5/1/2017

NABANITA GHOSH

CUSTOMER ANALYTICS FOR RETAIL SHOPPING

**INTRODUCTION:**

The retail dataset chosen consists of information about purchase pattern of households who visit a specific retail chain. The information ranges from household demographics to campaigning offer types to product level details. The applications of this dataset is widely spread but not limited to analyzing sales, customer profiling, and product clustering and campaign management. We plan to pursue customer analytics i.e. understand the driving factors leading to the customer spend along with forecasting the sales for the retail chain in the next 30 days so that we may devise the best possible market strategy and solution to target the right customer profiles to increase the total sales in the future.

**DATA:**

The dataset comprises of 7 CSV files as follows.

1. Demographic - 8 variables and 801 observations

2. Transaction Data- 12 variables and 2.5 M observations

3. Products Data- 7 variables and 90K observations

4. Coupons Data- 3 variables and 124K observations

5. Coupon Redemption – 4 variables and 2.3 K observations

6. Campaign- 3 variables and 7.2 K observations

7. Campaign Description – 4 variables and 30 observations

The dataset consists of Day Level Transactions for period of 2 years. We plan to aggregate household sales at day level in order to obtain 711 time instances for time series forecasting of daily sales of the retail chain. The transaction dataset which is our primary dataset has been obtained by recording POS terminal sales of purchase by customers.

All other information including customer details, marketing details have been derived from company’s data warehouse.

**RESEARCH QUESTION:**

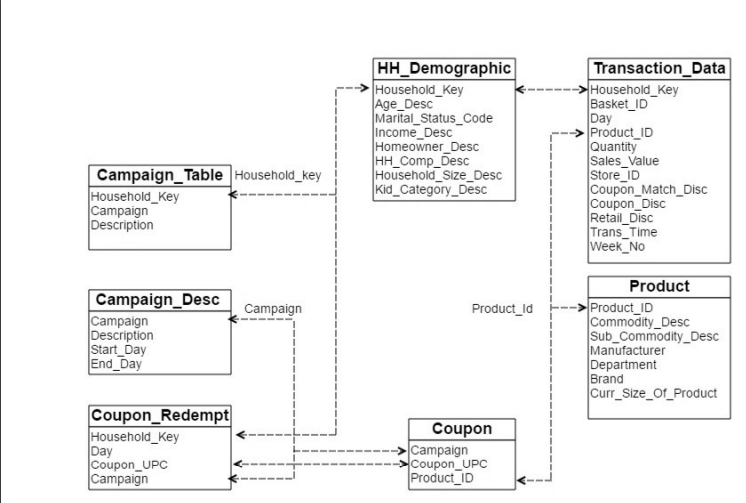
Q-> What are the major factors that influence customer spends?

To answer above question, we will look at three broad aspects of purchase parameters- Monetary Value, Quality and Quantity of Products Purchased, and Frequency of Visits.

Q -> What are the forecasted daily sales for the next 30 days after 12th December 2015

**ANALYTICAL APPROACH:**

The data is detailed across multiple CSV files and we began our analysis by merging all the data sets using common columns like household\_key as detailed in the ER Diagram below:



As we are performing customer analytics, we are not considering campaign information anywhere in our analysis. The data is transactional and detailed at household, day, product level. As we aimed to perform customer analytics and profiling, we began by aggregating data at household level. We created a master data set that had aggregated information about total sales from each household, total products purchased, general coupon discounts offered by retailer, number of stores visited, number of orders placed, total visits made to retailer, number of national brand products purchased (expressed as a percentage), recency of visit to retailer, number of coupons(these are coupons from targeted marketing campaigns of retailer) redeemed by household, visits made over a weekend (expressed as a percentage), Average Basket Size (average number of products purchased in a visit). This data set comprised of 2500 observations of 14 variables. Demographic information was not available for all the households but only for 801 households, hence we performed two analyses- one without considering demographic parameters (for all 2500 households) and another in consideration of demographic parameters (for 801 households).

