A Study of Demand and Sales Forecasting Model using Machine Learning Algorithm

Aneesh Tony[1], Pradeep Kumar[2], Rohith Jefferson[3], Subramanian[4]

[1,2,3] Student, Electronics and Communication Engineering Department, National Engineering College, K. R. Nagar, Kovilpatti

[4] Assistant Professor (SG), Electronics and Communication Engineering Department, National Engineering College, K. R. Nagar, Kovilpatti

ABSTRACT

Sales forecasting is a fundamental assignment in retailing. In such manner, the usage of machine-learning models for sales predictive investigation were studied. The goal is to review scientific literature and distinguish if there are advantages over conventional statistical techniques. The detailed study and examination of predictive models is to improve future sales predictions are completed in this study. Millions of reviews are being created daily which makes it hard for a consumer to make a good decision on whether to shop for the product. Investigating this huge number of opinions is hard and tedious for product manufacturers. This work considers the issue of arranging reviews by their overall semantic (positive or negative) In this paper, the proposed work uses three machine learning algorithms namely Linear Regression, Decision Tree (DT) and Random Forest (RF) in sales prediction and Logistic Regression for classifying the reviews. The forecasting precision of each scenario is evaluated with the Root Mean Square Error (RMSE). The examinations found that the best model is Random Forest Algorithm, which shows greatest precision in forecasting and in future sales prediction as it had a lower mean absolute error than the other two models.

Keywords
Machine Learning, Prediction, Sales forecasting, Sentimental analysis.

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

I. Introduction

The objective of every supermarket store is to shape benefit. This is accomplished when more products are sold, and hence the turnover is high. A significant test to expand the deals of a grocery store lies in the capacity of the manager to estimate sales and know promptly heretofore when to arrange and recharge inventories just as plan for labour and staffs. One of the preeminent significant resources a store can have is that the information created by clients as they communicate with different stores. Inside these information lies significant patterns and factors that can be displayed using a machine learning algorithm and this can prompt to a serious level of precision correctly forecast sales. Ventures should adjust their inventory network to front line advances and strategies to improve their performance, lessen costs and to serve better. A few of these new forward leaps have been permitted by Machine Learning, giving answers for complex issues that were as of recently hard to address, or improving the results of previous methods. Demand forecasting might be a fundamental part of production planning and give chain management, affecting competitiveness and productivity, giving basic data to buying choices, production, stock levels, accounts and marketing. In the ware area, where items are burned-through quickly, sales estimation becomes considerably more basic to business.

Because of quick weakening, some quick customer merchandise like meats, organic products, vegetables and dairy items have a short time span of usability, which are exceptionally transient. Different items, such as electronic products and design clothing, have short lifecycles, fast out of date quality, are habitually refreshed, and have many contending choices. In the retail food industry by and large, the most clarification for squandered items and stockouts is that the incorrectness of sales forecasting results in incorrect requests. All the more explicitly inside the fresh foods industry, including the refrigerated ones, similar to the dairy, leafy foods fragments have short timeframe of usability and wish to deal with quality inside the capacity and circulation measures make sales forecasting precision a significant factor for planning production, limiting lost deals because of an absence of items, diminishing returns because of the closeness of the expiry date, and improving accessibility of items to clients.

Numerous grocery stores now-a-days don't have an estimate of their yearly deals. This is because of the lack of abilities, assets and information to shape sales estimation. There are a few strategies to forecast sales and numerous grocery stores have depended on the conventional models. The utilization of conventional measurable technique to forecast supermarket sales has left a ton of difficulties unaddressed and generally bring about the production of predictive models that perform ineffectively. Nonetheless, machine learning has become an essential subject matter science that has made progress on account of its high predictive and forecasting. The period of enormous information including admittance to monstrous compute power has made machine learning go to for sales forecast. To accurately estimate a future occasion, a machine learning model is trained on past information from which it learns designs that predict future forecasts. An accurate forecasting model can significantly expand supermarket income and it improves benefit additionally and also gives experiences into the manner in which clients can be better served.

As online commercial places have gotten famous in ongoing many years, web venders and dealers are requesting their clients to share their perspectives on the things they have bought. Millions of reviews on various goods, services and places are created every day across the Internet. This has made the Internet the most popular source of a product or a service to get ideas and opinions. However, as the measure of surveys accessible for an item builds, it turns out to be hard for a forthcoming purchaser to settle on a legitimate choice about whether to buy the item. On the one hand various sentiments on a similar item and opposing criticism then again leave shoppers more unsure in asking the correct choice. Here the necessity for analysing these contents seem crucial for all e-commerce businesses. For all online business firms, the need to inspect these contents here seems important. In recent years, machine learning methods have become popular for their simplicity and accuracy in semantic and review investigation.

The motivation behind this investigation is twofold: firstly, to call attention to the most recent ML trends used in Demand & Sales Forecasting over conventional techniques; and furthermore, a strategy for surveying buyer audits to an item. A technique of forecasting demand by employing a model for grasping the fluctuation of sales, by putting away a majority of models of neural network, and furthermore by feeding sales results into a model of neural network to make it learn by the brief time frame such as by the week, and a recording medium to obtain the results. A strategy for evaluating customer audits to an item, including the means of (a) gathering the information, (b) pre-processing the collected data, (c) categorizing the reviews as positive or negative. Thus, the question to be answered by this study is: What are the advantages of applying Machine Learning to demand forecasting? The objective is to review scientific literature and identify if there are advantages over traditional techniques.

In section 2, the various literature reviews regarding sales forecasts are stated. In section 4, the visual representation of the data pre-processing techniques and predictions methods are highlighted. Discussion of consumer review categorizing process is done in section 5. Using various machine learning prediction algorithms the performance evaluation is calculated. Finally, the result is concluded by analysing the research summarization and future scope.

II. LITERATURE SURVEY

The writing audit performed in this research has identified important papers related to Machine Learning applied to demand prediction, focusing on the benefits achieved, business sectors addressed and the possible advantages over conventional statistical techniques. The utilization of conventional technique to figure grocery store deals has left a ton of difficulties unaddressed and for the most part bring about the formation of prescient models that perform ineffectively. To accurately figure a future occasion, an AI model is prepared on information from which it learns designs that are wont to anticipate future cases. Sales prediction is vital piece of present business intelligence. It is often an opulent issue, particularly within the case absence of information, missing information, and therefore the presence of anomalies. To be adequately equipped and to get higher income, business associations are continually looking for an obviously better model or method for information handling and support of basic information. During this exploration, the definite examination and investigation of understandable prescient models are endorsed to improve future sales prediction.

More reliable forecasting of demand for fast-moving goods, especially in the apparel, technology, and fresh foods sectors, may be a competitive advantage for manufacturers and retailers. The findings of the research paper [1] indicate that Machine Learning approaches, as well as a mixture of different techniques used in market forecasting models, will help fast-moving consumer goods producers and retailers. The most advantage noticed was improved accuracy in demand forecasting. When compared to conventional predictive methods, demand estimation was better, the flexibility to manage a larger range of data factors was better, and the capacity to process huge data volumes was better. Predicting demand is more like a regression problem than a mathematical problem. When opposed to predictive techniques, regression approaches for sales forecasting will produce improved outcomes. The research paper [2] explored a stacking method for creating a regression ensemble of single models. The findings suggest that predictive model efficiency for sales time series forecasting can be enhanced by using stacking techniques. We used a relative mean absolute error (MAE) for error estimation, which is measured as error = MAE/mean (Sales) 100 %. The most commonly used approaches for estimating demand are time series forecasting strategies. Exponential smoothing, Holt Winters model, Box & Jenkins model, regression simulations, and ARIMA are examples of well-known mathematical techniques. Their effectiveness is highly dependent on the implementation area, the projected target, and the user interface. The well-known Root Mean Square Error is used to determine the prediction precision of each case (RMSE). As anticipated, improving forecast accuracy, i.e., lowering the RMSE, leads to improved results for both the consumer and the manufacturer [3].

