Finding *n*-Most Demanding Products for Real-Estate Industry

Aditya Chimbalkar
Research Scholar, VIIT,
University Of Pune
Pune, Maharashtra 411048,
India

Swati Patil
Assistant Professor, VIIT,
University Of Pune
Pune, Maharashtra 411048,
India

Mridul Jain
Research Scholar, VIIT,
University Of Pune
Pune, Maharashtra 411048,
India

Ganesh Kate
Research Scholar, VIIT,
University Of Pune
Pune, Maharashtra 411048, India

Priyank Dwivedi
Research Scholar, VIIT,
University Of Pune
Pune, Maharashtra 411048, India

ABSTRACT
In this paper, real estate problem, n-most demanding products (n-MDP) discovering, is formulated. A set of customers demanding a real estate product with multiple features, there are existing products in market, preferred products set which is offered by the company, and a positive integer n, which will be found to help the company to select n products from preferred products such that expected number of the total customers for the n products is maximized. Here it is shown the problem is NP-hard when the number of features for a product is 3 or more. Bit Map Index (BMI) is developed on the customer requirements and products of real estates. The Output of the BMI is given as an input to the algorithm like Apriori. An Apriori algorithm proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. An Apriori Algorithm is provided to find the optimal solution of the problem by using the pruning strategy. The n value is determined by the algorithm which will help the company to maximize the product.

General Terms
Data mining, Preferred Product, Customer preferences

Keywords
Data mining, Algorithm, Apriori, Real Estate, Decision Support, Query Processing, Bit Map Index.

1. INTRODUCTION
MICROECONOMICS is the branch of economics that deals with the interaction of consumers and producer as how they make decisions in the market. Customer preference is an important factor in making decisions of product sales which is one of the key factors in Microeconomics. Also Microeconomics looks at the smaller picture and focuses more on basic theories of supply and demand and how individual businesses decide how much of something to produce and how much to charge for it. People who have any desire to start their own business or who want to learn the rationale behind the pricing of particular products and services would be more interested in this area. Taking competition into consideration, the problem studied in this paper is to identify the production plan with the highest utility for a company, where the utility of a production plan is evaluated according to the expected number of the total customers for the selected products in the plan.

There is a need to have a lookup into product's popularity, a product that satisfy the maximum number of customer preferences and is admired by most of the people .Here in this, paper, we are going to address this kind of product as N product that is the result of intensive data traversal and comparisons.

2. LITERATURE SURVEY
Data analysis is important to businesses will be an understatement. In fact, no business can survive without analyzing available data [2].Customer preference is an important factor in making decisions of product sales, which thus becomes one major concern in microeconomics. Kleinberg et al. [1] claimed that several microeconomic problems can be solved by data mining techniques, which motivate the researchers in the database community to deal with the microeconomic problems. Data Mining (DM) is the extraction of new knowledge from large databases. Many techniques are currently used in this fast emerging field, including statistical analysis and machine learning based approaches [3]. Using the found merits to promote the product should have the higher opportunity to attract more customers' attention than the manner in the first type. Nevertheless, the works in this type focus on an existing product whose characteristics are fixed, and it is possible that most customers are not interested in the product.

Nevertheless, the attention of the product advantages discovery is centered on the product whose characteristics have been known, and consequently the product may not satisfy the customers even though its merits are known. In recent years, new studies in [7], [8], and [9] appeared that tackled the issue of product positioning strategies. The purpose of the studies in this type is to help companies develop new products satisfying the needs of the customers within the target market, which is also the goal of this paper. Extended from [7], suppose there are numerous companies with their respective profit constraints and a set of customer requirements, by taking competition into consideration, the goal of [9] is to find one product with the maximum expected number of the customers for each company, which satisfies the profit constraint of the company. In summary, the found product in [7] and [9] has to satisfy the profit constraint of the company, which may be difficult to specify. Moreover, to

attract more customers, a company may choose to offer multiple products at the same time. . Given a set of customer requirements and the profit constraint of a company, the problem addressed in [7] is to identify the product dominating the largest customer requirements, which satisfies the profit constraint of the company.

Looking ahead, the market will continue to respond with an increasing tilt toward "easy to use" data discovery tools that offer flexibility years ahead of traditional data mining products, as well as less prohibitive costs, maintenance requirements and skilled resource demands [2]. The majority of research [1], [4], [3], [5], [10] relevant to microeconomic problems has focused on the potential customers finding. This is to help a company find out the potential customers who may be interested in its specified product, and then the company can advertise the properties product to the potential customers.

