**DSA3101 Data Science in Practice**

**Group Assignment 3**



**Group 3:**

Coleen Toh Xuanting (A0188875A)

Colin Ng Chenyu (A0189871J)

Dai Linghao (A0177874J)

Daniel Seah Jieh Tzen (A0180335L)

Divyadarshini Thirugnanasambandan (A0184488L)

Douglas Wei Jing Allwood (A0183939L)

**DSA3101 Data Science in Practice Group Assignment 3**

**Abstract**

A news article from Food Industry Asia (FIA) dated in May 2019 reports how even though the majority of Malaysian consumers want healthy food, the cost of such consumption is a deterrence. In the report, it shows that 99% of Malaysians are interested in maintaining a healthy diet and are actively trying to improve their consumption habits. However, a significant majority (71%) identified cost as a key barrier to achieving a healthy diet. Supermarkets tend to be a common source for many families in Malaysia to purchase their day-to-day needs for cooking. As such, this example highlights the significance of the need for supermarkets to promote healthier choices at more affordable prices.

**Introduction**

This project aims to **increase our client’s revenue** by **recommending an optimal food basket** (in terms of cost and nutritional value) for our client’s customers and at the same time, **promote healthier eating choices**.

Our group will assign health scores to each product which aid in analysing customer preferences. Based on the average heath scores of the products that the customers buy, we will segment the customers into two clusters: health conscious and not-health conscious. Following which we will perform Market Basket Analysis (MBA) to understand current consumption patterns and recommend healthier food replacements.

**Data**

We used four datasets:

1. DSA3101\_Hackathon\_Data.csv: *Panelist ID, Date of Transaction, Category of Product purchased, Product Pack Size, Volume of Product, Spend on Product*
2. DSA3101\_Hackathon\_Panalists\_Demographics.xlsx: *BMI, income, ethnicity, lifestage, strata, type of household and location*
3. DSA3101\_Hackathon\_Categories\_Information.csv: *Category, Calories/100g, Price per Volume*
4. Categories.csv: *Calories/100g, Price per Volume, Saturated Fats/100g, Sugar/100g, Sodium(mg)/100g*

This dataset is an extension of the Categories Information dataset in (c). We web-scrapped information about the products from <https://www.myfitnesspal.com/food/search> to get a better understanding of the individual products.

**Exploratory Data Analysis**

1. **Data Cleaning & Preparation**
2. Removal of Data where no Money was Spent  
   The presence of such data would affect the reliability of our models in the later part of the analysis. A total of 1291 rows of data were dropped
3. Calculation of Volume of Products Purchase

There were a few rows of data whereby the volume purchased was 0. We then estimate the volume purchased via spend / price per volume.

1. Deletion of Duplicate Rows

The presence of such data would affect the reliability of our models in the later part of the analysis. A total of 44,032 duplicated data were dropped.

1. **Manual Clustering of the Products**

With the addition of data that was web-scrapped, we now have a total of 4 attributes (i.e. calories/100g, fats/100g, sugar/100g and sodium/100g) for every product in our Categories dataset.

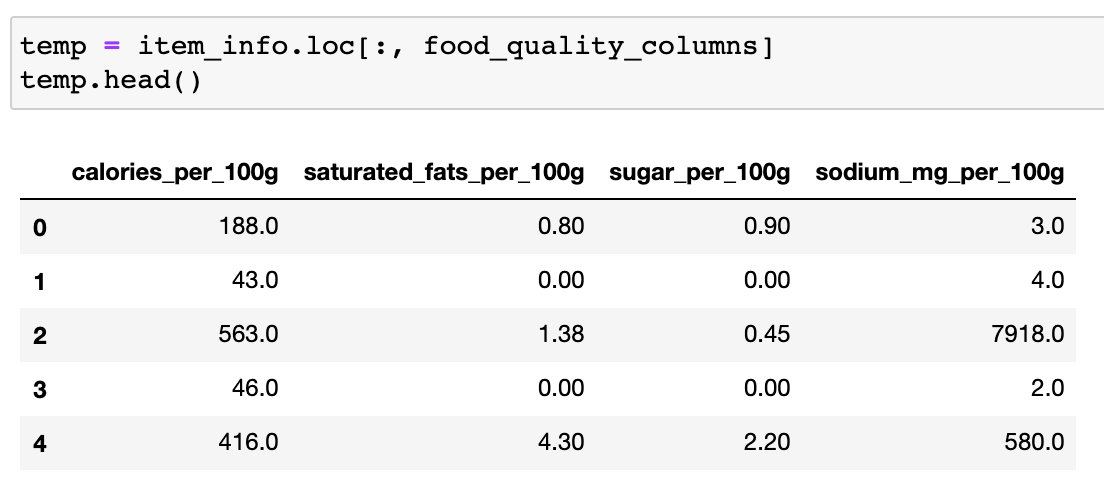
To ensure the relevance of our recommendations, we manually clustered the products into different categories as shown in *Table 1* (full list in annex).

