Turkish Journal of Computer and Mathematics Education *Vol.10 No.01 (2019), 631-637*
DOI: https://doi.org/10.17762/turcomat.v10i1.13559

Research Article

Implementation of Machine Learning Techniques for Big Mart Sales Forecast Analysis

Manika Manwal
Department of Computer Science & Information Technology, Graphic Era Hill University, Dehradun Uttarakhand India 248002

Abstract

The sales data for each individual item is now tracked by supermarket run-centers, Big Marts, in order to forecast possible consumer demand and revise inventory control. By mining the data store of the data warehouse, wide-ranging trends as well as anomalies are frequently found. The generated data may be utilised by merchants such as Big Mart that employ a number of machine learning techniques in order to forecast forthcoming sales volume. A prediction algorithm was developed to estimate the sales of an organisation like Big-Mart utilising Xgboost, Linear regression, Polynomial regression, also Ridge regression methodologies. It was shown that the model beats other models.

Keywords: Sales Data, Big Mart, Machine Learning Algorithms, Predictive Model.

1. Introduction

The competition among various shopping malls and huge supermarkets is becoming increasingly fierce and combative on a regular basis as a result of the fast growth of international malls and internet shopping. For the purpose of the company's stock control, transportation, plus logistical services, to pull in a huge number of clients in a short length of time and calculate the amount of sales for every item. Algorithm for machine learning currently in use are quite advanced and provides approaches for estimating or forecasting sales for possible type of company, that's extremely helpful to overcome low-cost methods utilised for prediction. Always more accurate forecasting is useful for creating and enhancing marketing plans for the market, which is also very beneficial.

Objectives:

The primary goal of our project is to efficiently preprocess the chosen dataset.

- To put machine learning into practise for greater performance.
- To determine the error rate, such as the mae, mse, and rmse
- To improve performance as a whole.

Numerous studies have been conducted using this important approach. This disparate outcome was brought about by the research's varied usage of methodologies. Due to all of these considerations, it is difficult to compare and select the approach that may be deemed the best. As a result, there is always potential for the establishment of improved methods that are appropriate for certain applications.

2. Literature Survey

In this essay, an analysis of a case study involving time series projections for monthly retail obtained by the US Census Bureau between 1992 and 2016 is presented. Two approaches are used to solve the modelling challenge. To begin with, the initial time series is de-trended utilizing moving windows averaging. Next, non-linear auto-regressive models are used to simulate the residual time series utilising neural networks, including feed-forward and neuro-fuzzy techniques. By computing the bias, the mae, with the rmse errors, forecasting prototype quality is formally evaluated. The traditional persistent model is serving as a guide in the final calculation of the model skill index. Results indicate that, when contrasted to the standard technique, utilising the proposed ways is more convenient. It is suggested to use a minimum volume ellipsoid framework to predict performance decline. However, the predicted run-to-failure rate for the system is low [1].

One approach that can improve the performance of weak classifiers is boosting, also Adaboost has been effectively used to solve many classification, detection, as well as data mining issues. Under this work, a novel parameter estimation technique called Adaboost-AC that acquires the weights of the weak classifiers using the accelerated good fitness function is described. Depending upon the UCI database, the novel algorithm has been contrasted to the established Adaboost, and the experimental results highlight its potential performance. It compares accurately with other approaches for categorization. The system's main flaw is its lack of efficiency [2].

In the world of business, information mining techniques are widely used to extract data from databases. Information mining involves using techniques like Utility Pattern Mining, which considers item sets while using time-based tactics. Utility Pattern Mining is appropriate for large datasets that evaluate successfully in pattern detection. Hierarchical High Average Utility Pattern Mining is suggested pertaining to the e-commerce and retail sectors in this research article. Unbounded stream data may produce consistent results that need to be updated dependent on the passage of time. The database's unbounded stream information was subjected to operations using HAUPM. When information with a greater effect than more current information is subjected to a state-of-the-art algorithm. By encouraging customers to purchase items that are popular in the market, these data sets produce beneficial results for the retail sector. H-HAUPM is preferred over other approaches because it can generate itemsets accurately, doesn't take up a lot of space when in use, is scalable, and maintains consistency. Comparatively, the proposed system is more trustworthy. The system's main flaw is that it is less efficient and does not produce ideal results [3].