The examination [4] utilizes AI calculations as an exploration approach to build up a house price expectation model. To improve the precision of housing price expectation, the paper examines the housing information of 5369 houses in Fairfax County, Virginia and built up a housing price forecast model dependent on AI calculations like C4.5, RIPPER, Naïve Bayesian, and AdaBoost and looked at their classification performance. The investigations show that the RIPPER algorithm, upheld precision, reliably beats the contrary models inside the presentation of lodging value forecast. One of the discoveries was that ML widens the span of D&SF, as it can deal with complex factors. More definitely, Fuzzy-ANN approaches showed great execution when managing uncertain information like climate factors, though DT and RF offered significant translation limit. Besides, Data Pre-preparing procedures demonstrated to considerably diminish the intricacy of the models, empowering both great precision and sensible figuring time. The outcomes show that ML genuinely outflanks conventional models in D&SF [5].

In this work [6], an insightful demand forecasting framework is created. The improved model depends on the examination and translation of historical information by using different forecasting machine learning algorithms. The proposed framework mixes nine distinctive time series techniques including moving average (MA), exponential smoothing, Holt-Winters, ARIMA strategies, and three Regression Models, SVR algorithm, and DL model including multilayer feedforward artificial neural network (MLFANN) are mixed by another combination methodology which is suggestive of boosting ensemble procedure. results demonstrate that the proposed framework presents recognizable precision enhancements for demand forecasting process in contrast with single prediction models.

In the paper [7], sales forecasting deals with three machine learning based algorithms (K-Nearest Neighbour, Gradient Boosting and Random forest) were researched. The outcomes show that the Random Forest algorithm performs better compared to the other two models, Gradient Boosted models effectively overfits to the dataset and K-Nearest Neighbour performs most unfortunate among the three. Subsequently it is seen that getting more information would increase the prediction accuracy of our models. The study paper [8] directed an outline of ongoing advancement in the field of sales forecasting with the emphasis on fashion and new product forecasting. For sales forecasting, statistical procedures, such as exponential smoothing, ARIMA, Box and Jenkins model, regression models or Holt-Winters model, are frequently applied. Traditional forecasting strategies face difficulties in delivering exact deals information for new items and customer arranged products. Within the recently presented approaches, hybrid forecasting models performs well.

In the paper [9], three machine learning based algorithms were examined which can be applied to prediction, like Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT). The exploration found that the Gradient Boost Algorithm is the best fit model, which exhibits greater precision in forecasting and prediction of future sales. Since internet business websites allow purchasers to leave feedback on various items, electronic trade is becoming progressively well known. Clients produce millions of reviews every day, making it unimaginable for item creators to monitor customer perceptions about their items. Thus, to extract useful information from a large assortment of data, it is important to classify such large and complex data. To conduct the study two different supervised machine learning techniques, SVM and Naive Bayes, has been attempted on beauty products from Amazon. Their precision has been compared. The outcomes showed that the SVM approach outperforms the Naive Bayes approach when the dataset information is bigger [10].

III. PROBLEM DEFINITION

A major challenge of a supermarket lies in the ability of the manager to forecast sales pattern and know readily beforehand when to order and replenish stocks as well as plan for manpower and staffs.

The main problems identified are

- Lack of forecasting accuracy
- Inability of traditional statistical methods to handle large data
- Poor performance of predictive models

The goal of this project is to accomplish two things:

- To begin, highlight the advantages of Machine Learning algorithm used in Demand & Sales Forecasting over traditional methods and
- secondly, a method of classification of reviews into positive and negative reviews based on the selected words.

IV. METHODOLOGY

Demand forecasting is a method of predictive analytics that aims to predict market demand. It's achieved by searching for trends and associations in statistical results. In order to predict demand, a number of methods and analysis have been used. Machine learning forecasting approaches may take a large number of data and features related to demand and use various learning algorithms to predict future demand and trends. As shown in fig. 1, the proposed work addresses three machine learning algorithms that can be used for predictive analytics: Linear Regression, Decision Tree (DT), and Random Forest (RF), and Logistic Classifier for sentimental analysis.

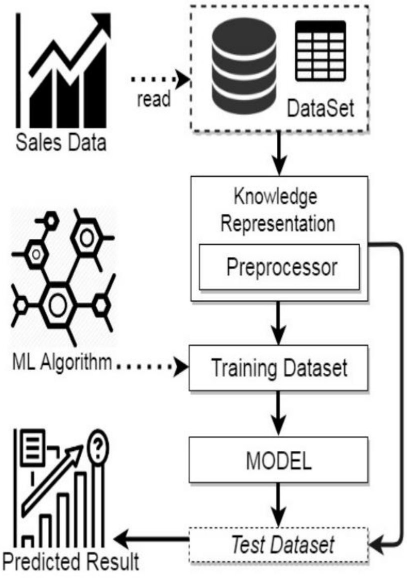


Fig. 1 System Architecture

A. Sales Prediction Algorithms

1) Linear Regression

Linear Regression is an algorithm that solves the Regression problem which is very simple to understand. When the attribute is constant, this is the typical form of regression. A linear relationship is formed between this variable and the independent variables in order to forecast it. In this case, the model creates a straight line that runs over the bulk of the data points, and this line functions as the forecast. The line with the least amount of error is known as the line of best fit. Nowadays, in the era of Machine Learning, where often complex statistical or tree-based algorithms are used to come up with highly accurate predictions, Linear Regression is one of those "classic" traditional algorithms that are adapted in Machine Learning, and thus the use of rectilinear regression in Machine Learning is profound. The equation for a linear regression line is Y=p+qX, where X is the explanatory variable and Y is the dependent variable. The line's slope is q, and the intercept is p.