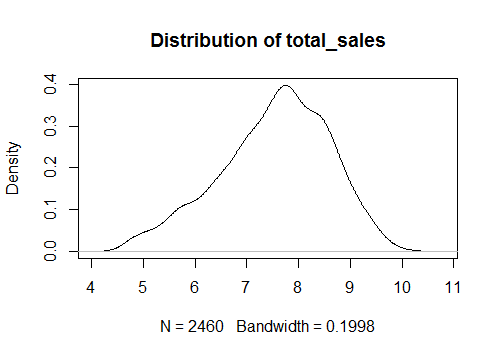
**DATA CLEANING:**

We observed some NAs in the data set in the column- No\_of\_coupon\_redemptions, which is acceptable as all households may not have been targeted for marketing campaigns by the retailer. We replaced all such NAs with zeros and progressed further.

**UNIVARIATE ANALYSIS:**

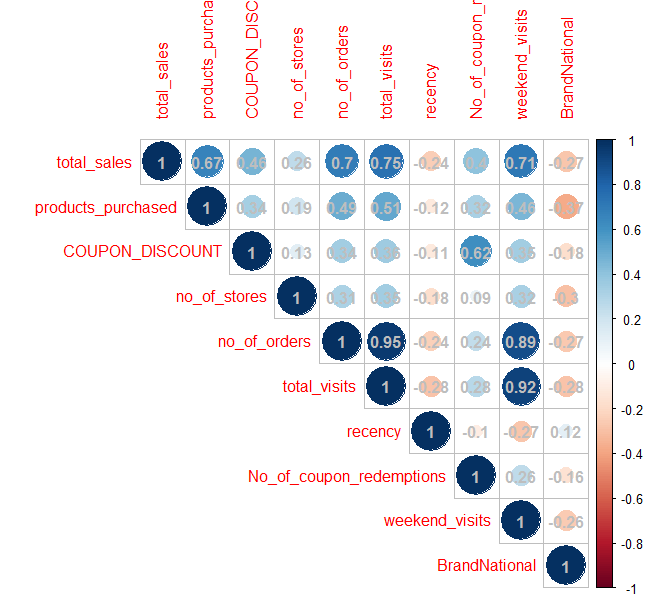
We began by performing univariate analysis and we see unexpectedly high range from 75th percentile to maximum for almost all the columns. We suspected something wrong and believed that there might be a few products that are purchased in very large quantities. To analyse, we aggregated quantity sold and sales produced values for all the products over the transaction and product merged dataset. Observing top 5 products, we see Gasoline-Reg-Unleaded produced total sales of 0.02 cents over a period of 2 years whereas quantity sold was extremely high (more than 2.5 million). This looked like an outlier and we planned to remove all transactions of Gasoline-Reg-Unleaded and recreated the Master data set. We further saw slightly better results for summary output.

Exploring further, we did not observe normal distribution for our dependent variable (total\_sales). The log transformation seemed to have normal distribution and thus we decided to use log transformed sales values for all the further analysis. Upon checking for outliers, few were observed and those outlier instances were removed from the data set. At the end of this step, we obtained a clean and data which has a normal distribution for our further analysis.



