Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 1, Janurary 2021: 315-322

Research Article

Retail Sales Prediction Using Machine Learning Algorithms

Dr. Bandaru Srinivasa Rao[1]
Dr. Kamepalli Sujatha[2]
Dr. Nannpaneni Chandra Sekhara Rao[3]
Mr. T. Nagendra Kumar[4]

[1, 2, 4] Vignan Foundation for Science, Technology and Research (VFSTR), Deemed to be University, Vadlamudi, Guntur District, Andhra Pradesh, India.
[3] VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India.

Abstract

Data and Data Science, both are playing a pivotal role in creating business intelligence. Machine learning became buzzword technology in data processing and analytics. Retailors in an unorganized sector take decisions intuitively whereas organized retail concerns use business intelligence for their business decisions. Retail sales prediction is common in the organized retail sector, which is highly useful for in time and strategic competitive decisions. In general, sales predictions are made based on past data using statistical and mathematical techniques, etc. The present paper is an attempt to predict retail sales using machine learning algorithms and to present the accuracy of results to the retailors, managers, and policymaker in the retail industry. In this paper three machine learning algorithms k-Nearest Neighbour regression model, Multinomial regression model, Ada Boost regression model were implemented on the training dataset. Multinomial and Decision Tree with Ada Boost regression models work with 100% accuracy. Decision Tree with Ada Boost regression model was implemented on the test set to predict the outlet sales. The regression model predicted the sales of a particular item indicated with the item identifier in a particular outlet indicated with the outlet identifier on the test set. The implications for retailers, managers, and policymakers in the retailing sector are provided based on the results.

Keywords: Retail, Sales, Prediction, Machine Learning Algorithms

1. Introduction

Business intelligence became a buzzword in the 21st century. Intuitional decision making is replaced by business intelligence in the Industrial evolution 4.0. Most of the economic decisions are happening to take into consideration data, data processing, results, and analytics. Information Technology plays a pivotal role in reducing data processing time and assurance of precession, reliability, and accuracy of results, which is highly useful for valuable predictions and business decisions. The retail sector is a combination of unorganized and organized concerns. Most of the unorganized retail concerns take decisions intuitively or based on the experience of proprietors whereas organized retail concerns prefer to make decisions based on data/information process results. Organized Food and grocery sector is playing a dynamic role in penetrating/occupying into

more and more unorganized food and grocery market. Organized food and grocery sector concerns are using contemporary technology for processing data to get better results in less time and applying data science for analytics for transforming the data into business intelligence for using a competitive advantage. In this scenario, academicians, researchers and practitioners contributions for retailers knowledge are: sales prediction in super markets (Thiesing, Middelberg, & Vornberger, 1995); food sales prediction (Meulstee & Pechenizkiy, 2008); simulation based sales forecasting in retail sector (Lv, Bai, Yin, & Dong, 2008), (Schwenke, Ziegenbalg, & Dresden, 2012) forecasting retail sales using neural network in (Y. F. Gao, Liang, Liu, Zhan, & Ou, 2009); using differential evolution (R. Majhi, Panda, Majhi, Panigrahi, & Mishra, 2009), neural network and VBA (Y. Gao, Liang, Tang, Ou, & Zhan, 2010), algorithm (Y. Gao, Liang, Zhan, Ren, & Ou, 2011), Linear Regression (Gopalakrishnan, Choudhary, & Prasad, 2018), Data Mining Techniques (İşlek, 2015),(Singh, Ghutla, Jnr, Mohammed, & Rashid, 2017), (Gaku & Takakuwa, 2015), two-Level Statistical Models (Punam, Pamula, & Jain, 2018), Grid Search Optimization (GSO) (Behera & Nain, 2019), Supervised Algorithms (Ohrimuk & Razmochaeva, 2020), machine learning techniques (Tsoumakas, 2018), (Arif, Sany, Nahin, Shahariar, & Rabby, 2019), (Krishna, Akhilesh, Aich, & Hegde, 2018), (Wang & Liu, 2019), Time Series (Ping, 2018), Xgboost technique (Behera, 2019), Neural Networks & GAS (Baba, Science, City, Prefecture, & Suto, 2000), PSO based adaptive ARMA model (B. Majhi, 2009), S V R mathematical model and methods (Yang & Huiyov, 2007), Structure Time Series (Wei, Geng, Ying, & Shuaipeng, 2014), Machine Learning Techniques (Cheriyan, Ibrahim, & Treesa, 2018), (Wu, Patil, & Gunaseelan, 2018), Fuzzy Inference System (Liang, Li, & Chen, 2019), strategies and technologies for marketers (B. Srinivasa Rao, 2018), recent applications of machine learning (Sujatha Kamepalli, Bandaru Srinivasa Rao, 2018, 2019), praiseworthy. These contributions disclose the role of mathematics, statistics, and econometrics and information technology in providing business intelligence for competitive advantage.

