Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432109

Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms

Ranjitha P[1]
Department of Computer Science,
Amrita School of Arts and Sciences, Mysuru
Amrita Vishwa Vidyapeetham, India
Email: ranjithapramod123@gmail.com
Spandana M[2]
Department of Computer Science,
Amrita School of Arts and Sciences, Mysuru
Amrita Vishwa Vidyapeetham, India
Email:Spandanasatishm@gmail.

***Abstract*— Currently, supermarket run-centres, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.**

Keywords—Linear Regression, Polynomial Regression, Ridge Regression, Xgboost Regression

I. Introduction

Everyday competitiveness between various shopping centres as and as huge marts is becoming higher intense, violent just because of the quick development of global malls also online shopping. Each market seeks to offer personalized and limited-time deals to attract many clients relying on period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. The current machine learning algorithm is very advanced and provides methods for predicting or forecasting sales any kind of organization, extremely beneficial to overcome low – priced used for prediction. Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful

II. RELEATED WORK

A great deal of work having been gotten really intended to date the territory of deals foreseeing. A concise audit of the important work in the field of big_mart deals is depicted in this part. Numerous other Measurable methodologies, for example, with regression, (ARIMA) Auto-Regressive Integrated Moving Average, (ARMA) Auto-Regressive Moving Average, have been utilized to develop a few deals forecast standards. Be that as it may, deals anticipating is a refined issue and is influenced by both outer and inside factors, and there are two significant detriments to the measurable technique as set out in A. S. Weigend et A mixture occasional quantum relapse approach and (ARIMA) Auto-Regressive Integrated Moving Average way to deal with every day food deals anticipating were recommend by N. S. Arunraj and furthermore found that the exhibition of the individual model was moderately lower than that of the crossover model.

E. Hadavandi utilized the incorporation of "Genetic Fuzzy Systems (GFS)" and information gathering to conjecture the deals of the printed circuit board. In their paper, K-means bunching delivered K groups of all information records. At that point, all bunches were taken care of into autonomous with a data set tuning and rule-based extraction ability. Perceived work in the field of deals gauging was done by P.A. Castillo, Sales estimating of new distributed books was done in a publication market the executives setting utilizing computational techniques. "Artificial neural organizations" are additionally utilized nearby income estimating. Fluffy Neural Networks have been created with the objective of improving prescient effectiveness, and the Radial "Base Function Neural Network (RBFN)" is required to have an incredible potential for anticipating deals.

Dataset: collected the dataset form the internet for the website called kaggle.com .In this work all having test dataset and train dataset in the test data set having a 5000 dataset and in the train data having a 8000 data

IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

set. Fig1shows the train data and Fig2 shows the sample of test dataset.

TABLE 1: Attributes Information

Attribute	Description
Item_Identifer	It is the unique product Id number.
Item Weight	It will include the product's weight.
Item_Fat_Content	It will mean whether the item is low in fat or not.
Item -Visibility	The percentage of the overall viewing area assigned to the particular item from all items in the shop.
Item -Type	To which group does the commodity belong
Item-MRP	The product's price list
Outlet-Identifier	a distinct slot number
Outlet-Establishment Year	The year that the shop first opened its doors.
Outlet-Size	The sum of total area occupied by a supermarket.
Outlet-Location	The kind of town where the store is situated.
Outlet-Type	The shop is merely a supermarket or a grocery store.
Item-Outlet-Sales	The item's sales in the original shop

