Price Prediction System using Data Anonymity and Top-k Products

Preethi.P
Department of Computer Science and Engineering
Saveetha Engineering College, Chennai, Tamil Nadu, India

Pradeep.R
Department of Computer Science and Engineering
Saveetha Engineering College, Chennai, Tamil Nadu, India

Abstract- Data Mining is an emerging research area in which useful patterns are discovered and extracted from a database. Since there are extensive data, there is an upcoming need for turning such a huge data into useful knowledge. In marketing, creating new product and fixing the price of it, is one of the major challenges. Predicting the price of new product has become an eminent need under the development of market economy and information technology. The enterprise observes price trends, news, reports, and technical specifications of competitor's products for predicting the price of their new product. This paper manifests that how to handle huge product dataset efficiently by incorporating three different concepts for predicting the price of new product. First, the original data set is anonymized using K-Anonymity which preserves the privacy of original data. Second, skyline processing is used to return only the relevant attributes which are needed for price prediction. Skyline computation is done to help users to take care huge available data by detecting some of the interesting data objects. It also minimizes the response time. Third, top-k products are extracted by analyzing dominance relationship among all tuples which enhances the overall performance. Then Price Predictor Algorithm (PPA) is implemented to compare a new product with top-k products that are extracted and not with all products available in a market. The experimental results also shows that the response time for comparing new product with top-k products is minimized with the help of PPA algorithm.

Keywords – K-Anonymity, Skyline Processing, Dominance Relation

I. INTRODUCTION

Data Mining is the practice of mining the useful patterns from huge dataset. Patterns are the collection of items whose occurrences are related to one another. The plenty of data tied with the data analysis tools results in data rich but information poor. Many organizations and companies release their products day by day. Data analyst or decision maker play a major role for fixing the price of new product. They compare all the products available in the market and then they predict the price of new product. Without some powerful tools, the human ability cannot do decision making with the tremendous data collected and deposited in huge and abundant data warehouses. Subsequently, decisions are made based on decision maker's observation and not on rich data stored in data warehouses. This is because the decision maker did not have the tools to mine the useful information entrenched in the massive data.

In order to make better decisions in business, data mining can be exploited to determine patterns and relationships that exist in data. One of the secrets to succeed in business is to fix the price of newly manufactured products correctly because people have concern about determining the real cost of their computers. Pricing the products correctly will find a way to sell many products as possible and creates the foundation for a business that will flourish. Historical prices are important information that can help decision makers to decide whether to fix the price of product that matches the requirements. They provide both a context to the analyst, and facilitate the use of prediction algorithms for forecasting future prices. For example, consider the process of predicting the price of newly introduced laptop. The first step in price prediction is to try to find at least ten laptops similar to new one. Then choose the laptops of with same brand and model number for comparison. HDD, RAM and speed of processor speed will be the significant factors in comparing prices thus checking those specifications when compiling new product's price will result in price prediction. There are some of the challenges in product price prediction while analyzing pricing history.

1. **Data Privacy**: There is no privacy on individual's information. Though the identifiers such as name and phone number are removed, there can exist some quasi-identifiers which also reveals personal information.
2. **Difficulty in handling numerous attributes**: The database consists of several attributes which may not be needed for prediction. Handling those attributes while forecasting leads to performance reduction and improves cost.
3. **High Cost and Less Efficient:** The cost of accessing data is high since the huge data with unimportant attributes are also handled in existing scenarios.

Sometimes data need to be shared for various purposes such as research, business and development of new system. Most sensitive applications such as Banking, Clinical data, etc., generate huge amount of transactions daily. To preserve data sensitivity of shared data, the data required to be sanitized before it can be circulated and examined. Data Anonymization is a common method for cleaning the data [1]. It is the process of swapping the contents of recognizable fields such as names, SSN numbers, IP addresses, and zip codes in a database with the more generalized or suppressed values. Thus, the values cannot be linked with any specific person, project or company. K-anonymity is a data anonymization technique that manipulates the data to achieve privacy goals. Manipulation is done by generalization and suppression. After the data has been cleansed, mining top-k patterns from this huge data set is a difficult task. Finding significant patterns may be very complex and the algorithm for efficiently solving this problem attracted a lot of attention in a data mining community. There are many data objects available in database, of which only few objects are needed for prediction. Skyline computation [11] discovers really important and needed data objects from potentially huge data set. It finds objects needed for the purpose of prediction and ignores irrelevant objects. Dominance Relationship Analysis (DRA) [7] is used to analyze the dominance relationship among tuples and results in top-k patterns. By analyzing dominance relationships, companies can position their products more effectively while remaining profitable. DRA and skyline processing are much important in decision making applications. Both are applied to our prediction algorithm to accurately predict the anonymous data. Thus, there is no compromise in security and accuracy of original dataset. To summary, the main contributions in this paper include:

- An effective Price prediction System (PPS) is introduced to predict the price for new product by analyzing the existing products
- Data anonymization technique is used to conceal the information which protects the privacy of data.
- Top-k Product Miner is introduced to reduce the data size by analyzing and returning skyline points and extracts top-k products through DRA.
- A new algorithm called Price Predictor Algorithm (PPA) is proposed which compares the new product with only top-k products and predicts the right price for new product. Thus it minimizes the response time and improves the overall performance of prediction system

The rest of this paper is organized as follows: Section 2 provides the literature in data anonymization, top-k pattern mining and price prediction. In section 3, architecture of Price Prediction System is discussed. Section 4 discusses the evaluation of proposed model and section 5 manifests the experimental results. Section 6 draws conclusion.

II. PROPOSED ALGORITHM

This section briefly delivers literature on price prediction, data anonymization, top-k pattern mining, skyline processing and Dominance relationship analysis.

A. Price Prediction

Data Mining can be used for performing analysis in various fields such as Finance, Business, Marketing etc., In Business, Finding the correct price for the products helps the business people to maximize their profits. Pricing Product [1] can be done by researching competitors' pricing and researching the demand for product. Various algorithms have been proposed for price prediction in various fields such as stock market, real estate, gold, crude oil and agricultural products. Stock market prediction [16] is regarded as a challenging task new and it is based on Logistic Regression (LR). Based on current month, stock price trend for next month is predicted. A new prediction model, GNP (Generic Network Programming) [17] is applied for searching for an optimal combination of two or more appropriate stock price indices, which is different from a conventional GA (Genetic Algorithm) or GP (Genetic Programming). In real estate [18], there is an insistent demand to establish an easy-operate and logical scientific prediction model. A real estate price prediction methodology based on BP (Back Propagation) neural network and Elman neural network is introduced [24], and approved that these two methodologies have a good accuracy. The combination of models such as Grey -Markov prediction model is established to estimate the urban

land price [26] of Tangshan city which results in higher accuracy. Since there are multiple properties, the foundation of the gold [19] price prediction is very complex. To reduce the randomness for upcoming price prediction, gray prediction method is used to establish a procedure for Chinese gold futures price data sample. A generalized Intelligent-agent-based fuzzy group forecasting model is proposed for oil price [20] prediction in which some single Intelligent-agent based predictors are fuzzified into some fuzzy prediction representations. At last, the aggregated fuzzy prediction is defuzzified into a brittle value as the final prediction results. A prototype decision support system of an agricultural product [21] market is designed and developed. It can extract online price information of a certain agricultural product from websites of agricultural wholesales, predict the product price in the future months, and provide further decision support on such issues as which cities the product should be sent to for sale and which cities should be in the transport route.

