Outlet Sales Analysis using R and Various Machine Learning Algorithms

Vidya Bhushan Singh and John Singh. K
School of Information Technology and Engineering,
Vellore Institute of Technology, Vellore, Tamil Nadu ,India
Email: pushkarnagbanshi@gmail.com

Abstract: In today's date data analysis is need for every data analytics to examine the sets of data to extract the useful information from it and to draw conclusion according to the information. Data analytics techniques and algorithms are more used by the commercial industries which enables them to take precise business decisions. It is also used by the analysts and the experts to authenticate or negate experimental layouts, assumptions and conclusions. In project ,prediction is based on item outlet sales by analyzing or exploring the Big mart sales data set. Inside the data frame or data set which we have used here , the counter variable "item outlet sales" is used by us for predictions. and, forecaster variables(eg-item weight, item visibility, item outlet count etc)are those by using which we can make forecasting on counter variable. various machine learning and data extraction models is considered for prediction are Linear regression, Decision tree, K-means , naïve bayes. Before prediction we have to explore and visualize the data because data exploration and visualization is an important stage of predictive modelling. The outcomes of this project is - High quality analysis, more flexibility and power as compared to others.

Keywords:R-Studio,Linear regression,Decision tree,K-means,Naive Bayes..

1. Introduction

As we know that in today's era data analysis is so important to everyone to make better decisions in their field. Analyzing the big data and extracting knowledge full information from the data is little bit tough. so, for mining of complex datasets we need a powerful and effective data mining tool to extract the information and take better decisions in future. We are using R here which is an open source free data mining tool and efficient too. R has several inbuilt packages which provides us efficiency like-ggplot2, VIM etc. R is an open-source data analysis environment and programming language .The process of converting data into knowledge, insight and understanding is Data analysis, which is a critical part of statistics. For the effective processing and analysis of big data, it allows users to conduct a number of tasks that are essential. R consists of numerous ready-to-use statistical modeling algorithms and machine learning which allow users to create reproducible research and develop data products. Although big data processing may be accomplished with other tools as well, it is when one steps on to the data analysis that R really stands unique, owing to the huge amount of built-in statistical formulae and third-party algorithms available. To create a powerful and reliable statistical model, data transformation, evaluation of multiple model options, and visualizing the results are essential. This is the reason why the R language has proven so popular: its interactive language uplifts exploration, clarification and presentation. Revolution R Enterprise gives the big-data support and speed to allow the data scientist to repeat through this process quickly.The dataset here is of Big Mart Sales-first phase of this analysis is to impute the missing values in dataset if any ,then comes the data exploration in which we have to explore the dataset and find the relation between various commodities in the big mart sales dataset, data exploration is an important key of predictive model. We cannot create a good predictive model until we know how to explore the dataset from beginning to last. This phase creates a solid base for data manipulation. Then comes the phase of using various machine learning algorithms to predict the outlet sales (response variable) like k- means, linear regression, multiple regression, random forest, decision tree , naïve bayes etc. The next step is to compare the prediction result of ,on basis of that we will get the outlet sales for each outlet present in the city.

2. Literature Survey:-

The author's of this paper[1] has main motive to analyze the electricity consumption data of Iran for the year 1991 to 2013 to predict the electricity consumption rate from 2014 to 2020.After research authors found that the growth rate of consumption from year 1999 to 2006 it was 73.53% where as from year 2006 to 2013 it was28.41%.After using regression model the result of prediction shows that the electricity consumption rate would increase by 22.28% by 2020 as compare to 2013.

Tanu jain, AK Sharma [2] interprets that the algorithms which are frequently used in the field of association rule mining are Eclat and Apriori (market basket analysis) algorithms. Both of these algorithms are mainly used for mining of primarily data sets and to find fraternity(associations) between these regular data sets using R which is a domain based language for data exploration, analysis and analytics. Several packages and libraries of R has been used by the authors to examine the performance of Eclat and Apriori algorithms on different item sets on the basis of execution time taken by both of the algorithms.

Author's of this paper [3] uses one of the most effective data processing and analyzation tool which is known as R Studio to analyze the RF(Random Forest) and LDA(latent dirichlet allocation) algorithms based on the outcome of large data sets to come up with more improved results which provides help to predicts some outcome in advance.

Nikhita awasthi and Abhay bansal [4] analyzes two advance machine learning algorithms known as ANN(Artificial Neural Network) and SVM(Support Vector Machine) on soil dataset to find hidden facts, information, various patterns etc.

