field of sales forecasting [4]. After 2000, computer and Internet-related industries have developed greatly, and the convenience they bring has allowed computer networks to quickly cover various industries around the world. Huge amounts of data are constantly being produced from global computer networks. This situation has made all organizations that collect information realize that if they can filter, guide and utilize this information, the rewards can be huge [5]. In order to effectively analyses and utilize this data, the technology of data storage and processing capabilities is also constantly improving. In this trend, the concept of "big data" emerged as the times require.

For the sales industry, "data processing through information systems to generate knowledge is crucial for decision makers" [4]. Therefore, sales forecasting needs to keep up with this trend and use it to support competitive advantage. "Big data analysis" has become one of the answers to new trends in the sales forecasting industry. Its precise definition is given by Hofmann [4], and its main purpose is to transform information into useful knowledge. Big data analytics (BDA) is a combination of big data and machine learning (ML) technologies that complement each other. This model is also recognized by the industry and serves as the main research and development direction in the future.

In this context, this article will collect three existing models that utilize machine learning algorithms for big data analysis and their practical applications in sales forecasting. The overall trends and situation of the current industry are then sorted out by analysing its inputs and results. This article will provide the model principles of the three mainstream models, as well as the actual parameters and results for subsequent analysis.

2. Basic descriptions

The definition of sales forecast is to predict a practical amount for the sales quantity and sales amount of one or more products within a specific period of time through a comprehensive analysis of historical data and the possible impact of various factors in the future [2]. When choosing to use artificial intelligence technology to assist sales forecasting, the three elements of "input", "output" and "model" need to exist to ensure that big data analysis can run. Through the definition of sales forecast, it can be determined that the "input" corresponds to the historical data and the factors that can affect the forecast results. However, according to the actual situation of the paper, different forecast target products will be affected by different factors. At the same time, the prediction performance of the model is also related to the influencing factors as variables. Therefore, the specific types of input data factors need to be determined based on the characteristics of the target products and the sales environment, and the best types of model input variables are determined based on subsequent test results [6-8]. "Output" should correspond to the realistic sales quantity and sales amount defined as the forecast result. "Model" should be an analysis method. In machine learning it should represent various algorithm models. The three algorithm models chosen to study in this article are Long Short Term Memory (LSTM), Artificial Neural Network (ANN), and Deep Neural Network (DNN). Next, this article will introduce the three models, bring in data and analyse the results.

3. LSTM

Long Short Term Memory (LSTM) is a neural network structure algorithm and a derived model of RNN. The purpose of its emergence is to solve the problem of the data dependence and gradient disappearance performed in the long-term tasks of RNN. The application of LSTM has performed relatively well in fields such as time series prediction and natural language processing. The ability of LSTM to adaptively

DOI: 10.54254/2755-2721/50/20241348

adjust the amount of historical memory and new information currently available is due to the structure of its cell state [9]. The central ideas behind the STM architecture are memory cells, which maintain their state over time, and nonlinear gating cells, which regulate the flow of information into and out of the cells [10]. A unit, an input gate, an output gate and a forget gate, these four parts are the key parts that make up an LSTM unit. The data for this research comes from the article LSTM-based Sales Forecasting Model. The purpose is to use real sales data of clothing products sold by Company A in South Korea from January 1, 2015 to December 31, 2019 to analyze the relationship between the sales and temperature changes, which use long short-term memory (LSTM) to make sales predictions [6].

The target products selected to validate the sales forecast model were shorts, flip-flop, and winter coats. The actual sales data used for input are the product date, product category, and quantity of items purchased for these three items. The input data for the influencing factors in this experiment uses the daily average temperature of the corresponding date. This part of the data is collected from the Korea Meteorological Administration. There are a total of 1825 days of data, the data used for training model will randomly select 1645 days of data from 1825 days, and the remaining 180 days of data were used to compare and analyze the prediction results. According to previous study, one can also understand that the prediction results of flip-flops and winter coats fluctuate more than the prediction results of shorts, which means that the prediction accuracy of the shorts model is higher. The reason is that the actual sales volume of shorts is much higher than that of the other two products, so its model has more data for training. This situation demonstrates the properties of the LSTM model, where using more sales data in training will lead to more refined sales predictions [6-10].