B. Data Anonymization

To protect the privacy of individuals, the outsourced data are anonymized before publishing. The most popular technique is K-Anonymity [5] was introduced, which guarantees that each person contained in the release cannot be distinguished from at least k-1 individuals. In K-Anonymity, the basic operations are generalization and suppression. Generalization can be conducted on the attribute level and cell level, while suppression can be applied to the tuple level and cell level. Besides the privacy concern, the K-Anonymity approach tries to reduce the information loss as much as possible. Several studies shows that [9] [10], 1) k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute. 2) k-Anonymity did not protect against attacks based on background knowledge. So, K-Anonymity is not enough for guaranteeing the privacy and hence, *l*-diversity [8] was proposed to address the problem. In *l*-diversity, each equivalence class has at least *l* well-represented sensitive values. However, *l*-diversity may be difficult and unnecessary to achieve. *l*-diversity is insufficient to prevent attribute disclosure and it does not consider the overall distribution of sensitive values. To overcome these drawbacks, new privacy technique called t-closeness [22] was proposed, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table.

C. Mining Top-K Patterns

Data mining promises to obtain valid and potentially useful patterns from a data base. Mining high utility items from the data base is an emerging topic in data mining [4], which refers to discovery of item sets with utilities higher than a user- specified minimum utility threshold. Actionability [12] addresses this problem in that, pattern is deemed actionable if the user can act upon it. Beyond top-k, bottom –k [13] correlative patterns are used to detect crime activity. Also, existing works on mining top-k patterns on the data streams are mostly for non-sequential patterns. Top-k sequential pattern mining allows the repetition of the same item and it emphasizes the order of sequence. A number of social Web sites provide tagging functionalities and also offer folksonomies within or across the sites. So, novel approach for mining and representing user interests with tagging practice [14] was proposed for clustering user-centric interests by analyzing tagging practices of individual users.

D. Skyline Processing

Currently, the storage and management of data are widely distributed. In order to help users to handle huge data efficiently, advanced query operators such as skyline queries are introduced to find set of interesting data objects. To minimize the response time, Skyline processing is introduced with several algorithms [11]. Skyline computing can be used in various applications such as search pruning, decision making and personalized services. To handle incomplete data [15], two algorithms are proposed namely, “Replacement” and “Bucket” that use traditional skyline algorithms. Skyline query processing in highly distributed environments [6] poses inherent challenges and demands and require non-traditional techniques due to the distribution of content and the lack of global knowledge. So, detail review done in existing approaches that are applicable for highly distributed environments, clarify the assumptions of each approach and provide a comparative performance analysis. Global skyline [27] is also an important variant of skyline that has been widely applied in multiple criteria such as decision making, business planning and data mining. So, Subspace Global Skyline (SGS) query is proposed for global skyline in ad hoc subspace. Recently, Skyline processing is applied on Big Data [6] for efficient computation.

E. Dominance Relationship Analysis

The concept of dominance has recently attracted much interest in the context of skyline computation. Dominance [7] can be used in business analysis from a microeconomic perspective. DRA aims to provide intuition to the dominant relationships between products and customers. Dominance analysis [2] can be useful when determining predictor importance if the independent predictor variables are correlated. Creating a new product which dominates all its competitors is one of the main objectives in marketing. Given a budget, the task [28] is to decide the best possible features of the new product that maximize its profitability. In general, a product is marketable if it dominates a large set of existing products, while it is not dominated by many.

Thus, the existing work does not provide privacy for individual information and fails to enlist the skyline computing paradigm in finding out the dominant parties for analysis purpose. Handling huge available data is tedious process in existing work. Hence to overcome the difficulty of handling huge data and to protect the privacy of individual data, the proposed work incorporates two different concepts for predicting the price of new product. Data anonymization provides privacy on individual's information. Techniques like Dominance analysis and Skyline Computation can be used to mine top-k products which reduce the response time and improves prediction accuracy.

Architecture of price prediction system

The goal of the project is to predict the price of new product by comparing all the products available in the market. Setting the right price for products may help to maximize profits and tend to maintain a noble relationship with customers. Pricing product effectively can avoid severe financial problems that may occur if prices are too high or low - if price is fixed too much company may price themselves out of the market, but if price is too little, the organization may be underpaid for their work. The price of product [1] is fixed based on some of the criteria such as researching competitors' pricing and researching the demand for corresponding product. The proposed work focuses on how to handle the huge dataset [Customer and product details] efficiently for price prediction of new product. It incorporate two different concepts such as Data Anonymity, Skyline Computing and Dominance Analysis

Fig. 3.1 shows the architecture of Price Prediction system. First of all the dataset is converted anonymized data. Then this dataset is subjected into top-k product miner to extract top-k products. In top-k miner, two operations are performed such as skyline computing and dominance relationship analysis. Skyline processing removes unwanted attributes and dominance analyzer analyses the relationship among all tuples and extracted top-k products. These products are stored in a database and decision maker will give features of new product. The price predictor implements Price Predictor Algorithm (PPA) to compare new product with top-k products and finally it returns the price of new products to the decision maker. Price Prediction System consists of several components such as data anonymizer, top-k product miner and price predictor which will be briefly discussed below.

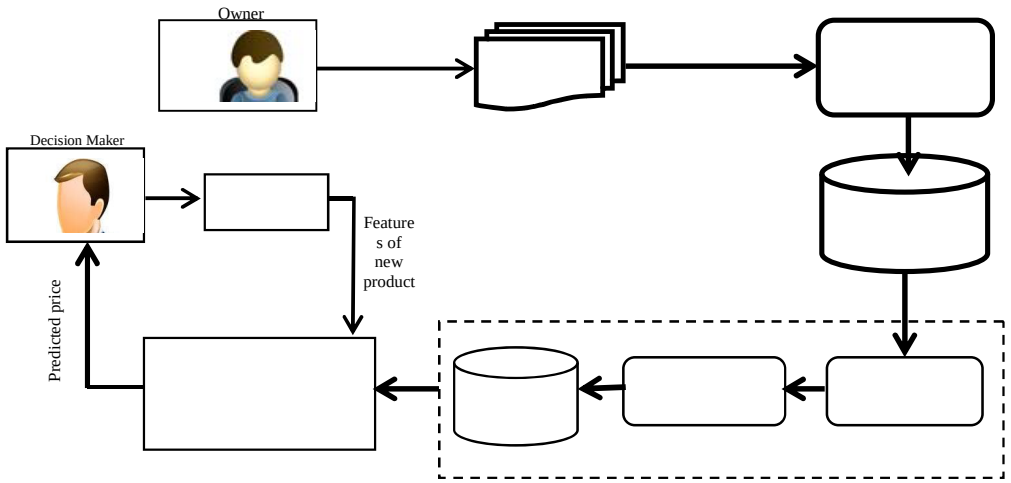

Fig 3.1 Architecture of Price Prediction System (PPS)