| **Category** | **Number of Products** |
| --- | --- |
| Drinks | 19 |
| Milk | 9 |
| Spreads | 10 |
| Snacks | 9 |
| Food | 4 |
| Cooking Essentials | 4 |

*Table 1. Summary of number of products in each Category*

1. **Calculation of Health Score for Products**

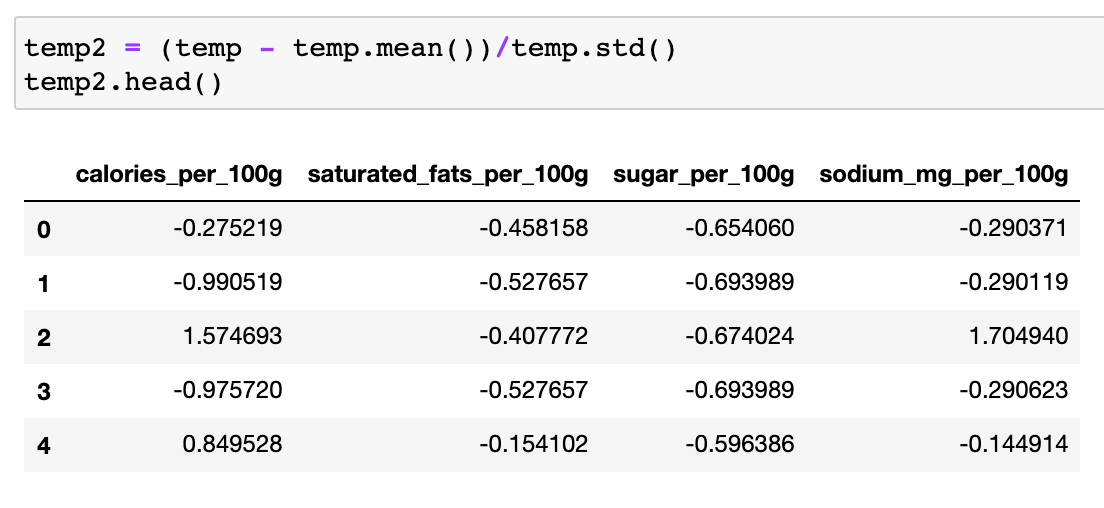
Using the values of the four nutritional attributes for each product as shown in *Figure 1* , we were able to manually calculate and provide a ‘Health Score’ that determined if a particular product is considered healthy or not.



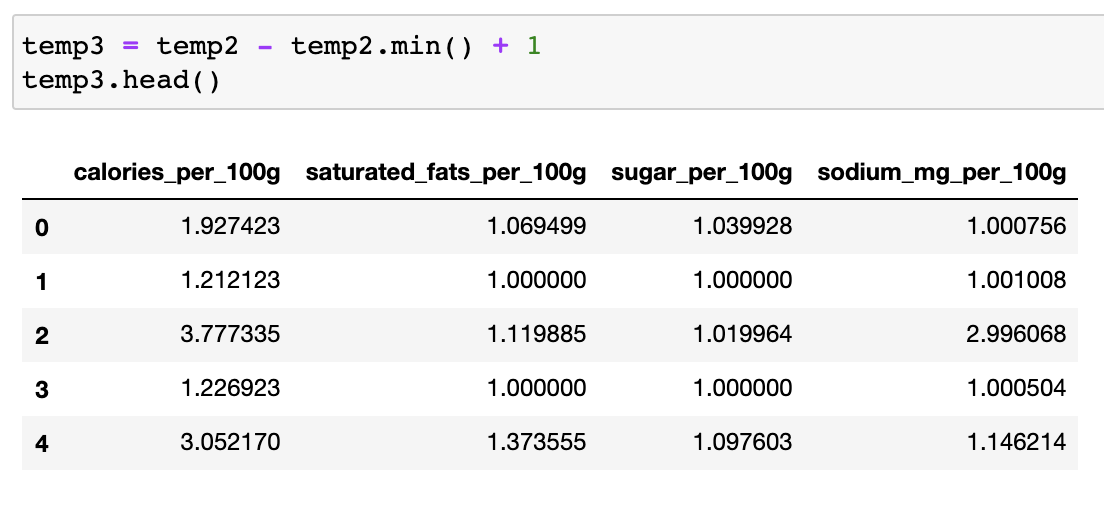
*Figure 1: Values of the 4 attributes of each product*

