Journal of Innovative Computing and Emerging Technologies

Comparative Analysis of Regression Algosrithms used to Predict the Sales of Big Marts

M.Saad Bin Ilyas[1], Atif Ikram[2], Muhammad Aadil Butt[3] and Iqra Tariq[4]

Received: 06 December 2022; Accepted: 24 January 2023; Published: 08 February 2023

Abstract: Sales predictions or forecasting can help in analyzing the current and future sales trends of a big mart company. Based on the sales prediction or forecast, a retailer company can plan its production, marketing and promotional activities. Using several machine learning techniques, the obtained data may then be utilized to predict possible sales for retailers. This paper investigates that which machine learning regression algorithm best predicts big marts sales and which technique has the highest correlation coefficient value and the lowest values of mean absolute error (MAE), relative absolute error (RAE), root mean squared error (RMSE), and root relative squared error (RRSE). A comparative analysis of various machine learning regression algorithms such as SMO regression, simple linear regression, linear regression, additive regression, multi-layer perceptron, random forest, and M5P will be provided in this paper. After the experiments are completed, a comparison of various cross validations and splitting ratios for training and testing data will be given.

Keywords: Regression, Sales Prediction, Machine Learning

I. INTRODUCTION

Retail sales prediction is the process of forecasting future sales for a retail store or a chain of stores. It is a crucial aspect of retail management and helps businesses make informed decisions about inventory, marketing, and overall strategic planning. There are many factors that can influence sales predictions, such as consumer spending patterns, economic conditions, promotional activities, and competition. Retail sales prediction can be done using a variety of techniques, including statistical models, machine learning algorithms, and time series analysis.

Big marts are very useful to the common people as they provide a one stop solution for all their shopping needs. They are especially suitable for people who do not have the time or patience to go to different shops for different items. In addition, big marts usually offer good deals and discounts which can save us a lot of money [1].

Every item is monitored across all of Big Mart's retail locations, including its supermarkets and shopping centers, in order to forecast customer needs for the foreseeable future and optimize inventory management. Real time Stock Management of Big Mart is the most important aspect of the store as the customers are provided with the items of their demand. Big Mart keeps a check on the inventory of every item in their store. It is necessary to have a proper inventory management system for the store to be able to meet the demands of their customers. The inventory of the store is tracked through a software which is updated every day so as to make sure that the store is not running out of any item that is demanded by the customers [2].

A. *Sales Predictions or Forecasting*

Sales predictions or forecasting will assist the retailer company to achieve its desired sales target. Sales prediction or forecasting can also help in analyzing the customer buying behavior. This will be beneficial

[1,3,4]Department of Computer Science, The University of Chenab,Gujrat,Pakistan, Email: muhammad.ilyas@cs.uol.edu.pk
[2]Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

for the retailer company to understand the customer needs and wants. Based on the sales prediction or forecast, a retailer company can offer customized products and services to its customers. This will aid the retailer company to build long term relationship with its customers. This will support the retailer company to develop strategies to stay ahead of its competitors. Based on the sales prediction or forecast, a retailer company can identify the areas where it needs to improve its performance. This will facilitate the retailer company to develop action plan to improve its sales [3].

B. *Sales Prediction using Machine Learning*

Sales prediction using machine learning techniques is a process of using algorithms to parse data and make predictions about future sales. The predictions are based on historical data and trends. This can be done by analyzing past sales data and using that information to train a machine learning model. The model can then be used to make predictions about future sales. This type of analysis can be used to improve sales forecasting and planning [4].

There are a number of machine learning techniques that can be used to predict the sales. The more common ones include: decision trees, artificial neural networks (ANN), regression, and support vector machines (SVM) [5]. The specific machine learning algorithm that is most appropriate for a given problem based on a number of aspects, including; the number of features, the type of data being used, the number of training samples, and the desired level of accuracy [6].

In general, more complex machine learning algorithms (such as ANN and SVM) require more data in order to learn the underlying patterns and achieve good accuracy. Simple machine learning algorithms (such as linear regression) can often be used with less data. The number of features is also an important consideration. If there are too many features, the machine learning algorithm may have difficulty learning the underlying patterns. If there are too few features, the algorithm may not be able to accurately capture the complexities of the problem [7].

The specific machine learning algorithm used can also impact the accuracy of the predictions. In general, more complex algorithms (such as ANN and SVM) tend to be more precise than simple algorithms (like linear regression). However, more complex algorithms also tend to be more computationally expensive, and may require more data in order to achieve good accuracy. The desired level of accuracy is also an important consideration. If the aim is to form a highly accurate model, then a more complex algorithm may be necessary. However, if the aim is to make a model that is fast and easy to use, then a simpler algorithm may be sufficient. The above considerations are just some of the aspects that need to be considered when choosing a machine learning algorithm for prediction. There is no single best algorithm for all problems, and the best algorithm for a given problem will depend on the specific details of the problem [8].

C. *Regression Techniques in Machine Learning*

Regression is a statistical technique used to predict the value of a dependent variable (DV) based on the value of one or more independent variables (IV). The IV can be linear or nonlinear, and the DV can be continuous or categorical. There are many different types of regression, and the type that is used depends on the data type and the type of analysis that is being performed [9].

Linear regression is the most common type of regression. It is used to forecast the value of a continuous DV based on the value of one or more IV. Linear regression is used when the DV is linearly related to the IV. Nonlinear regression is used to forecast the value of a DV based on the value of one or more IV. Nonlinear regression is used when the DV is not linearly related to the IV [10].

