Predictive Analysis for Big Mart Sales using Machine Learning Algorithms

Nayana R[1], Chaithanya G[2], Meghana T[3], Narahari K S[4], Sushma M[5]
[1,2,3,4] CSE Department, Sri Krishna Institute of Technology, B'lore-560090, India
[5] Assistant Professor CSE Department, Sri Krishna Institute of Technology, B'lore-560090, India

***Abstract*:- Currently, supermarket run-centers, Big Marts keep track of each individual item's sales data like item name, price, etc.. in order to Meet 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 many 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 centers and huge marts is becoming higher intense, violent just because of the quick development of global malls also online shopping. Each market 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 Various methods for predicting or forecasting sales of any kind of organization, extremely beneficial to overcome low – priced used for prediction.

The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results with respect to the task's data.

This can then further be used for forecasting future sales by machine learning algorithms such as the random forests and simple or multiple linear regression model.

II. BACKGROUND STUDY

There has been an increasing demand in the e-commerce market for refurbished products across India during the last decade. Despite these demands, there has been very little research done in this domain. The real-world business environment, market factors, and varying customer behavior of the online market are often ignored in the conventional statistical models evaluated by existing research work. In this paper, we do an extensive analysis of the Indian e-commerce market using the data-mining approach for the prediction of demand for refurbished electronics. The impact of the real-world factors on the demand and the variables are also analyzed. Real-world datasets from three random e-commerce websites are considered for analysis. Data accumulation, processing, and validation are carried out by means of efficient algorithms. Based on the results of this analysis, it is evident that highly accurate predictions can be made with the proposed approach despite the impacts of varying customer behavior and market factors. The results of the analysis are represented graphically and can be used for further analysis of the market and launch of new products.

In 2019 Wang, Haoxian Combination of Green supply chain management, green product deletion decision, and green cradle-to-cradle performance evaluation with Adaptive-Neuro-Fuzzy Inference System (ANFIS) to create a green system. Several factors like the design process, client specification, computational intelligence, and soft computing are analyzed and emphasis is given on solving problems of the real domain. In this paper, the consumer electronics and smart systems that produce nonlinear outputs are considered. ANFIS is used for handling these nonlinear outputs and offers sustainable development and management. This system offers decision making considering multiple objectives and optimizing multiple outputs. The system also provides efficient control performance and faster data transfer.

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.

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 are used for comparison between different algorithms which is not efficient so we can use the simpler algorithm for prediction.

This paper presents a case study concerning the forecasting of monthly retail time-series recorded by the US Census Bureau from 1992 to 2016. The modeling problem is tackled in two steps. First, original time series are de-trended by using a moving window averaging approach. Subsequently, the residual time series are modeled by Non-linear Auto-Regressive (NAR) models, by using both Neuro-Fuzzy and Feed-Forward Neural Networks approaches. The goodness of the forecasting models is objectively assessed by calculating the bias, the MAE, and the RMSE errors. Finally, the model

skill index is calculated considering the traditional persistent model as a reference. Results show that there is a convenience in using the proposed approaches, compared to the reference one.

In 2015 Xinqing Shu, Pan Wang Boosting is one of the algorithms which can boost the accuracy of weak classifiers, and Adaboost has been widely and successfully applied to classification, detection, and data mining problems. In this paper, a new method of calculating parameters, Adaboost-AC, which uses the accelerated good fitness function to acquire the weights of the weak classifiers is presented. The new algorithm is compared with the traditional Adaboost based on the UCI database and its promising performance is shown by the experimental results.

Das, P., Chaudhury Prediction of retail sales of footwear using feedforward and recurrent Neural Networks (2018) Prediction of retail sales of footwear using feedforward and recurrent neural networks used neural networks for prediction of sales. Using the neural network for predicting weekly retail sales, which is not efficient, So XG boost can work efficiently.

Makridakis,S., Wheelwrigh.S.C., Hyndman. R.J Forecasting methods and applications (2008) Forecasting methods and applications contain a Lack of Data and short life cycles. So, some of the data like historical data, consumer-oriented markets face uncertain demands, can be a prediction for an accurate result.

In 2012 O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail Regression analysis is used across business fields for tasks as diverse as systematic risk estimation, production and operations management, and statistical inference. This paper presents the cubic polynomial least square regression as a robust alternative method of making cost prediction in business rather than the usual linear regression. The study reveals that polynomial regression is a better alternative with a very high coefficient of determination.

In 2013 X. Yua, Z. Qi, Y. Zhao Advances in information technologies have changed our lives in many ways. There is a trend that people look for news and stories on the internet. Under this circumstance, it is more urgent for traditional media companies to predict print's sales than ever. Previous approaches in newspapers/magazines' sales forecasting are mainly focused on building regression models based on sample data sets. But such regression models can suffer from the over-fitting problem. Recent theoretical studies in statistics proposed a novel method, namely support vector regression (SVR), to overcome the over-fitting problem. This study, therefore, applied support vector regression to the newspaper/magazines' sales forecasting problem. The experiment showed that SVR is a superior method.

III. METHODOLOGY

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 and the normal probability plot .A transformation might be necessary if the assumptions made do not appear to be met.

