Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241487

Utilizing machine learning for sales forecasting in urban supermarkets

Yining Ma

Department of Economics, University of Washington, Seattle, USA

amamyn@uw.edu

Abstract. In contemporary society, the proliferation of supermarkets in densely populated urban areas is providing consumers with an abundance of choices and at the same time put supermarket companies into a highly competitive environment. In order to thrive in this market environment, it is of great importance for supermarket corporations to trace, record, and analyse their sales data so that they could have a better understanding of their customer behaviours. And machine learning is commonly applied in this sort of scenarios. By leveraging the insights generated through predictive analysis, machine-learning models can help companies to optimize inventory management, refine pricing strategies, and create effective marketing campaigns. This article applies three machine-learning algorithms to train and test the same data in order to find the model generating the best prediction of quantity sold for the target dataset. Linear Regression, Random Forest, and XGBoost are the three methods used to create three different machine-learning models. After evaluating the R-squared scores generated by three models, it turns out that the Random Forest Model displays the most precise prediction of the target variable, which generates an R-squared score of 99.27%. Thus, the Random Forest Model is the best model for sales prediction in this case.

Keywords: Machine Learning, Sales Forecasting, Predictive Analysis, Sales Data Analysis, Linear Regression, Random Forest, XGBoost.

1. Introduction

1.1. Research background

The proliferation of supermarkets in densely populated urban areas is experiencing a notable surge, accompanied by fierce and intense market competition. This burgeoning trend underscores the dynamic nature of the retail landscape in these areas, where consumers are presented with an abundance of choices, and retailers must continually innovate and strategize to thrive in this competitive environment.

Therefore, it is of great importance for the supermarket company to trace, record, and analyze their sales data. Additionally, studying customer behaviors and preferences is essential in order to fine-tune their sales strategies. The ultimate goal is to retain existing customers while simultaneously attracting new ones. In this ever-evolving landscape, the role of machine learning is steadily gaining significance. It has emerged as a potent tool for retail enterprises to harness, enabling them to forecast sales trends with increasing accuracy based on the analysis of historical sales data. By leveraging the insights

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/45/20241487

generated through predictive analysis, they can optimize inventory management, refine pricing strategies, and create effective marketing campaigns. This newfound capability allows them to stay ahead of market fluctuations and evolving customer preferences, ultimately propelling their business toward sustained growth and success.

1.2. Literature Review

The utilization of machine learning algorithms for data forecasting is widespread, with regression analysis being particularly prevalent in this field. In the book by Mentzer and Moon "Sales Forecasting Management", there is a comprehensive section dedicated to the discussion of regression analysis, including its applications and implications [1].

There are various approaches to performing a regression analysis. In the study conducted by Thomas Ng et al., the researchers employed four different leading indicator models to predict the housing supply in the Hong Kong private sector. These models included a linear regression analysis model, a genetic algorithms model, a GA-LRA, and a GA-LRA model with an Adaptive Mutation Rate [2]. In the research conducted by Zhang, the author utilized several regression models, including the linear regression model, Ridge regression, and XGBoost to forecast future sales at Walmart [3].

Two popular ways to compare difference forecasting models are the comparison of R-squared Error (RSE) and the Mean-squared Error (MSE) values of the models. An R-squared value higher than 0.7 could be considered as high precision, as stated in the research report by Smith et al [4]. Sometimes the predictions could be pretty precise, generating an accuracy rate up to 70% [3]. But the precision of the prediction on the same dataset can also be variable among different models, and a float of R-squared score from 0.5 to 0.95 is also possible [4]. In the research conducted by Thomas Ng et al., the authors provide two potential explanations for the errors and uncertainties generated by the predictive model. These explanations revolve around the interdependence of variables and the limited significance of certain economic indicators, which could be insightful for the interpretation of the results of this research [2].

