Supermarket Sales Prediction Based on Xgboost Classifier Model

Yuhao Yang *

Jinan University-University of Birmingham Joint institute, Jinan University, Guangzhou, China

* Corresponding Author Email: yyhjbji@stu2020.jnu.edu.cn

Abstract. As a matter of fact, it is necessary for the manager to have a better understanding of supermarket sales nowadays. With this in mind, it is significant to have a model to analyze the sales in supermarkets to help the manager. On this basis, in this article, the XGBClassifier model is used to do the prediction of supermarket sales. There are 750 training sets to fit the model and 250 testing sets to judge the performance. The accuracy of the model is 0.49 and the train score is about 62.67. The model is a normal way to do prediction, it is good to fit the data with big sets. This research is helpful for the people who want to manage a supermarket, it can give some important advice to improve the income of the supermarket. It will be more meaningful after training a big set of data because of higher precision.

Keywords: Supermarket sales; Xgboost; classifier.

1. Introduction

Contemporarily, a company's success depends heavily on its capacity to swiftly adapt to market developments. Accurate and timely sales forecasting is essential for companies in the manufacturing industry, other industries include wholesale and retail (enhancing profitability and service standards), marketing, logistics, and others [1].

Several forecasting strategies make use of data analytics and predictive modeling. By analyzing data, predictive modeling creates data models that predict outcomes. A few classifiers that may be used to predict sales include regression, KNN, random decision trees, demand forecasting, classification techniques, cluster, and Bayesian classification. The main classifier used in this study is regression. The regression model equation $Y = a + bX$ is rather straightforward because your sales would stand in for Y, the intercept would stand in for a, and the slope would stand in for b [2]. Sales forecasting typically use statistical techniques such random tree, ARIMA, the classification model, regression models. Innovative techniques like NN or data-mining algorithms are used in modern procedures. Hybrid models are developed more regularly to combine the strengths of many models in a novel way to enhance Sunitha et al. compare three models [3]. The results show that Gradient Boost Algorithm has a 0.98 overall accuracy rate, followed by Decision tree Algorithms with a 0.71 accuracy rate and Generalized Linear Model with a 0.64 accuracy rate [4]. Vito et al. contrast linear models found using the LSE method with fuzzy models found utilizing evolving approaches. The findings suggest that it may, given the right parameters, do well on smaller data sets [5].

Gopal and Nain use xgboost technology. They conclude that merchants or shopping malls are more adept at foreseeing daily variations in customer demand or product sales. At the organizational level, in-depth study is being done for accurate sales forecasting. Because a company's earnings are strongly tied to how effectively it forecasts sales, reducing losses for the firm, Big Marts desire a more accurate prediction system. At the organizational level, in-depth study is being done for accurate sales forecasting [6]. This study will construct Xgboost model to realize the classification.

2. Data and Method

The data is from Kaggle which is the historical record of sales data in 3 different supermarkets. It is collected by Aung Pyae. The ensemble modeling method known as "boosting" is to construct a strong classifier by merging all the different classifiers. A model is constructed by layering weak

models. First, a model is built using the training data set. The shortcomings of the original model are then addressed by the next model. The process of adding models will go on indefinitely, or until the maximum value of models be added or the whole training data has been correctly in prediction. One of the most popular ways of boosting is gradient boosting. Each prediction in gradient boosting corrects the mistake of the data before it. Each predictor in this method is training to use the previous residual mistakes as labels.

XGBoostClassifier is implanted by Gradient Boosted decision trees. Then, each decision tree's prediction scores are added up to determine.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \hat{F} \tag{1}$$

where, K is the quantity of trees, f is the functional space of $\hat{F}$, $\hat{F}$ is the data. Target function to do optimization for the above model is given by:

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where, first term is the loss function and the second is the regularization parameter. The model additive strategy:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) \tag{3}$$

The objective function of the above model can be defined as (Taylor series expansion)

$$obj^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + constant \tag{4}$$

Now, one defines the regularization term:

$$f_i(x) = \omega_{q(x)}, \omega \in R^T : R^d \to \{1,2,\dots,T\} \tag{5}$$

T is the quantity of leaves, w is scores in vector on each tree leave, and q is the function that correspond each data poinst to its matching leaf. The regularization term is [7]:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \tag{6}$$

and objective function [8]:

$$obj^{(t)} = \sum_{j=1}^{T} [G_j W_j + \frac{1}{2} (H_j + \lambda) \omega_j^2] + \gamma T \tag{7}$$

where $G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$. The score of measure how good the tree is:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{8}$$

This score is to evaluate the performance of the model.