In order to create effective adaptive forecasting frameworks for the short- and long-term prediction of the S&P 500 and DJIA stock indices, the current work offers novel clonal particle swarm optimization and PSO approaches. The main building block of the modeling techniques is an adaptive linear combiner, whose weights are adjusted repeatedly using learning rules based on PSO and CPSO. Technical indications are calculated using historical stock indexes and fed into the algorithms as input. In a simulation research, the Convergence rate's forecasting abilities, minimal mean square error, training time, also mean average percentage error are calculated for all prediction ranges using CPSO, PSO, and GA-based approaches. These findings show that the suggested CPSO with PSO-based paradigms outperform the GA one in terms of performance. However, when compared to the other two models, Performance-wise, the CPSO paradigm performs the best. The great efficiency of the suggested strategy is a benefit. However, the suggested approach performs poorly in terms of accuracy [4].

The publisher must decide how many copies of a new book should be printed for distribution to retailers. Producing too many copies is problematic since it results in an excess of inventory and a loss of investment, but printing too few copies will also have a detrimental effect on the economy. In this essay, we address the issue of forecasting overall sales so that to produce the appropriate number of books even before they are delivered to bookstores. Three stages of analysis were carried out: a preliminary exploratory analysis using strategies for data visualisation, a feature selection procedure that makes use of a variety of methodologies to pinpoint the elements that have the biggest impact on sales, as well as a regression or prediction stage, utilising a variety of machine learning techniques to develop developing book sales estimating methods. The developed models, which resemble basic decision trees, may forecast remarkably accurate sales predictions from pre-publication data. These may therefore be utilised as tools to help publishers make decisions, offering trustworthy assistance on the choice process for releasing a book. This is further demonstrated in the study by focusing on four sample examples of typical publishers with regards of their sales volume as well as the variety of books they produce. Good efficiency. However, the system's prediction results are not precise [6].

3. Proposed System

Currently, information mining techniques are widely used in the business world to extract data from databases. Information mining involves using techniques like Utility Pattern Mining, which considers item sets while using time-based tactics. Utility Pattern Mining is appropriate for large datasets that evaluate successfully in pattern detection. Hierarchical High Average Utility Pattern Mining is suggested for the e-commerce including the retail sector in this research article. Unbounded stream data may produce consistent results that need to be updated dependent on the passage of time. The database's unbounded stream information was subjected to operations using HAUPM. When information with a greater effect than more current information is subjected to a state-of-the-art algorithm. By encouraging customers to purchase items that are popular in the market, these data sets produce beneficial results for the retail sector. H-HAUPM is preferred over other approaches because it can generate itemsets accurately, doesn't take up a lot of space when in use, is scalable, and maintains consistency.

