Predicting Ice-Creams By Consumer Preferences

A Data Analytic Project for The Street Ice Cream Company Australia

Table of Contents

List of Tables

List of Figures

Introduction

In recent years, loyalty programs have become more numerous as companies wanted to attain new customers and retain them, boost up their spending, influence their spending behavior, and promote the purchase of extra products.

Customer retention is exceptionally vital for raising a sustainable business. According to Harvard school" increasing customer retention rates by 5 percent increase profits by 25 percent to 95 percent".

Some businesses do not use their loyalty scheme in a correct way and this fails them to bind existing purchaser records, in a way that they communicate inappropriate offers to consumer which drives them disengage.

Our company will provide assistance to your company to identify the products that the customers are more interested in to buying. Thus by employing our company you will be benefited with increase profit margin and customer satisfaction.

Aims and Objectives

Our aim is to identify the customers who bought some products from your store and than using their demographic and purchasing history we can predict more products that they would more likely to buy in feature.

Our objective is to implement a model that identify your loyal customers and predict the product that these customers are more interested in to buying. Thus based on the results, you offer them more suitable deals to keep their loyalty so that they made more transactions with you.

Background

Proper utilization of customer loyalty is vital for any business. Without knowing your customers, and lacking to do suitable research on their shopping, you are isolating your business from a great amount of profit.

In the past, reliability of customers was deliberated either on behavior or on attitude. But now the trend has changed and researcher insists to measure both at same time(Bennett and Bove 2002).

Neng and Bin states contentment and confidence relationship have a optimistic influence on behaviour and attitude loyalty of customer (Neng-Hui, Bin-Tsann et al. 2013).

There are different techniques that are being used to cluster objects that are alike, into one group. Some of them are

- Density based clustering
- Connectivity based clustering also known as hierarchal clustering
- Centroid based clustering also known as K-Mean Clustering

- Distribution based clustering also known as Gaussian Clustering

Young, Song, Si and Keun proposed a new method using K-mean clustering, based on Monetary, Recency and Frequency. Their finding was to trim down customer search effort and find out the items with high purchasability (Young Sung, Song Chul et al. 2012).

Similarly, Zhao and Zhang employed clustering using k-mean, in their study “Customer Segmentation on Mobile Online Behaviour”. They proposed a segmentation method that divides the customers behaviour sequence based on frequency dimension and calculate their similarity by difference of distribution and classify them based on similarity matrix(Zhao, Zhang et al. 2014).

In one study Hajiha, Radfar and Malayeri gave their recommendation to use K-Mean Algorithm by using RFM model to identify valuable customers(Hajiha, Radfar et al. 2011)

They have tested their findings on a large dataset, which was informative for company to understand their customers and improve their service quality.

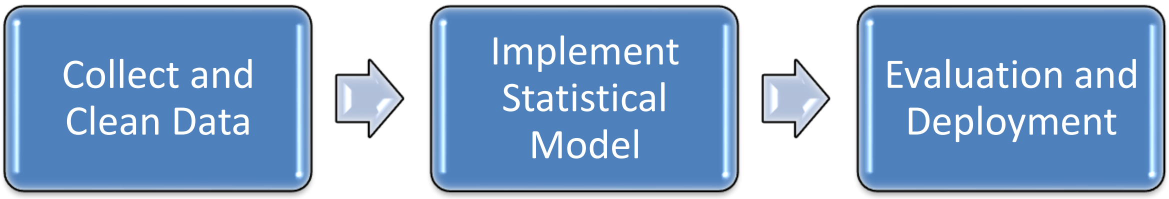
Gupzheng also recommend centroid technique in his study “Customer Segmentation based on survival Character”.He showed K-Mean technique with survival function to identify customer churn tendency(Guozheng 2007).

Based on these research and studies we suggest the best method will be to implement K-Mean/Centroid clustering technique into our project.

Data Analytic Scenario and Methodoloy

Data Mining Problem

The problem we are going to solve is to identify the customers based on their previous shopping and predict new items that customer would me more interested to buy in feature.Based on that knowledge your company would be able to offer a better deal to the customers.



