Profit Maximizing Recommendation using Market Basket Analysis

Kush Tekriwal
University of Washington
kusht@uw.edu

Mayur Gupta
University of Washington
mayu07@uw.edu

Vishwa Pardeshi
University of Washington
vishwp@uw.edu

Abstract

For our final project, we have partnered with MG Impex, a granite manufacturer in India. We leverage Market Basket Analysis to exploit customer purchasing patterns to recommend a profitable promotion campaign for each market segment. To mine association rules, we utilize the Apriori algorithm with tuned threshold values. We present and compare two approaches to determine the most profitable campaign.

1 Introduction

A key issue in traditional manufacturing or processing companies are their inability to leverage data for competitive advantage. Our sponsor for this project, MG Impex Pvt. Ltd., is a prime example. MG Impex is a 15 years old leading processor of natural stones in India. They process granite blocks and export slabs all around the world. Currently, the utilization of data is manual and heavily relies on the knowledge of a few key executives.

This dependence is inefficient, as marketing agents must consult higher officers before taking any decision. Another drawback is that marketing decisions are often based on domain knowledge without evidence from data. This approach will not be successful in the long term. To ensure that MG Impex continue to be pioneers in their field, it is important that they use data to tackle these problems.

The goal of the project is to increase profits and decision-making efficiency for MG Impex by exploiting customer purchasing behaviour and domain-specific knowledge to recommend a profitable promotion campaign. One key domain-specific knowledge shared by Subject Matter Experts relates to the role market segments plays in influencing a customer's purchasing behaviour. For example, Europe buys plain colours while the US buys vibrant colours.

Thus, the recommended profitable promotion campaign will contain market specific sequential recommendation of products which maximize profits. The sequences of products are obtained by mining market segment-specific association rules using Apriori Algorithm. For identification of sequences which maximize profits, two approaches - Nearest Neighbour and Greedy Approach are implemented and compared. These data-informed promotion campaigns will make the decision making process more robust and help MG Impex gain a competitive edge.

2 Previous Work

Market Basket Analysis (MBA) has found use in a lot of retail companies to devise cross-selling and promotional strategies by mining association patterns among the large amount of sales data generated. This statistical affinity analysis has been carried out by using various different algorithms such as Apriori, FP Growth, etc. However, in addition to being inefficient in its space utilization, market basket analysis might generate trivial rules due to inefficient handling of data granularity.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Inv. No	Date	Customer	Material	Owner	Colour	Block No.	2cm Slab	3cm Slab	2cm Area	3cm Area	Sale Value	Cost
295	3/7/20 0:00	Customer A	Gangsaw	M & G	Coffee Brown	2005	11	40.0	NaN	1182.244875	4.697295e+05	128247.0
295	3/7/20 0:00	Customer A	Gangsaw	M & G	Viscont White	1236	2	7.0	910.333008	1970.408125	7.828826e+05	456117.0
296	3/6/20 0:00	Customer B	Gangsaw	M & G	Thunder White	MIX	NaN	NaN	227.583252	NaN	3.871847e+04	356511.0
296	3/6/20 0:00	Customer C	Gangsaw	M & G	Viscont White	1994	22	NaN	1311.259716	NaN	1.001449e+06	452155.0

Figure 1: Data Sample Anonymized for Confidentiality

Previous work that we followed used MBA to derive association rule set and uses them in different ways. The paper [3] creates market basket networks and co-purchased product networks to analyzes the properties of the graphs. They compare different properties of networks to analyze the difference between co-purchased product network and market basket network is that the products in the latter network is linked when they are purchased at a same time, while the products in the former one is linked when they are purchased by a same customer only.

On the other hand the, paper[2] introduces methods to use the product affinities to predict ways to increase revenues, and estimate the magnitude of the possible increases as a function of customer price sensitivity and affinity saturation level. They also distinctly explain why using lift over support and confidence is a better way of approaching profit related problems. Their experiment with price sensitivity shows that if used correctly the association rules from Market basket analysis can impact the profits in a strong way.

In our report we combine Market Basket Analysis and Profit Maximization to identify the possible recommendation for MG Impex.

