**Machine Learning Techniques on BigMarketSales Data and Prediction of Sales of A Store**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | **PrabuddhaDasgupta**  Undergraduate Student  Computer Science and Engineering (Data Science)  School of Engineering and Technology –Jain University  Bangalore, Karnataka, India  [pdgproloy26@gmail.com](mailto:pdgproloy26@gmail.com) | **Atishay Jain**  Undergraduate Student  Computer Science and Engineering (Data Science)  School of Engineering and Technology –Jain University  Bangalore, Karnataka, India  [crazydiggerati@gmail.com](mailto:crazydiggerati@gmail.com) | **JitendraJaiswal**  Assistant Professor  Computer Science and Engineering (Data Science)  School of Engineering and Technology –Jain University  Bangalore, Karnataka, India  [jitendra.jaiswal@jainuniversity.ac.in](mailto:jitendra.jaiswal@jainuniversity.ac.in) | **Tanvir H Sardar**  Assistant Professor  Computer Science and Engineering (Data Science)  School of Engineering and Technology –Jain University  Bangalore, Karnataka, India  [tanvir.sardar@jainuniversity.ac.in](mailto:tanvir.sardar@jainuniversity.ac.in) | |  |  |

**Abstract**

The machine learning techniques are being implemented in big data for carrying number of analysis and predictions with different perspectives. Different machine learning algorithm may not be giving similar performance on a single dataset and the accuracy of prediction may differ. In this paper we have considereda big market sales data to analyze the performance of different ML techniques. Retail is a developing industry and data analytics is used to optimize business processes tasks. The purpose of this research is to analyze the sales of each product at a particular store by following different techniques and comparing the effectiveness of each algorithm. We will find the sales of a product by using various ML algorithms like linear regression and Random forestand then finding the accuracy of the prediction by comparing the results.

**Keywords:** Machine Learning, Data Analytics, Prediction, Big Market Sales, Regression.

**INTRODUCTION:-**

Machine Learning has been defined in various ways by different scientists, the most common definition of machine learning as coined by Tom Mitchell is “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. Machine learning can also be referred to as the implementation of statistics using programming**[7]**. Machine learning can be basically divided into various categories such as supervised learning, unsupervised learning and reinforced learning. The directory on which we are working here consists of labeled data and we would basically implement supervised learning algorithms. We will be implementing regression techniques using the machine learning algorithms. Here, we are taking Big Market sales data that is structured, and we will be working with that dataset and find the sales of a store. We have performed market forecasting which relates sales and market predictions. Sales prediction allows companies to predict achievable sales revenue, efficiently allocation of resources and a planning for the future growth of a product in the market. Usually what marketers do is they prefer web analytics and take a poll from various sources and then they analyze the data. A dataset is gathered in from web source (2013) and is used for sales prediction using Machine learning (ML) algorithms. We have used machine learning due to its ability to learn automatically from the dataset using an algorithm and improve them without any explicit coding. ML helps the algorithm to access data and uses it for their own personal learning. It is also much preferred because the predictions are much more accurate and provides a better precision towards certain prediction **[8]**.

**REVIEW OF LITERATURE:-**

There are many works reported in the literature using the algorithms specified. A few of the recent and novel works are reviewed and briefed below:

Bohdan **[1]** proposed machine-learning models for sales predictive analytics. The main goal of this work is to considered different machine-learning approaches for time series forecasting. The effect of machine-learning generalization has been considered. This effect can be used to make sales predictions when there is a small amount of historical data for specific sales time series in the case when a new product or store is launched. A stacking approach for building regression ensemble of single models has been studied. The use of regression approaches for sales forecasting can often give better results compared to time series methods. One of the main assumptions of regression methods is that the patterns in the historical data will be repeated in future. The results show that using stacking techniques, we can improve the performance of predictive models for sales time series forecasting.

Mikael and Halldén **[2]** investigate to find possibility to create a forecasting solution based on supervised learning. Two different methods are tested, Extreme Gradient Boosted Trees and Long Short Term Memory Neural Network. The two methods are evaluated against each other and compared to the current uplift model used by Caspeco. The data used for training the supervised learning methods is a combination of data provided by Caspeco, and data collected from the Swedish Meteorological and Hydrological Institute (SMHI). This is data such as temperature, minutes of sunshine, rainfall etc, all of which are known to have an impact on the sales of a restaurant. The results show that the models are dependent on the settings of the restaurants.

Wang and Li **[3]** applied machine learning algorithm into a real world problem – drug store sales forecasting. Given store information, and sales record we applied Linear Regression, Support Vector Regression(SVR) with Gaussian and Polynomial Kernels and Random Forest algorithm, and tried to predict sales for 1-3 weeks. Root Mean Square Percentage Error (RMSPE) is used to measure the accuracy. As it turned out, Random Forest outshined all other models and reached RMSPE of 12.3%, which is a reliable forecast that enables store managers allocate staff and stock up effectively. Among all models, Random Forest works the best, and provides a reliable prediction of the sales. Linear regression, SVR with Gaussian/Polynomial Kernels and RF all have their own strengths and limitations.

In the work of Yusuf and Alawneh **[4]**, the sales data is analyzed and predictions are made by using linear regression as implemented on the GPU to make the process faster. Sales forecasting is made possible by finding best fit line by linear regression techniques (i.e. linear convolution). To illustrate this process, simulated sales data was used. The sales forecasting with linear regression implementation using GPU was compared to the CPU implementation and a speedup of up to 7.557 xs was achieved.

