Insurance Analyst

Nikhitha Chandran¹, Saranya P S², Bhavya V R³, Roshni T A⁴
Department of Computer Science, Vidya Academy of Science and Technology, Thrissur, INDIA

ABSTRACT

Life insurance, basically a tool against protection of life or against death of individuals or any unforeseen event. It provides financial protection against such kind of risks. For every individual, the purpose of investments in life insurance might differ. According to the survey at present there are 23 private life insurance companies and one public life insurance company in India. Due to this large number of life insurance companies and having wide range of products there arise confusion among the investors as to which product and also of which company to purchase. So in order to form their updated marketing strategies, marketers are interested to know investment pattern of life insurance investors. It is possible to predict which type of customers are interested in certain kind of life insurance product and also developing new products based on company profitability (different products can be developed for different customers and marketing strategy can be made for them) using data mining techniques. Generally the Data mining techniques are used to extract hidden patterns, from large amount of data in the form of data-ware house. In the this research paper emphasis is on investor’s investment behavior in life insurance sector of India by using data mining techniques and implement big data with data mining concept. Here we consider two objective i.e., To find the list of potential customers and to find a reference plan in order to develop a new policy product.

Keywords---- Data mining (DM) ; Big data ; CRM

I. INTRODUCTION

This paper implements big data with data mining concept. Big data is a collection of dataset, so large and complex that it becomes difficult to process using traditional data processing application. Big data extract using data mining techniques. Data mining is the process of extracting knowledge from the large amount of data. This paper aim to present how data mining with big data is useful in insurance industry, how to produce good result and how data mining enhance in decision making using insurance data. The rapid growth of accumulation of digital data by various life insurance companies is attributed to the growth in the field of information technology. The lots of valuable information hidden in the data are hardly exploited.

The primary goal of data mining is to extract knowledge from data to support decision making. It is also known as technique of exploring data in order to discover previously unknown patterns. By performing data mining interesting regulates or high level information can be extracted from data bases and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management and query processing. In this paper using big data concept we develop a new data mining algorithm and it has been applied to find out our two objective i.e.,
- Potential customer i.e., which type of customers are interested in certain kind of life insurance product.
- New policy product i.e., with the help of data mining techniques by using the available data from different customers different products can be developed for different customers and marketing strategy can be made for them. Here an objective is to find a reference plan for developing a maximum profit new policy product.

a. Design Architecture
Steps to follow; Data mining refers to extracting or mining knowledge from large amount of data. Data mining as a synonym for another popularly used term, knowledge discovery from data or KDD.
- **Selection**: Selecting data relevant to the analysis task from the database.
- **Transformation**: Transforming data into appropriate forms to perform data mining
- **Data mining**: Choosing a data mining algorithm which is appropriate to pattern in the data,
extracting data patterns.

- **Interpretation/Evaluation:** Interpreting the patterns into knowledge by removing redundant or irrelevant patterns. Translating the useful patterns into terms that human understandable.

- Customer level Analysis
- Marketing and sales analysis
- Developing new product lines
- Reinsurance
- Financial analysis
- Estimating outstanding claims provision
- Detecting fraud

In the field of insurance industry decision support system plays an important role. Data mining used to support the controlling of policies, the administrative and management tasks, efficient management of organization and financial data.

### III. DATA MINING TECHNIQUES

Data mining techniques have applied to various insurance domains to improve decision making. Data mining use predictive modeling, market segmentation, market basket analysis to answer business questions with greater accuracy. Various data mining techniques used for the insurance industry development are Classification, Clustering, Regression and Association rules, summarization used for knowledge discovery from database.

**a. Classification.**

It is one of the most commonly used techniques, to develop models that can population records at large. The classifier training algorithm uses this technique for business development. Various classifiers are used for the classification algorithms such as Decision tree, Bayesian classifier, neural network, Support vector machine etc. Customer database can be segmented into homogeneous groups, classification maps data into predefined group into segments. Data mining classification algorithms are applied on insurance benchmark data set. Types of classification are

- Supervised classification
- Unsupervised classification clustering

**b. Clustering**

It used for identification of similar classes of objects. It’s used for grouping based on the customer’s behavior. It is applicable for customer segmentation and targeted marketing. Types of Clustering are

- Hierarchical agglomerative methods
- Partitioning methods
- Density based methods
- Grid based methods
- Model based methods

**c. Regression**

It can be used for prediction. Regression analysis used to model the relationship between one or more independent and dependent variables. In insurance firm, more complex
techniques needed to predict future values. Types of regression includes

- Linear regression
- Non-linear regression
- Multi-variant linear regression
- Multi-variant non regression

d. Association

Insurance companies faced a lot of problems on customer retention. Association used for this task, because it finds all the association where customers bought a frequent item set. Association helps business firms to make certain decisions. Market basket analysis and cross selling programs are typical examples for which association modeling is usually adopted.