Support vector regression is a method for regression that uses SVM. It is a supervised learning algorithm that can be used for both regression and classification. The support vector regression algorithm is very similar to the support vector machine algorithm for classification. The only difference is that in support vector regression, the aim is to find a function that best forecasts the value of a target variable, instead of a class label. The support vector regression algorithm works by mapping first the data points. Then, it finds a line or a hyperplane that best splits the data points. Finally, it projects the data points back into the original space and uses the line or hyperplane to make predictions. The support vector regression algorithm has a number of advantages. First, it is very robust to noise and outliers. Second, it is very efficient and can be used with large datasets. Finally, it can be used to learn non-linear functions.

The Random Forest algorithm is a supervised learning algorithm that can be used for both classification and regression. It is a bagging technique that joins several decision trees to make a more powerful model. The algorithm randomly selects a subset of the features to use for each tree, which results in a model that is less prone to overfitting.

K-NN Regression model takes the data point and predicts the DV of the unknown data point. It is a supervised learning algorithm. It means that it has both DV and IV. K-NN Regression model works on the principle of similarity. It means that it finds the data points that are similar to the unknown data points. The model then makes the prediction based on the similar data points. K-NN Regression model is an easy and simple in implementation. It is not computationally expensive. K-NN model of Regression can be implemented in both linear and non-linear relationships [11].

Related work is provided in section II, then methodology will be discussed in section III. In section IV and V experimentation and results will be provided and discussed whereas section V and VI are based on conclusion and references respectively.

II. Related Work

In order to estimate Wal-Marts sales, Jingru Wang provide an approach in this study [12] that combines a model based on the XG Boost framework with a model based on the Light GBM framework. Studies reveal that the XG Boost framework model and the Light GBM framework model outperform the conventional machine learning techniques and the combined model that results from combining the two models has the most predictive power. Majd Kharfan et al. improve the process of demand forecasting for recently introduced items in the fashion retailing industry, and suggests a data-driven approach based on machine learning methods [13]. Authors advised using the regression tree technique when dealing with complicated clusters that have various lifetime durations. They emphasized on utilizing linear regression and k-NN in their study for comparable clusters with larger sales volume and AUR, whereas random forests is recommended for clusters with mono-lifecycles. Zhou-zhou HE et al. integrates regression and consumer influence analysis; then provide EBMM, a technique for mining and predicting e-commerce business models. Several studies utilizing data from the Alibaba Group's online store have shown that EBMM is superior than modern techniques [14].

In this work, Ranjitha P and Spandana M examines the usefulness of several algorithms on the data based on profits and reviews. Authors infer that ridge and Xgboost regression provides superior predictions than linear and polynomial regression methodologies as compared with accuracy, MAE, and RMSE parameters. In the future, predicting sales and creating a sales plan may assist in preventing unpredicted cash flow and managing production, finance, and staffing requirements more efficiently [15]. I-Fei and Chi-Jie built prediction models based on clustering for sale of products related to computer using SOM, GHSOM, and K-means as clustering methods, whereas an ELM and SVR are two machine learning techniques. Comparing the model of GHSOM-ELM to the other forecasting models based on clustering, a single ELM, and a single SVR, the findings revealed that it performed the most effectively for predicting the sales of computer items [16].

This research by Chi-Jie Lu [17] builds sales forecasting models for computer devices, by combining MARS and SVR. The PSVR, PMARS, ARIMA, and GA-SVR prediction model results are contrasted with those of the anticipated MARS2VR sales forecasting system. The suggested model performed better than the opposing methods for the sale of computer products, according to experimental data, and produced improved forecast precision. Sunil K Punjabi et al. in this article [18] uses polynomial regression models to forecast sales of vehicles and used linear regression for sentiment analysis. Since the final model relies on the correctness of the sentiment analysis model, it must be very accurate.

The purpose of this study carried out by Gopalakrishnan T et al. [19] is to assess and forecast the sales of a major superstore in order to assist the retailer boost profitability, strengthen their brand and remain competitive in the marketplace while also increasing customer happiness. Linear Regression is a renowned algorithm in the area of Machine Learning, is the technique utilized for sales forecasting. The sales figures are from the years 2011 to 2013, and numbers for the year 2014 have been predicted. To calculate the prediction accuracy, real-time data for the year 2014 is also collected and compared to the expected data.

In this article Yi Yang et al. [20] suggests a e-SVR to forecast Chinese tobacco sales for a certain time period in the future. It also includes a representation of the sale trend in N. M. Saravana Kumar et al. [22] presents projected future sales in this research using data mining methods. Naive Bayes, Adaboost, Decision Tree with Naive Bayes, Particle Swarm Optimization, and Random Forest for prediction of sales are the algorithms that are compared. The particle swarm optimization approach using Naive Bayes has the minimum RMSE value. The best model is Naive Bayes with PSO, according to the algorithm's output. F. Jiménez et al. [23] choose features for online advertising sales prediction using the algorithm ENORA. In order to identify the best suitable model using an a-posteriori approach in the multi-objective setting, authors suggested a technique that uses feature selection for the assessment of model, decision making, and regression.

Meghana N et al. [24] proposed a framework that uses machine learning methods to forecast future sales using data from the previous year. Several machine learning models utilizing diverse algorithms, such as random forest regressor, linear regression, and XG booster techniques, were covered. These algorithms have been used to predict sales outcome. Carbonneau et al. [25] looked at the efficacy of sophisticated machine learning non-linear algorithms in predicting the inaccurate demand signals in the supply chain.