A. Data Anonymizer

Sometimes data must be shared for various purposes such as business and research. When publishing individuals' information, it is necessary to hide some personal data which involves legal implications. Thus the data anonymization came into existence. In order to protect sensitivity shared data, it should be anonymized before it can be distributed and analyzed. So, the owner of the data set takes original data set and converts it to anonymized data and subjects this data to data analyst

For example, table 3.2 shows the original employee table. If an adversary wants to find an individual or if he wants to misuse the information, he can directly use this table. Because within particular area(zipcode), only few male or female within the particular age group will be there and using phone no one can easily trace the person whom he wish to identify (i.e.) within area of zipcode 277516, a male with age of 26 can be easily identified and he can be misused by any adversary. The set of attributes such as age, gender, zipcode are called quasi-identifiers which can be matched with public dataset such as voter's data set or census data set for identifying an individual. Thus to provide privacy over individual data, data anonymization is used. Table 3.2 shows the anonymized data set. The replacement of values of the attribute with'*'is called suppression and the replacement with generalized values for the value of an attribute is called generalization. Consider the first record in a table. The age attribute value 20 is generalized as 20-30. The gender attribute value is suppressed using *. Similarly the attributes such as zipcode and mobile no is also suppressed to enhance privacy.

Initially, the owner of original data set converts it to anonymized form and subjects it to data analyst. The attribute Age, Mobile no, Country can be matched with some public data set such as voter's data set or senses data set to identify an individual. Thus, these attributes are anonymized to protect the privacy of individuals. This can be achieved through operations such as generalization and suppression. Then this anonymized data is subjected to data analyst or decision maker for predicting the price of new product.

Age	**Gender**	**Zipcode**	**Mobile No**
26	Male	277516	8466866676
35	Female	283311	9703662662
24	Female	215646	9966503503
37	Female	234721	9985687687

Table 3.2 An Employee table

Age	**Gender**	**Zipcode**	**Mobile No**
20-30	******	277***	84668*****
30-40	******	283***	97036*****
20-30	******	215***	99665*****
30-40	******	234***	99856*****

Table 3.2 An anonymized Employee table

B. Top-k Product Miner

The second step is to extract top-k patterns (top-k products) from huge data set. Top-k product miner is used to extract best or dominant products. There are huge patterns available which are less significant and sometimes irrelevant. Identifying and ignoring those irrelevant attributes are called skyline computation. When applying skyline computing, attribute size can be reduced and only relevant attributes that are needed for prediction will exist. When predicting the price of new product, comparing the new product with all the existing products available in the market will be time consuming process. In order to overcome this issue, dominant or best products are chosen from available products. It can be done by analyzing the dominance relationship exists among tuples. Once the top-k products are extracted, the price of new product is predicted by only comparing new product with top-k products. Thus it quickens the process and improves the performance by reducing the response time. As the manufacture's point of view, best or dominant product is always released as a new product. So selecting the top-k products and comparing them with new product does not result in inaccuracy.

C. Price Predictor

The last step is prediction of price of new product which can be implemented using Price Prediction Algorithm. When the Manufacturer introduces new product, it is necessary to fix the right price for it. From the top-k products, the price of new product is found by comparing new product with top-k products not with all products existing in the market. Top-k products are extracted such that Manufacturer has a thought of creating a product which should dominate all their competitors' product. After predicting the price of the product, the rating of product such as preferable or profitable result can be returned to the user. If a new product has to be introduced, the first thing is definitely the cost has to be predicted and this is done by judging all the dominant product's values in the market. Secondly, in the consumer perspective it should not be seen only as preferable. It should also be noted whether it is profitable. Some products will be preferred by some set of people who are affordable since the cost will be too high. For some set of people, it will not be worthful since they can't afford. Keeping this into consideration, rating of the product should be returned to user.

III. EXPERIMENT AND RESULT

The company monitors price trends, news, rumors, and technical specifications of competitor's products for predicting the price of their new product. The work focuses on how to handle the dataset efficiently for predicting the price of new product. To evaluate the proposed methodology, product data set is used. The data set consists of Customer details together with product (Laptop) details or technical specifications that have been bought by the customers. Note that the dataset have unique product details and it consists of 150 instances and 22 attributes. The sample dataset is shown in table 4.1. The attributes are Cust_Id, Age, Occupation, City, Zip code, Mobile No, Email Id, Prod_Brand, Model, Series, HDD, Display_Size, RAM, Processor, OS, Weight, Speed, Graphics Memory, Battery_Backup, USB_Ports, Color and Price.

Here the first step is to anonymize the customer details given in a dataset. In data mining applications like product marketing, people are not concerned on individual's data. Individual's information must be protected on behalf of privacy. An attacker can easily link individuals by comparing the group of attributes specified in the data set such as {age, gender, zipcode} with the public dataset available in the internet. So the individual information such as customer details are anonymized which is shown in Table 4.2. Basically two operations are performed: The first operation is Generalization. e.g. Age attribute. The value of age attribute in first tuple is 39 which can be generalized into 30-40. Similarly generalization is applied for all tuples in a table. The second operation is Suppression. E.g. zipcode attribute. The value of zipcode attribute in a first tuple is 277516 which can be suppressed into 277***. Similarly the value of Mobile no in a first tuple is 8680919478 which can be suppressed into 86809*****. Thus disclosure of individual's data is achieved and privacy of customer information is protected by anonymization process. All values in a gender attributes are suppressed to '*' values.