As shown above, we would undertake the following steps to do data mining for your company

- We will Collect all the data of your company and organize it with correct format and make that quantitative data to more qualitative data by clearing the missing values and predict the more suitable values for the missing one.And any changes we perceive may be crucial for your existing data set.

- We would employ the statistical or decision tree model on company data set to predict customer’s response of interest.
- Evaluate the results with our objectives.
- Deploy our results to your business.

Strategy for collecting and organizing the data

We will use loyalty card database and format the attributes in a more convenient way. As we follow IT Professional rules and ethics and we respect customer’s privacy. So, we would not use all the attributes from data base like their name.

Following are the attributes we would use from loyalty card and transaction database

- Gender- as a character like for Male ‘M’ and for Female ‘F’
- Age- as a number like ‘24’
- Post Code-as a number like 2000 is a post code of CBD
- Customer Id- as a number- it would be unique for each customer
- Transaction date- as a Date format dd-mm-yyyy
- Item type- as a number like splice with item number 0100
- Discount- as a number
- Transaction location- as a number (post code of store)

In next phase we will integrate this data base with customer transaction data base
To maintain accuracy we will eradicate incomplete records from the evaluation process. For example the row that does not contain either of transaction date or item type will be discarded.

Loyalty Card Customer Database							
Cust.Id	FirstName	LastName	Age	Gender	Address	Postcode	Email
1	Michle	Fardik	8	Male	21 memorial street	4213	mf@abc.com
2	Aena	Milson	17	Female	7 pyrmont road	1342	Aenamilson@somemail.com
3	Angus	Sutter	45	Male	10 happy Avenue	7252	-
4	David	Flyn	27	Male	1 Broadway	2000	david@ymail.com

Table 1- Loyalty Card Customers Database

As mentioned earlier, we are not going to use all fields in Loyalty Card Database. We will extract some relevant field from this database and integrate it with transaction table.

Transaction Table

Cust. Id	Sale No.	Date	Discount (%)	Product	Item Type	Location Id
1	8965	1/2/2014	0	Frutare	0001	4213
2	9763	14/4/2014	0	Cornetto	0010	1342
3	1034	28/4/2014	0	Golden GayTime	0011	2061
2	1045	7/5/2014	10	Splice	0100	2001
4	1165	9/6/2014	0	Paddle Pop	0101	2000
7	1587	17/6/2014	20	Blue Ribbon	0110	1342

Table 2- Transaction Database

We have created a binary evaluation criteria according to products as shown in table 3, that consumer may consider to buy in their next transaction. Please note this evaluation criteria has room to adjust more products. So, in future if you desire to add more products than it would not be a problem.

Ice Creams	Evaluation
Not interested in any Ice Cream	0000
Frutare	0001
Cornetto	0010
Golden GayTime	0011
Splice	0100
Paddle Pop	0101
Blue Ribbon	0110
Magnum	0111
Future Product	1000
Future Product	1001

Table 3- Products with likely outcome of the model

The likely outcome of the model will be the prediction of customer purchasing. For example if customer's predicted value is 0100 than it mean he/she will very likely to purchase Splice Ice Cream.

Proposed Methodologies

In order to predict Customers purchase using their purchasing history an effective model is required that consider three main factors that is speed, interpretability and accuracy.

We will implement Decision tree on our clustering data because it meets the three factors criteria mentioned above. A decision tree is a hierarchal model that display the results and their probable consequences.

To create a model we will adopt the following steps

1- Modify outliner data by cleaning noise, dealing with missing values and suggest inferable values to fill, in a way that it is more convenient to use and integrate data from other sources.

2- It is not necessary the customers who have loyalty card are actually loyal to you. To find out we will divide customers by their magnitude of transactions with you for a specific time range.

3- From step 2 we would have loyal customers. Than in this step on consumer behaviour and attitude towards a specific product will make a single cluster. And that common product will be linked with this group.

4- Decision tree will depict the customers who will make the transactions in the future.