In **[5]**, Jain et al., presents a use case of data mining for sales forecasting in retail demand and sales prediction. In particular, the Extreme Gradient Boosting algorithm is used to design a prediction model to accurately estimate probable sales for retail outlets of a major European Pharmacy retailing company. The forecast of potential sales is based on a mixture of temporal and economical features including prior sales data, store promotions, retail competitors, school and state holidays, location and accessibility of the store as well as the time of year. The model building process was guided by common sense reasoning and by analytic knowledge discovered during data analysis and definitive conclusions were drawn. The performances of the XGBoost predictor were compared with those of more traditional regression algorithms like Linear Regression and Random Forest Regression. Findings not only reveal that the XGBoost algorithm outperforms the traditional modeling approaches with regard to prediction accuracy, but it also uncovers new knowledge that is hidden in data which help in building a more robust feature set and strengthen the sales prediction model.

In honors thesis Yuchen Lin **[6]** have created different statistical models and used machine learning techniques for sales prediction.

**ENVIRONMENT AND THE ALGORITHMS USED:-**

The Machine Learning techniques will be implemented in RStudio that is basically scientific programming software

In this paper we will focus on the sales of each product at a particular store. We will find this by implementing two different machine learning algorithms like Linear Regression and random forest and perform the comparison of the results accordingly. For this we need a clear concept of Machine Learning algorithms and its mathematical implementation.

1. **Linear Regression**.

Linear regression is one of the mostly frequently used machine learning algorithms and works quite accurately and provides a relationship model between the input and output data. Linear regression has been studied since the late 1800s and has been researched from every possible ways and has been well known in various names.

The representation of linear regression combines specific set input variables (x) to which we have to find the predicted output (y) keeping both the datasets as numeric. In a single linear regression model the form of the model will be

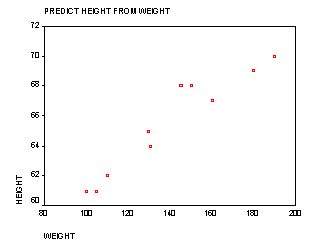
y= B0+B1x

When there is more than one input then the line is called a plane or a hyper-plane.

Example:- for example we are predicting weight (y) from height (x). Our linear regression model representation would look like

weight=B0+B1height

Here B0 is the bias co-efficient and B1 is the co-efficient of height. Let’s take B0 as 0.1 and B1 as 0.5 and then find the weight of a person when the height is given (Let height= 182). Thus the weight of the person would be 91.1. Thus we can put a bunch of heights and we can get the weights and thus we can get a graph as given below **[9].**



***(image courtesy – google images)***

1. **Random Forest**

Random forest or random decision forest is a method that operates by constructing multiple decision trees during training phase. The decision of the majority of the trees is chosen by the random forest as the final decision. The main question that arises is why we use random forest:

We use random forest for the following reasons:

1. No overfitting – use of multiple trees reduces the risk of overfitting and thus the training time is less.
2. Higher accuracy - runs efficiently on large database. For large data, it produces highly, accurate predictions.
3. Estimates missing data - Random forest can maintain high accuracy when large proportion of data is missing.

Random forest is a collection of decision trees and decision trees are a kind of predictive models

where there involves partitioning of data into subsets that contains instances with similar values.

Example- Suppose ehave a customer dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Age** | **Income** | **Gender** | **Interest** |
| Bob | 20 | $1000 | M | 1 |
| Susan | 40 | $600 | F | 0 |
| John | 60 | $1300 | M | 1 |
| Mary | 56 | $1100 | F | 0 |
| Josh | 39 | $1000 | M | 0 |
| Lauren | 28 | $200 | F | 1 |
| Kate | 31 | $1040 | F | 1 |

Here we have a small dataset. It has seven users it has similar attributes. It has a relationship

It has a relationship to one item that is interest. By 1 it means the user showed interest

and by 0 it means that the user didn’t. We will be applying the concept of random forest here.

The decision tree will be selecting the attribute that will maximize the difference in the results

of the interest attribute. So splitting will take place.

Interest

[outcomes: 1,0,1,0,0,1,1]

Age>35 ?

Susan(0)

John(1)

Mary(0)

Josh(0)

Bob(1),

Lauren(1)

Kate(1)

John(1)

Susan(0)

Mary(0)

Josh(0)

Income> $1200 ?

After doing this segregation we can build its predictive models by carrying various number of

hypothesis and then drawing a constructive inference.

**ANALYSIS AND RESULT:-**

For developing any prediction models we need to follow the given steps-

1. **Exploratory Data Analyis and data preprocessing**

First we need to understand and visualize the data using histograms, scatterplots, Boxplots, correlation matrices and seeing the head and tail attributes of the table. Then analysis of missing values and imputation will take place. Then from so many attributes we must select the best and the most important features.

1. **Feature Engineering**

We will be transforming and scaling features that includes identifying and deducing derived variables and also interaction variables.

1. **Mathematical modeling**

We will be building the mathematical models using machine learning in R and the model will be build using both linear based techniques (Linear Regression) and Tree based techniques (Random Forest).