Machine learning applications in data science are increasing day-by-day due to its capability of quick, precise, and accurate processing of data using mathematics, statistical and econometric applications, and providing business intelligence for business strategies development and execution. The present study is an attempt to efficiently predict retail sales using machine learning algorithms. The secondary objective of the study is to explore state-of-art tools, techniques, methods, and models used in sales prediction organized retail outlets. The primary objective is to predict organized retail sales using machine learning algorithms and comparing the results from implemented models. The scope of the study is limited to Big Mart sales prediction using machine learning algorithms. The study is expected to useful for retailers, decision-makers working for retail concerns, and policymaker for development, regulation, and control.

The work is organized by the identification of state-of-art technology for predicting sales in organized retail outs with precession. It follows methodology, data pre-processing, and building prediction model, implementing the prediction model on the training data set, identifying the best model for prediction, and implementing that model on a test set to predict outlet sales. Finally, it concludes the findings and implications.

2.Methodology

In this research paper, we used machine learning regressor models to predict the big mart outlet sales. Here we implemented three models on the training set. The selected best model implemented

on a test set to predict the outlet sales. The following diagram shows the flow of work followed in sales prediction.

Experimentation

The experiments were conducted by developing a simulation environment in python also using WEKA. Three machine learning algorithms k-Nearest Neighbour regression model, Multinomial regression model, Ada Boost regression model were implemented on the training dataset.

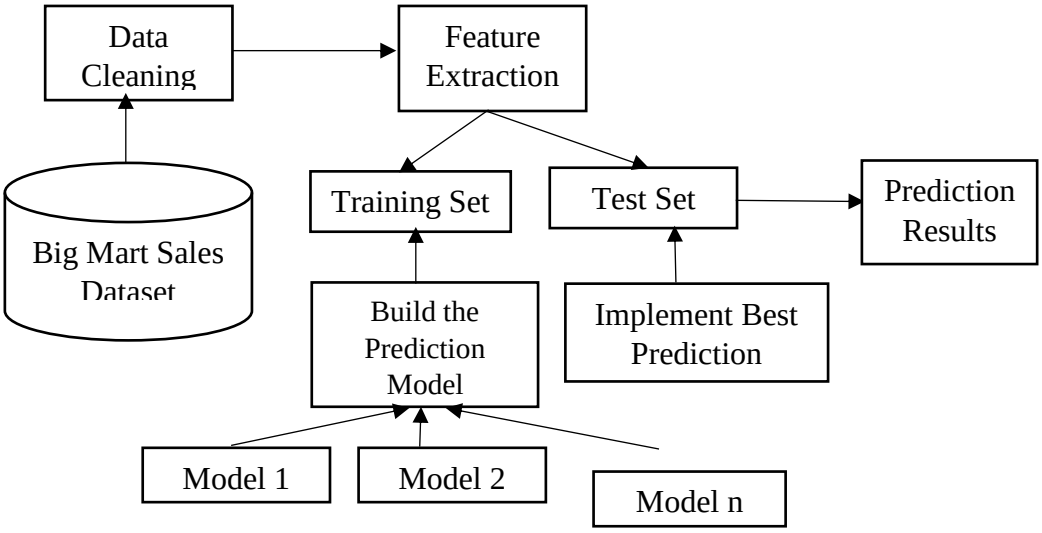

Fig 1. Proposed Flow of Work

Source: Proposed Flow of Work Represented Graphically

k-Nearest Neighbour (k-NN) Regression Model: The k-NN algorithm is a non-parametric strategy used for regression analysis. In this model the input consists of k nearest training examples in the feature space. The output value can be obtained by the average of k nearest neighbour values. In k-NN algorithm it uses three distance measures Euclidean Distance, Manhattan Distance and Minkowski Distance.