When the customers want to insure some policy then this technique helps us finding the associations between different items.

Classification - Predicting consumer behavior & Prediction
Predicting the likelihood of success of policies
Classifying the historical customer records
Prediction of what type of policy most likely to be retained, most likely to be left
Predicting insurance product behavior and attitude
Predicting the performance progress of segments throughout the performance period
Prediction to find what factors will attract new avenues in insurance sector
Classify trends of movements through the organization for successful/ unsuccessful customer historical records

e. Summarization

This technique which used for report generation provides better decision making for large volume of customer database with the help of visualization tools. It will provide more functionality in business decision making. For solving the business problems and making decision, this data mining techniques can be help to the organization but selecting the appropriate techniques can important for the organization.

IV. DATA MINING TASKS

Data mining is becoming common in both the insurance sectors like private and public. Data of the customer are one of the most valuable assets of any firm. The traditional methods, which were used for handling huge amounts of data generated by insurance transactions, are too complex. For transferring huge amount of data for decision making, data mining makes the methodology. Insurance firms use the data mining methodologies to enhance research and increase sales among the customers.

The data mining used for various tasks in the insurance sector as follows:

- Policy designing and policy selection
- Prediction
- Claims management
- Developing new product lines
- Underwriting and Policy management
- Risk management
- Reinsurance
- Fraud detection
- Trend analysis

V. TECHNOLOGIES USED

Hadoop:

Hadoop framework is designed to provide a reliable, shared storage and analysis infrastructure to the user community. There is a storage portion and an analysis functionality portion in this framework. The storage portion is provided by a distributed file system solution such as HDFS, while the analysis functionality is presented by MapReduce. Several other components are part of the overall Hadoop solution suite. The MapReduce is a tool for deep data analysis and the transformation of very large data sets. Hadoop allows the users to explore and analyze complex data sets by utilizing customized analysis scripts/commands. In other words, it can be said that through the customized MapReduce routines, unstructured data sets can be distributed, explored, and analyzed across thousands of shared-nothing processing systems/clusters/nodes. Hadoop’s HDFS uses replication technique on the data to safeguard the environment from any potential data-loss, by saving the data onto multiple nodes. HDFS (storage) and MapReduce (processing) are the two core components of Apache Hadoop.

The most important aspect of Hadoop is that both HDFS and MapReduce are designed with each other in mind and each are co-deployed such that there is a single cluster and thus provides the ability to move computation to the data. Thus, it seems that the storage system is not physically separate from a processing system. HDFS is a distributed file system that provides high-throughput access to data. It becomes possible by providing limited interface for managing the file system.

HDFS creates multiple replicas of each data block and distributes them on computers throughout a cluster to enable rapid and reliable access. The 2 main components of HDFS are NameNode and DataNode. NameNode is called the master of the system. It maintains the name system (files and directories) and manages the blocks which are kept on the DataNodes. DataNodes are the slaves which are deployed on each machine. It provide the actual storage and are responsible for running read and write requests for the clients.
Normally any set of loosely connected or tightly connected computers that work together as a single system is called Cluster. In simple words, a computer cluster used for Hadoop is called Hadoop Cluster. Hadoop cluster is a special type of computational cluster designed for storing and analyzing vast amount of unstructured data in a distributed computing environment. These clusters run on low cost commodity computers. Hadoop clusters are well known for boosting the speed of data analysis applications. They also are highly scalable ie, If a cluster’s processing power is overwhelmed by increased volumes of data, to increase throughput additional cluster nodes can be added .Hadoop clusters also are highly resistant to failure because each piece of data is replicated on other cluster nodes, which ensures that the data will not lost if there is a failure in one node.

