**BAN 671**

**Data Analytics with R**

**Sales Analysis**

**of a Retail Chain: Big Mart**

**Team Members:**

Kartik Amirneni

Ashish Solanki

Muhtasin Rahman

Juili Shete

Aditya Singh Dixit

California State University- East Bay

Hayward, California

**ABSTRACT**

This project aimed to analyze product level details, store specifics and annual sales information of a retail grocery chain named Big Mart across 10 of their stores across the country to increase sales. After exploring this data which needed proper cleaning and manipulation, through the analysis of this transformed data and building a range of predictive models such as regression and decision tree using R, we tried to examine what specific factors contribute to the sales of certain products and to what degree their significance holds. We found out that it was not simply price but many other variables at both the store and product level which affected the total sales of the product. The results of the project could be used by Big Mart to modify their sales practices and store systems to increase sales resulting in higher profits and satisfied customers.

**I. INTRODUCTION**

In the retail chain industry, like many others, there is constant need to analyze and quantify the impact of numerous factors concerning the business practices, store presentation, product pricing and their relationship to sales. In the most general economic sense, price is the key element which shapes the demand of products for value-minded customers. But there are numerous other aspects which influence why a certain item sells more than others. This can be product level details such as number of substitutes, its visibility within the store, the type under which it is classified etc. Store level features such as location, size and type are also important and therefore, should be considered for the analysis.

The data set, Big Mart Sales, was collected from the website Kaggle.com, which is an online repository for data sets and accompanying analyses. This data set has sales data for 1559 different products being sold across 10 stores located in various regions throughout the year 2013.

Through our analyses, we sought to explore questions such as which locations generate the most sales and whether fat content of a product shapes consumer demand. Through data visualization techniques we would be able to construct narratives about the business not immediately clear by simply looking at the numbers. By using R Studio to build predictive models based as linear regression, decision tree analysis and random forests we could identify significant predictors which affect the sales. We would get a clearer picture of what product features and store properties influence the level of sales of certain products. This can be used to further predict future sales of items by product type, location etc. and influence upcoming company marketing practices to produce more sales.

**II. RESEARCH QUESTIONS**

* Which predictive model has the most accuracy and lowest error when it comes to predicting sales?
* Will stores located in urban or Tier 1 cities have higher sales if we assume higher income levels of people residing in such locations? Or does the opposite hold true?
* Do the sales trend of items with high fat content indicate a preference for healthier alternatives among consumers? Does this change based on the location of the store?
* Do bigger stores stocking wider variety of products have higher sales since they act like one-stop shops and people would prefer the convenience of getting all their shopping done in a single location?
* Items with higher visibility and greater shelf space will catch the eyes of customers more than items stocked less or displayed at the back of the store. Will the location or visibility of a product within the store influence sales?
* Do older, more established stores have greater sales due to an assumed larger customer base compared to newer stores?
* Will better advertising of products help in higher sales and high demand?
* Does giving discounts and providing offers on particular items result in higher sale?
* It is necessary to stock up the store as per customer needs. Will it be beneficial to study the customer in the area and then stock-up on the goods in the store?
* Products used in day to day life are purchased by almost every customer. Do these daily products have more demand than other products?
* Should the store have a Marketing Department that helps in marketing all types of products to the Customer. The more the customer knows what products are available more they are tending to buy?
* If a particular store is located in the main marketplace which is densely populated, with multiple similar stores around, will the sales go up due to more customers or will they reduce due to multiple store options?

**III. PROJECT SIGNIFICANCE AND RESEARCH IMPLICATIONS**

In a hyper-competitive market in the age of big data, utilizing the technological innovations of new, complex tools and analysis techniques on the increasing amount of data available, businesses can distinguish themselves from others. Customers are always hunting for great deals as they are naturally inclined towards better products at lower costs. Product pricing and marketing practices not only drive sales but also increase typical basket size for most customers. There are also other benefits such as increasing store foot traffic and brand awareness.

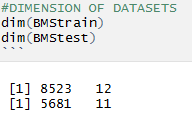
Even though our ultimate aim with the predictive model is to identify significant factors and use those to predict sales, the information we derive from the aforementioned data visualizations will also influence Big Mart and perhaps other businesses engaged in the same industry to adapt their business practices. Products with higher sales could have minimum level of stock maintained at all times, products with lower sales could be advertised more or discounted, store size and layout can be changed to grow customer presence, new stores could be built on popular locations to meet increased customer demand etc. Well-designed marketing and pricing strategies ensure long-term success, bringin more customers and ensure profitability for business.It’s also important to evaluate consumer behavior and purchase patterns. The goals are toimprove the company sales and customer experience, resulting in a positive outcome for all parties involved.

