Market Basket Analysis (MBA)

By

YASHI KESARWANI (Reg. no: 0000197)

Supervisor: Dr. Anirban Lakshman

A Thesis submitted to
Indian Institute of Information Technology Kalyani
for the partial fulfillment of the degree of
Bachelor of Technology
in
Computer Science and Engineering
December, 2019

Certificate

This is to certify that the thesis entitled “Market Basket Analysis (MBA)” being submitted by Yashi Kesarwani, an undergraduate student, Reg.No- 0000197, Roll No- 39/CSE/16042, in the Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, for the award of Bachelors of Technology in Computer Science and Engineering is an original research work carried by him under my supervision and guidance. The thesis has fulfilled all the requirements as per the regulations of Indian Institute of Information Technology Kalyani and in my opinion, has reached the standards needed for submission. The work, techniques and the results presented have not been submitted to any other University or Institute for the award of any other degree or diploma.

(Dr. Anirban Lakshman)
Assistant Professor
Indian Institute of Information Technology, Kalyani

Yashi Kesarwani
(Reg No: 0000197)
Department of Computer Science and Engineering
Indian Institute of Information Technology, Kalyani

Declaration

I hereby declare that the work being presented in this thesis entitled, "Market Basket Analysis (MBA)", submitted to Indian Institute of Information Technology, Kalyani in partial fulfillment for the award of the degree of **Bachelor of Technology** in Computer Science and Engineering during the period from July, 2019 to December 2019 under the supervision of Dr. Anirban Lakshman, Department of Computer Science and Engineering, Indian Institute of Information Technology Kalyani, West Bengal 741235, India, does not contain any classified information.

Yashi Kesarwani
(Reg No: 0000197)
Department of Computer Science and Engineering
Indian Institute of Information Technology, Kalyani

Acknowledgments

First of all, I would like to take this opportunity to thank and to express our gratitude to our supervisor Dr. Anirban Lakshman for his guidance, and instruction that he has given us from time to time. I have been doing thesis under his supervision through the completion of our undergraduate project. The skills we have able to learn from his during our thesis work have benefited us immensely and will continue to do so throughout our future endeavors.

Yashi Kesarwani
(Reg No-0000197)

Department of Computer Science Engineering
Indian Institute of Information Technology, Kalyani
Place: Kolkata
Date: 10/11/2018

Abstract

Companies nowadays are rich in vast amounts of data but poor in information extracted from that data. Big data is seen as a valuable resource and although the concept of data mining is still new and developing, companies in a variety of industries are relying on it for making strategic decisions. Facts that otherwise may go unnoticed can be now revealed by the techniques that shift through stored information.

Market basket analysis is a very useful technique for finding out co-occurring items in consumer shopping baskets. Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross-sale campaigns.

The main objective of the thesis is to see how different products in a beauty shop assortment inter-relate and how to exploit these relations by marketing activities. Mining association rules from transactional data will provide us with valuable information about co-occurrences and co-purchases of products. Such information can be used as a basis for decisions about marketing activity such as promotional support, inventory control and cross-sale campaigns.

CONTENTS

Chapters

Chapter-1
Introduction

1.1 Overview

The highly technological era that we live in has made it possible for companies to gather enormous quantities of data. Data mining is becoming more and more common for many businesses worldwide. The large amount of data that is being gathered on a daily basis captures useful information across different aspects of every business. The collection of data on a highly disaggregate level is seen as a raw material for extracting knowledge. While some facts can be revealed directly from disaggregate data, often we are interested to find hidden rules and patterns. Non-trivial insights can be generated through data mining. Data mining contains of various statistical analyses that reveal unknown aspects of the data. Mining tools have been found useful in many businesses for uncovering significant information and hence, providing managers with solutions for complicated problems.

Data mining is commonly seen as a single step of a whole process called Knowledge Discovery in Databases (KDD). According to Fayyad et.al, 'KDD is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data.' (Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, 1996) Data mining is a technique that encompasses a huge variety of statistical and computational techniques such as: association-rule mining, neural network analysis, clustering, classification, summarising data and of course the traditional regression analyses.

Data mining gained popularity especially in the last two decades when advances in computing power provided us with the possibility to mine voluminous data. Extracting knowledge and hidden information from data using a whole set of techniques found its applications in various contexts. Knowledge discovery is widely used in marketing to identify and analyse customer groups and predict future behaviour. Data mining is an effective way to provide better service to customers and adjust offers according to their needs and motivations.