2) Decision Tree

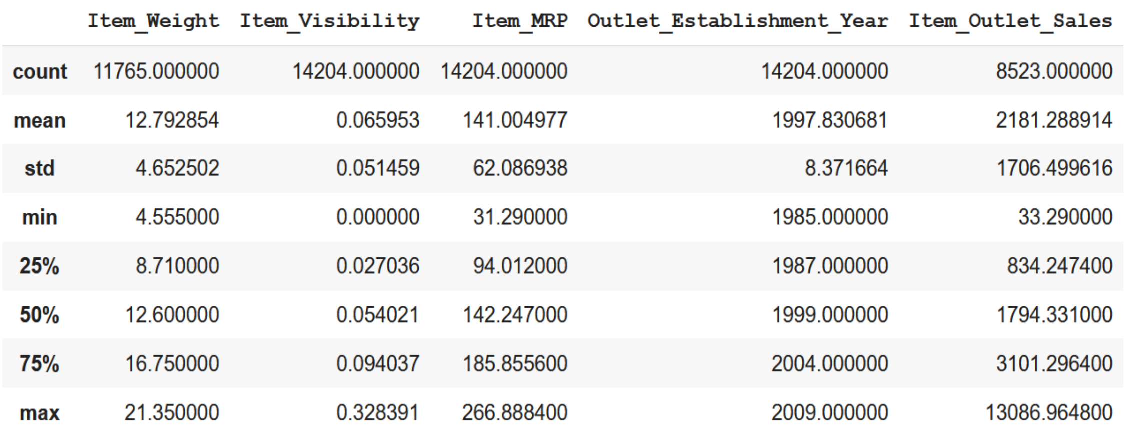
The supervised learning group contains the decision tree concept. They are often used to solve problems including regression and classification. The tree representation is used to solve the problem, with internal nodes representing dataset attributes, branches representing decision rules, and each leaf node representing the result. The accuracy of the decision tree classifier is usually good. The algorithm for predicting the type of a given dataset in a decision tree begins at the tree's base node. This algorithm supports the relation by comparing the values of the root attribute with the values of the record (real dataset) attribute, then following the branch and moving to the next node. The algorithm contrasts the attribute value with the opposite sub-nodes for each subsequent node before going on. It repeats the process until it hits the tree's leaf node.

3) Random Forest

Random forest is a versatile, easy-to-use machine learning algorithm that, in most cases, produces excellent results. Because of its simplicity and usability, it is perhaps one of the most commonly used algorithms. A random forest algorithm combines several machine learning algorithms (Decision trees) to obtain better accuracy. This is also called Ensemble learning. To boost precision, a random forest algorithm blends many machine learning algorithms (Decision trees). Ensemble learning is another name for this. In addition, the forest tends to be more resilient the more trees it includes. Similarly, in the random forest classifier, the larger the number of trees in the forest, the better the accuracy results. Although if some trees make false predictions, the rest of them would produce true predictions, improving the model's overall accuracy. Overfitting is an issue with the Decision Tree algorithm, which results in high precision on training data but low output on test data. Random forest, on the other hand, takes advantage of this vulnerability by allowing each tree to randomly sample from the dataset to produce various tree structures.

B. Data Collection and Preparation

The single most critical step in solving any machine learning problem is data collection. A data set is a set of information. In other words, an information set is the contents of a single database table or statistical data matrix, where each column of the table represents a specific variable and each row represents a specific member of the data set, as shown in Figure 2. A training data set is required for Machine Learning projects. It is the unique data set that is needed to train the model for different tasks. ML is highly dependent on data; without data, an algorithm would be unable to discover.



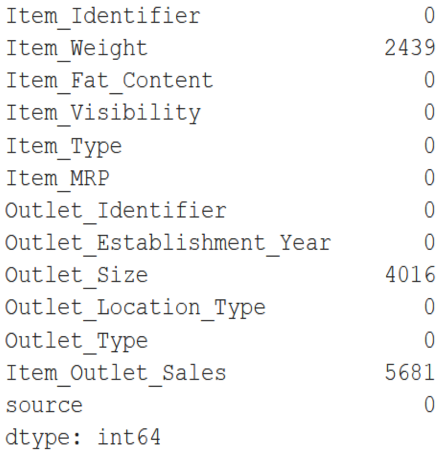
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	11765.000000	14204.000000	14204.000000	14204.000000	8523.000000
mean	12.792854	0.065953	141.004977	1997.830681	2181.288914
std	4.652502	0.051459	62.086938	8.371664	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.710000	0.027036	94.012000	1987.000000	834.247400
50%	12.600000	0.054021	142.247000	1999.000000	1794.331000
75%	16.750000	0.094037	185.855600	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Fig. 2 Numerical Data Summary – Numerical data summery is a number used to describe a specific characteristic about a dataset

It is known that all data sets are considered to be unreliable. At this point in the project, data processing is needed, which is an essential phase in the machine learning process. Data planning is basically the method of making a data set more suitable for machine learning. It is a series of procedures that take up the bulk of the time of machine learning programmes.

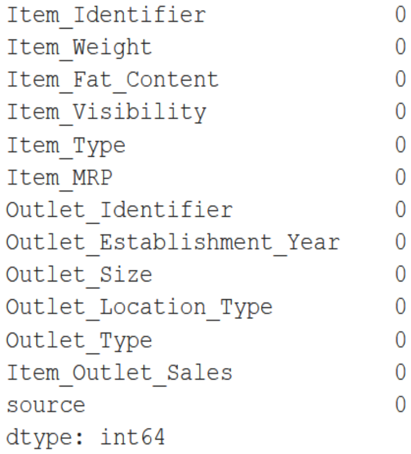
C. Data Pre-processing

The data that were required, diverse, and representative for the proposed work were collected at this point. Pre-processing involves extracting the necessary data from the entire data collection and creating a training set. In machine learning, good data pre-processing is the most crucial aspect that can make a big difference between a good model and a bad model. Data pre-processing can be described as the process of preparing data for use in a machine learning model. It is the first and most critical step in the development of a machine learning model.



```
Item_Identifier                  0
Item_Weight                   2439
Item_Fat_Content                 0
Item_Visibility                  0
Item_Type                        0
Item_MRP                         0
Outlet_Identifier                0
Outlet_Establishment_Year        0
Outlet_Size                   4016
Outlet_Location_Type             0
Outlet_Type                      0
Item_Outlet_Sales             5681
source                           0
dtype: int64
```

Fig. 3 Checking missing values



```
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_Size                  0
Outlet_Location_Type         0
Outlet_Type                  0
Item_Outlet_Sales            0
source                       0
dtype: int64
```

Fig. 4 Missing values cleaned

D. Dealing with Missing data

It is not unusual for a real-world data set to have any incomplete data, as seen in fig. 3. With missing or null data, most machine learning algorithms may fail. As a consequence, coping with lost data becomes crucial. The missing values are explained in Figure 4. The three fields with missing values are Item_Weight, Outlet_Size, and Item_Outlet_Sales. Missing values are filled in by imputing mean and mode to the columns.

E. Prediction

To represent the foremost likely value obtained for a given input the output of word prediction in machine learning is utilised. Events occurring in the future is named as prediction. Machine learning algorithms improves the intelligence of the system without manual intervention. Ethem Alpaydin [11] stated that "Machine Learning (ML) is used to optimize the performance criterion using sample data or the past experience" [11]. Historical data is used to train the model, then predicts a specific property of the info for brand spanking new inputs. Various areas use prediction as it allows us to make highly accurate guesses about many things. When the model is deployed, any user who does not have any understanding of the machine learning technique can use this model for his or her usages. Businesses can appropriately use big data for his or her benefit by successfully applying predictive analytics.

F. Sentiment Analysis:

Sentimental analysis is generally used to analyse whether the following sentences are positive or negative. It deals with areas of judgements, feelings that are generated from the text, data mining and web mining. It mainly deals with neural networks. The process is initially to get the sentences which is to be analysed and then identification of sentences takes place followed by feature selection after which the sentence gets identified and polarity is found. For review classification, the Logistic Regression algorithm is used.

