May-June 2020
ISSN: 0193-4120 Page No. 3555-3559

Predicting Supermarket Sales Data Using Machine Learning Algorithms

[1]Raghavendra Mokashi, [2]Revina Rebecca
[1]Research Scholar, REVA University, Bangalore
[2]Associate Professor, School of Computer Science and Applications,
REVA UNIVERSITY, Bangalore
[1]raghu.2205@gmail.com, [2]revinaprabhu1@gmail.com

Article Info
Volume 83
Page Number: 3555-3559
Publication Issue:
May-June 2020

Article History
***Article Received:** 19 August2019*
***Revised:** 27 November 2019*
***Accepted:** 29 January 2019*
***Publication:** 12 May 2020*

Abstract

The prediction of data in domains such as medical, marketing, sales, weather or stocks is having huge demand now a days. Prediction of sales is plays a vital role in todays world. In order to predict sales, an implementation of several algorithms is done in this work. The data for this work comprises of weekly sales data from different sectors in Supermarket, which exists all over the United States of America. The models implemented here for the prediction are Gradient Boosting, Random Forest and Extremely Randomized Trees Classifiers. To discover the best algorithm and the additional parameter values at which the best outcomes are obtained, a comparative assessment of the three algorithms is used.

***Keywords:** Marketing, Sales, Gradient Boosting, Random Forest, Extremely Randomized Trees*

1. Introduction

Accuracy in business predictions is more important today than ever before, owing to the evolving market demands and increase in complexity in making business decisions. Making important predictions in the fields of economic growth, weather etc. would be extremely useful to the country's growth and development ex: trends in the stock market.

Predicting has spread to a wide assortment of areas, for example, environment, climate, topography, sports and so forth, inferable from the worthwhile advantages and benefit making results of prediction. In the times when businesses are ready to spend money and the customer is covered with an innumerable of options, even a minor good feature would improve the business to an immense deal.

Sales prediction uses tendency recognized from heuristic data to predict future sales. This improves decision making including the current inventory, and efficiently managing the upcoming production Projecting includes trends from heuristic data to calculate future sales. Here we analyse the application for sales projection, based on investigating the results of a range of algorithms such as Random Forest [1].

Random forests is a strategy for classifying, degenerating and remaining tasks that work as building large number of decision trees at phase of examining and creating the value that is the mean of the characteristics (regression). Gradient Boosting[2], is also an collective method for regression, in that it minimizes the loss function by including regression-Trees using the gradient decreasing procedure, and Extra Trees[3], include randomizing tremendously both attribute and cut- point selection while splitting the tree's node.

2. Related Work

A study conducted by Ketchum and Hughes (1997) [12], identifying regions in Maine, perceive the issue of the endogenous area of Wal-Mart stores in more quickly developing areas. They endeavour to appraise the impacts of Wal-Mart on work and profit utilizing a distinction in-contrast aloof natures (DDD) estimator that looks at changes in retail business and income after some time in provinces in which Wal-Mart stores did and did not find, contrasted with changes for assembling and administrations. Nonetheless, essentially none of their assessed changes is factually noteworthy, so one cannot gain much from this information (and the information show up extremely loud). They ambiguously represent endogeneity, In spite of the fact that they do address the issue. Specifically, they report proof recommending that Wal-Mart area choices are autonomous of long haul economic growth paces of person provinces in their example, and that current and slacked development have no noteworthy impact on Wal-Mart's choice to enter.

In a later work, Basker (2005) [13] studies the impacts of Wal-Mart on retail business utilizing across the nation information. Basker endeavours to account expressly for endogeneity by instrumenting for the real number of stores opening in an area in a given year with the arranged number. The last depends on numbers that Wal-Mart appoints to stores when they are arranged; as per Basker, these store numbers show the request wherein the openings were wanted to happen. She at that point joins these numbers with data from Wal-Mart Annual Reports to gauge arranged openings in every area and year. Her outcomes show that district level retail work develops by around 100 in the time of Wal-Mart passage however decreases to an increase of around 50 employments in five years as other retail foundations agreement or close. Meanwhile, potentially in light of the fact that Wal-Mart smoothes out its inventory network, discount business decreases by 20 employments in the more extended term. The vital issue with this distinguishing proof methodology, nevertheless, is that the instrument is unconvincing.

