Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241543

# Analysis of implementation for advanced tools in sales prediction

**Yuan Huang**

Shenzhen College of International Education, Shenzhen 518060, China

s22404.huang@stu.scie.com.cn

**Abstract.** As a matter of fact, since the 20th century, companies have started using statistical methods to predict the future sales of their product in order to adjust sales strategy as well as optimize the unitality of the corporations. In recent years, with the help of machine learning thanks to the rapid development of computation ability, firms could predict the future sales of their products with a much higher precision based on the state-of-art scenarios and models. To be specific, machine learning models such as Decision Trees help firms predict the trend of their sales with high efficiency. On this basis and with this in mind, this study will investigate the implementation and application of advanced tools in sales prediction. With the appearance of neuron networks such as LSTM-RNN, firms are now able to predict with high accuracy. According to the analysis, this study concludes the result of research carried out on Decision trees, GBDTs, as well as RNNs. At the same time, the current limitations and prospects are demonstrated and illustrated.

**Keywords:** Sales prediction, RNN, LSTM, DT, GBDT.

## 1. Introduction

Sales prediction plays an important role in production, especially in goods with a short shelf life. Sales forecasting is one of the major tasks regarding sales planning. Regardless of the size of the enterprise and the number of sales personnel, the prediction of sales exerts effects on sales management in many aspects, such as planning, budgeting and the determination of sales. Sales forecast is to estimate the sales quantity and sales amount of all products or specific products in a specific time in the future. It is based on the full consideration of several influencing factors in the future, along with the sales performance of the company, through certain analysis methods to put forward feasible sales targets. This minimizes the stock of expired products. The technique first appeared in the 20th century, when companies started predicting sales based on past statistics. Sales forecasting gives insight into how a company should manage its workforce, cash flow and resources [1]. Therefore, predicted sales play an important role in a business's decision making and productions. This results in a need for machine learning algorithms that are able to carry out precise predictions. Data and precision could vary greatly depending on the model selected, and the possible loss of data and unpredictable accidents, such as natural disasters, makes the prediction even more difficult.

Despite the importance of sales forecasting, producing sales forecasts with a high quality is not an easy task. Understanding the factors that have an effect on sales forecasting is important before making a forecast and selecting the most suitable forecasting method. Some industries, such as food and beverages, have already started using industries. Moreover, an extensive portion of the goods sold in that market is sensitive to some form of seasonal change because of the different cultural habits, religious holidays, fasting and other factors. All these factors contribute to the fact that some types of goods are sold mostly during the limited period(s) of time [2]. Several popular methods of predicting sales are DT, GBT, and RNN.

## 2. Basic descriptions

Sales prediction often uses the statistics from past sales, including sales channel, revenue, the unit sold, as well as the supply and demand of the market and the data from the competitor. However, there could be other factors influencing its accuracy apart from the training data, such as unexpected extreme climate, as well as other natural or human-caused disasters that could potentially affect the amount of supply or demand of the product on the market.

At the same time, one also summarizes the research on sales forecasting based on machine learning, which is mainly divided into four aspects: machine learning, data, effect and business. The first aspect is the machine learning aspect, that is, machine learning may be more concerned with correlation. To make sales forecasts, one used only factor variables related to changes in sales, not causal variables. Some customers just use things like inventory, price and holidays, which have nothing to do with sales. However, it is not the default that high inventory will definitely affect the increase of sales, and low inventory will definitely affect the decrease of sales. Because the latter is a causal relationship, whereas one uses a correlation relationship for machine learning. The second aspect is the level of data. Without data or poor data quality, the effect will be poor. The third aspect is the effect level. Effectiveness is evaluating a model for good or bad. For sales forecasting, its effect may be mainly reflected in whether it increases the profit of the enterprise. However, this effect is not very good to comment, because it not only considers the accuracy of the prediction, the interpretability of the model and other algorithm effects, but also considers the supply chain of the enterprise, the overall ability and so on. The prediction effect of machine learning cannot be used as the sole criterion to measure whether a company has increased profits. The fourth aspect is the business level, i.e., the data preprocessing before the machine learning training, the model evaluation during the training and after the training all need a certain amount of business theory as a guide. If the business theory is weak, it may affect the whole modelling process and it will have an impact on its effectiveness. Scholars are from the perspective of algorithm, to solve the problem of algorithm, but the problem of algorithm, ultimately back to our business problem, in sales forecasting one has to go back to how to improve performance.