3 Data

3.1 Data Description

To understand the customers' purchasing behaviour, we were granted access to the following:

- Competitors' Information: This contains competitor's information such as selling customer, buying customer, product, and country of buyer.
- Sales Report : This includes invoice number, customer, product, sale value, cost and area of the material which was collected by MG Impex's Export team over the course of 3 years.

The Sales Report data is shown in Figure 1 captures a single product for a particular invoice per row. This sales data is aggregated by month in separate excel sheets.

3.2 Exploratory Data Analysis

To ensure that the Subject Matter Expert's insights regarding the role market segment plays in influencing a customer's purchasing behavior, we explored the hot-sellers for each market segment and as a whole. While it was noticed that the hot-sellers irrespective of the market segment included only 4 common products. These products can be observed in Figure 2 at the far right corner. This popularity across all market segments is due to the neutral colors - black and white of the product. However, beyond this the colors have a very specific market segment associated to it. Infact a few countries opt only for variants of the same color.

4 Methodology

4.1 Data Preprocessing

Primary preprocessing includes preparing the data for Market Basket Analysis Profit Maximization. The sales competitors' data from different excel sheets is merged together and cleaned by

1. Eliminating corporate identifier such as 'ltd','co', and 'inc'.
2. Typographical errors

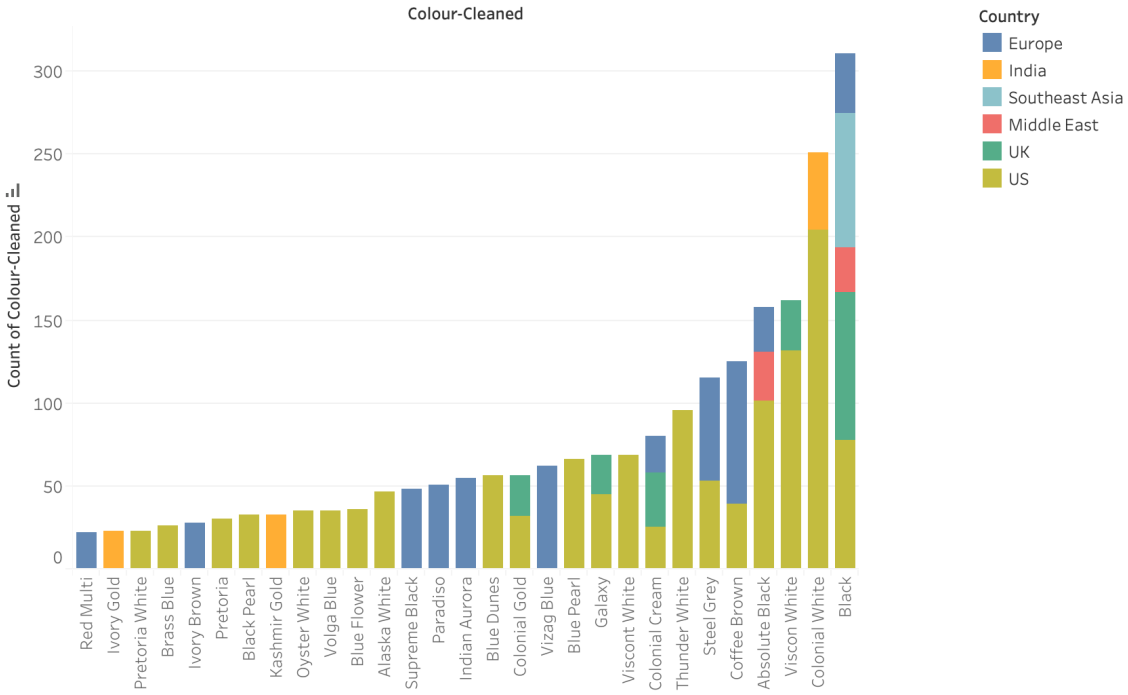

Figure 2: Product Frequency Plot for different Market Segment

3. Standardizing measurements units which varied across market segments.

To further prepare the sales data for Market Basket Analysis, we extracted customer's country by joining the sales data with competitors' data. The sales data was then transformed by grouping by invoice numbers to combine products sold together into a single string delimited by '|'. To ensure that the interesting rules don't go undetected, we adjusted the granularity of the sales data. This was achieved by creating a taxonomy of the products by grouping similar versions of products and market segment specific data chunks.

