Big Mart Sales Prediction System

Prof. Nidhi Nigam Prof. Rachna Bahrawat Jessica Chouhan Palak Jaiswal

(Department of Computer Science and Information Technology, Acropolis Institute of Technology and Research, Indore Madhya Pradesh, India).

(nidhinigam@acropolis.in , rachanabahrawat@acropolis.in , jessicachouhan20207@acropolis.in , palakjaiswal20126@acropolis.in)

Abstract: *Nowadays shopping malls and Big Marts keep the track of their sales data for each and every individual item for predicting future demand of the customer and updating the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data stored in the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine-learning techniques for the retailers like Big Mart. In this paper, we propose a predictive model using XG boost Regressor technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models. A retail company wants a model that can predict accurate sales so that it can keep track of customers' future demand and update them in advance of the sale inventory. In this work, we propose a Grid Search Optimization (GSO) technique to optimize the parameters and select the best tuning hyper parameters, the further ensemble with Xgboost techniques for forecasting the future sales of a retail company such as Big Mart and we found our model produces the better result.*

Keywords: Machine Learning, Data Exploration, Sales Forecast, Random Forest, Linear Regression.

INTRODUCTION

Big Mart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientists out there to help them create a model that can predict the sales, per product, for each store to give accurate results. Big Mart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities.

With this information, the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure the success of their business. Big Mart is a massive network of stores that spans the globe. Big Mart's trends are extremely important, as data scientists analyse them by product and location to identify potential centers. Using a computer to predict Big Mart sales allows data scientists to explore different patterns by shop and product to get the best results. Many businesses rely largely on their information base and require market forecasting. Forecasting involves evaluating data from a wide variety of sources, including consumer trends, buying behaviours, and other considerations. This research would also assist businesses in properly managing their financial means. And that is where machine learning can really be put to good use. In this paper, we employ data mining approaches including discovery, data transformation, feature development, model construction, and testing to forecast sales using various machine learning algorithms. This approach involves pre-processing raw data acquired by a large mart

for missing data, abnormalities, and outliers. After that, an algorithm will be trained to create a model depending on the data.

Everyday competitiveness between various shopping centers 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 a period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services.

RELATED WORK

(Cheriyan et al.) This study looks into the judgments that should be made experimental results and the insights gained via data visualization. It made use of data mining methods. The Gradient Boost method has been found to be the most accurate in predicting future transactions. (Armstrong J) Three modules, hive, R programming, and tableau, were used to forecast sales. By looking at the store's past, you may have a better knowledge of the income and make changes to the objective to make it more successful. To achieve the findings, key values are retrieved inside the diagram to decrease all intermediate values by lowering the intermediate key feature. (Panjwani et al.) The aim of the study is to provide appropriate findings for predicting a firm's future sales or needs using approaches such as Clustering Models and metrics for sales forecasts. The algorithmic approaches' potential is assessed and employed in further study as a result. (Manpreet Singh et al.) Inspection of data obtained from a retail store and projection of future store management techniques are carried out in this study. The impacts of numerous sequences of events, such as meteorological conditions, vacations, and so on, may genuinely change the status of various departments, therefore it also analyses and evaluates these effects and their impact on sales. (Fawcett, Tom and Foster J. Provost) The method of identifying suspicious behavior using an automated prototype is described in this study. For the purpose of completing this acceptable prototype, many machine-learning methods were used. Here, data mining and constructive induction approaches are used to uncover the disparity in cell phone owners' behaviour. (Demchenko et al.) To forecast sales, a generic linear method, a decision tree approach, and a decent gradient approach were employed. The original data set evaluated included a large number of entries, but the final data set utilized for analysis was significantly less than the original since it included non-usable data, duplicate entries, and unimportant sales data.

LITERATURE SURVEY

A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression (2018) Kadam, H., Shevade, R., Ketkar, P. and Rajguru. A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression used Random Forest and Linear Regression for prediction analysis which gives less accuracy. To overcome this, we can use XG boost Algorithm which will give more accuracy and will be more efficient. Forecasting methods and applications (2008) Makridakis, S., Wheelwrigh. S. C., Hyndman. R.J. Forecasting methods and applications contain a Lack of Data and short life cycles. So some of the datalike historical data, and consumeroriented markets face uncertain demands and can be predicted for accurate results. Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018) C. M. Wu, P. Patil, and S. Gunaseelan. Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data Used Neural Network for comparison of different algorithms. To overcome this Complex model

like neural networks is used for comparison between different algorithms which is not efficient so we can use simpler algorithms for prediction. Prediction of retail sales of footwear using feed-forward and recurrent Neural Networks (2018) by Das, P., Chaudhury. Prediction of retail sales of footwear using feed-forward and recurrent neural networks used neural networks for prediction of sales. Using a neural network for predicting weekly retail sales, is not efficient, So XG boost can work efficiently.