The health score for each product is generated according to the following procedure:

1. Standardize each of the four columns: “calories per 100g”, “saturated fats per 100g”, “sugar per 100g”, “sodium in mg per 100g”as shown in *Figure 2*.
2. Standardize then translate each column so that each of their minimum values is 1. (i.e. standardise then add a fixed constant to every value in the column, so that the minimum value of each column is 1) as shown in *Figure 3*.
3. For each item, multiply its final scores for “calories per 100g”, “saturated fats per 100g”, “sugar per 100g”, “sodium in mg per 100g” to generate its final health score.



*Figure 2: Standardization of the 4 attributes of each product*



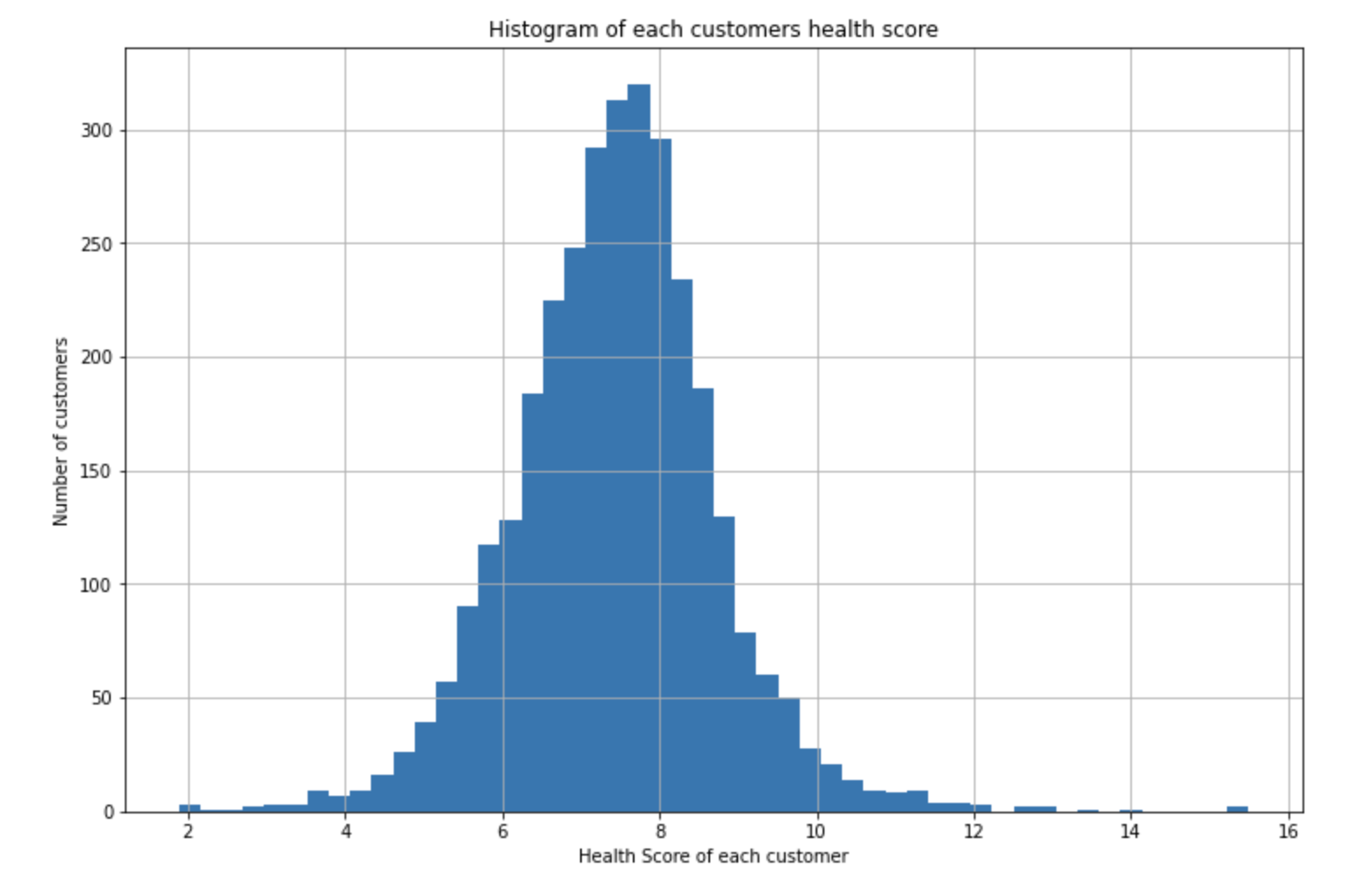
*Figure 3: Translation of each attribute by a fix constant*