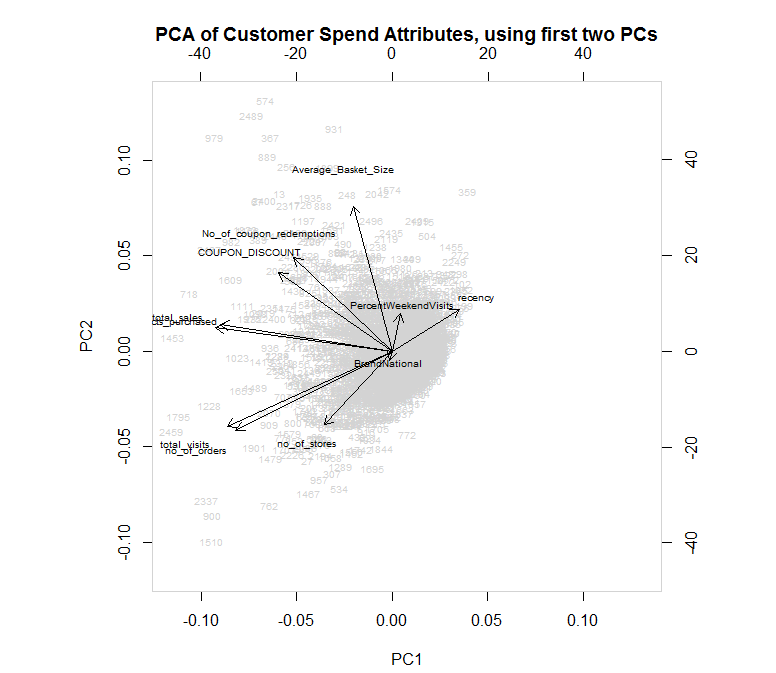
**BIVARIATE ANALYSIS:**

Total Sales seem to be highly positively correlated with total visits, no of orders, weekend visits, products purchased. No of stores, Recency and Brand National seem to be less correlated with total sales. This can be seen further in correlation plot below:

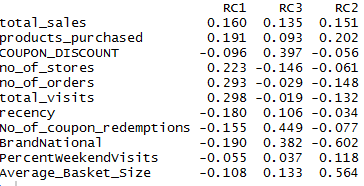


**PRINCIPLE COMPONENT ANALYSIS:**

We decided to perform Principle Component Analysis as our data had many columns. Data was scaled and parallel analysis suggested that 3 components are enough to explain variance in data.



Observing the plot, we see total\_sales and products\_purchased AND total\_visits and no\_of\_orders have similar weights of explaining variance. The weights of each of the variables as seen in this is as below:



As a key takeaway from this step, we decided to drop no\_of\_orders variable as similar information is being expressed by total\_visits.

**LINEAR REGRESSION ANALYSIS FOR 2500 HOUSEHOLDS:**

Several linear regression models were created to predict total sales for all 2500 households. Initially the model was created considering all variables from Master data set and then the model was enhanced by just including the significant variables. Later, we observed that products purchased was highly significant in determining total sales. So, we introduced interactions of all variables with products purchased to observe its influence on total sales and observed interactions between generally offered coupon discount and number of coupons redeemed (from targeted marketing campaigns’). The results have been summarized below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Parameters | Significant Interactions | Adj R Sq |
| 1 | All | No Interactions | 80% |
| 2 | Significant Variables- products purchased, no of stores, total visits, recency, BrandNational, Average Basket Size | No Interactions | 80% |
| 3 | Significant Variables- Average Basket Size, No of stores, recency, BrandNational | Products Purchased : Total Visits | 62% |
| 4 | Significant Variables- No of stores, recency, BrandNational, total visits | Products Purchased: Average Basket Size | 71% |
| 5 | Significant Variables- Average Basket Size, total visits, recency, BrandNational | Products Purchased: No of stores | 78% |
| 6 | Significant Variables- Average Basket Size, total visits, recency, BrandNational | Products Purchased: Coupon Discount | 77% |
| 7 | Significant Variables- Average Basket Size, total visits, recency, BrandNational, products purchased | Coupon Discount: No of coupon redemptions | 79% |
| 8 | Significant Variables- Average Basket Size, total visits, recency, BrandNational | Products Purchased : PercentWeekendVisits | 79% |
| 9 | Significant Variables- Average Basket Size, total visits, recency, BrandNational | Products Purchased: No of coupon redemptions | 77% |
| 10 | Significant Variables- Average Basket Size, no of stores, recency | Products Purchased: BrandNational | 75% |

Looking at above results, we further evaluated models 2, 5, 7, 8 and chose Model 8 for predicting total\_sales variable. The adjusted R square for this model is good as compared to others and interactions were significant to influence dependent variable. The residual plot is slightly left skewed. We tried to improvise Model 8 further by removing outliers but did not observe any noticeable improvement in performance.

Thus, to conclude for this analysis, we say that total\_sales is highly influenced by Average Basket Size, Percentage of National Product purchased, total visits made by household, and is negatively influenced by recency. The total sales is also influenced by number of products purchased in relation with percent weekend visits.

