**Marketing Analytics for a Grocery Store**

**Team B-11**

**Akshay Jagadale, Emily Leung, Chiranjeeve Mohapatra, Pradhija Sankarapandian**

**Executive summary**

In this paper, we discuss about how marketing campaigns impact consumer habits and bring benefits to both business and consumers. A common marketing problem case has been taken which deals with segmenting customers based on their response towards buying behaviour. A predictive model has been created by using different methodologies based on their behavior and can be used for predicting behaviour of new customers. The below analysis and findings shows number of ways to effectively implement marketing campaigns on customers.

**Introduction**

Marketing analytics is a field where market performance is measured, managed and analyzed. This is frequently used in order to increase return on investments (ROI) as well as enable a more effective marketing performance. With a good knowledge and application of marketing analytics, marketers are better able to be more efficient at their jobs and achieve significant costs savings in their marketing efforts.

Apart from the aforementioned benefits, marketing analytics provides a deeper insight regarding customer preferences and trends, which is very important for any customer-centric industry. However, even though marketing analytics can offer so solutions and benefits, many industries and organizations are not actually able, or capable, of utilizing marketing analytics to its fullest potential. According to a survey senior marketing executives published in the Harvard Business Review, "more than 80% of respondents were dissatisfied with their ability to measure marketing ROI [1]



Within this report we present a few solutions through the use of various marketing analytics techniques for a common marketing problem. A grocery store wants to quickly analyze marketing campaigns based on responses, revenue and other key metrics. Its main aim is to predict what its revenue will be for different customers based on the various market related parameters of such a customer. These market related parameters can be channels the customers use, customer’s demographic information, etc.

**Data and methodology**

We obtained the data from IBM Watson Analytic’ sample dataset [2]. The datasets here are primarily related to a marketing domain and helps in building models for marketing analytics. The source dataset contains a total of fifteen variables. AmountSpent is our objective and the dependent variable that we want to predict. Not all of the other fourteen variables are necessary in predicting the dependent variable, thus, we have excluded the following variables: storeid, customerid, week, CouponSequence and CarryOver. We have retained the following variables as independent variables: HealthyStore, StoreSize, StoreLayout, gender, WhoShoppingFor, Vegetarian, ShoppingStyle, CouponUser, CouponValue.

**Retained Variable Descriptions:**

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| HealthyStore | This is a categorical variable and has two possible values which are “Regular store” or “Marketed as health food store” |
| StoreSize | There are three possible store sizes “large”, “medium” or “small”. StoreSize is also a categorical variable. |
| StoreLayout | It is a categorical variable and has four possible values. These values are “Emphasizes bakery”, “Emphasizes deli”, “Emphasizes produce” or “No Emphasis”. |
| Gender | This variable indicates the customer’s gender with values “Male” or “Female”. |
| WhoShoppingFor | This variable indicates who the customer is shopping for and the possible values are “Self”, “Self and Spouse” or “Self and family”. |
| Vegetarian | This variable indicates if the customer is vegetarian or not. It has only two possible values “Yes” or “No” |
| ShoppingStyle | This variable indicates what the customer’s shopping style. The possible values are “Often; what’s on sale”, “Weekly; similar items” , or “Biweekly; in bulk”. |
| CouponUser | This variable indicates what type of coupon user the customer is. The possible values are “From newspaper”, “From mailings”, “From both” or “No”. |
| CouponValue | This variable indicates the coupon values. The possible values are “00 No value”, “05 percent”, “15 percent” or “25 percent” |

**Data Preparation**

The primary step in any Data Science project is to have the appropriate dataset to build the model on.

We searched the dataset for the missing values. Our findings were that the dataset was fully populated and there weren’t any missing values to be found. However, had there been any missing values, we could have imputed mean values or removed the entire records in the cases where missing values were present.

Outliers are values that are exceptionally larger or smaller than the expected value. This is typically an issue when dealing with continuous variables. The found outliers need to be excluded in order to yield appropriate results. In regards to our dataset, the data was primarily composed of categorical data so there were not many outliers to accommodate for.

**Methodology**

There are mainly two types of prediction problems: classification and regression. In this application, we are using regression because our objective is to predict the amount spent and we are not attempting to classify the data into groups or segments.

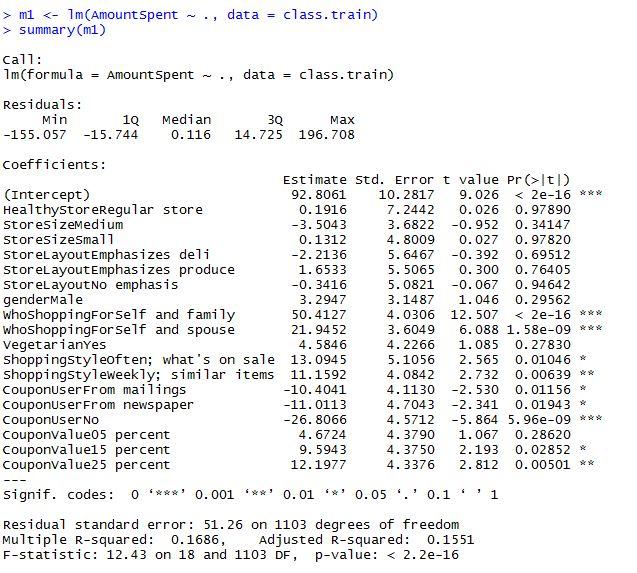
We divided the entire dataset into two parts, a training and a validation dataset. The training dataset is composed of 75% of the total dataset whereas validation dataset contains the remaining 25% of the dataset. We performed sampling in order to get random training cases and then we converted the required variables into categorical variables.