Author's of this paper [5] uses the list of top 10 machine learning algorithms to observe the influence of these algorithms which was published by IEEE in ICDM(International Conference on Data Mining) in the month of December 2006.List contains top 10 data mining algorithms which are mostly used by the research community are as follows:k-means algorithm, support vector machine, Apriori algorithm, Naive Bayes, Nearest neighbour(kNN), Decision tree, Expectation-Maximization(EM)algorithm, pageRank, AdaBoost, Eclat algorithm.

Author's [6] basically interprets that what is data analysis and how we can do it efficiently?.In this paper author recommends R for data analysis because of its tremendous capability of data exploration, several inbuilt packages, easy to implement several machine learning algorithms etc. As we know that R is a statistical language as well as programming language which helps in effective model prediction and better visualization techniques. So after survey authors found that the with R data analysis is much more efficient.

Hilda,jurgen and joseph [7] analyzes and compares three famous mechanisms or tools of data mining known as – Rapid Miner,Weka,R respectively to use their expertise in the area of structural health monitoring. This paper interprets several functionalities of R,Weka,Rapidminer in time series analysis for structural health monitoring like visualizing,filtering,applying statistical models etc.

Nitin, paul, nick and vasilious [8] provides an inexpensive approach for security analysis in big data by combining the functionalities of R and EmEditor using CI(Computational Intelligence) for desktop users of windows.This approach is applied on huge dataset and windows firewalls provides the data.The computational intelligence based security can also expand for other branches like web branches,application logs etc.

Author's of this paper [9] analyzes a free and emerging statiscal language known as R for mining big data . This paper provides information about implementing a clustering technique known as K-means with the help of R on a huge data set.In today's date data coming from several sources is immense and to mine such a huge data in not a simple task,but this paper shows us that we can do it efficiently using R.

The author's from Renmin university of china [10] explained about a system known as novel trigger system which would give better better prediction results as compared to single prediction model for various different types of items or products. After research the authors come at a point to conclude that the accuracy of novel trigger system is more than that of single prediction model. various enterprises can use this for better future prediction which would affect their sales.

3. Results and discussion concluded from literature survey:-

- As we can see in fig.2 where prediction is done from year 2014 to 2020,for year 2000 to 2006 the increase in consumption rate was 10% annually, where as it was 4% from 2007 to 2013.Hence from the prediction the authors came to point that the consumption rate will increase by 3.2% yearly from 2014 to 2020.

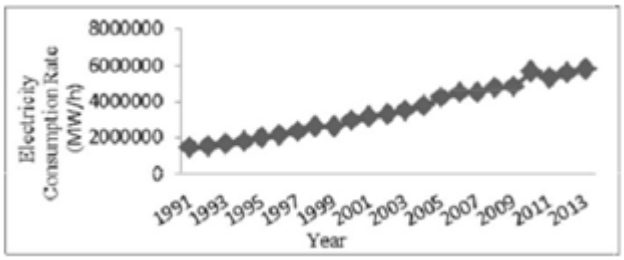

Fig 1: Electricity consumption rate from 1991 to 2013

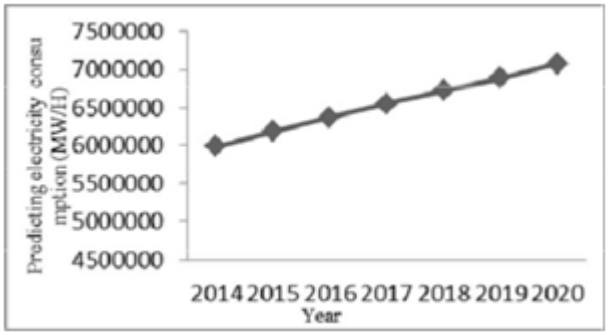

Fig 2: Predicted electricity consumption rate from 2014 to 2020

- From the table and diagram given below we can observe that the execution time of Eclat algorithm decreases as the records in the data sets increases as compare to the Apriori algorithm.

Data sets	Apriori	Eclat
5000	0.11	0.06
10000	0.13	0.12
20000	0.25	0.24
40000	0.56	0.47

Table 1: Execution time of éclat and apriori(ref-quant. analysis of apri. & éclat algo.)

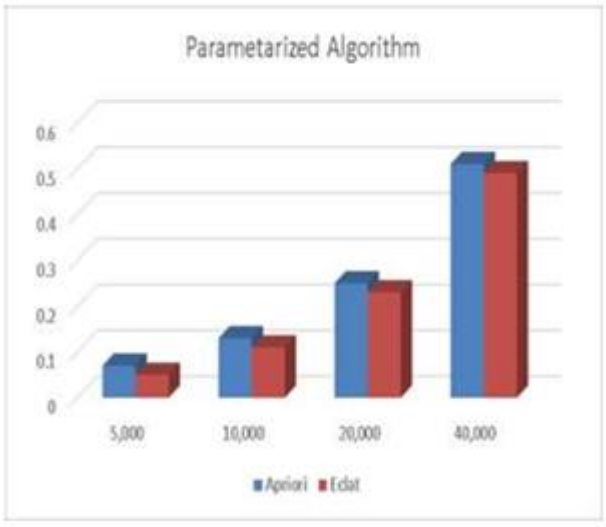

Fig 3:Execution time of éclat and apriori algo. for diff. datasets

- After observing various case studies and from the table given below we come to the conclusion that the accuracy of Random Forest algorithm is better than Latent Dirichlet algorithm.