4. ANN

An Artificial neural network (ANN) can be defined as a numerical model developed in a similar way with the structure of the biological nervous system [7, 11]. Therefore, its overall structure is very similar to that of the brain using a network of interconnected cells called neurons. It is precisely because of this special structure that artificial neural networks can learn complex non-linear relationships [12]. It is a fairly flexible algorithm that can tolerate errors and missing data and easily adapt to possible changes in the model. The ANN will change depending on its chosen activation function, network topology, training algorithm and other parameters. The role of an activation function is to convert a neuron's net input signal into a single output signal for further broadcast in the network [11]. Network topology or architecture describes the number of neurons and layers (i.e., groups of neurons) in the model and how they are connected [11]. The training algorithm specifies how to set connection weights so that neurons are inhibited or excited proportionally to the input signal [11]. Therefore, the specific structure of ANN needs to be selected according to the actual target.

The research data for this link comes from an external article on the practical application of ANN models [7]. The target product of this sales forecast is sports shoes in Istanbul shopping mall stores. According to the characteristics of the target product, the study sets air temperature, special holidays, product prices, and product discounts as independent variables, and product sales as the dependent variable. The data range is weekly sales data from 2014-2017. In order to avoid unbalanced or extreme data from affecting the results of the model, these data will first be processed using normalization. The training data will contain 80% of the processed data, and the training data will be the remaining 20%. Subsequently, multiple artificial neural network structures were tested, and the best learning model was obtained through error judgment: By comparing the MSE and MAPE of multiple test results, one found that the ANN structure in the figure above can obtain the best prediction performance in this experiment. Using the reserved test data to compare with the predicted data, the results are shown in Fig. 1. The results showed that when analysing store sales based on weekly averages, most stores had an error rate of less than 10%, which means that the ANN model is suitable for sales prediction of the stores in the dataset.

DOI: 10.54254/2755-2721/50/20241348

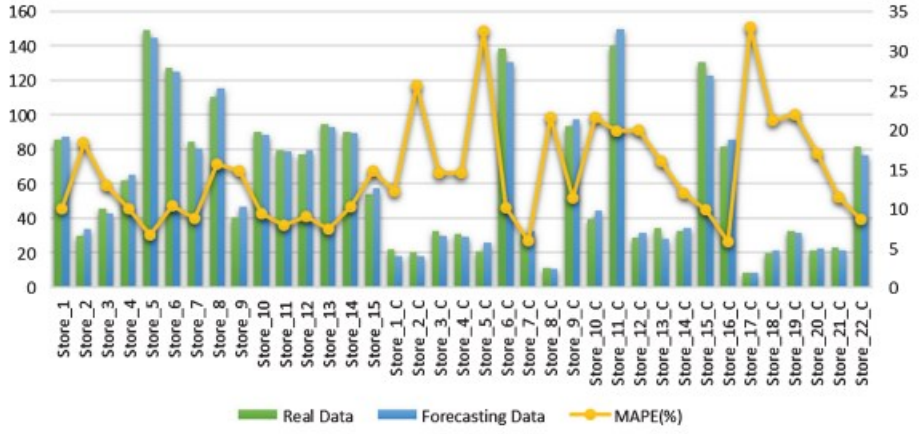

Figure 1. Prediction results [7].

5. DNN

DNN can be understood as a neural network with many hidden layers, and its structure is much like a complex artificial neural network [9]. The core idea is to automatically learn a large amount of observation data through the virtual neuron layer to identify the underlying patterns and classify the data [8]. The structure diagram of DNN is shown in the Fig. 2, which consists of an input layer, two or more hidden layers, and an output layer [13]:

- Input layer: identifies basic elements from the received raw data and then sends them to the hidden layer.
- Hidden layer: transfers the extracted data representation to the next layer.
- Output layer: Classifies the received data into predetermined categories.