Train data set

	A	B	C	D	E	F	G	H	I	J	K	L
1	Item_Iden	Item_Wei	Item_Fat_	Item_Visil	Item_Type	Item_MRF	Outlet_Id	Outlet_Es	Outlet_Siz	Outlet_Lo	Outlet_Ty	Item_Outlet_Sales
2	FDA15	9.3	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermar	3735.138
3	DRC01	5.92	Regular	0.019278	Soft Drink	48.2692	OUT018	2009	Medium	Tier 3	Supermar	443.4228
4	FDN15	17.5	Low Fat	0.01676	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermar	2097.27
5	FDX07	19.2	Regular	0	Fruits and	182.095	OUT010	1998		Tier 3	Grocery St	732.38
6	NCD19	8.93	Low Fat	0	Householc	53.8614	OUT013	1987	High	Tier 3	Supermar	994.7052
7	FDP36	10.395	Regular	0	Baking Go	51.4008	OUT018	2009	Medium	Tier 3	Supermar	556.6088
8	FDO10	13.65	Regular	0.012741	Snack Foo	57.6588	OUT013	1987	High	Tier 3	Supermar	343.5528
9	FDP10		Low Fat	0.12747	Snack Foo	107.7622	OUT027	1985	Medium	Tier 3	Supermar	4022.764
10	FDH17	16.2	Regular	0.016687	Frozen Fo	96.9726	OUT045	2002		Tier 2	Supermar	1076.599
11	FDU28	19.2	Regular	0.09445	Frozen Fo	187.8214	OUT017	2007		Tier 2	Supermar	4710.535
12	FDY07	11.8	Low Fat	0	Fruits and	45.5402	OUT049	1999	Medium	Tier 1	Supermar	1516.027
13	FDA03	18.5	Regular	0.045464	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermar	2187.153
14	FDX32	15.1	Regular	0.100014	Fruits and	145.4786	OUT049	1999	Medium	Tier 1	Supermar	1589.265
15	FDS46	17.6	Regular	0.047257	Snack Foo	119.6782	OUT046	1997	Small	Tier 1	Supermar	2145.208
16	FDF32	16.35	Low Fat	0.068024	Fruits and	196.4426	OUT013	1987	High	Tier 3	Supermar	1977.426
17	FDP49	9	Regular	0.069089	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermar	1547.319
18	NCB42	11.8	Low Fat	0.008596	Health an	115.3492	OUT018	2009	Medium	Tier 3	Supermar	1621.889
19	FDP49	9	Regular	0.069196	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermar	718.3982
20	DRI11		Low Fat	0.034238	Hard Drinl	113.2834	OUT027	1985	Medium	Tier 3	Supermar	2303.668
21	FDU02	13.35	Low Fat	0.102492	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermar	2748.422
22	FDN22	18.85	Regular	0.13819	Snack Foo	250.8724	OUT013	1987	High	Tier 3	Supermar	3775.086
23	FDW12		Regular	0.0354	Baking Go	144.5444	OUT027	1985	Medium	Tier 3	Supermar	4064.043

Fig1: Shows the sample of train data

Test dataset

	A	B	C	D	E	F	G	H	I	J	K
1	Item_Iden	Item_Wei	Item_Fat_	Item_Visil	Item_Type	Item_MRF	Outlet_Id	Outlet_Es	Outlet_Siz	Outlet_Lo	Outlet_Type
2	FDW58	20.75	Low Fat	0.007565	Snack Foo	107.8622	OUT049	1999	Medium	Tier 1	Supermarket Type1
3	FDW14	8.3	reg	0.038428	Dairy	87.3198	OUT017	2007		Tier 2	Supermarket Type1
4	NCN55	14.6	Low Fat	0.099575	Others	241.7538	OUT010	1998		Tier 3	Grocery Store
5	FDQ58	7.315	Low Fat	0.015388	Snack Foo	155.034	OUT017	2007		Tier 2	Supermarket Type1
6	FDY38		Regular	0.118599	Dairy	234.23	OUT027	1985	Medium	Tier 3	Supermarket Type3
7	FDH56	9.8	Regular	0.063817	Fruits and	117.1492	OUT046	1997	Small	Tier 1	Supermarket Type1
8	FDL48	19.35	Regular	0.082602	Baking Go	50.1034	OUT018	2009	Medium	Tier 3	Supermarket Type2
9	FDC48		Low Fat	0.015782	Baking Go	81.0592	OUT027	1985	Medium	Tier 3	Supermarket Type3
10	FDN33	6.305	Regular	0.123365	Snack Foo	95.7436	OUT045	2002		Tier 2	Supermarket Type1
11	FDA36	5.985	Low Fat	0.005698	Baking Go	186.8924	OUT017	2007		Tier 2	Supermarket Type1
12	FDT44	16.6	Low Fat	0.103569	Fruits and	118.3466	OUT017	2007		Tier 2	Supermarket Type1
13	FDQ56	6.59	Low Fat	0.105811	Fruits and	85.3908	OUT045	2002		Tier 2	Supermarket Type1
14	NCC54		Low Fat	0.171079	Health an	240.4196	OUT019	1985	Small	Tier 1	Grocery Store
15	FDU11	4.785	Low Fat	0.092738	Breads	122.3098	OUT049	1999	Medium	Tier 1	Supermarket Type1
16	DRL59	16.75	LF	0.021206	Hard Drinl	52.0298	OUT013	1987	High	Tier 3	Supermarket Type1
17	FDM24	6.135	Regular	0.079451	Baking Go	151.6366	OUT049	1999	Medium	Tier 1	Supermarket Type1
18	FDI57	19.85	Low Fat	0.054135	Seafood	198.7768	OUT045	2002		Tier 2	Supermarket Type1
19	DRC12	17.85	Low Fat	0.037981	Soft Drink	192.2188	OUT018	2009	Medium	Tier 3	Supermarket Type2
20	NCM42		Low Fat	0.028184	Householc	109.6912	OUT027	1985	Medium	Tier 3	Supermarket Type3
21	FDA46	13.6	Low Fat	0.196898	Snack Foo	193.7136	OUT010	1998		Tier 3	Grocery Store
22	FDA31	7.1	Low Fat	0.10992	Fruits and	175.008	OUT013	1987	High	Tier 3	Supermarket Type1
23	NCJ31	19.2	Low Fat	0.182619	Others	239.9196	OUT035	2004	Small	Tier 2	Supermarket Type1