The cleaned transformed sales data was additionally preprocessed to enable profit maximization by generating a product to profit mapping. This mapping for a particular product is generated by calculating the unit profit. Unit profit is the difference between the unit cost and unit price of the product's material. Here, unit cost unit price is calculated by dividing the cost price of the product's material by the area of the purchased material. Since there is variability for a product's profit due to discounts or damage, we averaged the unit profit across transactions.

4.2 Market Basket Analysis

Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence [1]. A co-occurrence is when two or more things take place together. The antecedent is the condition and the consequent is the result. The association rule has three measures that express the degree of confidence in the rule, Support, Confidence, and Lift.

1. Support: Measure of how frequently the collection of items occur together as a percentage of all transactions.
2. Confidence: Confidence of the rule denotes how often the consequent appears in transactions that contain the antecedent only.
3. Lift: It is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of the consequent.

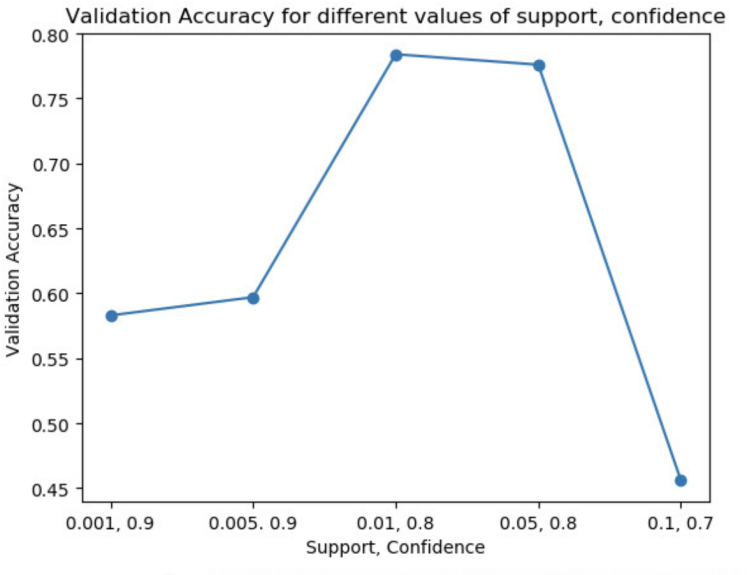

Figure 3: Validation Accuracy for Different Values of Support Confidence

Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

Traditional market basket analysis, however, yields limited insights due to data sparsity. In practice, this is considered one of the major drawbacks of association rule mining: if the support is set high, fewer rules are generated but most of them might be obvious and hence useless. On the other hand, if the support is set low, too many rules are generated and the domain experts have to evaluate the generated rules and identify the ones that are useful.

Thus, we run Apriori for every market to integrate customer information and context and, consequently, generate non-trivial association rules. Our several experiments in fine-tuning thresholds to generate sufficient interesting rules are observed in Figure 3. We determine support = 0.01, confidence = 0.8, max length = 4 to be appropriate values for our analysis.

An important step in this section is the evaluation of our approach. We split our cleaned data set into a train and validation set using a 80/20 split. We run Apriori on the training set for each market and measure coverage on the validation set. Specifically, we count the number of rules in the validation set that are created from any of the Apriori runs. We observed that 0.78 of all rules in the validation set are covered. Following this result, we reran our analysis using the entire data set.

4.3 Profit Maximization

After mining association rules, with profit as our objective function we select rules which can be recommended to our sponsor. MG Impex and other companies in the industry often accept orders in spaced time intervals rather than in bulk. This is done due to high raw material cost and to assist in procurement planning. Additionally, due to the high volume of customers and limited capacity, it is important to meet maximum customer's requirements. For this use case, a sequence is more useful than a basket of recommendations.

To provide a series of recommendation, we decided to form a chain of recommendations for the client. The chain created at each step takes into consideration the previously provided rules. For example, if the previous recommendation is A -> B, the next recommendation will be based on the rule described by A,B. Unfortunately, the combination of A,B is not always found in the association rules set. Therefore we propose two alternate methods to provide a series of three recommendations.