1.3. Research aim

In this research, three machine-learning algorithms would be applied to train and test the same target data set in order to find the model which generates the best prediction of the target variable: the quantity sold. The three models would be: (1) Linear Regression Model, (2) Random Forest Model, and (3) XGBoost model. The way to compare and evaluate the accuracy of the prediction of each model would be a comparison of the value of R-squared score for each model.

A precise prediction of the target variable would help the supermarket company to accurately trace and record their sales data, enabling them to make more informed decisions across various aspects of their business.

2. Data source

The dataset “supermarket_sales” was obtained from Kaggle.com. It is a historical sales dataset of a supermarket company that consists of three branches of supermarkets in three different locations in Myanmar: Yangon, Mandalay, and Naypyitaw, which are the largest, second largest and third largest cities in the country [5].

The dataset covers sales data for 3 months, from January to March 2019. Each datum is recorded by the time when the order is placed. Thus, even when the entire dataset consists of less than a hundred days, there are 1000 rows of data recorded as each datum will be accurate to the second based on the order time. The dataset has 1000 rows of data and 17 columns of variables in total, including Invoice ID, Brance, City, etc. Among them, the variable “Quantity”, which stands for the quantity bought by customers, or the quantity sold by the vendor, is the target variable for this research.

For more details about all the variables of the original dataset, please refer to Kaggle.com for the concrete data.

DOI: 10.54254/2755-2721/45/20241487

3. Methodologies

In this research, three supervised learning methods were used for data modeling and prediction, which are Linear regression, Random Forest, and XGBoost regression.

3.1. Linear regression

Linear regression is a fundamental machine learning and statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. It is a supervised learning method, primarily employed for regression tasks, where the goal is to predict a continuous numerical output [6].

Linear regression is one of the most basic machine learning algorithms that is commonly used for data modeling and prediction. As the features of this research perfectly fit the requirements of this method (one target variable and multiple predictors), linear regression would be applied to create the first model.

3.2. Random Forest

Random Forest is a versatile and powerful machine-learning algorithm that belongs to the ensemble learning family. It is widely used for both classification and regression tasks. The fundamental idea behind Random Forest is to build multiple decision trees during the training phase and combine their predictions to achieve more accurate and robust results [7].

Because of Random Forest's exceptional robustness and reduced susceptibility to overfitting in comparison to individual decision trees, as well as its ability to handle both numerical and categorical data effectively, it will be employed as the secondary method for data training and prediction in this research.

3.3. XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is a popular and powerful machine learning algorithm used for regression tasks. It is an optimized and scalable version of the gradient boosting algorithm, designed to deliver high predictive accuracy and efficiency [8]. This method would be the last method for modeling in this research.

3.4. Data processing

Given that the target dataset, which is named “spmkt”, comprises 1000 rows of data and a total of 17 variable columns, it is assumed that there should be some missing values or incomplete data contained. However, a thorough examination of the dataset revealed that "supermarket_sales.csv" stands as a remarkably well-maintained and organized dataset with no instances of missing values within any of the 17 variable columns, which affirms the dataset's overall data integrity and completeness.

To streamline the dataset and optimize it for further analysis, several actions were taken. Specifically, within the original dataset, denoted as "spmkt," some of the 17 variables appeared unrelated to the target variable, "Quantity." Consequently, these irrelevant variables were excluded during the creation of a new dataset, aptly named "spmkt_data."

Additionally, to enhance data clarity and prevent redundancy, variables with overlapping meanings were consolidated. Notably, "Date" and "Time" were merged into a single variable, as were "Branch" and "City."

Moreover, to facilitate data analysis, categorical variables such as "Customer type," "Gender," and "Payment" underwent an encoding transformation. They were converted into dummy variables, a suitable format for analytical purposes.

As a result of these modifications, the refined dataset is now referred to as "spmkt_data_encoded." These adjustments collectively aim to provide a more focused and efficient dataset for subsequent data analysis tasks.

