E-COMMERCE SALES PREDICTION

[1]Abhijeet Patre, [2]Rumit Jain, [3]Pranita Darveshi, [4]Rovina D'Britto, [5]Sanketi Raut

[1,2,3]Student, [4,5]Assistant Professor
Department of Information Technology,
Universal College of Engineering, Vasai, India.

Abstract: A sales prediction is an essential tool for managing a business of any size. With the help of prediction, the production of the particular product will be minimal or in required amount. It will be helpful for the seller to analyze the sales and give the proper or approximate order to the manufacturers for the production purpose. Prediction also helps the sellers to analyze which brand or product is liked by the customers and according to that the seller can pass the order of the production. This will help in proper utilization of the resources and simply avoid the wastage of economy, resources and time. Ecommerce sales prediction predicts the sales of the particular product on the basis of past history and multiple factors such as trend, seasonal, state and festive using Multiple Linear Regression (MLR). We have also compared different algorithms like Random Forest, Naïve Bayes, SVR, KNN and Multiple Linear Regression and found out that MLR gives the accurate results.

***Keywords* - E-commerce, sales prediction, machine learning, regression.**

I. INTRODUCTION

The world is full of data. Data is everywhere around us. It's all stored in the form of data, right from managing monthly budgets, storing information on mobile phones, buying items from the shops. Humans got to deal with a lot of data in their everyday lives. Such data could be as small as managing the monthly budgets or big ones as data from a multinational corporation [2]. Online shopping complexes have a lot of data to work upon. Handling inventory, managing manufacturer orders, handling inventory prices, handling data on products, handling their sales, and much more. It's a huge task to work on such a big dataset [2]. A lot of work needs to be done on the dataset to analyze and predict it. All of this work is done to test the current sales position and to assess the expected future sales, so that the problems such as over production, improper utilization of resources can be solved [4]. The purpose of e-commerce website is to help customers narrow down their broad ideas and enable them to finalize the products they want to purchase. It intends to solve the problems of the sellers by getting the approximate sales of the products they need using machine learning algorithm. It has seen many e-commerce business owners forecast the profits of their business approximately, simply because experience counts for more than any data or figures. Previously the sales forecast was not accurate due to less factors involved but, in this paper, different factors are taken into consideration to increase the accuracy. In this paper, the methodology for predicting sales of the Ecommerce websites for next year's using concepts of machine learning is described. According to the characteristics of the data, we use the method of multiple regression analysis, random forest, KNN, Naive-Bayes and SVR (support vector regression) methods to forecast the E-commerce sales. The result showed that the best performance among the methods used was multiple linear regression than other approaches.

II. LITERATURE REVIEW

The following research is persuasive on various applications of Predictive analysis in the machine learning domain:
The analysis of the sales of a big superstore, and to predict their future sales for helping them to increase their profits and make their brand even better and competitive as per the market trends by generating customer satisfaction as well using Linear Regression algorithm [1]. More factors should have been considered to increase the accuracy.

This paper [3] focuses on the field of prediction models to develop an accurate and efficient algorithm to analyze the customer spending in the past and output the future spending of the customers with same features. In this, different machine learning techniques such as regression and neural network to develop a prediction model are implemented and a comparison is done based on their performance and accuracy of prediction. To improve the results, a dataset with sufficient features and increase in quantity must be obtained.

This [4] paper provides the processing of the visualization of transaction data and prediction model using regression methods on flight ticket sales on travel agents. The real data is used to predict the profit with predictive analytics. The regression methods that used linear regression, multilayer perception (MLP), and m percentage model (M5P). The Technique used is limited to make predictive value. It would be more interesting to use predictions by normalizing data and forming classes into other techniques such as SVM and ARIMA.

III. METHODOLOGY

The proposed approach was sorted out into three phases, first is information assortment, which incorporates gathering information and changing it into handled information. At that point, it incorporates demonstrating the information for expectations utilizing Machine learning strategies (concentrating on straight relapse calculation). Also, at last approving and execution of our outcomes utilizing exactness and precision methods. At long last extraordinary Machine learning algorithms are likewise contrasted to examined the RMSE (Root Mean Squared Error) esteem. The information is first gotten to by the client from different sources. For the most part, the information isn't accessible in prepared structure so information change is utilized to purify the dataset. At that point, utilizing suitable modes, demonstrating is done on the dataset and results are determined as needs be. What's more, at long last, the information is approved utilizing exactness and precision systems and conclusive outcome are executed. A point by point conversation on every one of the accompanying advances will be finished further.