**LOGISTIC REGRESSION ANALYSIS FOR 2500 HOUSEHOLDS:**

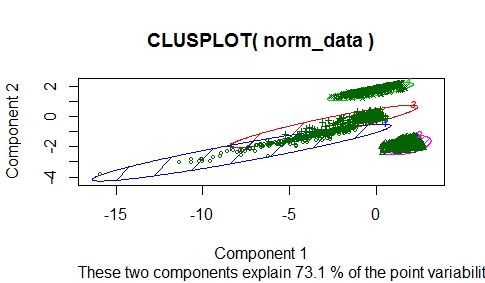
We created a new factor variable named customer spend with values High and Low. Households that spend more than mean value of total\_sales are categorized as High and the ones with values less than the mean value are categorized as Low. We did this to understand parameters that influence higher customer spend values. We started by creating a model considering all parameters and improvised in next step by just considering significant variables. We noticed better results as AIC obtained for subsequent model was less than the previous one. The model that we obtained showed that the log of odds of high customer spend are influenced by products purchased and percentage of National Brand products purchased. We used the interaction information from Model 8 above and used the same for this model. The interactions were significant to affect customer spend but the overall model was not better than the model without interactions, AIC value was very high. So, we chose the model without any interaction. We further evaluated the model by checking multicollinearity and the results indicated that the variables- products purchased and percentage of National brand products purchased are not correlated. DurbinWatsonTest was also performed and the p-value indicated that the residuals are not correlated. The pseudo R square value for this model is nearly 90% which indicated that this is a good model.

**ANALYSIS USING DEMOGRAPHIC PARAMETERS FOR 801 HOUSEHOLDS:**

All the analysis performed until now was without considering household demographic parameters like Age, Income, Marital Status, Size of Household, and more. We merged the derived Master data set with household demographic data set using household\_key. We further performed Chi square test to check correlation between customer spend and each of the demographic parameter. The results of Chi square test indicated that no correlation exists between customer spend and any of the demographic parameter. We also created a logistic regression model by just considering customer spend and all the demographic parameters and the results indicated neither of the variables are significant to predict customer spend.

**CLUSTER ANALYSIS FOR 2500 HOUSEHOLDS:**

Initially we started with significant variables from regression model and sales to cluster the customers using Kmeans. The scree plot had elbow points between 4 and 6. Several models were built using k=4, changing nstart from 100 to 1000 and changing the seed value. The best model for k=4 with nstart=500 as it gave distinct clusters with good amount of variability being explained.



**KMEANS MODEL with K=4 ,NSTART-500**

Next model with k=5 and k=3 were tried and both yielded less distinct clusters and lower variability being explained. Changing the seeds did not change total withinss of the model much.

Next thing which we tried was whether transformation techniques made any impact on data. We had initially made the cluster using min max transformation and total withinss was in the range of 97 to 120 for k=4. We tried scale function (mean, standard deviation transformation) on k=4 and total withinss was observed to be much higher than min max transformation. This led us to use min-max transformation for the subsequent models.

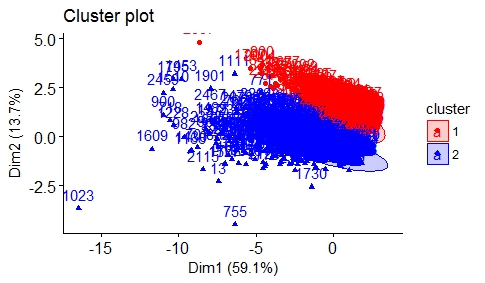
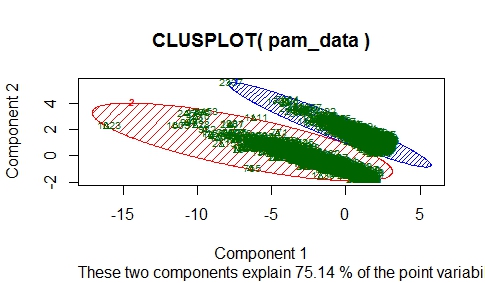
Next, we tried the PAM method and the scree plot suggested k=2. Using K=2 average width silhouette was the highest and on plotting clustering, two nice distinct clusters were formed with 75% variability explained (the highest variability achieved till now). We finalized our model to be PAM with k=2.

