Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

PREDICTION SYSTEM FOR BIGMART SALES USING MACHINE LEARNING

Mrs. S. Suganya[1], Assistant Professor,
Dept. Of Computer Science and Engineering, SSM Institute of Engineering and Technology.
S. Santhoshkumar[2], G. Saravanakumar[3], J. Shiffin Paul[4], J. Vishwa Bharathi[5], Students,
Dept. Of Computer Science and Engineering, SSM Institute of Engineering and Technology.
Email: suganselva01@gmail.com[1], santhoshselvaraj666@gmail.com[2],
saravanaganesan2001@gmail.com[3], shiffinpaulj@gmail.com[4], vishwabharathi46@gmail.com[5].

Abstract
The aim is to build a predictive model that analyse the sales of each product at a particular outlet and predict their future sales for helping them to increase their profits and make their brand even better and competitive as per the market trends by generating customer satisfaction as well. The resulting data can then be used to prediction potential sales volumes for retailers such as Big Mart through machine learning. The estimate of the system proposed should take account of price tag, outlet and outlet location. The technique used for prediction of sales are Linear Regression Algorithm and Random Forest algorithm, which is a supervised algorithm in the field of Machine Learning that offers an efficient prevision of Big Mart sales based on gradient.

Keywords: Machine Learning, Linear Regression, Random Forest Algorithm.

I. Introduction

Due to the rapid development of malls and online shopping, competition between different shopping centres and large marts is growing more heated and violent on a daily basis. 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. Related 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 individual model's exhibition was slightly less than the crossover model's.

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, In a publication market the executives environment, sales estimation of newly

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

distributed books was carried out using computational tools. Additionally, "artificial neural organisations" are used for income estimation. 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 in Test Dataset and Training Dataset a 5500 dataset and in the train data having a 8500 data.

Dataset:
The dataset consists of 8523individual data. There are 12 columns in the dataset, which are described below.
1.**ItemIdentifier** - Unique product ID
2.**ItemWeight** - Weight of product
3.**ItemFatContent** - Whether the product is low fat or not
4.**ItemVisibility** - The % of the total display area of all products in a store allocated to the particular product
5.**ItemType** - The category to which the product belongs
6.**ItemMRP** - Maximum Retail Price (list price) of the Product
7.**OutletIdentifier** - Unique store ID
8.**OutletEstablishmentYear** - The year in which the store was established
9.**OutletSize** - The size of the store in terms of ground area covered
10.**OutletLocationType** - 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.**ItemOutletSales** - Sales of the product in the particular store. This is the outcome variable to be predicted.

Train data set:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
2	FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
3	DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
4	FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
5	FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
6	NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
7	FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
8	FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
9	FDP10		Low Fat	0.127468857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636
10	FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002		Tier 2	Supermarket Type1	1076.5986
11	FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007		Tier 2	Supermarket Type1	4710.535
12	FDY07	11.8	Low Fat	0	Fruits and Vegetables	45.5402	OUT049	1999	Medium	Tier 1	Supermarket Type1	1516.0266
13	FDA03	18.5	Regular	0.045463773	Dairy	144.1102	OUT046	1997	Small	Tier 1	Supermarket Type1	2187.153
14	FDX32	15.1	Regular	0.1000135	Fruits and Vegetables	145.4786	OUT049	1999	Medium	Tier 1	Supermarket Type1	1589.2646
15	FDS46	17.6	Regular	0.047257328	Snack Foods	119.6782	OUT046	1997	Small	Tier 1	Supermarket Type1	2145.2076
16	FDF32	16.35	Low Fat	0.0680243	Fruits and Vegetables	196.4426	OUT013	1987	High	Tier 3	Supermarket Type1	1977.426
17	FDP49	9	Regular	0.069088961	Breakfast	56.3614	OUT046	1997	Small	Tier 1	Supermarket Type1	1547.3192
18	NCB42	11.8	Low Fat	0.008596051	Health and Hygiene	115.3492	OUT018	2009	Medium	Tier 3	Supermarket Type2	1621.8888
19	FDP49	9	Regular	0.069196376	Breakfast	54.3614	OUT049	1999	Medium	Tier 1	Supermarket Type1	718.3982
20	DRI11		Low Fat	0.034237682	Hard Drinks	113.2834	OUT027	1985	Medium	Tier 3	Supermarket Type3	2303.668
21	FDU02	13.35	Low Fat	0.10249212	Dairy	230.5352	OUT035	2004	Small	Tier 2	Supermarket Type1	2748.4224
22	FDN22	18.85	Regular	0.138190277	Snack Foods	250.8724	OUT013	1987	High	Tier 3	Supermarket Type1	3775.086
23	FDW12		Regular	0.035399923	Baking Goods	144.5444	OUT027	1985	Medium	Tier 3	Supermarket Type3	4064.0432