3. PROBLEM STATEMENT

In this section, it's been formally defined the n-MDP discovering problem for Real-Estate Industry that would help the builders to know the customer's preferences and to provide them with the features that are actually demanded so as to maximize sale.

4. PROPOSED SOLUTION

4.1 Building Bitmap Index

Consider a situation of real estate market where the most important decision of customer on which he/she is going to buy a property in his budget. Budget contains their maximum investment & minimum investment for a property. For making a maximum profit decision the real estate company has collected the requirement i.e. the budget range of the customer. Now assume the real estate company has many properties which they want to sale. Each property has a specific cost. The real estate company wants to compete their all properties to decide on which they can earn a maximum profit. It is assumed that customer will buy a property which will satisfy his/her requirement. If one or more properties will satisfy their requirement then the property will be selected on their implicit choice of the customer. It is assumed that the probability of selecting between two or more properties which are satisfying the requirement is equal.

Let us consider a scenario where there are products P1, P2, P3... And Customers C1, C2, C3...Products are having certain cost range (in millions) and customers have their budget depending upon the cost range. Now we plot a graph, where X-axis has Customers Budget (in millions) and Y-axis has Products Cost (in millions).We plot points depending upon the range of the product and customer. Here it is considered that customer C1 who has certain amount of range (in millions) has so we define that particular range with dotted line. The Graphs has been plot for various customers with their budget range. The graph helps in identifying the products which have their foothold in the given range by customers. The products which lie outside the range are not satisfying the budget. The following graphical representation helps more clearly understand the concept.

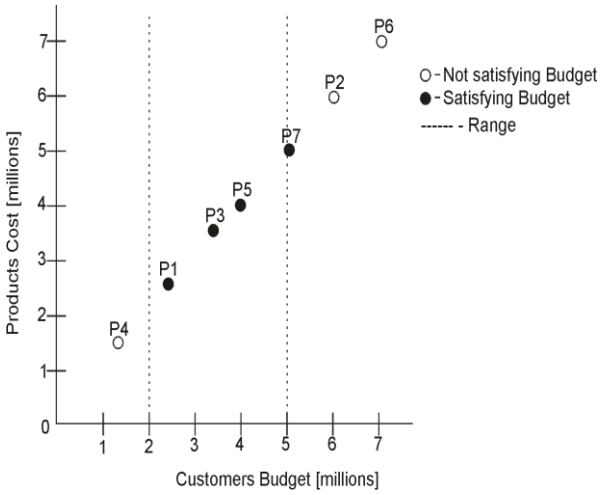

Fig. 1 (a) Customer Budget in range 2-5 (in millions)

This graph depicts the budget range of C1 customer so there are 4 products which are in vicinity of the required budget.

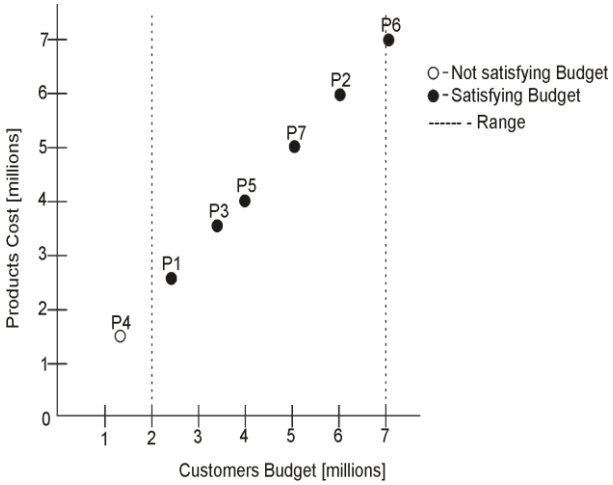

Fig.1 (b) Customer Budget in range 2-7 (in millions)

Now here in Fig.1 (a) are 6 products satisfying the budget range of the customer C2. Such type of graphs are developed for various customer budgets.

Based on the above graphs for each customers the properties which are satisfying the requirement of the customer i.e. the properties lying in the budget range of the customer are said to be 1 in the bit map index and which doesn't are said to be 0. So according to each customer the bit map index is created for all the customers and all the competing properties as shown in Table 1.