We use the concept of using graphs to express association rules from [3], where they represent the association rules using undirected graph with support and confidence values on the edges. We introduce a novel approach by weighting edges by lift. Lift is our choice of metric as its a symmetric

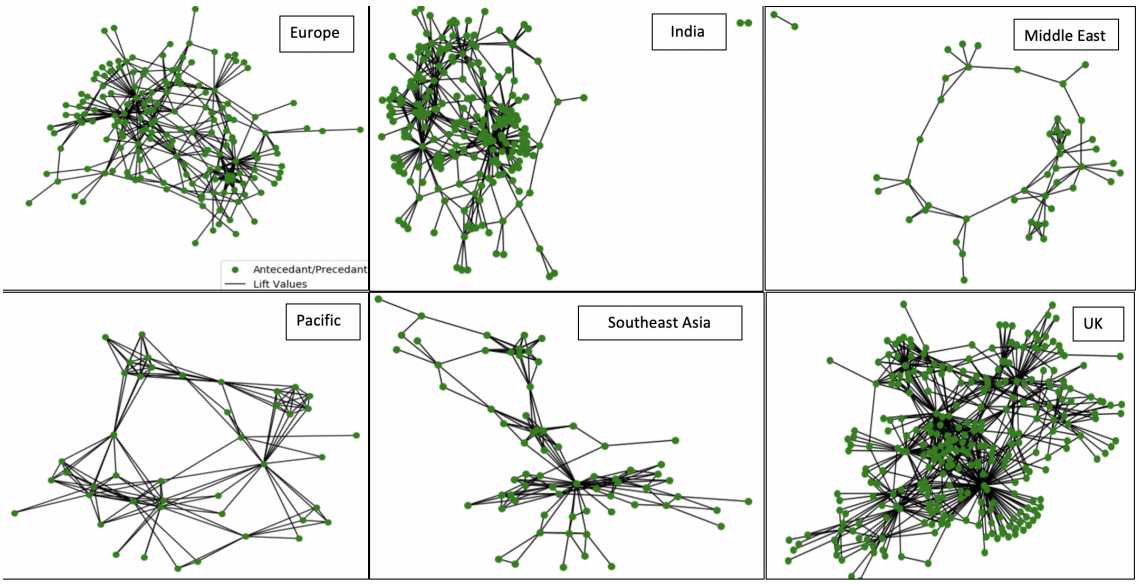

Figure 4: Undirected-Grpah for different Market Segment

metric and hence can be easily described using an undirected graph. As described in [2], finding pairs of items with high lift could lead to increase revenue by selling more items, as support and confidence include misleading information about the nature of the affinity.

The graphs for the market segments are shown in the Figure 4. These graphs and how they differ largely from one another in structure and number of rules further helps bolster our assumption that market segment influences purchasing behavior.

4.3.1 Profit Calculation

Before we describe the two approaches, we define two possible cases of profit calculation that has been used through the course of the project. As described in the previous section, we have created a product to profit mapping. Generally, if we have T, total recommendations and P is the sum of profit contribution of T recommendations, then the total profit is defined as $\frac{P}{T}$

Additionally, when defining the profit for association rules set, the calculation of profit needs to be altered to accommodate the different rule set from Apriori algorithm. The Apriori algorithm provides rules associated with a single product as well as combination of products, therefore for both the cases the profit is defined as follows:

- Profit calculation for a single product: Let us assume a recommendation of product A with profit R associated with it. Now when the product is encountered in the course of calculation of profit, we use the profit value R as it is, without any alteration.

- Profit calculation for a combination product: Let us assume a recommendation of combination product [A,B,C] with profit associated with each product in combination being $[R_1, R_2, R_3]$ respectively. Now when the product is encountered in the course of calculation of profit, we calculate the profit contribution for the combination product as $\frac{R_1+R_2+R_3}{3}$. Generally, it can be defined as if the product is a combination of n items with profit for each product being $R_1, R_2, R_3, ...R_n$, then the profit contribution of the product is defined as $\frac{R_1+R_2+...+R_n}{n}$

The normalization for combination of products is important so that multi-product rules do not outweigh the single product rules.