For the instrument to be substantial, two conditions must hold. The first is that arranged store openings ought to be related with (prescient of) genuine openings; this condition is not tricky. The subsequent condition is that the variety in arranged openings produces exogenous variety in real openings that is uncorrelated with the in-undisclosed determinants of work that endogenously influence area choices.

Common ways to find the solution for the problem in machine learning are by using Artificial Neural Networks (ANNs) [4] and numerical methods using Autoregressive Integrated Moving Average (ARIMA) [6]. Usually ANNs displayed good improvement in anticipating because of their capacity to describe non-direct information with great precision [5]. The results can be enhanced by adding additional parameters.

A combination of drifting Trend Breakdown utilizing Loess and Autoregressive Integrated Moving Average (STL+ ARIMA) has been used in timely basis for predicting and these yielded improved results [7].

The paper entails three algorithms namely, Random Forest, Gradient Boosting, and Extra Trees, that are executed on the Supermarket dataset. The algorithms were implemented using Python 3.7. The performance of each algorithm was compared to emphasize the best results.

3. Data set

Here we pick the dataset from supermarket sales data holding data related to an American retail organization, Supermarket. With the data of 45 supermarket department stores, focussing on their sales on a weekly basis. Each row/entry has the following attributes: the associated store (number), the corresponding department (number), the date of the starting day in that week, departmental weekly sales, the store size, and a Boolean value specifying a holiday in that week[9].

Along with the above, is a parallel set of features including Consumer Price Index, unemployment rate, temperature, fuel price, and promotional markdowns. These values are generated from the given training data, for cross-validation, and final testing.

4. Algorithms

In this research work, three projecting models were developed based on the following machine-learning algorithm: Random Forest, Gradient Boosting, and Extremely Randomized Trees (Extra Trees).

For this analysis, the existing algorithms such as Naïve Bayes and Adaptive Boosting were used, but the result was not as expected, so it was not considered for further discussion [11].

Random Forest

The description of Random Forest architecture is Figure 1 [9]. With more trees, the Random Forest calculation enhances more haphazardness to the model. It looks for the best element from an irregular subset of highlights instead of scanning for the most important element while parting a hub. This gives an increasingly exact model, as it prompts an a lot more noteworthy decent variety. Thus, in Random Forest, only the algorithm for diverging a node considers a random subset of the features. Trees can be made more random by using random thresholds for each element of searching for the best thresholds.

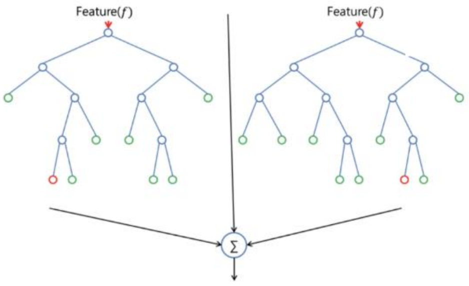

Figure 1: Random Forest Architecture

The features used for training the model were week number, store number, department number, the holiday flag, Consumer Price Index, unemployment rate, temperature, fuel price and store size. The algorithm was carried out using Python’s RandomForestRegressor function present in the scikit-learn class. In the Python implementation, Mean Absolute Error (MAE), mean-squared error

(MSE) and R^2 score are calculated for the predicted values.

Figure 2 shows evaluation of the Predicted qualities and the genuine estimations of the weekly sales with the hyper parameters set at the optimized values.

