iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** XII **Month of publication:** December 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Applications of Data Mining Techniques In Health Insurance

Peeyush Vyas
CE/IT Department, Vadodara Institute of Engineering, Vadodara.

Abstract —Data mining is used to extract patterns in a huge data set. The hidden information can easily be fetched out using the different techniques of data mining. Market Basket Analysis is such a widely used technique which is basically used to get frequent data set. The meaning of basket is collection of transaction data items taken in a transaction item set. It uses mathematical measures like Confidence, Support and Lift. In this paper there is a small effort to make use of Market-Basket Analysis or Association Rule Mining for knowledge discovery in the Health Insurance Company to support decision making using knowledge as premeditated factor. In this paper, also there is a use of different data mining techniques like association rules, clustering, classification etc. for knowledge discovery in health insurance business. These anticipated data mining techniques and the decision- maker may explore the extension of health insurance activities to give the more strength in the present health insurance sector.
Keywords- Data Mining, Market-Basket Analysis, Frequent sets, Association, Clustering, Health-Insurance

I. INTRODUCTION

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. This paper is emphasizing the use of one of the techniques of Data Mining called 'Clustering' along with its various algorithms for the training data set. Here, data extracted can be used to find similar types of data items in a group that are different from other data items in another group.

Also, there is a small try to make use of the Market-Basket Analysis or Association Rule Mining for knowledge discovery in a Health Insurance Company. How Health Insurance Companies can increase their revenue while offering various health insurance plans as per the need of a customer. There are various plans available related to the health insurance and the main goal here is to provide the best health insurance plan according to the customer's need and budget. It will be a useful method for discovering customer's purchasing patterns by the available association. Also, it will make the use of Clustering and Classification on the available data. Market Basket Analysis technique of Data Mining can help the health insurance company for raising its business revenue. While applying this technique, company can come to know the buying pattern of customers for their available combo-pack of different diseases, duration and benefits to the family. It also discusses the actual need of the customer and can design the new health insurance package just to meet the customer's requirements. This way, company can attract new customers in this competitive market and can increase its profit.

Data mining techniques can make the use of linear regression and correlation for business solutions because a health insurance company requires easily accountable models and model criteria. Data mining can help to improve models by detecting nonlinear relationship using important variables and identifying interaction terms. It leads to predict relationships and behaviors more accurately and effectively to get greater profits with reduced costs.

Data Mining can be helpful to the health insurance company in the following business practices like-

To seek for new customers.

To retain existing customers

Maintaining proper classification.

II. WHAT IS HEALTH INSURANCE?

The term Health Insurance or Medical Insurance or Mediclaim is a type of insurance that covers medical expenses of the insurer. Health insurance comes in handy in case of severe emergencies. It provides a financial coverage for medical and hospitalization expenses, in case of an illness, disease or accident. A health insurance policy is a contract between an insurance company and an

individual. Sometimes it is associated with covering disability and custodial needs. The contract is renewable annually. Health insurance is affordable and carries the assurance and freedom from insecurities that threaten normalcy now and then. The type and amount of health care costs that will be covered by the health plan are specified in advance. Health plans are available in two formats, individual and group plans. In an individual policy you are personally the owner of the policy. While in a group plan, the sponsor owns the policy and the people covered under it are called its members.

III. TO SEEK FOR NEW CUSTOMERS

It has always been a great challenge for a company in adding new customers to improve goodwill of company and to improve business revenue. Many traditional and conventional methods are applied by the sales department to improve this but sometimes they are not able to produce the desired results or they are able to target a particular segment only. 'Clustering' is one of the widely used data mining techniques which can be used in the health insurance companies to identify and recognize market segment related to youth, males, females, businessmen, businesswomen or executives. In Clustering, most similar types of data are grouped together and the domain experts are required to define the meaning of created clusters. So, Health Insurance Company can target the untouched consumers.

Example 1

Health insurance companies can focus on the different types of customer segments as shown below as a sample data in the table 1. Using these, company can prepare various groups or clusters and can target the potential customers with the relevant preparation on marketing strategies.

