

market basket analysis

Data Mining (Fall 2019)



December 28, 2019

IBA

A Term Project

Presented To

Dr. Sajjad Haider

Faculty of Computer Science

IBA Karachi

In Partial Fulfillment

Of the Requirements

for the Course

Introduction to Data Mining

by

Maheen Sohail(14801)

Merry Areej(14885)

Abdul Wahhab(14949)

Afzal Zubair(12981)

Fall 2019

Letter of Transmittal

28th December 2019

Dr. Sajjad Haider

Instructor of Introduction to Data Mining

Institute of Business Administration, Karachi

Subject: Submission of Project Report

Dear Dr. Haider,

We feel immense pleasure in presenting to your good self, the term report as part of our course requirement. We found this report to be truly challenging in many aspects, indeed very interesting in relation to the various interpretational and engrossing exercises. Writing this report was a truly comprehensive learning experience. Our research for the project was both primary and secondary. We also used different internet sources for our project.

The purpose of this project was to discover most trending association rules in the sales department of a bakery and analyzing them in a way that helps the bakery increase their sales. This increased our learning and we hope that it proves to be helpful for others too.

Thank you

Best Regards,

Team MBA

Abstract

Market basket analysis (MBA), also known as association rule learning, is an important and prominent component of Data Analytics that can be used in various fields such as marketing, bioinformatics, education field, nuclear science, etc. MBA is one of the key techniques that are used by retailers to discover relationships and associations between different items, to analyze customer purchase behavior, to determine the placement of goods, designing sales promotion to improve sales distribution techniques and increase business. It generates frequent itemsets by analyzing purchasing patterns of customers which then makes it easy to find the popular itemset and the worst combination of items from the large transactional data.

Contents

[**Introduction**](#_glm102l1shui) **6**

[**2. Business Understanding**](#_3sawpf8mzryg) **7**

[Background](#_6f7xpd1b7yis) 7

[Motivation of the study](#_i13rgbetf6d5) 7

[Challenges faced by the company](#_hh3ism3frg67) 8

[**3. Data Understanding**](#_clsp2zhwnu8t) **9**

[Data Description](#_1oqegnocic3v) 9

[Data Exploration](#_kmy6qjbvf64c) 9

[**4. Data Preparation**](#_5ap5jgn7b9i7) **10**

[**5. Modelling/ Evaluation**](#_nfhmje6n9c9v) **11**

[5.1 Analyzing Trends](#_fb9d7lbcvuwq) 11

[5.2 Modelling](#_lm1wehzgpeq) 13

[5.3 Description of the Machine and libraries Used](#_7yicmxsuv769) 16

[**6. Findings**](#_litvx0mxli1s) **17**

[Insights](#_d770be5p23g) 17

[Limitations](#_unsdz7an9ogi) 17

[Advice to the Bakery](#_4vfc4j63vxpo) 18

[**7. Deployment**](#_7slp58459g76) **19**

[**References**](#_2zk6oojnyb13) **20**

# Introduction

The market basket is described as an item set sold together by using a client on a particular visit to the shop. In our visit to the shop, we have a tendency to shop for quite a few products from multiple categories and placed them all collectively in a single basket which is considered a single transaction. Market basket analysis is the analysis of those baskets altogether.

Market basket analysis contains a huge set of analytic techniques geared toward uncovering the associations and connections among particular objects, coming across similar behaviors and relationships among items. In trade, the rules are made on the idea that if a customer bought a certain group of products, what are the chances that he/she will buy another certain group of products. For example, it's known that after a purchaser buys meat, in most cases, buys soft drinks also. These behaviors in purchases are something that the companies selling their products are interested in. The sellers are interested in analyzing which items are bought collectively in an effort to create new marketing/sales strategies that may be beneficial in enhancing the benefits of the company in addition to customer experience.

Most of the retail markets are extra centered on the “what” their purchaser’s purchase. But they ignore the reality about “when” they buy it, which is equally significant in determining their buying behavior. This project is centered on not only “what” the consumer buys but additionally “when” they buy. [5]

**Keywords**

**Support**: The percentage of transactions that contain all of the items in an itemset. The higher the support the more frequently the itemset occurs.