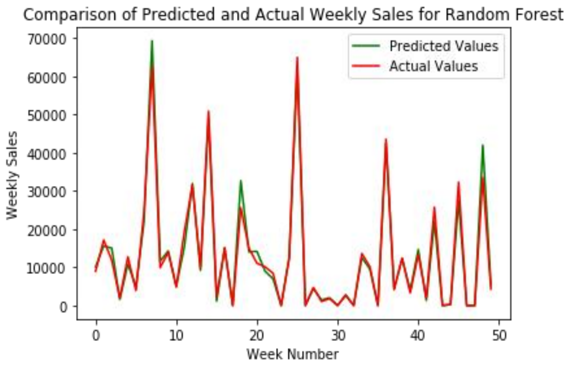

Figure 2: Performance Visual of Random Forest

Gradient Boosting

According to Fried manetal, Gradient Boosting sequentially fits a straightforward Parameterized capacity to current pseudo-residuals by least-squares at every repetition by building additive regression models. The gradient of the loss purpose is minimalized with respect to the classical values at each preparation point, referred as the pseudo-residuals in the model. Figure 4 illustrates the operation of the Gradient Boosting algorithm, at various iterations.

The features used for training the model were the same as, in the Random Forest classification. The algorithm was implemented using Python's Gradient BoostingRegressor function from the scikit-learn class, and the mean absolute error, mean squared error score and R were calculated for the predicted values.

Figure 3 shows correlation of the anticipated qualities and the real values of the week-by-week sales with the hyperactive parameters set at the improved values.

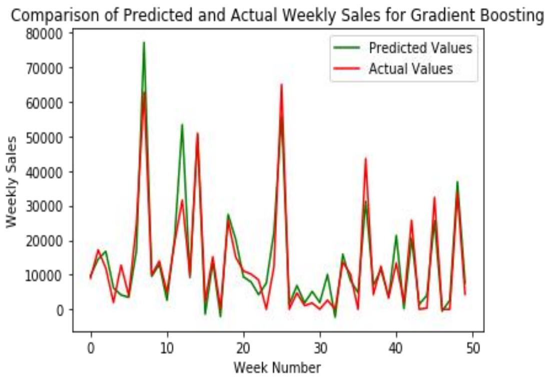

Figure 3: Performance Visual of Gradient Boosting

Extra Trees

The Random Forest and Extra Trees algorithms are almost the same. In the Random Forest algorithm, the tree-splitting phenomenon is more deterministic in nature whereas in Extremely Randomized Trees, the split of the trees is a completely random one. In other words, during the process of splitting, the algorithm chooses the best split among random splits in the selected Variable for the current decision tree.

Python's Extra Trees Regressor function from The scikit-learn class was used to execute The algorithm and the various performance metrics computed for the previous methods are evaluated and reported.

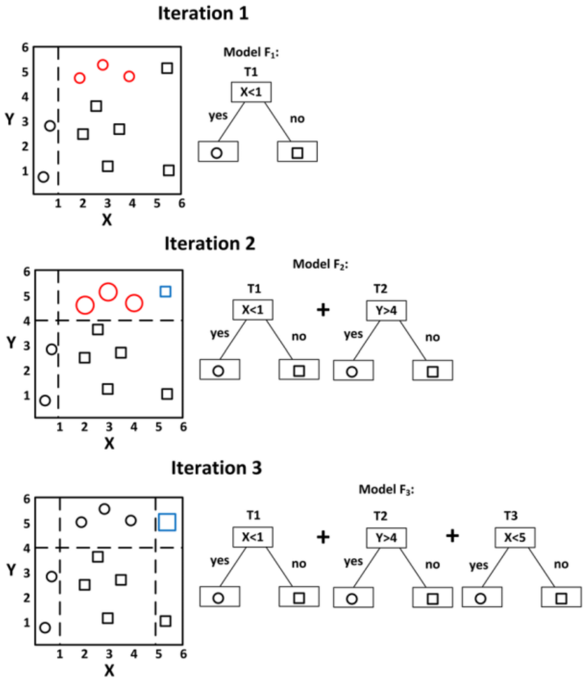

Figure 4: Operation of Gradient Boosting

Figure 5 draws a comparison of the predicted and actual values of the weekly sales with the hyper parameters set at the optimized values.