**MapReduce:**

MapReduce is a framework for performing distributed data processing using the MapReduce programming paradigm. In this programming paradigm, each job have 2 phases. A map phase followed by a reduce phase. These 2 are user-defined phases. Map phase is a parallel, share nothing processing of input. In reduce phase output of the map phase is aggregated. For these phases we have 2 algorithms, called mapper and reducer algorithms. The mapper will be having actual data as input and sorted form of input as output, which in turn will be the input for the reducer algorithm. HDFS is the storage system for both input and output of the MapReduce jobs. There are mainly 2 components in MapReduce. They are JobTrackers and TaskTrackers. JobTracker can be termed as the master of system which manages the jobs and resources in the cluster. It tries to schedule each map as close to the actual data being processed. TaskTrackers also termed as slaves which are deployed on each machine. They make sure that the map and reduce tasks are take place as instructed by the JobTracker.

MapReduce have complicated algorithms. So in order to make it simple, languages like pig, hive were invented. These language converts our scripts into MapReduce format.

**Pig:**

Pig is a dataflow system which runs on top of Hadoop. It offers SQL like language, Pig Latin to express the data flow and transfer the dataflow to map reduce jobs. Also, it supports user defined functions which can currently be implemented in Java, Python, JavaScript and Ruby. Though it is developed to run on top of Hadoop, it can also be extended to run on top of other systems. Pig allows three execution modes: interactive mode, batch mode and embedded mode. In interactive mode, the user issues commands through an interactive Grunt shell. Only when a store command is given, Pig would execute all the commands. In batch mode, the user runs a pre-written query script, which typically ends with Store. Embedded mode allows user to embed Pig query in java or other programs.

**CentOS:**

CentOs (abbreviated from Community Enterprise Operating System) is a linux distribution that attempts to provide a free,enterprise-class, community-supported computing platform which aims to be functionally compatible with its upstream source, Red Hat Enterprise Linux (RHEL). In January 2014, CentOS announced the official joining with Red Hat while staying independent from RHEL, under a new CentOS governing board.

**VI. DEVELOPMENT**

This project includes 5 phases. Data Collection, Selection, Transformation, Data Mining and finally the Evaluation phase. In the first phase, data collection process is carried out. The Insurance Data, both structured and unstructured. This second phase is the selection phase. In this phase the required data are selected from the whole data. The third phase is the Transformation phase. In this phase the selected data’s are transformed into appropriate form and sort it. The fourth phase is the Data Mining phase. In this phase the data’s are analyzed. In the final phase the analyzed data’s are evaluated into the final result.

- **1st PHASE:** Here, from the unstructured data, data collected and preprocessed into CSV files. A CSV is comma separated values files which are stored in tabular data form. Pre-processing is required to improve the input data quality and to generate enhanced data for further processing.
- **2nd PHASE:** In this phase From the CSV Files, the required data were selected. For e.g.: we can select Policy Details, Plans, Customers, and so on as required by the query.
- **3rd PHASE:** Gives table format of data. That is, the selected data were transformed into table format.
- **4th PHASE:** Here, we analyze the transformed data based on query. Our 1st query is Finding Potential customers

Here we are getting the list of customers who are fit to join or choose the policy that is to be introduced.

Next query is New Policy Product

This gives us the maximum profit policy in the company so that it can be used as a reference for designing a new policy.

- **FINAL PHASE:** This phase gives the visualization of results. The analyzed results were stored and published into a portal. All the results
can be visualized as charts, bar diagrams or tables etc.

• For implementing this we have to install centos

6.1 Installation in CentOS

• Disable UEFI Boot in new Systems
• Disable secure boot if required
  – (In windows 8 systems)
• Install CentOS Minimal Installation
• Enable LAN
  – /etc/sysconfig/network-scripts/ifcfg-eth0
  o ONBOOT=yes
  o NM CONTROLLED=no
  – /etc/sysconfig/selinux
  o SELINUX=disabled
• Install wget and nano
  – yum install wgetnano -y
• Download Java
  – wget-c http://www.xeoscript.com/java.rpm
    • Install Java
  – rm -ivsjava.rpm
    • Download Cloudera Repository
    • Install Cloudera Repository
  – rpm -ivscleoudera-repo.rpm
  • Install pig
  – yum install pig -y
  • Install Cloudera Manager
  • Download parcels
  • Setup Cluster

6.2 Algorithm

Here we have 2 objectives as described above and hence developed 2 algorithm two find out the result for them.