TABLE I *(Sample Data)*

Sex	*Age*	*Income*
Male	≤18	1000
Male	>18	≥15000
Male	≥45	≥100000
Female	≤18	1000
Female	>18	15000
Female	>45	≥100000
Male	≥65	<50000
Female	≥65	<50000

TABLE I.I

Pre-Existing Disease	Yes/No
Injury/illness in Last 48 Months	Yes/No
Filed any previous claim?	Yes/No
Proposal for health insurance declined in past?	Yes/No
Tobacco User?	Yes/No
Medical Tests required?	Yes/No

As far as marketing strategy of a health insurance company is concerned, it can target its customers on the basis of common available attributes like Age, Sex, Income, Occupation, Education, previous health history, etc. Company does not have any predefined for this label. Based on the outcome of the grouping, marketing and advertising strategies can be decided for a specific type of health insurance package. Here each and every attribute is important of its own while deciding the type of health insurance package. For example, while planning for long term health insurance with low premium then only the children below age 18 are to be targeted. For a middle aged person the attributes like age along with income and simple medical tests are to be done. If there is group of only old persons then company has to concentrate on age, previous disease, medical history, any injury case and all the medical tests are required very vigorously.

IV. CLUSTERING AND k-MEANS ALGORITHM

It is a data mining technique to find groups of objects in such a way that that the objects in a group will be similar or related to one another than the objects in other groups.

Given a database D = {T1,T2,…,Tn } of tuples and an integer value k, the clustering problem is to define a mapping f : D → {1,…,k} where each Ti is assigned to one cluster kj, $1 \le j \le k$. A cluster kj, contains precisely those tuples mapped to it i.e. kj = { Ti | f(Ti) = kj, $1 \le i \le$ n, and ti Є D}

The k-means algorithm proceeds as follows:

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining

objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges.

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached.

Input:

D= {t1, t2, t3,...,tn} //Set of elements

k //Number of desired clusters

Output:

k //set of clusters

Algorithm:

assign initial values for means m1, m2, …, mk;

repeat

assign each item ti to the cluster which has closest mean;

calculate new mean for each cluster;

until convergence criteria is met.

V. TO FIND NEW CUSTOMERS AND RETAIN EXISTING CUSTOMERS

In today's throat cut competition, finding a new customer and retaining of old customers has become a challenging and tedious task for any of the companies. Also, facilities provided differ from company to company so customers have ample number of options to compare. Basically, there are different riders and services available like sum assured, cash less hospitalization, pre and post hospitalization expenses, cash allowance, lifetime renewal etc.

In this paper, it is decided to set frequent item sets based on predefined support and there is a need to find out all the association among the frequent data items.

Example 2

Market Basket Analysis can be effectively used in a health insurance company. The analysis of data depends upon, which type of health insurance packages are being purchased by a consumer. Some association rules can be generated showing that which riders of health insurance are purchased together. On these available facts, companies can decide some sort of available association between different health insurance plans that are sold for various purposes. The different transactions with sample data can be shown as below. For simplicity we have included the most commonly used riders and facilities. The abbreviations used for the same are as follows –

Sum Assured : SA(SMS), Cash Less Hospitalization : CLH(LOCAL), Pre & Post Hospitalization Expenses : PPHE(2G), Cash Allowance : CA(3G), Life Time renewal : LTR(STD), Duration: DU(International).

TABLE II

Sno.	Transaction	Item
1	t1	CA, SA, CLH
2	t2	PPHE, SA,CLH
3	t3	CA,CLH
4	t4	CA,SACLH,LTR,DU
5	t5	CA,SA,CLH,LTR
6	t6	PPHE,CLH
7	t7	CA,CLH,LTR
8	t8	PPHE,SA,CLH,LTR
9	t9	CA,LTR,DU
10	t10	CA,CLH,LTR,DU

The health insurance company can target youth while providing a health insurance package containing more facility related to sum assured. Similarly, youth &females can be targeted for pre & post hospitalization expenses and cash allowance along with sum

assured. Males can be offered for cash allowance, cash less hospitalization and life time renewal. Females can be targeted while providing pre & post hospitalization expenses with sum assured and cash less hospitalization and executives can be targeted while providing Cash less hospitalization, life time renewal, and duration along with cash allowance.