TABLE I. DESCRIPTION OF DATASET USED FOR SALES PREDICTION

Sr. No.	Attributes	Description
1	Item-weight	Weight of Product
2	Item-fat-content	Fat Product Contains
3	Item-visibility	Total percentage of are allocated to this store
4	Item-type category	Product belong to which category
5	Item-MRP	Product Rate or Price
6	Outlet-identifier	Unique ID of Store
7	Outlet-Establishment-year	Established year of store
8	Outlet-size	Size of Store
9	Outlet-location-type	Type of located cities
10	Outlet-type	Supermarket sort
11	Item-outlet-sales	Product sales in particular area

The advanced approaches did not significantly outperform more conventional procedures for the simulated data set, while generally producing better results. The more sophisticated data mining approaches (SVM and RNN) provide greater benefits for the actual foundries data, nevertheless. On the foundries test set, SVM and RNN provide the better results. Xiaodan Yu et al. forecasts newspaper and magazine sales with accuracy. The SVR method was applied to forecast sales. The results of the experiment demonstrated the effectiveness of SVR [26].

In this research work authors described analysis based on regression as a data mining approach and created a platform for using financial institution data, in particular. A forecasting model has been developed that use data mining to provide recurring predictions about stock market values. S Abdulsalam Sulaiman Olaniyi et al. applied regression analysis as a data mining approach in this work to characterize the patterns of prices of stock market and forecast the stock market future values [27].

This research predicts the sales of the various product categories. For the purpose of creating precise sales projections, models such as SVR, ANN linear regression, and nonlinear regression were implemented [28]. Nelson F. and M. A. R. Biswas used linear regression analysis for use in the residential sector, particularly with regard to single-family homes. For the understanding of the better usage of regression analysis for forecasting housing energy consumption, the energy signatures and conditional demand analysis was also explored [29].

In this study [30] Kumari Punam1et al. proposed a two-level statistical model that lowers the MAE value up to 39.17 percent. Authors suggested that the two-level statistical model produced better predictions for the huge mart dataset than the other single model predictive approaches.

III. METHODOLOGY

Since it is easy to estimate sales while working, there are various issues the sector confronts in the absence of predictive data. If the forecasts are correct, there is a chance for success. This article emphasizes data exploration and visualization in addition to preparation processes. Pre-processing activities including data exploration, and feature engineering. Then, for precise prediction analysis, the machine mining algorithm is used. The best algorithm for predictive analysis of big-mart sales is then determined after all algorithms have been evaluated with correlation coefficient, RMSE, MAE, RAE, and RRSE. Complete methodology diagram is shown in Figure 1.

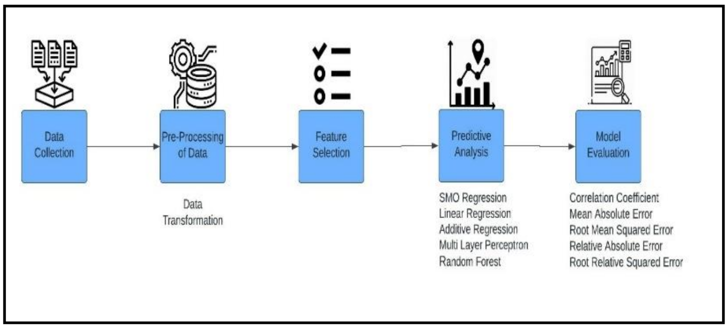
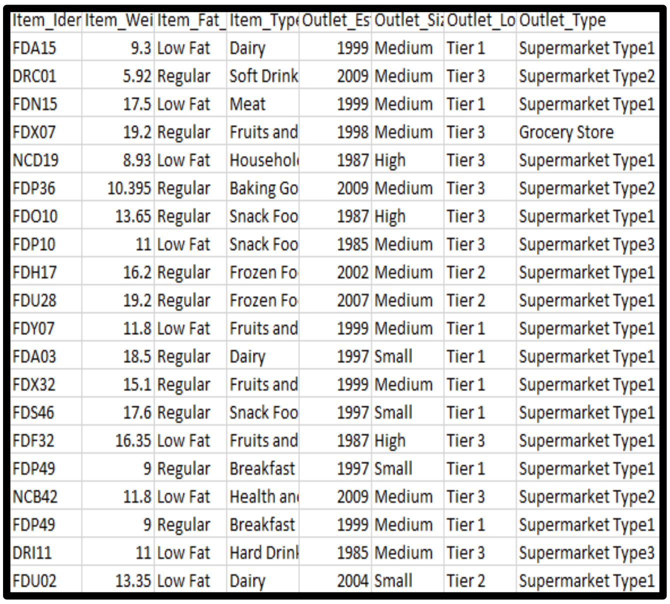


Figure 1: Methodology Diagram for Sales Prediction

A. Dataset Interpretation

Big Mart sales data were utilized as the dataset and downloaded from Kaggle [31]. There are 11 attributes in the dataset as shown in Table 1.