3.1 Information Assortment

Information is gathered from different sources and sorted out in a single record. Next, the information purging procedure is applied. Information purging is the way toward identifying and adjusting off base or outdated records from a record set, table, or database and alludes to distinguishing inadequate, off base, off base or superfluous pieces of the information and afterward supplanting, adjusting, or erasing the messy or coarse information. Information purging might be performed intuitively with information wrangling devices, or as cluster preparing through scripting. Subsequent to purging, an informational collection ought to be steady with other comparable informational indexes in the framework. The irregularities recognized or on the other hand expelled may have been initially brought about by client section mistakes, by debasement in transmission or capacity, or by various information word reference meanings of comparable substances in various stores. Likewise, different clamors inside the information is too evacuated. Alongside that, the information is classified appropriately and every single clear space or unessential data inside information is likewise evacuated so that examination could be performed effectively on the information.

3.2 Model Building using Multiple Regression

The handled information is utilized for prescient displaying so that proper outcomes can be produced from it. This prescient displaying is finished utilizing a strategy called Machine Learning. It is characterized as a PC's capacity to learn without being unequivocally modified". At its generally essential, Machine Learning utilizes modified calculations that get furthermore, break down information to foresee yield esteems inside a satisfactory range. As new information is taken care of to these calculations, they learn and enhance their tasks to improve execution, creating 'insight' after some time. There are four types of machine learning algorithms: supervised, semi supervised, unsupervised and reinforcement. Out of which, we will look only on supervised learning in this paper. Supervised learning- In supervised learning, the machine is taught with the help of examples. The user provides the ML algorithm with a dataset that includes desired inputs and outputs, and the algorithm finds a method to determine how to arrive at those results.