PROPOSED SYSTEM

Data Processing and Methodology

- Data Collection: We have collected the data securely in accordance with an agreed methodology. The procedure for the collected data may differ from client to client and is dependent on the type, quantity, availability, and need of data.
- Data Cleaning and Pre-processing: The collected data is passed through a 'cleaning' process, so as to make sure that the data is segregated properly and identified gaps in the data are filled with the appropriate information, making data compatible and also fixing errors in storage systems which can cause data redundancy.
- Data Modelling: This is primarily a process in which the given dataset and the objects in it are analyzed to get a clear view of the requirements that may help us support our business model. Based on the analysis of patterns present in the data, models are then created on the established flow of the project. This flow offers better assistance in the utilization of the previously agreed upon the semi-formal model that showcases the features of the project. It also provides guidance to follow the relation between the data objects and other objects.
- Data Prediction: Machine Learning prediction mod using the data. This will then be applied to the pre-processed dataset. Some of the Models to be used for the prediction are:
  - Linear Regression
  - Random Forest
  - Decision tree
  - XG Boost Regressor
- Data Visualization: Data Analyzed is then further picturized for customers and the admin to reach out conclusions and take effective decisions.

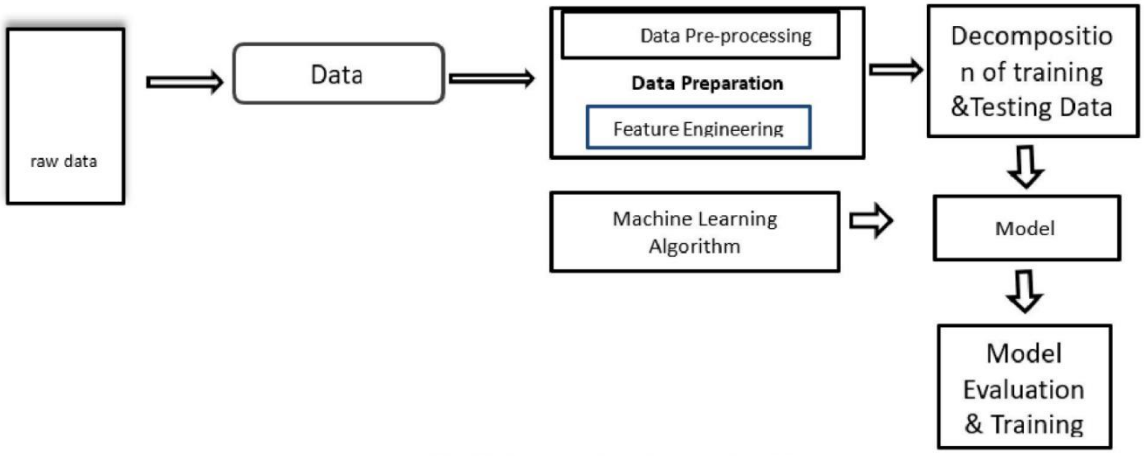


Fig:-Working procedure of proposed model

Results

To forecast BigMart's revenue, simple to advanced machine learning algorithms have been implemented, such as Linear Regression, Ridge Regression, Decision Tree, Random Forest, XGBoost. It has been observed that increased efficiency is observed with XGBoost algorithms with lower RMSE rating. As a result, additional Hyperparameter Tuning was conducted on XGBoost with Bayesian Optimization technique due to its quick and fairly simple computation, which culminated in the acquisition of the lowest RMSE value and making the model better matched to the underlying results. The submission file detailing Item Outlet Sales for Item based on the Model is resulted.

Platform and Language:

Python is a general purpose, interpreted-high level language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed as the 'batteries included language' for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient.