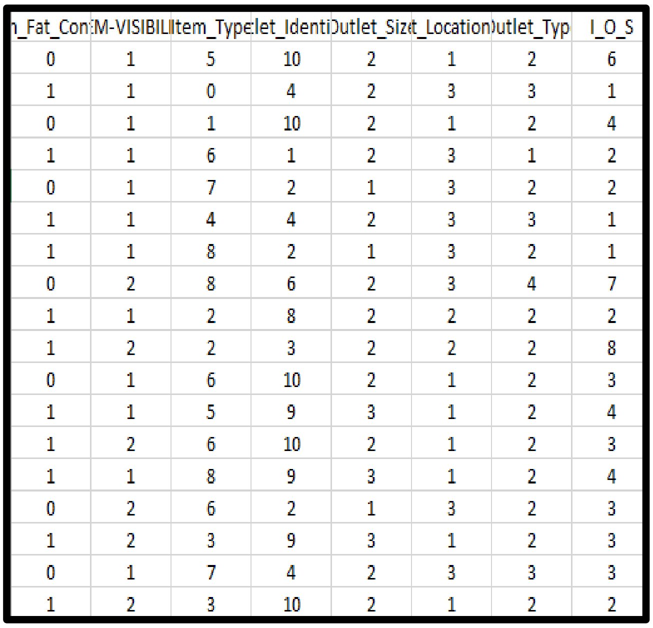
Multiple data from a few places are included in the dataset. Dataset contains numerical and categorical values shown in figure 2, categorical values are converted to numerical values in order to perform the regression analysis. For this purpose, we use ordinal encoding scheme.



Item_Ider	Item_Wei	Item_Fat_	Item_Typ	Outlet_Es	Outlet_Si:	Outlet_Lo	Outlet_Type
FDA15	9.3	Low Fat	Dairy	1999	Medium	Tier 1	Supermarket Type1
DRC01	5.92	Regular	Soft Drink	2009	Medium	Tier 3	Supermarket Type2
FDN15	17.5	Low Fat	Meat	1999	Medium	Tier 1	Supermarket Type1
FDX07	19.2	Regular	Fruits and	1998	Medium	Tier 3	Grocery Store
NCD19	8.93	Low Fat	Househol	1987	High	Tier 3	Supermarket Type1
FDP36	10.395	Regular	Baking Go	2009	Medium	Tier 3	Supermarket Type2
FDO10	13.65	Regular	Snack Foo	1987	High	Tier 3	Supermarket Type1
FDP10	11	Low Fat	Snack Foo	1985	Medium	Tier 3	Supermarket Type3
FDH17	16.2	Regular	Frozen Fo	2002	Medium	Tier 2	Supermarket Type1
FDU28	19.2	Regular	Frozen Fo	2007	Medium	Tier 2	Supermarket Type1
FDY07	11.8	Low Fat	Fruits and	1999	Medium	Tier 1	Supermarket Type1
FDA03	18.5	Regular	Dairy	1997	Small	Tier 1	Supermarket Type1
FDX32	15.1	Regular	Fruits and	1999	Medium	Tier 1	Supermarket Type1
FDS46	17.6	Regular	Snack Foo	1997	Small	Tier 1	Supermarket Type1
FDF32	16.35	Low Fat	Fruits and	1987	High	Tier 3	Supermarket Type1
FDP49	9	Regular	Breakfast	1997	Small	Tier 1	Supermarket Type1
NCB42	11.8	Low Fat	Health an	2009	Medium	Tier 3	Supermarket Type2
FDP49	9	Regular	Breakfast	1999	Medium	Tier 1	Supermarket Type1
DRI11	11	Low Fat	Hard Drin	1985	Medium	Tier 3	Supermarket Type3
FDU02	13.35	Low Fat	Dairy	2004	Small	Tier 2	Supermarket Type1

Figure 2: Dataset without Ordinal Encoding

The process of ordinal encoding is simple. We simply convert each value in the column to a numerical value. For example, if we have a column with three values, A, B, and C, we can convert them to 1, 2, and 3. After applying this method whole dataset attribute values are converted to numeric values as shown in Figure 3.



n_Fat_Con	M-VISIBILI	Item_Type	let_Identi	utlet_Siz	t_Location	utlet_Typ	I_O_S
0	1	5	10	2	1	2	6
1	1	0	4	2	3	3	1
0	1	1	10	2	1	2	4
1	1	6	1	2	3	1	2
0	1	7	2	1	3	2	2
1	1	4	4	2	3	3	1
1	1	8	2	1	3	2	1
0	2	8	6	2	3	4	7
1	1	2	8	2	2	2	2
1	2	2	3	2	2	2	8
0	1	6	10	2	1	2	3
1	1	5	9	3	1	2	4
1	2	6	10	2	1	2	3
1	1	8	9	3	1	2	4
0	2	6	2	1	3	2	3
1	2	3	9	3	1	2	3
0	1	7	4	2	3	3	3
1	2	3	10	2	1	2	2

Figure 3: Dataset with Ordinal Encoding

A. *Pre-processing*

This research included two steps of pre-processing: data exploration, pre-processing, and cleaning. Data pre-processing involves converting raw data into an articulate form. Incomplete and inaccurate data are often found in raw data. Finding the categorical values, missing values, and separating the data into a train set and test set. Whereas the Table 2 shows that which category has been replaced with what numerical value.

TABLE II. ATTRIBUTES NORMALIZATION VALUES

Attribute		**Attribute**		**Attribute**	
Item_Type		***Outlet_Type***		***Outlet_Identifier***	
Baking Goods	1	Grocery Store	1	OUT010	1
Breads	2	Supermarket Type2	2	OUT013	2
Breakfast	3	Supermarket Type3	3	OUT017	3
Canned	4	Supermarket Type4	4	OUT018	4
Dairy	5	***Outlet_Size***		OUT019	5
Frozen Foods	6	High	1	OUT027	6
Fruits and Vegetables	7	Medium	2	OUT035	7
Hard Drinks	8	Small	3	OUT045	8
Health and Hygiene	9	***Item_Fat_Content***		OUT046	9
Household	10	Low Fat	1	OUT049	10
Meat	11	Regular	2		
Starchy Foods	12	***Outlet_Location_Type***			

Seafood	13	Tier 1	1		
Snack Foods	14	Tier 2	2		
Soft Drinks	15	Tier 3	3		
Others	16				

B. *Feature Selection*

The selection of a subset of features to use in a machine learning model is called feature selection. Feature selection is a process that can automatically select those features in the data that contribute most to the interested prediction variable or output. There are three main benefits of performing feature selection before modelling machine learning data: It reduces overfitting that is less redundant data means less opportunity to make decisions based on noise, it improves accuracy because a model with fewer features is simpler and therefore, often more accurate and, it reduces training time as fewer features also mean that training a machine learning model will take less time.