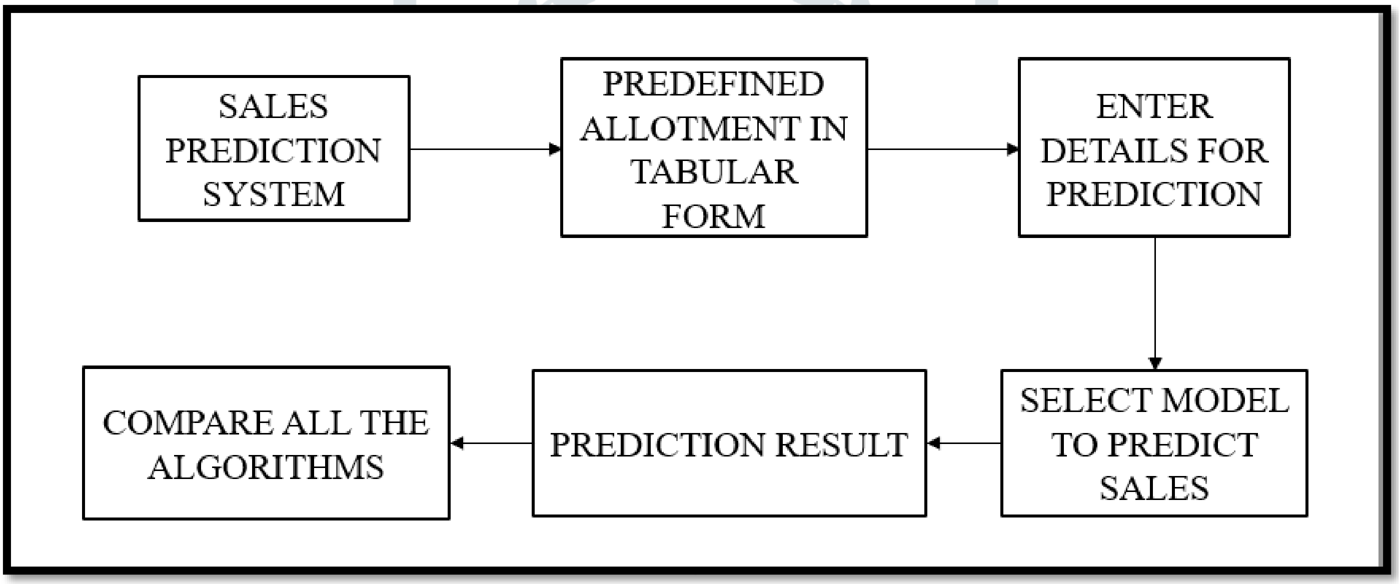

SALES PREDICTION SYSTEM
PREDEFINED ALLOTMENT IN TABULAR FORM
ENTER DETAILS FOR PREDICTION
SELECT MODEL TO PREDICT SALES
PREDICTION RESULT
COMPARE ALL THE ALGORITHMS

Figure 1: System Architecture

The predictive analytics in this paper is done using Multiple Regression [10] algorithm, which will be discussed further now.

3.3 Multiple Linear Regression

Linear regression, while a useful tool, has significant limits. As it's name implies, it can't easily match any data set that is non-linear. It can only be used to make predictions that fit within the range of the training data set. And, most importantly for this article, it can only be fit to data sets with a single dependent variable and a single independent variable. This is where multiple linear regression comes in. While it can't overcome all three of those weaknesses of linear regression, it is specifically designed to create regressions on models with a single dependent variable and multiple independent variables.