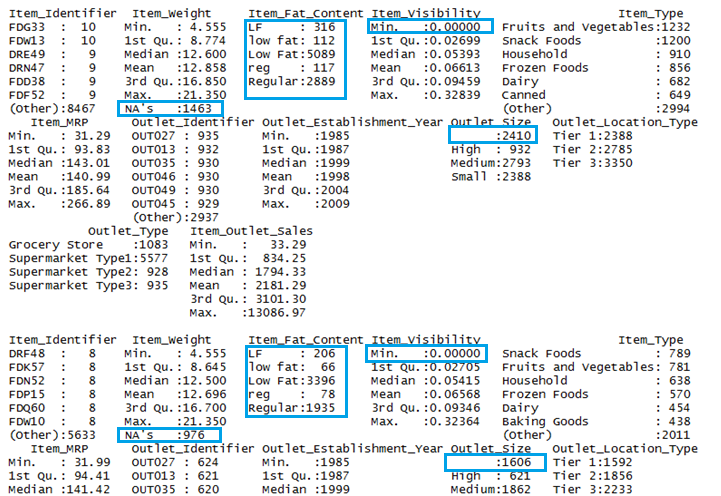
**IV. DATA DESCRIPTION**

We have 8523 observations in training dataset and 5681 observations in test data set, train data set has both input and output variables. We need to predict the sales for test data set.

|  |  |
| --- | --- |
| 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 particular store. (TARGET VARIABLE) |

**Table: Variables and Descriptions**





**V. SUMMARY STATISTICS**

Here are some quick inferences drawn from variables in train data set:

* Item\_Fat\_Content has mis-matched factor levels.
* Minimum value of item\_visibility is 0. Practically, this is not possible. If an item occupies shelf space in a grocery store, it ought to have some visibility.
* Item\_Weight has 1463 missing values.
* Outlet\_Size has an unmatched factor levels.

DATA MANIPULATION

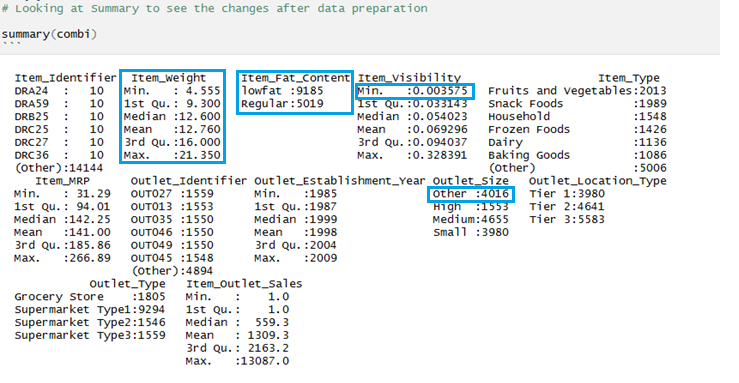
We combined the training and testing dataset so that we don’t have to prepare data separately for both, also we added target variable in test dataset so that both can be appended.

We impute the missing value of Item\_weight variable and also impute missing value by median. We are using median because it is known to be highly robust to outliers.

Then we impute the zero(0) value of Item\_visibility with median.

We proceed by converting different variations of fat to consistent values- low fat and Regular fat content. Then we impute the records of Outlet\_size with others.

We summarize as below:



Thus, we were able to manipulate the data (steps shown above with comments) and made it ready for visualization and modelling and perform further analysis regarding our research questions.

**VI. Methodologies required to complete the analysis**

* Data Imputation.

The initial acquired data consisted of missing values which hinders the stability of the data. Hence, we used the imputation method to make the data more consistent. This resulted in more predictable results compared to the original data. With the help of R Programming, we have summarized the data in term of categories which gives us an overview of the type of product in this database. The summary also gives us the count of particular type of product category-wise. The summary also categorizes the products based on their fat content, so that they can be recommended separately for health-conscious people. The summary also shows an overview of the stores in which the said product is being sold.

* Hypothesis testing.

For the research questions stated above, we will attempt to run hypothesis test.

First, we assume an outcome of the analysis after which we run the actual dataset and match the results from the hypothesis. There is a set percentage of error called “P-Value” which acts as the threshold for hypothesis decision. If calculated error is more than the P-Value, then the assumption done for the database outcome is not correct.

Our data consists of a catalog of edible product that are sold in various stores nationwide. It is sold in small scale stores to big marts situated in rural or urban areas. Our hypothesis testing aims to confirm sale trends of the product in rural and urban areas. We are going to analyze which category of product is more famous in a particular area.

After the data is ready, we start the predictive models using Linear Regression, Decision Tree, Random Forest. We start with the baseline model to predict the sales.

* Linear Regression using random forest and decision tree.

Here we run Linear regression using one or two variables which explains the degree of relationship between the two using a best fit line or plane. Linear Regression can cause overfitting or under-fitting, so we need to tune the parameters for better results.

Regression using decision tree plotted using all the variables gives us a bigger model but with overfitting but tuning the parameters can help. Decision tree on the other hand will provide us with the optimal decision path in selecting best possible product to be sold in a store. Decision tree covers all possible scenarios with the dataset and shows how much profit and loss the store incurs with each selection of product. Also, the decision tree shows the probability of success and failure of each decision with the amount they will gain in case of success or the loss they will bear in case of failure.

* R-Square and RMSE comparison.

Various methods will be used to gain a R-Square value as well as RMSE (error). The method with the highest R-Square value and lowest RMSE is selected as the most optimal test for data analysis.