Fig2: Shows the sample of test data

III. METHODOLOGY

Fig3 shows the architecture Diagram of the proposed model where they focus on the different algorithm application to the dataset. Where we are calculating the Accuracy, MAE, MSE, RMSE and final concluding the best yield algorithm. Here are the following Algorithm are used.

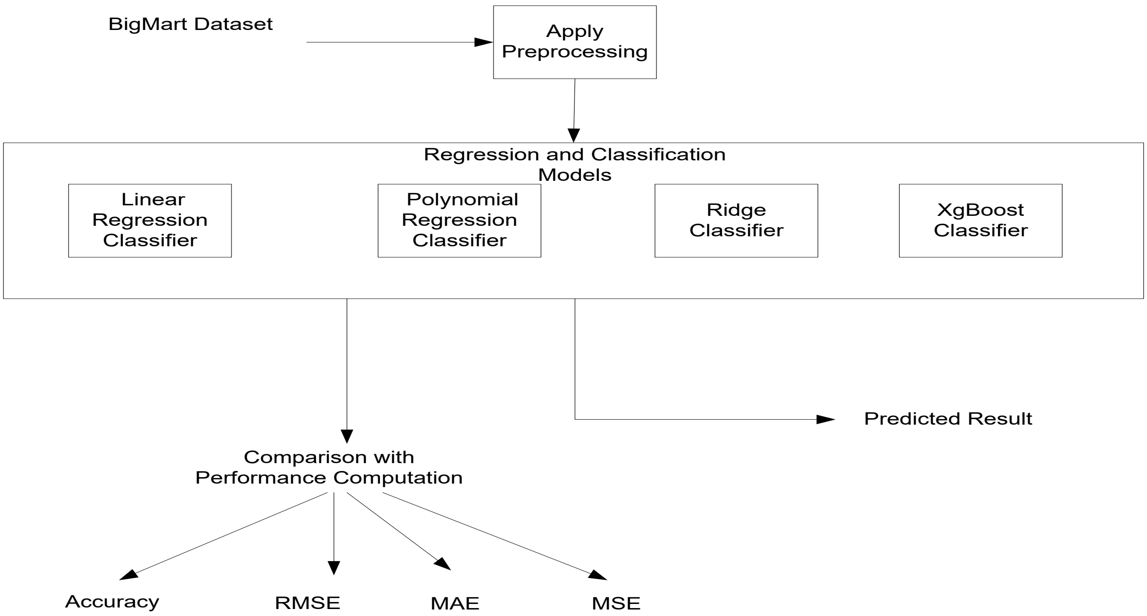

Fig3: Shows the proposed Architecture Diagram

A. Linear Regression

- Build a fragmented plot.1) a linear or non-linear pattern of data and 2) a variance (outliers). Consider a transformation if the marking isn't linear. If this is the case, outsiders, it can suggest only eliminating them if there is a non-statistical justification.
- Link the data to the least squares line and confirm the model assumptions using the residual plot (for the constant standard deviation assumption) and the normal probability plot (for the normal probability assumption) A transformation might be necessary if the assumptions made do not appear to be met.
- If required, convert the data to the-least square using the transformed data, construct a regression line.
- If a change has been completed, return to the previous process 1. If not, continue to phase 5.
- When a "good-fit" classic is defined, write the least-square regression line equation. Consist of normal estimation, estimation, and R-squared errors.

Linear regression formulas look like this:

$$Y = o_1 x_1 + o_2 x_{2+} \dots\dots\dots o_n x_n$$

R-Square: Defines the difference in X (depending variable) explains the total variance in Y (dependent variable) (independent variable). This can be expressed mathematically as

$$R - Square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})^2}{\sum(Y_{actual} - Y_{mean})^2}$$

B. Polynomial Regression Algorithm

- Polynomial Regression is a relapse calculation that modules the relationship here among dependent(y) and the autonomous variable(x) in light of the fact that as most extreme limit polynomial. The condition for polynomial relapse is given beneath: $y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+...... b_nx_1^n$
- It is regularly alluded to as the exceptional instance of various straight relapse in ML. Since we apply some polynomial terms to the numerous straight relapse condition to change it to polynomial relapse adjustment to improve accuracy.
- The informational collection utilized for preparing in polynomial relapse is of a non-straight nature.
- It uses a linear regression model to fit complex and non-linear functions and datasets.

C. Ridge Regression