Using this method of calculation, the Health Score for each product would be ≥ 1 since the smallest value possible is 1×1×1×1=1. The rationale here is that a multiplicative Health Score will compound the effects of every contributor. For instance, a product with 3 low and 1 high attributes will be lower than a product which has 2 low and 2 high attributes. (e.g. 1×1×1×10 =10 is less than 1×1×5×5 = 25). **The lower the health score, the healthier the product.**

1. **Segmentation of Customers into Health Conscious and Non-Health Conscious**

We calculated the Health Score of a customer by taking the average of the Health Scores of all the items the customer had bought. To ensure that the customers are well segmented, we labelled the top 30 percentile as our Non-Health Conscious customers (Cluster 1), i.e. those with Health Scores above 8.04. The bottom 30 percentile are our Health Conscious customers in (Cluster 0), i.e. those with Health Scores below 6.85.

With this segmentation, we have an equal number of 971 customers in both categories.

****

*Figure 4: Histogram of each customer’s Health Score*

**Methods and Analysis**

1. **Using Decision Tree to Predict whether a Customer is Health Conscious**

Based on the segmentation done above, we aim to predict whether a customer is health conscious so that we can provide appropriate recommendations that are suited for the customer.

To ensure that the decision tree can be applied to new customers, we used only the demographics data as the input features since new customers can be asked for this data when signing up to any app or shopping platform. First, we created indicator random variables for our categorical variables. 1456 of the data were used for training and the remaining 486 were part of our test set.

| Prediction \ Actual | **Health Conscious** | **Non-Health Conscious** |
| --- | --- | --- |
| **Health Conscious** | 171 | 60 |
| **Non-Health Conscious** | 69 | 186 |

*Table 2: Confusion Matrix of Decision Tree*

| **Accuracy** | 0.7346 |
| --- | --- |
| **Sensitivity** | 0.7561 |
| **Specificity** | 0.7125 |
| **F1-Score** | 0.7344 |
| **AUC** | 0.7343 |

*Table 3: Performance of our DT model*

1. **Customer Segmentation on Health and Non-Health Conscious Customers**

Our main objective is to recommend an optimal food basket to the customers, this means recommending a healthy food basket that is targeted towards each customer. In line with our objectives, we identify that narrowing our target demographic to the health conscious customers would result in the largest conversion rates as they would be the most receptive to receiving healthier recommendations and alternatives.

