CS634 Data Mining

Big Mart Sales Data Prediciton

Param Trivedi
Computer Science
New Jersey Institute Of Technology

Jash Parekh
Computer Science
New Jersey Institute Of Technology

***Abstract*—This report showcases the comparison between Random Forest, Linear Regression and Decision tree models and gives an overview of the data mining tasks that were performed.**

I. INTRODUCTION

In today's world the need for new products and better-quality products are increasing at a rapid rate. Supermarket or grocery store are becoming a go to place to access these products and as a reason it highly needed that these supermarket chains be able to forecast future sales to make better decisions which would lead them to higher profits. This report focuses on predicting the sales of products which are located at 10 different outlets which belongs to the Big Mart chain.

II. DATA PREPROCESSING

A. *Data Description*

The Big Mart sales data consists of 8523 rows and has 12 variables. The variables are described in the Table 1.1:

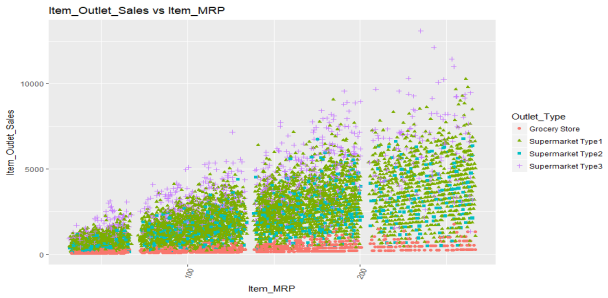
Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

Table 1.1

B. *Insights*

1. Item Outlet Sales vs Item MRP

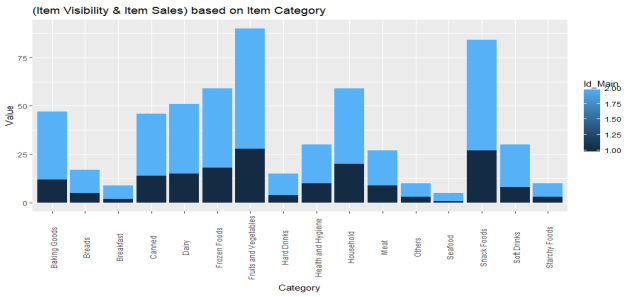
It can be seen in graph 1.1 that in sales Supermarket Type 1 dominates compared to other type of outlets. However, it is interesting to see the gaps in the prices around 60,130 and 200. There could be many reason for why there are gaps in the prices, it could be because prices for different categories differ and which led to the gaps.



Graph 1.1 Item Outlet sales vs Item MRP

2. Item Visibility and Item Sales vs Item Type

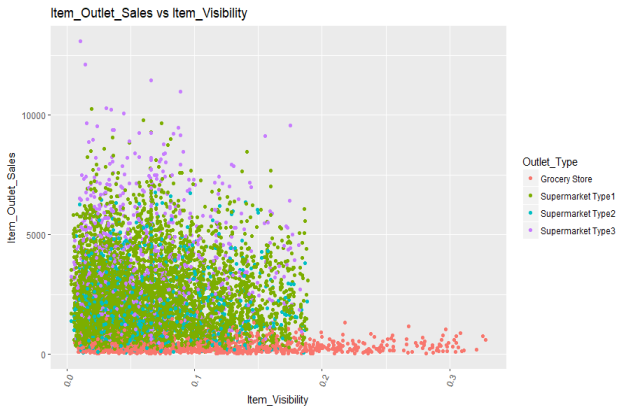
Graph 2.2 indicates that when Item Visibility and for a item category is directly proportional to aggregate sales for that particular category. Those category, who had the highest visibility had the highest sales. This finding could help sell those categories which have lower sales and have high MRP. Categories like dairy could benefit as they are highly consumed by general population and have a varied range of prices.



Graph 1.2 Item Visibility vs Item Type

3. Item Visibility vs Item Outlet Sales

Our observations in graph 1.2 do not hold when we see at the sales of Items in different outlets. Here the grocery store does have a uniform distribution even for items which have higher visibility. Even for Supermarket Type 3 some of Items with high sales have lower visibility compared to other items.