**Confidence:** The probability that a transaction that contains the items on the left-hand side of the rule (in our example, pencil, and paper) also contains the item on the right-hand side (a rubber). The higher the confidence, the greater the likelihood that the item on the right-hand side will be purchased

**Lift:**  The probability of all of the items in a rule occurring together (otherwise known as the support) divided by the product of the probabilities of the items on the left and right-hand side occurring as if there was no association between them.[1]

# 2. Business Understanding

## Background

Marketbasket analysis can be used to learn more about customer behavior. The of market basket analysis in the bakery is to discover the trends within the methodology transactions of bakery items. The bakery that we have selected for our study sells more than 90 different items and has been in the business for more than 4 years now.

## Motivation of the study

The bakery wants to analyze the customer purchase behavior by way of locating associations between the distinctive objects that consumers bought during their visit to the bakery. Such an analysis can help the sales growth and can help the bakery develop advertising and marketing techniques through gaining insight into which items are often bought together by consumers. The strategies can also include:

• Reorganizing store layout according to trends.

* Targeted Marketing by sending out coupons to customers for products related to items they usually purchase.

• Analyzing the greatest sales time periods.

• Designing a catalog.

• Identifying trending items.

* Customized offers to regular customers.

## Challenges faced by the company

The problem many retailers face is how to extract important information from their vast customer and product feature databases in order to gain a competitive advantage. They are unaware of the purchasing habits of the customer so they don’t know which items should be placed together in their bakery. A retailer must know the needs of customers and adapt to them. With the help of this market basket analysis bakery managers can determine the strong relationships between the items which ultimately helps them to put products that co-occur together close to one another. Also decisions like which item to stock more, cross-selling, up-selling, store shelf arrangement are determined.

# 3. Data Understanding

## Data Description

The data for this project was collected from Kaggle. It is a bakery’s dataset which contains it’s customer transactions for the period from October 2016 to April 2017. Every record in this dataset corresponds to a transaction made at the bakery, specifying exactly what item was purchased, when did the transaction take place and the transaction ID. 21293 transactions consisting of 9531 unique customers and 95 unique products in the dataset.

Number of Attributes: 4

1. Date: Date of the transaction
2. Time: Time of the Transaction
3. Transaction: Transaction ID corresponding to the transaction.
4. Item: Item purchased in that particular transaction.

## Data Exploration

No null values were found in the dataset, but some records had an item value ‘NONE’ in them, which meant no items were purchased in that particular transaction, thus it was of no concern to us.

# **4.** Data Preparation

Before doing anything with the data as soon as we read in the data in jupyter notebook, we start exploring it. Data exploring is the most important step for any analysis. While exploring the data the issues that needed fixing were as follows:

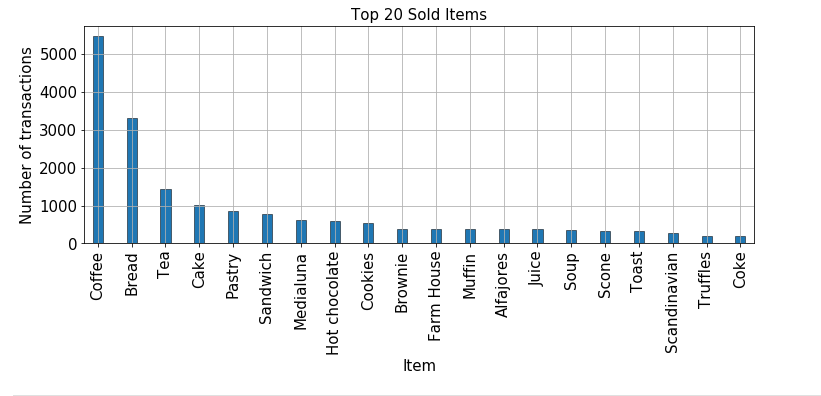
1. Removing ‘None’ values: There were many such records which had no items corresponding to them, instead a ‘NONE’ was fed in the database. This was of no use to our analysis, so we disregarded it by simply removing those records from our dataset.
2. Matrix formation: There were repeating values in the dataset for the transaction IDs. Each Item bought by a single customer was recorded in a separate row, so we had to bring it in a way that all the items bought by a unique customer are attributed to their Transaction ID. So we transformed our data into a matrix form. For a specific transaction i, if an item j is purchased then the matrix position (i,j) is set to be 1. If the item is not purchased then it is set to be 0. [3]
3. Merging date and time: Time is a very crucial factor for the analysis of trends here, that is why we merged the date and time of the transaction into a single column named datetime. [5]

