International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

PREDICTIVE ANALYSIS FOR BIG MART SALES USING LINEAR AND XGBOOST ALGORITHMS

P.Vetrivel
Department Of Computer Science and Engineering,
Sree Sowdambika College of Engineering,
Aruppukottai, India

R.Jayanthi
Department Of Computer Science and Engineering,
Sree Sowdambika College of Engineering,
Aruppukottai, India
rjayanthi205@gmail.com

H.Indira
Department Of Computer Science and Engineering,
Sree Sowdambika College of Engineering,
Aruppukottai, India
indirah2001@gmail.com

S.Gayathri
Department Of Computer Science and Engineering,
Sree Sowdambika College of Engineering,
Aruppukottai, India
gayathri6717@gmail.com

Abstract: *Currently, supermarkets, Big Marts tracks sales data for each item to anticipate potential consumer demand and review asset management. Confusion and common trends are often found by digging into a warehouse data store. For vendors like Big Mart, the resulting data can be used to predict future sales volume using various machine learning strategies such as the big mart. The forecasting model was developed using Xgboost and Linear retrieval techniques to predict business sales such as Big -Mart, and it was found that the model was superior to existing models.*

Keywords: *Data mining, Machine learning, Continuous values prediction, Stacking, Sales Prediction*.

I. INTRODUCTION

The daily competition between the various shopping malls as and as the large marts grow exponentially, is violent simply due to the rapid development of the world's largest shopping malls and online shopping. Each market seeks to offer personal and limited time deals to attract more time-dependent customers, so that the volume of each item is balanced in managing organizational stock, transportation and planning services. The current machine learning algorithm is very advanced and offers ways to predict or predict sales of any type of organization, it is more profitable to win - the lower the costs used to predict. Better forecasting is always helpful, in developing and refining market marketing strategies, which are also very helpful.

The main purpose of our project is to,

- Successfully preview selected data.
- Introduce machine learning for better performance.
- To calculate error rate such as, MAE, MSE, RMSE.

Improving overall performance

Problem Statement:

There has been a lot of research done on this relevant process. This different result was due to differences in the different aspects of the methods used in the study. Because of all these factors, it is not easy to compare and choose the one that can be described as the best. Therefore, there is always room for the development of a better method that is suitable for a particular application.

II. LITERATURE REVIEW

Predicting the Monthly Trading Time Series: Case Study, 2018

Author: Giuseppe Nunnari, Valeria Nunnari

Methodology:

This paper presents an example of the prediction of a monthly trading time series recorded by the US Census Bureau from 1992 to 2016. The modeling problem is addressed in two steps. First, the real-time series is reliably removed using a moving window measurement method. Next, the remaining time series modeled Non-linear Auto-Regressive (NAR) models, using both Neuro-Fuzzy and Feed-Forward Neural Networks methods. The beauty of predictive models, equally tested by calculating bias, mae errors and rmse. Finally, the model skills model is calculated by considering a continuous traditional model as a reference. The results show that there is more ease in using the proposed methods, compared to the reference.

Advantages:

• A low-volume ellipsoid model is proposed to undermine performance.

Disadvantages:

• Rushing to failure forecasts is low.

An advanced Adaboost algorithm based on uncertain tasks, for 2015

Author: ShuXinqing, Wang Pan

Methodology:

Boosting is one of the algorithms that can improve the accuracy of weak sections, and Adaboost has been widely and successfully used in classification, discovery and data mining problems. In this paper, we introduce a new method of calculating parameters, Adaboost-AC, which uses a good speed-up function to determine the weights of weak dividers. The new algorithm is compared to the Adaboost culture based on the UCI website and its promising performance is reflected in the test results.

Advantages:

• Makes accurate class comparisons with other methods.

Disadvantages:

• It does not work well.

Extraction of Patterns through Mining Methods Through Window Creation, 2020

Author: Suresh K, Praveen O

Methodology:

Data mining techniques are widely used in commercial areas to classify information on a website. The information mine includes the use of Utility Pattern Mining shows about the time to apply strategies in the processing of object sets. Utility Pattern Mining (UPM) makes sense for important data that successfully evaluates pattern identification. In this research paper, Hierarchical High Average Utility Pattern Pattern Mining (HAUPM) is proposed for the e-commerce and retail industries. Unlimited streaming data may produce continuous results required for periodic updates. Hierarchical High Average Utility Pattern Mining (HAUPM) is used to perform tasks in unlimited broadcast information on a website. Which modern algorithm is made on information that has a greater impact than the latest information. These data sets provide profitable results in the retail industry based on making customers buy trendy products in the market. H-HAUPM has been chosen over other strategies to obtain high-impact materials based on the accuracy of production sets, which do not consume much space to use, measure and maintain consistency.