Table 1. Depicts the Bitmap Index of the satisfying products for a particular budget range

Customers / Properties	C1	C2	C3	C4	C5	C6	C7
P1	1	1	1	0	0	1	1
P2	0	1	0	1	0	1	0
P3	1	1	1	0	1	1	0
P4	0	0	1	0	0	1	1
P5	1	1	1	1	1	1	0
P6	0	1	0	1	0	0	0
P7	1	1	1	1	1	1	0

So, now the set of the products will be {P1, P2, P3, P4, P5, and P6}. And the C1 can select P1, P3, P5 and P7 out of all properties, so the probability of choosing P1 by C1 is 1/4. Similarly we can get the probability of each customer from C2 to C7 for choosing P1 product. Finally, adding the probabilities of all customers for C1 is as follows:

P1= $(\frac{1}{4}+\frac{1}{6}+\frac{1}{5}+0+0+\frac{1}{6}+\frac{1}{2})$ = 1.27

Similarly for the P2 is as follows:

P2= $(0+\frac{1}{6}+0+\frac{1}{4}+0+\frac{1}{6}+0)$ = 0.57

Similarly for the P3 = 1.1, P4= 0.86, P5= 1.35, P6= 0.41and P7= 1.35. Now as have got highest probability of Products getting selected. So, we can say most demanding product is P5 and P7. So by this we can get the top products from which a real estate company can earn a maximum profit. The performance evaluation of the Bit Map Index is explained in the Section 5.

4.2 Algorithm for Optimal Solutions

The most intuitive method to get the optimal solution is to perform exhaustive search. In other words, the subsets of Preferred Product (PP) with n preferred products are enumerated. The optimal solution is the set of n preferred products with the highest expected number of the total customers in C. The Apriori algorithm is designed based on providing some pruning strategies to reduce the search space of the optimal solution.

4.2.1 Apriori-Based Algorithm

The APR algorithm generates all the sets containing a single preferred product first. Let S denote a set of l referred products, where $1 \le l < n$. For any nPP which contains S, denoted nPPS, the main idea of the APR algorithm is to estimate the upper and lower bounds of E (nPPS; C). The bound values are used to prune the sets of l preferred products whose supersets are impossible becoming the optimal solution of the n-MDP discovering problem. In the next iteration, the remaining sets of l preferred products ($1 \le l < n$) are combined to generate the sets of (I+1) preferred products. The above process will repeat until the sets of n preferred products are generated to discover the n-MDP.

Let S denote a set of I preferred products, where $1 \le l < n$, and N(S, c) denote the number of preferred products in S satisfying customer c. Besides, N(PP, c) denotes the number of preferred products in PP satisfying customer c. It is supposed that N(PP,c) and N(S,c) are known. Let U denote the set of preferred products, whose cardinality is (n - l), which are inserted into S to form a set nPPS of n preferred products. For any set nPPS which contains S, the upper bound and the lower bound of N(nPPS; c), denoted UB_N(nPPS; c) and B_N(nPPS; c), respectively, are estimated according to N(PP; c) and N(S,c) as follows:

Upper bound of N (nPPS, c). An upper bound of N (nPPS,c) occurs when all the preferred products in U satisfying c. Moreover, since nPPS is a subset of PP, it is impossible that N (nPPS,c) is larger than N (PP,c). Therefore, N (PP,c) is another upper bound of N (nPPS, c). To get a tighter upper bound, UB_N (nPPS, c) is got as follows:

UB_N (nPPS, c) = Min (N(S,c) + (n - l),N(PP,c)).

Lower bound of N (nPPS,c). Among the (|CP| - |S|) preferred products which are not in S, there are (N (PP,c) - N(S,c)) products satisfying customer c. In other words, there are ((|PP|) - |S|) - (N(PP,c) - N(S,c))) preferred products which are not in S and do not satisfy customer c. Since the cardinality of U is (n - l), if ((|PP|) - |S|) - (N(PP,c) - N(S,c))) is larger than or equal to (n - l), a lower bound of N(nPPS,c) occurs when none of the preferred products in U satisfy c. That is, LB_N(nPPS,c) = N(S,c). Otherwise, there are at least ((n - l) - ((|PP| - |S|) - (N(PP,c) - N(S,c)))) products in U satisfying c. Therefore, LB_N(nPPS,c) is got as follows:

LB_N(nPPS,c) = Max(N(S,c),N(S,c) + ((n - l) - ((|PP - |S|) - (N(PP,c) - N(S,c))))).