In our model we use ReliefFAttributeEval as an attribute evaluator and ranker method for attribute selection. After the evaluation of all the training data, three attributes found with lower ranks shown in table 3 were reduced or removed from the dataset. Based on the result further experimentation will be performed on the remaining seven attributes.

TABLE III. Ranked Attributes by Attribute sellection

Rank	**Value**	**Attribute No.**	**Attribute Name**
1	0.00422	8	Outlet_Maturity
2	0.0029	7	Outlet_Identifier
3	0.00268	11	Outlet_Type
4	0.00246	9	Outlet_Size
5	0.00237	10	Outlet_Location_Type
6	0.00226	6	Item_MRP
7	0	3	Item_Fat_Content
8	-0.00354	4	Item-Visibiity
9	-0.005	5	Item_Type
10	-0.011	1	Item_Weight

C. *Regression models*

The most popular data mining technique is the decision tree, which takes values of attribute as input and outputs a result in the form of a Boolean. The result is one of the methods that uses a tree-like structure, in which each route begins by forming the root node towards the leaf node that characterizes the data order by breaking up the data until the result is reached in Boolean form [33]. One of the techniques used to estimate a single decision model from many decision trees is random forest. The method builds a forest of decision trees using the bagging principle. It provides the most precise output prediction since it uses results from several decision trees. The forecast for the regression model is based on the average value of all previous predictions [34].

SMOReg is a SMO-Based Regression method which implements the Support Vector Regression method for linear and non-linear models. The method is based on sequential minimal optimization (SMO) which is a method for solving complex optimization problems by decomposing them into a series of smaller subproblems. SMO is an effective process for resolving problems with a large number of variables and constraints. The method has been useful to a diversity of problems including regression, classification, and feature selection. The SMO-based regression method has several advantages over other methods for regression. The method is computationally efficient and can be applied to problems with variables having large numbers and constraints. The method is also flexible and can be used for both linear and non-linear models [35].

Simple Linear Regression is a statistical method that allows us to forecast a incessant outcome variable based on one or more forecaster variables. The modest form of linear regression is suitable for a straight line of the data. However, this approach is often not appropriate, particularly when there is non-linearity in the relationship between the predictor and outcome variable or when the variance of the outcome variable is not constant. Linear Regression assumes that the data is linear, which shows that there is a linear relationship between the DV and the IV [36].

Additive regression is a statistical approach to modelling the relationship between a DV and a linear combination of IV (predictors). Additive regression is a special case of multiple linear regression, where the IV are assumed to be linearly independent. Additive regression is commonly used in settings where there is a need to model the affiliation between a DV and a large number of IV. For example, in marketing research, additive regression may be used to model the relationship between sales of a product and a large number of marketing mix variables (e.g., price, advertising, promotion, etc.). Additive regression can also be used in time series analysis to develop the relationship between a DV and a linear combination of lagged values of the DV and values of IV [37].

A feedforward ANN model called a multilayer perceptron (MLP) translates sets of input data to a collection of suitable outputs. Each layer of nodes in an MLP is completely linked to the one below it in a directed graph. Each node, with the exception of the input nodes, is a neuron that employs a nonlinear activation function. Backpropagation is a supervised learning method used by MLP to train the network [38].

The M5P decision tree algorithm is a supervised learning algorithm. It is a non-parametric method used for classification and regression. The algorithm is used to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The M5P decision tree algorithm is a generalization of the ID3 algorithm. The M5P algorithm can handle both categorical and numerical data. It builds a tree by recursively splitting the data into subsets based on the values of the features. The M5P algorithm uses the information gain criterion to split the data. Information gain is a measure of how well a particular feature can predict the target variable. The M5P algorithm also prunes the tree to avoid overfitting. The M5P decision tree algorithm has a number of advantages. It is simple to understand and implement. It is also robust to data with missing values. [39].

All the machine leaning models discussed above will be used to perform the regression on the available dataset and obtained results will be analyzed in the results and discussion section of this paper.

IV. EXPERIMENTATION

All the models extend the preceding techniques by using 10-fold and 5-fold cross-validation and data splitting of 66% for training and 34% for testing, and also 80% for training and 20% for testing. In essence, cross-validation provides insight into a model's ability to generalize to new data. The focus of this work has been on modelling different data mining methods.

In our paper SMO regression, simple linear regression, linear regression, additive regression, multi-layer perceptron, random forest, and M5Pare the models used to predict the sales of big marts. Performance evaluation of these models based on correlation coefficient, MAE, RMSE, RAE, and RRSE has been provided in this section [40].

Regression analysis uses the correlation coefficient to gauge how strongly the IV and DV are related linearly. The absolute amount of the difference between the predicted value and the actual value is what is meant by one of the most often used metrics in regression analysis, known as MAE.

The standard deviation of the residuals is measured by the regression model standard error, or RMSE. To put it another way, it is a gauge for how much the forecasts depart from the observed values. The fit is better the lower the RMSE.

The absolute error for each data point is first determined in a regression analysis, and the average of these absolute errors is then computed. The difference between the projected value and the actual value is the absolute error. The squared difference between a model's projected value and the actual value for a given data point is added together to get the RRSE, which is then divided by the total number of data points.