Benefits:

• More Reliable.

Disadvantages:

• It works well and does not give a complete result.

In the development of advanced flexible models to effectively predict stock indicators using clonal-PSO (CPSO) and PSO strategies, 2018

Author: R. Majhi, G. Panda, G. Sahoo, and A. Panda

Methodology:

The current paper introduces clonal particle swarm optimization (CPSO) and PSO methods to develop effective short-term and long-term forecasting models for the S&P 500 and DJIA stock indicators. The basic structure of the models is a flexible line component whose weight is repeatedly updated with PSO and CPSO-based learning rules. Technical indicators are calculated from previous stock indicators and are used as inputs for models. Using experimental performance simulation studies, MSE error, training time and intermediate percentage error (MAPE) for CPSO, PSO and GA-based models found in all predictive grades. Comparison of these results shows that CPSO and PSO-based models produce higher performance compared to the single GA. However the CPSO model offers much better performance compared to the other two.

Advantages:

• High performance.

Disadvantages:

• Low accuracy and precision.

Support Reduction of Vector / Magazine Sales Forecast Vector, 2013

Author: Xiaodan Yu, Zhiquan Qi, YuanmengZhaoc

Methodology:

Advances in information technology have changed our lives in many ways. There is a tendency for people to look at news and news online. Under this scenario, it is more urgent for traditional media companies to predict

sales (i.e. newspapers / magazines) than ever before. Previous methods of predicting newspaper / magazine sales are mainly focused on building retrospective models based on sample data sets. But such reduction models can face the problem of over-equality. Recent research on mathematical theory has proposed a novel approach, which is support vector regression (SVR), to overcome the problem of over-equality. In contrast to the standard retreat model, the SVR's goal is to achieve a minimum structural risk rather than a minimum operational risk. Therefore, this study used vector regression in predicting the sale of newspaper / magazine sales. Tests have shown that SVR is a superior approach to this type of activity.

Advantages:

• High performance.

Disadvantages:

• This method can not only take longer to test and has a higher cost.

• The destructive method model is high accuracy, but has very high parameters, high complexity, and a large number of calculations.

III. EXISTING SYSTEM

In existing environments, data mining techniques are widely used in commercial areas to classify information on a website. The information mine includes the use of Utility Pattern Mining shows about the time to apply strategies in the processing of object sets. Utility Pattern Mining (UPM) makes sense for important data that successfully evaluates pattern identification. In this research paper, Hierarchical High Average Utility Pattern Pattern Mining (HAUPM) is proposed for the e-commerce and retail industries. Unlimited streaming data may produce continuous results required for periodic updates. Hierarchical High Average Utility Pattern Mining (HAUPM) is used to perform tasks in unlimited broadcast information on a website. Which modern algorithm is made on information that has a greater impact than the latest information. These data sets provide profitable results in the retail industry based on making customers buy trendy products in the market. H-HAUPM has been chosen over other strategies to obtain high-impact materials based on the accuracy of production sets, which do not consume much space to use, measure and maintain consistency.

DISADVANTAGES

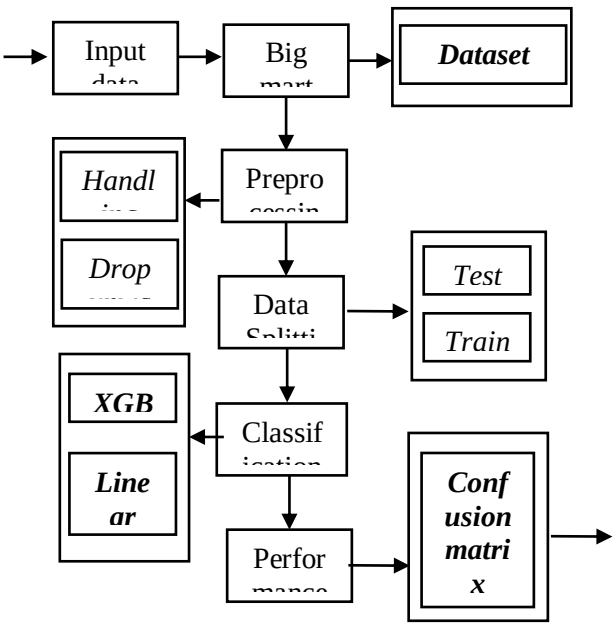
• Results are low compared to the proposed.
• Time consumption is high.
• Theoretical limitations.

