A Sales Prediction Method Based on XGBoost Algorithm Model

Kunluo Li*

College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China

*Corresponding author: likunluo@email.szu.edu.cn

Abstract. Reasonable and accurate sales forecasting is an important issue for large chain stores. Forecasting short- and long-term product sales helps companies develop marketing strategies and inventory turnover plans. In today's ever-changing business environment, the application of artificial intelligence technology allows for more efficient processing of large amounts of data while taking into account many external factors such as the climate, consumer patterns, and financial situation. An XGBoost linear regression model for the Kaggle competition was trained using the dataset of Ecuadorian Favorita chain stores that was made available. The suggested prediction model seeks to address the seasonality and data scarcity issues. In the context of machine learning, producing several samples for both training and testing aids in our ability to assess the model's efficacy. The most popular technique for detecting overfitting and underfitting issues is to create various samples of data for training and testing models. The experimental findings demonstrate that the XGBoost linear regression model can reasonably provide scientifically based predictions for chain store sales and has a high prediction accuracy.

Keywords: XGBoost; Forecast of Future Sales, One-hot Encoding, R-Squared, Time Lag.

1. Introduction

When operating a large chain of stores, companies need to provide customers with appropriate goods while preserving the proper product inventories at the right time. To prevent out-of-stocks from affecting product sales, companies usually keep a large percentage of safety stock [1-3]. Thus, reducing safety stock means that the cost of inventory and warehousing can be reduced, and accurately forecasting customer demand contributes to the profitability of large chain stores, while allowing upstream manufacturers to adjust production. The effectiveness of safety stock depends on accurate forecasting of product sales [4]. In order to improve the competitiveness of a company, managers should be able to use available information to accurately forecast future sales of various products so that the company can grasp market demand and reasonably arrange the inventory levels of different products in stores [5,6].

There are two main types of current forecasting techniques: classical methods based on statistical models and the use of newly emerging artificial intelligence techniques. In the current dynamic and ever-changing business environment, the application of AI techniques allows for more efficient processing of large amounts of data while taking into account many external factors, such as weather, shopping trends and the economic environment [7,8]. Therefore, in this paper, the linear regression model of XGBoost will be used to analyze the data, along with sales data and external influences for a large chain of stores in Ecuador, in order to obtain more accurate sales forecasts [9].

2. Data and Variables

2.1 Data Description

The datasets in this project are daily merchandise sales data from 54 Favorita stores located in Ecuador, daily prices of oil, and holiday data from Ecuador. The time of All datasets is between January 1, 2011 and August 15, 2017. There are 3,000,888 samples and 17 features in the datasets, of which 11 features are categorical features and 6 are numerical features (table 1).

Table 1. The Datasets' Statistic Information

Datasets	Samples	Raw Features
Train	3000887	6
Holiday_events	168	6
Oil	1218	1
Stores	54	5
Transactions	83488	3

2.2 Explanations for Research Variables

The sales of goods from the Favorita shops are influenced by a variety of variables, including changes in the macroeconomic climate in Ecuador, variations in household income, holidays and working days, and different cities. Given that Ecuador is an oil-dependent nation and that its economy is extremely susceptible to shocks in oil prices, the shift in prices also affects sales.

By analyzing the sales under different dates, it can be found that Saturday is the biggest day of sales per week, as well as some specific kinds of goods have obvious seasonality. And by analyzing oil prices and sales volume, it is proved that oil prices and sales volume are negatively correlated (table 2).

Table 2. Key Features and Corresponding Explanation

Feature Name	Explanation
Store_nbr	The retail location where the goods are sold
Family	The types of goods
Transactions	The total sales for a family of products at a specific store on a specific date
Dcoilwtico	Daily Crude Oil Prices: West Texas Intermediate (WTI)
Holidays_events	Holidays and Events in Ecuador
Holidays_locale	National and Local Holidays
City	The store's city
State	The state in which the store is situated

3. Methodology

3.1 One-hot Encoding

By using one-hot encoding, categorical variables can be transformed into a format that machine learning algorithms can use to make predictions more accurately. This method essentially creates a vector with a length that corresponds to the total amount of categories within data collection. The ith component of this vector, which is given a value of 1, is assigned a value of 1 if a data point falls into the ith category. This makes it possible to keep track of the categories in a way that makes sense in terms of numbers. One-hot transforms each categorical value into a fresh categorical column, and each of these columns is assigned a binary of either 1 or 0. The integer values of each categorical column are represented by a binary vector. All values will be represented by zero and the index will be represented by 1.