6.2.1 Algorithm for finding potential customer

Inputs - Income of customers and plan details.
Data.csv contains all data (in csv format) and plan details in plan.txt.
Step1: Start
Step2: LOAD all PLAN_DETAILS from plan.txt
Step3: LOAD all WHOLE_DATA from data.csv
Step4: Group the WHOLE_DATA by contact number by using FOREACH operator and enclose schema in parenthesis using FLATTEN.
Step5: For all the unique data generate annual_income, name, contact_number, id, expiry, monthly_income as
WITH_MONTHLY_INCOME_VALIDITY
(Expiry = start date + (plan_validity+1)*60*60*24*30) & Monthly_income = annual_income / 12 )
Step6: Find the cross product of WITH_MONTHLY_INCOME_VALIDITY and PLAN_DETAILS as WITH_PLAN_DATA
Step7: For all WITH_PLAN_DATA generate plan_name, name, contact_number, expiry, plan_sum_assured, plan_amount, plan_policy_profit, plan_validity, has_expired, bal_amount, monthly_income as
WITH_HAS_EXPIRED_BAL_AMOUNT
( has_expire = expiry - (plan_start * 60 * 60 * 24 * 30) & bal_amount = (monthly_income - 1200) - (plan_amount * 2) )
Step 8 : Filter
WITH_HAS_EXPIRED_BAL_AMOUNT by conditions has_expired > 0 && bal_amount > 0 as EXPIRED_USERS
Step 9: For all EXPIRED_USERS generate plan_name, name, bal_amount, monthly_income, contact_number as POTENTIAL_USERS
Step 10 : Display POTENTIAL_USERS

(VOTE : Algorithm takes input plan_start in seconds. Hence in order to do further calculations convert it by multiplying (60*60*24*30).

6.2.2 Algorithm for New policy product

Inputs – Profit details and No. of customers
Step1: Start
Step2: Load WHOLE_DATA from newplan.txt using pigstorage as default
Step3: Generate new item set GENERATED_DATA from WHOLE_DATA using FOREACH operator
Step4: Calculate income as net premium*number of customers
Step5: GENERATE_DATA WITH _ EXPENSE from GENERATED_DATA using FOREACH operator
Step6: Calculate each expense value as income*expense/100
Step7: Calculate expenditure from GENERATE_DATA_WITH_EXPENSE using FOREACH operator
Step8: Calculate profit as a income-total _expense
Step9: DUMP profit
Step7: Stop

VII. CHALLENGES

In insurance organizations, processing large quantities of data during data mining has some of the
challenges. Data mining system faces a lot of problems and pitfalls when handling customer’s data such as Noisy data High volume and high complexity for different kinds of data. Hybrid one or more techniques Corrupted values Missing attribute values. One of the biggest challenges that insurance faces is improve the customer retention and higher revenue.

VIII. CONCLUSION

In the present Scenario in Life Insurance sector in India, the companies can have competitive edge over other companies by managing analyzing their customer data base efficiently. By using data mining techniques the important information can be extracted from these customer data bases and the life insurance companies can gain the customers and new products can be developed which suits to the customers, which in turn increases the profitability of the company.

Potential Customer: - By using the data mining techniques the life insurance companies can segment the customer data base. i.e. Analysis of the customers can be made depending upon session/ geographical / financial base. So with the help of mining techniques typical profiles of life insurance customers with respect to the various Insurance products can be find out i.e. which type of customers are interested in certain kind of life insurance product. Here result is a list of potential customers.

Developing New Products: - The profitability of the life insurance companies can be increased by identifying the sound financial customer segment and marketing strategy can be made for them accordingly. If the life Insurance companies fail to deliver the right product to right customers the company will show downward trend. So with the help of data mining techniques by using the available data from different customers different products can be developed for different customers and marketing strategy can be made for them. Here result is a reference plan for developing a maximum profit new policy product. It is unlikely that Big Data alone will ever become a top priority for insurers. Like other technologies, it will continue to be an enabler for businesses. It can however influence each and every part of the insurance value chain and will offer newer insights to businesses at a swift speed. Big Data can help significantly in understanding customer preferences and concerns to help insurers devise newer products and refine existing service levels to stay competitive and profitable.

IX. REFERENCES

[14] Nanthawadee SUCHARITTHAM,Thanaruk Theeramunkong,Choochart Haruechaiyasak ,Bao Tu Ho, Dam Hieu Chi, ”Data Mining for Life Insurance
Knowledge Extraction: A Survey,” School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand
[15] Gabor Melli, Osmar R. Zaane, Brendan Kitts,”Introduction to the Special Issue on Successful Real-World Data Mining Applications,” SIGKDD Explorations Volume 8, Issue 1
[20] StackIQ White Paper, Capitalizing on Big Data Analytics for the Insurance Industry, 4225 Executive Square, Suite 1000, La Jolla, CA 92037, 858.380.2020, info@stackIQ.com