IV. PRPOSESD SYSTEM

Much work has been done with the intention that so far the venue is predictable. A brief overview of the important function in the big_mart deals field is shown in this section. Many other Measurable methods, for example, by regression, (ARIMA) Auto-Regressive Integrated Moving Average, (ARMA) Auto-Regressive Moving Average, have been used to create predictive estimates for a few deals. However, expectations are an advanced problem and are influenced by both external and internal factors, and there are two important risks of measurable approach as stated in A. S. Weigend et Occasional combination of quantum multiplication method and (ARIMA) Automatic-Regressive Integrated Moving Average method of dealing with the expected daily feasts recommended by N. S. Arunraj also found that the performance of each model was lower than that of the crossover model.

ADVANTAGES

- Works well on large numbers of databases.
- The test result is high compared to the existing system.
- Time consumption is low.
- Provide accurate prediction results



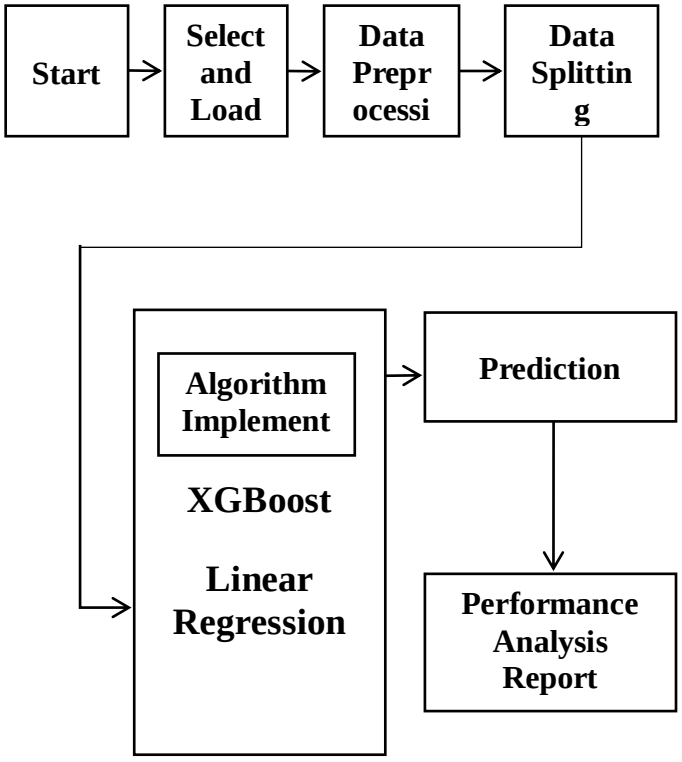


Figure 2 Flow Diagram

V. EXPERIMENTAL RESULTS AND DISCUSSION

DATA SELECTION:

• The input data was collected in an online database form of a website called kaggle.com.
• In this project all have a test set and train data set in the 5000 database test set and the 8000 data train set.
• In our collected database read this process using pandas.

```
****************** data selection**********************
**********Train  data ************
  Item_Identifier  Item_Weight  ...        Outlet_Type  Item_Outlet_Sales
0           FDA15         9.30  ...  Supermarket Type1          3735.1380
1           DRC01         5.92  ...  Supermarket Type2           443.4228
2           FDN15        17.50  ...  Supermarket Type1          2097.2700
3           FDX07        19.20  ...      Grocery Store           732.3800
4           NCD19         8.93  ...  Supermarket Type1           994.7052

[5 rows x 12 columns]
**********Test  data *************
  Item_Identifier  Item_Weight  ... Outlet_Location_Type        Outlet_Type
0           FDW58       20.750  ...               Tier 1  Supermarket Type1
1           FDW14        8.300  ...               Tier 2  Supermarket Type1
2           NCN55       14.600  ...               Tier 3      Grocery Store
3           FDQ58        7.315  ...               Tier 2  Supermarket Type1
4           FDY38          NaN  ...               Tier 3  Supermarket Type3

[5 rows x 11 columns]
```

Figure 3.Data Selection

DATA PREPROCESSING:

• Pre-data processing is the process of extracting unwanted data from the database.