Further analysing the clusters, we looked at proportion of members, sales, products, recency, % of national products, visits etc. in each cluster to understand the characteristic of both the clusters.

Cluster 1- lower sales, less products, less recent, more % of national products, less discounts

This cluster has potential to grow and can be targeted so that customers move from 1 to cluster 2

Cluster 2- Cream members present in this cluster with high values for sales, products, visits, redemptions, discounts. But they have low % of national products. We have earlier found national products drive sales but this cluster is buying more of private products. Business can decide to target this cluster with more offers for national products since their redemption is much better than clusters.

**TIME SERIES ANALYSIS AND FORECASTING:**

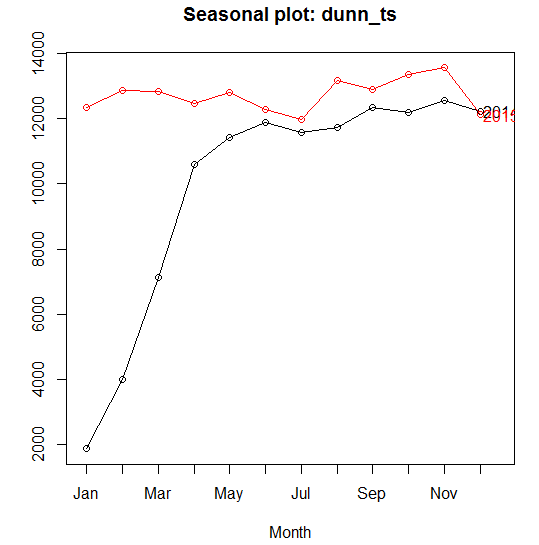
Time series data description: Aggregated Sales of the retail giant at Day level.

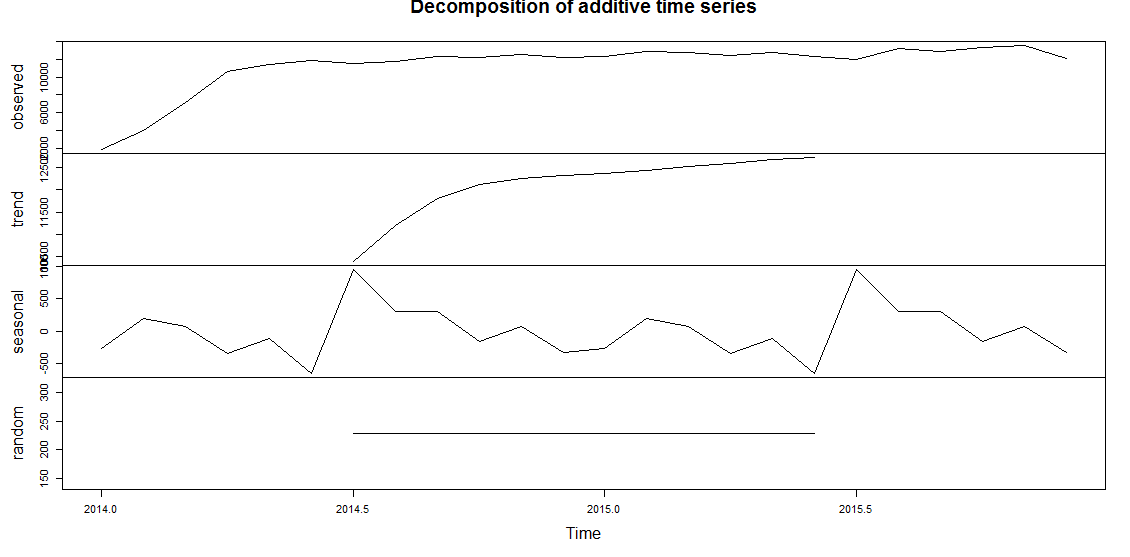
There are 711 data points or in other words 711 days of sales in the time series dataset.