1. **Hyperparameter tuning**

Hyperparameters are certain parameters whose values are set before starting the algorithm. These parameters cannot be learned using regular training process. These parameters express higher level properties of the model such as its complexity and the learning rate of the algorithm. In hyperparameter tuning we will be adjusting these parameters. Ex - number of leaves or depth of the tree in random forest.

1. **Ensembling**

Ensembling methods provides higher accurate predictions rather than from constituent learning algorithms alone. In ensembling we are stacking the regressor with linear regression and random forest **[10]**.

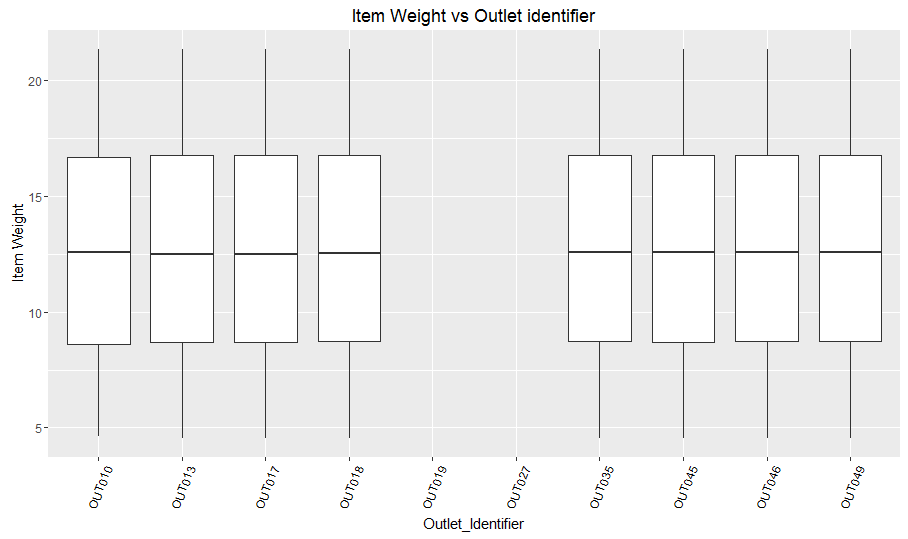
**Data Exploration and Preparation**

Fat Content

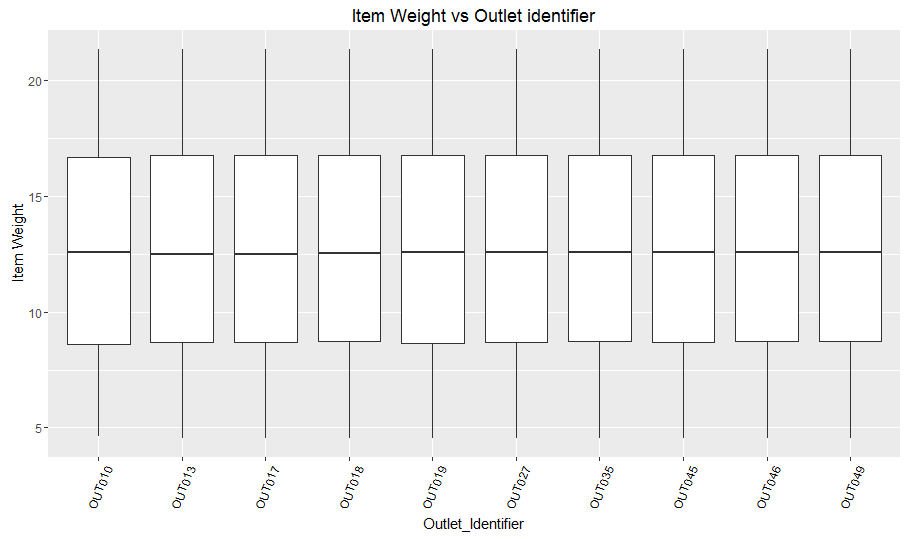
The original data contain five different levels for the fat content: LF, low fat, Low Fat, reg, and Regular. Clearly, LF, low fat, and Low Fat are the same, as are reg and Regular. Hence, we replace LF and low fat by Low Fat and reg by Regular. Further, certain types of non-consumables, i.e. those in the categories Health and Hygiene, Household and Others are either Low Fat or Regular according to the data. Clearly, this makes no sense. Hence, we introduce a new fat level None for non-consumables.

Item Weights

Now we are checking for the missing values in this attribute and we see that 2439 entries are missing in the category Item\_Weight. From the boxplot we can see that OUT019 and OUT027 have not reported any weight data.