1.2 Business use of data mining

Companies nowadays are rich in vast amounts of data but poor in information extracted from that data. Big data is seen as a valuable resource and although the concept of data mining is still new and developing, companies in a variety of industries are relying on it for making strategic decisions. Facts that otherwise may go unnoticed can be now revealed by the techniques that sift through stored information. When applying mining tools and techniques we seek to find useful relationships, patterns and anomalies that can help managers make better business decisions.

Data mining tools perform analyses that are very valuable for business strategies, scientific research and getting to know your customers better. Managerial insights are no longer the only factor trusted when it comes to decision-making. Data driven decisions can lead to better firm performance.

Data-based implications are gaining popularity while the gut instinct of managers is remaining in the background. Analysing data not only improves firm performance but gives us accurate insights on different aspects of the business.

Data mining is widely used in marketing for spotting sales trends, developing better marketing campaigns and finding the root cause of specific problems like customer defection or fraudulent transactions, for example. It is also used for prediction of behaviour: which customers are most likely to leave us (customer churns) or what are the things that an individual will be most interested to see in a website.

1.3 Research problem description

In the recent years analyzing shopping baskets has become quite appealing to retailers. Advanced technology made it possible for them to gather information on their customers and what they buy. The introduction of electronic point-in sale increased the use and application of transactional data in market basket analysis. In retail business analysing such information is highly useful for understanding buying behaviour. Mining purchasing patterns allows retailers to adjust promotions, store settings and serve customers better.

Identifying buying rules is crucial for every successful business. Transactional data is used for mining useful information on co-purchases and adjusting promotion and advertising accordingly. The well-known set of beer and diapers is just an example of an association rule found by data scientists

Chapter-2

Objective & Literature Review

2.1 Objective:

The main objective of the thesis is to see how different products in a grocery store interrelate and how to exploit these relations by marketing activities. Mining association rules from transactional data will provide us with valuable information about co-occurrences and co-purchases of products. Some shoppers may purchase a single product during a shopping trip, out of curiosity or boredom, while others buy more than one product for efficiency reasons.

Over the past two decades a lot of attention has been devoted to the subject of data mining. While retailers are involved in this topic because of the absolute utility of market basket data, market analysts are interested because of the research and technical challenges they face while analyzing the data.

Increasing amount of data is being generated every second and this allows experts to search for meaningful associations among customer purchases. Customers make purchase decisions in several product categories on a single shopping trip. Interdependencies among products have faced increased attention recently as retailers are trying to improve their businesses by applying quantitative analyses to their data.

It is very important for retailers to get to know what their customers are buying. Some products have higher affinity to be sold together and hence the retailer can benefit from this affinity if special offers and promotions are developed for these products. It is also important to the retailer to cut off products from the assortment which are not generating profits. Deleting loss-making, declining and weak brands may help companies boost their profits and redistribute costs towards aspects of the more profitable brands. (Kumar, 2009) This is yet another reason why data mining is seen as a powerful tool for many businesses to regularly check if they are selling too many brands, identify weak ones and possibly merge them with Page | 7 healthy brands. Data mining techniques are highly valued for the useful information they provide so that the retailer can serve customers better and generate higher profits.

Chris Anderson in his book ‘The long tail: Why the future of business is selling less of more’ explains a concept of the ‘98% rule’, which is quite contrasting to the well-known 80/20 rule. In other words, 2% of the items a retailer sells are frequent, while 98% of the items have very low frequencies, which create a long tail distribution. This is why the presence of this ‘98% rule’ in the retail business created the need for data mining software and made quantitative analysis a must for retailers. (Anderson, 2006)