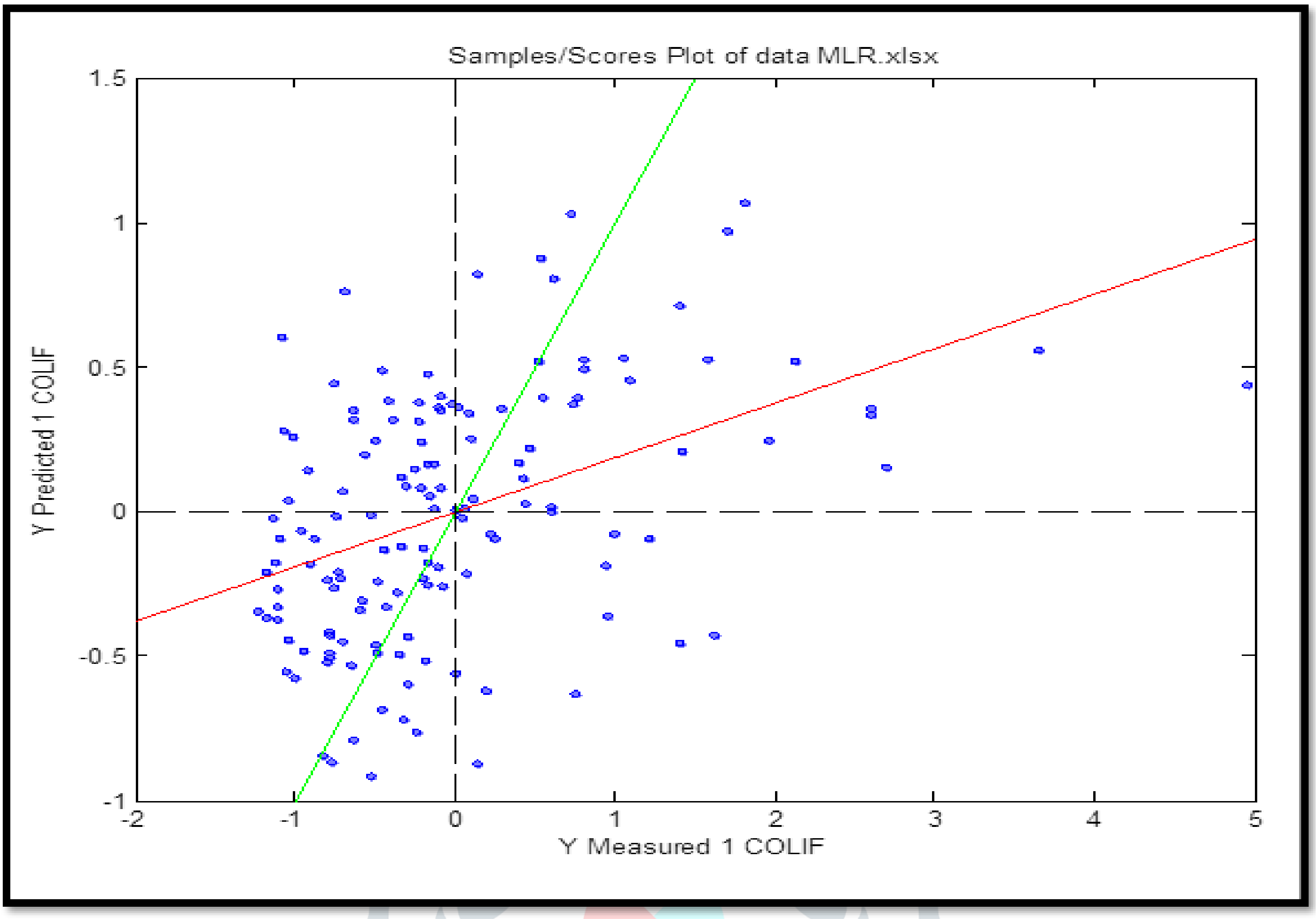


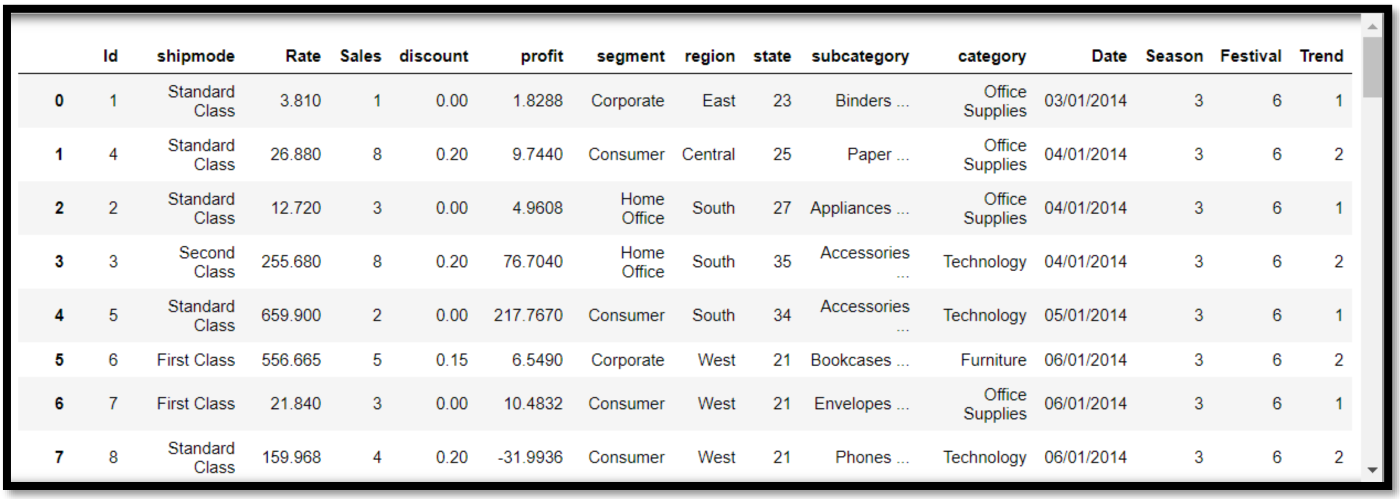
Figure 2: Sample Multiple Linear Regression Model for Scores

IV. IMPLEMENTATION

Initially, predictive analysis are formed using the Multiple Regression Algorithm and later on different Machine Learning Algorithms are compared on the basis of RMSE(Root Mean Square Error) value.

