An Analysis of Machine Learning Algorithms to Predict Sales

Varshini S[1], Dr. D. Preethi [2]

[1]Student, Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, India

[2]Assistant Professor, Department of Computer Science, Mount Carmel College, Autonomous, Bengaluru, India

Abstract: ***Machine Learning plays a vital role in the area of sales and marketing. Therefore sales forecast is used to predict the future sales based on the past sales data. It also allows corporations to efficiently allocate resources including flow of cash, production, and make better informed business plan and decisions. In this research, we offer an efficient and accurate sales forecasting model based on several Machine Learning approaches that can handle enormous volumes of data, such as the Big Mart dataset, which contains a high number of customer data and individual data item attributes. ML models such as the XG Boost Regressor, Random Forest Regressor, ANN, and SVR are also used in a complete analysis.***

Keywords: sales; prediction; bigmart dataset; regression

1. Introduction

In today's digitally linked world, every store wants to know what their customers want ahead of time so that they don't run out of sale items during a particular season. To do this, we utilise sales forecasting, which is a technique for predicting future sales based on previous sales data. The ability to estimate sales of items is critical for organisations in the retail chain. Accurate forecasting can assist businesses in maximising their investment, lowering inventory costs, increasing sales and profitability, and avoiding hazards. [7]
In addition, sales forecasts are critical inputs for numerous managerial decisions, including price, store allocation, listing/delisting, ordering, and inventory management. Previous sales prediction research has always relied on a single prediction model. However, a single model cannot be the most effective for all types of products. [10] Predicting sales is also vital for offline firms. To estimate future sales and plan for the company's sales, predictions are often generated using statistical methodologies such as regression or a range of various models. [10]

The use of machine learning and artificial intelligence techniques to predict the sales has become an increasing trend, In case of sales predictionML has proved to be a boon. [1]Previous sales prediction research has always relied on a single prediction model. However, a single model cannot be the most effective for all types of products. [10]. Can unlock the secret of large data using machine learning, allowing retailers to better understand themselves and their competitors, adjust sales planning, and remain unstoppable. [12].

The purpose of this research is mainly to help the organization to predict their target and modify their strategy to improve their business productivity in future. In our proposed method we have first done data preprocessing to filter and remove out the outliers and we do complete data analysis using the dataset to know the factors that are affecting the sales and we have used five algorithm and done a comparative study to conclude the best performing algorithm based on RMSE, R2_score and MAPE and such metrics an also done the sales prediction of a particular product.

The rest of this research paper is structured as following, In the section II we have discussed few of the previous research papers, Section III we discuss about the algorithms used and also detailed description of our work. Section IV compares our method to other approaches and provides the experimental findings and analyses. Finally, we discuss some areas where our work could be enhanced in the future.

2. Literature Review

As sales is one of the most important factor of any retailer industry so the forecasting of that sales and analyzing it as been done by many researchers which is summarized:

An accurate sales prediction model using ML and feature engineering is conducted by XieDairu et al [1] also eXtreme Gradient Boosting (XGBoost) is used for utilizing these features to forecast the sales amount, various algorithms where compared and result concluded that xgboost performs better with RMSE 0.655 also it performed well with less computing time and for memory resources. A grid search optimization technique was suggested by GopalBehera et al [2] to optimizing the parameters and selecting the best tuning hyper parameter also used XG Boost (before tuning parameter) and (After tuning parameter) hence the model is performing better when the hyper parameters are tuned.

During the experiment inYiyangNiu et al [3]implemented detailed feature engineering procession such as memory compression, statistical features, temporal feature selection and it presents the feature engineering is more effective for training model. In SunithaCheriyan [4]Presents an investigation of how decisions are made based upon data and insights that are gained by visualising data, as well as the usage of data mining tools. The gradient boost method has been demonstrated to have the highest level of accuracy in predicting future transactions. Purvika Bajaj [5] sets out to predict the pattern of the company sales and the quantity of things to be sold based on a few key characteristics. To acquire an understanding of the data, analysis and exploration of the collected data were also performed. In Prabhat Sharma et al [6] prediction is done for the turnover of a automobile industry during the COVID - 19 era using ML. Jingru Wang [7] as suggested light GBM framework alone with xgboost framework to build a sales model. In Kunhui Lin et al [8], a new SVR that combines linear regression and time series prediction was used to forecast the time series. The novel model divides time series into linear and nonlinear components and uses SVR to forecast the nonlinear component.