The Euclidean Distance can be represented as

$$\text{Distance} = \sqrt{\sum_{j=1}^{k} \left(x_j - y_j\right)^2}$$

The Manhattan Distance can be represented as

$$\text{Distance} = \sum_{j=1}^{k} \left|x_j - y_j\right|$$

The Minkowski Distance can be represented as

$$\text{Distance} = \left(\sum_{j=1}^{k} \left(|x_j - y_j| \right)^q \right)^{1/q}$$

Multinomial Regression Model: Multinomial Logistic Regression generalizes logistic regression to multiclass issues. This model was utilized to foresee the probabilities of the various potential results of a completely disseminated dependent variable, given a lot of independent factors. Multinomial Logistic Regression models how multinomial reaction variable Y relies upon a lot of k informative factors, X= (X1, X2, ... Xk).

Decision Tree Regression-Ada Boost Model: An AdaBoost regressor is a meta-estimator that starts by fitting a regressor on the first dataset and afterward fits extra duplicates of the regressor on the equivalent dataset yet where the loads of occasions are balanced by the error of the current forecast. All things considered, resulting regressors focus more around troublesome cases.

Dataset Description: The dataset was collected from Analytics Vidhya web portal. It is of two parts training set and testing set. The training set contains 8523 samples with 12 attributes and test set contains 5681samples. The following table gives attribute description of the training set.

Table 1. Dataset Description

S. No.	Variable	Description
1	Item_Identifier	Unique product ID
2	Item_Weight	Weight of product
3	Item_Fat_Content	Whether the product is low fat or not
4	Item_Visibility	The % of total display area of all products in a store allocated to the particular product
5	Item_Type	The category to which the product belongs
6	Item_MRP	Maximum Retail Price (list price) of the product
7	Outlet_Identifier	Unique store ID
8	Outlet_Establishment_Year	The year in which store was established
9	Outlet_Size	The size of the store in terms of ground area covered
10	Outlet_Location_Type	The type of city in which the store is located
11	Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
12	Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Source: https://www.analyticsvidhya.com/

The following diagram shows the attribute visualization of training set.

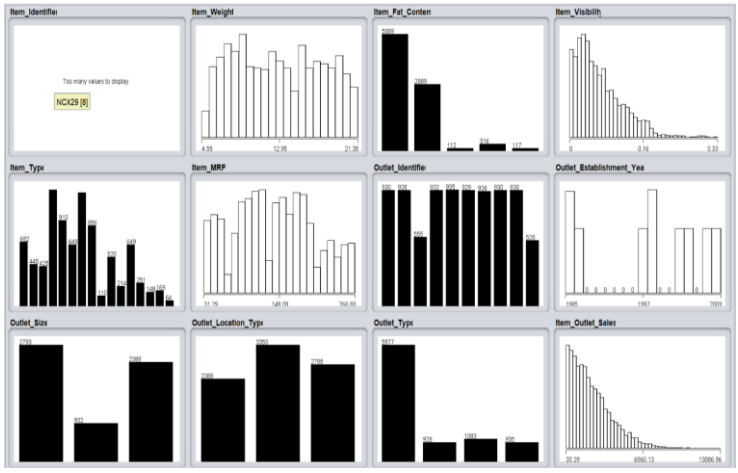

Fig 2. Attribute Visualization of Training set

Source: Experimental Setup (WEKA3.8.3)

3. Results and Discussions

The three models k nearest neighbour Regression Model, Multinomial Regression Model and Decision tree Regression- Ada Boost Model were implemented on training set. Three models work with accuracy levels 0.85, 1.00, 1.00 respectively. The following table shows the accuracy of three models on the training dataset.

Table 2. Comparison of Accuracy of Models

S. NO.	Regression Model	Accuracy
1	k nearest neighbour Regression Model	0.85
2	Multinomial Regression Model	1.0
3	Decision tree Regression- Ada Boost Model	1.0

Source: Experimental Results

The following diagram shows the graphical representation of comparison of accuracy of three models.