Graph 1.3 Item Outlet Sales vs Item Visibility

C. *Data Cleaning.*

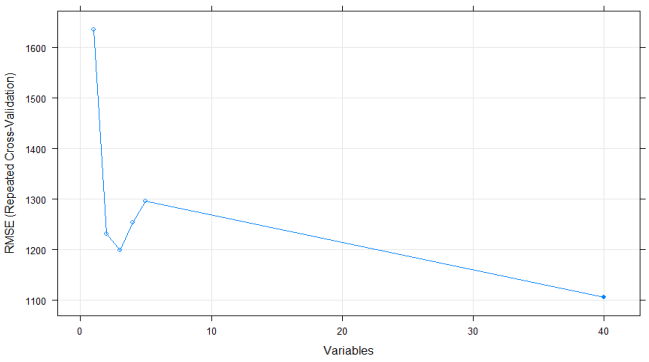
Before we implement the algorithms on the data we have make sure that data is clean so that we can get appropriate results. In the Big mart sales data following columns shoes missing values: "*Item_Weight*", "*Outlet_Size*". Following steps were taken to get rid of the missing values. For "*Item_Weight*" we replace the missing values with mean of the column and for "*Outlet_Size*" we replace the values with the mode of column.

D. *Feature Engineering*

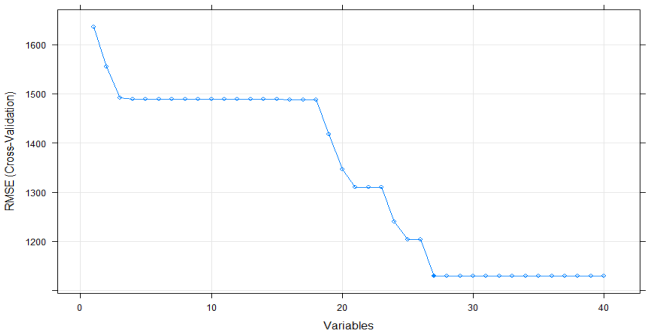
Attribute "*Item_Fat_Content*" is a categorical attribute which had two categories: Low Fat and Regular. However, the data had Low Fat, low fat, LF, reg and Regular which were then renamed Low Fat and Regular respectively. Attribute "*Outlet_Establishment_Year*" did not had much intuitive meaning and hence it was replaced with how old the store is. This might help us determine better sales because if store is

E. *Recursive Feature Engineering*

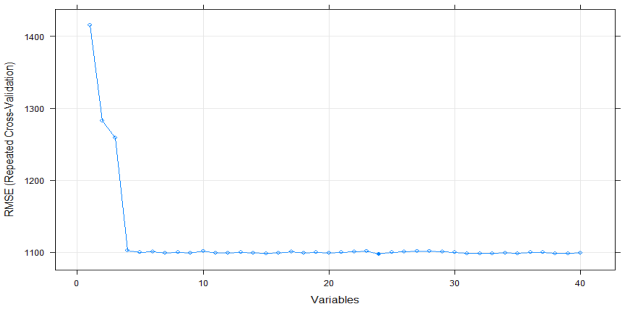
This algorithm is one of the popular methods for feature selection. This method, ranks the importance of attributes and could help us determine which attributes should be eliminated. This method creates subsets of data where each subset contains attributes number from 1 to n and desired algorithm is implemented (E.g. Random Forest). Graph 2.1 ,2.2 and 2.3 the performance of attributes for Random Forest, Linear Regression and Decision Trees. Here the data is converted into One Hot Form and has 41 attributes. In this method, we have chosen RMSE values as metric to see the performance.



Graph 2.1 RFE for Random Forest



Graph 2.2 RFE for Linear Regression



Graph 2.3 RFE for Decision Trees

It can be seen from graph that we get the best score when all the attributes are taken into consideration.

III. ALGORITHMS

This section will showcase the implementation of Random Forest, Linear Regression and Decision tree method on the Big Mart sales data. There are two models

1. Predicts the Sales of Item for that store
2. Predicts the quantity of Item sold instead of Item Sales. The idea behind this implantation is that quantities sold might make more sense than the sale of item. In the final step while checking the accuracy Item Sold is multiplied with Item MRP.