4.3.2 Nearest Neighbors

The first profit calculation approach using the undirected association rules graph used Nearest Neighbors. We find the three nearest neighbours with maximum profit using the profit calculation described above. Our final recommendation is the product and its neighbors that will generate the maximum sum of profit contributions.

For experimental purposes, we also find the nearest neighbor with maximum lift. This offers an insightful comparison between lift and profit recommendations. We present two cases to describe this comparison:

- Case 1: Different products recommended by lift and profit recommendations
  We ran our algorithm for the product "steel grey satinato" in the UK market. The recommendations while considering the maximum lift were 'indian aurora', 'red multi' and 'black pearl' whereas the profit maximization gave the recommendations of 'indian aurora', 'colonial white' and 'coffee brown'. Tables 1 and 2 present the profit values.
  We observe a profit of 16.90 Indian Rupees/sq. feet [$\frac{105.88-77.12+21.96}{3}$] using lift recommendations and a profit of 60.61 Indian Rupees/sq. feet [$\frac{105.88+47.83+28.12}{3}$]using profit recommendations.
  When we presented these results to the company, they preferred the profit recommendation over the lift recommendations. They agreed to the fact that lift recommendations aligned better with subject matter experts views but the profit margin was too large to ignore. The company also concurred that the profit recommendation are accurate and feasible to sell in the market.
- Case 2: Same product recommended by lift and profit recommendations
  There are cases when a product has exactly 3 or less number of association rules. When there are fewer rules than neighbours to find, the recommendations from both lift and profit are the same.

Table 1: Recommendations Using Lift Based Nearest Neighbors for Case 1

	indian aurora	red multi	black pearl
Lift	58.0	29.0	14.5
Profit Contribution	105.88	-77.12	21.96

Table 2: Recommendations Using Profit Based Nearest Neighbors for Case 1

	indian aurora	colonial white	coffee brown
Lift	58.0	4.46	8.28
Profit Contribution	105.88	47.83	28.12

The advantage of the process is that we have strong connection between the initially bought item and the three recommendations we have provided. The downside of the process is that often a product only has a single rule with other product or combination of product due to data sparsity. Additionally, this approach does not solve our use case as to recommending a sequence since the three nearest neighbors may be all in a single basket. In such cases, we are not able to provide the correct recommendation to the company.

4.3.3 Greedy Approach

The second approach finds all possible paths of length three from a product. This greedy approach takes products sequentially based on the profit value and create a chain of recommendations, that can be provided one after the other. Again, we can repeat this process using the maximum lift value for experimental purposes.

The process to find the lift based greedy sequence and profit based greedy sequence are very similar with a small difference.

- Lift-based Greedy Approach: For the lift based greedy approach, we use tree traversal based on the weights on the edge (lift). Starting from an antecedent, the algorithm iteratively selects the next recommendation based on the maximum lift edge between two nodes. The process is run until we have a path of length 3 or, in other words, 3 sequential recommendations.

- Profit-based Greedy Approach: For the profit based greedy approach, we use tree traversal based on the condition of maximizing the profit.

Hence, the difference in the two process lies in the fact that we use the values of lift between two products obtained from the association rules from Apriori algorithm, whereas the profit based greedy approach assumes all edges to unit weight and uses the profit at each node as the metrics for selection.

An assumption for the approach is that if A->B and B->C, then there exist a relation between A and C, which diminishes as the number of nodes increases. To address this assumption, we chose to recommend a series of three products, such that the relation between the first and the last recommendation is not too weak.

To prove the above assumption let's first observe Bayes rule with extra conditioning.

$$\begin{aligned} P(a,z|b) &= \frac{P(a,z,b)}{P(b)} \\ &= \frac{P(z,b)P(a|z,b)}{p(b)} \\ &= \frac{P(b)P(z|b)P(a|z,b)}{P(b)} \\ &= P(z|b)P(a|z,b) \\ &= P(a|z,b)P(z|b) \end{aligned}$$

The intuition behind the assumption is based on the above. Now we consider a path A->B->C->D, and try to define the recommendation D. Since we are using lift as metric of evaluation, mathematically the trust in recommendation D can be written $\frac{Confidence((A,B,C)->D)}{P(D)}$. Now let us see how D depends sequentially on each predecessor.