In each layer, neurons perform complex nonlinear calculations, and the results are assigned a weight as output. The next layer will get the weighted outputs that combined through a linear transformation [8]. Complex data will be well handled by this combination of structures, but it is also prone to overfitting problems, which can be alleviated by using regularization techniques [13].

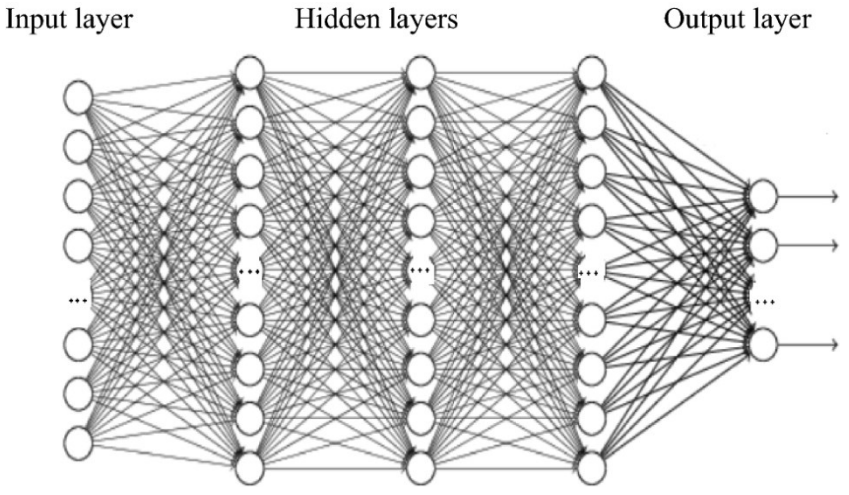

Figure 2. A sketch of DNN [13].

The data for this experiment comes from the article researching sales forecasting using Google data by DNN model [8]. The forecast target is the car sales of Ford and Honda in Taiwan. The actual data used is provided by Google, and the time range is from 2009 to 2015. The model framework used by DNN in this experiment is a deep feedforward network, which is suitable for the characteristics of the

DOI: 10.54254/2755-2721/50/20241348

nonlinear one-dimensional structure of the input data. In addition to DNN, the study also used the prediction results of several other algorithmic models for comparative testing. The estimation accuracy is evaluated using the MAPE. The study uses actual data from 2009-2013 as training data to train the model, and the actual data from 2013 will be used as validation input data to predict sales in the 12 months of 2014. The validation data set is used to prevent model overfitting. Actual data from 2014 will be used to determine the optimal forecast model. Finally, the actual data in 2014 is used as input data to predict the sales volume in the 12 months of 2015, and the corresponding actual data is used to compare and determine the results. According to the analysis, the loss of train and validation MAPE almost overlap after 800 epochs, over-fitting didn't happen [8].

6. Limitations and prospects

Under the trend of big data, data has become one of the most valuable resources. Sales need to extract relevant information to gain competitive advantage, so sales forecasting has an increasingly strong need for the ability to quickly process large amounts of data and extract needed information. Machine learning has an important role in this industry due to its ability to effectively process large amounts of data, and has also shown considerable value and potential. According to the prediction results of the three models in this article, the prediction accuracy of the machine learning model trained with sufficient data is very high, and it is better than the traditional method in this regard. It can also handle data with complex variables, which expands the scope of sales forecasting. It can even identify hidden patterns in demand that can be used as a baseline for identifying new market trends [4]. Although it performs well in terms of accuracy, shortcomings in other aspects still exist. The high accuracy of machine learning requires the support of a large amount of training data. If there is a lack of reliable data basis, the deviation of the prediction results will be large. Therefore, applying machine learning to sales forecasting now requires strong data storage capabilities. It also requires certain data processing capabilities and professional related workers to assist in establishing appropriate models. These requirements require a lot of resources to implement and maintain.