With the help of these concepts, the merchant can quickly build up his retail shop and develop the business in the future with the help of these strategies [9]. Market basket analysis provides a common item set, i. e. association rules may simply express customer purchasing behaviour, allowing the merchant to quickly expand his retail store and expand his business in the future. PanjwaniMansi's [10] goal is to deliver acceptable results for predicting a firm's future sales or requests using approaches such as Clustering Models and sales prediction metrics. As a result, the potential of algorithmic approaches is calculated and used to future study.

3. Methodology

The main aim of this research project is to analyze and predict the future sales using various ML techniques to produce models which are comprehensive and reliable.

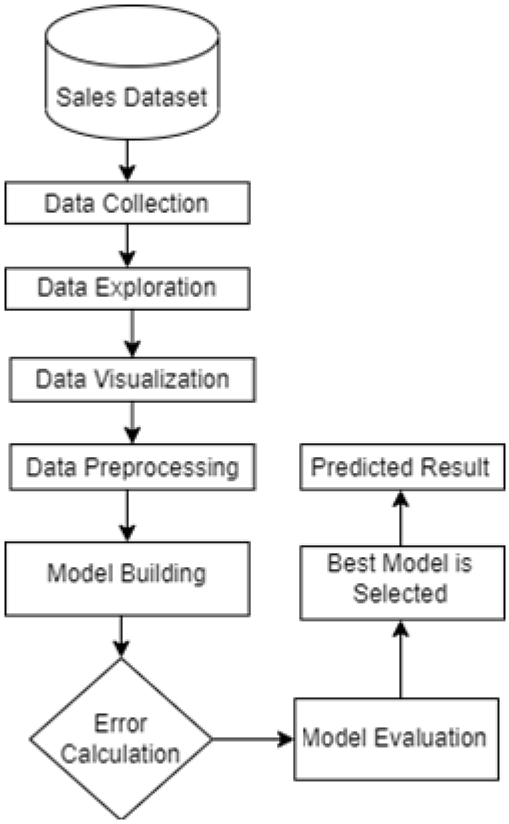

Figure 1: System Framework

a) Data collection and Preparation

The data plays an vital role for any model to predict the future sales in a retailers environment accurately. The dataset was collected from a website called kaggle, For our research we collected the BigMart sales dataset, it as12 features with 8523tupels, the data set has both input and output variables.

Table 1: Attributes Information

Item__Identifier	ID of the product
Item__Weight	Each products weight
Item_Fat_Content	It says whether the product as low fat or no
Item_Visibility	It refers to the percentage of a store's total display area that is devoted to all products.
Item_Type	It describes the food category to which the product belongs
Item__MRP	This is the product's maximum retail price.
Outlet_Identifier	Idof the stores
Outlet__Establishment_Year	Store establishment year
Outlet_Size	The store size
Outlet__Location_Type	The location of the store in relation to the size of the city
Outlet__Type	It tells whether the outlet is grocery or any supermarket type
Item_Outlet_Sales	This is the variable that has to be predicted. It contains the product's sales in the specific retail stores

b) Exploratory Analysis

An exploratory study was undertaken after data preparation to better understand the nature of our data [4], which allows decision makers to see analytics displayed visually, allowing them to grasp difficult ideas or uncover new patterns.

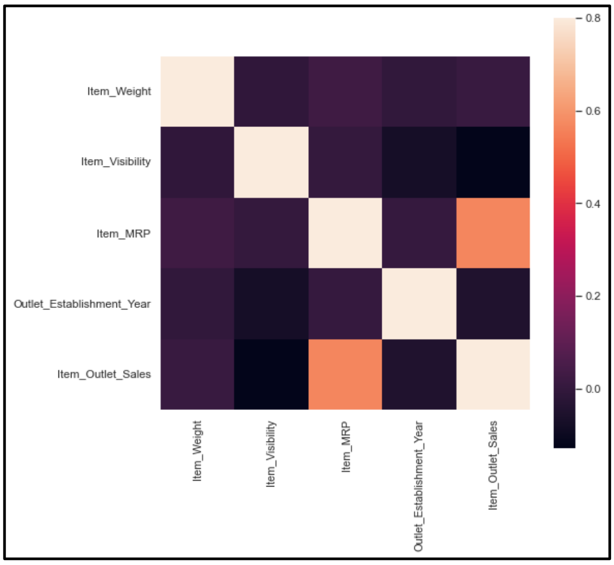

Figure 2: Correlations between numerical variables and the target variable

Fig.2. We can see that the Item Visibility feature has the lowest association with our goal variable based on the present numeric variables. As a result, the higher the price, the less apparent the product is in the store. Furthermore, Item MRP has the strongest positive association.