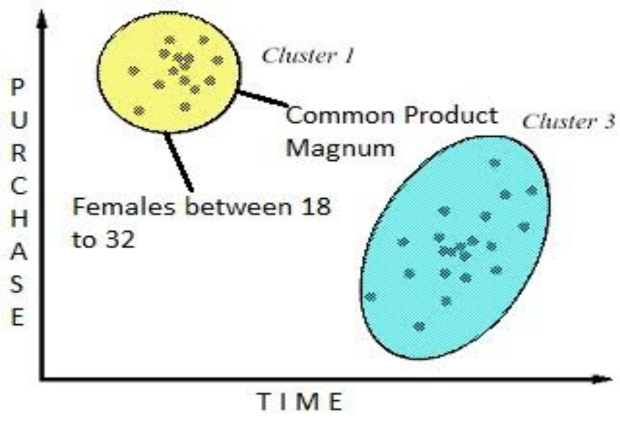


Figure 1- Clustering of data based on Age and Gender

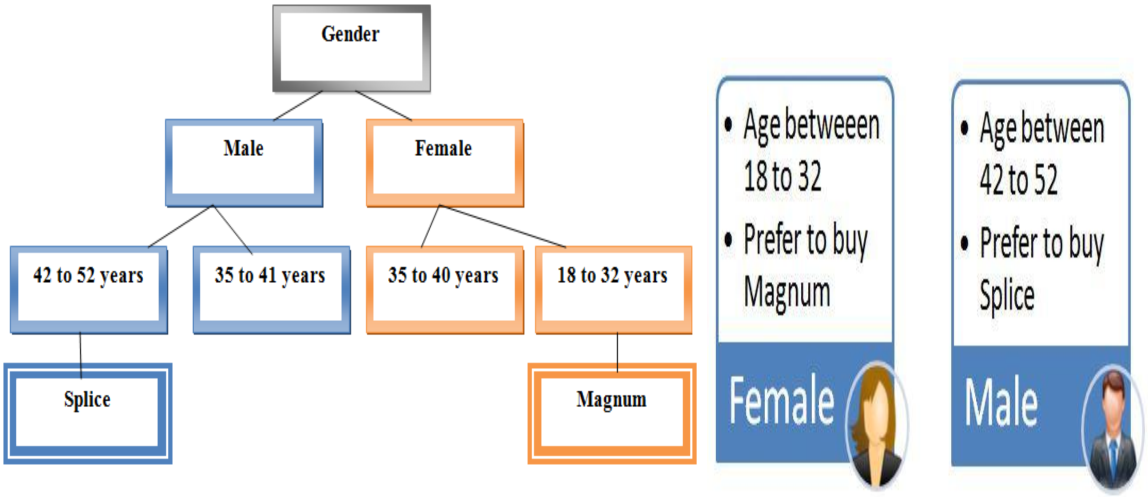
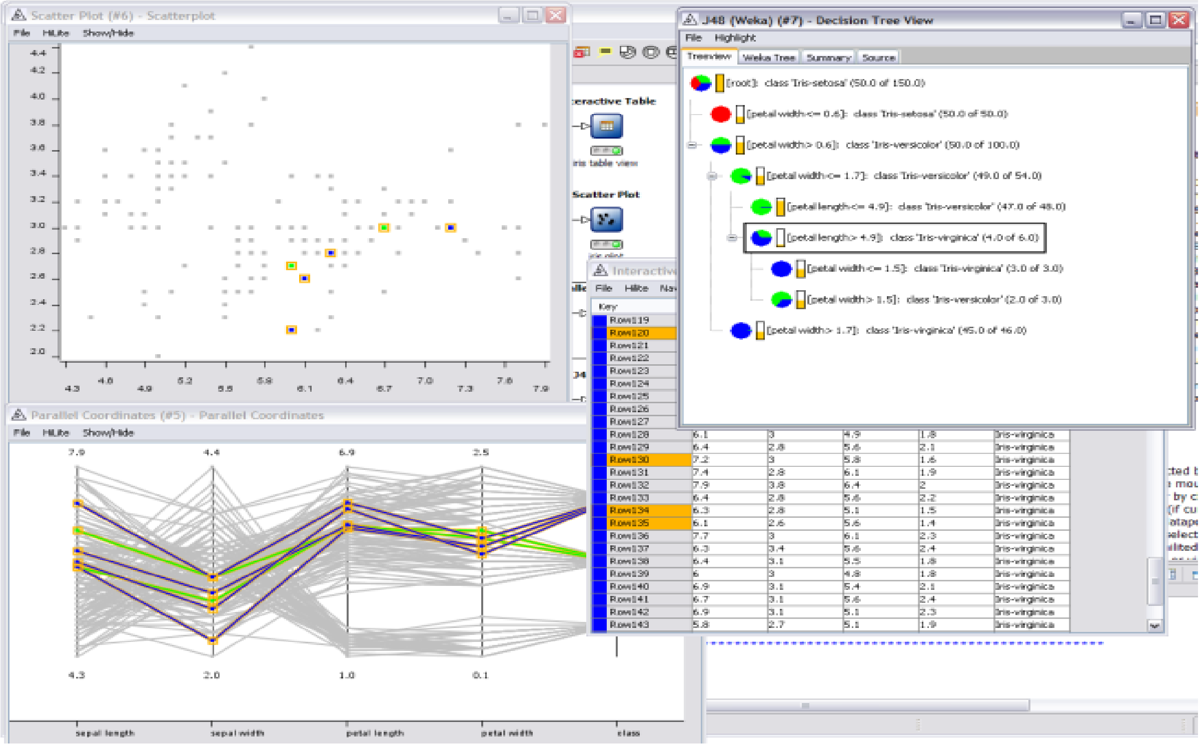