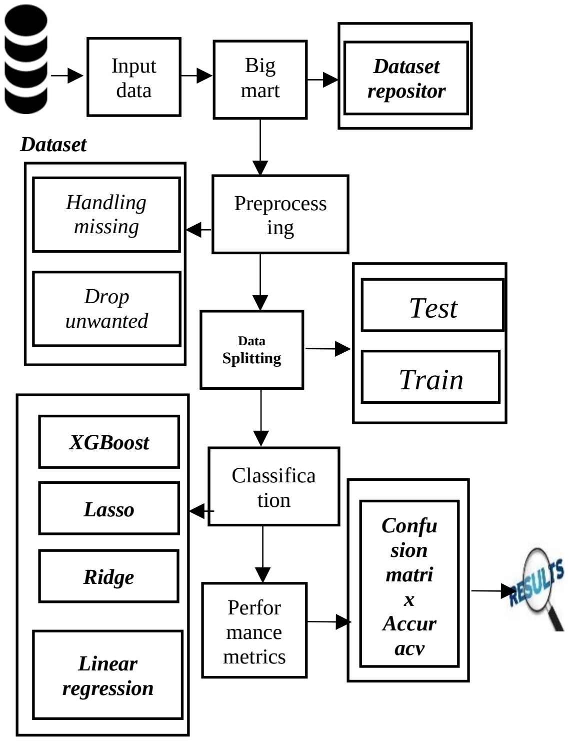


Fig 1: System Architecture

Up till now, a lot of effort has been accomplished that was actually intended for the discipline of transaction planning. A quick summary of the key research on big-mart agreements is given in this section. A few deals prediction standards have been developed using a variety of additional Measurable techniques, including regression, Auto-Regressive Integrated Moving Average, and Auto-Regressive Moving Average. A combination occasional quantum relapse method along with Auto-Regressive Integrated Moving Average are two key drawbacks in comparison to the given measurement method in A. S. Weigend et al. Deals envisioning is a complex topic which is impacted through internal as well as external factors. N. S. Arunraj recommended a typical approach to handling daily food bargains and discovered that the individual framework's exhibition was somewhat worse than the crossover framework's. The following are some of the proposed approach's major benefits:

- It is efficient for a sizable number of datasets.
- When compared to the current system, the experimental outcome is excellent.
- Low time commitment.
- Deliver precise forecast outcomes.

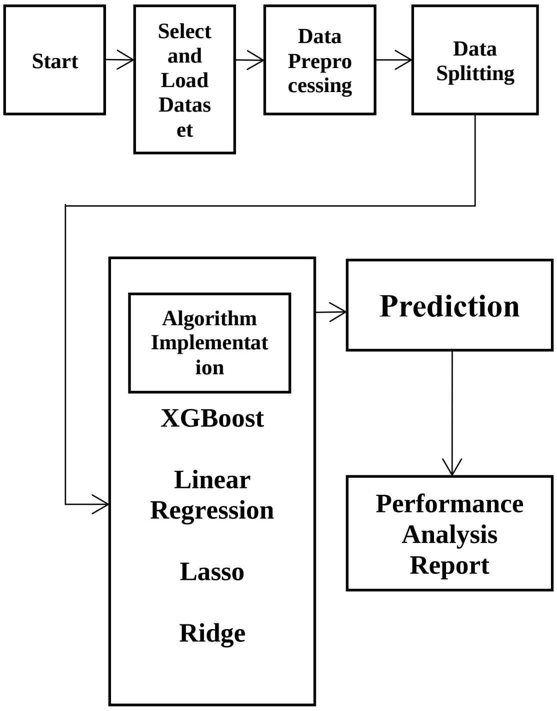


Fig 2: Flow Diagram

The following part provides an explanation of the many steps required in putting the proposed system, as depicted in Fig. 2, into operation:

1. **Data selection**
   The dataset for the website kaggle.com was compiled using input data that was gathered from the internet. In this study, there are two sets of data: a test data set with 5000 data points and a train data set with 8000 points. This technique used pandas to read data from our obtained dataset.
2. **Data pre-processing**
   Getting rid of extraneous data from the dataset is known as pre-processing. To create a dataset with a machine learning-friendly structure, pre-processing data transformation techniques are applied. The dataset is made more effective at this step by being cleaned of any inaccurate or superfluous data that can reduce the dataset's accuracy. Remove missing data: In this process, the null values, which including missing values and Nan values, are changed to 0. Data was cleared of any errors and missing values as well as duplicates. Encoding Data that may be categorised: Variables with a finite set of label values are regarded as categorical data. Most machine learning algorithms prefer numerical input and output variables.
3. **Data splitting**
   For machine learning to be successful, data must be available. Test data are required in addition to the training data in order to evaluate how efficiently the algorithm works, although in this case, the training and testing dataset are distinct. We must separate training and testing in our process into x train, y train, x test, and y test. The process of breaking accessible data into two pieces, often for cross-validator needs, referred to as data splitting. A portion of the data is applied to create a prediction model, while another portion is utilised to assess the effectiveness of the model.
4. **Regression Algorithms**
   We must incorporate machine learning algorithms like, in our method.