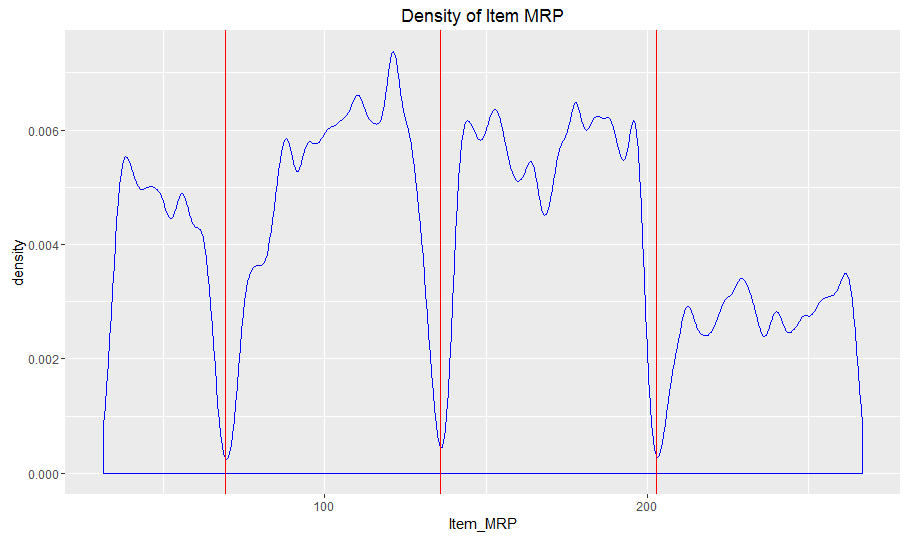


Thus we have to work with this missing values and we taking the values from other stores as we see that all the wares are available elsewhere also. Thus after successfully working on this we get.



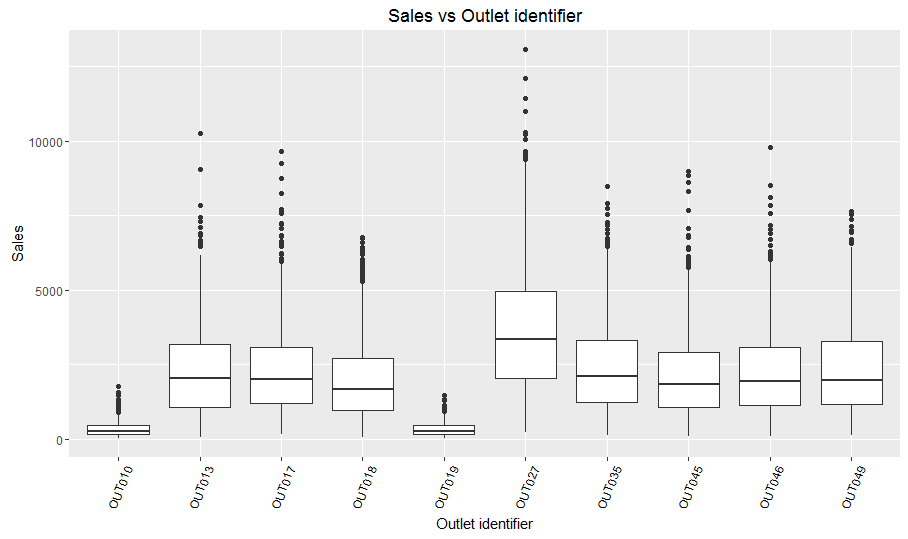
Density of the price

Looking on the plot we can see that there are four categories of the prices as seen below. To differentiate between the factors we introduce four factors low, medium, high, very high.

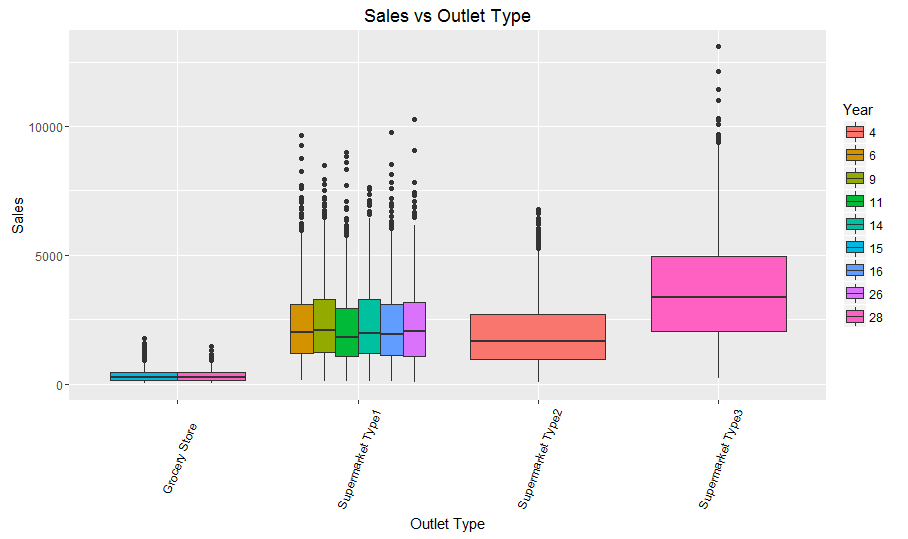


Sales and outlet

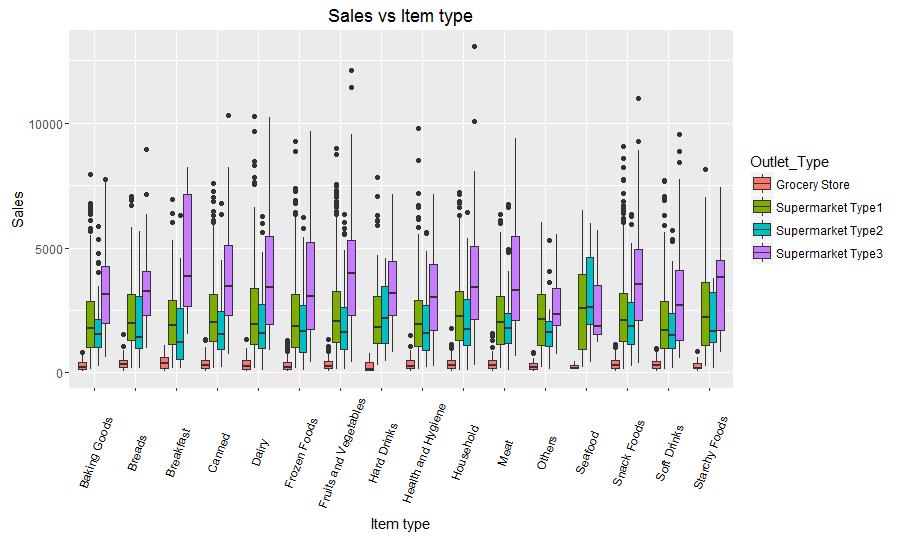
We need to explore the sales with respect to various outlets, we need to count how many sales were reported in each outlet. If we compare the outlet sales we can see that OUT10 and OUT19 has fewer sales compared to others.