In the lth iteration of the APR algorithm ($1 \le I \le n$), the set of all the sets of I preferred products, denoted PPSl, is generated by combining the sets of (l - 1) preferred products remained in the previous iteration.

Let S1 and S2 denote two sets of l preferred products in PPSl. If LB_E(nPPS2; C) is larger than UB_E(nPPS1; C); E(nPPS1; C) must be less than E(nPPS2; C). In other words, none of the sets of k preferred products containing S1 will become the optimal solution of the n-MDP discovering problem. Consequently, S1 can be pruned such that it is not necessary to generate the longer sets containing S1. The iterative process is repeated until the sets of n preferred products are generated. Finally, the n-MDP is discovered by selecting the set nPP of n preferred products with the highest E(nPP,C).

The APR Algorithm

Input: N_vec(P,C), set C of Customer requirements, the set PP of preferred products, and the value of n.

Output: a set of n preferred products.

1. For each preferred product pp in PP
2. Compute the satisfaction bit string of pp;
3. PPS1=null
4. For each preferred product pp in PP
5. PPS$_1$ = PPS$_1$ U{{pp}};
6. For (l=1;l<n;l++)
7. { MAX=0;
8. For each S in PPS$_1$
9. { compute UB_E(nPPs,C);
10. compute LB_E(nPPs,C);
11. If(LB_E(nPPs,C) > MAX) MAX = LB_E(nPPs,C); }
12. PPS$_1$ = PreferredCheck(MAX, PPS$_1$);
13. PPS$_{l+1}$ = Apriori_gen(PPS$_1$); }
14. MAX=0;
15. For each S in PPSn
16. { compute E(S,C);
17. If (E(S,C) > MAX)
18. { MAX = (S,C);
19. nPP = S; }}
20. Return nPP;

Function PreferredCheck(MAX,PPSl)

1. For each S in PPSl
2. If (MAX > UB_E(nPPs,C))
3. PPSl = PPSl – {S};
4. Return PPSl

Function Apriori_gen(PPSl)

1. PPS$_{l+1}$ = PPSl ⊗ PPSl; //{A ⊗ B|A,B ∈ PPSl, |A ∩ B| = l-1}
2. For each S in PPS$_{l+1}$
3. If ((any subset of S with size l) ∉ PPSl) PPSl = PPSl ~{S};
4. Return PPS$_{l+1}$;

5. OBSERVATIONS

To identify or find the effectiveness, efficiency and memory requirement of the proposed Apriori algorithm. The algorithm was implemented in JAVA language and was run on a PC with processor of Intel core 2 duo, 2GB of main memory and under the Microsoft Windows XP Professional.

A publicly available data generator given by the Borzsonyi was used to generate the data sets on which the algorithm runs. The data generator provides three types of data distribution, i.e., independent, correlated, and anti-correlated distributions.

5.1 Performance of Bitmap Index

To Evaluate the efficiency of the bit map index to calculate the value in the N_vec(P,C), a sorting approach is used and implemented as the main criterion. The execution time to get the value of the N_vec(P,C) from the Bit Map Index and sorting based approach is used in the implementation.

In the sorting-based approach, for each quality describer, data of customers and products are sorted in increasing order according to their corresponding values on the quality describer. Then the value of N_vec(P,C) for a customer C is obtained by counting the number of products whose ranks on all the quality describers are less than those of the customer. After getting the value of N_vec(P,C) for each customer C, N vec(P,C) is also obtained by the sorting-based approach. The execution efficiency of using the BMI index structure is 15.93, 36.99, 57.14, 125.69, and 250.43 times faster than that of performing the sorting-based approach on 20K, 40K, 60K, 80K, and 100K customers, respectively. The execution time taken by the bit map index is shown in the Fig 2.