Cust_ID	Age	Gender	Occupation	Country	ZipCode	Mobile_No	Email	Prod_Bran	Model	Series	HDD	Display	RAM	Processor	OS	Weight	Speed	Graphics	Battery	USB	Color	Price
12101	39	Female	Adm-clerical	United-States	277516	8680919478	musicgal_16@hotmail.com	Acer	Aspire	C720	320	11.6	2	Intel-I5	Chrome-OS	1.25	1.4	1	8.5	2	Black	20,659
12102	50	Female	Exec-managerial	United-States	283311	9444678903	mitchell_1995@gmail.com	Acer	Aspire	7736G	320	17.3	4	Intel-I5	Win-7	2.5	1.1	1	3.5	4	Black	36490
12103	38	Male	Handlers-cleaners	United-States	215646	9840284945	midoban@hotmail.com	Acer	Aspire	P3171	500	11.6	2	Intel-I3	win-8	0.79	1.4	1	6	4	White	57590
12104	53	Male	Handlers-cleaners	United-States	234721	9444789656	lonely12@hotmail.com	Acer	Travelmate	8481T	320	14	6	Intel-I5	Win-7	2.1	1.6	2	5.4	3	Black	77896
12105	28	Male	Prof-specialty	Cuba	338409	9368344575	lovingdunli@hotmail.com	Acer	Aspire	8951G	1536	18.4	8	Intel-I7	Win-7	3.8	2	2	5.5	4	Black	34580
12106	37	Male	Exec-managerial	United-States	284582	9791517460	aeon-t-flair@hotmail.com	Acer	Aspire	5750G	640	15.6	4	Intel-I7	Win-7	2.6	2	1	4.5	3	Blue	36227
12107	49	Male	Other-service	Jamaica	160187	8883771971	affid_1995@hotmail.com	Acer	Aspire	S3391	500	13.3	4	Intel-I7	Win-7	1.36	1.9	1	5.5	2	Black	50000
12108	52	Female	Exec-managerial	United-States	209642	9786523451	ahomeart@hotmail.com	Acer	Aspire	4810TZ	500	14	4	Intel-pentium	Win-7	1.9	1.3	1	8	3	Black	36945
12109	31	Female	Prof-specialty	United-States	245781	9984672341	kinwei1021@hotmail.com	Acer	Aspire	V5471P	500	14.1	4	Intel-I5	Win-8	2.1	2.6	1	5	3	Misty	46100
12110	42	Male	Exec-managerial	United-States	159449	9840482345	alex-cheah@hotmail.com	Acer	Aspire	V5122P	500	11.6	6	AMD-A	Win-8	1.38	1	1	3.5	2	Silver	32267
12111	37	Male	Exec-managerial	United-States	280464	9436857843	alwaystruetou@gmail.com	Acer	Aspire	E1572	500	15.6	4	Intel-I5	Win-8	2.1	1.6	1	5.3	3	Black	23270
12112	30	Female	Prof-specialty	India	141297	9436857841	angel263_5566@gmail.com	Acer	Aspire	E1531	500	15.6	4	Intel-pentium	Linux	2.45	2.2	1	4.5	4	Silver	23270
12113	23	Female	Adm-clerical	United-States	122272	9436857824	aowjiarong@hotmail.com	Acer	Aspire	R7571	500	15.6	8	Intel-I5	Win-8	2.4	1.8	2	6.7	3	Silver	89894
12114	32	Male	Sales	United-States	205019	9436857820	apple_girl0123@gmail.com	Acer	Gateway	NE56R	500	15.6	2	Intel-pentium	Linux	2.6	2.26	2	5	4	Black	21425
12115	40	Male	Craft-repair	Cuba	121772	8466882882	aqua_pig@hotmail.com	Acer	Aspire	5745C	500	15.6	2	Intel-I3	Linux	2.6	2.53	1	3	4	Black	32000
12116	34	Male	Transport-moving	Mexico	245487	8466855855	ash_bio@hotmail.com	Apple	MacBook	Pro 15	253	15.4	8	Intel-I7	MacOs	2.02	2	2	8	2	Black	160051
12117	25	Male	Farming-fishing	United-States	176756	8466866676	samy_ang@hotmail.com	Apple	MacBook	Air2013	256	11.6	4	Intel-I5	MacOs	1.08	1.3	1	9	2	White	83975
12118	32	Female	Machine-op-inspct	United-States	186824	9703662662	bab_sher11@hotmail.com	Apple	MacBook	Pro 2011	750	17	4	Intel-I7	MacOs	2.99	2.2	1	7	3	Silver	146616
12119	38	Male	Sales	United-States	228887	9966503503	baka-tevin@hotmail.com	Asus	Transformer	TX300	500	13.3	4	Intel-I7	Win-8	1.3	1.9	1	5	2	Black	41376
12120	43	Male	Exec-managerial	United-States	292175	9985687687	bellcrossljz@hotmail.com	Asus	Transformer	V550CA	1024	15.6	6	Intel-I7	Win-8	2.65	2	2	3.1	3	BlackAlu	41486
12121	40	Female	Prof-specialty	United-States	193524	9000055287	burned_pig@hotmail.com	Asus	Transformer	G750JX	1024	17.3	8	Intel-I7	Win-8	4.8	2.4	2	4	4	Black	145400
12122	54	Female	Other-service	United-States	302146	9000055692	californiasnoo@gmail.com	Asus	Zenbook	U500	256	15.6	8	Intel-I7	Win-8	2.2	2.5	2	5	3	Aluminu	121929
12123	35	Female	Farming-fishing	United-States	276845	9000055704	charles_ex@gmail.com	Asus	VivoBook	S400C	500	14	4	Intel-I3	Win-8	1.8	1.8	1	3.5	3	BlackAlu	62497
12124	43	Female	Transport-moving	United-States	117037	9000055126	cas1303@hotmail.com	Asus	VivoBook	S56CA	750	15.6	4	Intel-I5	Win-8	2.39	1.7	1	4	3	Silver	51858
12125	59	Male	Tech-support	United-States	109015	9985687687	cutez_haha@hotmail.com	Asus	Taichi	21A	256	11.6	4	Intel-I7	Win-8	1.25	1.9	1	5	3	Black	155577
12126	56	Male	Tech-support	United-States	216851	9985645645	daryl_lim_jz@hotmail.com	Asus	VivoBook	S200	500	11.6	4	Intel-I3	Win-8	1.38	1.4	1	3	3	Grey	36298
12127	19	Male	Craft-repair	United-States	168294	9505907010	elene5566@hotmail.com	Asus	Zenbook	UX31A	256	13.3	4	Intel-I7	Win-7	1.3	1.9	1	3	2	Silver	150390
12128	54	Male	Other-service	South	180211	9948808010	faith_giritz95@hotmail.com	Asus	Zenbook	UX32A	500	13.3	4	Intel-I3	Win-7	1.5	1.4	1	3.5	3	Silver	88144
12129	39	Male	Exec-managerial	United-States	367260	9948811711	freakopig@hotmail.com	Asus	Zenbook	N73SV	640	17.3	6	Intel-I5	Win-7	3.4	2.3	1	3.1	3	Black	22294
12130	49	Female	Craft-repair	United-States	193366	9948999312	ginpnix_95@hotmail.com	Asus	Zenbook	N55SF	640	15.6	6	Intel-I7	Win-7	2.7	2.2	1	2.1	4	Black	15556
12131	23	Female	Protective-serv	United-States	190709	9436857794	huiqi999@hotmail.com	Asus	Zenbook	G74SX	1536	17.3	8	Intel-I7	Win-7	4.28	2	2	4.5	4	Black	75000
12132	20	Male	Sales	United-States	266015	9436857787	isabel_95@gmail.com	Asus	Automobili	VX7	1536	15.6	8	Intel-I7	Win-7	3.82	2	2	5.9	4	Black	35555
12133	45	Male	Exec-managerial	United-States	386940	9436857764	jis178@hotmail.com	Asus	EeePC	1018P	250	10.1	1	Atom	Win-7	1.1	1.5	0.654	9	3	Black	11032
12134	30	Female	Adm-clerical	United-States	259951	9436857732	jiawen1995@hotmail.com	Dell	Inspiron	15R5521	500	15.6	4	Intel-I3	Win-8	2.2	1.8	2	4	4	Silver	44990
12135	22	Female	Other-service	United-States	311512	9436857704	jiawen95@hotmail.com	Dell	Latitude	E6430S	250	14.1	2	Intel Dual Cor	Win-7	2.1	2.53	1	9	4	Black	94139
12136	48	Male	Machine-op-inspct	Puerto-Rico	242406	9000055457	joshuangye@hotmail.com	Dell	Latitude	XPS12	750	12.5	4	Intel-I5	Win-8	1.52	1.7	1	3	1	Black	134095
12137	21	Male	Machine-op-inspct	United-States	197200	9000066085	joycexinping@hotmail.com	Dell	Adamo	XPS	256	13.4	4	Intel Dual Cor	Win-7	1.46	1.4	1	3	2	Silver	15475
12138	19	Male	Adm-clerical	United-States	544091	9000066924	glendakoh@hotmail.com	Dell	Inspiron	Duo	320	10.1	2	Intel Dual Cor	Win-7	1.54	1.5	1	4	2	Black	39053
12139	31	Male	Sales	Cuba	284154	9000066251	laurw1311@hotmail.com	Dell	Inspiron	Mini1018	250	10.1	1	Intel Single co	Win-7	1.33	1.66	1	7	2	Black	20852
12140	48	Female	Prof-specialty	United-States	265477	9000066452	pulnicol@hotmail.com	Dell	Inspiron	1564S	250	15.6	2	Intel-I3	Win-7	2.73	2.3	1	3.4	3	Black	23689
12141	31	Male	Machine-op-inspct	United-States	507875	9948999342	rs_pigsy@hotmail.com	Dell	Latitude	E6410ATG	320	14.1	4	Intel-I7	Win-7	1.93	1.2	1	4	4	Black	23451
12142	53	Male	Prof-specialty	United-States	288506	9948999274	tanhanzhou@gmail.com	Dell	Studio	1535S	250	15.4	4	Intel Dual Cor	Win-7	2.8	2.1	1	6	4	JetBlack	44880
12143	24	Female	Tech-support	United-States	172987	9966644766	timothy4195@gmail.com	Dell	Inspiron	3537S	750	15.6	6	Intel-I5	Win-8	2.2	1.6	2	3	4	Black	42890
12144	49	Female	Adm-clerical	United-States	294638	9912222898	venetia_kok@gmail.com	Dell	Vostro	1700S	256	17.1	4	Intel Dual Cor	Vista	3.5	1.8	1	4	3	Black	56349
12145	25	Female	Handlers-cleaners	United-States	289980	9912555664	weiyang.song@gmail.com	Dell	Vostro	3555S	500	15.6	4	AMD-E2	Win-7	2.47	2.4	1	4	3	Black	121000
12146	57	Female	Prof-specialty	United-States	337895	9436857685	jasminefih@hotmail.com	Dell	Latitude	E5430	320	14	8	Intel-I5	win-7	2.04	2.5	2	4	3	Black	40640
12147	53	Male	Machine-op-inspct	United-States	144361	9436857682	yendi@singnet.com.sg	Dell	Vostro	2520S	750	15.6	4	Intel-I5	Win-7	2.36	3.2	1	4.5	3	Grey	34581
12148	44	Male	Exec-managerial	United-States	128354	9436857675	cas1303@hotmail.com	Dell	Inspiron	N5110	500	15.6	4	Intel-I3	Win-7	2.49	2.2	1	4	4	Black	49594
12149	41	Male	Craft-repair	United-States	101603	9436857668	faith_giritz95@gmail.com	Dell	Studio	1558S	320	15.6	4	Intel-I5	Win-7	2.6	2.53	1	3.3	3	Black	77923
12150	29	Male	Prof-specialty	United-States	271466	9436857664	dinuochan@hotmail.com	Dell	Latitude	D630	250	14.1	4	Intel Dual Cor	Win-XP	2.3	1.8	1	5	3	Black	39318