Figure 2- Possible outcome through Modelling Technique

One of the examples of clustering and decision tree based on gender and age is shown above. As we can see in Figure 1 and Figure 2 that Female with age from 18 to 32 years old are likely to buy Magnum while in case of Male with the age of 42 to 52 years old they are likely to buy splice ice cream.

We will use a KNime software to build clusters and depict their logical flow using the same tool. The possible output sample of the tool is shown in figure 3.



```
[root]: class 'Iris-setosa' (50.0 of 150.0)
    [petal width <= 0.6]: class 'Iris-setosa' (50.0 of 50.0)
    [petal width > 0.6]: class 'Iris-versicolor' (50.0 of 100.0)
        [petal width <= 1.7]: class 'Iris-versicolor' (49.0 of 54.0)
            [petal length <= 4.9]: class 'Iris-versicolor' (47.0 of 48.0)
            [petal length > 4.9]: class 'Iris-virginica' (4.0 of 6.0)
                [petal width <= 1.5]: class 'Iris-virginica' (3.0 of 3.0)
                [petal width > 1.5]: class 'Iris-versicolor' (2.0 of 3.0)
        [petal width > 1.7]: class 'Iris-virginica' (45.0 of 46.0)
```

Key					
Row119					
Row120					
Row121					
Row122					
Row123					
Row124					
Row125					
Row126					
Row127					
Row128	6.1	3	4.9	1.8	Iris-virginica
Row129	6.4	2.8	5.6	2.1	Iris-virginica
Row130	7.2	3	5.8	1.6	Iris-virginica
Row131	7.4	2.8	6.1	1.9	Iris-virginica
Row132	7.9	3.8	6.4	2	Iris-virginica
Row133	6.4	2.8	5.6	2.2	Iris-virginica
Row134	6.3	2.8	5.1	1.5	Iris-virginica
Row135	6.1	2.6	5.6	1.4	Iris-virginica
Row136	7.7	3	6.1	2.3	Iris-virginica
Row137	6.3	3.4	5.6	2.4	Iris-virginica
Row138	6.4	3.1	5.5	1.8	Iris-virginica
Row139	6	3	4.8	1.8	Iris-virginica
Row140	6.9	3.1	5.4	2.1	Iris-virginica
Row141	6.7	3.1	5.6	2.4	Iris-virginica
Row142	6.9	3.1	5.1	2.3	Iris-virginica
Row143	5.8	2.7	5.1	1.9	Iris-virginica

Figure 3- KNIME Software Output

Evaluation of Results:

Assessment of outcomes in the company is necessary in terms of accuracy level to maintain the successful business objectives.

The outcomes of the analysis will be evaluated with collected information over a time period and compare it with previously collected information to confirm its accuracy intensity. To perform this task we will simulate project result at ordinary time periods per year and guarantee the customers securing the expected purchase. Sequentially to assess the practical business value will evaluate the profit from purchases done by customers.