Random forest	Latent Dirichlet Allocation
Confusion matrix & statistics reference, prediction N0N345 2 O 3.37	Confusion matrix & statistics reference, prediction N0N329 39 O 19.0
Accuracy:0.9871	Accuracy:0.8501
95% CI:(0.9701,0.9958)	95%CI:(0.8106,0.8842)
No information rate:0.8992	No information rate:0.8992
p-value[Acc>NIR]:1.972e-12	p-Value [Acc>NIR]:0.9991
Kappa:0.9295	Kappa:-0.0707

Table 2: RF vs LDA(ref:-Big data predictive analysis)

- So from the below tables we can observe that the highest prediction percentage for ANN model is 55% and 74% for SVM model.

	Prediction percentage	Training steps	RMS
ANN with 1 node	48	3378	31.34
ANN with 5 node	52	61268	19.45
ANN with 7 node	55	73073	15.8

Table 3:comparison of ANN models(ref:Application of data mining classification technique on soil data)

Kernel type	Prediction(%)	Support vectors	Training error
Polynomial	68	597	0.35
Radial basis	74	543	0.2
Hyperbolic tangents	43.50	519	0.62

Table 4: comparison of SVM models(ref: Application of data mining classification technique on soil data)

- Fig.4 shows us three different colors – sky blue,blue and red which means it shows us three different clusters of three different cateogries of petal length and petal width of iris dataset.

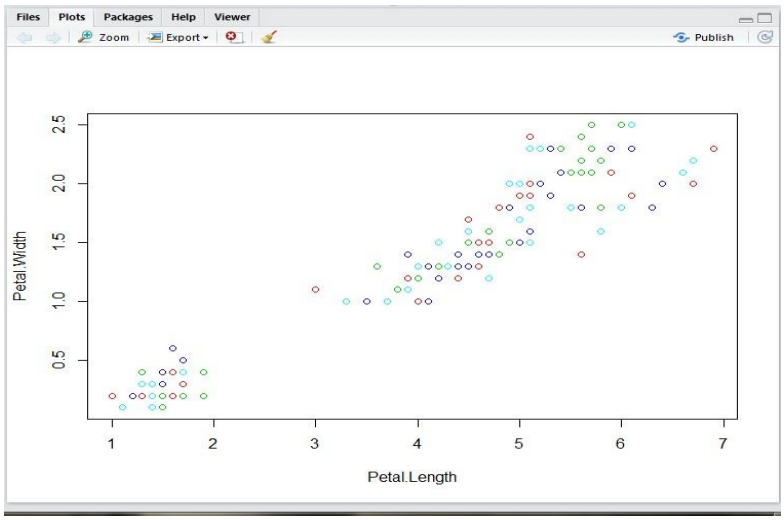

Fig-4: K-means clustering on iris dataset

4. Proposed Work:-

So, we have taken the dataset of big mart sales on which we are going to do the operations like exploration , visualization, cleaning, imputation and prediction after importing the dataset in R-Studio. After importation we have to explore and visualize the dataset to find out missing values and irregularities in it. If , the irregularities and missing values are present in the dataset then we have to clean it and impute the data at the place of missing value , after that we can apply several data mining algorithms for prediction.