The first technique we applied was multiple linear regression to verify which variables were actually important in predicting the amount spent. Following this, we then applied stepwise regression in both directions in an attempt to obtain a lower RMSE value.

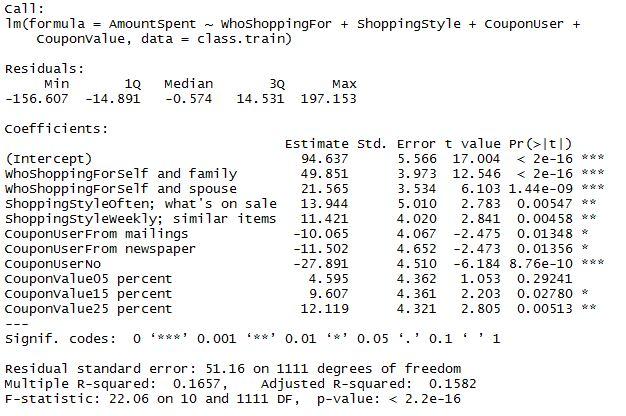
**Analysis**

The following is the analysis of the two techniques that we applied on training dataset:

Multiple Linear Regression:



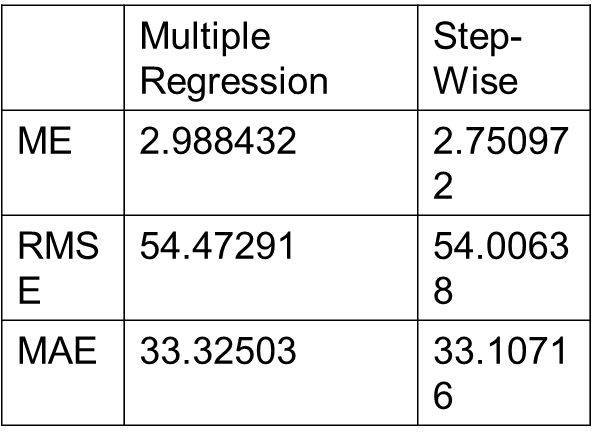
Stepwise Regression:



We applied our model using stepwise regression on test dataset to get our predictions and RMSE value.

**Results**

The overall model on training dataset using stepwise regression is significant since p-value is smallest. The RMSE value for stepwise regression is lower than it is for multiple regression. We even achieved lower ME, MAE and more significant F-value in stepwise as compared to multiple regression. The table shown below displays results achieved from the two regression models:



We can infer that stepwise regression creates a better model for prediction than linear multiple regression model. In addition to this, the stepwise regression model shows more number of significant variables as compare to multiple regression model since it eliminated all non-significant variables. In other terms, stepwise regression model has taken into account variables which are significant and impacts more on the output rather than insignificant variables in multiple regression. The adjusted R-squared value for stepwise regression is found to be superior than the value displayed in the linear multiple regression model. In our attempts, it is found that about 15-16% of the variation can be explained by the stepwise regression model.

**Actionable Marketing Implications**

For any brand to thrive, a positive attitude of customers towards brand should be the ultimate goal of every company. And to bring positive attitude, marketing campaign holds very important role to play. Here we have put few of our contemplation. From the regression model above, it is evident that customers shopping for themselves, family and spouse are highly significant and have strong impact on revenue. Customers who are buying often and weekly have also good impact on revenue. Customers who are not using any coupons have a negative impact on the model, this implies the more number of customers without coupons, the less revenue the store makes. We’ve also found that coupon with values of 25% off seems to have greater impact than other coupon values.

These are some of the insights we gained from our analysis that will help to understand customer needs and preferences, in turn will help to generate more revenues and drive marketing campaigns accordingly. From the findings, we can say that the high negative coefficient is one of the reasons for low revenue and the company should focus more on those customers who don’t opt for any coupons since the strategy can help increase their revenue significantly. Secondly, the company should further analyze the buying patterns of customers who tend to buy for themselves, their spouses and their families. This can aid customer retention by providing these analyzed customers more attractive offers. In addition to this, marketing team should always keep in mind factors such as market demand, competition etc.

**References**

[1] <http://www.wordstream.com/marketing-analytics>

[2] <https://community.watsonanalytics.com/guide-to-sample-datasets/>