In addition to recommending a healthy food basket, we are also aiming to maximise profits from these recommendations. Following this line of thought, we will segment the customers by spending, into three categories, ‘High Spenders’, ‘Medium Spenders’ and ‘Low Spenders’. This allows us to gain better understanding of the specific customer groups and hence more specific recommendations appropriate for the customer. In addition, this separation also allows us to identify key customer segments that will have the greatest impact on our profits, as observed by the 80-20 rule, a majority of the profits are likely to be caused by a minority. These three customer groups will be further investigated in this report. Do note that our marketing campaigns will not focus on the low spenders since this customer segment offers the least potential for profits.

|  | **No. of Customers** | **No. of Transactions** |
| --- | --- | --- |
| **High Spenders** | 319 | 79024 |
| **Medium Spenders** | 435 | 159168 |
| **Low Spenders** | 217 | 79494 |

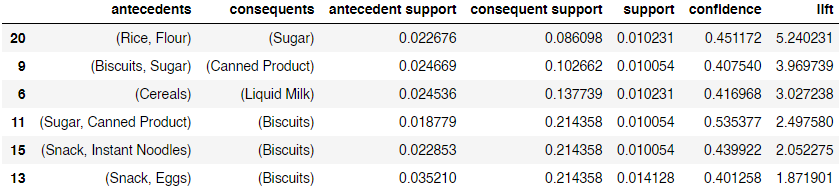
*Table 4. Segmentation of Health-Conscious Customers*

1. **Market Basket Analysis on Segmented Customers to Recommend Optimal Food Basket**

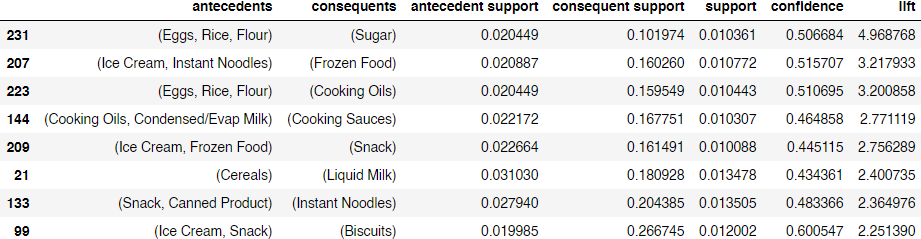
Our market basket analysis will be performed on the health conscious customers subset of data, further segmented using the customer segmentation of spending as explained in the previous section. We will use the following three metrics to assess the quality of an association:

* **Support**: Denotes the frequency of a rule within transactions. A high value means that the rule is involved in a large portion of transactions in the given data. Choosing an appropriate *support value* ensures that the association rules reported are actionable. After we analysed the distribution of support values of association rules across various clusters by studying their histograms and median values, we came to the conclusion to use the minimum support of 0.01. Even though the number may seem low, in the bigger picture with close to 80,000 transactions per sub-cluster, this support value provides us with roughly 800 transaction observations which is sufficient to reach a conclusion.
* **Confidence**: Measures the percentage of times that an item (e.g. item Y) is purchased given that another item (e.g. item X) was purchased. This was a secondary variable of consideration that we used to ensure that the generated association rules were actionable.
* **Lift**: Tells us how much better a rule is at predicting a result than the confidence of an item. This was our primary metric for discovering actionable insights as high Lift values would allow our client to promote the sale of consequent items by launching marketing campaigns (e.g. discounts, bundle packs) for the antecedent items.

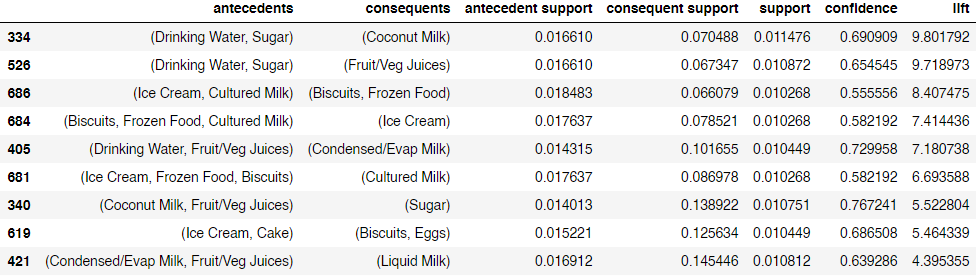
The following tables will show the association rules for the different customer segments, the three customer segments being ‘High Spending Health Conscious Customers’, ‘Medium Spending Health Conscious Customers’ and ‘Low Spending Health Conscious Customers’.



*Figure 5. MBA of High Spending Health Conscious Customers*



*Figure 6. MBA of Medium Spending Health Conscious Customers*

**

*Figure 7. MBA of Low Spending Health Conscious Customers*

From the above obtained association rules for each of the customer segments, we can then identify the key association rules through observing a high lift value for the appropriate rule. Across the board, we can see that the lift values for the obtained association rules for all customer segments are high, therefore we can be confident that the customers are likely to be receptive to the recommendations made with these association rules.