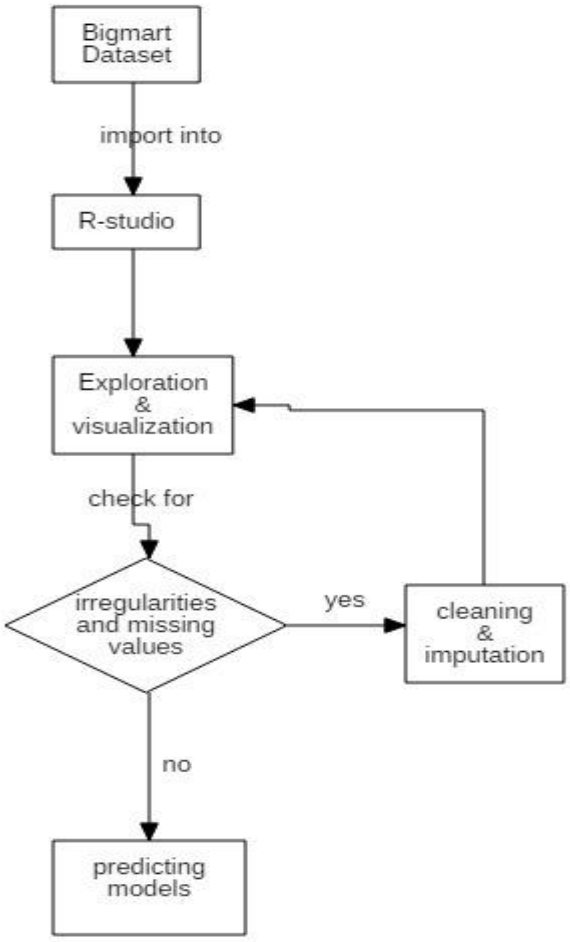

Diagram 1: Architecture diagram

R is a programming language as well as statistical language and it is a very powerful language for data exploration, visualization and prediction. It provides the facility of various inbuilt functions which makes our job easier.

company name	use of R
Facebook	For behavior analysis
Google	For advertising effectiveness
Twitter	For data visualization
Microsoft	xbox matchmatching service
Uber	For statistical analysis
IBM	Joined R Consortium Group
ANZ	For credit risk modeling

Table 5:use of R

a) Exploration and visualization:-

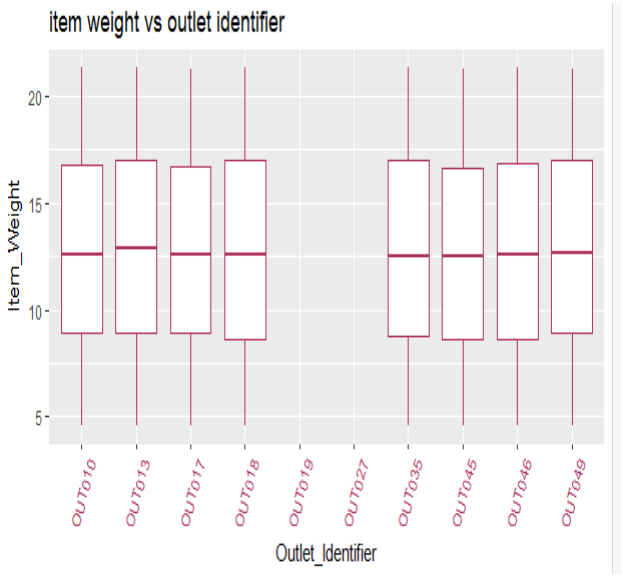

Fig 5: item weight vs outlet identifier

- As it is clear from above diagram that outlet out019 and out0027 does not have any item weight,so we have to impute those missing values of item weight for the two outlets 0027 and 0019.

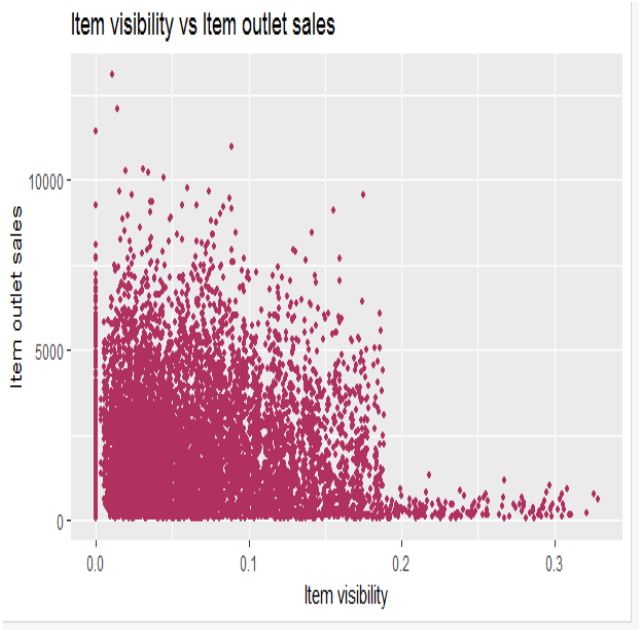

Fig 6.Item visibility vs Item outlet sales

- As we can see item visibility less than 0.2 have majority of sales and visibility more than 0.2 have minority of sales.