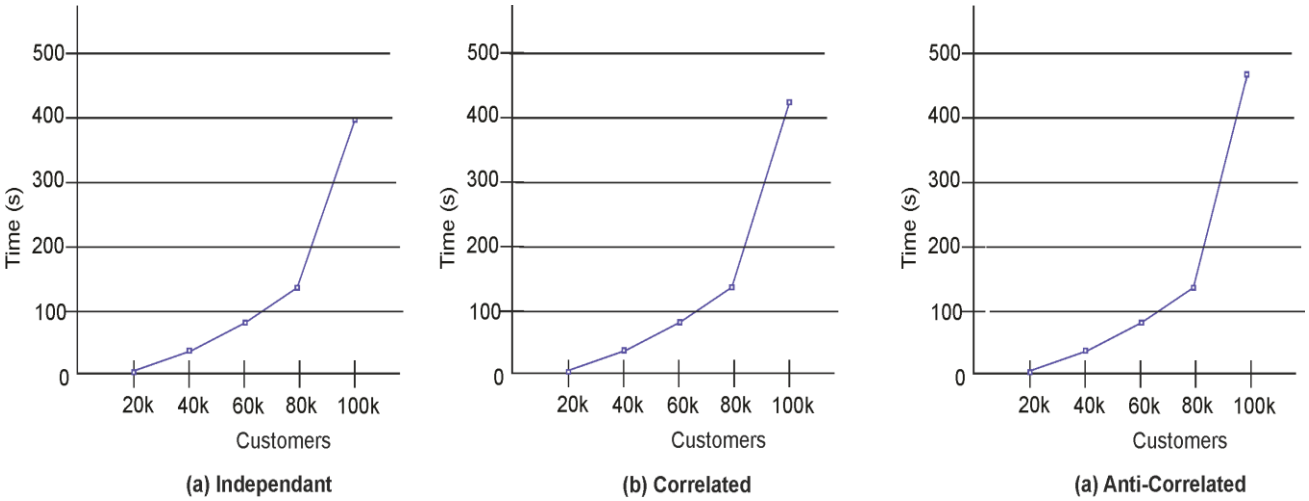

Fig 2: Execution time of Bit map Index and performing the sorting based approach

5.2 Performance of Algorithm

Initially, it has been identified the pruning rate of the Apriori algorithm and that is obtained by adding the number of times to compute the upper and lower bounds of N_vec(P,C) and to compute the expected number of the total customers for the remaining sets of k products. The Table 2 shows the pruning rate for the varying products

Now, execution time is evaluated on varying number of the products and the results are shown in fig 3. The products have been listed from 40 to 80 on the x-axis and time is from 0.1s to 100000s on the y-axis. After this we have evaluated the memory requirement for the varying number of products and the results are shown in fig 4. The products have been listed from 40 to 80 on the x-axis and memory requirement is from 5500KBs to 13500KBs on the y-axis. And at last we have evaluated the execution time on varying number of customer and the results as shown in fig 5. The customers have been listed from 20k to 100k on the x-axis and time is from 0.1s to 100000s on the y-axis.

Table 2: Pruning Rate on Varying Products (Independent Distribution)

Products / Algorithm	40	50	60	70	80
Apriori	87.3947%	90.3821%	92.2317%	93.4878%	95.3778%