V. RESULTS AND DISCUSSIONS

In table 4 results of all the models are compared with each other by applying the big marts sales dataset by using 66% of data for training and remaining 34% data for testing. The result shows that M5P is the better model among the others in terms of errors calculation and has better correlation coefficient value.

TABLE IV. SPLIT 66.0% TRAIN, REMAINDER TEST

Performance Measures/ Algorithm	CC	MAE	RMSE	RAE	RRSE
SMO Regression	0.7018	1.4320	1.9416	0.6885	0.7273
Simple Linear Regression	0.5607	1.6370	2.2120	0.7871	0.8286
Linear Regression	0.7023	1.4328	1.9053	0.6889	0.7137
Additive Regression	0.7321	1.3761	1.8192	0.6616	0.6814
Multi-Layer Perceptron	0.7608	1.5432	2.1264	0.7421	0.7965
Random Forest	0.6627	1.4313	2.0505	0.6882	0.7681
M5P	**0.7712**	**1.2047**	1.7047	**0.5792**	**0.6385**

Figure 4 shows the graphical representation of the results obtained from the experiments performed on data using test mode; split 66% train and 34% for testing.

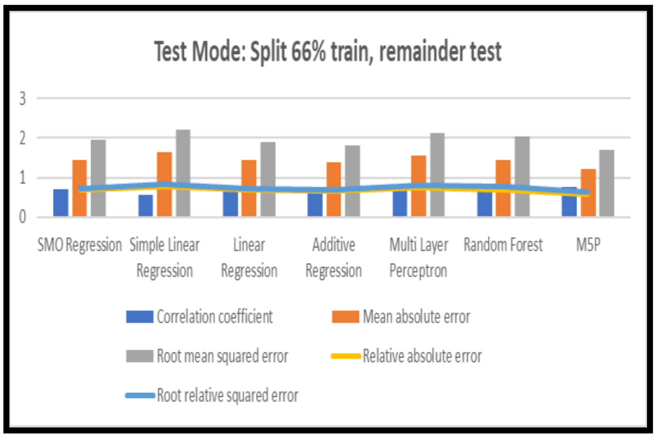


Figure 4. Split 66% train, remainder test

Table 5 depicts the results after execution of the same data set but with different split ratio for training and testing i.e., 80% of data being used for training and 20% for testing. M5P has the better results again in this case and has better value of correlation coefficient. Whereas the graph of the stated results is shown in figure 5.

TABLE V. SPLIT 80.0% TRAIN, REMAINDER TEST

Performance Measures/ Algorithm	CC	MAE	RMSE	RAE	RRSE
SMO Regression	0.7055	1.4180	1.9278	0.6813	0.7208
Simple Linear Regression	0.5754	1.6115	2.1904	0.7742	0.8191
Linear Regression	0.7058	1.4260	1.8976	0.6851	0.7095
Additive Regression	0.7296	1.3808	1.8291	0.6634	0.6839
Multi-Layer Perceptron	0.7534	1.2703	1.8238	0.6103	0.6819

Random Forest	0.6630	1.4201	2.0558	0.6823	0.7687
M5P	**0.7690**	**1.2021**	**1.7112**	**0.5775**	**0.6398**

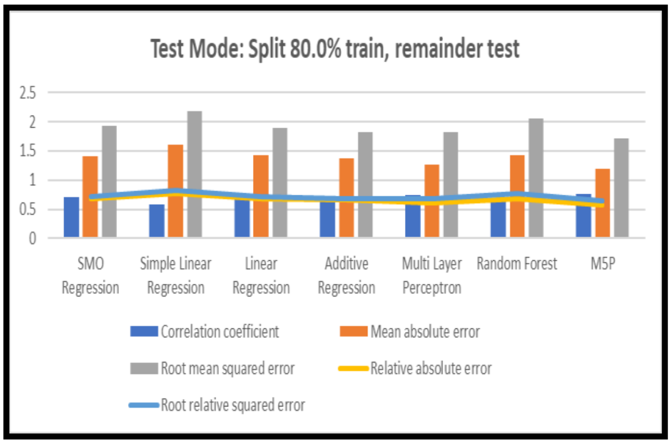


Figure 5. Split 80.0% train, remainder test

Cross validation of 10-folds is used to evaluate the models and the results shown in table 6 describes that M5P has the best results among others in terms of correlation coefficient and errors. These results are also represented in the graphical form in figure 6.

TABLE VI. 10-FOLD CROSS-VALIDATION

Performance Measures/ Algorithm	CC	MAE	RMSE	RAE	RRSE
SMO Regression	0.7011	1.3942	1.8697	0.6777	0.7202
Simple Linear Regression	0.5632	1.5936	2.1447	0.7746	0.8262
Linear Regression	0.7012	1.4047	1.8505	0.6828	0.7128
Additive Regression	0.7268	1.3608	1.7856	0.6614	0.6878
Multi-Layer Perceptron	0.6677	1.5823	2.0317	0.7691	0.7826
Random Forest	0.6697	1.3921	1.9982	0.6766	0.7697
M5P	**0.7665**	**1.1794**	**1.6670**	**0.5733**	**0.6421**

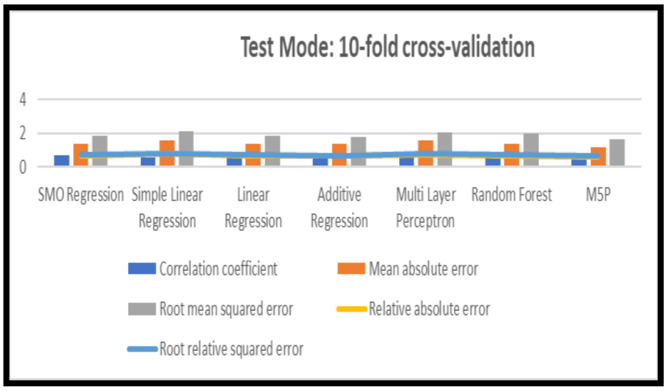


Figure 5. 10- fold cross-validation

In figure 7 comparison of different algorithms using 5-fold cross validation has been observed after evaluating the data set, M5P performs well as shown in table 7 among the other algorithms.