4.1 Predictive Analysis

In this paper, the dataset is collected from the various online websites such as Kaggle, UCI Machine Learning Repository, Data.gov. Predictive analysis is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future. After this, the collected data was manipulated, unwanted information was removed in excel and finally it was stored in the form of CSV (Comma Separated Value) file as the model is built in Jupyter notebook (Python). After this dataset was imported to jupyter notebook using python library known as Pandas which provide an easy way to create, manipulate and wrangle the data. The dataset consists of both categorical as well as numerical features, but we need numerical values so that the machine can understand and process the values into machine learning format. For this the dataset was manipulated initially in excel. After this the required columns were retained and remaining were dropped from the datasets. As the dataset include date column and during execution the string values were not getting converted into floats so later on Label Encoder was used to conquer this issue and complete the execution. The exported dataset is shown in the figure:



	Id	shipmode	Rate	Sales	discount	profit	segment	region	state	subcategory	category	Date	Season	Festival	Trend
0	1	Standard Class	3.810	1	0.00	1.8288	Corporate	East	23	Binders ...	Office Supplies	03/01/2014	3	6	1
1	4	Standard Class	26.880	8	0.20	9.7440	Consumer	Central	25	Paper ...	Office Supplies	04/01/2014	3	6	2
2	2	Standard Class	12.720	3	0.00	4.9608	Home Office	South	27	Appliances ...	Office Supplies	04/01/2014	3	6	1
3	3	Second Class	255.680	8	0.20	76.7040	Home Office	South	35	Accessories ...	Technology	04/01/2014	3	6	2
4	5	Standard Class	659.900	2	0.00	217.7670	Consumer	South	34	Accessories ...	Technology	05/01/2014	3	6	1
5	6	First Class	556.665	5	0.15	6.5490	Corporate	West	21	Bookcases ...	Furniture	06/01/2014	3	6	2
6	7	First Class	21.840	3	0.00	10.4832	Consumer	West	21	Envelopes ...	Office Supplies	06/01/2014	3	6	1
7	8	Standard Class	159.968	4	0.20	-31.9936	Consumer	West	21	Phones ...	Technology	06/01/2014	3	6	2

Figure 3: Data represented in Jupyter notebook

After this the features were co-related with each other and are visualized using Seaborn and Matplotlib libraries which is shown in the figure:

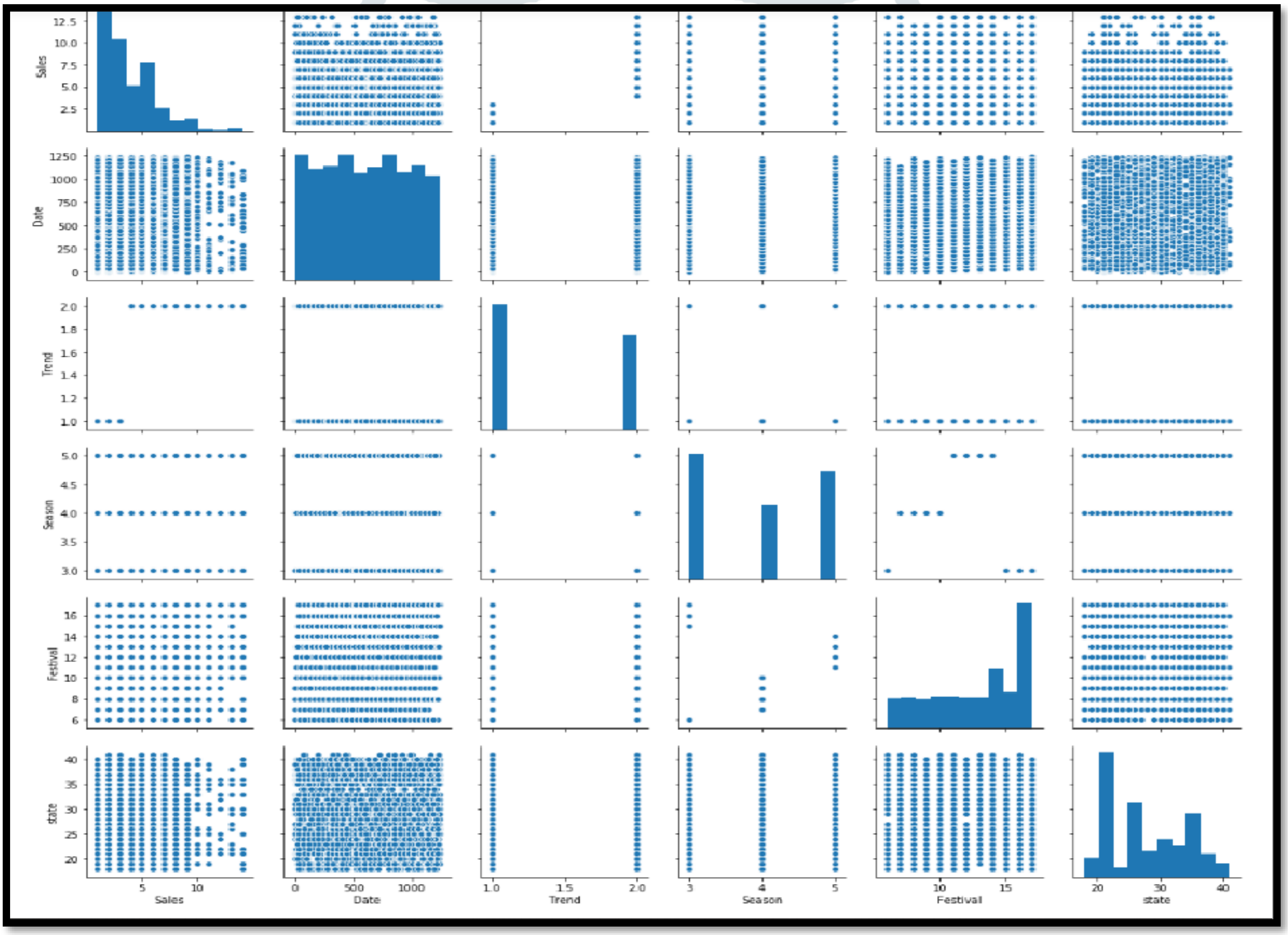


Figure 4: Data visualization in Jupyter notebook

Later on, the data was split using train-test split. In this he data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. Finally the Multiple Regression was imported from sklearn model and after fitting the dataset into the model, predictions were made using model.predict function which is used in machine learning to predict the further values or sales.

4.2 Comparison

In this paper, different Machine Learning Algorithms are compared to analysis that which one gives the best results and these algorithms are compared on the basis of RMSE (Root Mean Square Error) Value. The model with RMSE value closer to 1 indicates the best fit model. Various algorithms are elaborated further.

4.2.1 K-Nearest Neighbour (KNN):

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. A supervised machine learning algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data. A classification problem has a discrete value as its output. A regression problem has a real number (a number with a decimal point) as its output. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

4.2.2 Random Forest (RF):

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach. It can be used for both classification and regression tasks. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won’t overfit the model.

4.2.3 Naive Bayes (NB):

This lets us examine the probability of an event based on the prior knowledge of any event that related to the former event. The equation of Bayes’ theorem says that if A and B are two events and, P(A—B): the conditional probability that event A occurs, given that B has occurred. This is also known as the posterior probability. P(A) and P(B): probability of A and B without regard of each other. P(B—A): the conditional probability that event B occurs, given that A has occurred.

4.2.4 Support Vector Regression (SVR):

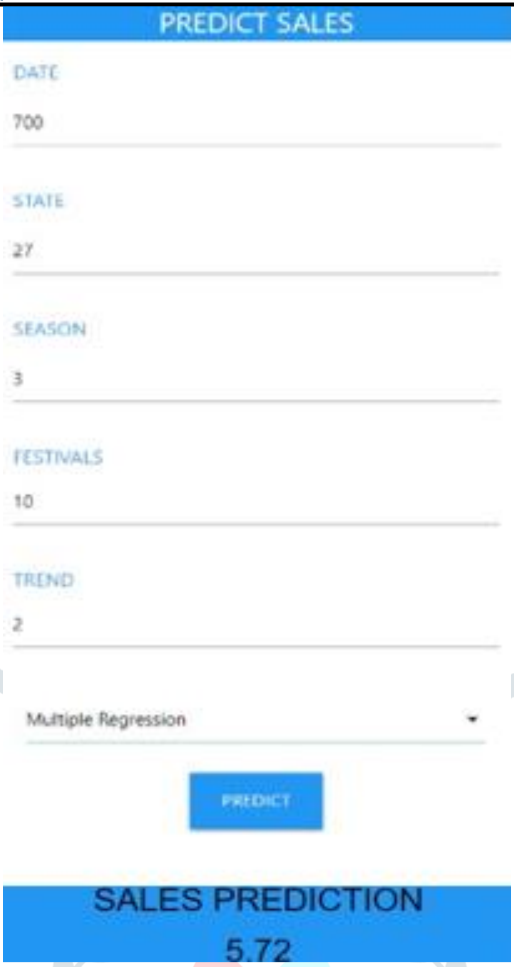
Support Vector Regression (SVR) is the combination of a Support Vector Machines and Regression. Support Vector Machines (SVMs) are used for classification. The goal of an SVM is to define a boundary line between the 2 classes on a graph. We can think of this as “splitting” the data in the best possible way. This boundary line is called a hyperplane.