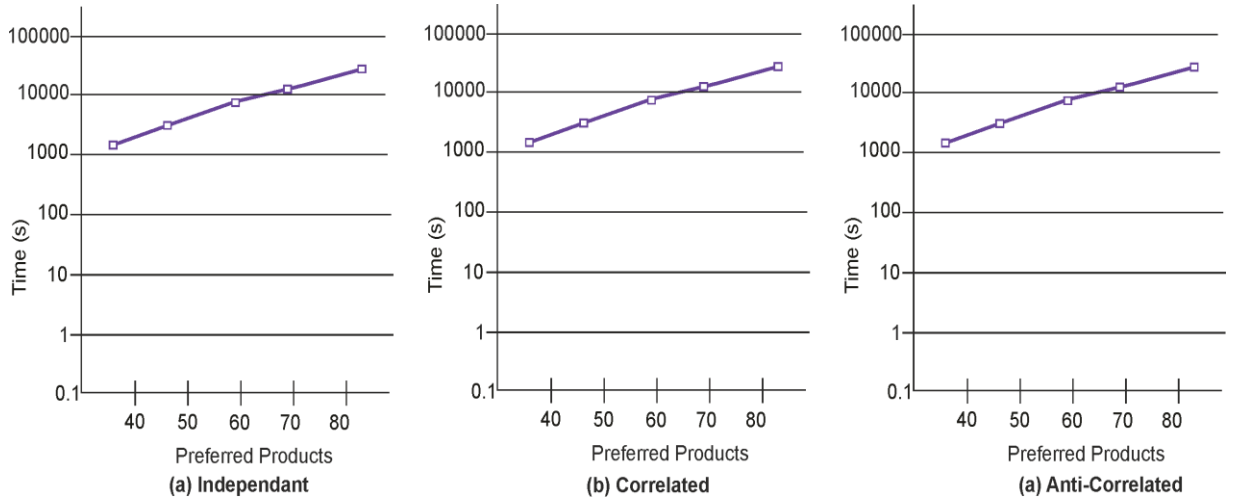

Fig 3: Execution time on a varying Preferred Product

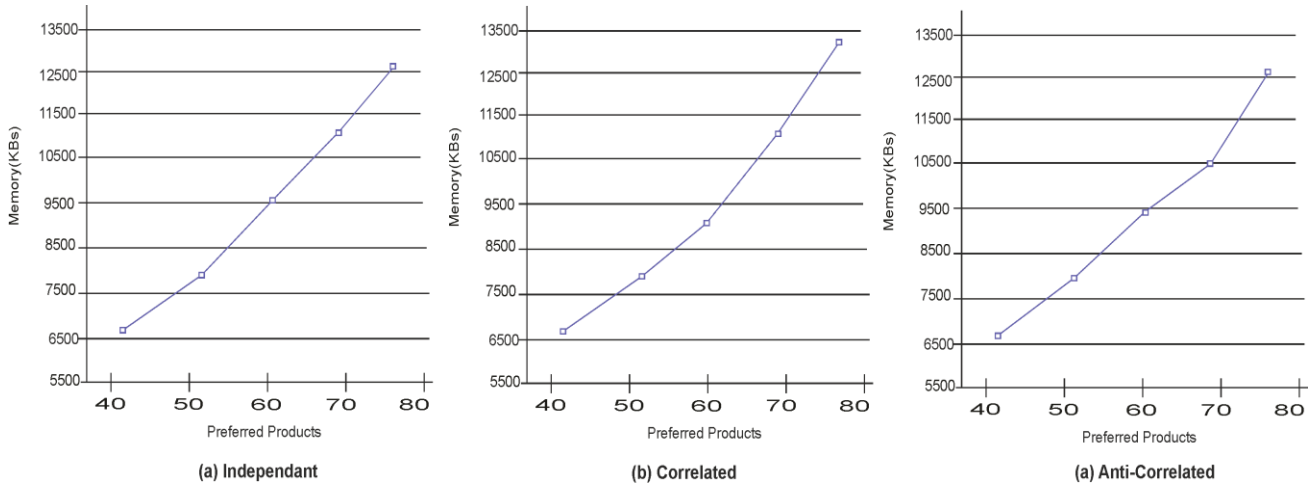

Fig 4: Memory requirement on a varying Preferred Products

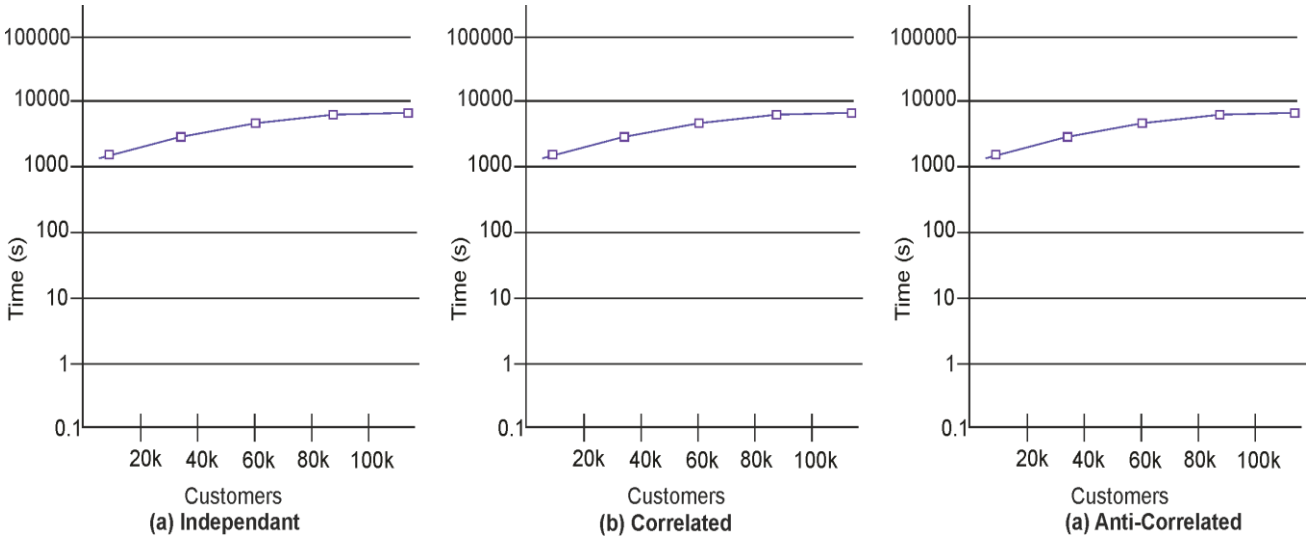

Fig 5: Execution time on a varying Preferred Customers

6. CONCLUSIO

In this paper, it has been formulated that *n*-MDP discovering problem for determining *n* most demanding products with the highest expected number of the total customers. Real Estate Industry in India is facing many complex problems resulting into loss for industry & expensive home for buyers .Thus finding the n-most demanding product gives an optimal solution to the builder as now he has the insight of buyers expectations and would only invest in what the customer actually wants .Customer preferences are taken into consideration and algorithms are implemented giving the valuable information about customer's opinion. The output is the result of intensive calculations over a recorded data and results are shown statistically.

7. ACKNOWLEDGMENTS

We take this opportunity to express our profound gratitude and deep regards to our guide Asst. Prof .Mrs. Swati Patil for her exemplary guidance, monitoring and constant encouragement throughout the course of this thesis. The blessing, help and guidance given by her time to time shall carry us a long way in the journey of life on which we about to embark. We are also thankful to Mr.Sachin Sakhare, HOD IT, VIIT, Pune who encouraged us in this project.

We also take this opportunity to express a deep sense of gratitude to Mr. Atul Dhaygude, CEO and HEAD, Orion Innovations Pvt. Ltd, for his cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are obliged to Dr. Sachin R. Sakhare, for the valuable guidelines provided by him on data mining. We are grateful for their cooperation during the period of our assignment.