G. Logistic Regression:

Under the Supervised Learning technique, one of the most commonly used Machine Learning algorithms is logistic regression. It is a technique for estimating a single variable using a number of independent variables. A classification algorithm is logistic regression. It is normal to forecast a binary result based on a series of independent variables. A binary result is one in which there are only two options: the occurrence occurs (1) or it does not occur (0). As seen in Fig. 5, Linear Regression is used to solve regression problems, while Logistic Regression is used to solve classification problems. A threshold is often set to predict which class a knowledge belongs to. Centered on this threshold, the obtained estimated probability is assessed into classes by considering this threshold. There are two types of decision boundaries: linear and non-linear.

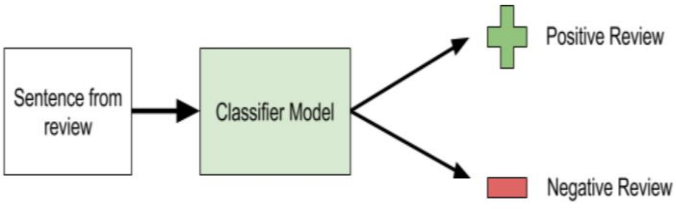


Fig. 5 Logistic Classifier

H. Data Loading:

To understand if a product review is positive or negative, it is important to identify it. The goods are given a 5-star ranking. They can be used to compare. The first step is to build a word dictionary that the model would use to identify.

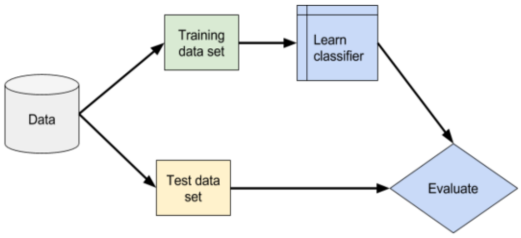


Fig. 6 Review Evaluation

I. Data Analytics:

This is the main step where the reviews are classified into positive or negative. In this step ML algorithm (Logistic Regression) is used to get the results. As sentiment analysis is being carried out, it is important to tell the model what positive sentiment is and what a negative sentiment is. In the rating column, there are ratings from 1 to 5. 1 and 2 can be defined as bad reviews and 4 and 5 as good reviews. Positive sentiments are defined as 1 and negative sentiments as 0. To decide whether a review is positive or negative, a sentiment classifier is established. The words (wordcount column) and reviews from the training data will be used to build a model to analyse the feedback, as seen in Fig. 6. Based on the model's expected feedback, the most favourable and negative responses can be presented.

V. RESULTS & DISCUSSION

The study's findings are summarised in this segment. The accuracy value represents the percentage of the testing data set that the model correctly identified.

A. Sales Prediction

The algorithms' precision is the most important factor in their success. The Root Mean Square Error is calculated, and the average of the errors is shown as the Error Rate in Fig. 7.

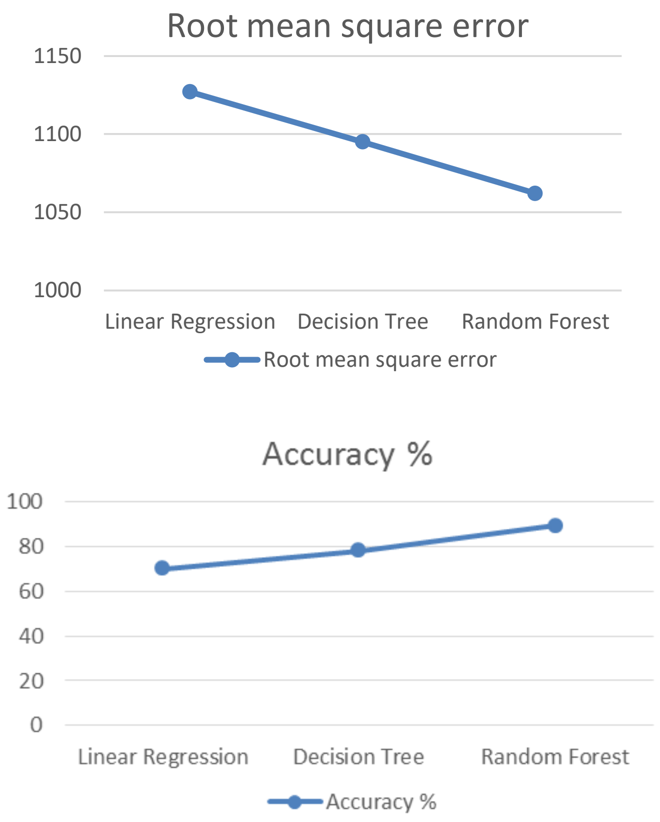


Fig. 7 Comparative performances of different forecasting algorithms

As predicted, an improvement in forecast accuracy, i.e., a reduction in the RMSE, leads to improved results not just for the retailer but also for the maker, as seen in Fig. 7. The past data is used to train the model, and then selected property of the data for new inputs is predicted. From fig. 8, it is inferred that low fat occupies the first place followed by regular fat. Fig. 9 provides the information that the occupancy of snack foods is more or less like the fruits and vegetables at the high rate and sea food occupies the least rate. From fig. 10, it is known that the sales are high at supermarket type 1 in comparison with all other markets.