After all, predicting sales is a business problem, and it is challenging to address this problem at the algorithmic level merely based on the data. One thinks this effect has been better, in fact, than the kind of big data precision marketing, accurate prediction or there is still a distance to go. This is the outlook and plan for using machine learning to do industry landing.

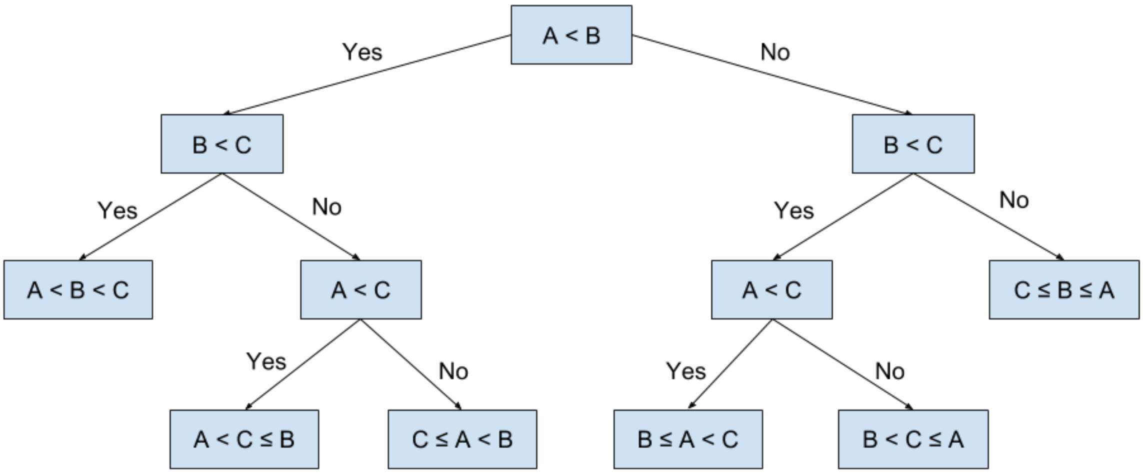

**Figure 1.** Example of a decision tree [3].

## 3. Models

A DT (decision tree) algorithm is a supervised machine learning algorithm that can be used for classification and regression. It is easy and fast use compared to neural networks although it's functions could be often limited. Each node in a decision tree contains a question relative to a particular attribute. Leaf nodes are groups of instances that receive the same class label [4]. A sketch is shown in Fig. 1.

Decision trees are commonly used to estimate risk as well as forecasting. For example, decision trees could be used to estimate potential activities based on the weather, humidity, temperature of the day, as well as any other real-life data. In sales forecasting, the possible training data could be obtained from historical sales data. Gradient boosting is another common algorithm used in classification and regression in machine learning. This model builds on the idea that, by using the boosting method, a combination of weak students will generate a strong student [3]. In the case where decision trees act as the week learners, the strong student is called a Gradient boosted tree(GBT). DT is generated by a form that is typically represented as a statistical classifier. This model can act as a regressor as well as a classifier. A typical DT includes Nodes and Branches. Each node requires problems that are based on one or more properties, i.e., comparing an attribute value with a constant or using other functions to compare more than one property [5]. Decision Trees used for sales prediction are commonly trained using past sales data, and predicts results based on the group the values of the different features belongs. A past sales data could be obtained. An algorithm that the Decision tree generates are shown in Fig. 2 and some of the results are shown in Table 1 [6].