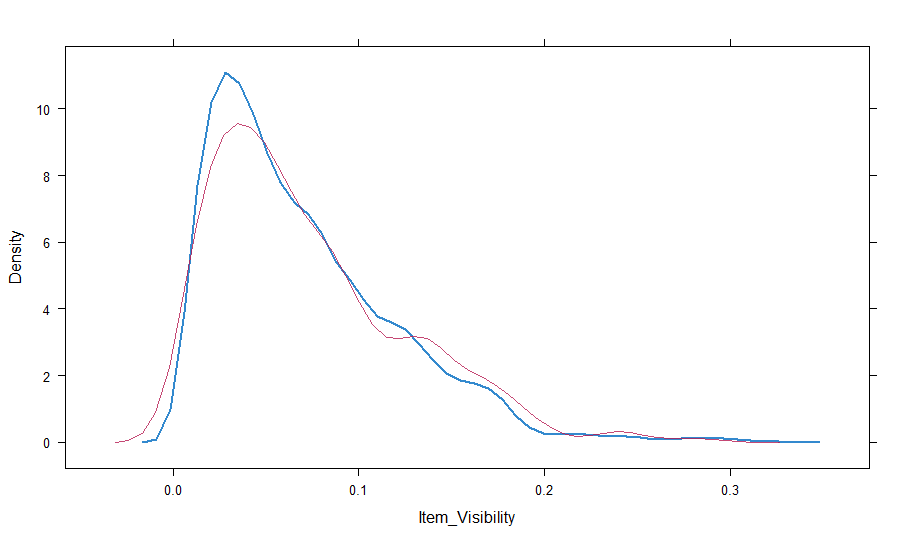


When we see the sales vs. its outlet type we can see that. With the given below graph we can explore the year of sale, the type of store from where the sales where predicted and the sales growth.

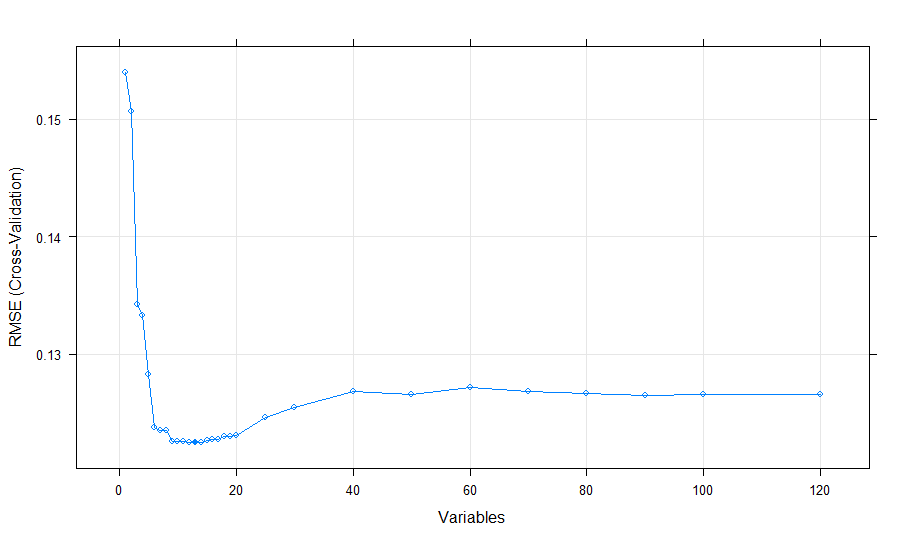


Now we will see the sales vs. the item type graph that represents how much the sales and the type of products sold and from which type of market.