Deployment

In the beginning, will establish and explore the project in terms of monthly reports which can help to improve the business strategies. Tracking process is to verify the project accuracy sufficiently. Furthermore, to monitor the effectiveness of strategies, we will follow up by correlating the predicted outcomes in terms of real life data.

It required no initial changes after one year will enclose the project and deliver to the client. Also will instruct one of the employ of yours company about how to utilize and formulate the project plan to take better decision.

Project Plan and TimeLine

By using CRISP-DM Methodology we have divided our tasks based on developer's hourly rates with expected completion time.

The budget for this project is $22,500 which is very competitive as the hourly rate of developer is $ 110 in the market. Also to have long terms relations we are not adding any contingency cost which is normally 2 % of total cost.

Phase	Task	Expected Duration Days	Developer Hours	Cost per Hour of a Developer ($)	Cost ($)
Business Understanding	Prepare Business Objectives	1	10	100	1000
	Asses Situation	1	10	100	1000
	Determine Data Mining Goals	1	10	100	1000
	Produce Project Plan	1	10	100	1000
Data Understanding	Collect Initial Data	1	10	70	700
	Verify Data Quality	1	10	70	700
Data Preparation	Select Data	1	10	70	700
	Clean Data	1	10	70	700
	Format Data	1	10	70	700
	Integrate Data	1	10	70	700
Modelling	Select the Modelling technique	1	10	100	1000
	Generate Test Design	2	20	100	2000
	Build the Model	3	30	100	3000
	Asses the Model	1	10	90	900
Evaluation	Evaluate Results	2	10	100	1000
Deployment	Plan Deployment	1	10	80	800
	Plan monitoring and Maintenance	1	10	80	800
	Produce Final Report	1	10	80	800
	Review Project	1	10	100	1000

Total	$ 22,500

Table 4- Project Plan

The minimum completion time of the project is almost 27 days. In below table you can see the project timeline with individual task completion.

Year/Month	April 2015																					
Days	Start of Month (1st to 10th)							Mid of Month (11th to 20th)							End of Month (21th to 30th)							
Prepare Business Objectives	🟥	🟥																				
Asses Situation			🟥	🟥																		
Determine Data Mining Goals					🟥																	
Produce Project Plan					🟥	🟥																
Collect Initial Data							🟥	🟥	🟥	🟥	🟥	🟥										
Verify Data Quality									🟥	🟥	🟥	🟥										
Select Data											🟥	🟥	🟥	🟥								
Clean Data												🟥	🟥	🟥								
Format Data													🟥	🟥	🟥							
Integrate Data													🟥	🟥								
Select the Modelling technique													🟥	🟥								
Generate Test Design															🟥	🟥	🟥	🟥				
Build the Model															🟥	🟥	🟥	🟥				
Asses the Model																	🟥	🟥				
Evaluate Results																	🟥	🟥	🟥			
Plan Deployment																			🟥	🟥		
Plan monitoring and Maintenance																			🟥	🟥		
Produce Final Report																				🟥	🟥	

Review Project																					🟥	🟥

Table 5- Project TimeLine

References

Bennett, R. and L. Bove (2002). Identifying the key issues for measuring loyalty. Australasian Journal of Market Research. **9:** 27-44.

Reichheld, F. F. (1993), "Loyalty-based management," *Harvard Business Review* (March-April), 64-73.

<http://hbswk.hbs.edu/archive/1590.html, Harvard school >

Guozheng, Z. (2007). Customer Segmentation Based on Survival Character. Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on.

Hajiha, A., et al. (2011). Data mining application for customer segmentation based on loyalty: An iranian food industry case study. Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on.

Neng-Hui, W., et al. (2013). Relationship Quality and Customer Loyalty in Taiwan -- A Longitudinal Aspect. Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013 Seventh International Conference on**:** 773-775.

Young Sung, C., et al. (2012). Implementation of personalized recommendation system using k-means clustering of item category based on RFM. Management of Innovation and Technology (ICMIT), 2012 IEEE International Conference on.

Zhao, H., et al. (2014). Customer segmentation on mobile online behavior. Management Science & Engineering (ICMSE), 2014 International Conference on**:** 103-109.