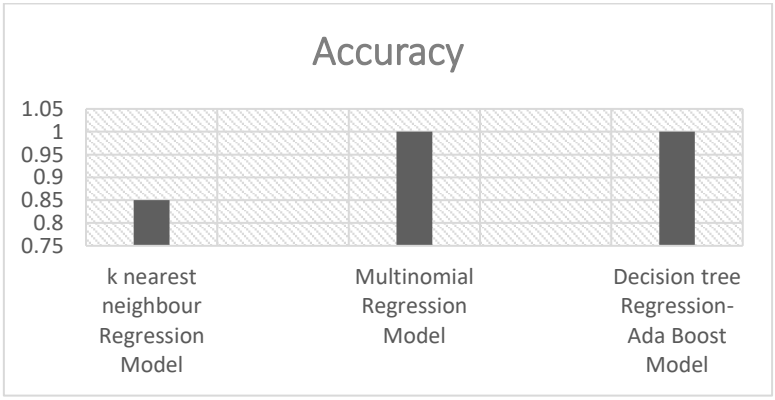

Fig 3. Graphical Representation of Comparison of Accuracy of Models

Source: Experimental Results

Multinomial and Decision Tree with Ada Boost regression models results in with an accuracy of 100%. So, Decision Tree with Ada Boost regression model was implemented on test set to predict the outlet sales. The following table shows the sample output. Here we randomly considered some outlets to visualize the output and we considered two digits after decimal point in the predicted values.

Table 3. Sample Output from Decision Tree with Ada Boost regression model

S. No.	Item Identifier	Outlet Identifier	Item_Outlet_Sales
1	FDW58	OUT049	1827.07
2	FDY38	OUT027	6373.43
3	FDC48	OUT027	2133.95
4	FDQ56	OUT045	1445.86
5	DRL59	OUT013	837.94
6	DRC12	OUT018	2973.97
7	FDG52	OUT046	788.68
8	FDX22	OUT010	499.05
9	FDE21	OUT035	1959.97
10	NCR06	OUT018	523.79
11	FDC26	OUT013	1837.78
12	NCS41	OUT035	3135.96
13	FDU34	OUT046	2088.55
14	FDM03	OUT013	1769.50
15	FDV44	OUT027	5146.45
16	FDT04	OUT013	611.97
17	FDW12	OUT035	2500.40
18	NCY42	OUT027	3495.73
19	FDA14	OUT013	2267.96
20	FDU58	OUT013	3087.11
21	FDB23	OUT046	3767.15
22	FDF47	OUT027	6236.36
23	NCA30	OUT045	3029.83
24	DRM48	OUT035	669.02
25	FDS08	OUT049	2729.96
26	FDC53	OUT019	263.36

Source: Experimental Results

From the above table it is clear that the regression model predicted the sales of a particular item indicated with item identifier in a particular outlet indicated with outlet identifier. Hence this model best predicts the outlet sales prediction. Using these results, Big Mart will try to understand the properties of products and outlets which play a key role in increasing sales.

4.Conclusion

Machine learning applications are increasing day-by-day in business data processing and analytics

area. In this work the results derived from the machine learning based algorithms are more precise, accurate and confidently use for decision making than other models. Compared to three models the last two models are giving hundred per cent accurate results. The results of this study can boost confidence of retailers to implement machine learning in their business data processing and analysis. It also useful for the managers associated with retail sector for developing suitable competitive marketing strategies. Last but not least it helps the policy makers to make different estimations and make suitable policies relating to retail sector.