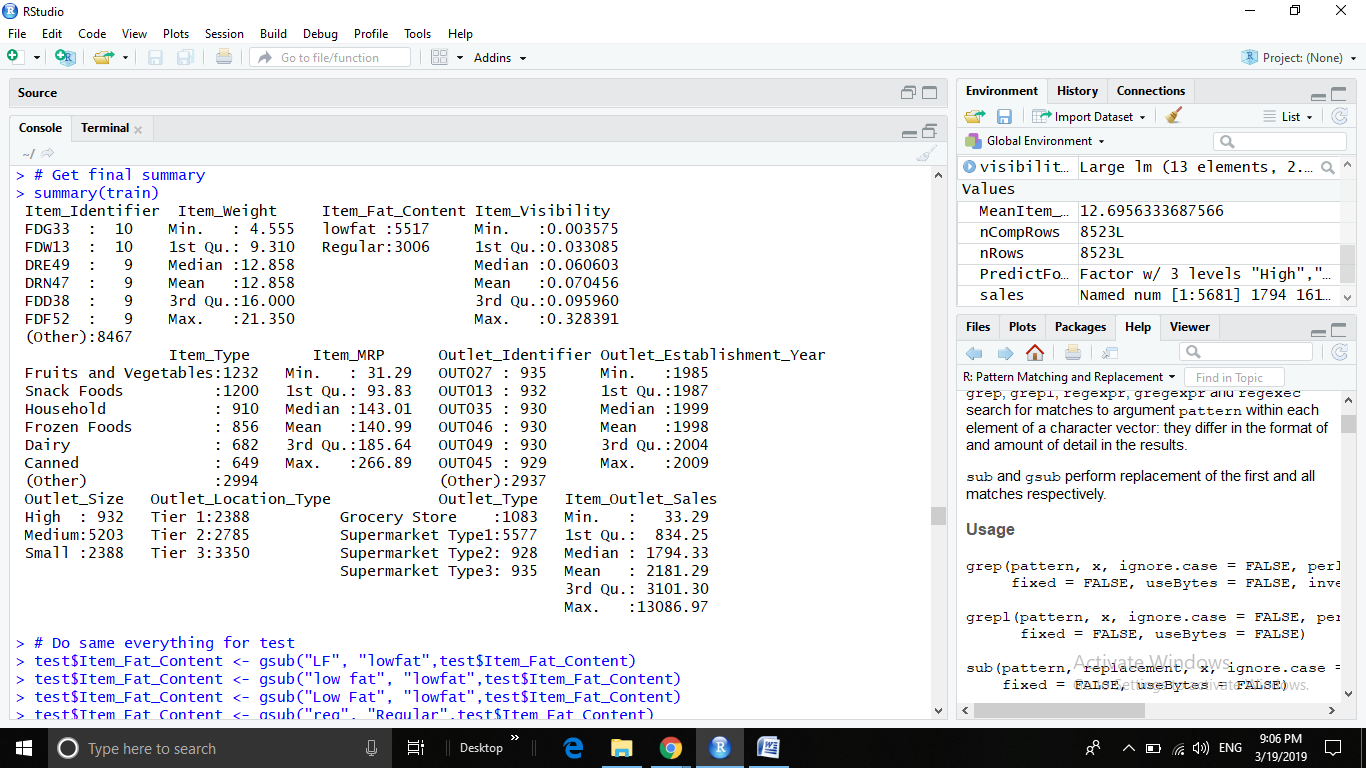
When we compare the zero and the imputed values then we can see our output ie. the given values vs. the predicted values graph as shown below (with the help of linear regression) -