Linear regression formulas look like this:

$$Y=o1x1+ o2x2+\dots\dots onxn$$

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= b0+b1x1+ b2x12+ b2x13+..... bnx1n

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. $Min(\|Y - X(theta)\|^2 + \lambda\|theta\|^2)$

The usual regression equation forms the base which is written as: $Y = XB + e$

D. XGBoost:

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.

IV. IMPLEMENTATION

For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequence of steps as mentioned in Figure 1 and in this work we propose a model using Xgboost technique. Every step plays a vital role for building the proposed model. After preprocessing and filling missing values, we used ensemble classifier using Decision trees, Linear regression, Ridge regression, Random forest and Xgboost. Both MAE and RSME are used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it was found that the model will predict best using minimum MAE and RSME.

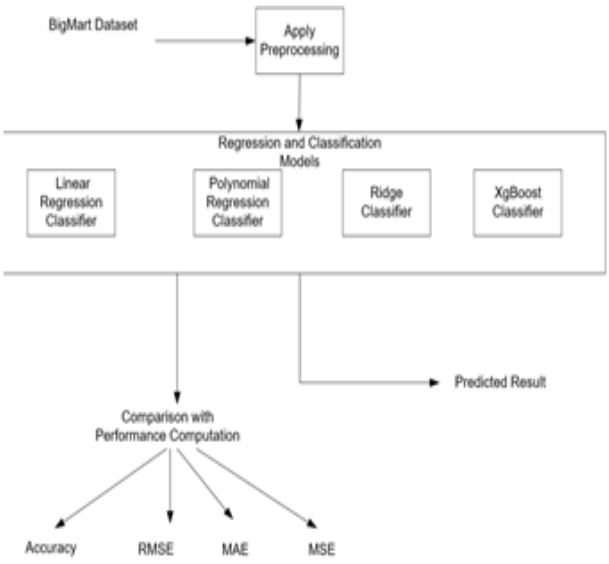


Figure 1: Architecture of the System

V. RESULTS

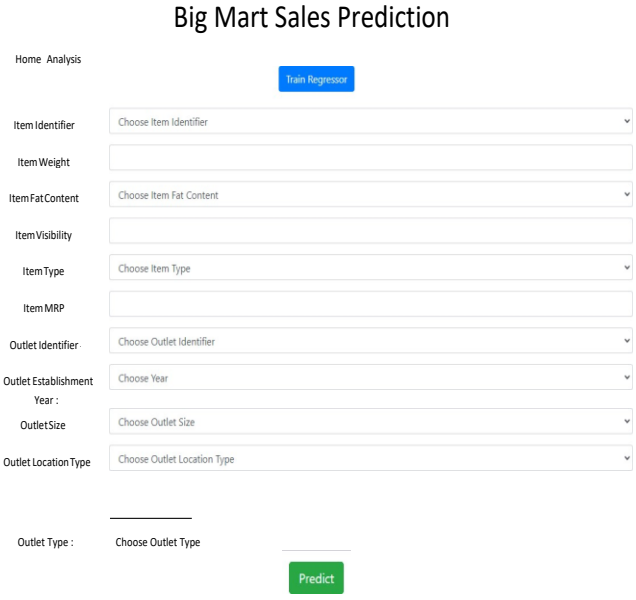
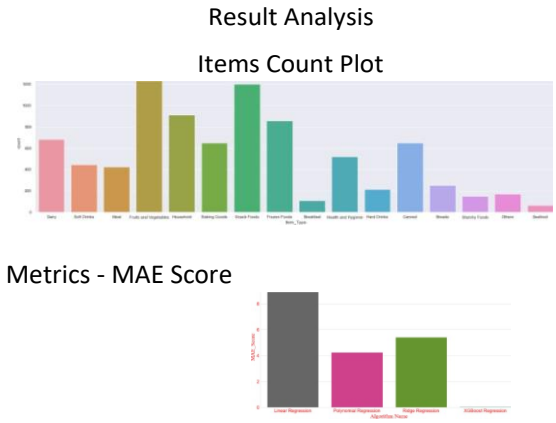


Figure 2: Big Mart Sales Prediction



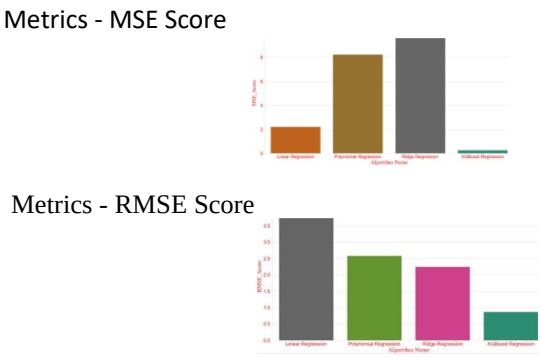


Figure 3: Result Analysis

VI. CONCLUSION

In this project, 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 a better prediction with respect to Accuracy, MAE, and RMSE than the Linear and polynomial regression approaches.

VII. ACKNOWLEDGEMENT

We would like to thank Assistant Professor Sushma M for his valuable suggestion, expert advice and moral support in the process of preparing this paper.

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. 217231, 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.

[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): 101110

[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 Linear Regression 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.