8. REFERENCES

[1] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "A Microeconomic View of Data Mining," Data Mining and Knowledge Discovery, vol. 2, no. 4, pp. 311-322, 1998.

[2] International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.6, Nov-Dec. 2012 pp-4657-4663 ISSN: 2249-6645 "Data Mining: Future Trends and Applications" by Annan Naidu Paidi *Asst .Prof of CSE Centurion University, Odisha, India.*

[3] "https://www.cgi.com/sites/default/files/white-papers/business-intelligence-white-paper.pdf

[4] E. Dellis and B. Seeger, "Efficient Computation of Reverse Skyline Queries," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 291-302, 2007.

[5] X. Lian and L. Chen, "Monochromatic and Bichromatic Reverse Skyline Search over Uncertain Databases," Proc. 27th ACM SIGMOD Int'l Conf. Management of Data, pp. 213-226, 2008.

[6] A. Vlachou, C. Doulkeridis, Y. Kotidis, and K. Norvag, "Reverse Top-k Queries," Proc. 26th Int'l Conf. Data Eng.,pp.365-376,201

[7] [7] X. Lian and L. Chen, "Monochromatic and Bichromatic Reverse Skyline Search over Uncertain Databases," Proc. 27th ACMSIGMOD Int'l Conf. Management of Data, pp. 213-226, 2008.

[8] Q. Wan, R.C.-W. Wong, I.F. Ilyas, M.T. Ozsu, and Y. Peng, "Creating Competitive Products," Proc. 35th Int'l Conf. Very Large Data Bases, pp. 898-909, 2009.

[9] Z. Zhang, L.V.S. Lakshmanan, and A.K.H. Tung, "On Domination Game Analysis for Microeconomic Data Mining," ACM Trans.Knowledge Discovery from Data, vol. 2, no. 4, pp. 18-44, 2009.

[10] W.C. Wang, E.T. Wang, and A.L.P. Chen, "Dynamic Skylines Considering Range Queries," Proc. 16th Int'l Conf. Database Systems for Advanced Applications, 2011.