3.2 XGBoost Model

XGBoost began as a terminal application that could be set using the libsvm configuration file, originally as a research study established by Tianqi Chen as a member of the Distributed (Deep) Machine Learning Community (DMLC) group. After then, it gained popularity in ML competition circles after being incorporated into the Higgs Machine Learning Challenge's winning answer.

An unregularized XGBoost algorithm is generally:

Training set $\{(x_i, y_i)\}_{i=1}^{N}$, a differentiable loss function $L(y, F(x))$, several weak learners M, and learning rate α are the inputs.

Algorithm:

1.start the model off with a fixed value:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_{i,\theta}). \tag{1}$$

2. For $m = 1$ to M :

(1) Calculate the "hessians" and "gradients":

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{2}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2} \right]_{f(x)=\hat{f}_{(m-1)}(x)} \tag{3}$$

By resolving the following optimization issue using the training set $\left\{ x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} \right\}_{i=1}^{N}$, you can fit a base learner (or weak learner, such as a tree):

$$\hat{\Phi} = \arg\min_{\Phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \Phi(x_i) \right]^2. \tag{4}$$

$$\hat{f}_{(m)}(x) = \alpha \hat{\Phi}_m(x). \tag{5}$$

(3) Modify the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x). \tag{6}$$

3.Output

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x). \tag{7}$$

The veracity of the claim that XGBoost is an effective technique for developing supervised regression models may be determined by comprehending its (XGBoost's) objective function and the underlying learner. The loss function and normalization are both parts of the objective function. The disparity between the model results and the actual values, or the difference between the actual and anticipated values, is revealed. Reg:logistic is the most used loss function for binary classification, whereas reg:linear is the most popular loss function for regression problems in XGBoost. One of the integrated learning techniques is XGBoost, which entails training and merging various individual models (referred to as base learners) to produce individual predictions [10].

4. XGBoost Model Implementation

4.1 Data Preprocessing

There are two ways to deal with missing data on oil prices. The two options are to either remove the missing values or replace them with additional values. It would be more acceptable to replace the missing data with the oil price data from the preceding date of the oil price because oil prices are very volatile. Afterwards, in order to investigate possible serial dependencies in a time series, a "lagged" copy of that series is created. A time series is said to be lagging when its values are advanced by one or even more time steps, or, alternately, when the time in its index is advanced by one or more steps. Selecting the appropriate lag for an oil price characteristic is done by calculating the partial autocorrelation oil price. Therefore, the lag = 3 is chosen by us.

In the holiday dataset, there are some features that are classified data. This means that I need to convert these categorical features into integer codes, so I perform the data encoding step, by converting discrete category information into one-hot encoding form. After converting the multi-category variables of k categories into k-1 variables in binary form, each in binary form represents the size of the relative reference class, and finally each categorical feature is converted into an integer feature in the format 0 or 1.

4.2 Experiment Result

The generation of various samples for training and testing in the context of machine learning aids in our ability to assess model performance. Overfitting is a relatively prevalent issue while training models. This happens when a model works exceptionally well on the massive quantity of data used to train it, but struggles to generalize to new, unknown data points. Underfitting, on the other hand, is when a model does badly on the data used to train it. Underfitting typically happens when the model does not match the issue you are trying to solve. The most popular technique for finding such issues is to generate several samples of data for model training and testing. This allows us to assess how effectively the model generalizes to new, unknown data by utilizing the training data set to train our model and the test set as a collection of data points thereafter.

A training set and a test set can be created by using the skict-learn method train_test_split(), which is probably the most commonly used method. A training set is created with 0.8 samples and a test set with 0.2 samples. A random state (random_state) corresponding to the seed is also defined so that the results can be repeated. The test set is then created by simply removing the corresponding index from the original data frame now contained in the training set.