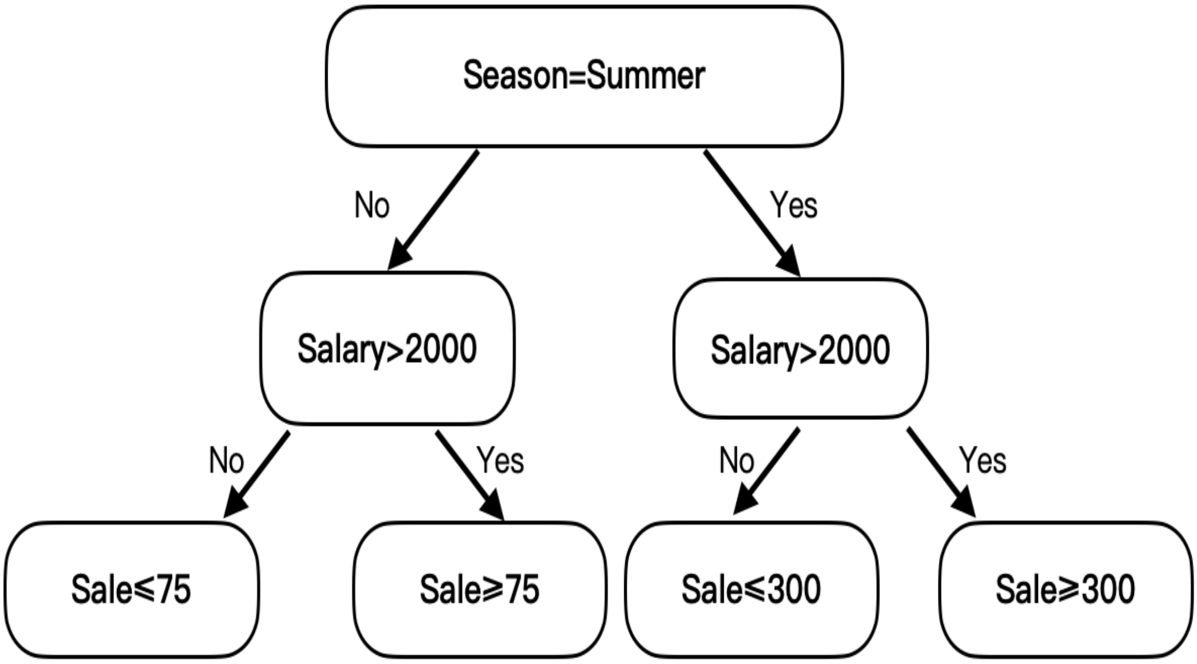

**Figure 2.** Algorithm of a decision tree [6].

**Table 1.** Analysis results.

| Year | Average salary | Season | Number sold |
|---|---|---|---|
| 2017 | 2000 | Summer | 300 |
| 2018 | 3000 | Summer | 450 |
| 2019 | 3000 | Autumn | 100 |
| 2020 | 1000 | Autumn | 50 |

Recurrent Neuron Networks(RNN) is one of the two broad types of neuron networks. It is a bi-directional neuron network meaning that it allows the output from some nodes to affect the inputs of these nodes. This example of a decision tree takes the season and the average salary of it's potential customers as an input. With this tree, one could predict the possible sales. For example, if such a product is to be sold in Summer, and the average salary of potential customer is 400, then a possible predicted number of sales could be 500. Compared to traditional decision trees, gradient-boosted decision tree is a collection of many weak decision trees. The core of GBDT is to accumulate the results of all trees as the final result [6]. Therefore, it takes in the same parameters as an input as traditional decision trees. As it is a strong learner made up of many weaker learners, it is capable of outperforming decision trees.

Compared to decision trees, RNNs require more parameters such as the number of hidden layers while training. RNNs are able to capture sequential information by carrying results from previous computations or states into the next states [7]. However, neuron networks take up considerably more time to train compared to decision trees.

GBDT is commonly used for regression and classification. In a research conducted by Zhou et al. using Walmart sales data, their LightBGM model, which is a model based on the GBDT model framework, produces the most accurate result. The RMSE(Root Mean Squared Error) result is 2.07, while the traditional linear regression model had a RMSE result of 3.20, and their support vector model produces a result of 2.85. The LightGBM has the advantage of a higher training efficiency, supports the processing of large-scale data, and has higher prediction accuracy [8]. A similar research done by Deng et al. found that LightGBM produced the least error compared to linear regression and SVM [9]. Recurrent neuron network models such as Long short-term memory and echo-state network are extremely popular in forecasting, due to its high accuracy. According to a research done by Feng and Chen, whose experiment consists of a comparison of the accuracies between different machine learning models on sales forecasting, the LSTM model has a higher accuracy compared to XGBoost, although it is less stable [10]. Chandrah and Naraganahalli 'sresearch on the forecast of automobile parts sales in 2021 shows that the RNN-LSTM model has a better performance compared to the SES, Croston, SBA, TSB and modified SBA while forecasting the sales on automobile parts using the data from the Norwegian Road Federation [11]..