Using the above equation we can write $Confidence((A,B,C)->D)$ as:

$$P(D|C,B,A) = \frac{P(D,C,B|A)}{P(C,B|A)}$$

For simplicity assume Z = (C,B) and b = A. Let us assume $X = B|A$, therefore

$$P(D|C,B,A) = \frac{P(D,C,X)}{P(C,X)}$$

By conditional probability we know $P(I,J) = P(I|J)P(J)$, hence the equation converts to:

$$P(D|C,B,A) = \frac{P(D|C,X)P(C,X)}{P(C,X)}$$
$$P(D|C,B,A) = P(D|C,X)$$

The above equation implies that $P(D|C,B,A)$, is simply the probability of D given C and B given A. But this proof does not show the clear relation between C and $B|A$.

Now applying equation one again to $P(D|C,X)$, we get

$$P(D|C,B,A) = P(D|C,X)$$
$$P(D|C,B,A) = \frac{P(D,C|X)}{P(C|X)}$$

Assume $C|X = Y$, we have:

$$P(D|C,B,A) = \frac{P(D,C|X)}{P(C|X)}$$
$$P(D|C,B,A) = \frac{P(D,Y)}{P(Y)}$$

According to Bayes theorem this is nothing but $P(D|Y)$, hence

$$P(D|C,B,A) = \frac{P(D,Y)}{P(Y)}$$
$$P(D|C,B,A) = P(D|Y)$$

Therefore from above equation we can say that $P(D|C,B,A)$, is nothing but probability of D given C given B given A. Lift is nothing but $\frac{P(D|C,B,A)}{P(D)}$. Hence we can say that there is a connection of D from not only C, but also from B and A

This approach has the advantage of maintaining connections, although weak, with all the nodes in the recommendation and provide us with the desired number of recommendation. Another benefit is that the more the customer selects the recommended product, a pattern establishes and also indirectly supports the accuracy of our recommendation.

To help better understand the Greedy approach, we use an example where nearest neighbor fails to provide us the appropriate number of recommendation.

We run the nearest neighbor approach for UK market on the product combination 'colonial gold,colonial white'. The recommendations for the combination are 'black premium', 'colonial cream'. As mentioned in the nearest neighbor section, any product with less than 3 association rules will have same lift and profit recommendation. Considering the problem of the company, providing only two recommendations puts them in a loss. Therefore we needed a new way to get the desired number of recommendations. We run the greedy approach for the same product combination 'colonial gold,colonial white' and derive the results for both lift and profit recommendations.

The greedy approach with lift maximization suggest the sequence of products 'black premium', 'colonial cream', 'absolute black,ivory gold'. The recommendation can be interpreted as 'black premium' is the immediate next node. The lift values of each product with all its predecessors are given in the Table 3. We list the individual profit contributions in Table 4. The total profit for the recommendation is 53.98/per sq feet $[\frac{119.16+73.22+23.52}{4}]$. Note that we divide by 4, as the final recommendations contains 2 products.

Table 3: Lift Values Using Lift Based Greedy Approach

	colonial gold,colonial white	black premium	colonial cream	absolute black,ivory gold
colonial gold,colonial white	NA	47.0	5.87	4.7
black premium	47.0	NA	23.5	5.875
colonial cream	5.87	23.5	NA	9.44
absolute black,ivory gold	4.7	5.875	9.44	NA

Table 4: Profit Contributions Using Lift Based Greedy Approach

	black premium	colonial cream	absolute black,ivory gold
Profit Contribution	119.16	73.22	23.52

Based on the lift values in Table 3, we can see that our assumption of connection between first and last node still holds.

Similarly, the greedy approach with profit maximization suggests the sequence 'black premium', 'colonial cream', 'steel grey satinato'. The lift values of each product with its predecessors are given in the Table 5. The individual profit contributions are listed in Table 6. The total profit for the recommendation is 119.04/per sq feet $[\frac{119.16+73.22+164.75}{3}]$.