TABLE VII. 5-FOLD CROSS-VALIDATION

Performance Measures/ Algorithm	CC	MAE	RMSE	RAE	RRSE
SMO Regression	0.7011	1.3937	1.8695	0.6775	0.7202
Simple Linear Regression	0.5633	1.5935	2.1444	0.7746	0.8261
Linear Regression	0.7013	1.4045	1.8502	0.6827	0.7128
Additive Regression	0.7276	1.3587	1.7839	0.6604	0.6872
Multi-Layer Perceptron	0.7288	1.5812	2.0253	0.7686	0.7802
Random Forest	0.6711	1.3894	1.9925	0.6754	0.7676
M5P	**0.7675**	**1.1776**	**1.6636**	**0.5724**	**0.6409**

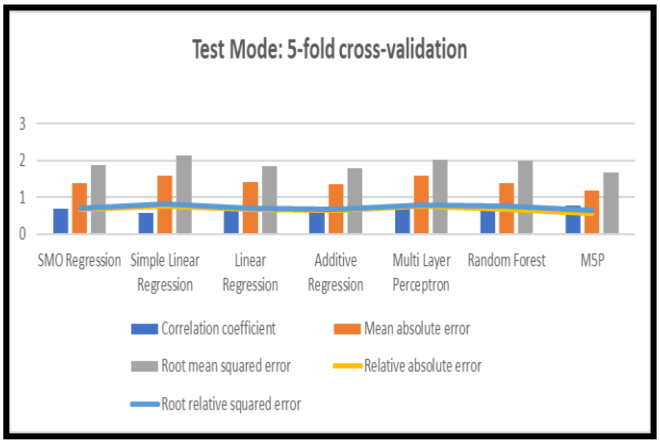


Figure . 5- fold cross-validation

VI. CONCLUSION

To prevent a shortage of sale products throughout any season, every shop in the modern, digitally linked world wants to know the client needs in advance. Day to day stores or firms are making more precise predictions, many scientists are engaged in this effort to get an accurate sales forecast. A company's revenue is directly proportionate to its profit. The Big marts are demanding the precise sales forecasts and more precise forecast methods to prevent losses on their commitment. In this paper, we have compared different models for predicting future sales in the Big Mart dataset of a certain Big Mart location or shop. Experimental investigation discovered that the M5P performs well among the other algorithms with minimum MAE, RMSE, RAE, RRSE and maximum correlation coefficient value.

VII. FUTURE WORK

There are several areas where future work could be done to improve sales prediction for a retail store such as Big Mart. Some of the future directions are Incorporating more data sources: Currently, sales prediction models may be based on a limited set of data, such as sales history and demographic information. In the future, incorporating data from other sources such as weather, social media sentiment, and website traffic could provide a more complete picture and improve the accuracy of predictions. Incorporating external events and promotions: Sales predictions may also be improved by considering external events and promotions that can affect the demand for products. This can be achieved by including information about holidays, local events, and promotions in the prediction model. Enriching with store-specific data: Another area of improvement could be to enrich the data with store specific information. For example, store layout, store-specific promotions or discounts, store-specific sales history could be used to improve the prediction.

REFERENCES

[1] Kumar, Navneet, and Suraj Choudhari. "BigMart Sale Prediction using Machine Learning." *International Journal of Innovative Science and Research Technology, Volume 6, Issue 9,* (2021).

[2] Silva, Ana Lúcia, and Margarida GMS Cardoso. "Predicting supermarket sales: The use of regression trees." *Journal of Targeting, Measurement and Analysis for Marketing* 13.3 (2005): 239-249.

[3] Behera, Gopal, and Neeta Nain. "Grid search optimization (GSO) based future sales prediction for big mart." *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2019.

[4] Khan, Muhammad Adnan, et al. "Effective demand forecasting model using business intelligence empowered with machine learning." *IEEE Access* 8 (2020): 116013-116023.

[5] Beheshti-Kashi, Samaneh, et al. "A survey on retail sales forecasting and prediction in fashion markets." *Systems Science & Control Engineering* 3.1 (2015): 154-161.

[6] Pavlyshenko, Bohdan M. "Machine-learning models for sales time series forecasting." *Data* 4.1 (2019): 15.

[7] Chu, Ching-Wu, and Guoqiang Peter Zhang. "A comparative study of linear and nonlinear models for aggregate retail sales forecasting." *International Journal of production economics* 86.3 (2003): 217-231.

[8] Hirt, Robin, et al. "How to learn from others: transfer machine learning with additive regression models to improve sales forecasting." *2020 IEEE 22nd Conference on Business Informatics (CBI)*. Vol. 1. IEEE, 2020.

[9] Kavitha, S., S. Varuna, and R. Ramya. "A comparative analysis on linear regression and support vector regression." *2016 online international conference on green engineering and technologies (IC-GET)*. IEEE, 2016.

[10] Catal, Cagatay, et al. "Benchmarking of regression algorithms and time series analysis techniques for sales forecasting." *Balkan Journal of Electrical and Computer Engineering* 7.1 (2019): 20-26.

[11] Raizada, Stuti, and Jatinder kumar R. Saini. "Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting." *International Journal of Advanced Computer Science and Applications* 12.11 (2021).

[12] Wang, Jingru. "A hybrid machine learning model for sales prediction." 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI). IEEE, (2020).

[13] Kharfan, Majd, Vicky Wing Kei Chan, and Tugba Firdolas Efendigil. "A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches." Annals of Operations Research 303.1 (2021): 159-174.

[14] He, Zhou-zhou, et al. "E-commerce business model mining and prediction." Frontiers of Information Technology & Electronic Engineering 16.9 (2015): 707-719.

[15] Ranjitha, P., and M. Spandana. "Predictive analysis for big mart sales using machine learning algorithms." 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021.

[16] Chen, I-Fei, and Chi-Jie Lu. "Sales forecasting by combining clustering and machine-learning techniques for computer retailing." Neural Computing and Applications 28.9 (2017): 2633-2647.

[17] Lu, Chi-Jie. "Sales forecasting of computer products based on variable selection scheme and support vector regression." *Neurocomputing* 128 (2014): 491-499.

[18] Punjabi, Sunil K., et al. "Sales prediction using online sentiment with regression model." *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020.

[19] Gopalakrishnan, T., Ritesh Choudhary, and Sarada Prasad. "Prediction of Sales Value in online shopping using Linear Regression." 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018.

[20] Yang, Yi, et al. "SVR mathematical model and methods for sale prediction." Journal of Systems Engineering and Electronics 18.4 (2007): 769-773.

[21] Pavlyshenko, Bohdan M. "Using bayesian regression for stacking time series predictive models." 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP). IEEE, 2020.

[22] Kumar, NM Saravana, et al. "A Study On the Forecasting Bigmart Sales Using Optimized Data Mining Techniques." Science in Information Technology Letters 1.2 (2020): 52-59.

[23] Jiménez, Fernando, et al. "Multi-objective evolutionary feature selection for online sales forecasting." Neurocomputing 234 (2017): 75-92.

[24] Meghana, N., et al. "Improvizing big market sales prediction." Journal of Xi'an University of Architecture & Technology Volume XII, Issue IV, (2020): 4307-4313.

[25] Carbonneau, Real, Kevin Laframboise, and Rustam Vahidov. "Application of machine learning techniques for supply chain demand forecasting." European Journal of Operational Research 184.3 (2008): 1140-1154.

[26] Yu, Xiaodan, Zhiquan Qi, and Yuanmeng Zhao. "Support vector regression for newspaper/magazine sales forecasting." Procedia Computer Science 17 (2013): 1055-1062.

[27] Olaniyi, S. Abdulsalam Sulaiman, Kayode S. Adewole, and R. G. Jimoh. "Stock trend prediction using regression analysis–a data mining approach." ARPN Journal of Systems and Software 1.4 (2011): 154-157.

[28] Aktepe, Adnan, Emre Yanık, and Süleyman Ersöz. "Demand forecasting application with regression and artificial intelligence methods in a construction machinery company." Journal of Intelligent Manufacturing 32.6 (2021): 1587-1604.

[29] Fumo, Nelson, and MA Rafe Biswas. "Regression analysis for prediction of residential energy consumption." Renewable and sustainable energy reviews 47 (2015): 332-343.

[30] Punam, Kumari, Rajendra Pamula, and Praphula Kumar Jain. "A two-level statistical model for big mart sales prediction." 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE, 2018.

[31] https://www.kaggle.com/datasets/devashish0507/big-mart-sales-prediction accessed on 1-Jul-2022

[32] Potdar, Kedar, Taher S. Pardawala, and Chinmay D. Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers." International journal of computer applications 175, no. 4 (2017): 7-9.

[33] Gupta, Bhumika, Aditya Rawat, Akshay Jain, Arpit Arora, and Naresh Dhami. "Analysis of various decision tree algorithms for classification in data mining." International Journal of Computer Applications 163, no. 8 (2017): 15-19.

[34] Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. J. O. G. R. Chica-Rivas. "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines." Ore Geology Reviews 71 (2015): 804-818.

[35] He, Liang, Gierad Laput, Eric Brockmeyer, and Jon E. Froehlich. "SqueezaPulse: Adding interactive input to fabricated objects using corrugated tubes and air pulses." In Proceedings of the eleventh international conference on tangible, embedded, and embodied interaction, pp. 341-350. 2017.

[36] Marill, Keith A. "Advanced statistics: linear regression, part I: simple linear regression." Academic emergency medicine 11, no. 1 (2004): 87-93.

[37] Stone, Charles J. "Additive regression and other nonparametric models." The annals of Statistics 13, no. 2 (1985): 689-705.

[38] Tang, Jiexiong, Chenwei Deng, and Guang-Bin Huang. "Extreme learning machine for multilayer perceptron." IEEE transactions on neural networks and learning systems 27, no. 4 (2015): 809-821.

[39] Zhan, Chengjun, Albert Gan, and Mohammed Hadi. "Prediction of lane clearance time of freeway incidents using the M5P tree algorithm." IEEE Transactions on Intelligent Transportation Systems 12.4 (2011): 1549-1557.

[40] Kavitha, S., S. Varuna, and R. Ramya. "A comparative analysis on linear regression and support vector regression." In 2016 online international conference on green engineering and technologies (IC-GET), pp. 1-5. IEEE, 2016.