The problem with the existing system is handling data is difficult and high cost and performance issues. For simplicity, 150 instances are taken. If there are more than 15000 instances, handling those data will become a difficult and challenging task. To handle those dataset efficiently, top-k products are extracted. The dataset may consist of numerous attributes which may not be needed for prediction. Neglecting those attributes and processing dataset will enhance the overall performance and reduces the response time. For larger datasets, it is impossible to get results quickly and scanning such a huge data is expensive and it may take several hours. So

people become annoyed to wait for results upon several hours. Skyline operation results in a set of interesting patterns from multi-dimensional data items by ignoring items that are not dominated by others. This operation involves low overhead to diminish input/ output cost efficiently. In our dataset, only few attribute actually needed for prediction.

Cust_ID	Age	Gender	Occupation	Country	ZipCode	Mobile_No	Email
12101	30-40	******	Adm-clerical	US	277***	86809*****	musicgal_16@hotmail.com
12102	40-50	******	Exec-managerial	US	283***	94446*****	mitchell_1995@gmail.com
12103	30-40	******	Handlers-cleaners	US	215***	98402*****	midoban@hotmail.com
12104	50-60	******	Handlers-cleaners	US	234***	94447*****	lonely12@hotmail.com
12105	20-30	******	Prof-specialty	Cuba	338***	93683*****	lovingdunii@hotmail.com
12106	30-40	******	Exec-managerial	US	284***	97915*****	aeon-t-flair@hotmail.com
12107	40-50	******	Other-service	Jamaica	160***	88837*****	affid_1995@hotmail.com
12108	50-60	******	Exec-managerial	US	209***	97865*****	ahomeart@hotmail.com
12109	30-40	******	Prof-specialty	US	245***	99846*****	kinwei1021@hotmail.com
12110	40-50	******	Exec-managerial	US	159***	98404*****	alex-cheah@hotmail.com
12111	30-40	******	Exec-managerial	US	280***	94368*****	alwaystruetou@gmail.com
12112	20-30	******	Prof-specialty	India	141***	94368*****	angel263_5566@gmail.com
12113	20-30	******	Adm-clerical	US	122***	94368*****	aowjiarong@hotmail.com
12114	30-40	******	Sales	US	205***	94368*****	apple_girl0123@gmail.com
12115	30-40	******	Craft-repair	Cuba	121***	84668*****	aqua_pig@hotmail.com
12116	30-40	******	Transport-moving	Mexico	245***	84668*****	ash_bio@hotmail.com
12117	20-30	******	Farming-fishing	US	176***	84668*****	samy_ang@hotmail.com
12118	30-40	******	Machine-op-inspct	US	186***	97036*****	bab_sher11@hotmail.com
12119	30-40	******	Sales	US	228***	99665*****	baka-tevin@hotmail.com
12120	40-50	******	Exec-managerial	US	292***	99856*****	bellcrossljz@hotmail.com
12121	30-40	******	Prof-specialty	US	193***	90000*****	burned_pig@hotmail.com
12122	50-60	******	Other-service	US	302***	90000*****	californiasnoo@gmail.com
12123	30-40	******	Farming-fishing	US	276***	90000*****	charles_ex@gmail.com
12124	40-50	******	Transport-moving	US	117***	90000*****	cas1303@hotmail.com
12125	50-60	******	Tech-support	US	109***	99856*****	cutez_haha@hotmail.com