In this work, the Python libraries of NumPy, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Panda's tool of Python has been employed for carrying out data analysis. Random forest regressor is used to solve tasks by assembling random forest method. As a development platform, Jupyter Notebook, which proves to work great due to its excellence in 'literate programming', where human friendly code is punctuated within code blocks, has been used.

Data Modelling and Observations:

Correlation is used to understand the relation between a target variable and predictors. In this work, Item-Sales is the target variable and its correlation with other variables is observed. Considering the case of Item-Weight, the feature item weight is shown to have a low correlation with the target variable Item-Outlet-Sales in Fig.

	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	
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	
24	FDG52	13.65	LF	0.065630844	Frozen Foods	47.7402	OUT046	1997	Small	Tier 1	Supermarket Type1	
25	NCL19		Low Fat	0.027447057	Others	142.347	OUT019	1985	Small	Tier 1	Grocery Store	
26	FDS10	19.2	Low Fat	0.035178935	Snack Foods	180.7318	OUT035	2004	Small	Tier 2	Supermarket Type1	
27	FDX22	6.785	Regular	0.038455125	Snack Foods	209.4928	OUT010	1998		Tier 3	Grocery Store	

When a predictive model generated from any supervised learning regression method is applied to the dataset, the process is said to be data scoring. The above model score clearly infers about Data Scoring. The probability of a product's sales to rise and sink can be discussed and understood on the basis of certain parameters. The vulnerabilities associated with a product or item and further its sales are also necessary and play a very important role in our problem□solving task. Further, a user authentication

mechanism should be employed to avoid access from any unauthorized users and thus ensuring all results are protected and secured.

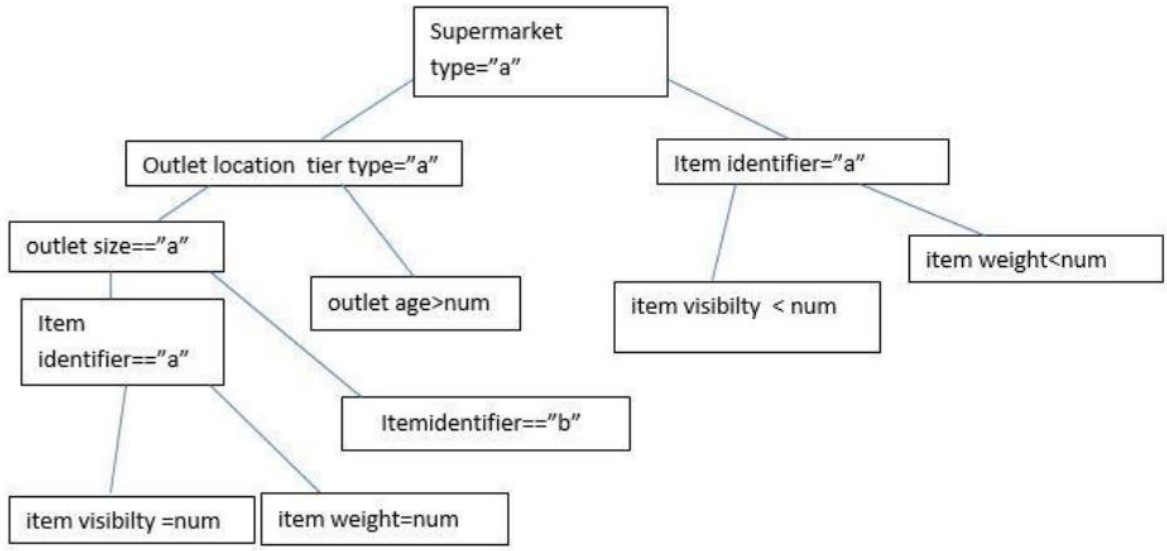
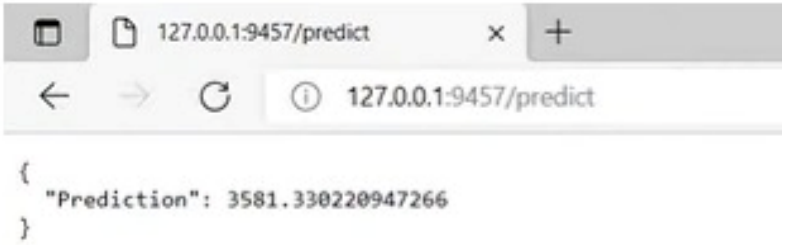


Fig 10: Flowchart for division of dataset on various factors (having proper leaves after pruning)