# 5. Modelling/ Evaluation

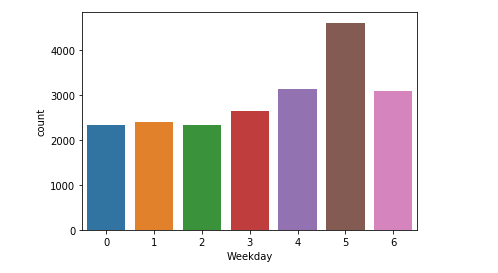
### 5.1 Analyzing Trends

Before applying the Market Basket Analysis algorithms, it was important to analyze the trends in the sales of the bakery items by what information we had.

We plotted a graph for the *top 20 products sold* at the bakery, bringing it to our attention that coffee was the most desired item at this bakery.

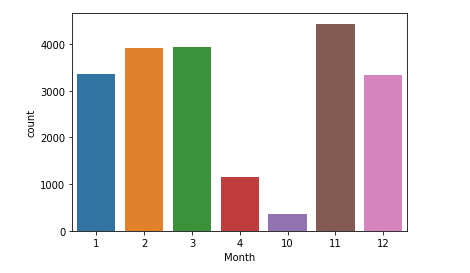


Therefore, we used the datetime column in certain ways, to get an overview of the weekly, hourly and monthly sales.

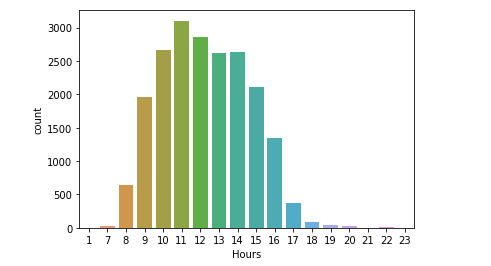
The *weekly sales* trend was as follows: 

This showed us that the 6th day of the week, i.e. Saturday is the most profitable day for the sales accountable for at least 22% of the total number of transactions throughout the week.

The *monthly sales* trend was as follows:



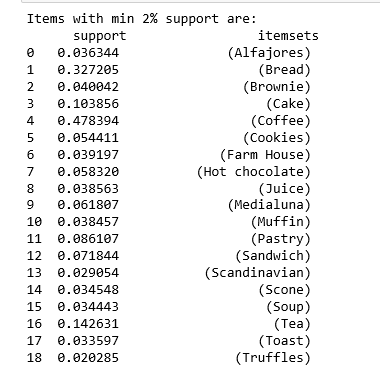
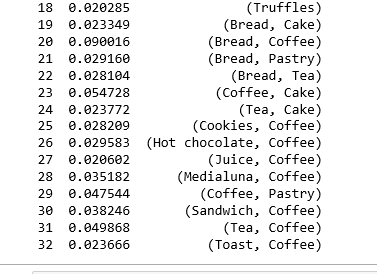
This graph shows us a significant difference between the sales of the month of APRIL and OCTOBER as compared to other months, but that was because our data did not have the details for all 30 days for these two months. The data for APRIL was of only the first 9 days, and the data for OCTOBER was of only the last two days. Apart from that, we saw that the most number of sales were in the month of November while the least number of sales were in the month of December.