2.2 Literature Review:

I read some scientific papers related to our topic and got some informations from those papers. Saurabh Malgaonkar *et al* [1] describe a perfect method to extract the data from the huge sets of marketing datasets efficiently by using different techniques of Association Rule Mining (ARM). The systems satisfied the following objectives – to make more informed decision about product placement, pricing, promotion and profitability. Also, Mahmoud Houshmand *et al* [2] especially focused on customer behaviour. Another paper written by Xiaohui Yu, *et al* [3] find out which product should be crosssold. Saurabh Malgaonkar *et al* [1] also describe that their system identifies customer purchasing habits. It provides insight into the combination of products within a customer's baskets. The term 'basket' normally applies to a single order. However, the analysis can be applied to other. We often compare all orders associated with a single customer. Ultimately, the purchasing insights provide the potential to create cross sell propositions. Moreover Julie Marcoux, *et al* [4] put emphasize on sales forecasting. Sales forecasting is an important part of business management since it provides relevant information that can be used to make strategic business decisions. Forecasting can be divided in three categories: future forecasts, environmental forecasts and industry forecasts. Sales forecasting, considered as a company forecast, aims at assessing the performance of a given company regardless of its competitors. Abbas, et al [5] discussed in their paper that one of the problems regarding data mining is to search for meaningful relations in computer purchase data. They also discussed that how differently arranging products in shops decreased the sales of particular items. So, they are using market basket analysis to identify purchasing pattern to help retailers to make a better arrangement of the products.

Chapter-3

System Implementation & Design

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

This application is basically used to find the buying pattern of consumer based on the associations of products in previous transactions.

3.1 Why Association Analysis?

In today's world, there are many complex ways to analyze data (clustering, regression, Neural Networks, Random Forests, SVM, etc.). The challenge with many of these approaches is that they can be difficult to tune, challenging to interpret and require quite a bit of data prep and feature engineering to get good results. In other words, they can be very powerful but require a lot of knowledge to implement properly.

Association analysis is relatively light on the math concepts and easy to explain to non-technical people. In addition, it is an unsupervised learning tool that looks for hidden patterns so there is limited need for data prep and feature engineering. It is a good start for certain cases of data exploration and can point the way for a deeper dive into the data using other approaches.

As an added bonus, the python implementation in MLxtend should be very familiar to anyone that has exposure to scikit-learn and pandas. For all these reasons, I think it is a useful tool to be familiar with and can help you with your data analysis problems.

One quick note - technically, market basket analysis is just one application of Association Analysis. In this post though, I will use association analysis and market basket analysis interchangeably.

3.2 Data Science – Apriori Algorithm in Python

Data Science Apriori algorithm is a data mining technique that is used for mining frequent item-sets and relevant association rules.

Apriori algorithm is a classical algorithm in data mining that is used for mining frequent item-sets and association rule mining. The Apriori algorithm is used for the purpose of association rule mining. Now, what is association rule mining? **Association rule mining is a technique to identify frequent patterns and associations among a set of items.** Understanding customer buying habits. By finding correlations and associations between different items that customers place in their 'shopping basket,' recurring patterns can be derived.

Let us understand the purpose of Market Basket Analysis with the help of an example?

Say, Joshua goes to buy a bottle of wine from the supermarket. He also grabs a couple of chips as well. The manager their analyzes that, not only Joshua, people often tend to buy wine and chips together. After finding out the pattern, the manager starts to arrange these items together and **notices an increase in sales.**

This process of identifying an association between products/items is called association rule mining. To implement association rule mining, many algorithms have been developed. Apriori algorithm is one of the most popular and arguably the most efficient algorithms among them.

What is an Apriori Algorithm?

Apriori algorithm assumes that any subset of a frequent itemset must be frequent. Say, a transaction containing {wine, chips, bread} also contains {wine, bread}. So, according to the principle of Apriori, if {wine, chips, bread} is frequent, then {wine, bread} must also be frequent.

The key concept in the Apriori algorithm is that it assumes all subsets of a frequent itemset to be frequent. Similarly, for any infrequent itemset, all its supersets must also be infrequent.

Example of working of the Apriori Algorithm-

Here is a dataset consisting of six transactions in an hour. Each transaction is a combination of 0s and 1s, where 0 represents the absence of an item and 1 represents the presence of it.

Transaction ID	Wine	Chips	Bread	Milk
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

We can find multiple rules from this scenario. For example, in a transaction of wine, chips, and bread, if wine and chips are bought, then customers also buy bread.

{wine, chips} => {bread}

In order to select the interesting rules out of multiple possible rules from this small business scenario, we will be using the following measures:

- **Support**
- **Confidence**
- **List**
- **Conviction**

Support

Support is the relative frequency that the rules show up. In many instances, you may want to look for high support in order to make sure it is a useful relationship. However, there may be instances where a low support is useful if you are trying to find “hidden” relationships.

Support of the item *x* is nothing but the ratio of the number of transactions in which the item *x* appears to the total number of transactions.