In regression models, the R-Squared statistic (also known as the coefficient of determination, or R^2) is used to calculate the percentage of variance within the dependent variable that can be accounted for by the independent variable. R-squared measures how closely the data matches the regression model, to put it another way (goodness of fit). The value of r-squared can range from 0 to 1. A low r-squared is typically a negative scenario for predictive models, whereas a greater r-squared shows that the model explains more variability. After importing the collated data into the XGBoost regression model, a model was trained using the data. Its R-Squared was then calculated and a score of 0.92 was obtained, which indicates that our model has explained enough variation in the observed data. As shown in Fig.1, the transactions predicted by the model are close to the validated data (figure 1).

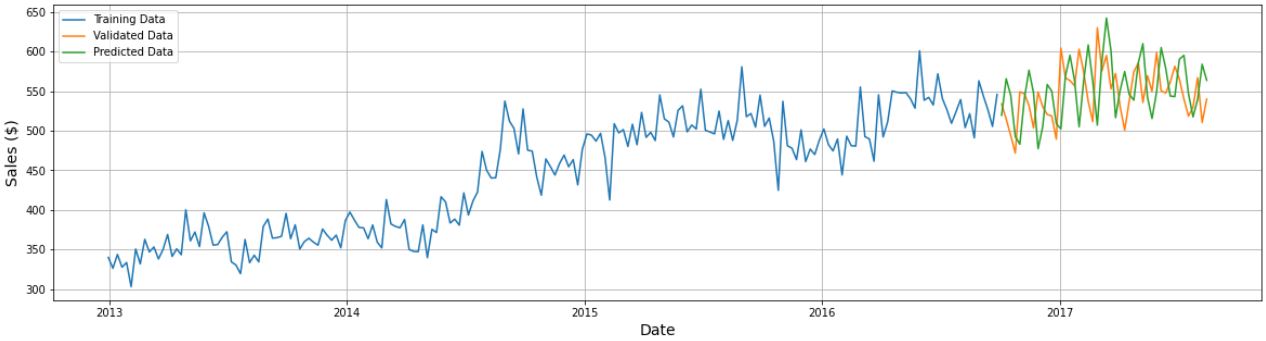

Fig 1. Prediction Using XGBoost Model

5. Conclusion

In this paper, sales are predicted over time based on merchandising data from 2013-2017 for a chain of stores located in Favorita, Ecuador, and their association with local oil prices and holidays is also analyzed. By training an XGBoost model, future sales trends and sales volumes can be predicted more accurately. The use of XGBoost model can maximize the advantages of the prediction model, which helps business managers to make decisions and can help companies to develop long-term marketing strategies, which has important business value for store companies.

References

[1] Xia M, Wong W K. A seasonal discrete grey forecasting model for fashion retailing[J]. Knowledge-Based Systems, 2014, 57: 119-126.

[2] Chang P C, Wang Y W. Fuzzy Delphi and back-propagation model for sales forecasting in PCB industry[J]. Expert systems with applications, 2006, 30(4): 715-726.

[3] S. Lam, M. Vandenbosch, M. Pearce, Retail sales force scheduling based on store traffic forecasting, J. Retailing 74 (1) (1998) 61–88.

[4] S.H. McIntyre, D.D. Achabal, C.M. Miller, Applying case-based reasoning to forecasting retail sales, J. Retailing 69 (4) (1993) 372–398.

[5] H.S. Shih, E.S. Lee, S.H. Chuang, C.C. Chen, A forecasting decision on the sales volume of printers in Taiwan: an exploitation of the analytic network process, Comput. Math. Appl. 64 (6) (2012) 1545–1556.

[6] M.M. Florance, M.S. Sawicz, Positioning sales forecasting for better results, J. Bus. Forecast. 12 (4) (1993) 27–28.

[7] A.V. Iyer, M.E. Bergen, Quick response in manufacturer–retailer channels, Manage. Sci. 43 (4) (1997) 559–570.

[8] K.L. Donohue, Efficient supply contract for fashion goods with forecast updating and two production modes, Manage. Sci. 46 (11) (2000) 1397–1411.

[9] C.W. Chua, G.P. Zhang, A comparative study of linear and nonlinear models for aggregate retail sales forecasting, Int. J. Prod. Econ. 86 (3) (2003) 217–231.

[10] Ji S, Wang X, Zhao W, et al. An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise[J]. Mathematical Problems in Engineering, 2019, 2019.