With the help of random forest or random feature elimination we get the result-



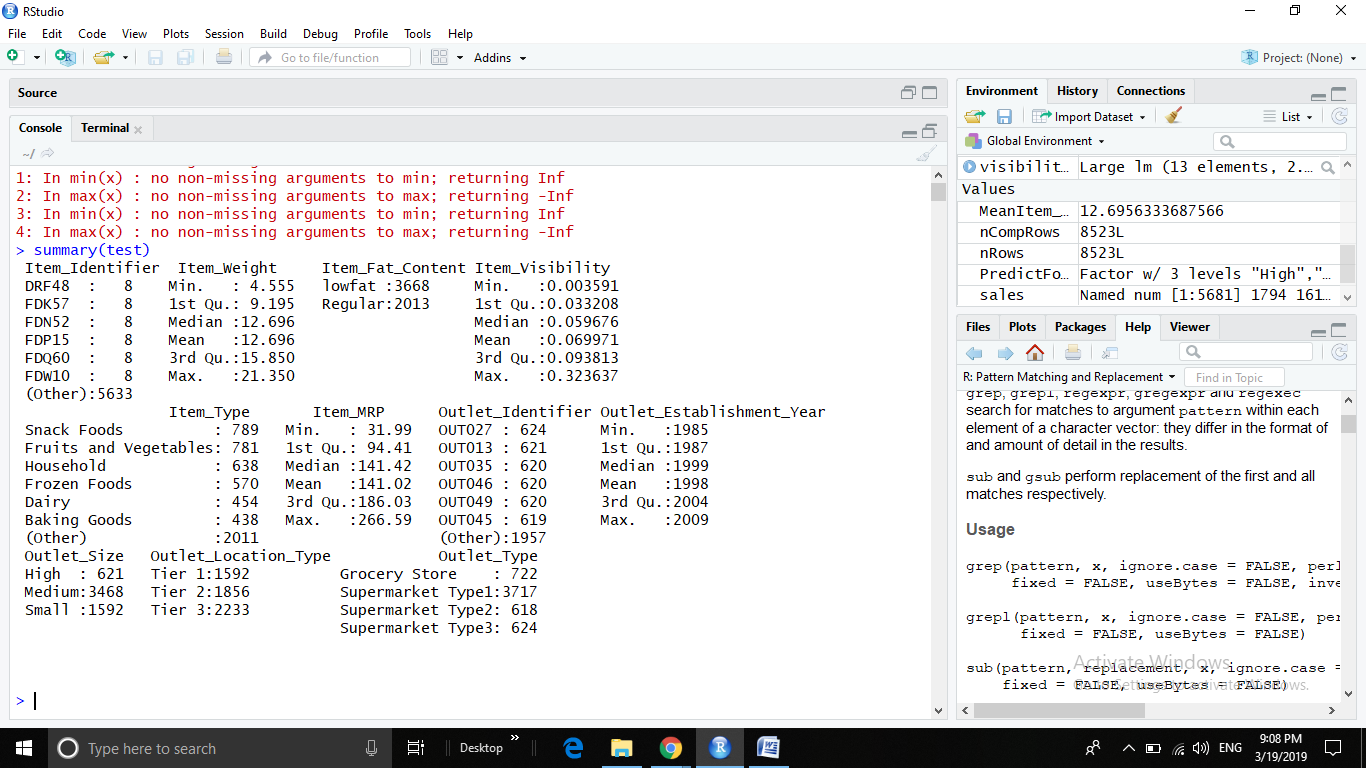
The training data summary:-

In the training data we have 8523 observations of 12 variables.