Support(x) = Number of transactions in which item x occurs / Total number of transactions

Support(wine) = 4/6 = 0.66667

Confidence

Confidence is a measure of the reliability of the rule. A confidence of .5 in the above example would mean that in 50% of the cases where Diaper and Gum were purchased, the purchase also included Beer and Chips. For product recommendation, a 50% confidence may be perfectly acceptable but in a medical situation, this level may not be high enough.

Confidence (*x* => *y*) signifies the likelihood of the item *y* being purchased when the item *x* is purchased. This method takes into account the popularity of the item *x*.

i.e.,

Conf ({wine, chips} => {bread}) = support (wine, chips, bread)/support(wine, chips)

Conf ({wine, chips} => {bread}) = (2/6)/(3/6) = 0.667

Lift

Lift is the ratio of the observed support to that expected if the two rules were independent. The basic rule of thumb is that a lift value close to 1 means the rules were completely independent. Lift values > 1 are generally more "interesting" and could be indicative of a useful rule pattern.

Lift (*x* => *y*) is nothing but the 'interestingness' or the likelihood of the item *y* being purchased when the item *x* is sold. Unlike confidence (*x* => *y*), this method takes into account the popularity of the item *y*.

i.e.,

lift ({wine, chips} => {bread}) = $\frac{support(wine,chips,bread)}{support(wine,chips)}$

lift ({wine, chips} => {bread}) = $\frac{\frac{2}{6}}{\frac{3}{6}*\frac{4}{6}}$ =1

- **Lift (*x* => *y*)** = 1 means that there is no correlation within the itemset.
- **Lift (*x* => *y*)** > 1 means that there is a positive correlation within the itemset, i.e., products in the itemset, *x* and *y*, are more likely to be bought together.

- **Lift (*x* => *y*)** < 1 means that there is a negative correlation within the itemset, i.e., products in itemset, *x* and *y*, are unlikely to be bought together.

Conviction

Conviction of a rule can be defined as follows:

Conv (*x* => *y*) = $\frac{1-supp(y)}{1-conf(x=>y)}$

i.e.,

Conv ({wine, chips} => {bread}) = $\frac{1-supp(bread)}{1-conf(\{wine,chips\}=>\{bread\})}$ = $\frac{1-\frac{4}{6}}{1-\frac{2}{3}}$ =1

Its value range is [0, +∞].

- **Conv (*x* => *y*) = 1** means that *x* has no relation with *y*.
- Greater the conviction higher the interest in the rule.

Let us understand the meaning of support, confidence and lift with the help of an interesting rule.

Rule: Milk, Bread => Butter, Support = 0.5, Confidence = 0.846 and Lift = 1.241.

The support value for this rule is 0.5. This number is calculated by dividing the number of transactions containing 'Milk,' 'Bread,' and 'Butter' by the total number of transactions.

The confidence level for the rule is 0.846, which shows that out of all the transactions that contain both "Milk" and "Bread", 84.6 percent contain 'Butter' too.

The lift of 1.241 tells us that 'Butter' is 1.241 times more likely to be bought by the customers who buy both 'Milk' and 'Butter' compared to the default likelihood sale of 'Butter.'

Chapter –4
Evaluation & Results

4.1 About Dataset:

The specific data of Online Retails comes from the UCI Machine Learning Repository and represents transactional data from a UK retailer from 2010-2011. This mostly represents sales to wholesalers so it is slightly different from consumer purchase patterns but is still a useful case study.

The top 5 rows of the dataset are as follows:

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kin

Then we do some data cleaning and do one-hot encoding on the whole data over the dataset.

After the cleanup, we need to consolidate the items into 1 transaction per row with each product 1 hot encoded. For the sake of keeping the data set small, I'm only looking at sales for France. However, in additional code below, I will compare these results to sales from Germany. Further country comparisons would be interesting to investigate.

There are a lot of zeros in the data but we also need to make sure any positive values are converted to a 1 and anything less the 0 is set to 0. This step will complete the one hot encoding of the data and remove the postage column (since that charge is not one we wish to explore)

Now that the data is structured properly, we can generate frequent item sets that have a support of at least 7% (this number was chosen so that I could get enough useful examples):

The final step is to generate the rules with their corresponding support, confidence and lift

```
In [10]: rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
         rules.head()
```

Out[10]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.096939	0.102041	0.073980	0.763158	7.478947	0.064088	3.791383
1	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.102041	0.096939	0.073980	0.725000	7.478947	0.064088	3.283859
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
4	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.102041	0.094388	0.073980	0.725000	7.681081	0.064348	3.293135

That’s all there is to it! Build the frequent items using apriori then build the rules with association_rules.