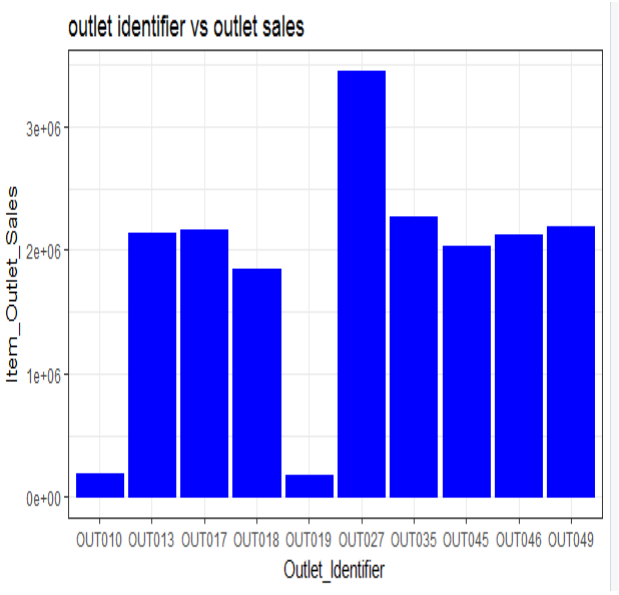

Fig 7: outlet identifier vs outlet sales

- As from the above figure we can analyze that the OUT027 contributes most to the total sales followed by OUT35 and so on.

b] Cleaning and imputation

- As it is clear from the fig.5 that outlet out019 and out0027 does not have any item weight,so we have to impute those missing values of item weight for the two outlets 0027 and 0019.

- If we go through the train and test dataset we get that the test data has 11 columns whereas train data has 12 columns.

- So,we have to add one extra column with value 0ne(1) in test dataframe that is missing.

  >Test_market$Item_Outlet_Sales<-1
  >View(Test_market)

- Inorder to simplify I am going to combine the Train_market and Test_marketdataframes.

  >combine<-rbind(Train_market,Test_market)
  >View(combine)

- Now, if we look into the combine data we can see that the column outlet size has NA(empty) values.so,for now i am going to impute “any” for the empty cells.

  >levels(combine$Outlet_Size)[1]<-"any"

- So,we can see that only Item_Weight has NA values,after imputing in column Outlet_Size.

  >combine$Item_Weight[is.na(combine$Item_Weight)]<-
  >median(combine$Item_Weight[!is.na(combine$Item_Weight)])

- So,here I have imputed the missing values in column item weight by median of non missing values.

  >table(is.na(combine$Item_Weight))
  >table(is.na(combine))

- After that we analyzes that there is irregularities in the column iten fat content.

```
>levels(combine$Item_Fat_Content)

>library("plyr", lib.loc="~/R/win-library/3.4")

>combine$Item_Fat_Content<-
revalue(combine$Item_Fat_Content,c("LF"="Low Fat" ,"low
fat"="Low Fat","reg"="Regular"))

>table(combine$Item_Fat_Content)
```

- If,we go through the data again,we will see that the item visibility column has value “0”,which is not acceptable.so,we have to impute median value for all the zeroes in item visibility.

```
>combine$Item_Visibility<-
ifelse(combine$Item_Visibility==0,median(combine$Item_Visibility
),combine$Item_Visibility)
```

- Here we are going to create a dataframe fatcount which will show the fat level for each item type.

```
>fatcount<-
as.data.frame(setNames(aggregate(combine$Item_Fat_Content,by=li
st(Category=combine$Item_Type,Category=combine$Item_Fat_Con
tent),FUN=length),c("Item_Type","Item_Fat_Content","total")))

>View(fatcount)
```

- After running fatcount we will able to analyze that the item type like “household”, ”others”,”health & hygiene” also has fat level like low fat,regular etc, which makes no meaning to them.so, I am going to introduce a new fat level category “none” for these item types.

```
>levels(combine$Item_Fat_Content)<c(levels(combine$Item_Fat_C
ontent), "None")

>combine[which(combine$Item_Type=="Health and Hygiene")
,]$Item_Fat_Content<- "None"

>combine[which(combine$Item_Type=="Household"),]$Item_Fat_
Content<- "None"

>combine[which(combine$Item_Type=="Others")
,]$Item_Fat_Content<- "None"
```

```
>combine$Item_Fat_Content<-factor(combine$Item_Fat_Content)

>library(plyr)

>library(dplyr)
```

dplyr package makes data manipulation quite effortless

- count of outlet identifiers

```
>a <-combine%>%group_by(Outlet_Identifier)%>%tally()

>names(a)[2] <-"Outlet_Count"

>head(a)

>combine <- full_join(a, combine, by ="Outlet_Identifier")
```

- count of item identifier

```
>cnt<-combine%>%group_by(Item_Identifier)%>%tally()

>names(cnt)[2]<-"item_count"

>head(cnt)

>combine<-merge(cnt,combine,by="Item_Identifier")
```

- age in year of every outlet

```
>combine$year<-as.factor(2017-
combine$Outlet_Establishment_Year)
```

- “DR”, are mostly eatables. “FD”, are drinks. And, “NC”, are non consumable
- substr(), gsub() function to extract and rename the variables respectively.

```
>newtype<-substr(combine$Item_Identifier,1,2)
>newtype<-gsub("FD","Food",newtype)
>newtype<-gsub("DR","Drinks",newtype)
>newtype<- gsub("NC","NonConsumable",newtype)
>table(newtype)
```

- Let's now add this information in our data set with a variable name 'Item_Type_New.