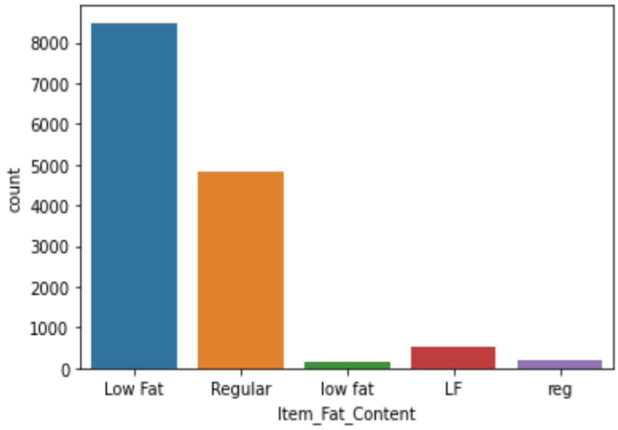


Fig. 8 Variable distribution of item fat content

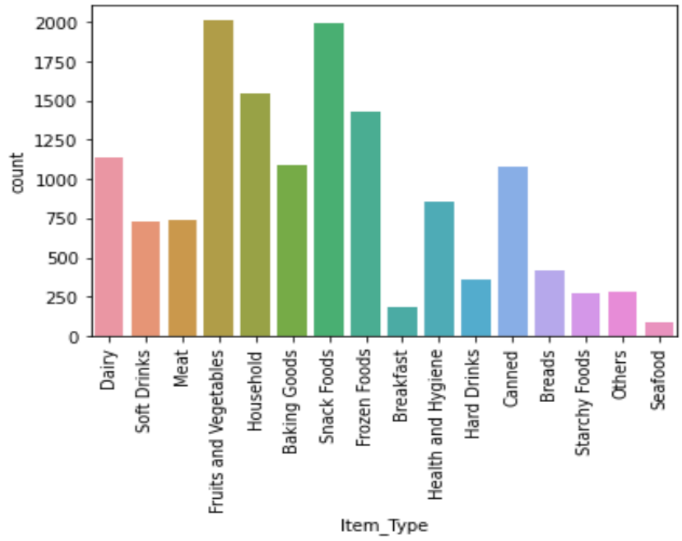


Fig. 9 Variable distribution of item type

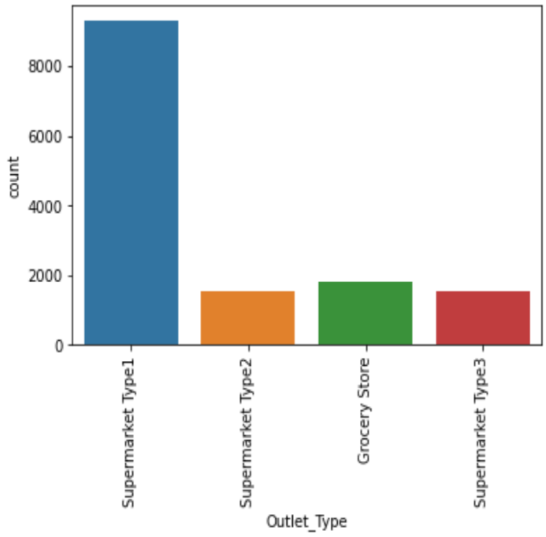
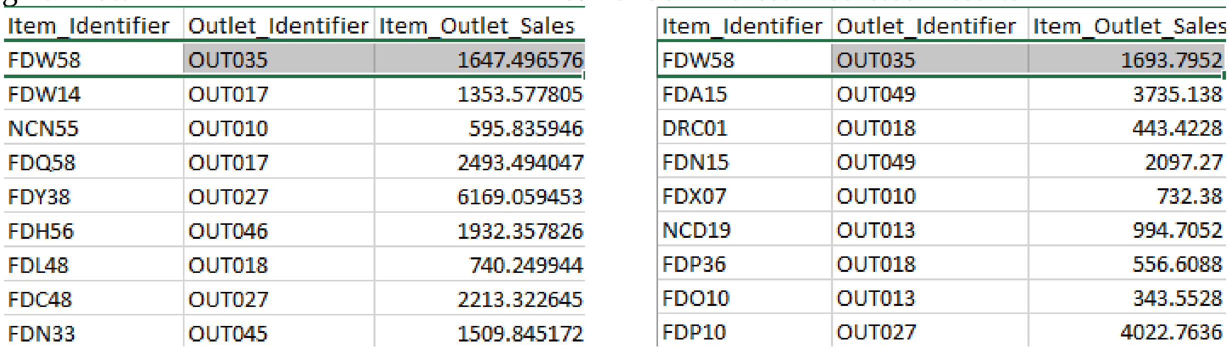


Fig. 10 Variable distribution of outlet type

Table 1 Comparative performances of different forecasting algorithms

Sl. No.	Algorithm	Root Mean Square Error	Accuracy %
1	Linear Regression	1127	70
2	Decision Tree	1095	78
3	Random Forest	1062	89



a. Original Data

Item_Identifier	Outlet_Identifier	Item_Outlet_Sales
FDW58	OUT035	1647.496576
FDW14	OUT017	1353.577805
NCN55	OUT010	595.835946
FDQ58	OUT017	2493.494047
FDY38	OUT027	6169.059453
FDH56	OUT046	1932.357826
FDL48	OUT018	740.249944
FDC48	OUT027	2213.322645
FDN33	OUT045	1509.845172

b. Random Forest Predicted Results

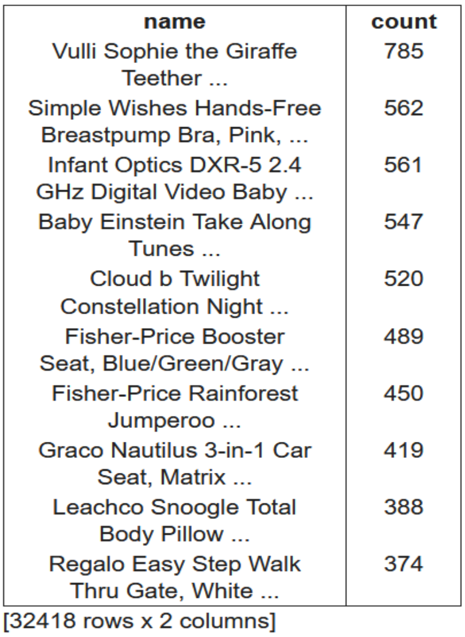
Item_Identifier	Outlet_Identifier	Item_Outlet_Sales
FDW58	OUT035	1693.7952
FDA15	OUT049	3735.138
DRC01	OUT018	443.4228
FDN15	OUT049	2097.27
FDX07	OUT010	732.38
NCD19	OUT013	994.7052
FDP36	OUT018	556.6088
FDO10	OUT013	343.5528
FDP10	OUT027	4022.7636

Fig. 11 Original Data Vs Predicted Results

Table 1 shows the comparative analyses of the three algorithms focused on prediction efficiency, and fig. 11 shows their projected outcomes. Random Forest Algorithm has 89 percent accuracy, Decision Tree Algorithms is second with approximately 78 percent accuracy, and Linear Regression Model is third with 70 percent accuracy, according to the data. Finally, depending on the precision of the three algorithms, the Random Forest algorithm is the best suited model.

B. Sentimental Analysis

1) Count of each product



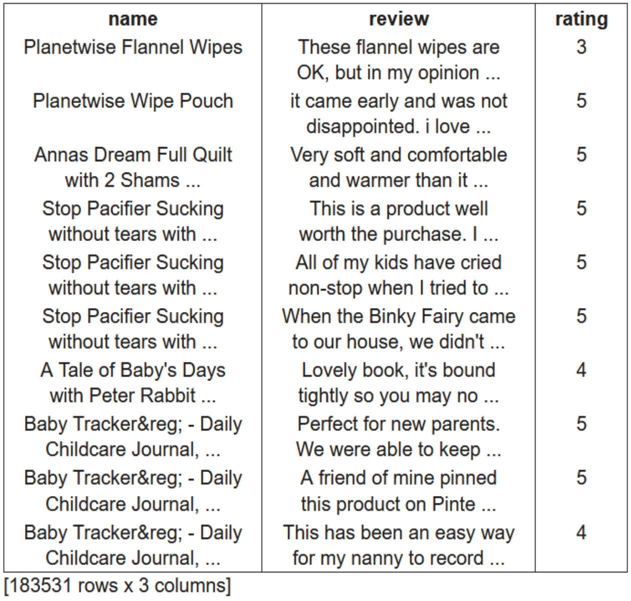
name	count
Vulli Sophie the Giraffe Teether ...	785
Simple Wishes Hands-Free Breastpump Bra, Pink, ...	562
Infant Optics DXR-5 2.4 GHz Digital Video Baby ...	561
Baby Einstein Take Along Tunes ...	547
Cloud b Twilight Constellation Night ...	520
Fisher-Price Booster Seat, Blue/Green/Gray ...	489
Fisher-Price Rainforest Jumperoo ...	450
Graco Nautilus 3-in-1 Car Seat, Matrix ...	419
Leachco Snoogle Total Body Pillow ...	388
Regalo Easy Step Walk Thru Gate, White ...	374

[32418 rows x 2 columns]

Fig. 12 Sorting the dataset based on count value

Fig. 12 represents the count of the product in the datasets which is taken for the analysis and they are sorted based on their count value. This Output gives us a clear view about the count of the individual product and their count.