• Pre-processing data processing functions are used to convert data into a structure suitable for machine learning.

• This step includes cleaning up the database by removing unimportant or damaged data that may affect the accuracy of the database, making it more efficient.

• No data deletion

• Coding category data

• Missing data deletion: In this process, empty values such as deficit values and Nan values are replaced by 0.

• Non-duplicate values are extracted and data is purified from any abnormalities.

• Category data encoding: That category data is defined as a variable with a limited set of label values.

• That most machine learning algorithms require numerical input and output variables.

```
******************Handling Null Values From Train data******************
*****************Before Preprocess*********************
Item_Identifier                0
Item_Weight                 1463
Item_Fat_Content               0
Item_Visibility                0
Item_Type                      0
Item_MRP                       0
Outlet_Identifier              0
Outlet_Establishment_Year      0
Outlet_Size                 2410
Outlet_Location_Type           0
Outlet_Type                    0
Item_Outlet_Sales              0
dtype: int64
```

Figure 4. Before Data Preprocessing

```
******************After Preprocess Train data*******************
Item_Identifier               0
Item_Weight                   0
Item_Fat_Content              0
Item_Visibility               0
Item_Type                     0
Item_MRP                      0
Outlet_Identifier             0
Outlet_Establishment_Year     0
Outlet_Location_Type          0
Outlet_Type                   0
dtype: int64
```

Figure 5.After Data Preprocessing

DATA SPLITTING:

• During the machine learning process, data is required for learning to take place.

• In addition to the required training data, test data is required to test the performance of the algorithm but here we have a set of training and testing data separately.

• In our process, we should classify as training and testing into x_train, y_train, x_test, y_test.

• Data segregation is the act of dividing available data into two parts, usually for verification purposes.

• One part of the data is used to improve the prediction model and the other is used to evaluate the performance of the model.

```
*********Data Splitting**************
*********train**************
   Item_Identifier  ...  Outlet_Location_Type
0              156  ...                     0
1                8  ...                     2
2              662  ...                     0
3             1121  ...                     2
4             1297  ...                     2

[5 rows x 9 columns]
0    1
1    2
2    1
3    0
4    1
Name: Outlet_Type, dtype: int32
```

Figure 6. Data Spilting

REGRESSION ALGORITHMS:

- In our process, we should use machine learning algorithms like these

  1) XGBoost Regression
  2) Linear Regression

XGBoost Regression:

• "Extreme Gradient Boosting" is similar but very effective in the grading process. It has both a line model solution and a tree algorithm.

• Allows xgboost in any event faster than current slope stabilization actions. Supports a variety of targeted skills, including repetition, ordering and balancing. Since "xgboost" is very high in scientific power but often delayed by organization, it is worth some competition. It is similarly useful for counter-authorization to obtain important features.

```
**********XGBoost Algorithm**********
MAE: 0.580135325891538
MSE : 0.5525050985149149
RMSE : 0.7433068669902861
```

Figure 7. XGBoost Algorithm

Linear Regression:

• Create a separate line or non-line pattern for data and variations (outliers). Consider the change if marking is out of line. If so, outsiders, it may be advisable to eliminate them only if there is a non-statistical excuse.

• Connect the data to a small square line and confirm the guessing of the model using the residual area (by the normal deviation guess) and the normal probability structure (by the normal probability estimate). A change may be required if the speculation made does not appear to be met.

• If necessary, convert data into very small squares using converted data, forming a regression line.

• When the change is complete, return to previous process 1. If not, proceed to step 5.

• When defining an old "well-matched" type, write a small square number of retreat line. It contains standard measurements, estimates, and duplicate errors..

```
**********Linear Regression Algorithm**********
MAE : 0.4192722032738482
MSE : 0.24209906998602537
RMSE : 0.49203563893891406
```

Figure 8.Linear Regression Algorithm

RESULT COMPARISION:

• The Final Outcome will be generated based on all predictions. The effectiveness of this proposed method is tested using measures such as,

1) MAE

2) MSE

3) RMSE

• In this process we compare the three results above to produce a graph..