Therefore, there should be two general development directions for machine learning in sales forecasting in the future. The first is to continue to improve model performance. For example, a combination of multiple algorithm models can be developed. This method can use the advantages of different algorithms to compensate for their respective shortcomings, which can significantly improve the performance of the model. Better-performing models can handle more, more complex situations. The second direction is to reduce application costs. One of the reasons why machine learning is difficult to apply is how companies store large amounts of sales data suitable for training. In order to reduce costs in this area, you can consider developing relevant data processing technology to extract suitable variable parameters from huge data and delete redundant data. In this way, computing costs can be reduced and the pressure on data processing can be reduced.

7. Conclusion

To sum up, the purpose of this article is to study the implementation of machine learning models in the field of sales forecasting based on big data environment. With the support of huge data streams of big data, the prediction accuracy of machine learning models is better than traditional methods. Moreover, compared with traditional models, machine learning models can process complex data, which allows various external environmental factors to be input into the model as variables in the forecast, which expands the forecast range of sales forecasts and help predict results to be more comprehensive. In addition, there are many excellent online tools that can assist in the establishment of machine learning models, making the implementation of the model less complex and difficult to establish than traditional models. However, the cost of maintaining machine learning models is still high. Data storage, data processing and professional talents all require considerable resources. Therefore, in the future, machine learning models need to continue to improve performance while considering how to reduce their costs to help them expand in the field of sales prediction. This article demonstrates the current application of big data analysis in the field of sales forecasting, and shows the actual implementation of the three

DOI: 10.54254/2755-2721/50/20241348

models. Let readers understand the advantages and disadvantages of machine learning in this field, which can help readers make a choice between traditional methods and machine learning, and provide insights into subsequent development directions.

References

[1] Chase C W Jr 1997 Selecting the appropriate forecasting method. The Journal of Business Forecasting Methods and Systems vol 16(3) p 23
[2] Chase C W Jr 1999 Editorial: Sales forecasting at the dawn of the new millennium? The Journal of Business Forecasting Methods and Systems vol 18(3) p 2.
[3] Mentzer J T and Kahn K B 1997 State of sales forecasting systems in corporate america. The Journal of Business Forecasting Methods and Systems vol 16(1) pp 6-13.
[4] Martins E and Verardi G N 2023 Sales forecasting using machine learning algorithms. Revista de Gestao e Secretariado vol 14(7) pp 11294–11308.
[5] Dewey J 2022 Big data. Salem Press Encyclopedia vol 1.
[6] Hong J K 2021 LSTM-based Sales Forecasting Model. KSII Transactions on Internet and Information Systems vol 15(4) pp 1232.
[7] Caglayan N, Satoglu S I, Kapukaya E N and Kahraman, C 2020 Sales forecasting by artificial neural networks for the apparel retail chain stores-An application. Journal of Intelligent and Fuzzy Systems vol 39(5) pp 6517–6528.
[8] Yuan F C and Lee C H 2020 Intelligent sales volume forecasting using Google search engine data. Soft Comput vol 24 pp 2033–2047.
[9] Hwang S, Yoon G, Baek E and Jeon B K 2023 A Sales Forecasting Model for New-Released and Short-Term Product: A Case Study of Mobile Phones. Electronics (Basel) vol 12(15).
[10] Shakti G and Rahul B 2020 Impact of Uncertainty in the Input Variables and Model Parameters on Predictions of a Long Short Term Memory (LSTM) Based Sales Forecasting Model. Machine Learning and Knowledge Extraction vol 2(14) pp 256–270.
[11] Türkbayrağí M G, Dogu E and Esra A Y 2022 Artificial intelligence based prediction models: sales forecasting application in automotive aftermarket. Journal of Intelligent and Fuzzy Systems vol 42(1) pp 213–225.
[12] Correia A, Lopes C, Costa S E, Monteiro M and Lopes R B 2020 A multi-model methodology for forecasting sales and returns of liquefied petroleum gas cylinders. Neural Computing and Applications vol 32(16) pp 12643–12669.
[13] Loureiro A L, Miguéis V L and Silva L F 2018 Exploring the use of deep neural networks for sales forecasting in fashion retail. Decis Support Syst vol 114 pp 81-93.