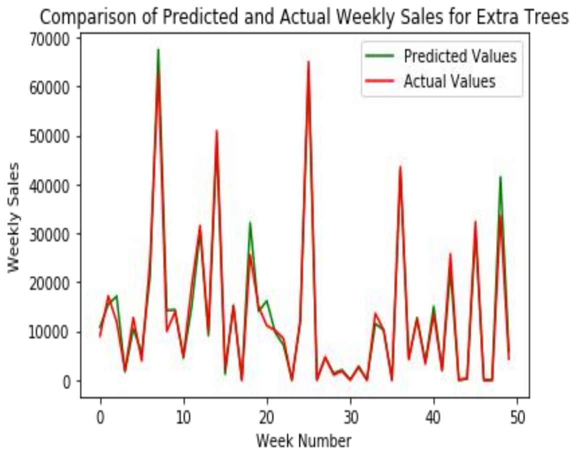

Figure 5: Performance Visual of Extra Trees

5. Output Results

The Mean Absolute Error (MAE) of around 2130 was calculated with the sample dataset. The predicted values of weekly sales are 0’s, the MAE is found to be about 22000. In this study, the last 15% of the considered dataset was used as the local test data set.

The Gradient Boosting algorithm was taken as a base and the MAE was found to be 5771.5, with a R^2 score of 0.80 that indicates that 80% of the predicted values were accurate. These were the best results obtained with the n_estimator additional parameter, which refers to the number of decision trees, which are used for regression.

The remaining additional parameters were set to their default values. Table 1 refers to different values given to the additional parameters and the results that obtained.

Table 1: Gradient Boosting performance

No.of Estimator	Minimum Sample Split	Minimum Sample Leaf	MAE	R^2 Value
101	2	1	6721.5	0.75
101	3	5	6722.2	0.74
150	2	1	6134.1	0.78
150	3	5	6134.5	0.78
200	3	1	5771.5	0.80

The Random Forest algorithm performs much efficient than Gradient Boosting in that it’s MAE was calculated as 1979.4, with a R^2 score of 0.94. The performance metrics were the best achieved with the n_estimators additional parameter set at 150.

Table 2 refers to different values given to the parameters and the results that obtained.

Table 2: Random Forest performance

No.of Estimator	Minimum Sample Split	Minimum Sample Leaf	MAE	R2 Value
51	2	1	1996.8	0.93
51	3	5	2051.5	0.93
100	2	1	1985.2	0.93
150	3	5	2047.3	0.93
150	2	1	1979.4	0.94

The exceptionally Randomized Trees algorithm works comparatively better than the Random Forest. This increase in performance may be attributed to higher randomization in the sample data procedure. The n_estimators parameter was set to 150, while the min_samples_split and min_samples_leaf parameters were placed at 2 and 1 respectively, to obtain the best results wherein the MAE was 1965.5 and R^2 score was 0.94. Table 3 refers to various values given to the parameters and the results that obtained.

Table 3: Extra Trees performance

No.of Estimator	Minimum Sample Split	Minimum Sample Leaf	MAE	R^2 Value
71	2	1	1976.8	0.93
71	3	5	1976.8	0.94
115	2	1	1968.8	0.94
115	3	5	1968.8	0.94
150	2	1	1965.5	0.94

In addition, increasing the number of regression trees randomly is not advised as it leads to overloaded computational requirement resulting in a big amount of time spent in analysing the model without benefitting its correctness.

Table 4 presents the best results obtained from each machine-learning algorithm applied to the dataset.

Table 4: Assessment of ML Algorithms

Algorithm	Mean Absolute Error	Mean Square Error	R^2 Value
Gradient Boosting	5771.5	9.87E+07	0.80
Random Forest	1979.4	3.00E+07	0.94
Extra Tress	1965.5	2.96E+07	0.94

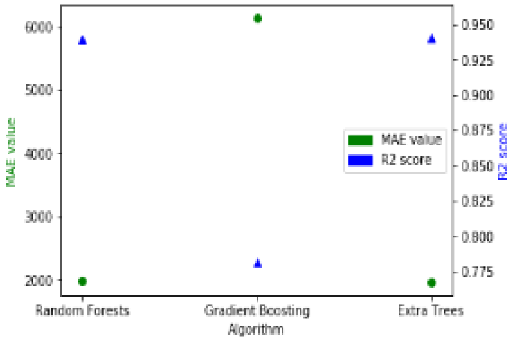

Figure 6: represents the comparative performance of each of the algorithms executed with the hyper parameter, n_estimators, set at 150.