Table 5: Lift values Using Profit Based Greedy Approach

	colonial gold,colonial white	black premium	colonial cream	steel grey satinato
colonial gold,colonial white	NA	47.0	5.87	3.91
black premium	47.0	NA	23.5	5.875
colonial cream	5.87	23.5	NA	9.44
steel grey satinato	3.91	4.7	7.83	NA

Table 6: Profit Contributions Using Profit Based Greedy Approach

	black premium	colonial cream	steel grey satinato
Profit Contribution	119.16	73.22	164.7481527

From the lift values in Table 5, we observe that our assumption about connection between first and last node holds. We highlight the difference in profit, which MG Impex feel is too high to ignore.

5 Results

In this section, we present the final recommendations for each market after employing the method described above. When there are sufficient neighbours, we prefer to report the nearest neighbour results. We also use the profit based approach, as this is preferred by MG Impex.

1. Europe: 'marina blue' -> 'emerald pearl' -> 'vizag blue' -> 'colonial cream'
2. UK: 'colonial gold, colonial white' -> 'black premium' -> 'colonial cream' -> 'steel grey satinato'
3. India: 'red multi colour' -> 'red multi colour' -> 'oyster white' -> 'belvedre' -> 'premium black'
4. Middle East: 'blue flower' -> 'galaxy' -> 'ivory fantasy' -> 'pretoria'
5. South-east Asia: 'colonial white, steel grey' -> 'blue dunes' -> 'absolute black' -> 'black satinato'
6. USA: 'galaxy, pretoria' -> 'colonial white' -> 'colonial gold,viscont white'
7. Pacific: 'coffee brown' -> 'belvedre' -> 'emerald pearl' -> 'brass blue'

Out of all markets, only Middle East recommendations was generated from nearest neighbours. The rest were generated using the greedy approach.

6 Discussion

The profitable promotion campaign was recommended by using profit-based greedy approach. This method was preferred by the company's subject matter expert as well. However, we assume fixed pricing rather than adjusting prices for item bundles. A final consideration must include the effect of multiple pairings of items on the customer's choice. Research has shown that too many choices will distract the customer and might lead either to no decision or to a sub-optimal decision. This is another benefit of our method, as the number of pairings were kept relatively small or localized, so each customer is provided with a limited choice.

7 Conclusion

Our work is projected to improve their profit margin by INR 270/ sq. feet. This was calculated by summing profit projected for all the rules, assuming that the company chooses to implement a promotional strategy including all the rules generated for that particular promotional strategy.We hope the results of our work inspire other processing and traditional manufacturing companies to see the value of data. We would be open to conducting a similar analysis for other companies in the future.

8 References

[1] (n.d.). Retrieved from https://infocenter.informationbuilders.com/wf80/index.jsp?topic=/pubdocs/RStat16/source/topic49.htm

[2] Hoanca, B., Mock, K.J. (2011). Using Market Basket Analysis to Estimate Potential Revenue Increases for a Small University Bookstore.

[3] Kim, H. K., Kim, J. K., Chen, Q. Y. (2012). A product network analysis for extending the market basket analysis. *Expert Systems with Applications,* **39**(8), 7403–7410. doi: 10.1016/j.eswa.2012.01.066

9 Appendix

In this section, we outline the mathematical background for our algorithms.

9.1 Apriori

The Apriori algorithm is a two pass approach that limits the need for main memory. The key idea is that if an item i does not appear in s baskets, then no pair including i can appear in s baskets. In the first pass, we count the number of occurrences of each individual item. Items that appear more than given support threshold are considered frequent items. In the second pass, we only count pairs in main memory where both elements are frequent. We repeat this iterative process until the specified max length of combinations is reached.

9.2 Market Basket Analysis

Market Basket Analysis utilizes multiple metrics to asses the association rules. The rules are defined as:

$$\text{Support(X -> Y)} = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ Transaction}$$

$$\text{Confidence(X -> Y)} = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

$$\text{Lift(X -> Y)} = \frac{Confidence(X->Y)}{Support(Y)}$$