An association rule in a database can be viewed as a set of tuples and each tuple containing data items. In the above examples there are ten transactions and six data items: {PPHE,CA,SA,CLHLTR,DU} as {S1,S2,S3,S4,S5,S6}

Here, the main problem is to find out all the instances where customers who bought a subset of a frequent item set, most of the time also bought remaining data items in the same frequent set. For a given frequent item set, say {S1,S2,S3}, if a customers who buys a subset formed by S1 and S2 also buys S3 75% of times then there is some sense to apply the rule. This percentage is called confidence of the rule. Confidence of rule "B given A" is a measure of how much more likely it is that B occurs when A has occurred. It can be defined as the ratio of the number of transactions that include all data items in a particular frequent item set to the number of transactions that include all items in the subset.

$$\text{Confidence } I = \frac{P(X \cap Y)}{P(X)}$$

Range [0, 1]

If I=1 then most interesting; If I=0 then least interesting

On the other hand, a support measures how often the collection of items in an association occur together as a percentage of all the transactions.

$$\text{Support } I = \frac{(X \cap Y)}{N}$$

Range [0, 1]

If I=1 then most interesting; If I=0 then least interesting

To clear the above rules, let us consider the following examples where we want to find out the required association rule.

Support = 30% - only the items that are bought together by at least 3 customers are considered.

Confidence = 90% - In 90% of the transactions, the association rule is to be true.

Case 1: (S2, S4)→S3

(S2, S4) was bought by 5 customers but only 3 of them also bought S3, so the Confidence is 60%

Case 2: (S5, S6)→S2

(S5, S6) was bought by 3 customers and all 3 of them bought S3 as well.

So, Confidence is 100% so this rule has strong confidence and it is to be considered as it is more than 90%.

VI. APRIORI ALOGORITHM

Apriori Algorithm Pseudocode

Join Step: C_k is generated by joining $L_{k-1}$with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-I itemset

Pseudo-code:

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 = {frequent items};

for(k = 1; L_k !=∅; k++) do begin

C_{k+1} = candidates generated from L_k;

for each transaction t in database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return$\cup_k L_k$;

VIII. CLASSIFICATION

To improve predictive accuracy, the database can be further subdivided into more similar groups. The segmentation can be done

using different attributes like renewal charges on sum assured, No claim bonus, ambulance charges, ICU charges and buying behavior of the customer. It helps in the further classification and the classification maps data into predefined groups. On these attributes, classes are defined and used in the classification algorithms. Hence, health insurance companies can more accurately predict the suitable health insurance packages for an appropriate person depending upon Age, Sex, income and medical history.

Example **-** Using occupation, Age, income and Medical history, health insurance company can find a segment and these patterns can be stored in the databases. It will help in selling a health insurance package depending upon these attributes. These attributes can be compared with the stored patterns in the databases by the executive and a suitable health insurance package can be suggested to the customer.

TABLE III

Age	Occupation	Income (Monthly)	BP	SUGAR	HEART	KIDENY	OTHER DIESESE	MEDICAL TEST REQUIRED
≤18	Student	<=1000	OK	OK	OK	OK	NONE	NO
18-25	Service	1000-30000	OK	OK	OK	OK	NONE	NO
18-25	Business	1000-30000	YES	NO	NO	NO	NONE	YES
25-35	Service	30000-50000	NO	YES	NO	NO	NONE	YES
25-35	Business	30000-50000	YES	YES	NO	NO	NONE	YES
35-40	Service	50000-100000	NO	NO	NO	NO	NONE	NO
35-45	Business	50000-100000	YES	YES	NO	NO	LEVER	YES
45-50	Service	100000-150000	YES	NO	YES	NO	NONE	YES
45-50	Business	100000-150000	YES	YES	YES	NO	NOE	YES
50-55	Service	150000-200000	NO	NO	NO	NO	NONE	YES
50-55	Business	150000-200000	YES	NO	NO	YES	NONE	YES
55-60	Service	200000-500000	YES	YES	YES	NO	STONE	YES
55-60	Business	200000-500000	YES	YES	YES	YES	NO	YES
>=60	Retired	>=500000	YES	YES	NO	NO	NO	YES