  >combine$Item_Type_New <-newtype

- drop variables not required in modeling
  >library(dplyr)
  >combine <- select(combine, -c(Item_Identifier, Outlet_Identifier, Outlet_Establishment_Year))

- divide data set
  >new_train_market <- combine[1:nrow(Train_market),]
  >new_test_market <- combine[- (1:nrow(Train_market)),]

b) List of implemented algorithms

1. Linear Regression

2. Decision tree

3. K-Means

4. Naïve Bayes

1. Linear Regression:-

Linear regression falls under the cateogry of supervised learning and it is one of the simplest algorithm in the field of machine learning which is used for building the relationship between independent variable(x) and dependent variable(y).

Working algorithm:-

Here we are considering item outlet sales as dependent variable and other variables as independent variable.

After calculating the correlation of dependent variable with every independent variable in the dataset we come to the conclusion that the item mrp has strong correlation with item outlet sales as compared to other independent variables.

i)Item mrp vs Item outlet sales

>cor.test(new_train_market$Item_Outlet_Sales,new_train_market$Item_MRP)

>linear_model<lm(new_train_market$Item_Outlet_Sales~new_train_market$Item_MRP)
so,the equation for linear regression is=>y=9.4892x+(-19.4130)

y=9.4892x-19.4130,where x=item mrp,slope=9.4892,intercept=-19.4130.

so,increase of 1 rs. in item mrp leads to increase of rs. 9.4892 in item outlet sales.

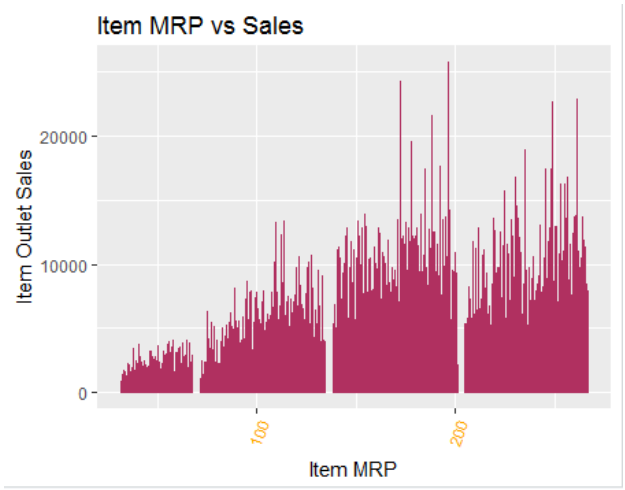

Fig.8:item mrp vs sales

So,we can analyze that if item mrp will increase item outlet sales will also increase and vice versa.

ii)Item visibility vs Item outlet sales

>model2<lm(new_train_market$Item_Outlet_Sales~new_train_market$Item_Visibility)

equation->y=mx+c
y= -2833.27x+1521.35
where x=item visibility,slope= -2833.27,intercept=1521.35

so,we can conclude that every 0.1 visibility increment in the item visibility after 0.2 leads to decrement in item outlet sales by -2833.27.

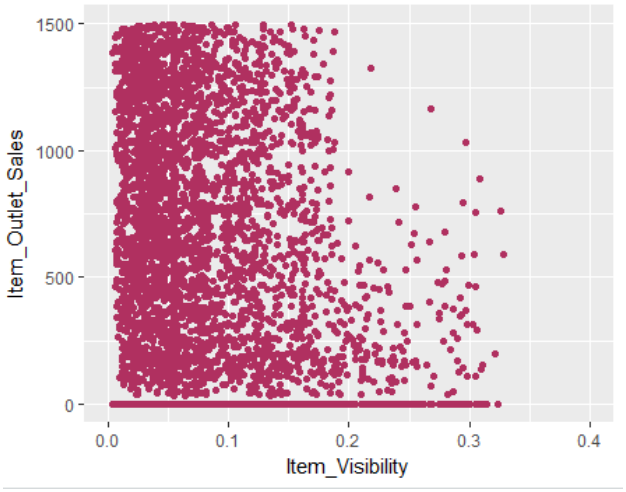

Fig.9:item visibility vs item outlet sales

We can observe from the above figure that if item visibility will be more than 0.2 then item outlet sales will decrease.