References

1. Arif, A. I., Sany, S. I., Nahin, F. I., Shahariar, A. K. M., & Rabby, A. (2019). Comparison Study : Product Demand Forecasting with Machine Learning for Shop.
2. B. Srinivasa Rao (2018), Butterfly Customers: Strategies and Technology for Marketers, International Journal of Engineering & Technology, 7 (3.24) (2018) 512-516
3. Baba, N., Science, I., City, K., Prefecture, O., & Suto, H. (2000). for Constructing an Intelligent Sales Prediction. 565–570.
4. Behera, G. (2019). A Comparative Study of Big Mart Sales Prediction A Comparative Study of Big Mart Sales Prediction. (October).
5. Behera, G., & Nain, N. (2019). Grid Search Optimization (GSO) Based Future Sales Prediction For Big Mart. 172–178. https://doi.org/10.1109/SITIS.2019.00038
6. Cheriyan, S., Ibrahim, S., & Treesa, S. (2018). Intelligent Sales Prediction Using Machine Learning Techniques. 53–58.
7. Dr. Sujatha Kamepalli and Dr. Srinivasa Rao Bandaru (2018) Implementation Framework of
8. Artificial Intelligence in Financial Services, International Journal of Research and Analytical Reviews, November 2018, Volume 5, Issue 4.
9. Gaku, R., & Takakuwa, S. (2015). Big data-driven service level analysis for a retail store. (2008), 791–799.
10. Gao, Y. F., Liang, Y. S., Liu, Y., Zhan, S. Bin, & Ou, Z. W. (2009). A neural-network-based forecasting algorithm for retail industry. Proceedings of the 2009 International Conference on Machine Learning and Cybernetics, 2(July), 919–924. https://doi.org/10.1109/ICMLC.2009.5212392
11. Gao, Y., Liang, Y., Tang, F., Ou, Z., & Zhan, S. (2010). A demand forecasting system for retail industry based on neural network and VBA. 2010 Chinese Control and Decision Conference, CCDC 2010, 3786–3789. https://doi.org/10.1109/CCDC.2010.5498506
12. Gao, Y., Liang, Y., Zhan, S., Ren, X., & Ou, Z. (2011). Realization of a demand forecasting algorithm for retail industry. Proceedings of the 2011 Chinese Control and Decision Conference, CCDC 2011, 4227–4230. https://doi.org/10.1109/CCDC.2011.5968968
13. Gopalakrishnan, T., Choudhary, R., & Prasad, S. (2018). Prediction of Sales Value in online shopping using Linear Regression. 5–10.
14. İşlek, İ. (2015). A Retail Demand Forecasting Model Based on Data Mining Techniques. 55–60.
15. Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (2018). Sales-forecasting of Retail Stores using Machine Learning Techniques. 160–166.
16. Liang, Y., Li, J., & Chen, M. (2019). Online Shop Daily Sale Prediction Using Adaptive Network-Based Fuzzy Inference System.
17. Lv, H. R., Bai, X. X., Yin, W. J., & Dong, J. (2008). Simulation based sales forecasting on retail small stores. Proceedings - Winter Simulation Conference, (1968), 1711–1716. https://doi.org/10.1109/WSC.2008.4736257
18. Majhi, B. (2009). Efficient sales forecasting using PSO based adaptive ARMA model. 1333–1337.
19. Majhi, R., Panda, G., Majhi, B., Panigrahi, S. K., & Mishra, M. K. (2009). Forecasting of retail sales data using differential evolution. 2009 World Congress on Nature and Biologically Inspired Computing, NABIC 2009 - Proceedings, 1343–1348. https://doi.org/10.1109/NABIC.2009.5393740
20. Meulstee, P., & Pechenizkiy, M. (2008). Food sales prediction: "If only it knew what we know." Proceedings - IEEE International Conference on Data Mining Workshops, ICDM Workshops 2008, 134–143. https://doi.org/10.1109/ICDMW.2008.128

21. Ohrimuk, E. S., & Razmochaeva, N. V. (2020). Study of Supervised Algorithms for Solve the Forecasting Retail Dynamics Problem. 441–445.
22. Ping, X. (2018). Particle Filter Based Time Series Prediction of Daily Sales of an Online Retailer.
23. Punam, K., Pamula, R., & Jain, P. K. (2018). A Two-Level Statistical Model for Big Mart Sales Prediction. 2018–2021.
24. Schwenke, C., Ziegenbalg, J., & Dresden, D.-. (2012). Simulation based Forecast of Supermarket Sales Chair for Technical Information Systems.
25. Singh, M., Ghutla, B., Jnr, R. L., Mohammed, A. F. S., & Rashid, M. A. (2017). Walmart ' s Sales Data Analysis- A Big Data Analytics Perspective. 114–119. https://doi.org/10.1109/APWConCSE.2017.00028
26. Sujatha Kamepalli, Bandaru Srinivasa Rao (2019), International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8, Issue-6C2, April 2019
27. Thiesing, F. M., Middelberg, U., & Vornberger, O. (1995). Short term prediction of sales in supermarkets. IEEE International Conference on Neural Networks - Conference Proceedings, 2, 1028–1031. https://doi.org/10.1109/icnn.1995.487562
28. Tsoumakas, G. (2018). A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review. https://doi.org/10.1007/s10462-018-9637-z
29. Wang, J., & Liu, L. (2019). A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry. 317–320.
30. Wei, D., Geng, P., Ying, L., & Shuaipeng, L. (2014). A Prediction Study on E-commerce Sales Based on Structure Time Series Model and Web Search Data. i, 5346–5351.
31. Wu, C. M., Patil, P., & Gunaseelan, S. (2018). Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data.
32. Yang, Y., & Huiyov, C. (2007). S V R mathematical model and methods for sale prediction. 18(4), 769–773.