1) XGBoost Regression
2) Lasso Regression
3) Linear Regression
4) Ridge Regression

XGBoost Regression: "Extreme Gradient Boosting" is similar to the gradient boosting approach but is substantially more effective. It has a tree algorithm in addition to a linear model solver. Which makes "xgboost" far faster than how slope boosting is currently implemented. It provides a number of goal capacities, including rating, ranking, and relapse. "Xgboost" is suited for some rivalry since it has a very high prescient force but is often slow with organisation. Additionally, it is helpful for cross-approval and identifying important elements.

Lasso Regression: A type of shrinkage-based linear regression is lasso regression. This specific type of regression is well suited when models show when there is a need to automate important steps in the model selection process or when there is a significant level of multicollinearity, like as variable selection and parameter removal. Less absolute shrinkage and selection operator, also known as lasso or LASSO, In order to increase the predictability and comprehension of the final statistical model, the LASSO regression analysis approach, employed in statistics and machine learning, selects variables and regularises them.

Build a fragmented linear or non-linear pattern of data and variance using linear regression (outliers). When the marking is not linear, take a transformation into account. If so, only situations with a non-statistical foundation could removal of foreigners be advised.

Lasso Regression: A type of shrinkage-based linear regression is lasso regression. This specific type of regression is well suited when models show there is a lot of multicollinearity or you want to automate some steps in the model selection process, including variable selection with parameter removal. Less absolute shrinkage and selection operator, also known as lasso or LASSO, is a machine learning and statistical regression analysis approach that selects variables and regularises them to make the final statistical model more predictable and understandable.

Linear Regression: Create a fragmented linear or nonlinear data pattern with a variance (outliers). A transformation should be considered if there is no linear marking. If yes, only scenarios without a statistical basis would be considered removal of foreigners be advised. Use the residual plot (under the assumption of constant standard deviation) and the normal probability plot to connect the data to the least squares line and validate the model premises (for the normal probability assumption). If the presumptions seem to be unfounded, a change could be required. Create a regression line using the modified data and, if necessary, transform the information to least squares. Return to step 1 of the previous procedure if a modification has been completed. Instead, proceed to stage 5.

Ridge Regression: To assess any data that is multicollinear, utilise the model tuning method of ridge regression. The L2 regularisation process was done using this technique. The least squares are impartial and the variances are significant when multicollinearity problems occur, leading to a large gap between the predicted and actual values.

5. Result Comparison

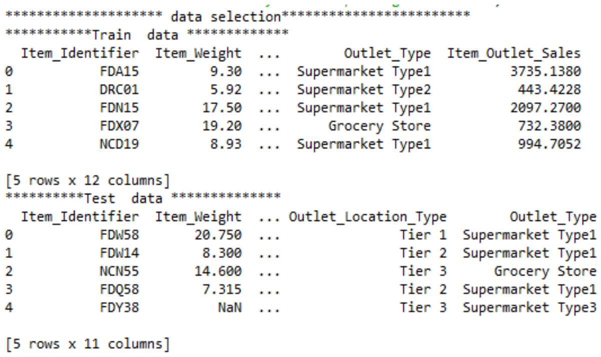
The total forecast will be used to create the Final Result. Some measures, such as, are used to gauge the efficacy of this advised course of action:

1) MAE
2) MSE
3) RMSE

To create the graph in this method, we compare the three results mentioned above.