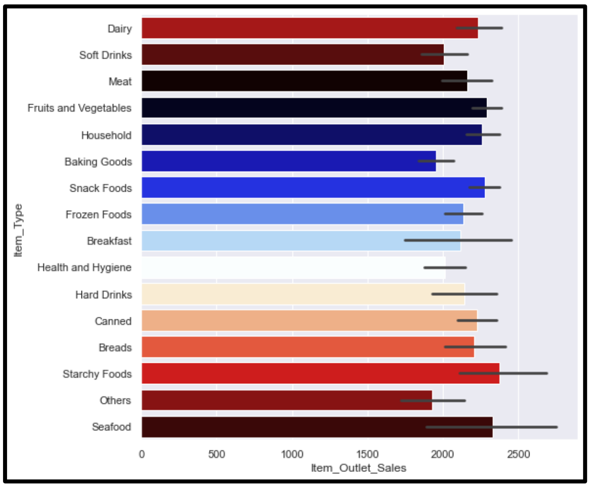

Figure 3: Bar plot for item types and item outlet sales

In Fig.3Fruits - Veggies and Snack Foods were available, but sales of Seafood and Starchy Foods appeared to be higher, suggesting that sales may be boosted by stocking products that are frequently purchased by customers.

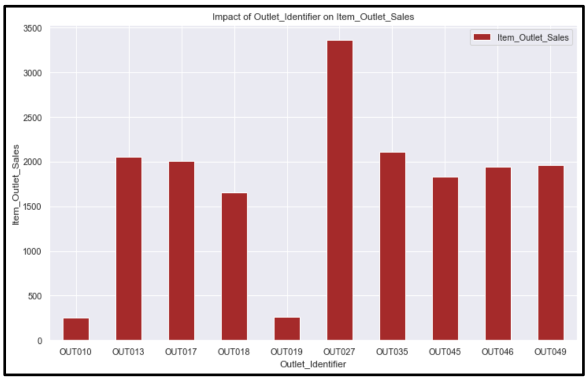

Figure 4: Shows the impact of outlet_identifieron item_outlet_sales

In Fig.4, From the graph, eight different stores have been identified. Two of the locations (OUT010 and OUT019) have lower sales than the others. OUT027 has the best sales, while the others are average.

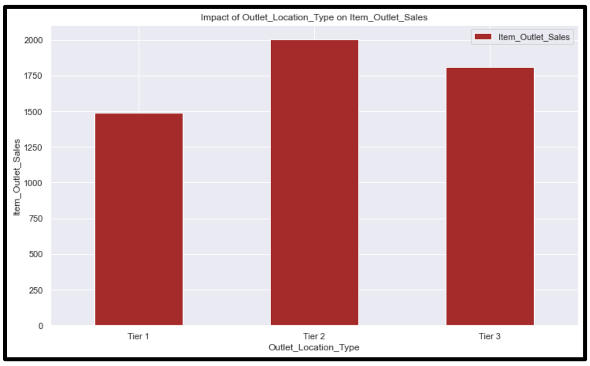

Figure 5: Shows the impact of Outlet_Location_Type on Item_Outlet_Sales

In Fig.5 we can see the highest sales is in Tier 2 then followed by Tier 3 and Tier 1.

c) Data Preprocessing

In data preprocessing we had to handle missing values, there are few missing values in item_weight and outlet_size columns. Which was replaced by mean and mode. Also there were unwanted data so we removed it and Encoding categorical data was done, [13]Extracted independent and dependent variables, Dependent variables is the final targets or output variables that must be evaluated and compared to one another. Creating two datasets: one for training and one for testing: Two different datasets are not imported for train and test to avoid over fitting. As a result, the splitting is done within a single dataset. The training dataset is the information that will be used to train the model. Test dataset are those that can be used to predict a test's outcome.

d) Model Building

Following the completion of the preceding phases, the dataset is now ready to be used to create a predictive model to forecast Big Mart's sales. We reviewed five machine learning algorithms that can be used to solve prediction problems in this research, including XGBoost, ANN, Random forest, SVR, and Cat Boost Regression.