The *hourly sales* were as follows: 

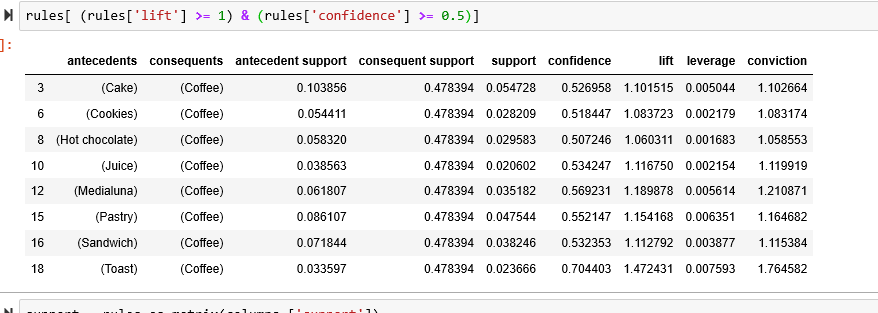
This showed us that the peak hours for the bakery were between 10 o’ clock in the morning till 2 o’ clock in the afternoon. To further analyze the sales by the times throughout the day, we classified the data in the following categories and found out the exact number of sales in those time periods:

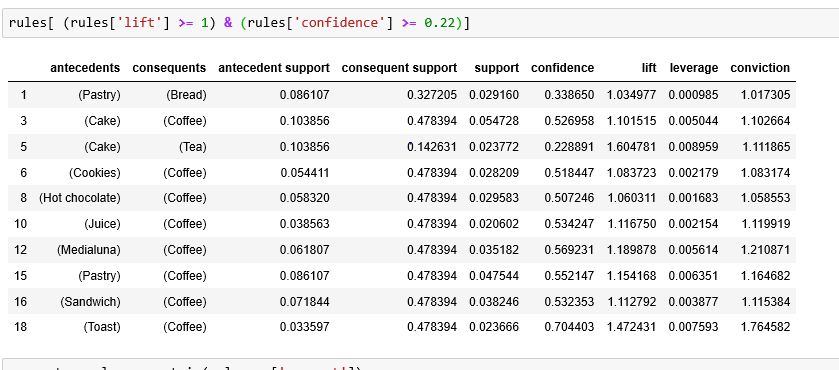
* Early Morning( 06:00:00 till 09:00:00): 669 transactions
* Late Morning(09:00:00 till 12:00:00): 7734 transactions
* Early Afternoon(12:00:00 till 15:00:00): 8111 transactions
* Late Afternoon(15:00:00 till 18:00:00): 3826 transactions
* Night(18:00:00 till 21:00:00): 152 transactions

### 5.2 Modelling

After analyzing the data in the different trends mentioned above, we applied the Market Basket Algorithm on it. Apriori algorithm and association rules were imported from the mlxtend library. The apriori algorithm was applied on the dataset, to get the most frequent itemsets. Minimum support that we used was 0.02 i.e. 2% and after we ran the algorithm we got 32 most frequent items in the dataset 

After that, we applied lift to the items which have minimum support value of 2%, keeping the minimum threshold of lift to 1 as a lift greater than 1 means products A and B are more likely to be bought together than they have the likelihood of being bought separately. Through this process, we got 20 rules such that their support was more than 2% and their lift was greater than 1. 

For further precision, we applied the metric of confidence on those 20 rules, keeping the minimum threshold of the confidence level to 50%, giving us top 8 association rules from the dataset.

All of these 8 rules, had the item Coffee in them, so to get more rules apart from the coffee ones, we set the confidence level to 22%, which gave us 2 more rules which were Pastry → Bread and Cake → Tea. The rule with the most confidence was found to be Toast → Coffee with a confidence level of 70% and a lift of 1.47 

### 5.3 Description of the Machine and libraries Used

All of the above work was done on Jupyter using Python, and the machine we used for the project was:

Machine: Personnel Laptop (core i5)

Ram: 4 GB Hard Disk: 500 GB HDD

Libraries used: numpy, pandas, matplotlib.pyplot, os, seaborn, datetime, mlxtend.frequent\_patterns