Prediction results and Conclusion:



Discussion

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analysed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life [1]. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects. In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results [2]. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key

To forecast BigMart's revenue, simple to advanced machine learning algorithms have been implemented, such as Linear Regression, Ridge Regression, Decision Tree, Random Forest, XGBoost. It has been observed that increased efficiency is observed with XGBoost algorithms with lower RMSE rating. As a result , additional Hyperparameter Tuning was conducted with Bayesian Optimization technique due to its quick and fairly simple computation, which culminated in the acquisition of the lowest RMSE value and making the model better matched to the underlying results. The submission file detailing Item Outlet Sales for Item based on the Model is resulted.

Various machine learning algorithms like Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Decision Tree, AdaBoost, XGBoost have been built to predict the sales revenue of Big Mart. It's been found that the most efficient algorithm to predict the sales revenue of Big mart is observed with Gradient Boosted Decision Tree and Random Forest algorithms having the least RMSE value among other algorithms.

CONCLUSION

In this work, the effectiveness of various algorithms on the data on revenue and review of, best performance here proposes a software to using regression approach for predicting the sales centred 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. We have designed a predictive model using ensemble techniques with this algorithm in the Big Mart dataset for forecasting future sales of a particular store or outlet of Big Mart. help big marts to refine their methodologies and strategies which in turn helps them to increase their profit. The results predicted will be very useful for the executives of the company to know about their sales and profits. them the idea for their new locations or Centre's of Big-mart.

ACKNOWLEDGMENT

We felt great pleasure in submitting this paper on Big Mart Sales Prediction Using Machine Learning. First, we could like to express our gratitude to Almighty God, Creator of the whole universe. A huge thank you to Prof. Rachna Bahrawat and Prof. Nidhi Nigam, Department of Computer Science and Information Technology, for your supreme support, guidance, and patience. We would like to express our sincere gratitude and appreciation to all our colleagues who have helped us in one way or another in the writing of this research paper.

REFERENCES

1) Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad (2021, June). Big Mart Sales Prediction using Machine Learning.2021 International Journal of Research Thoughts (IJCRT).
2) Inedi. Theresa, Dr. Venkata Reddy Medikonda,K.V.Narasimha Reddy. (2020, March). Prediction of Big Mart Sales using Exploratory Machine Learning Techniques 020 International Journal of Advanced Science and Technology (IJAST).
3) https://www.javatpoint.com/machine-learning-random-forest-algorithm
4) https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

5) Bohdan M. Pavlyshenko (2018, August 25). Rainfall Predictive Approach for La Trinidad, Benguet using Machine Learning Classification. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP).
6) Pavan Chatradi, Meghana, Avinash Chakravarthy V, Sai Mythri Kalavala , Mrs.Neetha KS (2020), Improvizing Big Market Sales Prediction, Volume 12 Issue 4 (https://www.xajzkjdx.cn/gallery/423- april2020.pdf)
7) K. Punam, R. Pamula, and P. K. Jain, "A two-level statistical modelfor big mart sales prediction," in 2018 International Conference on Computing, Power and Communication Technologies (GUCON).
8) IEEE, 2018, pp. 617–620.
9) P. Das and S. Chaudhury, "Prediction of retail sales of footwear usingfeedforward and recurrent neural networks," Neural Computing and Applications, vol. 16, no. 4-5, pp. 491–502, 2007.
10) C.-W. Chu and G. P. Zhang, "A comparative study of linear andnonlinear models for aggregate retail sales forecasting," International Journal of production economics, vol. 86, no. 3, pp. 217–231, 2003.
11) S. Beheshti-Kashi, H. R. Karimi, K.-D. Thoben, M. Lutjen, and M. Teucke, "A survey on retail sales forecasting and prediction in fashion markets," Systems Science & Control Engineering, vol. 3,no. 1, pp. 154–161, 2015.
12) S. Asur and B. A. Huberman, "Predicting the future with socialmedia," in Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent TechnologyVolume 01. IEEE Computer Society, 2010, pp. 492–499.
13) V. Dhar and E. A. Chang, "Does chatter matter? the impact of usergenerated content on music sales," Journal of Interactive Marketing, vol. 23, no. 4, pp. 300–307, 2009.
14) J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stockmarket," Journal of computational science, vol. 2, no. 1, pp. 1–8