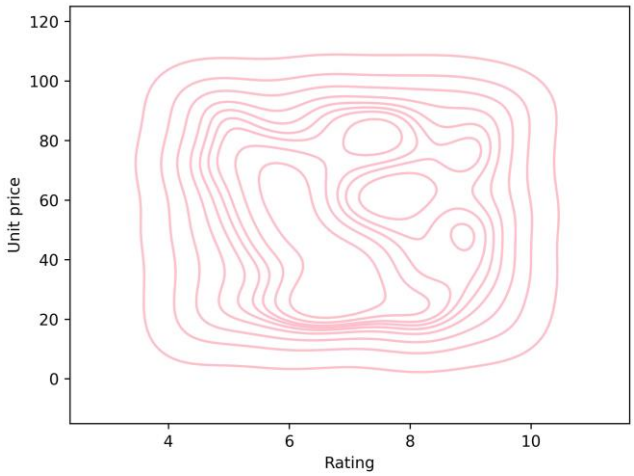


Fig. 1 Probability density between Unit price and Rating.

3. Results and Discussion

3.1. Feature Engineering

This study denotes the data of markets with different conditions, each condition can affect the total sales of the market. It is clear to analyze the data which can affect supermarket sales. It is helpful to take it into the prediction model. As a matter of fact, the frequency of unit price, quantity and rating is steady. Frequency of the other 5 variables is not uniform. It is useful to determine which variable is more important in this analysis. Based on the correlation analysis, it can be helpful to analyze the closeness of each two variables. If the value is closed to 1, the two variables are closed. Conversely, if the value is close to 0, which means that these two variables with less relation. The most relative features are Tax 5%, cogs and gross income. It concludes that rating is almost independent with other variables in this data. Kernel Density Estimate, often known as the KDE Plot, is used to display the probability density of a continuous variable. It visually displays the probability density at different levels in a continuous variable. A common method of illustrating the distribution of data based on a five-number summary is shown in Fig. 1. "minimum", "first quartile (Q1)", "median," "third quartile (Q3)," and "maximum"). Visual data analysis is beneficial.

3.2. Models Evaluation

Because variables with non-numeric data cannot be applied to prediction model, it is necessary to transform them into numeric data.

Randomly choose 750 sets as the training data, and the other sets as the testing data. Let the data in the label Gender as y_train, and the data in other labels as x_train. Then fit x_train and y_train in the model. To judge the performance of the model, it is necessary to take the results of the model to predict x in the testing sets. The simple process of this prediction model is first construct a decision tree of the data, then work with the determining tree and do pruning to improve the efficiency. After classification, there will be the prediction value of each variable in the model. Then using the score standardized equation to gain the score of this prediction model in testing sets. This score denotes the performance of this prediction model. There is also the precision and accuracy of the model in results.

By using the XGBClassifier model, it will get reports about accuracy, confusion matrix and training score after prediction. The training score is about 62.67, it is not very high and not very low, which means that the model is not bad. The results are shown in Table 1 and Table 2 with training score of 62.67.

Table 1. The classification report after doing prediction by XGBClassifier model.

	Precision	Recall	F1-score	Support
0	0.49	0.56	0.52	124
1	0.49	0.41	0.45	126
Accuracy			0.49	250
Macro Avg	0.49	0.49	0.49	250
Wighted Avg	0.49	0.49	0.49	250

Table 2. The classification confusion martrix.

	Predict Positive	Predict Negative
Real Positive	70	54
Real Negative	74	52

3.3. Explanation and Implication

First do regularization to reduce the sensitivity, then construct a tree and pruning to prevent overfit. Dividing the data into blocks that can be used in parallel to create the trees or for other computations. By attempting both routes in a split and determining a default direction by calculating the gain,

XGBoost manages data sparsity. In detail, The XGBClassifier model is used to fit the training sets and judge the testing sets in this project. In this project, choosing 750 training data and 250 testing data randomly. In this model, fitting the Gender label other labels. After constructing a decision tree, it is easy to gain the prediction value of the data in the model. Then calculating the score of this prediction model in testing sets using the score standardized equation. The performance of this prediction model is shown by this score. The findings of the model's precision and accuracy are another factor.The estimated training score is 62.67. It implies that it is possible to predict the supermarket sales with different conditions. It will be more accurate if there is enough data and conditions. It is useful to give some simple advice for the manager of a supermarket. And there are also many methods to improve the performance of XGBClassifier model like introduce data with bigger size. It is useful to help the manager to know the supermarket sales in the future easily.