Table 4.2: Anonymized dataset

These attributes are called Skyline Points. The skyline points in our dataset are as follows: Prod_Brand, HDD, Display_Size, RAM, Processor, OS, Weight, Speed, Graphics Memory, Battery_Backup and Price. The skyline processed data is shown in table 4.3. Next step is mining Top-k products. Dominant tuple calculation is used to extract top-k products. Developing a new product which dominates all its competitor is one of the main objectives of Marketing. In product dataset, the best feature of laptop is considered for mining dominant products. For example, the features such as Hard disk size should be greater than 500 GB, RAM should be greater than or equal to 4 GB, Processor should be Intel i5 or above, OS can be Windows 7 or above, Graphics Memory is greater than or equal to 1 GB, Speed should be greater than 1.50 GHz, Weight should be within 1.5 to 3 kg, Battery time must be greater than 4 hours and Price should ranges from 35000 to 50000. These are some important requirements for having best laptop as it comes under preferable and profitable.

Prod_Brand	HDD	Display	RAM	Processor	OS	Weight	Speed	Graphics	Battery	USB	Price
Acer	320	17.3	4	Intel-i5	Win-7	2.5	2.2	1	3.5	4	36490
Acer	500	11.6	2	Intel-i3	win-8	0.79	1.4	1	6	4	57590
Acer	320	14	6	Intel-i5	Win-7	2.1	1.6	2	5.4	3	77896
Acer	1536	18.4	8	Intel-i7	Win-7	3.8	2	2	5.5	4	34580
Acer	640	15.6	4	Intel-i7	Win-7	2.6	2	1	4.5	3	36227
Acer	500	13.3	4	Intel-i7	Win-7	1.36	1.9	1	5.5	2	50000
Acer	500	14	4	Intel-pentium	Win-7	1.9	1.3	1	8	3	36945
Acer	500	14.1	4	Intel-i5	Win-8	2.1	2.6	1	5	3	46100
Acer	500	11.6	6	AMD-A	Win-8	1.38	1	1	3.5	2	32267
Acer	500	15.6	4	Intel-i5	Win-8	2.1	1.6	1	5.3	3	23270
Acer	500	15.6	4	Intel-pentium	Linux	2.45	2.2	1	4.5	4	23270
Acer	500	15.6	8	Intel-i5	Win-8	2.4	1.8	2	6.7	3	89894
Acer	500	15.6	2	Intel-pentium	Linux	2.6	2.26	2	5	4	21425
Acer	500	15.6	2	Intel-i3	Linux	2.6	2.53	1	3	4	32000
Apple	253	15.4	8	Intel-i7	MacOs	2.02	2	2	8	2	160051
Apple	256	11.6	4	Intel-i5	MacOs	1.08	1.3	1	9	2	83975
Apple	750	17	4	Intel-i7	MacOs	2.99	2.2	1	7	3	146616
Asus	500	13.3	4	Intel-i7	Win-8	1.3	1.9	1	5	2	41376
Asus	1024	15.6	6	Intel-i7	Win-8	2.65	2	2	3.1	3	41486
Asus	1024	17.3	8	Intel-i7	Win-8	4.8	2.4	2	4	4	145400
Asus	256	15.6	8	Intel-i7	Win-8	2.2	2.5	2	5	3	121929
Asus	500	14	4	Intel-i3	Win-8	1.8	1.8	1	3.5	3	62497
Asus	750	15.6	4	Intel-i5	Win-8	2.39	1.7	1	4	3	51858
Asus	256	11.6	4	Intel-i7	Win-8	1.25	1.9	1	5	3	155577

Table 4.3: Dataset after Skyline Computing

So the tuples that satisfies these requirements are extracted as Top-k products. Since the concept of Dominance is very much useful in customer perspective for selecting the products they like. Then the new Product details such as Brand, HDD, RAM, Processor, OS, Weight, Speed, and Battery Backup are obtained. These data is compared with only the dominant products that are extracted form huge dataset. Thus it improves the performance and it saves time by only comparing with few dominant products and not with all products available in the Market. Top-14 tuples that are extracted from 150 tuples is shown in a following table.

Prod_Brand	HDD	RAM	Processor	OS	Weight	Speed	Graphics	Battery	Price
Acer	500	4	Intel-i5	Win-8	2.1	2.6	1	5	46100
Dell	500	4	Intel-i3	Win-6	2.2	1.8	2	4	44990
HP	1024	4	Intel-i5	Win-8	2.4	2.6	1	3	46490
Lenovo	500	8	Intel-i7	Win-8	1.78	2.1	2	5.9	59464
Lenovo	1024	8	Intel-i7	Win-8	2	2	2	8	57693
Lenovo	1024	8	Intel-i7	Win-8	2.2	2.2	2	3	46515
Samsung	1024	6	Intel-i5	Win-7	2.5	3.1	1	3	53990
Samsung	1024	6	Intel-i5	Win-8	2.5	3.1	1	3	55500
Samsung	750	4	Intel-i5	Win-8	2.33	2.6	1	6	40600
Samsung	1024	6	Intel-i5	Win-8	1.79	2.6	1	6.5	57800
Samsung	1024	4	Intel-i5	Win-8	2.3	3.1	1	6	41600
Samsung	1024	6	Intel-i5	Win-8	2.5	2.6	2	5	59580
Sony	500	4	Intel-i5	Win-8	2.39	3.3	1	4.5	46490
Toshiba	500	4	Intel-i5	Win-7	1.99	2.6	1	9	55229

Table 4.4: Top-14 Products

Then, the new product details are obtained from decision maker for predicting the price of it. After obtaining the input, the price predictor algorithm compares it with dominant products or top-k products and price of the new product is predicted with greater accuracy. Then the price of new product is returned to decision maker. The Price Prediction Algorithm is as follows

A. Price Prediction Algorithm (PPA)

```
Input:   top-k products where k is the number of products
         features of new product, F
Output: predicted price, P
repeat until k
        for each feature f in F
                for each feature f1 in top-k product
                        if f is equal to f1
                                return P from f1
                        else if f is less than f1
                                return price P as less compared to top-k products
                        else
                                return price P as greater compared to top-k products
                        end if
                end for
        end for
end repeat
```

This algorithm returns the price after comparing with only top-k products and not with all the products available in the market. If new product's features are equal to the features of top-k products, the price of that top-k product is returned. If new product's features are less compared to top-k products, then price of new product is returned which is less to all top-k products' price. If new product's features are larger than top-k products, then price of new product is returned which is greater than all top-k products' price. Thus, the overall response time is minimized with top-k products.