Fig1: Shows the sample of train data

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, May : 2023

	A	B	C	D	E	F	G	H	I	J	K
1	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type
2	FDW58	20.75	Low Fat	0.007564836	Snack Foods	107.8622	OUT049	1999	Medium	Tier 1	Supermarket Type1
3	FDW14	8.3	reg	0.038427677	Dairy	87.3198	OUT017	2007		Tier 2	Supermarket Type1
4	NCN55	14.6	Low Fat	0.099574908	Others	241.7538	OUT010	1998		Tier 3	Grocery Store
5	FDQ58	7.315	Low Fat	0.015388393	Snack Foods	155.034	OUT017	2007		Tier 2	Supermarket Type1
6	FDY38		Regular	0.118599314	Dairy	234.23	OUT027	1985	Medium	Tier 3	Supermarket Type3
7	FDH56	9.8	Regular	0.063817206	Fruits and Vegetables	117.1492	OUT046	1997	Small	Tier 1	Supermarket Type1
8	FDL48	19.35	Regular	0.082601537	Baking Goods	50.1034	OUT018	2009	Medium	Tier 3	Supermarket Type2
9	FDC48		Low Fat	0.015782495	Baking Goods	81.0592	OUT027	1985	Medium	Tier 3	Supermarket Type3
10	FDN33	6.305	Regular	0.123365446	Snack Foods	95.7436	OUT045	2002		Tier 2	Supermarket Type1
11	FDA36	5.985	Low Fat	0.005698435	Baking Goods	186.8924	OUT017	2007		Tier 2	Supermarket Type1
12	FDT44	16.6	Low Fat	0.103569075	Fruits and Vegetables	118.3466	OUT017	2007		Tier 2	Supermarket Type1
13	FDQ56	6.59	Low Fat	0.10581147	Fruits and Vegetables	85.3908	OUT045	2002		Tier 2	Supermarket Type1
14	NCC54		Low Fat	0.171079215	Health and Hygiene	240.4196	OUT019	1985	Small	Tier 1	Grocery Store
15	FDU11	4.785	Low Fat	0.092737611	Breads	122.3098	OUT049	1999	Medium	Tier 1	Supermarket Type1
16	DRL59	16.75	LF	0.021206464	Hard Drinks	52.0298	OUT013	1987	High	Tier 3	Supermarket Type1
17	FDM24	6.135	Regular	0.0794507	Baking Goods	151.6366	OUT049	1999	Medium	Tier 1	Supermarket Type1
18	FDI57	19.85	Low Fat	0.05413521	Seafood	198.7768	OUT045	2002		Tier 2	Supermarket Type1
19	DRC12	17.85	Low Fat	0.037980963	Soft Drinks	192.2188	OUT018	2009	Medium	Tier 3	Supermarket Type2
20	NCM42		Low Fat	0.028184344	Household	109.6912	OUT027	1985	Medium	Tier 3	Supermarket Type3
21	FDA46	13.6	Low Fat	0.196897637	Snack Foods	193.7136	OUT010	1998		Tier 3	Grocery Store
22	FDA31	7.1	Low Fat	0.109920138	Fruits and Vegetables	175.008	OUT013	1987	High	Tier 3	Supermarket Type1
23	NCJ31	19.2	Low Fat	0.182619235	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.