1) XGBoost Regression

The XG Boost approach is based on decision trees and gradient boosting. The algorithm was built to be as efficient as possible in terms of computing time and memory resources. Boosting is a sequence method relay on the concept of the ensemble. [14] this involves a group of low - achieving students and improves accuracy. model variables are weighted at each time based on the impacts of the preceding instant. Rightly calculated findings are given less weight, whereas mistakenly calculated results are given more. This method is used by the XGBoost model to execute internal stepwise ridge regression, which chooses features and removes multiple regression. [15]

2) ANN

An Artificial Neural Network (ANN) is a computing system that replicates the behaviour of a neural network. When the number of input variables is enormous and the relationships between them are complicated, such as in big data scenarios, an ANN learns from these weights and provides the output. As a result, the results are purely dependent on the ANN's learning based on the real world data presented, rather than assumptions about variables. [13]

3) Random Forest

The random forest algorithm is a useful technique for predicting sales. It's simple to use and understand for predicting the outcomes of ML projects. Random forest are employed in sales prediction because they have decision tree - like hyper parameters. The tree model looks like decisions making aid. A random forest model is created for each learner using a random collection of rows and a few randomly determined parameters. [16] The final prognosis could be based on all of the predictions made by the pupils. In a regression problem, the final estimate might be the average of all previous projections.

4) SVR

Support Vector Regression (SVR) uses the same technique as SVM. Support Vector Machines (SVMs) which is well known for classification problems. These types of models are known as SVR It gives us the flexibility to define how much error is been found, Support Vector Machines (SVMs) are well - known in classification problems. These types of models are known as Support Vector Regression models (SVR). SVR allows us to determine how much error in our model is acceptable and matches the data with the proper line (or hyper plane in higher dimensions). SVR is a sophisticated technique that allows us to establish our error tolerance, both in terms of an acceptable error margin and tuning our tolerance for mistakes that fall outside of that acceptable error rate. [17]

e) Evaluation Metrics

i) **RMSE**

The root - mean square error (RMSE) is the standard deviation of the residuals (prediction error). The distance between a data point and the regression line is referred to as "residual. " The RMSE is a measure of how evenly these residuals are dispersed. You can see how densely the data is grouped around the ideal line, to put it another way. Mean squared error is commonly used to validate experimental results in climatology, prediction, and regression analysis.

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

ii) **R2_SCORE**

The coefficient of determination, often known as the R2 score, is used to evaluate the efficacy of a linear regression model. An input - independent variable can be used to forecast the amount of output - dependent attribute change. Based on the percentage of total deviation of the model's reported results, it is used to determine how effectively the model's observed outcomes are reproduced.

R2 = 1 –Ssres/Sstot

Here, SSres stands for the sum of squares of residual errors. The cumulative sum of all errors is SStot.

iii) **MAPE**

The Mean Absolute Percent Error (MAPE), also known as Mean Absolute Percent Deviation (MAPD), is a statistic that measures the accuracy of a prediction system. The average absolute percent error is widely used as a loss function in regression problems and model evaluations because it can be visualised reasonably simply in terms of relative error. The formula is widely used to express accuracy as a ratio:

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

4. Result and Discussion

For the given dataset, the various models mentioned before were employed to accomplish the prediction. The models were assessed using the RMSE, R - squared, and MAPE metrics.

Table 2: Comparison table

Algorithm	RMSE	R2_score	MAPE
XGboost	1203.069	0.53	0.61
ANN	1256.80	0.49	0.85
RandomForest	1171.429	0.55	0.54
SVR	1255.93	0.46	0.91

From the table above, Random Forest and XG Boost regressors have performed well compared to other algorithms.

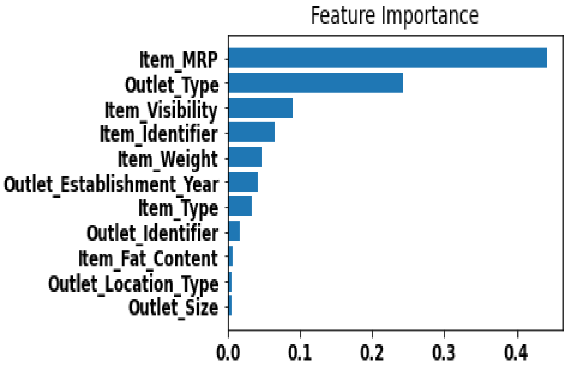

Figure 6: Feature importance

Ensemble learners frequently outperform traditional machine learning algorithms, and ensemble learners performed well in this experiment in terms of RMSE, R2, and MAPE. The most essential features for such algorithms are MRP and outlet_type.