4. Limitations and Prospects

XGBClassfier in this project gains a score of about 62.67, which is not good. A better predictive model than XGBClassifier will exist. When training large models, XGBoost could need a lot of processing power, which makes it less suitable for systems with constrained resources. XGBoost can overfit, especially if a small dataset was used for training or if the model has an excessive number of trees. CatBoost and LightGBM, two programs that are comparable to XGBoost, can handle category characteristics natively. Prior to employing XGBoost, categorical features must be encoded or converted. Because there are so many hyperparameters in XGBoost that may be altered, hyperparameter tweaking is crucial to optimal performance. However, it could require some work and knowledge to identify the ideal confluence of features. XGBoost can use a lot of memory when working with large datasets, making it less appropriate for systems with low memory. Additionally, in order to increase the model's accuracy, larger data sets are required [9]. Additionally, in this model, XGBClassifier performs worse than random forest classification. The data's insufficient size is the cause, after all.

XGBoost has a history of consistently delivering excellent results in a wide range of machine learning workloads. Because it is designed for scalable and efficient machine learning model training, XGBoost is excellent for large datasets. A wide range of hyperparameters are available in XGBoost and may be changed to optimize performance. For concurrent learning, XGBoost also offers a block structure. On multicore machines or clusters, it makes scaling up simple. Furthermore, it makes use of cache awareness to reduce memory usage when training models with huge datasets. Finally, by using disk-based data structures rather than ones that are stored in memory while doing calculations, XGBoost offers out-of-core computing capabilities. Prediction is useful in many different disciplines. If the data is more detailed, it will be more accurate to do prediction. It is also useful to improve the efficiency of XGBoost by introducing some new parameters. It will be more accurate and efficient as the theory updates [10]. it also needs people to collect more relative data of supermarket sales to train the XGBoost model, then it will perform better and gain a good score measure. There are still also many challenges of supermarket sales prediction in the future. For instance, the accuracy of data, a complex model would cause uncertainty in prediction. It needs to be searched new methods to do prediction.

5. Conclusion

In conclusion, this study is supermarket sales prediction by XGBClassifier model. It shows the relation of each variable in this data after Feature engineering. In this model, there are randomly 750 training data and 250 testing data. By fitting the data of label gender and data in other labels, and taking the model in the testing data, it will have the prediction values of the testing data. By using the score measure equation, the results are that the training score of XGBClassifier is about 62.67. The results are not good and not bad, it shows that a normal score and accuracy. The limitation is that it

needs a lot of computer power, and it is susceptible to overfitting. It will perform better if introducing some parameters to improve accuracy. From this research, it can provide an estimate of supermarket sales for the manager. It can make their work easier and more efficient. And XGBClassifier is also useful in many fields like computer science and biology. It is a great prediction model to solve problems, it can give people advice and make management easy in their works. The trends of supermarket sales prediction is that the rising of big data, the growth of online sales and the development of machine learning. The big data and great machine learning can help the performance of the model.

References

[1] Meulstee P, Pechenizkiy M. Food sales prediction:" If only it knew what we know". 2008 IEEE International Conference on Data Mining Workshops. IEEE, 2008: 134-143.

[2] Kohli S, Godwin G T, Urolagin S. Sales prediction using linear and KNN regression. Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019. Singapore: Springer Singapore, 2020: 321-329.

[3] Beheshti-Kashi S, Karimi H R, Thoben K D, et al. A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering, 2015, 3(1): 154-161.

[4] Cheriyan S, Ibrahim S, Mohanan S, et al. Intelligent sales prediction using machine learning techniques. 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, 2018: 53-58.

[5] Dovžan D, Logar V, Škrjanc I. Solving the sales prediction problem with fuzzy evolving methods. 2012 IEEE International Conference on Fuzzy Systems. IEEE, 2012: 1-8.

[6] Behera G, Nain N. A comparative study of big mart sales prediction. Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part I 4. Springer Singapore, 2020: 421-432.

[7] Liew X Y, Hameed N, Clos J. An investigation of XGBoost-based algorithm for breast cancer classification. Machine Learning with Applications, 2021, 6: 100154.

[8] Carmona P, Dwekat A, Mardawi Z. No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. Research in International Business and Finance, 2022, 61: 101649.

[9] Ma M, Zhao G, He B, et al. XGBoost-based method for flash flood risk assessment. Journal of Hydrology, 2021, 598: 126382.

[10] Li C, Zheng X, Yang Z, et al. Predicting short-term electricity demand by combining the advantages of ARMA and XGBoost in fog computing environment. Wireless Communications and Mobile Computing, 2018, 2018.