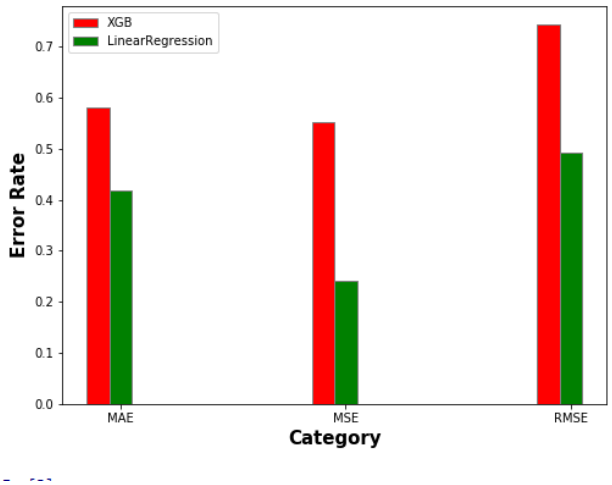


Figure 9. Comparsion Graph

VI. CONCLUSION

In this work, the efficiency of various algorithms in revenue data and updates, the best performance algorithm, here propose a drop-down software to predict sales based on sales data from the past the accuracy of linear prediction can be improved in this way, polynomial retreat , Ridge reverse, and Xgboost reverse can be determined. Therefore, we can conclude that ridge retreat and Xgboost provide better predictions about Accuracy, MAE and RMSE than Linear and polynomial retrieval methods.

VII. FUTURE ENHANCEMENT

In the future, predicting sales and building a sales system can help prevent unexpected cash flows and manage productivity, labor and financial needs more effectively. In future work we can also consider an ARIMA model that shows a time series graph.

REFERENCES

1. Ching Wu Chu and Guoqiang Peter Zhang, “Comparative study of direct and indirect models of integrated sales forecasting models”, Int. Journal Production Economics, vol. 86, pages 217-231, 2003.

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

3. Suma, V., and ShavigeMalleshwara Hills. "Predicting Demand-Based Data Mining in India Indian Renewable Energy Market." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101- 110

4. Giuseppe Nunnari, Valeria Nunnari, "Predicting a Monthly Monthly Sales Season: Case Study", Proc. of IEEE Conf. in Business Informatics (CBI), July 2017.

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

6. Zone-Ching Lin, Wen-Jang Wu, “Multi-Line Analysis of Model Spatial Model Specification”, IEEE Trans. in Semiconductor Manufacturing, vol. 12, no. 2, pages 229 - 237, May 1999.

7. O. Ajao Isaac, A. AbdullahiAdedeji, I. Raji Ismail, “Polynomial Regression Model of Make Cost Prediction In Mixed Cost Analysis”, Int. Journal of Mathematical Theory and Modeling, vol. 2, no. 2, pages 14 - 23, 2012.

8. C. Saunders, A. Gammerman and V. Vovk, "The 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". HuanXu, Constantine Caramanis, Member, IEEE, and ShieMannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration. ”Advanced Adaboost algorithm based on uncertain tasks”. of Technology Wuhan, China.

10. XinqingShu, Pan Wang, "Advanced Adaboost Algorithm based on uncertain tasks", Proc. of Int. Conf. in Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.

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

12. N. S. Arunraj, D. Ahrens, A combined annual average rate of automatic movement and quantile regression of daily food sales forecasts, Int. J. Production Economics 170 (2015) 321-335P

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

14. X. Yua, Z. Qi, Y. Zhao, Vector Postpacking of News / Magazine Sales Forecast, Procedia Computer Science 17 (2013) 1055–1062.

15. E. Hadavandi, H. Shavandi, A. Ghanbari, Advanced Methodology for Predicting the Integration of Comprehensive Genetics and Data Collection: Printed Regional Board Study Case, 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, Using computational intelligence methods to predict the sale of newly published books in a real business management environment, Knowledge-Based Systems 115 (2017) 133-151.

17. R. Majhi, G. Panda and G. Sahoo, “Development and performance evaluation of a FLAN-based stock market forecast model.” Expert Systems with Applications, vol. 36, issue 3, part 2, pages 6800-6808, April 2009.

18. Pei Chann Chang and Yen-Wen Wang, "Fuzzy Delphi and back-to-back sales forecasting model in the PCB industry", Application systems systems, vol. 30, p. 715-726, 2006.

19. R. J. Kuo, Tung Lai HU and Zhen Yao Chen "radial basis function function for neural trading forecasting networks", 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 advanced models of Adaptive Prediction Efficient Prediction of Stock Indices using Clonal-PSO (CPSO) and PSO Techniques”, International Journal of Business Forecasting and Market Intelligence, vol. 1, no. 1, pages 50-67, 2008