5. Conclusion

In today's world, every retailer company and business wants to know in advance what their customers want, to avoid a shortage of seasonal items for sale. Daily forecasts from businesses and retailers are more accurate Demand for the sale of products to enable companies to obtain a higher return on investment. A company's profits are directly proportional to the exact sales forecast that Big Marts wants. A more accurate prediction method to avoid investment losses.

In our research we have used four algorithms out of which Random forest regressor performs well compared to the other three algorithms with RMSE as 1171.429 and R2_score as 0.55. And the Actual value of the first product is 3735.138 and the predicted result with this algorithm is 4035.473. Here by we Have predicted our future sales for each item, [18] which would be helpful for the retail companies to take informed decision and As part of future work we aim to enhance the result by using any other latest and efficient techniques.

References

[1] X. dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost, " 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp.480 - 483.

[2] G. Behera and N. Nain, "Grid Search Optimization (GSO) Based Future Sales Prediction for Big Mart, " 2019 15th International Conference on Signal - Image Technology & Internet - Based Systems (SITIS), 2019, pp.172 - 178

[3] Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering, " 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), 2020, pp.458 - 461

[4] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques, " 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2018, pp.53 - 58

[5] Purvika Bajaj, Renesa Ray, ShivaniShedge, ShravaniVidhate, Prof. Dr. NikhilkumarShardoor, ”Sales Prediction Using Machine Learning Algorithms”, 2020 International Research Journal of Engineering and Technology (IRJET)

[6] P. Sharma, S. Khater and V. Vashisht, "Sales Forecast of Manufacturing Companies using Machine Learning navigating the Pandemic like COVID - 19, " 2021 2nd International Conference on Computation, Automation and KnowledgeManagement (ICCAKM), 2021, pp.1 - 5.

[7] J. Wang, "A hybrid machine learning model for sales prediction, " 2020 International Conference on Intelligent Computing and Human - Computer Interaction (ICHCI), 2020, pp.363 - 366,

[8] K. Lin, Q. Lin, C. Zhou and J. Yao, "Time Series Prediction Based on Linear Regression and SVR, " Third International Conference on Natural Computation (ICNC 2007), 2007, pp.688 - 691.

[9] Mr. SohamPatangia, Mr. Kevin Shah, Mrs. MadhuraMokash, Ms. RachanaMohite, Mr. Gaurav Kolhe, Mrs. PrajaktaRokade, ” Sales Prediction of Market usingMachine Learning”, 2020, International Journal of Engineering and Technical Research

[10] PanjwaniMansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar and RupaliHande. “Sales Prediction System Using Machine Learning. ”, 2020, No.3243. EasyChair.

[11] Matthias Ulrich, Hermann Jahnke, Roland Langrock, Robert Pesch, Robin Senge, ” Classification - based model selection in retail demand forecasting”, 2022, International Journal of Forecasting Volume 38,, Pages 209 - 223

[12] Zhaoweijie, Hujiangmin, Chenliang” Forecast Rossmann Store Sales Base on Xgboost Model”, 2020, 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)

[13] Sharma, S. K., Chakraborti, S. &Jha, T. Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. InfSyst E - Bus Manage 17, 261–284 (2019).

[14] R. P and S. M, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms, " 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp.1416 - 1421.

[15] Jiyu Wu, Wei Wang, Yujun Cao, BiliangZhong, Zhenkun Chen, Zhenzhang Li, ” Product Marketing Prediction based on XGboost and LightGBM Algorithm Yunxin Liang”, 2019, Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition, Pages 150–153.

[16] Bohdan M, Pavlyshenko “Machine - Learning Models for Sales Time Series Forecasting” 2019, IEEE Second International Conference on Data Stream Mining & Processing (DSMP)

[17] Y. Feng and S. Wang, "A forecast for bicycle rental demand based on random forests and multiple linear regression, " 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, pp.101 - 105.

[18] Naveenraj R, VinayagaSundharam R, ”Prediction of Big Mart Sales Using Machine Learning”, 2021, International Research Journal of Modernization in Engineering Technology and Science