B. Result analysis

To evaluate the performance of proposed methodology, product dataset is used. This dataset consists of customer details and technical specifications of products. The proper protection of personal information is increasingly becoming an important issue in an age where misuse of personal information and identity theft are widespread. So disclosure of personal information can be achieved using k-anonymity technique.

The dataset is anonymized using Flash Anonymization tool kit. The attributes such as age, gender, zipcode and Mobile number are anonymized using generalization and suppression operations of k-anonymity. The distribution of values before and after anonymization is depicted in a following graph.

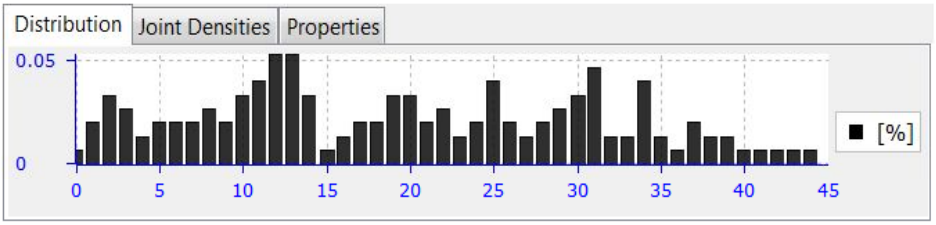

Fig 5.1: Distribution of values of Age attribute before Anonymization

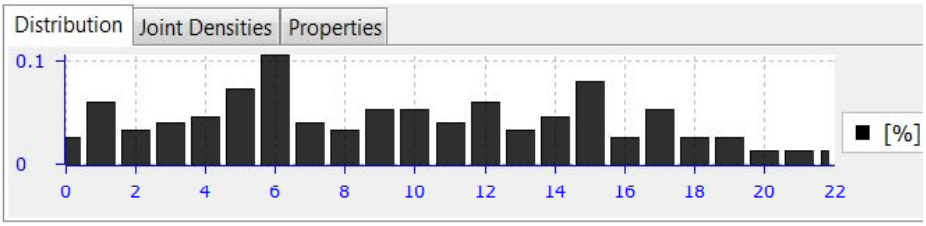

Fig 5.2: Distribution of values of Age attribute after Anonymization

For example gender attribute, there are only two values such as male and female. Distribution of these two values before anonymization is shown in fig 5.3 where these two values are represented in two rectangles. Since these values are suppressed with '*' value, the entire distribution is represented in single rectangle which is shown in fig 5.4. Similarly for other attributes also, the distribution is represented.

Fig 5.3: Distribution of values of Gender attribute before Anonymization

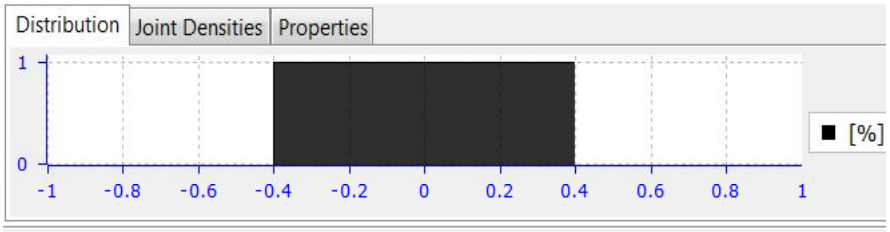

Fig 5.4: Distribution of values of Gender attribute after Anonymization

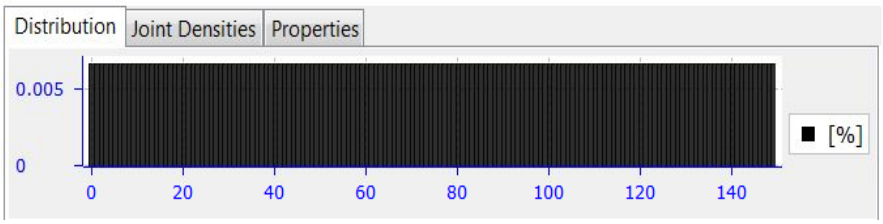

Fig 5.5: Distribution of values of Zipcode attribute before Anonymization

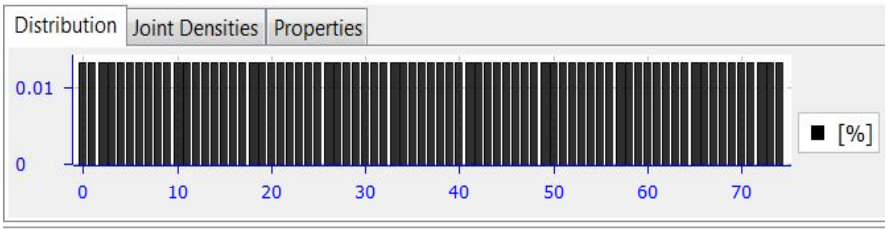

Fig 5.6: Distribution of values of Zipcode attribute after Anonymization

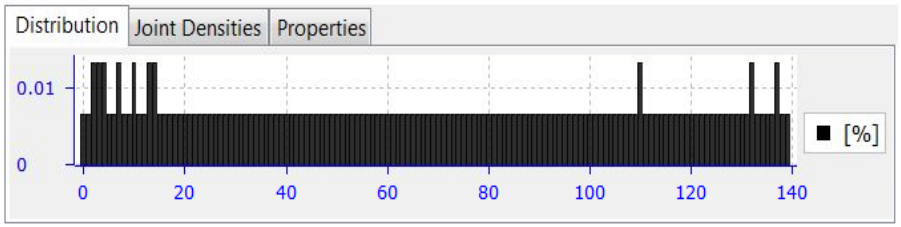

Fig 5.7: Distribution of values of Mobile No attribute before Anonymization

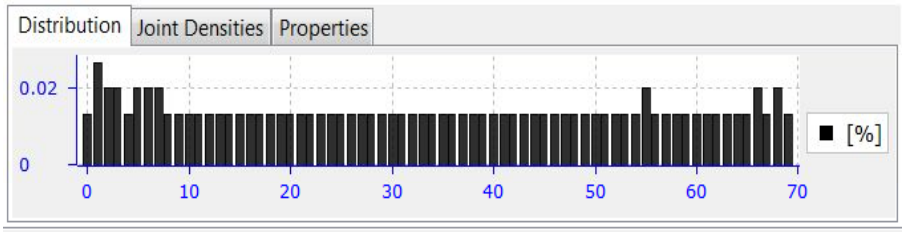

Fig 5.8: Distribution of values of Mobile No attribute after Anonymization

C. Performance Analysis

The proposed work also enhances the performance of price prediction system as it uses top-k or best products for feature comparison. So performance measurement can be depicted with top-k products and without top-k products. Our dataset has 150 instances from which only top-14 products are extracted. These top-14 products are compared with new product and the price of new product is predicted.