4.2.5 Multiple Linear Regression (MLR):

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).

V. RESULT ANALYSIS

The results of the predictive analysis are observed on the UI (User Interface) after the required fields are filled by the users. The inputs of all the categorical fields are given in numerical form as well which helps the model to execute faster. The categorical fields which are converted into numbers are represented on user interface, so the user just have to input the numerical values to get the desired output. The figure below shows the sales prediction on user interface.



PREDICT SALES

DATE
700

STATE
27

SEASON
3

FESTIVALS
10

TREND
2

Multiple Regression

SALES PREDICTION
5.72

Figure 5: Sales Prediction

Later on, comparison of various machine learning algorithms was performed and their results are presented in the tabular in as well as graphical format.

Table 1: Comparison of experiment results of five algorithms.

Serial Number	Algorithm	RMSE
1	MLR	1.32
2	KNN	2.58
3	NB	7.02
4	SVR	2.35
5	RF	2.51

where,
MLR is MULTIPLE LINEAR REGRESSION
KNN is K-NEAREST NEIGHBOURS
NB is NAIVE BAYES
SVR is SUPPORT VECTOR REGRESSION
RF is RANDOM FOREST

As the table and graph shows that Multiple Regression has the lowest RMSE value of 1.32 which is closer to 1 and indicates the best model as compared to others in this case.