A. *Decision Trees*

Decision trees is a machine learning technique that are used for classification and regression problems. The idea behind this algorithm goes in a top-down approach where you all the train cases at the node and then you split the tree in to branches until you the reach the leaf node. Decision tree uses Gini Index / Entropy to split the nodes. Gini Index measures the impurity of the attributes and choses the attributes which are the purest. The attribute with Gini score 0 is the purest.

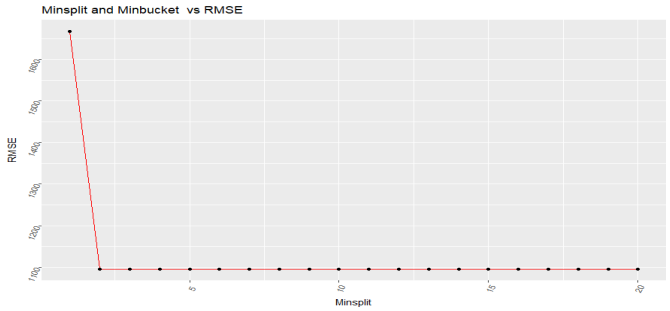
Parameter Adjustment:

A decision tree algorithm is applied on two models and we incremented the values of "*Minsplit*" (the minimum number of observations that must exist in a node in order for a split to be attempted) and "*Minbucket*" (Minimum number of observation at leaf node). The value of "*Maxdepth"* is set to 4. The value of "*Minbucket*" is shown in equation 1.

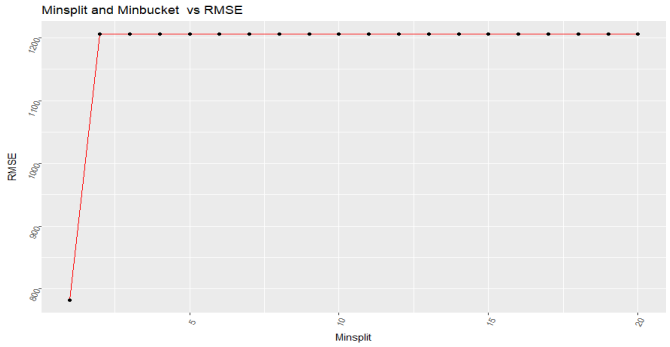
$$Minbucket = round(\frac{Minsplit}{3}) \quad (1)$$

Results:

Graph 3.1 and 3.2 shows the performance of Model 1 and 2 respectively. It can be seen in the graphs that Model 1 performance better as the values of "*Minsplit*" and "*Minbucket*" are increased, compared to Model 2. The optimal RMSE score for Graph 1 is 1096.45 and Graph 2 is 782.45. The value of parameter "*Minsplit*" and "*Minbucket*" for Model 1 are set to 20, and 7 and for Model 2 are set to 1 and 1.



Graph 3.1 Minsplit and Minbucket vs RMSE for Decision Tree for Model 2



Graph 3.2 Minsplit and Minbucket vs RMSE for Decision Tree from Model 2

B. *Random Forest*

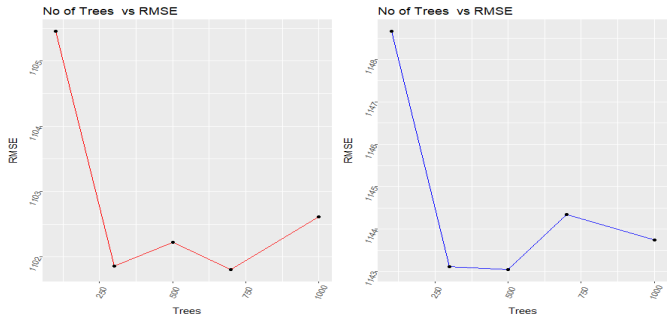
Random Forest is a popular ensemble learning method. As the name suggests it creates a forest of decision trees and out of those trees the one which has the highest majority is chosen as a final model which will be used for prediction. Random forest takes N attributes form the dataset and then it splits the data into edges, just like decision trees which uses Gini Index or Entropy to determine the split points Random Forest also considers those metrics to choose the best split point. It will create N number of trees with each tree is made on subset of data and in the end it calculates the votes that each tree has and chooses the one which has the majority votes.