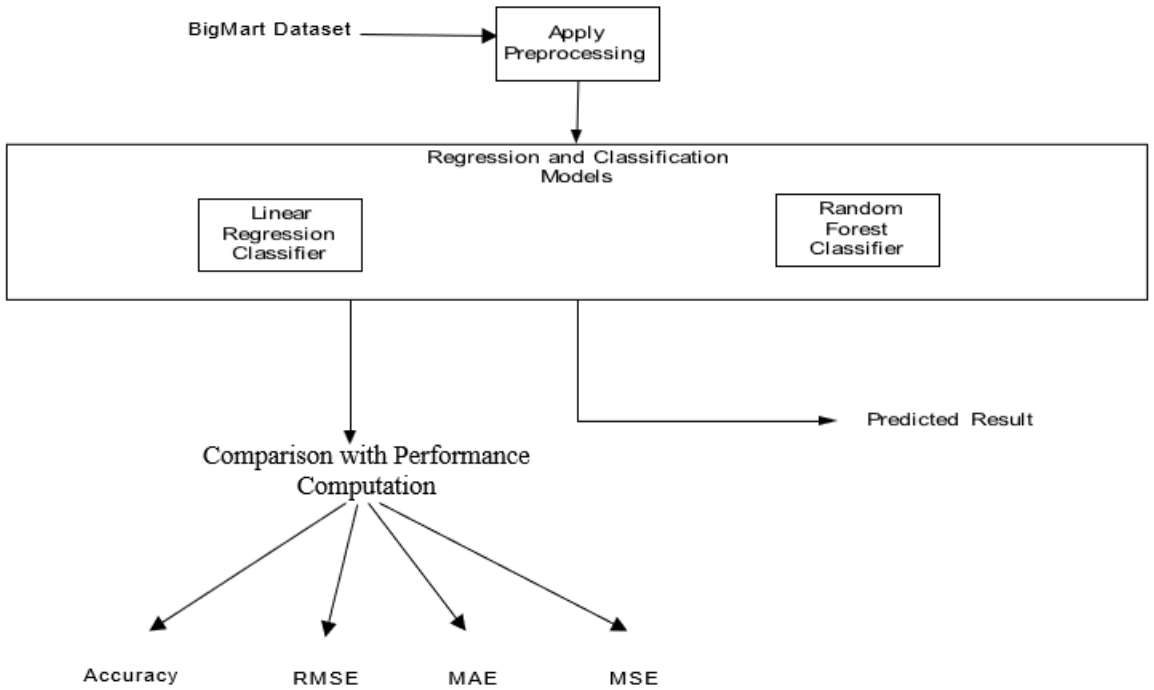

Fig3: Shows the proposed Architecture Diagram

A. Linear Regression:

• Create a shattered plot. A data pattern, either linear or nonlinear, and a variance (outliers). If the marking isn't linear, think about a transformation. Outsiders can recommend only removing them if there is a non-statistical rationale if this is the case.

• Use the residual plot (for the constant standard deviation assumption) and the normal probability plot (for the normal probability assumption) to connect the data to the least squares line and validate the model assumptions. If the presumptions seem to be unfounded, a transformation might be required.

• If necessary, change the data to the-least-squares form a regression line using the converted data.

• Write the least-square regression line equation when a "good-fit" classic is specified. If a change has been finished, go back to step 1; if not, move on to phase 5. consist of R-squared errors, estimation, and normal estimation.

B. Random Forest Regression:

• Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

IV. Result And Discussion

Linear Regression
TABLE 1: Shows the linear regression result on the various parameter

Parameter	value
MSE	1162.4412631603452
MAE	880.99990440845
R^2	0.5041875773270634

Random Forest regression
TABLE 2: Shows the Random Forest regression result onthe various parameter

Model	MSE	MAE	R^2
Linear Regression	1162.44	880.999	0.504
Random Forest Regression	0.547	782.253	1110.49

TABLE 3: Comparison of MAE, MSE, RMSE with the model

Model	MSE	MAE	R^2
Linear Regression	1162.44	880.999	0.504
Random Forest Regression	0.547	782.253	1110.49

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 using regression approach for predicting the sales based on past sales data, this method can improve the forecasting accuracy of linear regression. We can therefore conclude that Linear Regression and Random Forest regression provide better predictions than linear and polynomial regression approaches in terms of accuracy, MAE, and RMSE. you can forecast sales and create sales plans to avoid unexpected cash flows and manage production, staffing, and financing needs more effectively. In future work, ARIMA models showing time series plots can also be considered.

References

[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]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.

[6] 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.

[7] 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.

[8] "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.

[9] 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.

[10] A.S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.
[11] 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
[12] D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.
[13] X. Yua, Z. Qi, Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, Procedia Computer Science17 (2013) 1055–1062.
[14] E. Hadavandi, H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems anddata clustering: a Case study of the printed circuit board, Expert Systems with Applications 38 (2011) 9392–9399.