4.2 Results

Now, the tricky part is figuring out what this tells us. For instance, we can see that there are quite a few rules with a high lift value which means that it occurs more frequently than would be expected given the number of transaction and product combinations. We can also see several where the confidence is high as well. This part of the analysis is where the domain knowledge will come in handy. Since I do not have that, I’ll just look for a couple of illustrative examples.

We can filter the dataframe using standard pandas code. In this case, look for a large lift (6) and high confidence (0.8)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
17	(SET/6 RED SPOTTY PAPER PLATES)	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.127551	0.132653	0.102041	0.800000	6.030769	0.085121	4.336735
18	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.137755	0.127551	0.122449	0.888889	6.968889	0.104878	7.852041
19	(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.127551	0.137755	0.122449	0.960000	6.968889	0.104878	21.556122
20	(SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETRO...	(SET/6 RED SPOTTY PAPER PLATES)	0.102041	0.127551	0.099490	0.975000	7.644000	0.086474	34.897959
21	(SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY...	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.122449	0.132653	0.099490	0.812500	6.125000	0.083247	4.625850
22	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	(SET/6 RED SPOTTY PAPER CUPS)	0.102041	0.137755	0.099490	0.975000	7.077778	0.085433	34.489796

In looking at the rules, it seems that the green and red alarm clocks are purchased together and the red paper cups, napkins and plates are purchased together in a manner that is higher than the overall probability would suggest.

What is also interesting is to see how the combinations vary by country of purchase. Let's check out what some popular combinations might be in Germany.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(PLASTERS IN TIN CIRCUS PARADE)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.115974	0.137856	0.067834	0.584906	4.242887	0.051846	2.076984
6	(PLASTERS IN TIN SPACEBOY)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.107221	0.137856	0.061269	0.571429	4.145125	0.046488	2.011670
11	(RED RETROSPOT CHARLOTTE BAG)	(WOODLAND CHARLOTTE BAG)	0.070022	0.126915	0.059081	0.843750	6.648168	0.050194	5.587746

It seems that in addition to David Hasselhoff, Germans love Plasters in Tin Spaceboy and Woodland Animals.

In all seriousness, an analyst that has familiarity with the data would probably have a dozen different questions that this type of analysis could drive. I did not replicate this analysis for additional countries or customer combos but the overall process would be relatively simple given the basic pandas code shown above.

Chapter – 5
Applications

1. Find products with affinity to be sold together.

A lot of research has been done in marketing to show that there are demand interdependencies among certain related products within a single store. Retailers tend to exploit this tendency by adjusting price promotions in a profit-maximising way. They can also exploit these product associations by incorporating them into promotional strategies. Analysing purchases in multiple categories allows retailers to benefit from promotion and other marketing activities. Incorporation of product interdependencies into a pricing strategy is an effective way of boosting profits.

For example, Mulhern and Leone study the impact of price promotions on cake mix and cake frosting. Their main objective is to evaluate the overall profitability of implicit price bundling. Reducing the price of cake mix increase purchases of both cake mix and frosting and the overall profit improves. The study shows how promotions have positive impact on the sales of a complementary product.

Finding associations between product purchases is an effective way to adjust price promotions better and make better predictions on the effect of price bundling. Also, it is important to keep product complementarities in mind when making promotions. Complementary products often sell well together but this does not mean that they are a pair and a price increase in one of the sets will not affect sales of the other one. Complementarity gives managers control over their customers' buying behaviour, but co-occurrence of specific product categories in a single shopping basket is less controllable. Market basket analysis reveals all the underlying patterns of buying behaviour that cannot be simply observed. Analyzing shopping baskets also shows multi-category dependencies across products which allows retailers to bundle new products that have not been discovered yet as a set.

2. Improve in-store settings and optimize product placement.

Gaining insight on product interdependencies can help retailers optimize store layout. It is an important aspect of retailing business because in-store settings may help increase sales if done right. It also influences buying behaviour, store traffic and the whole shopping atmosphere. If market basket analysis reveals that certain products are often purchased together, it is of great interest for the retailer to put these two items or categories of products close to each other to facilitate the customer.