DOI: 10.54254/2755-2721/45/20241487

3.5. Exploratory data analysis

In order to explore the correlations between the encoded 14 variables in the target dataset "spmkt_data_encoded", a correlation matrix was constructed, and a heatmap was applied to visualize the matrix. The rest 14 variables were "Branch", "Unit price", "Quantity", "Tax 5%", "gross income", "Rating", "cogs", "Customer type_member", "Customer type_normal", "Gender_Female", "Gender_Male", "Payment_Cash", "Payment_Credit card", and "Payment_Ewallet".

Initially, the correlation matrix was created wishing to find which variables have the highest correlation with the target variable "Quantity". However, through the heatmap (Figure 1), it is shown that most of the variables do not display an obvious correlation with quantity sold as most of the correlation values are close to zero, and the color of the heatmap is overall pale. The only three variables that display a relatively strong correlation with quantity sold are "Tax 5%", "gross income", and "cogs". The result is reasonable since the target variable is related to transactional relationships, and as such, it is expected to exhibit a stronger association with variables that would influence transactional relationships, such as the cost and selling price of the products.

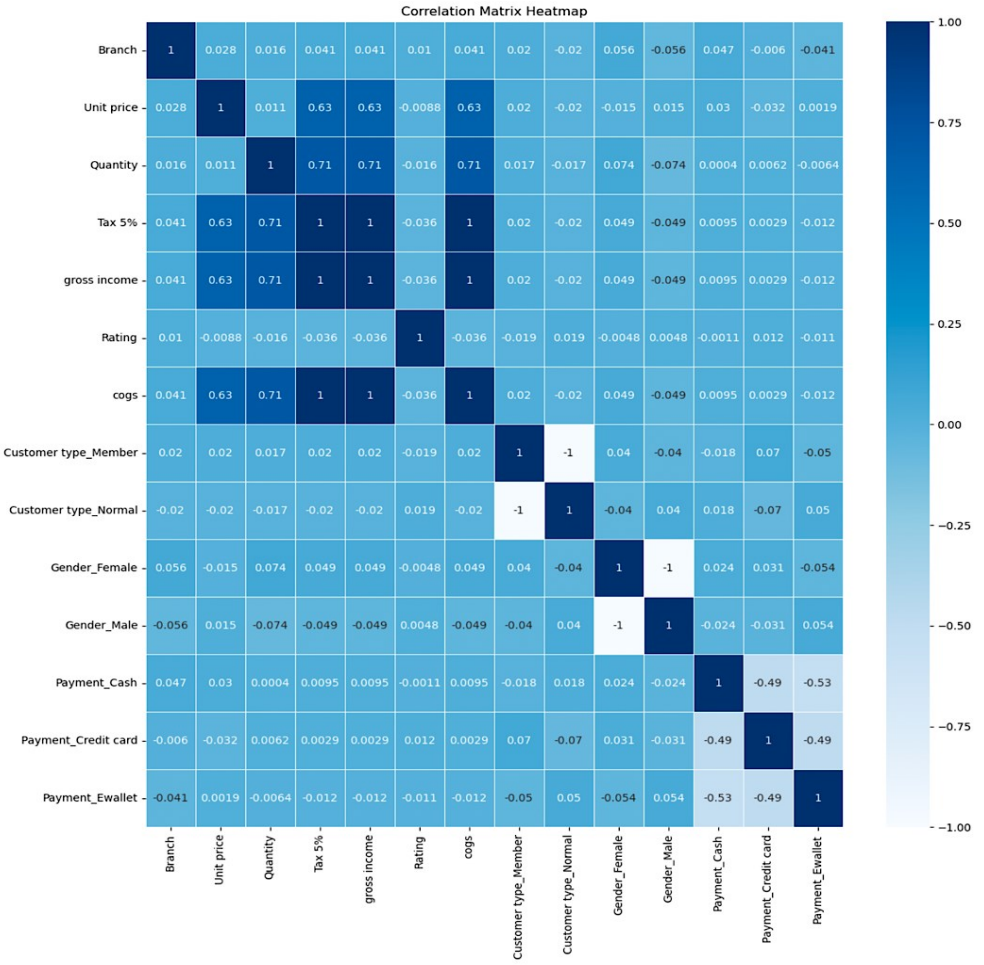

Figure 1. The correlation matrix heatmap of the selected 14 variables.