Ridge regression is a model tuning tool used to evaluate any data that suffers from multicollinearity. This method performs the L2 regularization procedure. When multicollinearity issues arise, the least squares are unbiased and the variances are high, resulting in the expected values being far removed from the actual values.

The cost function for ridge regression:

$$\text{Min}(||Y – X(\text{theta})||^2 + \lambda||\text{theta}||^2)$$

D. XGBoost Regression

"Extreme Gradient Boosting" is same but much more effective to the gradient boosting system. It has both a linear model solver and a tree algorithm. Which permits "xgboost" in any event multiple times quicker than current slope boosting executions. It underpins various target capacities, including relapse, order and rating. As "xgboost" is extremely high in prescient force however generally delayed with organization, it is appropriate for some rivalries. It likewise has extra usefulness for cross-approval and finding significant factors.

IV. RESULT AND DISCUSSION

Liner Regression

TABLE 2: Shows the linear regression result on the various parameter

Parameter	value
MSE	7.4631
MAE	1.166
RMSE	2.731

Polynomial regression

TABLE 3: Shows the polynomial regression result on the various parameter

Parameter	value
MSE	6.120
MAE	2.968
RMSE	7.823

Ridge regression

TABLE 4: Shows the Ridge regression result on the various parameter

Parameter	value
MSE	3.671
MAE	8.289
RMSE	1.916

XgBoost Regression

TABLE 5: Shows the Xgboost regression result on the various parameter

Parameter	value
MSE	0.001
MAE	0.029
RMSE	0.032

Frequency of item_fat_content

TABLE 6: Shows the Xgboost regression frequency of item fat content

Parameter	value
Low Fat	5089
Regular	2889
LF	316
reg	117

TABLE 7: Comparison of MAE, MSE, RMSE with the Model

Model	MSE	MAE	RMSE
Linear Regression	7.4631	1.166	2.731
Polynomial Regression	2.0364	7.002	1.427
Ridge Regression	3.6712	8.289	1.916
Xgboost Regression	0.001	0.029	0.0321

V. CONCLUSION

In this work, the effectiveness of various algorithms on the data on revenue and review of, best performance-algorithm, here propose a software to using regression approach for predicting the sales centered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, polynomial regression, Ridge regression, and Xgboost regression can be determined. So, we can conclude ridge and Xgboost regression gives the better prediction with respect to Accuracy, MAE and RMSE than the Linear and polynomial regression approaches. In future, the forecasting sales and building a sales plan can help to avoid unforeseen cash flow and manage production, staff and financing needs more effectively.In future work we can also consider with the ARIMA model which shows the time series graph.

REFERANCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, “A comparative study of linear and nonlinear models for aggregate retails sales forecasting”, Int. Journal Production Economics, vol. 86, pp. 217-231, 2003.

[2] Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2. Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of D

[3] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110

[4] Giuseppe Nunnari, Valeria Nunnari, “Forecasting Monthly Sales Retail Time Series: A Case Study”, Proc. of IEEE Conf. on Business Informatics (CBI), July 2017.

[5]https://halobi.com/blog/sales-forecasting-five-uses/. [Accessed: Oct. 3, 2018]

[6] Zone-Ching Lin, Wen-Jang Wu, “Multiple LinearRegression Analysis of the Overlay Accuracy Model Zone”, IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 – 237, May 1999.

[7] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, “Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis”, Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.

[8] C. Saunders, A. Gammerman and V. Vovk, “Ridge Regression Learning Algorithm in Dual Variables”, Proc. of Int. Conf. on Machine Learning, pp. 515 – 521, July 1998.IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO. 7, JULY 2010 3561.

[9] ”Robust Regression and Lasso”. Huan Xu, Constantine Caramanis, Member, IEEE, and Shie Mannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration.”An improved Adaboost algorithm based on uncertain functions”.Shu Xinqing School of Automation Wuhan University of Technology.Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China.

[10] Xinqing Shu, Pan Wang, “An Improved Adaboost Algorithm based on Uncertain Functions”, Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.

[11] A. S. Weigend and N. A. Gershenfeld, “Time series prediction: Forecasting the future and understanding the past”, Addison-Wesley, 1994.

[12] N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, Int. J. Production Economics 170 (2015) 321-335P

[13] D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.

[14] X. Yua, Z. Qi, Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, Procedia Computer Science 17 (2013) 1055–1062.

[15] E. Hadavandi, H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: a Case study of the printed circuit board, Expert Systems with Applications 38 (2011) 9392–9399.

[16] P. A. Castillo, A. Mora, H. Faris, J.J. Merelo, P. GarciaSanchez, A.J. Fernandez-Ares, P. De las Cuevas, M.I. Garcia-Arenas, Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment, Knowledge-Based Systems 115 (2017) 133-151.

[17] R. Majhi, G. Panda and G. Sahoo, “Development and performance evaluation of FLANN based model for forecasting of

stock markets".Expert Systems with Applications, vol. 36, issue 3, part 2, pp. 6800-6808, April 2009.

[18] Pei Chann Chang and Yen-Wen Wang, "Fuzzy Delphi and back propagation model for sales forecasting in PCB industry", Expert systems with applications, vol. 30,pp. 715-726, 2006.

[19] R. J. Kuo, Tung Lai HU and Zhen Yao Chen "application of radial basis function neural networks for sales forecasting", Proc. of Int. Asian Conference on Informatics in control, automation, and robotics, pp. 325- 328, 2009.

[20] R. Majhi, G. Panda, G. Sahoo, and A. Panda, "On the development of Improved Adaptive Models for Efficient Prediction of Stock Indices using Clonal-PSO (CPSO) and PSO Techniques", International Journal of Business Forecasting and Market Intelligence, vol. 1, no. 1, pp.50-67, 2008.

[21]Suresh K and Praveen O, "Extracting of Patterns Using Mining Methods Over Damped Window," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 235-241, DOI: 10.1109/ICIRCA48905.2020.9182893.

[22] Shobha Rani, N., Kavyashree, S., & Harshitha, R. (2020). Object Detection in Natural Scene Images Using Thresholding Techniques. Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020, Iciccs, 509–515.

[23] https://www.kaggle.com/brijbhushannanda1979/bigmartsales-data. [Accessed: Jun. 28, 2018].