On further inspection, it was observed the monthly sales volume for first 90 days is in the range of 1K whereas rest of the months have sales volume between 10K to 13K.

The next step was to understand the trend and seasonality in the dataset.

Seasonal and decomposition plots show peaks around July and November for both 2014 & 2015.



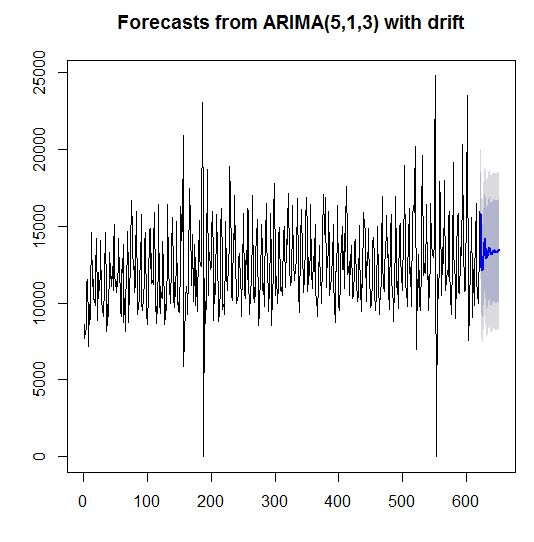


This indicates data is seasonal. The trend looks stationary which is a good sign for not doing further transformation to make the data stationary. On performing ADF test and fitting regression line on the data, it is further confirmed that data is stationary.

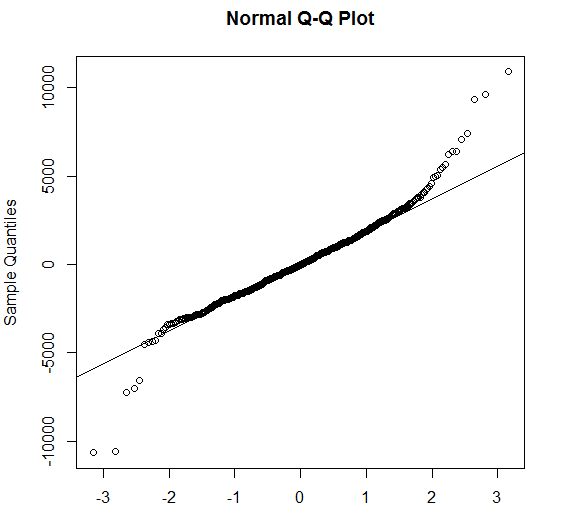
Next step was to try different time series model and select the best model.

Based on the RMSE value, we chose Auto ARIMA (5,1,3) since it had the lowest root mean square and the residuals passed the normality test indicating a robust model.

Looking at the forecast for next 30 days, we observed a dip in sales in next 10 days and again the sales picking up for the next 20 days. The retail giant should be able to use this information to take steps to improve the sudden dip by giving some offers to customers.

**ROOT MEAN SQUARE ERROR from Different Forecasted MODELS**

**The Auto ARIMA model had the least RMSE and the final forecasting for next 30days is shown**



**The residuals of Auto ARIMA models seem to be normally distributed**

**CONCLUSIONS:**

* Customer Spend is not influenced by Coupon Discount and Number of Coupon Redemptions suggesting improvement in coupon offers
* Purchase value is highly dependent on the percentage of national brand products as opposed to private brand products
* The number of products purchased in interaction with percentage of weekend visits significantly influence sales
* There is no correlation of any of the demographic parameters and product sales
* Cluster 2 has high volumes of sales and higher percentage of coupon redemption. This cluster is responding well to marketing campaigns. But the percentage of national brand products bought them are low. More campaigns focused on national products can be targeted on them to drive sales since national product is a significant variable in driving sales
* Cluster 1 has low sales volume and low coupon redemptions. They should be the prime focus of marketing campaign team to drive sales to migrate them from 1 to cluster 2
* Time Series ARIMA model indicates drop in daily sales in the next 10 days requiring immediate attention to give offers to customers specially cluster 1 customers to drive sales