3.6. Data training and testing

Since the size of the dataset is not large enough for further filtering, the cleaned and encoded dataset "spmkt_data_encoded" is used for training and testing. The training and testing sizes were set to the most common splitting sizes, with 80% of the entire dataset being the training set and the rest 20% of the data being the testing set. The random state is set to 42 to make sure that the same data are

DOI: 10.54254/2755-2721/45/20241487

assigned to the train and test data each time. The training and testing set for independent variables and the target variable are named: X_training, X_testing, y_training, and y_testing.

The three models were created individually on the same range of datasets, with each applying the method of Linear Regression, Random Forest, and XGBoost.

3.7. Model Evaluation and Comparison

Three diagrams were created for each model to visualize the difference between the actual quantity and the quantity predicted by each model. To better compare the accuracy of predictions generated by the three models, the R squared score and adjusted R squared value were calculated. The formula of the calculation applied is as follows:

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)} = 1 - \frac{(\Sigma(y_i - \hat{y})^2)}{(\Sigma(y_i - \bar{y}_i)^2)} \quad (1)$$

Since the R-squared score is a measure of the error in the model, a higher R-squared score corresponds to a lower occurrence of error, which indicates a more accurate prediction and a better model.

4. Results

4.1. Linear Regression

The computed R-squared score for the Linear Regression model is 0.8258, while the adjusted R-squared score stands at 0.8136. This means that roughly 82.58% of the variance in the target variable can be attributed to the features incorporated into the linear regression model. With such a relatively high R-squared score, it's reasonable to describe the predictions as moderately accurate.

The diagram showing the difference between the actual and predicted value of the Linear Regression model is shown below (Figure 2). Since the variable on the x-axis (Quantity) is a discrete variable, the diagram seems a little bit different from the most common diagrams showing the gap between the actual and predicted values.

In Figure 2, it can be inferred that the predictions are closest to the actual value when the Quantity is around 6, and the error gets larger when the actual quantity gets both larger and smaller. The shape of the diagram is roughly symmetrical, with the center of symmetry located at around (6, 6).

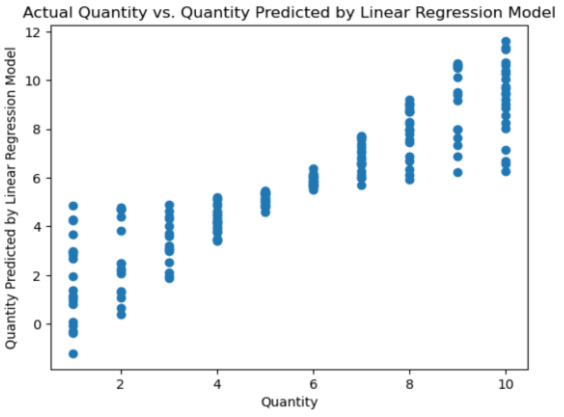

Figure 2. The scatter plot of The Actual Quantity vs. Quantity Predicted by Linear Regression Model.

As for the coefficients of the attributes in the linear regression model, it is shown that all the coefficients can be considered as "small" coefficients as their magnitudes are all significantly smaller than 1 (Table 1). That is, the value of all these coefficients is between -0.1 and 0.1. Therefore, all of

DOI: 10.54254/2755-2721/45/20241487

these variables generate relatively small impact on our target variable, and an increase in one unit of each of these features leads to a change in our target variable less than 0.01 unit.