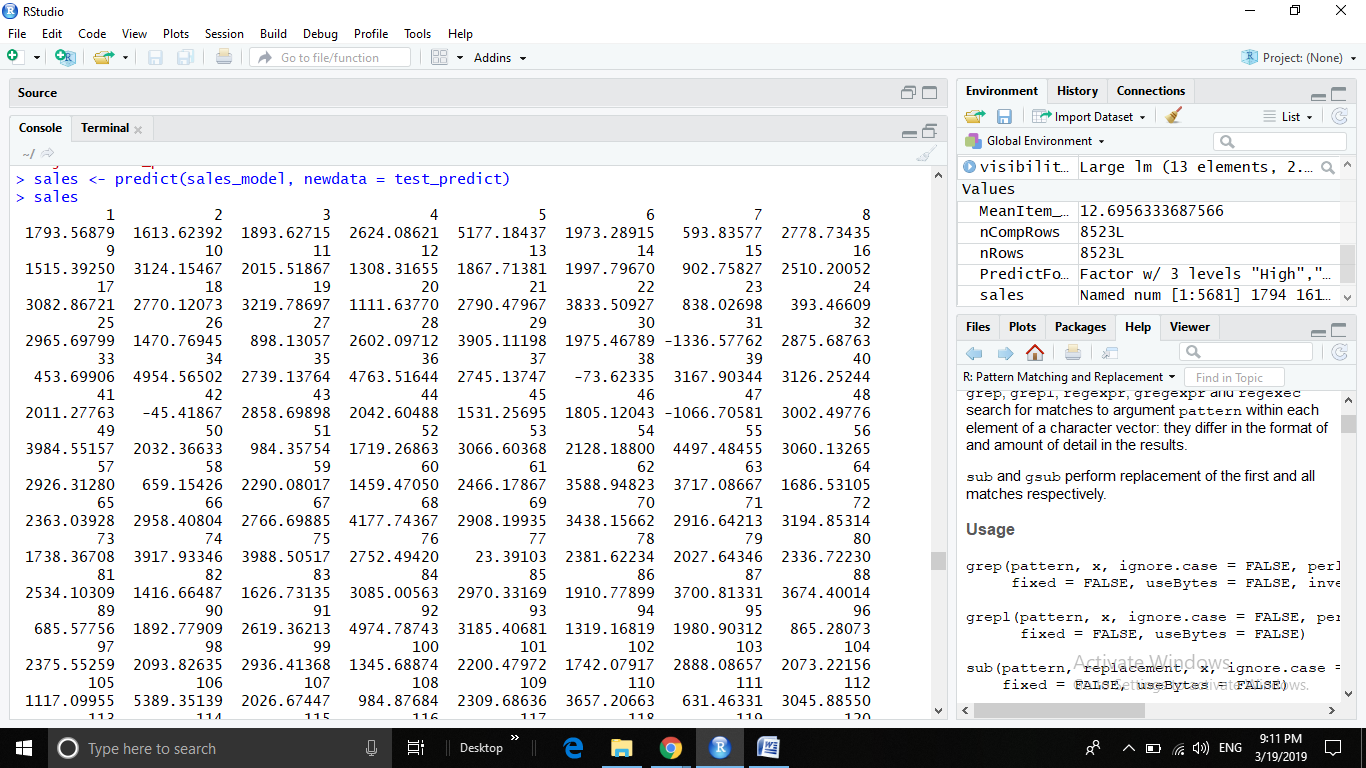


The testing data summary:-

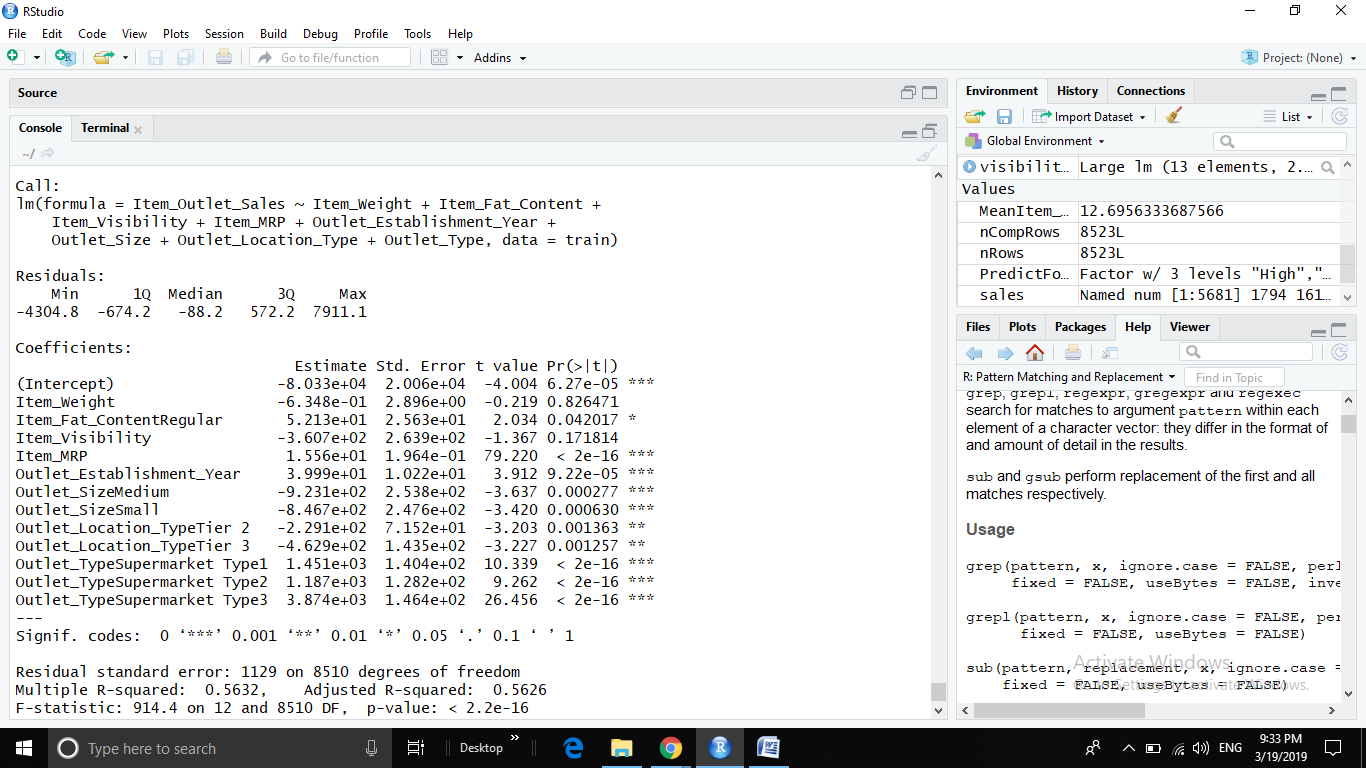
In the testing data we have 5681 observations of 11 variables.



sales(up to 112 observation) :-



The sales model:-



**CONCLUSION:-**

In order to find a decent model to predict sales we performed an extensive search of various machine learning models available in R. In this analysis we have analyzed the data from both train and test models and have trained the algorithm with some training values. This development have been done with the use of the ml algorithms, we have also worked accordingly for the missing values and then opted the final summary. Thus from the above results we can draw inferences that when it comes to data extrapolation the linear model gives better results than tree based algorithms and provides a better score, it can also be used for anomaly detection due to this property of data extrapolation. Thus in this model Linear Regression model will be giving much better result than random forest

**REFERENCES:-**

1. **[1]** Pavlyshenko, Bohdan M. "Machine-Learning Models for Sales Time Series Forecasting." *Data* 4.1 (2019): 15.
2. **[2]** Holmberg, Mikael, and Pontus Halldén."Machine Learning for Restaurant Sales Forecast." (2018).
3. **[3]** Chenghao Wang, Yang Li, “Drug Store Sales Prediction”, Report by Stanford(2016), online at [cs229.stanford.edu/proj2015/216\_report.pdf](http://cs229.stanford.edu/proj2015/216_report.pdf)
4. **[4]** Ayomide Yusuf, ShadiAlawneh, “GPU Implementation of Sales Forecasting with Linear Regression”, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), Volume-6, Issue-4, July-2018 DOI: 10.21276/ijircst.2018.6.4.1, ISSN: 2347-5552.
5. **[5]** Jain, Ankur, Manghat Nitish Menon, and Saurabh Chandra."Sales Forecasting for Retail Chains." (2015):.<https://pdfs.semanticscholar.org/76a2/44f4da1d29170a9f91d381a5e12dc7ad2c0f.pdf>
6. **[6]** Auto Car Sales Prediction: A Statistical Study Using Functional Data Analysis and Time Series. Honors Thesis by Yuchen Lin, University of Michigan. <https://deepblue.lib.umich.edu/bitstream/handle/2027.42/112123/lycumich.pdf?sequence=1&isAllowed=y>
7. **[7]**  <https://machinelearningmastery.com/what-is-machine-learning/>
8. **[8]** <https://bigdata-madesimple.com/machine-learning-sales-forecasting-tackle-insufficient-data-issue/>
9. [**9]** <https://www.statisticssolutions.com/what-is-linear-regression/>
10. **[10]** <https://medium.com/diogo-menezes-borges/project-1-bigmart-sale-prediction-fdc04f07dc1e>
11. Reference :-
12. <https://www.quora.com/When-are-random-forests-RFs-better-than-linear-regression-models-Do-RFs-fit-nonlinear-data-better-from-the-get-go-without-a-plotting-transform-dance-Are-there-other-cases-where-I-should-use-RF-instead-of-trusty-old-linear-models>
13. <https://www.quora.com/How-does-random-forest-work-for-regression-1>
14. <https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/>