In this example, suppose we want to classify the customers based on the attributes like Age, Occupation, Income, BP, Sugar, Heart, Kidney etc. then the different health insurance package classification with the prediction of medical tests required or not can be done very easily.

IX. CLASSIFICATION ALGORITHM

A. K Nearest Neighbors

It is a non-parametric technique used for pattern recognition and Classification. Here, k closest training examples in the feature space are taken as input. Feature space is an abstract space defined by feature extraction procedure that transforms raw data into sample vectors i.e. an n-dimensional vector of numerical features that represent some object. Output is a class membership. An object is classified by a majority vote of its neighbors with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In this algorithm, distance of new item in the training data set must be determined. The new item is then placed in the class that contains the most common items from the (K) closest set.

Input:

/* T Training data, K Number of neighbors, t Input tuple to classify */

Output: c //class to which t is assigned

Algorithm:

N = ø

//Find the set of neighbors, N, for t

For each d € T do

If |N| ≤K, then N = N Ụ {d};
else
if u € N such that sim(t,u) ≤ sim(t,d), then
begin
N = N – {u}; N = N Ụ {d};
end
//Find class for classification
C=class to which the most u € N are classified;

X. CONCLUSION

Data mining techniques can be used in decision making in a health insurance company to increase company's business. Here, data mining techniques like classification and market basket analysis are required to understand by the health insurance company. These techniques are certainly helpful to retain existing customers and to grab new customers in the competitive era. While using Association rule new combinations are created according to the mathematical measured like 'Confidence' and 'Support' that can help the company to sell new health insurance packages to the existing customers according to the customer's need. Classification, on the other hand, can be used to seek and mark new customers or scheming new combo of health insurance packages.

REFERENCES

[1] Alex Berson and Stephen J. Smith, "Data Warehousing, Data Mining, And OLAP", McGraw–Hill, 1997.

[2] Jiawei Han, Laks V. S. Lakshmanan and Raymond T.NG, "Constraint-Based Multidimensional Data Mining", IEEE, August 1999. Chen, Y.-L., Chen, J.-M., & Tung, C.-W. (2006).

[3] A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. Decision Support Systems, 42 (2006), 1503-1520.

[4] J. Blanchard, F.Guillet, and H. Briand.Exploratory visualization for association rule rummaging. In KDD 03WorkshoponMultimediaDatAMining(MDM-03),2003

[5] L.Cavique. A scalable algorithm for the market basket analysis. Journal of Retailing and Consumer Services,14(6):400407, 2007