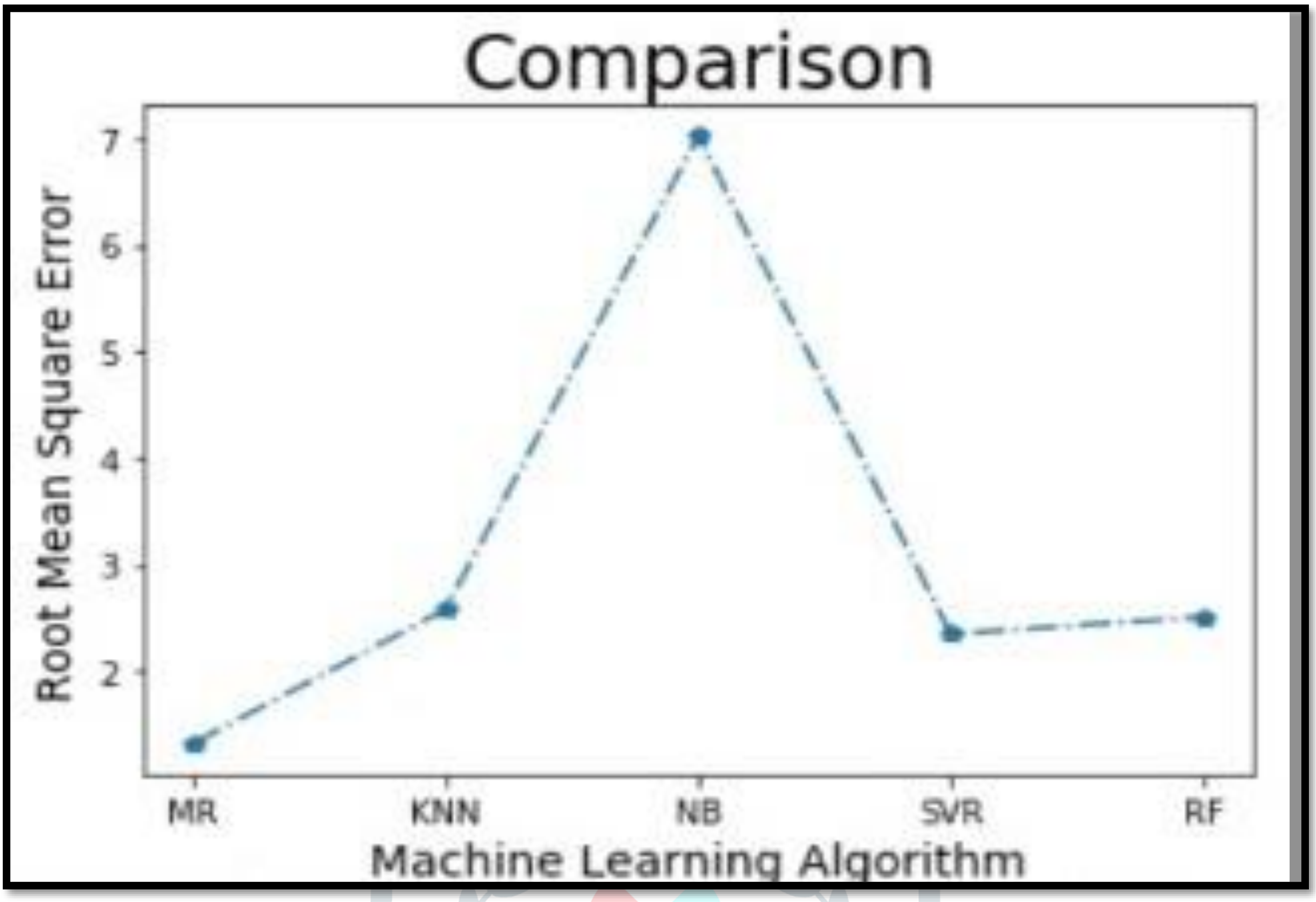


Figure 6: Accuracy Comparison

VI. CONCLUSION

Thus, the main objective of predictive analysis to predict the sales in performed. We also compared the RMSE (Root Mean Square Error) value of various algorithms to check the accuracy of the models and we concluded that Multiple Regression has the lowest RMSE esteem. The reason for estimating exactness was to approve our forecast with the genuine result. This is a significant advance to produce trust of target crowd so they can accept to our expectations are right and take important activities as needs be. With this model, future sales can be predicted which will help the sellers and manufacturers to analyze the production of the product. With our forecasts, they can refine their approaches and methodologies to increment their deals of items. Further multiple products as well as factors can be added to increase the efficiency and accuracy. Multiple algorithms can be used together and built a prediction model.

REFERENCES

[1] Sanket Rai, Aditya Gupta, Abhinav Anand, Aditya Trivedi and Saumya Bhadauria, “Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods,” ABV Indian Institute of Information Technology and Management Gwalior, Madhya Pradesh, India, 2019.

[2] Ching-Seh (Mike) Wu,Pratik Patil, and Saravana Gunaseelan, “Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data,”San Jose State University San Jose, CA 95192, 2019.

[3] Gopalakrishnan T,Ritesh Choudhary, and Sarada Prasad, “Prediction of Sales Value in online shopping using Linear Regression,” Manipal University Jaipur, Rajasthan, India, 2018.

[4] Rahmatika Pratama Santi and Masayu Leylia khodra, “Profit prediction using regression model for travel agents,” School of Electrical Engineering and Informatics Institut Teknologi Bandung, 2018.

[5] YouLi Feng , ShanShan Wang, “A Forcast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression,” Inner Mongolia university Inner Mongolia , China, 2017.

[6] LIJun, “Multiple linear regression method based on factor analysis and its application in stock prediction[D],” Nanjing University, 2014.

[7] LinBin, “Multiple linear regression analysis and its applicatio[N],” CHINA SCIENCE AND TECHNOLOGY INFORMATION May, 2010.

[8] Wang, Huiwen, Meng, Jie, “Multiple linear regression prediction modeling method[J],” Journal of Beijing University of Aeronautics and Astronautics, 2007.

[9] Wei, Zhijing, liu, XiYu, Zhao, QingZhen,“ The Analysis Based on Statistic Software SPSS and Multiple Linear Regression Analysis[N],” Information technology and information, 2005.

[10] Camus R,Cantarella G E,Inaudi D, “Real-time estimation and prediction of origin-destination matrices per time slince[J],” International journal of Forecasting,1997.

[11] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[12] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987.