# 6. Findings

## Insights

The results obtained in this project are satisfactory. Our main objectives as we mentioned at the start of the business problem are achieved. The association rules discovered will provide a competitive advantage to the Bakery owner. It will surely help him over its competitors. It will help the Bakery Owner to take decisions more effectively since he’ll be aware of the customer behavior, most trending products and at what time periods do they make more profit. A deep analysis of multiple trends of the transactions have led us to the fact that among other days of the week, Saturday is the day with the highest transaction rate, thus the Bakery Owner can launch new products on weekends and as for the other days, he can come up with promotions or discount offers in order to increase sales. Furthermore, we found out that most of the sales throughout the day were between the late morning and early afternoon times so the manager can manage the staff for those time periods accordingly. We found out that for this bakery, their coffee is the most sold product and most of the association rules also lie around it making it the most important factor the bakery should be focusing on. They can bring in more varieties to their coffee and also get reviews from their customers as to what changes they can bring to their other products as well in order to increase their sales too.

## Limitations

* The data was only for 7 months. If the data set was very large (i.e. for more than one year). It would have been very difficult for a normal computer to analyze the data faster. Hence, we had to stick to a smaller data set.
* Thresholds for support and confidence doesn’t have any specific criteria to decide.
* Learning a new methodological tool or statistical technique requires hands-on experience. Though we provided a description of how to use MBA through an illustrative dataset, there are several complications when you are practically analyzing the data. It requires several iterations to get to a better result.[5]

## Advice to the Bakery

The bakery can use this market basket analysis model in order to manage the placement of items in their bakery layout. Related products should be placed together in such a manner that customers can logically find items he/she might buy which increases customer satisfaction and hence the profit. Although most of the association rules found were with respect to the sales of Coffee, but the bakery can take the chance of highlighting other rules as well for example Pastry → Bread and Cake → Tea, since they had significant confidence and lift as well.

The bakery can also use these association rules to promote targeted marketing and give out coupons to customers for products related to items they usually purchase. The identification of the most trending items can be used by the bakery and they can stock their pantry with such products beforehand. The analysis of the peak hours and peak business months can be beneficial for the bakery since they can prepare for extra workload beforehand.

For future data collection, We would advise the bakery to keep records of the previous purchases of the regular customers, so that it can help identify a particular customer’s behavior. The bakery can also keep records of the reviews given by their customers regarding their most sold products or the newly launched ones. This information can help us build a model regarding the sentiment analysis of the customers. The bakery can also keep a numerical record of the value of the transactions made so that a regression model can also be applied to get more insights into the data as well as predict future sales.

# 7. Deployment

The system after deployment should work in accordance with the bakery’s traditional ways.

* The system should be beneficial for the business and easy to implement. If the business is not benefiting it means that more business understanding is required and the system needs improvements. Crisp-DM will be followed iteratively to make the system more effective.
* Training sessions should be organized for the staff so that they should know how the system works.
* Processes should be tracked and monitored on a regular basis whether system is showing the desired behavior or not. This can be done by outsourcing a third party that can monitor it or the bakery owner can also hire new IT-related staff members.

# 8. References

[1] "Market Basket Analysis: Understanding Customer Behaviour"-[Lynsey McColl](https://select-statistics.co.uk/our-consultants/lynsey-mccoll/) <https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/>.

[2] "Market Basket Analysis" - Nagesh Singh Chauhan

<https://en.wikipedia.org/wiki/Association_rule_learning>.

[3] "Market Basket analysis for a Supermarket based on FrequentItemset Mining" - Loraine Charlet Annie & Ashok Kumar

<https://www.academia.edu/38191712/Market_basket_analysis>

[4] "What is the CRISP-DM methodology?"- Smart Vision

<https://www.sv-europe.com/crisp-dm-methodology/>

[5]"Market basket analysis: trend analysis of association rules in different time periods"-Sohaib Zafar Ansari

<https://pdfs.semanticscholar.org/cb40/b283ecc2f0a9e082354f62819ed9ff8aa367.pdf?_ga=2.185628961.905049762.1577444928-577594838.1577444928>