Parameter Adjustment:

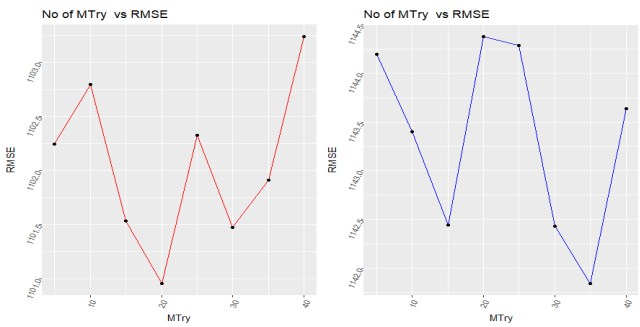
Random Forest model was implemented with different values of parameter like "*NTrees*" (No of decision trees), "*MTry*" (No of nodes that are randomly sampled at each split) and "*Maxnodes*" (Maximum no of nodes leaf nodes a tree can have).

Results:

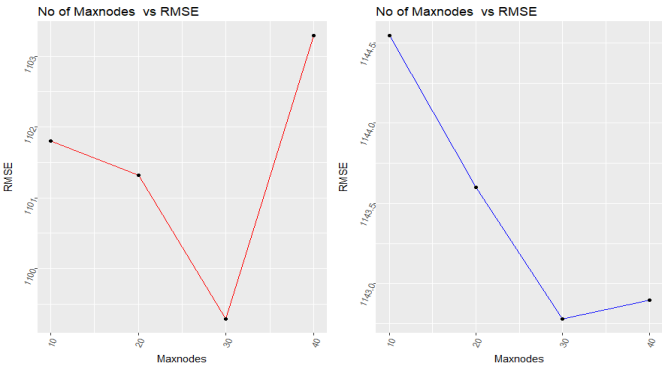
Graph 3.3, 3.4, 3.3 shows the plot of "*NTrees",* "*MTry*", "*Maxnodes*" for Model 1 and Model 2 vs RMSE.Random Forest algorithm was implemented for different values of "*NTrees",* "*MTry*", "*Maxnodes*" and the best score for Model 1 was 1099.288 which was obtained when "*NTrees",* "*MTry*" and "*Maxnodes*" are set to 700, 20 and 30 and for Model 2 was 1142.778 which was obtained when parameters value are set to 500, 35 and 30 respectively.



Graph 3.3 No of Trees vs RMSE for Model 1 (Left) and Model 2 (Right)



Graph 3.4 No of MTry vs RMSE for Model 1 (Left) and Model 2 (Right)



Graph 3.5 No of Maxnodes vs RMSE for Model 1 (Left) and Model 2 (Right)

C. *Multivariate Regression*

Multiple linear regression or commonly known as Multivariate regression is a popular technique for prediction. The main idea behind this is to visualize the data as an equation where each variable is assigned a weight.

$$y = \alpha + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon \qquad (2)$$

Equation 2 shows the generalized formula of multivariate regression where β is the assigned weights to the attributes. This algorithm has tendency to be biased toward the attributes with higher weights.

Results:

This algorithm was implemented on the Big Mart Sales data and as multivariate regression does not take in to account categorical attribute they were converted in to the "*onehot"* encoding format. The RMSE score for model 1 was 1136.461 and model 2 was 1094.424.

IV. Summary

To finally summarize the results of the three algorithms we can that the optimal RMSE score for Model 1 and Model 2 obtained using the Decision Trees, Random Forest and Multivariate Regression are 1096.45 and 782.45, 1099.288 and 1142.778, 1136.41 and 1094.424 respectively. We can say that for Model 1 Decision Tree performs better than other algorithms and for Model 2 Multivariate Regression performance is better compared to other algorithms.

After implementing the algorithm's, we submitted the solution for the Test data set that was provided by Analytics Vidya Big Mart Sales Competition. Random Forest performed better for Model 1 and Model 2 with RMSE score of 1152.916 and 1145.759

References

[1] https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3518362/

[3] https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674

[4] Hands-On Machine Learning with Scikit-Learn and TensorFlow By Aurelien Geron. ISBN 9781491962299