## 4. Limitations and prospects

Machine learning could provide a rather high accuracy on forecasted sales, and is therefore frequently used by companies and organizations worldwide. However, there are several limitations. Models used for predicting could only be used to predict sales that follow a certain trend or pattern, making it predictable. Sales that are random could not be predicted accurately. The main difference between e-commerce sales prediction and entity sales prediction is in user experience. One calls e-commerce sales prediction online prediction, while entity sales prediction, such as clothing sales prediction, pharmacy sales prediction and stationery sales prediction, belongs to offline prediction. For online prediction, in addition to the impact of the business environment itself (price, inventory, quality, evaluation, etc.), there is also a part of the factors affecting sales is the user behavior data on the Internet, mainly including browsing, clicking and collecting data. These user behavior data are mainly active through the distribution code technology (by deploying the crawled code on the web or PC). At the same time, it is also necessary to consider the lag of user behavior data, that is, a user has collected the product, but it takes a certain period of time to place an order to buy it, at this time, it is necessary to split the variable with lag. Then, one added it to the machine learning model. The sample size depends on the prediction target, short prediction target, the need for less training set, if the prediction target is longer (more than 5 days), generally need more than 1 year of historical data. As for accuracy, it depends on the data. The accuracy indicator is usually RMSE, and the smaller the RMSE, the better the accuracy. Apart from this, the actual results could be affected by unexpected events which could have a huge impact on actual sales, and these possible changes could not be foreseen by models.

## 5. Conclusion

To sum up, all models take in past sales data as input. According to the analysis, decision Trees have the least accuracy among the models discussed in this paper. RNN models produces the most accurate results compared to most other regression algorithms, however, it is less time efficient. Gradient Boosted Trees excel in time efficiency and produces more accurate results that traditional Decision Trees, however it is less accurate compared to RNNs. From the whole points, these results and analysis pave a path for further analysis, evaluations as well as developments for sales prediction in terms of the machine learning scenarios.

## References

[1] Cheriyan S, Ibrahim S, Mohanan S and Treesa S 2018 Intelligent Sales Prediction Using Machine Learning Techniques International Conference on Computing Electronics and Communications Engineering (iCCECE) pp 53-58.

[2] O'dell C and Grayson C J 1998 Food Sales Prediction: If Only It Knew What We Know. California management review vol 40(3) pp 154-174.

[3] Wisesa O, Adriansyah A and Khalaf O I 2020 Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm 2nd International Conference on Broadband Communications Wireless Sensors and Powering (BCWSP) p 15.

[4] Thomassey S and Fiordaliso A 2006 A hybrid sales forecasting system based on clustering and decision trees Decision Support Systems vol 42(1) pp 408–421.

[5] Charbuty B and Abdulazeez A 2021 Classification Based on Decision Tree Algorithm for Machine LearningJournal of Applied Science and Technology Trends vol 2(01) pp 20-28.

[6] Qian H, Wang B, Yuan M, Gao S and Song Y 2022 Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree Expert Systems with Applications vol 190 p 116202.

[7] Saxena P, Bahad P and Kamal R 2020 Long short-term memory-RNN based model for multivariate car sales forecasting Int J Adv Sci Technol vol 29 pp 4645-4656.

[8] Zhou Y, Cheng Y, Lin Y and Mengqiu T 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) p 21.

[9] Deng T, Zhao Y, Wang S and Yu H 2021 Sales Forecasting Using GBDT Based Model And Data Mining Method IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) pp 11-18.

[10] Feng C and Chen Z D 2019 Sales Forecasting Based on LightGBM Computer Systems and Applications vol 28(10) pp 226-232.

[11] Chandriah K K and Naraganahalli R V 2021 Application of Weighted Combination Model Based on XGBoost and LSTM in Sales Forecasting Multimedia Tools and Applications vol 80(17) pp 26145-26159.