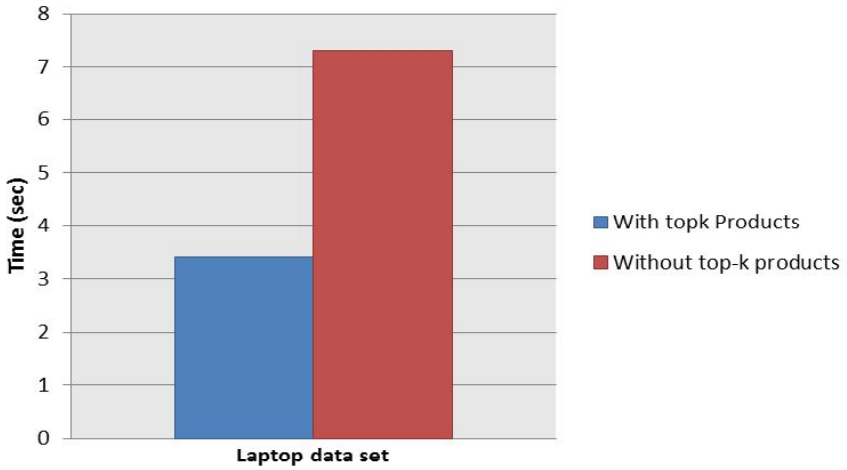

Fig 5.9: Response time for price prediction with and without top-k products

Suppose consider there are n products available in the market. The Price Prediction Algorithm compared with top-14 products. The result is depicted in a graph. The response time for returning price of new product with top-k products is 3.5ms and the response time for returning the same without top-k products is 7.3ms. It shows the execution time for price prediction with top-k products is less than execution time for price prediction without top-k products. Thus the performance of price prediction system has been improved.

IV.CONCLUSION

In this paper, to handle huge dataset efficiently for price prediction, few important concepts were incorporated. Price prediction has done in various fields such as stock market, gold, crude oil and agricultural products. This paper performs price prediction on electronic products such as laptops. In recent years, there is always a need for securing personal information when they are published. So data anonymization technique was used to preserve privacy of individual's data. Irrelevant and unimportant attributes were ignored through skyline computing. Due to extensive presence of data, the discovery of frequent patterns becomes imminent need for many applications. Finally top-k products were extracted. Our Price Prediction Algorithm (PPA) compares the new product with the top-k products and returns the price of new product.

REFERENCES

[1] http://searchsecurity.techtarget.com/tip/Comparing-enterprise-data-anonymization-techniques.
[2] www.cse.psu.edu.uan-k-anon-cluster.ppt
[3] http://arx.deidentifier.org/anonymization
[4] WeiWu C. and Vincent S. (2012) 'Mining Top-k High Utility Item sets' in proc., ACM 978-1-4503-1462
[5] Sweeney L. (2002) 'k-anonymity: A model for protecting privacy' International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.
[6] Han X and Wang J. (2013) 'Efficient skyline computation on Big Data' IEEE Transactions on Knowledge and Data Engineering, vol. 25, No. 11
[7] Tung H. and Wang .S (2006) 'DADA: A Data cube for Dominance relationship Analysis 'ACM 1-59593-256
[8] Machanavajjhala A. and Venkitasubramaniam M. (2007) '*l*-diversity: Privacy beyond k-anonymity' ACM Transactions in Knowledge Discovery Data, vol. 1, no. 1
[9] Park H. and Shim K. (2010) 'Approximate algorithms with generalizing attribute values for k-anonymity' Inf. Syst., vol. 35, no. 8, pp. 933–955
[10] Zhang Z. and Anthony T. (2013) 'K- Anonymity for Crowdsourcing Database' IEEE Transactions on Knowledge and Data Engineering.
[11] Liao W. and Choudhary A. (2005) 'A Fast High Utility Item sets Mining Algorithm' ACM 1-59593-208
[12] Jiang Y. and Tuzhilin A. (2006) 'Mining Actionable Patterns By ole Models' 22nd International IEEE Conference on Data Engineering

[13] Phillips P. and Lee I. (2009) 'Mining Top-k and Bottom-k Correlative Crime patterns through Graph epresentation' IEEE 978-1-4244-4173, Jun 2009.
[14] Lin Jiang and Hang Chung (2010) 'Mining Top-k Sequential Patterns in the Data Stream Environment' International IEEE Conference on Technologies and Applications of Artificial Intelligence
[15] Lucchese C. and Orlandoy S. 'Mining Top-K Patterns from Binary Datasets in presence of Noise' SIAM, 2009.
[16] Jibing Gong and Shengtao Sun (2009) 'A New Approach of Stock Price Trend Prediction Based on Logistic Regression Model' International Conference on New Trends in Information and Service Science
[17] Mori S. and Hirasawa K. (2004) 'A Stock Price Prediction Model by Using Genetic Network Programming' SICE annual Conference
[18] Xiaolong H. and Ming Z. (2010) 'Applied Research on Real Estate Price Prediction by the Neural Network' 2nd Conference on Information Technology
[19] Guiyang Xu (2011) 'China Gold Future Price Prediction Model' 2nd IEEE Conference
[20] Lean Y. and Keung L (2008) 'A Generalized Intelligent-Agent-based Fuzzy Group Forecasting Model for Oil Price Prediction' IEEE International Conference on Systems, Man and Cybernetics
[21] YUC-Yan, MA Jun and ZHAOY- Yan (2009) 'Online Price Prediction and Decision Support for Agricultural Products' International Conference on Information Management
[22] Ninghui L. and Suresh S. (2007) 't-closeness: Privacy beyond k- Anonymity and l-diversity' IEEE 23rd International Conference on Data Engineering
[23] Mamoulis N. and David C. (2013) 'Dominance Relationship Analysis with Budget Constraints' Publication of Knowledge and Informative System
[24] Jichun C. and Fengfang W. (2008) 'Application of BP Neural Network in the prediction of real estate price's chronological sequence' Statistic and Decision, pp. 42-43
[25] Ouyang Jiantao, (2009) 'The application for real estate investment price by nonlinear gray forecast model' Industrial Technology & Economy, Vol. 24, No. 5, pp78-80.
[26] Yuelong Y. and Jizhou Z. (2011) 'Prediction of Urban Land Price based on Grey-Markov Model' International Conference on Computer Science and Network Technology
[27] Junchang X. and Guoren W. (2013) 'Subspace Global Skyline Query Processing' ACM EDBT/ICDT
[28] Shen G. and Leong H. (2013) 'Dominance Relationship Analysis with Budget Constraints' IEEE international conference