Calculation of RMSE(root mean square error):-

>linear_model<- lm(log(Item_Outlet_Sales) ~ ., data = new_train_market)

>rmse(new_train_market$Item_Outlet_Sales,
exp(linear_model$fitted.values))

hence,rmse= 2091.777

we can see that the root mean square we get here is high so, we are going to implement decision tree to look that root mean square error will improve or not.

2] Decision tree:-

Decision tree, as we can see that its name contains tree which means , it is a tree-shaped structure used to take specific decisions to finds out solution of several complex scenarios. From the figure 10 we can easily observe its structure like a probability tree which will help us to take appropriate decisions.

Working algorithm:-

```
>fit <- trainControl(method = "cv", number = 5)
>cart_Grid <- expand.grid(.cp= (1:50)*0.01)

#decision tree >
>tree_model <- train(Item_Outlet_Sales ~ ., data = new_train_market,
```

17

```
method = "rpart", trControl = fit, tuneGrid = cart_Grid)

>print(tree_model)

>tree <- rpart(Item_Outlet_Sales ~ ., data = new_train_market, control =
rpart.control(cp=0.01))
>prp(tree)
```

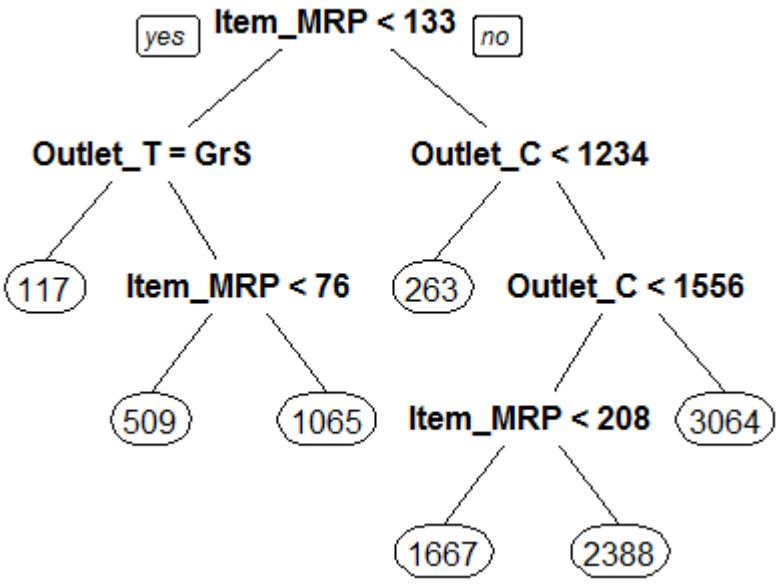

Fig 10:Decision tree

We can now able to take decisions based on this decision tree. It shows that the variable item mrp plays a major role in predicting item outlet sales.

Calculation of RMSE:-

```
> score <- predict(tree, type = "vector")
> rmse(new_train_market$Item_Outlet_Sales, score)
```

hence,rmse= 1520.095

so, the root mean square error we get here is 1520.095 which is better than that of linear regreesion, but still it is too high. So for further improvement we can try other models like random forest etc, which we are leaving for future work.

3] K-means:-

K-means is a clustering technique of type unsupervised learning which is used to group elements of same characteristic into a category or cluster.It is also known as Lloyd's algorithm which is implemented for partioning of data into K clusters defined according to centroids.

Working algorithm:-

```
>newframe<data.frame(new_train_market$item_count,new_train_market$
Outlet_Count,new_train_market$Item_Weight,new_train_market$Item_V
isibility,new_train_market$Outlet_Type)

>copy<-newframe
>copy$new_train_market.Outlet_Type<-NULL
>pred<-kmeans(copy,3)
>View(pred)
>pred

>table(newframe$new_train_market.Outlet_Type,pred$cluster)
```

	1	2	3
Grocery Store	**0**	**521**	**550**
Supermarket Type1	**5584**	**0**	**0**
Supermarket Type2	**931**	**0**	**0**
Supermarket Type3	**937**	**0**	**0**

From the above result we can clearly analyze that all the supermarket type 1 ,type 2, type 3 respectively belongs to cluster 1,whereas for grocery store it is approx 50-50 belongs to cluster 2 and cluster 3 respectively.

4] Naive Bayes:-

Naive bayes is a classification algorithm which uses Baye's theorem to classify various objects.It uses probability theory for classification of data.In this algorithm events probability can be adjusted, if in future a new data will be added.