6. Conclusion

The greater part of the shopping markets intend to pull in the customers to the store and make benefit to the most extreme degree by them. When the clients enter the stores they are pulled in then unquestionably they shop more by the exceptional offers and get the ideal things, which are accessible in a good cost and fulfil them. On the off chance that the items are according to the need of the customers, at that point it can make most extreme benefit. The retailers can likewise roll out the improvements in the tasks, goals of the store that cause misfortune and effective strategies can be applied to acquire benefit by watching the historical backdrop of information the current stores an away from of deals can be realized like regularity pattern and irregularity. Deals drop is awful thing gauging deals assists with dissecting it and it can defeat through the business drop to stay in the opposition conjecture assumes a crucial job. Wal-Mart is the main retailer in the USA and it additionally works in numerous different nations all around the globe and is moving into new nations as years cruise by.

There, are different organizations who are continually ascending also and would give Walmart an intense rivalry later on if Walmart does not remain to the highest point of their game. So as to do as such, they should comprehend their business inclines, the client needs and deal with the assets astutely.

In this paper analysis and the implementation

Of three algorithms namely, Random Forest, Gradient Boosting, and Extra Trees, on the Supermarket dataset and a comparing the results of to do the optimization study to determining the optimistic algorithm.

Among the three Random Trees was very efficient model in projecting sales data. Extra Trees, is an extension of Random Forest, also demonstrated very good corrective result. These algorithms may possibly produce even improved results if they are provided with better Graphics Processing Units (GPUs).

7. Future Scope

Future extension would incorporate further investigating Extra Trees and building up the model to consider inadequate markdown information. It likewise incorporates the advancement of the extra parameters of the counterfeit up to improve the effectiveness of forecast. It additionally includes joining the models to make a get together of preparing models that could speak to even the moment subtleties present in the information.

There are very skilful calculations to foresee deals in enormous, medium or little associations, and utilization of such calculation lead to better dynamics.

References

[1] Kulkarni, Vrushali Y., and Pradeep K. Sinha. "Random forest classifiers: a survey for research directions." *Int J Adv Comput* 36.1 (2013): 1144-53.

[2] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." *Machine learning* 63.1 (2006): 3-42.

[3] Friedman, Jerome H. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38.4 (2002): 367-378.

[4] Zhang, Guoqiang, B. Eddy Patuwo, and Michael Y. Hu. "Forecasting with artificial neural networks." *International journal of forecasting* 14.1 (1998): 35-62.

[5] Allende, Héctor, Claudio Moraga, and Rodrigo Salas. "Artificial neural networks in timeseries forecasting: A comparative analysis." *Kybernetika* 38.6 (2002): 685-707.

[6] Adebiyi, Ayodele Ariyo, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. "Comparison of ARIMA and artificial neural networks models for stock price prediction." *Journal of Applied Mathematics* (2014), Article ID 614342, 7 pages, 2014. doi:10.1155/2014/614342.

[7] James J. Pao, Danielle S. Sullivan, “Time Series Sales Forecasting”, Final Year Project, 2017. Accessed at http://cs229.stanford.edu/proj2017/final-reports/5244336.pdf

[8] Sun, Zhan-Li, et al. "Sales forecasting using extreme learning machine with applications in fashion retailing." *Decision Support Systems* 46.1 (2008): 411-419.

[9] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Boca Raton, FL: Chapman & Hall, 1984.

[10] Kaggle. “Supermarket Data”. https://www.kaggle.com/aungpyaeap/supermarket-sales