2) *Head of the Datasets*



name	review	rating
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4
Baby Tracker&reg; - Daily Childcare Journal, ...	Perfect for new parents. We were able to keep ...	5
Baby Tracker&reg; - Daily Childcare Journal, ...	A friend of mine pinned this product on Pinte ...	5
Baby Tracker&reg; - Daily Childcare Journal, ...	This has been an easy way for my nanny to record ...	4

[183531 rows x 3 columns]

Fig. 13 Head of Sframe

Fig. 13 represents the head of the dataset. This output also shows the column present in the datasets.

3) *Distribution of rating values*

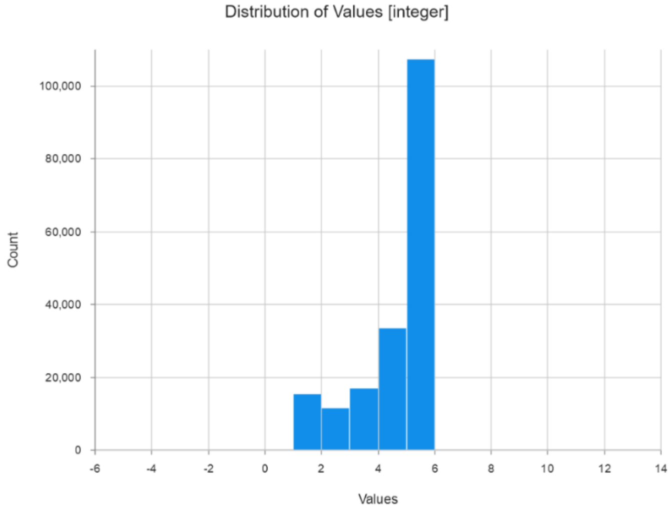


Fig. 14 Variable distribution of rating values

Fig. 14 explains the distribution of ratings over their count value.

4) *Product Sentiment Count Graph*

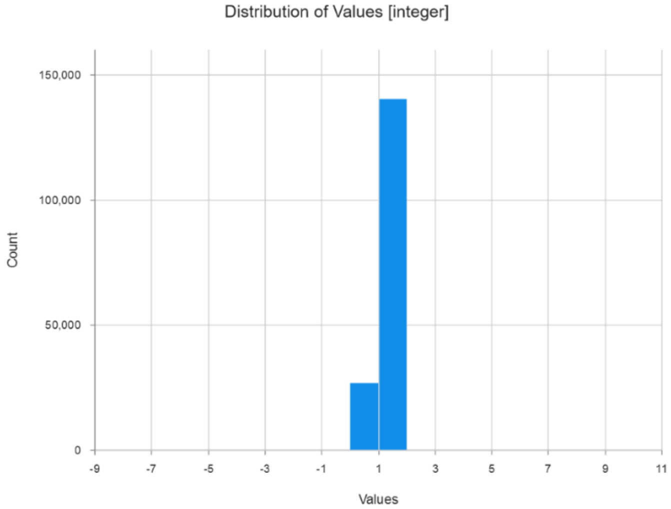
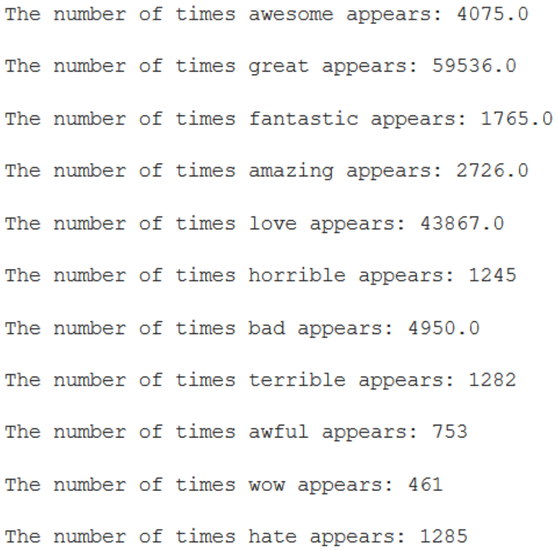


Fig. 15 Variable distribution of sentiments

In this step, the Product Id was removed first and then the reviews with the rating as 3 were removed since people who feel between good and bad will give rating as 3 and finally replaced the value as 1 for the ratings 4 and 5 and the value as -1 for the ratings 1 and 2. Fig. 15 represents the variable distribution of sentiments.

5) Word count of selected words

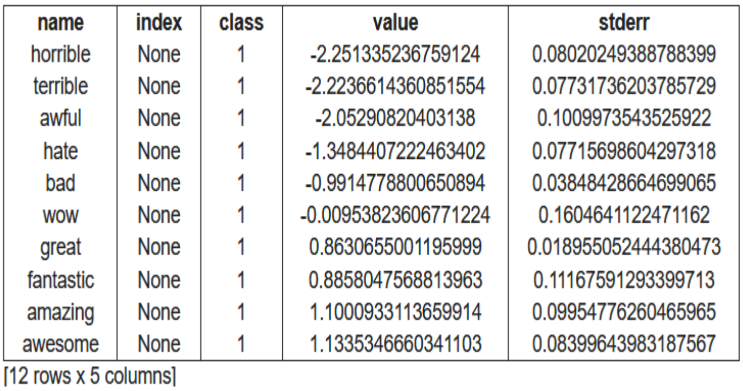


```
The number of times awesome appears: 4075.0

The number of times great appears: 59536.0

The number of times fantastic appears: 1765.0

The number of times amazing appears: 2726.0

The number of times love appears: 43867.0

The number of times horrible appears: 1245

The number of times bad appears: 4950.0

The number of times terrible appears: 1282

The number of times awful appears: 753

The number of times wow appears: 461

The number of times hate appears: 1285
```

Fig. 16 Word count of selected words

Several words were selected to make the data to train the model. For that some of the selected words were given using which the Positive and Negative review were classified. Fig. 16 represents the number of times the selected words appeared in the reviews which were given by the reviewers.

6) Allocating weight to each word



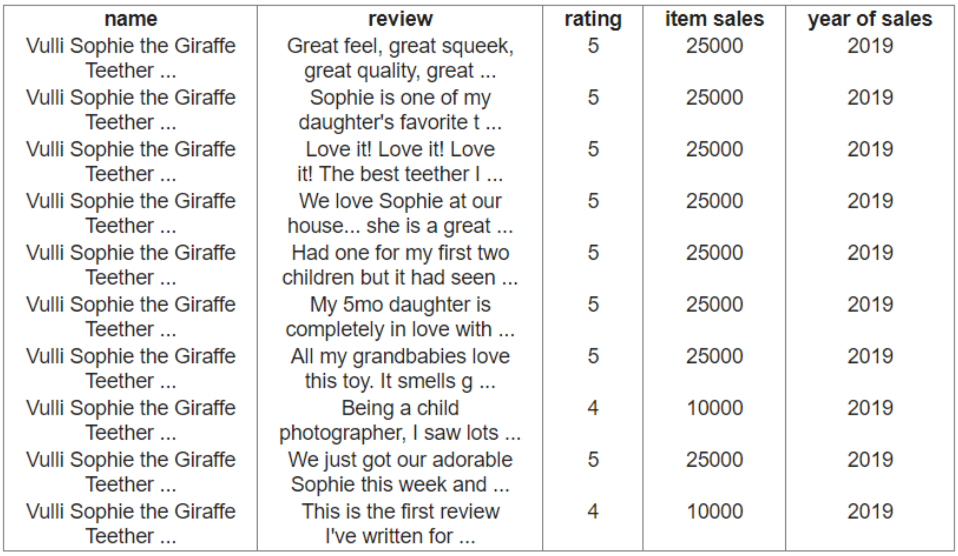
name	index	class	value	stderr
horrible	None	1	-2.251335236759124	0.08020249388788399
terrible	None	1	-2.2236614360851554	0.07731736203785729
awful	None	1	-2.05290820403138	0.1009973543525922
hate	None	1	-1.3484407222463402	0.07715698604297318
bad	None	1	-0.9914778800650894	0.03848428664699065
wow	None	1	-0.00953823606771224	0.1604641122471162
great	None	1	0.8630655001195999	0.018955052444380473
fantastic	None	1	0.8858047568813963	0.11167591293399713
amazing	None	1	1.1000933113659914	0.09954776260465965
awesome	None	1	1.1335346660341103	0.08399643983187567