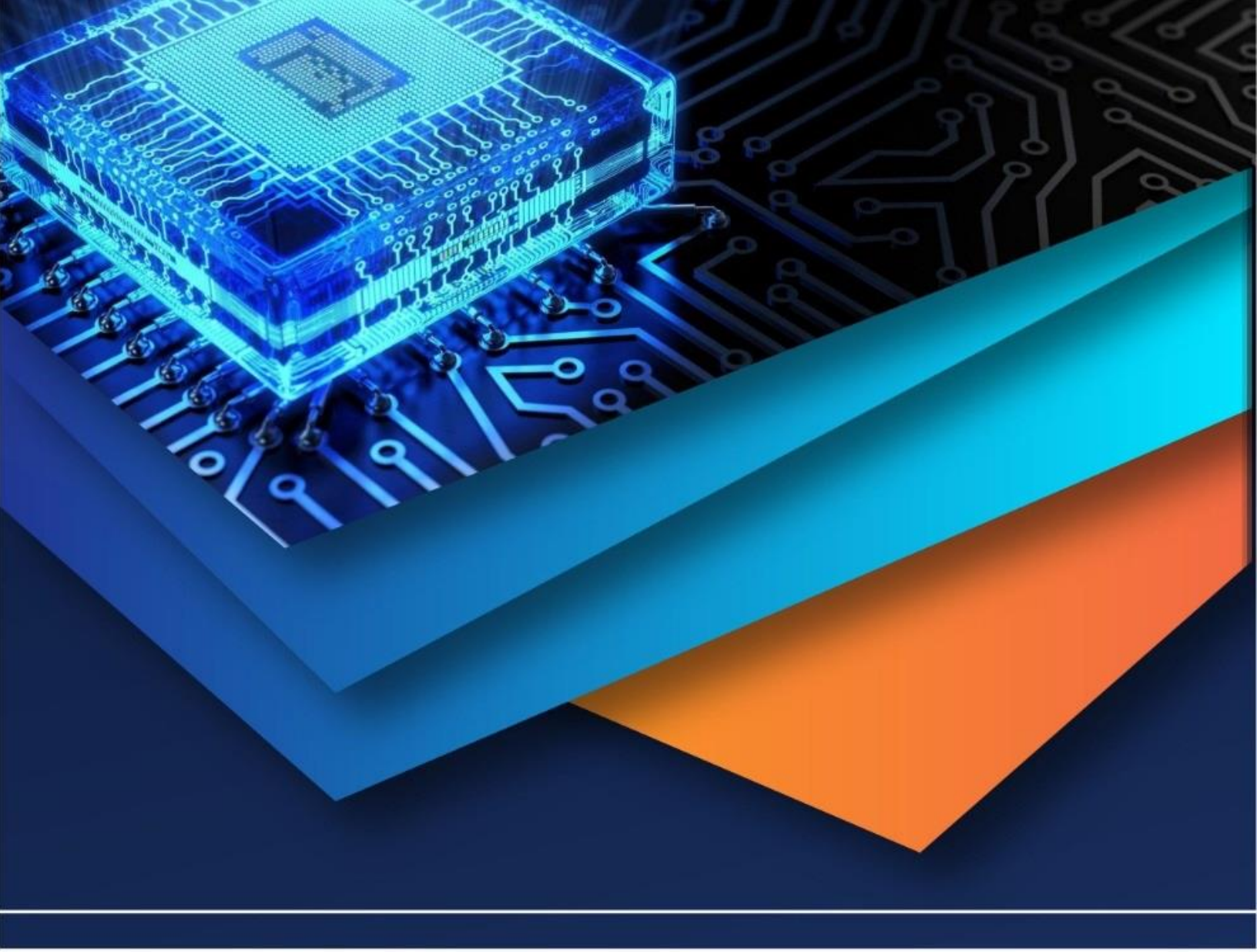
[6] R.Agrawal and R.Srikant.Fast algorithms for mining association rules in very large databases. In Proceedings of the 20th International Conference on VLDB, pages 487499, Santiago, Chile, 1994.

[7] R.Agrawal and R.Srikant.Fast algorithms for mining association rules in very large databases. In Proceedings of the 20th International Conference on VLDB, pages 487499, Santiago, Chile, 1994.

[8] Mr. A. B. Devale and Dr. R. V. Kulkarni "A REVIEW OF DATA MINING TECHNIQUES IN INSURANCE SECTOR" Golden Research Thoughts Vol- I , ISSUE - VII [January 2012]

[9] Martin Staudt, AncaVaduva and Thomas c, "Metadata Management and Data Warehouse", Technica Report, Information System Research, Swiss Life, University of Zurich, Department of Computer Science , July 1999, vaduva@ifi.unizh.ch

[10] Neha Khandelwal et.al " Climatic Assessment Of Rajasthan's Region For Drought With Concern Of Data Mining Techniques" in International Journal Of Engineering Research and Applications (IJERA) ISSN: 224-9622 www.ijera.com Vol. 2, Issue 5, September- October 2012, pp.1695-1697



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY