Predictive Analysis for Having Loyal Costumer in an Insurance Company

Karuna Midha[1] Himanshu Garg[2] Shubham Sharma[3] Rohan Nagpal[4]
[1]Assistant Professor [2,3,4]Student
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Maharaja Agrasen Institute of Technology, India

Abstract— In today's world as health issues among people increases, people become more aware for the health insurance. It's a positive thing for the health companies but as the no. of customers increases, it is come into light that people are not punctual for paying the premium of the policy. This paper helps the policy companies to highlight and point out the defaulters who haven't paid their premium. Mostly people forget about it, and some of them not paying the premium on time. In this research paper, I tried to understand the consumer behaviour in Insurance sector. The main objective of this paper to identify customers behaviour of paying the policy premiums and will they pay their next premium on time or not. Even will they pay their premium or not irrespective of time. Data was collected by various sites and some previous years data of some policy companies. Frequencies, Tabulation and some Data Science models have been used for the analysis. The objective of this project is to summarize is to make a predicting algorithm that can be used in real life applications to derive meaningful and accurate prediction based on the various aspects of data that is accessed.

Keywords: Insurance Company, Customers Behaviour

I. INTRODUCTION

Insurance sector plays a very important role in the development of the economy. Insurance company relies completely on its customers. Customer and Insurance company relationship is completely dependent on each other. An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that you pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a medical insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalized in that year, the insurance provider company will bear the cost of hospitalization etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalization cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalized that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is life insurance where every year you pay a premium of certain amount to insurance provider company so that in case of unfortunate event of your death, the insurance provider company will provide a compensation (called 'sum assured') to your immediate family. Similarly, there can be a variety of insurance products for different kinds of risks. As you can imagine, if a large number of customers do not pay the premium on time, it might disrupt the cash flow and smooth operation for the company. A customer may stop making regular premium payments for a variety of reasons - some may forget, some may find it expensive and not worth the value, some may not have money to pay the premium etc.

Doing a research to whether a customer would make the premium payment can be extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers who are less likely to pay and convince them to continue making timely payment.

Now, in order to predict, whether the customer would pay the next premium or not, we have information about past premium payment history for the policyholders along with their demographics (age, monthly income, area type).

II. RELATED WORK

Literature Review is required to take the matter into considerations that can't be cleared in the past researches. Many researchers try to interpret various kind of conclusions and to improve those past results literature review is needed. Many professors studied and worked on various factors but didn't predict with very understandable score to come into any conclusion. The present literature serves many varied interesting features, which forms the vital background for the study and conducted a consideration.

Viswanadham (2015) studied claims settlement operations of insurance companies with the objectives of evaluating performance in terms of both maturity and death claims before and after IRDA period. Claim settlement processing time expressed in speed ratios and adjudicatory measures of the corporation to redress the grievances of policyholders in settlement of claims. The study concluded that corporation should provide efficient service with courtesy in the matters of claim settlements. It should create highest trust in the minds of policyholders by establishing open and transparent grievance redressal procedure. As, satisfied customer will be a brand ambassador for the insurance company; claim settlement should be given more importance

K.Swathi, R.Anuradha (2017) their studies concludes that Rising healthcare costs can punch a big hole in your pocket. Paying a small health insurance premium is the simplest way to mitigate the financial losses and to achieve peace of mind. Avail health insurance to save yourself from worries of hospitalization. One of the main reasons for the low penetration and coverage of health insurance is the lack of competition in the sector. The Insurance Regulatory Authority of India (IRDA) which is responsible for insurance policies in India can create health circles, similar to telecom circles to promote competition.

All rights reserved by www.ijsrd.com

.Government should still conduct awareness programmes to inform the people about the benefits of health insurance.

Neuman (2017) worked on various factors on which a company evaluates the customers. Customers mostly forget about the premium payment as after due date they usually ignore it or postpone it for a time. But after sometime, they didn't bother about the payment as well. Due to this companies usually marked this kind of customers as defaulters. This study concluded that irresponsible nature of customers is the major cause for the customers to not to pay the premium on time.

Albert (2013) noted in his article that age of the customers played the most important role. As the customers of older age forgets to pay the premium on time. As age is the factor, memory becomes weaken day by day. Even at the time they forget about details and sometimes forget for what kind of policy they have registered. This study concluded that customers with higher age forgets to pay policy because of the priority issues and sometimes mental conditions also considered as well.

Jagendra Kumar (2015) by his study, revealed that customers of different age level and maturity levels, take different approaches to pay the premium. Customers at the starting years of the policy pay the premium timely and frequently and as years passes, they become less responsible. Moreover, people of age group of 30-45 take the policy more seriously. In all these responses emotional factor of people to their family and their working schedules are the main reasons for their approach of payment of premiums.

Yusuf, Gbodamasi and Hamadu (2019) conducted an empirical study on the customers that income of these customers is somehow plays a role on their manner of paying the premium. It is observed customers having low income sometimes deliberately tries to avoid the payment of premium. On the other hand, people having high income or we can say, constantly having higher income rate pays the premium on time and sometimes in advance. This can be observed that income plays vital role in payment of premiums of policy.

III. PROBLEM STATEMENT

Insurance Company wants to analyze the behaviour of their customers for paying the premium. The company wants to analyze how likely the customers will pay their next premium. Moreover, if customers will pay the premium, then will they pay the premium on time or not. Part of our research works on the behaviour of all the customers of the insurance company.

IV. SCOPE & MOTIVATION

This project is really beneficial for the companies who really are on the verge of take down as they can't make enough revenue due to their customers. It can really help them analyze where they are lacking in terms of man power or some strategy and they can improve the revenue of their company and end their relationship with the default customers of their company.

A. Project Objective

The main objective of the research paper is to identify the customers who are likely to pay the next premium in future and are valuable for the company. The factors on which their premium is being paid and other factors also taken into consideration on which prediction have been done.

V. RESEARCH DESIGN

Research has been done on various stages as listed below.

1) Decision of necessary factors required for the research
2) Collection of Data samples
3) Data Visualization
4) Data Exploration

1) Factors Required:

Marking all the factors and listing them as necessary and unnecessary for our project based on our hypothesis. Hypothesis testing is the part where we picturize all our requirements from the data before looking at the actual data. Below factors are taken into consideration.

1) Unique ID of the customers
2) Age in days of policy holder
3) Monthly income of policy holder
4) No. of premiums late by 3-6 months
5) No. of premiums late by 6-12 months
6) No. of premiums late by more than 12 months
7) Total premiums paid on time

2) Collection of Data Samples:

"Collection of Data" is the most important step on the working of the research. In this study, the researcher has aimed to predict the quality customers for the insurance company. Customers who aren't beneficial in long term for the company are marked as defaulters. This study was based on primary information and the data were quantitative in nature. Samples were taken by adopting convenient sampling technique. Various insurance companies' websites and informative closed-end databases were used as a tool of data collection. Descriptive studies of individuals have been done. Standard factors are taken into consideration for the collection of data. Permission for accessing of databases is asked by respective companies, whenever needed. People of every category, age-group and sex are included and not any particular kind of group is excluded on any basis. Data, to as much extent, is cleaned and there are not many missing values taken into consideration while doing the research.

3) Data Visualization:

The raw data has been collected through different techniques as discussed above, needed editing and processing. In the first step, information has been carefully edited and incomplete answer has been evaluated according to the research theme. These collected information have been analyzed according to the objectives of research. Meaningful tables have been generated from the process of data. The data has been analyzed by using both qualitative and quantitative techniques. In this process description, explanation and generalization have been made using statistical tests. Similarly, discourse and statistical analysis has been linked with tables and figures have been used to achieve the objectives. For this study, the descriptive statistics: frequency, correlation, and cross-tabulation were used in

All rights reserved by *www.ijsrd.com*

order to meet the research objectives and thereby to answer the defined research.

B. Data Exploration:

The data thus collected was collected was tabulated, interpreted & analyzed with a view to make the study meaningful. In the present study, percentage, a frequency & cross tabulation method has been used for analysis. For this study, the descriptive statistics: frequency, correlation, and cross-tabulation were used in order to meet the research objectives and thereby to answer the defined research.

VI. ANALYSIS OF DATA

A. Explanation of factors and sample data

	0	1	2	3	4	5	6
id	110965	112858	236547	21564	32154	87546	21698
age_in_days	12500	22210	13221	15621	18961	13002	9554
income	65475	69974	265470	225661	456321	789650	775620
perc_premium_paid_by_cash_credit	94.52%	32.54%	32.21%	54.71%	32.14%	100.00%	23.11%
Count_3-6_months_late	0	3	2	1	7	6	0
Count_6-12_months_late	3	0	2	0	1	0	3
Count_more_than_12_months_late	0	1	0	0	0	0	0
no_of_premiums_paid	4	12	11	5	17	1	10
target	1	1	1	1	1	1	1

Table 1:

Table 1 depicts the sample data of our complete dataset. In the table we have taken sample dataset of all the factors on which our research is done. These are the top 6 rows of all the factors. We have considered our target as the required result. When the target is 0 it means customer will not pay the next premium and when the target is 1, it means customer is likely to pay his/her next premium. Thing that is to be noted here we haven't considered the health issues and particular gender of any individual. The complete calculations and observations have been done only on this dataset (complete one).

	count	mean	std	min	25%	50%	75%	max
id	500.0000	129532.8240	150397.8095	154.0000	48484.5000	87546.5000	163882.0000	1663646.0000
age_in_days	500.0000	18490.5480	5252.5236	1764.0000	14961.2500	17739.0000	23000.0000	29998.0000
income	500.0000	359176.8240	1312122.4551	16541.0000	364673.0000	780221.0000	859430.0000	3745630.0000
perc_premium_paid_by_cash_credit	500.0000	0.6219	0.3724	0.0000	0.3213	0.9600	1.0000	1.0000
Count_3-6_months_late	500.0000	2.5940	2.3694	0.0000	0.0000	2.0000	4.0000	9.0000
Count_6-12_months_late	500.0000	2.0540	1.8356	0.0000	0.0000	2.0000	4.0000	8.0000
Count_more_than_12_months_late	500.0000	0.8620	1.0813	0.0000	0.0000	0.0000	1.0000	5.0000
no_of_premiums_paid	500.0000	11.1420	4.0602	1.0000	8.0000	11.0000	14.0000	40.0000

Table 2 depicts various values as explained. Count is the total number of values i.e. 500 are taken into consideration for the observation. Mean of all the factors is calculated to find some particular average values of all the factors. And the minimum and maximum values exist in the dataset. Moreover, we also calculated the quartile deviation at 25,50 & 75 percentile. These all values are calculated to find the basic scenario of the data. For instance, we can observe that age of customers here has the average value of 18490.54 i.e. of 50 years. Maximum age of the customer is 82 years. Now based on his age various things will be considered and various observations has been done. Based on all this data, the part of research that this individual is likely to pay the premium or not is being done.

B. Explanation of range of customers who paid the percentage premium amount by cash credit or full payment at once

S. No	% premium paid by customers by cash	Frequency
1.	0-10	60
2.	10-20	40
3.	20-30	25
4.	30-40	60
5.	40-50	20
6.	50-60	10
7.	60-70	40
8.	70-80	20
9.	80-90	25
10.	90-100	200

Table 3:

This frequency table 3 shows that the maximum number of customers have paid the full amount in cash. More precisely, 200 customers are the one who prefers to pay the 100% amount by cash credit or at once. Even there are 100 customers are from the category who paid less than 20% amount in cash credit. This data concluded that company have mor of the cash flow so that they can pay the amount to another customer if someone claims. Cash flow is manageable by the company. It shows a positive sign for company's requirements.

C. Explanation of income of customers

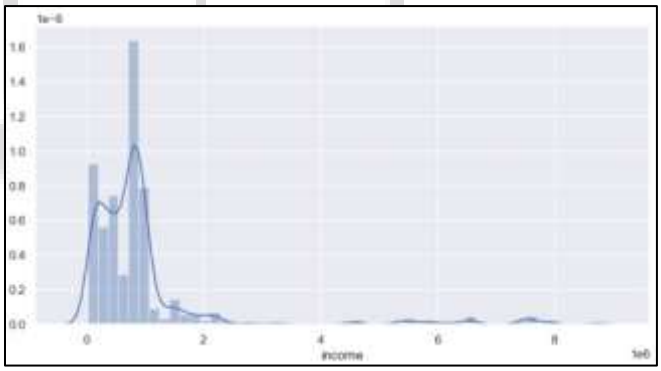

Fig. 1:

In the above graph i.e. figure 1, we have observed that income of most of the customers is within the amount of 10 lakh. Some of the customers lies in the category of outliers where it can be observed that the value of these outliers is much higher than the normal range. These outlier values are the reason that some of the values lies in the area of 80 lakh. These uneven values in the income can lead to misinterpretation of the data. Data analysis can't be more accurate with these outlier values. Exact values of these outliers and how they can lead to misinterpretation is shown in the table 4.

All rights reserved by *www.ijsrd.com*

count	500.0000
mean	950179.8340
std	1312123.4551
min	16541.0000
0%	16541.0000
10%	147885.3000
20%	246688.0000
30%	456321.0000
40%	549068.8000
50%	756321.0000
60%	789779.8000
70%	874563.9000
80%	987168.2000
90%	1458596.6000
99%	7541331.9900
99.9%	8321925.1080
100%	8745630.0000
max	8745630.0000

Table 4:

In table 4, we have observed that in total 500 entries we have taken into consideration the mean is 950179, whereas most of 90% values lies in the range of 10 lakhs. This is because of the outlier values. From the above table we can observe that 80% of the values lies within the 987168. 90% lies within 15 lakhs approx. From the above table we can conclude that the most of the customers company have had the income within 15 lakhs approximately. Company have to set their policies and target accordingly. It isn't necessary that the customer having higher income will pay the next premium definitely. Income can be a factor but it isn't the most important factor.

So far, we have considered age and income of the customers as a crucial factor but it simply doesn't provide us the required results. Moreover, percentage of premium paid by cash by the customers can also show the number of customers who contribute in the cash flow of the company but they will pay their next premium or not is not commendable. So, for that we need some important correlations between the factors so that we can find some important research from these factors. These correlations must be between target values and other factors as target is the ultimate factor which we need to examine. Dependencies of target values on other factors can conclude our research proficiently. To find these dependencies we prepare a Heatmap of the correlation between other factors with the target values.

VII. HEATMAP AND CONCLUSIONS

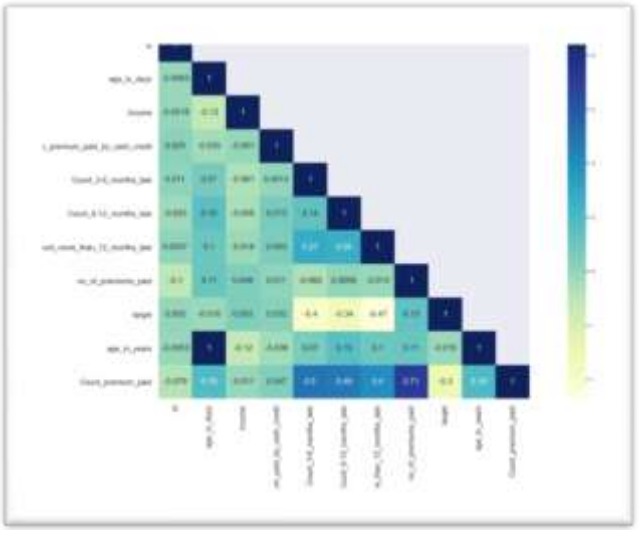

Fig. 2:

From figure 2 we can conclude the following points:

1) ID of the customers is nearly 0.002 correlated
2) i.e. almost 0. So, customers having different id numbers is not going to affect the target of the company.
3) Age in days is negatively (-0.016) correlated with target. It implies that there are less chances that a customer with higher age will pay the premium on time. But the correlation is not so strong so this factor affects the company but not to that much extent.
4) Similarly, we have converted the age in years from age in days. Therefore, correlation between these two with respect to target value remains the same.
5) Income of the customers is positively related
6) w.r.t. target. Customers with income relatively higher than others are more likely to pay the premium. Correlation is somewhat better than above two factors so this will bother the insurance company.
7) Premium paid by cash credit is positively correlated (0.032) w.r.t. target value. It simply implies that company is not bothered by much extent as this only help them to improve the cash flow for the business.
8) Count 3-6 months is strongly negatively correlated (-0.4) w.r.t. target. This factor plays important role as customers who haven"t paid their premium on time in past have the image that they will not pay the premium on time or not at all in future.
9) In the same manner, count 6-12 months late and count more than 12 months late have the correlation -0.34 and -0.47 respectively w.r.t. target values. These factors can give the prediction of customers who will not pay the premium on time in near future.
10) Count premium paid is the overall summation of all three factors of paying the premium late. It can be observed it is negatively strongly correlated (-0.3) w.r.t the target values. Similarly, it will also affect the nature of the customers for paying the next premium on time.

All rights reserved by *www.ijsrd.com*

ACKNOWLEDGMENT

I thank my Mentor Miss Karuna Midha for analyzing the data and for advising on all aspects related to this paper, along with other concepts too. Later for validating the experimental result and reviewed the paper.

REFERENCES

[1] Yusuf, H. O., Gbadamosi, A. &Hamadu, D. (2009). Attitudes of Nigerians towards insurance services: An empirical study. African Journal of Accounting, Economics, Finance and Banking Research, 4 (4), 334-346.
[2] Viswanadham, P. (2005). Claims settlement operations performance evaluation of LIC of India, The Indian Journal of Commerce, Vol. 58 (2), 80-90.
[3] Yadav, R. K. &Mohania, S. (2014). Claim settlement process of life insurance services – A case study of ICICI prudential life insurance Application of Cost Reduction … 44 company, International Letters of Social and Humanistic Sciences, 24, 26-32.
[4] Kalani, M., Salunkhe, H. A., &Ahirrao, M. B. (2013). Comparative study of claim settlement ratio of LIC with other insurance companies in India, Indian Journal of Applied Research, 3 (5), 389-391.
[5] Bates, I. & Atkins, B. (2007). Management of Insurance Operations. London, Global Professional Publishing.
[6] Brear, S. (2004). Chartered insurance institute (CII) course book, UK, Personal lines Insurance, CII learning solutions, pp 14/9-14/17. Butler, S. & Francis, P. (2010). Cutting the Cost of Insurance Claims, taking control of the process. Booz & Co.
[7] Namasivayam, N., S. Ganesan and S. Rajendran (August, 2006), "Socio economic Factors Influencing the Decision in Taking Life Insurance Policies", Insurance chronicle (The ICFAI University Press), page 65-70
[8] Rao, Bn. Vankateswara (June, 2006), "LIC- New Business Lacks Vigor", Insurance chronicle (The ICFAI University Press), page33-40
[9] Ms Sunayna Khurana, Lecturer, "Customer Preferences in Life Insurance Industry in India", ICFAI National College.

All rights reserved by *www.ijsrd.com*