Table 1. The coefficients of attributes of the Linear Regression Model

Attribute	Coefficients
Branch	-0.0226
Unit Price	-0.0813
Tax 5%	0.0007
Gross Income	0.0007
Rating	0.0340
COGS	0.0146
Customer type_Member	0.0234
Customer type_Normal	-0.0234
Gender_Female	0.0120
Gender_Male	-0.0120
Payment_Cash	0.0371
Payment_Credit card	-0.0996
Payment_Ewallet	0.0625

4.2. Random Forest

The R-squared score for the Random Forest model calculated is: 0.9927, and the adjusted R-squared score is: 0.9922. Therefore, approximately 99.27% of the variance in the dependent variable is explained by other features in this Random Forest model.

Compared to the R-squared score generated by the Linear Regression model, the R-squared score of the Random Forest model is even higher and is very close to 100%. Therefore, this Random Forest model generates a very precise prediction of the data.

A visualization of the conclusion drawn above is shown below (Figure 3), which shows the gap between the actual quantity and the quantity predicted in the Random Forest model. It is apparent that the error is much smaller than the Linear Regression model (Figure 2) since the diagram shows a roughly straight line with a width approximately smaller than 1.

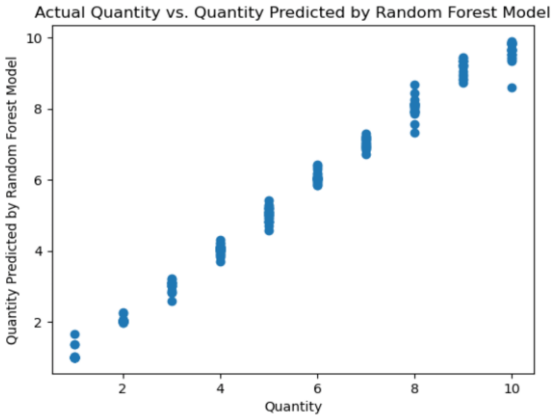

Figure 3. The scatter plot of The Actual Quantity vs. Quantity Predicted by Random Forest Model.

4.3. XGBoost

The R-squared score for the XGBoost model calculated is: 0.9917, and the adjusted R-squared score is: 0.9911. Therefore, approximately 99.17% of the variance in the dependent variable is explained by other features in the XGBoost model.

DOI: 10.54254/2755-2721/45/20241487

The R-squared score for this model is much higher than that of the Linear Regression model but is slightly lower than that of the Random Forest model. It is reasonable to say that the XGBoost model generates a pretty good result, yet not the best in this scenario.

The conclusion above is also supported by the diagram below (Figure 4) illustrating the disparity between the predicted and actual values of the target variable within the XGBoost model. In Figure 4, it can be observed that the distribution follows a somewhat linear pattern, resembling that of the Random Forest model. However, the data points in the XGBoost model display slightly greater dispersion compared to those in the Random Forest model (Figure 3). Thus, it can be inferred that the prediction of the XGBoost model is slightly less accurate than that of the Random Forest model.

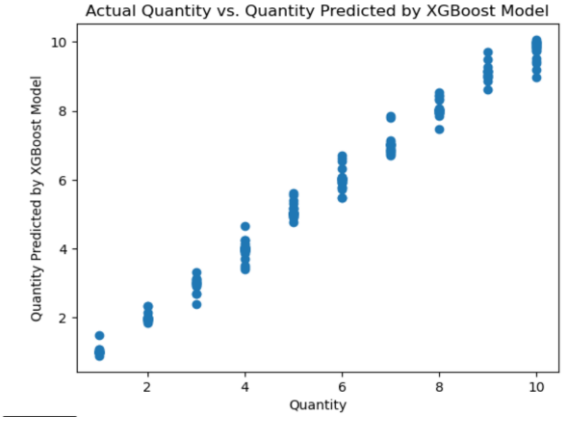

Figure 4. The scatter plot of The Actual Quantity vs. Quantity Predicted by XGBoost Model.