Optimization of in-store settings may help improve shopping experience by reducing congestion and saving time for customers. With the right space planning the store benefits from increased cross product sales and impulse purchases. Moreover, store layout and atmosphere have a very strong impact on customer perceptions. Various dimensions of store layout have positive effect on customers' purchase intentions and loyalty. This is why it is so crucial to extract knowledge from data so one can adjust store settings in order to improve customers' shopping experience.

3. Improve layout of the catalogue of e-commerce site.

Visual displays of products apply also to the catalogue of the firm online site. Ecommerce website interface plays significant part of customers' perceptions. A key success factor for profitable e-commerce site is the layout. In order to be able to determine an optimised layout for website it is important to know the interdependencies among different products.

A lot of research has been done in finding an optimal location, colouring and design for catalogues of e-commerce sites. The last step of successfully implementing a website strategy is to know how to place different products in order to maximize cross-sales. For instance, if we know which products have affinity to be sold together, we have to make sure that they are side by side on the same page on the website. It is also possible to provide discount in the form of shipping benefits for a group of products that have higher probabilities of selling together.

4. Control inventory based on product demand.

For the recent years, with more powerful analytical software it is possible to predict almost everything. It is now feasible to predict product demand based on data from past purchases, for example. For this objective it is important to know which products are related in terms of cross-sales.

Being able to find the probability of purchase for each product or a certain set of products is essential for controlling inventory. analyze the nature of the relationship which exists between price, promotion, sales and consumption. The authors' main finding is that price promotions encourage stock-pilling, while on the other hand stock-pilling rationally leads to increase in consumption.

Chapter - 6

Conclusion & Future Scope

6.1 Conclusion

Taking everything into account, by summing up the entire, we think this framework will have an efficient effect on marketing and sales analysis that can be used to make strategic business decision. This application can be extended to other fields like – sales tracking, product tracking, discount and pricing calculation etc. In future this method can be applied to very large databases where memory space is valuable and requires optimization. It can be further tuned for better performance and efficiency.

Limitations:

The limitation first of all and most importantly we face about getting real life dataset. We have gone to every single big super shops where ever possible for their transaction data to conduct our thesis but they refused to give us. So, with the help of our supervisor sir we work with the available online retail dataset.

6.2 Future work

In future we will try to implement new and advanced mining algorithm along with apriori, fp growth and for better performance and fast result for sparse dataset. We will also build a user-friendly web application where the user will be able to select as many items they wish and will be able to see the corresponding support count on a bar chart. From this they will be able to visualize and realize the relationship among particular items. The organizer of a super shop may use this application to arrange their products in the shop to boost their sell. Beside market basket data association analysis can be applied in other fields such as bio informatics, medical diagnosis and scientific data analysis

References

[1] Saurabh Malgaonkar, Sakshi Surve and Tejas Hirave. “Use of Mining Techniques to Improve the Effectiveness of Marketing and Sales” IEEE Trans. Paper id-63.

[2] Mahmoud Houshmand, Mohammad Alishahi. “Improve the classification and Sales management of products using multi-relational data mining”. IEEE Journals,2011. 978- 1-61284-486-a2/111

[3] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, Aijun An. “Mining Online Reviews for Predicting Sales Performance: A Case Studyin the Movie Domain”. IEEE Transactions On Knowledge and Data Engineering, VOL. 24, NO. 4, APRIL 2012.

[4] Julie Marcoux, and Sid-Ahmed Selouani, Universit«e de Moncton, Campus de Shippagan, Canada. “A Hybrid Subspace-Connectionist Data Mining Approach for Sales Forecasting in the Video Game Industry. 2009 World Congress on Computer Science and Information Engineering.

[5] Wan Faezah Abbas, Nor Diana Ahmad, Nurlina Binti Zaini, “Discovering Purchasing Pattern of Sport Items Using Market Basket Analysis”, International Conference on Advanced Computer Science Applications and Technologies, 2013.

[6] Han, J., Kamber, M. “Data Mining Concepts andTechniques”, Morgan Kaufmann Publishers, 2006.

[7] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "New algorithms for fast discovery of association rules," in Third International Conference on Knowledge Discovery and Data Mining, 1997.

[8] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Ïntroduction to Data Mining”