Working algorithm:-

```
>naive<data.frame(new_train_market$item_count,new_train_market$Outlet_Count,new_train_market$Item_Weight,new_train_market$Item_Visibility,new_trainmarket$Item_MRP,new_train_market$Item_Outlet_Sales,new_train_market$year,new_train_market$Item_Fat_Content)

>naive$new_train_market.Item_Fat_Content<ifelse(naive$new_train_market.ItemFat_Content=="Regular",1,0)

>set.seed(1234)
>vid<-sample(2,nrow(naive),replace=T,prob = c(0.8,0.2))
>train<-naive[vid==1,]
>test<-naive[vid==2,]
>naive$new_train_market.Item_Fat_Content<as.factor(naive$new_train_


market.Item_Fat_Content)
>model<-naive_bayes(new_train_market.Item_Fat_Content~.,data=naive)

#prediction

>predic<-predict(model,train,type='prob')
>head(cbind(predic,train))

#confusion matrix for train dataset

>predic1<-predict(model,train)
>(tab1<-table(predic1,train$new_train_market.Item_Fat_Content))
```

predic1	0	1
0	3999	2803
1	17	13

```
#misclassification

>1-sum(diag(tab1))/sum(tab1)
[1] 0.4127635

#confusion matrix for test dataset

>predic2<-predict(model,test)
>(tab2<-table(predic2,test$new_train_market.Item_Fat_Content))
```

predic2	0	1
0	997	686
1	4	4

#misclassification

>1-sum(diag(tab2))/sum(tab2)
[1] 0.4080426

So, the misclassification error for train data is nearly 41% and for test data it is nearly 40% which is a big misclassification error, for improvement we have to find the correlation between every independent variable and if correlation is high ,we have to drop those variables, so for now we are leaving it for future work.

Conclusion-

In this paper, we have used four data mining algorithms-linear regression, decision tree, k-means, naive bayes. We have implemented linear regression for finding out the relation of item outlet sales with other variables like item mrp , item visibility etc. for prediction purposes and that we have calculated the root mean square error which is equal to 2091.777 for linear regression. To improve this root mean square error(r m s e) and to make decisions we have implemented a decision tree and after implementation the r m s e reduces to 1520.095. As we can see that still the root mean square error is high, we can implement random forest algorithm, éclat algorithm etc which we will do in our future work. K-means is also used by us for clustering the dataset according to their categories and at last we have implemented naive bayes classifier for variable item fat content .The main motive of this paper is to show you how to tackle or deal with such giant dataset which has missing values as well as regularities.

References:-

1] Wenjie H., Qing Z., Wei X., Hongjiao F., Mingming W., and Xun L. 'A Novel Trigger Model for Sales Prediction with Data Mining Techniques', Data Science Journal, 14: 15, 2015,pp. 1–8.

2]Pooja K., A.R. K., 'International Journal of Computer Applications', National Seminar on Recent Trends in Data Mining 2016 ,pp.0975 – 8887.

3] Eric T.,Manish G.,Praveen K,⁎,1,' 'The Role of Big Data and Predictive Analytics in Retailing', journal of retailing 93,2017,pp.79-95.

4] Nitin N., Paul J., Nick S. and Vasilios , 'Big Data Security Analysis Approach Using Computational Intelligence Techniques in R for Desktop Users', 2016,pp.1-16.

5] Hilda K., J¨u., Josef K., 'Using R, WEKA and RapidMiner in Time SeriesAnalysis of Sensor Data for Structural Health Monitoring', IEEE comp. society 2011,pp.1529-4188.

6] Noorollah K., 'Analysis and predicting electricity energy consumption using data mining techniques- A case study I.R. Iran -Mazandaran province', 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA 2015) March 11-12, 2015,pp-4799-8445.

7]SanchitaP.,'Big Data Analytics Using R',International Research Journal of Engineering and Technology (IRJET)e-ISSN: 2395 -0056 Volume: 03 Issue: 07 | July-2016 ,pp.78-81.

8]Xindong W., Vipin K., J. Ross Q., Joydeep G., Qiang Y., Hiroshi M.,Geoffrey J.,McLachlan ,Angus Ng ,Bin g L.,Philip S. Y., Zhi-Hua Z., Michael S., David J. Hand · Dan Steinbe r., 'Top 10 algorithms in data mining', Knowl Inf Syst (2008) 14: pp.1–37 DOI 10.1007/s10115-007-0114-2.

9] Tanu Jain* Dr. A.K Dua Varun Sharma, 'Quantitative Analysis of Apriori and Eclat Algorithm for Association Rule Mining' International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 4 Issue 10 Oct 2015, Page No. 14649-14652.

10]NIKHITA A., ABHAY B., 'APPLICATION OF DATA MINING CLASSIFICATION TECHNIQUES ON SOIL DATA USING R' International Journal of Advances in Electronics and Computer Science, ISSN:pp. 2393-2835 Volume-4, Issue-1, Jan.-2017.

11] Priyanka P., Kavita S., Oza Ass., K. Kamat., 'Big Data Predictive Analysis:Using R Analytical Tool',International conference on I-SMAC 2017,pp.839-842