4.4. Comparison and discussion

Creating a table that puts all R-squared values of the three models together and sorts them in descending order (Table 2), it is shown clearly that the Random Forest model generates the best prediction as it has the highest R-squared score as well as the adjusted R-squared score. The computed R-squared score for the Random Forest model is roughly 99.27%, the XGBoost Regression model is 99.17%, and the Linear Regression model is 82.58%. Therefore, it can be seen that the Random Forest model is the best model for predicting the quantity sold for the dataset "spmkt_data_encoded".

Table 2. R-squared Score and the Adjusted R-squared

Model	R-squared Score	Adjusted R-squared Score
Random Forest	99.2746	99.2239
XGBoost	99.1655	99.1072
Linear Regression	82.5796	81.3621

Though the R-squared score seems unreasonably high, with all of the values higher than 80% and the best one close to 100%, after comparing with predictions conducted by other researchers on kaggle.com, these values seem to be commonly shared by various prediction methods and should be good for this research. For instance, one researcher employed the CatBoost Regressor, achieving nearly identical training and testing scores of approximately 100% [9]. Similarly, the ExtraTrees Classifier yielded a training score of 100%. Furthermore, another researcher employed the Gradient Boosting Classifier, attaining a highly respectable training score of 88.75% [10].

However, it's worth noting that different methods led to varying levels of accuracy. Some methods, such as the AdaBoost Classifier, returned a lower training score of 67.0%. Additionally, the Naive Bayes method yielded a precision of approximately 55.10% [10]. Therefore, it becomes apparent that prediction accuracy can fluctuate depending on the chosen approach.

Overall, the results obtained from the three models in this research appear to be reasonably consistent and promising.

DOI: 10.54254/2755-2721/45/20241487

5. Conclusions and further research opportunities

In this research, three distinct machine-learning algorithms were applied to train and evaluate the same dataset, with the goal of identifying the most effective model for predicting the quantity sold in the target dataset. These three methods include Linear Regression, Random Forest, and XGBoost, each of which yielded a unique machine-learning model.

Among the three machine learning models created in this research, the Random Forest model returns the best prediction of the target variable "Quantity". It achieves an impressive R-squared score of approximately 99.27%, signifying that roughly 99.27% of the variance in the dependent variable can be accounted for by the model's features. With an R-squared score closely approaching 100%, the Random Forest model demonstrates an exceptionally high level of predictive precision. The XGBoost Regression model scored 99.17%, and the Linear Regression model attained an R-squared score of 82.58%.

Nevertheless, given that the target dataset comprises only 1000 values in total, the conclusions drawn may lack sufficient generalizability. It's possible that the strong performance of the models is partly attributed to the dataset's limited size. A larger and more comprehensive dataset might yield a broader and potentially less accurate prediction.

References

[1] Mentzer, J. T., & Moon, M. A. (2004). Sales forecasting management: a demand management approach. Sage Publications.

[2] Ng, S. T., Skitmore, M., & Wong, K. F. (2008). Using genetic algorithms and linear regression analysis for private housing demand forecast. Building and Environment, 43(6), 1171-1184.

[3] Zhang, S. L. (2021, January). Machine learning model for sales forecasting by using XGBoost. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 480-483). IEEE.

[4] Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. Journal of neuroscience methods, 220(1), 85-91.

[5] "Supermarket Sales." N.p., n.d. Web. 21 Sept. 2023.

[6] Sharma, G. "5 Regression Algorithms You Should Know - Introductory Guide!" Analytics Vidhya. N.p., 26 May 2021. Web. 25 Sept. 2023.

[7] ER, S. (2021). Understand Random Forest Algorithms With Examples (Updated 2023). Analytics Vidhya, 17.

[8] Quinto, B., & Quinto, B. (2020). Introduction to Machine Learning. Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More, 1-27.

[9] "Supermarket sales_Pred_EDA." N.p., n.d. Web. 24 Sept. 2023. https://kaggle.com/code/himanshukumar7079/supermarket-sales-pred-eda

[10] "Supermarket Sales Prediction." N.p., n.d. Web. 30 Sept. 2023. https://www.kaggle.com/code/aryantiwari123/supermarket-sales-prediction