4. Results

Currently, supermarket run-centers, Big Marts, gather sales data for each product to predict potential demand from consumers also adjust inventory management. The data store of the data warehouse is frequently mined to identify anomalies and broad patterns. With the use of different machine learning algorithms, merchants like Big Mart may utilise the collected data to predict future sales volume. Here we present a programme to use regression methodology for forecasting the sales concentrated on historical sales information. Using this approach, By determining polynomial regression, Ridge regression, plus Xgboost regression, it is possible to increase the precision of linear regression prediction. So, based on accuracy, MAE, and RMSE, it can be concluded that ridge and Xgboost regression offer better estimates than linear and polynomial regression techniques.

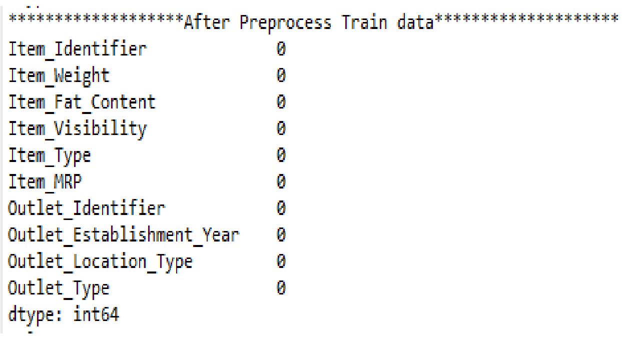


```
******************** data selection************************
***********Train  data *************
  Item_Identifier  Item_Weight  ...        Outlet_Type  Item_Outlet_Sales
0           FDA15         9.30  ...  Supermarket Type1          3735.1380
1           DRC01         5.92  ...  Supermarket Type2           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]
```

Fig 3: Data Selection



```
******************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
```

Fig 4: After Pre-process Train Data

```
**********************Comparision Graph*************************
```

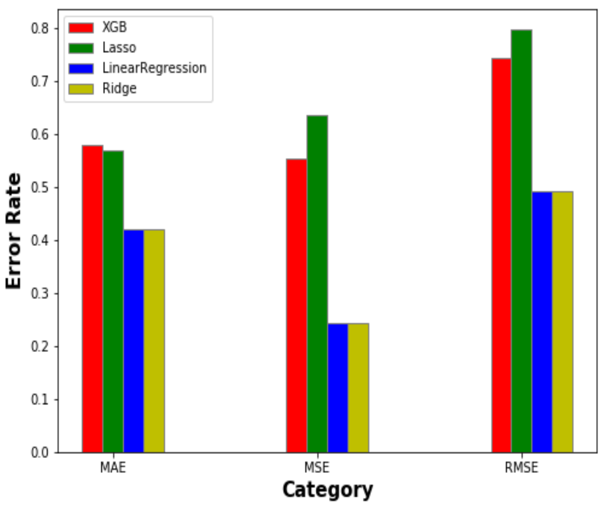


Fig 5: Result Comparison Graph

5. Conclusion

Here we present a programme to use regression methodology for forecasting the sales concentrated on historical sales information. With this method, the accuracy of linear regression prediction can be strengthened, and polynomial regression, Ridge regression, and Xgboost regression may be determined. Therefore, we could say that, in regards to accuracy, ridge and Xgboost regression provide better forecasts over linear and polynomial regression techniques, mean absolute error, and root mean square error.

6. Future Enhancement

In the future, creating a sales plan and predicting sales might be beneficial to avoid unforeseen cash flow and improve the management of production, staffing, and the necessary resources. Future studies may think about the ARIMA model as well that may present time series graphs.

Reference

[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.

[3] C. M. Wu, P. Patil and S. Gunaseelan, quot;Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data,quot; 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2018, pp. 16-20.

[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/.

[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".

[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. Krishna, A. V, A. Aich and C. Hegde, quot;Salesforecasting of Retail Stores using Machine Learning Techniques,quot; 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 160-166.

[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.