**Data-driven Recommendation and Impact Analysis**

1. **Recommended healthier food options to high spending health conscious return customers**

Nowadays, customers are becoming increasingly aware of their nutritional needs. As a result, there has been an increase in the number of health-conscious consumers. As mentioned in our method and analysis part, based on the purchase transaction records of the customers, we separated them into two groups, health conscious customers and non-health conscious customers. What we are going to do is to **recommend healthier food items to high spending health conscious customers**, as they are more willing to accept the items although they might be more expensive.

Our targeted customers, for example, would be more likely to purchase sugar given that they have had rice and flour in their baskets (shown in *Figure 5*). Instead of recommending sugar to them, we recommend honey as an alternative since honey has a lower health score (i.e it is a healthier option) and it also has a higher price per volume as compared to sugar. Similarly, for those with biscuits and sugar in their baskets, we would recommend instant soup to them rather than canned products. With this method we create a group of itemscalled the **healthy food bundle which consists of rice, flour, honey, eggs and instant soup**. We will recommend these items to our target customers.

**Impact Analysis**

The high spending health conscious customers make around 80,000 transactions with the store every year. Amongst these transactions, approximately 24,000 transactions contain at least one item in the “healthy food bundle”.

Assuming we give a 10% discount if all the items in our “healthy food bundle” are bought together and 2% of our target customers decide to take up the offer recommended to them (i.e., the bundle) in every transaction they make, then, the revenue the bundle generates per year would be .

**Impact Analysis of recommending healthier alternative (Example of just one item):**

In this example, we consider recommending Honey as an alternative to Sugar.

An approximate average of 600 transactions made by healthy high-spender customers per year contain sugar. By **recommending honey as an alternative**, and assuming 2% of the transactions accept our recommendation, then our client could make a profit of per annum.

Difference in price between sugar and honey = 38.38 - 2.94 = 35.44

Profit =

Of course, this is an impact analysis for the substitution of only one product. Recommending healthier, more expensive alternatives on say 30 products in total could result in an average profit increase of per annum.

1. **Recommended optimal basket to new customers by predicting whether they will be health conscious or not with Decision Tree**

The way we segment customers to be health conscious or not is based on their past transaction records. However, if the customer is new, we don't have this data to help us classify them. Thus, we used the demographic data of our existing customers to determine their health consciousness with the help of Decision Trees. We got a fairly acceptable model with an accuracy of 73.46%.

When new customers come to the store, we would conduct a survey to collect their demographic information to segment them into health conscious and non-health conscious groups. For the health conscious groups, we will provide the same packet we generated on the above to them. For the non-health conscious groups, we would directly provide the food basket using MBA.

**Impact Analysis**

From the data, an average of 9.68 customers join the shop each month. With a 73% prediction accuracy we would be able to predict 7 of their health conscious statuses correctly and target them with the appropriate optimal food basket of either healthy products for health conscious customers or any MBA targeting for non-health conscious customers. On an annual basis, this would be 84 correctly labelled customers per year. Note that this value ignores the existing customer base which would already be easy to label because our client already has their transaction data and can tell which customers are health-conscious and which are not.

To assess the financial impact of this predictive algorithm, assume that a correct prediction results in a 10% success rate for the marketing campaign, then given that each targeted marketing attempt results in an average of $2 in increased spending per visit to the shop and given that most customers visit the shop every 7 to 14 days, we expect the following annual growth in spending:

. Note that this will add every year if no customers churn.

Additionally, the data we have collected is obviously only a subset of the actual customers who visit our clients shops. This is as we only have 3236 customers on record over 3 years which is bound to only be a fraction of the total customer base. Hence, for a more accurate estimate, if the regular customer base is *k* times higher than 3236, then the actual annual growth in spending should be .

Lastly, as a non-quantifiable metric, the current churn rate is about 9.31 customers per month. Perhaps by offering customers items that suit their health preferences with 73% accuracy, customers will have a better shopping experience and the churn rate might decrease.

**Annex**

The manual clusters below aid us in recommending healthier options for the MBA.