[12 rows x 5 columns]

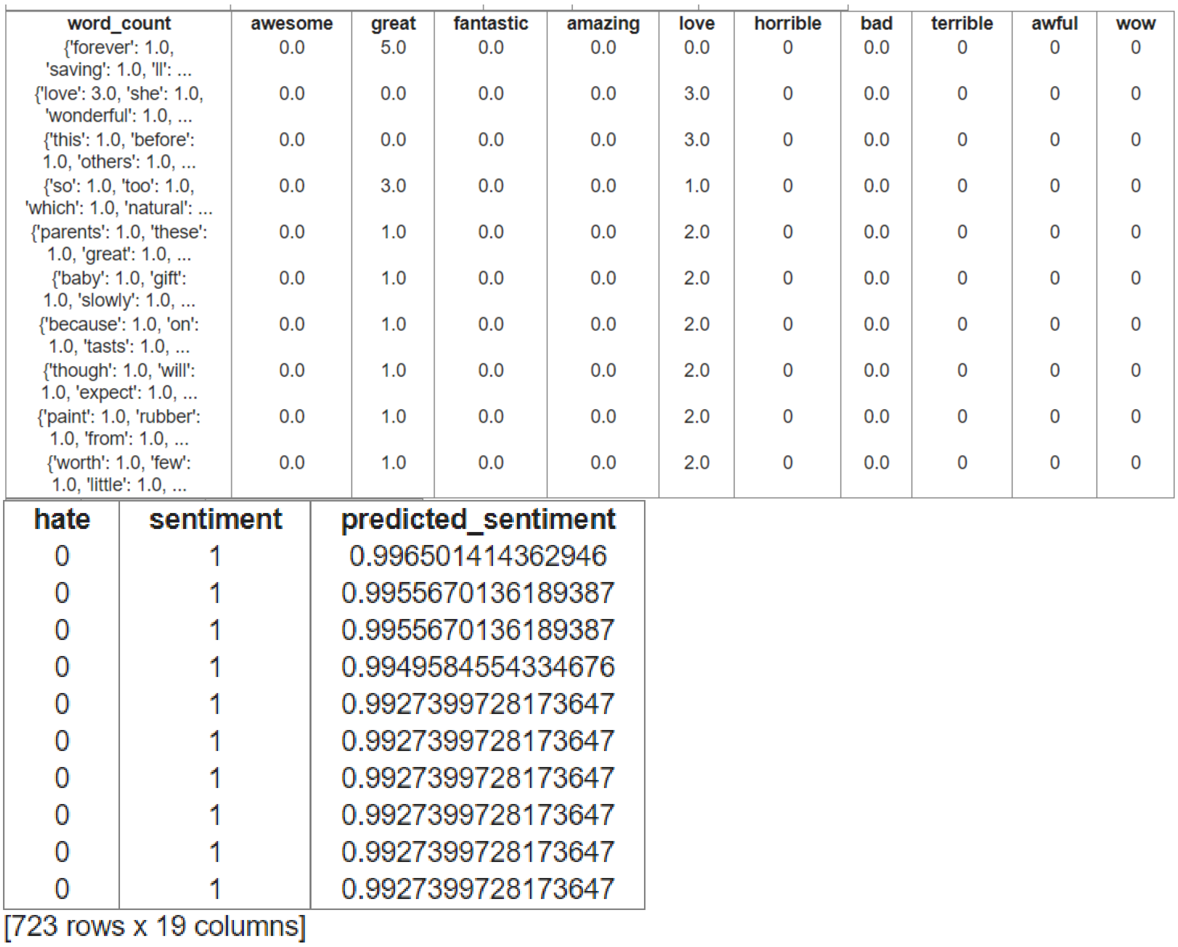
Fig. 17 Allocating weight to each word

Each of the selected words were given weigh according to their positive or negative value so that using the values the reviews can be classified. Fig. 17 shows the value assigned to each of the selected words.

7) *Evaluation*



name	review	rating	item sales	year of sales
Vulli Sophie the Giraffe Teether ...	Great feel, great squeek, great quality, great ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	Sophie is one of my daughter's favorite t ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	Love it! Love it! Love it! The best teether I ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	We love Sophie at our house... she is a great ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	Had one for my first two children but it had seen ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	My 5mo daughter is completely in love with ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	All my grandbabies love this toy. It smells g ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	Being a child photographer, I saw lots ...	4	10000	2019
Vulli Sophie the Giraffe Teether ...	We just got our adorable Sophie this week and ...	5	25000	2019
Vulli Sophie the Giraffe Teether ...	This is the first review I've written for ...	4	10000	2019



word_count	awesome	great	fantastic	amazing	love	horrible	bad	terrible	awful	wow
{'forever': 1.0, 'saving': 1.0, 'll': ...	0.0	5.0	0.0	0.0	0.0	0	0.0	0	0	0
{'love': 3.0, 'she': 1.0, 'wonderful': 1.0, ...	0.0	0.0	0.0	0.0	3.0	0	0.0	0	0	0
{'this': 1.0, 'before': 1.0, 'others': 1.0, ...	0.0	0.0	0.0	0.0	3.0	0	0.0	0	0	0
{'so': 1.0, 'too': 1.0, 'which': 1.0, 'natural': ...	0.0	3.0	0.0	0.0	1.0	0	0.0	0	0	0
{'parents': 1.0, 'these': 1.0, 'great': 1.0, ...	0.0	1.0	0.0	0.0	2.0	0	0.0	0	0	0
{'baby': 1.0, 'gift': 1.0, 'slowly': 1.0, ...	0.0	1.0	0.0	0.0	2.0	0	0.0	0	0	0
{'because': 1.0, 'on': 1.0, 'tasts': 1.0, ...	0.0	1.0	0.0	0.0	2.0	0	0.0	0	0	0
{'though': 1.0, 'will': 1.0, 'expect': 1.0, ...	0.0	1.0	0.0	0.0	2.0	0	0.0	0	0	0
{'paint': 1.0, 'rubber': 1.0, 'from': 1.0, ...	0.0	1.0	0.0	0.0	2.0	0	0.0	0	0	0
{'worth': 1.0, 'few': 1.0, 'little': 1.0, ...	0.0	1.0	0.0	0.0	2.0	0	0.0	0	0	0

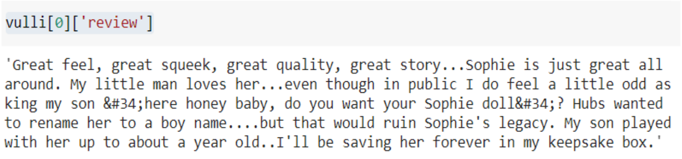
hate	sentiment	predicted_sentiment
0	1	0.996501414362946
0	1	0.9955670136189387
0	1	0.9955670136189387
0	1	0.9949584554334676
0	1	0.9927399728173647
0	1	0.9927399728173647
0	1	0.9927399728173647
0	1	0.9927399728173647
0	1	0.9927399728173647
0	1	0.9927399728173647

[723 rows x 19 columns]

Fig. 18 Evaluating one product – Vulli Sophie the Giraffe Teether

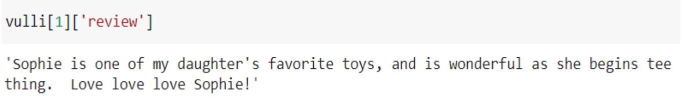
At last, reviews for one product are separated and analyzed for classification. Based on the given weight values for the selected words and also based on their repetition, the reviews are classified into positive and negative reviews. The sentimental study is illustrated in Figure 18, and the top two favourable and critical ratings are as follows.

8) *Most positive review*



```
vulli[0]['review']
```

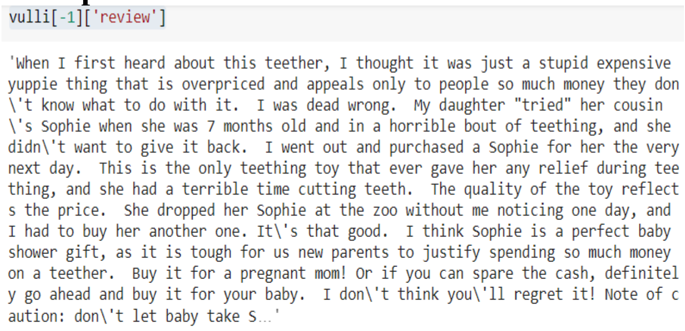
```
'Great feel, great squeek, great quality, great story...Sophie is just great all
around. My little man loves her...even though in public I do feel a little odd as
king my son &#34;here honey baby, do you want your Sophie doll&#34;? Hubs wanted
to rename her to a boy name....but that would ruin Sophie's legacy. My son played
with her up to about a year old..I'll be saving her forever in my keepsake box.'
```



```
vulli[1]['review']
```

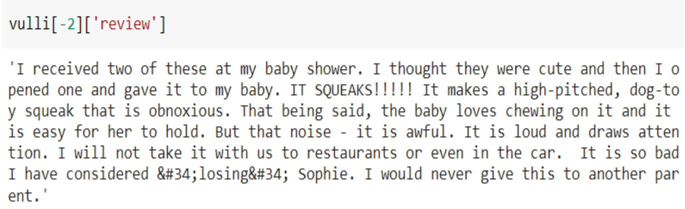
```
'Sophie is one of my daughter's favorite toys, and is wonderful as she begins tee
thing.  Love love love Sophie!'
```

9) *Most negative review*



```
vulli[-1]['review']
```

```
'When I first heard about this teether, I thought it was just a stupid expensive
yuppie thing that is overpriced and appeals only to people so much money they don
\'t know what to do with it.  I was dead wrong.  My daughter "tried" her cousin
\'s Sophie when she was 7 months old and in a horrible bout of teething, and she
didn\'t want to give it back.  I went out and purchased a Sophie for her the very
next day.  This is the only teething toy that ever gave her any relief during tee
thing, and she had a terrible time cutting teeth.  The quality of the toy reflect
s the price.  She dropped her Sophie at the zoo without me noticing one day, and
I had to buy her another one. It\'s that good.  I think Sophie is a perfect baby
shower gift, as it is tough for us new parents to justify spending so much money
on a teether.  Buy it for a pregnant mom! Or if you can spare the cash, definitel
y go ahead and buy it for your baby.  I don\'t think you\'ll regret it! Note of c
aution: don\'t let baby take S...'
```



```
vulli[-2]['review']
```

```
'I received two of these at my baby shower. I thought they were cute and then I o
pened one and gave it to my baby. IT SQUEAKS!!!!! It makes a high-pitched, dog-to
y squeak that is obnoxious. That being said, the baby loves chewing on it and it
is easy for her to hold. But that noise - it is awful. It is loud and draws atten
tion. I will not take it with us to restaurants or even in the car.  It is so bad
I have considered &#34;losing&#34; Sophie. I would never give this to another par
ent.'
```

VI. CONCLUSION

The researchers finalised the statement that to handle enormous volume of data, the business organisations require an intelligent sales prediction system. Sales forecasting process is very complex because there are lots of factors that should be taken into consideration. To implement achievable goals and successfully implement them, supermarkets chains always want to forecast sales. For sales forecasting in this analysis, three machine learning algorithms (Linear Regression, Decision Tree, and Random Forest) were used, RF performed better, as it had a lower mean absolute error and a higher accuracy than the other two models. It is also observed that getting more data would generally increase the predictive power of the models.

There have been made several attempts for review classification till date. This paper proposes a general framework to classify reviews. Sentiment analysis, also known as opinion mining, is a branch of science that explores people's behaviours, thoughts, and sentiments toward particular individuals. Now-a-day's technology is growing day by day and there are numerous website and application available within the online market by which they sell their product. Every product contains hundreds of reviews and on basis of these reviews, user buys the product most of the time. The Logistic Regression algorithm will help the user to pay for the right product by classifying the product's reviews.

Huge records are used in our experiments to compare algorithms. Since handling such a vast number of documents is challenging and time-consuming, some records were discarded during the review process. The fields and attributes used in the study were inadequate for further analysis. It was the most daunting hurdle encountered during the study. Big data can be used as a method for predictive analytics in sales forecasts to speed up the latest studies. Big data analysis and forecasting are considered essential areas of today's business world.

References

[1] Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine Learning in Predicting Demand for Fast-Moving Consumer Goods: An Exploratory Research. IFAC-PapersOnLine, 52(13), 737-742.

[2] Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

[3] Thomassey, S. (2010). Sales forecasts in clothing industry: The key success factor of the supply chain management. International Journal of Production Economics, 128(2), 470-483.

[4] Park, B., & Bae, J. K. (2015) Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. ELSEVIER - Expert Systems with Applications, 42(6), 2928-2934.

[5] Cadavid, J. P. U., Lamouri, S., & Grabot, B. (2018, July). Trends in Machine Learning Applied to Demand & Sales Forecasting: A Review. hal-01881362.

[6] Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Atak Bulbul, B., & Ekmis, M. A. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. Complexity, 2019

[7] Rising Odegua (2020) Applied Machine Learning for Supermarket Sales Prediction. Research gate.

[8] Beheshti-Kashi, S., Karimi, H. R., Thoben, K. D., Lütjen, M., & Teucke, M. (2015). A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering, 3(1), 154-161.

[9] Machine Learning, Tom Mitchell, McGraw Hill, 1997.

[10] Sepideh Paknejad (2018). Sentiment classification on Amazon reviews using machine learning approaches. diva2:1241547

[11] Ethem Alpaydin. (2004). Introduction to Machine Learning (Adaptive Computation and Machine Learning), The MIT Press.