**A Study on The Probability That A Customer Will Renew Their Insurance**

**Synopsis**

Ujjwal Agarwal -190921

Dhruv Singhal -190294

Harsh Kumar Gautam -190361

Mahaveer Bana -190467

Anay Nagar -190130

Dheeraj Bana -190286

Kapil -190415

Nitin Jangir -190568

Pavan Meena -190595

Ashok Kumar Jyani -190194

**Abstract**

In this paper, we analyze the probability of a customer defaulting on their insurance premium payment. Insurance is a mean of protection from sudden unforeseen financial loss and risk profiling is an important aspect of it. Keeping customers satisfied and being able to custom target various demographics is crucial to maintain and increase customers.The dataset consists of 17 parameters for 79853 customer observations with a combination of Indicator and continuous variables. Data mainly covers customers’ demographic information, premium payment related behavior and risk profile information. Random tree and Logistic regression have been used for modelling. Their performance measures have then been compared to decide which is better suited for the dataset and hence used to draw meaningful conclusions. We are building a model to predict the probability that a customer will default the premium payment and hence in our analysis ‘*renewal’* would be the target or the response variable i.e. the dependent variable and other variables would be independent or the predictor variables. Through this study we will try to identify key factors that influence the timely payment of premium by customers.

**Introduction**

Insurance is a risk management tool which reduces the impact or consequences of an uncertain loss. The entity offering protection against the risk is called ‘Insurer’ and the customer taking the protection is called ‘Insured’. For purchasing the protection (‘Insurance’) the Insured has to pay a premium termed as ‘Insurance premium’ or simply ‘premium’ which is usually periodic in nature. While the Insured looks at insurance as a safety tool, insurers earn profit by spreading the risk across large number of insured parties with varied risk profile.

The insured has the right to claim compensation under the Insurance policy as long as the premiums are duly paid. However, if the premium is not paid even after the due date or in other words the Insured default the premium payment, the Insurance policy gets lapsed and any further claim cannot be made under the Insurance policy.

At present, Life Insurance, Health Insurance, Vehicle Insurance and General Insurance (Non-Life) are the commonly used risk management tools in Indian market. Insurers are governed by autonomous regulatory bodies (Insurance Regulatory and Development Authority in India) to protect the interest of the Insured and to prevent mis-selling and other unfair practices.

Developed countries have significantly higher insurance penetration but developing countries are also catching up with increasing awareness, income levels and affordable costs of purchasing insurance.

From a commercial point of view, premium paid by the customer is the major revenue source for Insurer. Default in premium payments results in significant revenue losses as and hence Insurer would like to know upfront which type of customers are likely to default the premium payments.

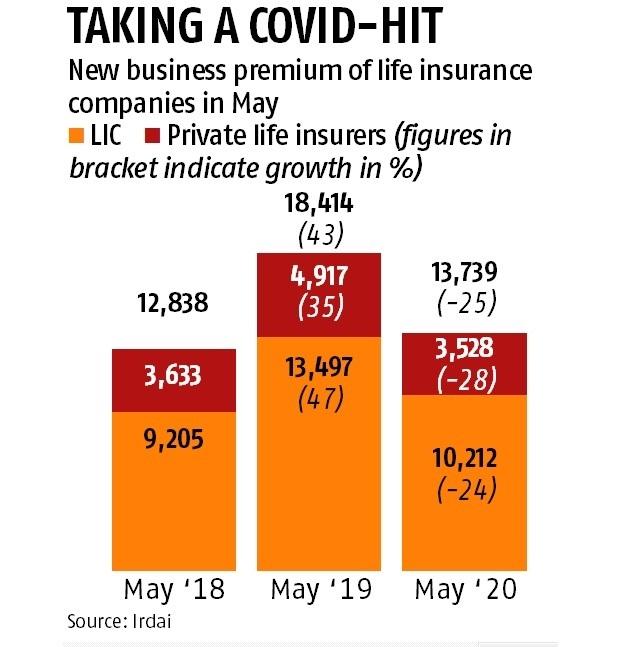
There are numerous type of factors that have an effect on insurance renewal like psychographic, socio-economic factors such as financial inclusion(taking a loan from a bank or getting a bank account), demographic factors such as gender or education levels of the household head etc.

**Effect on insurance due to COVID-19**

Due to the outbreak of coronavirus, an increasing number of people have become more aware of insurance. Many of them consider insurance as a necessity to be prepared for any unforeseen circumstance in the future. Prior to the spread of COVID-19 in India, only about 10% of people showed an interest in buying insurance to cover medical emergencies including pandemics and infectious diseases. Now, however 71% of people consider it as a necessary safety tool.

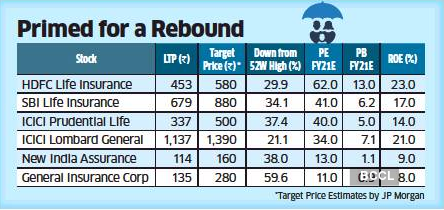
But, due to the pandemic most people have lost their jobs and many middleclass people were unable to pay their premium. As a result, they couldn’t renew the premium and they had to discontinue their policies which had a large negative impact on companies as well.

“The disruption caused by the coronavirus’ spread and the pandemic-induced lockdown resulted in Life insurance industry losing around four million policies and premiums of around Rs 45,000 crore”, said Raj Kumar, managing director of LIC.



Life insurers faced high mortality claims due to COVID as a loss to insurance companies and also of low interest rates. But after the Lockdown phase, it seems to have got back on track from July as premium collection has turned positive, boosted by increased interest in insurance.

“As of August 31, LIC’s premium income is back to previous year’s level. So, we have covered the gap that was there in April, May when the premium went down by about 32 per cent,” said Kumar, managing director of LIC.



**Literature Review-**

* **Fuzzy-probabilistic multi agent system for breast cancer risk assessment and insurance premium assignment-(Farzaneh Tatari, Mohammad-R. Akbarzadeh-T, Ahmad Sabahi) (2012)**

In their study, they have done risk assessment on the development of breast cancer. According to studies one in nine women will develop breast cancer at some point in her life. As it has been found that not all factors equally increase the chance of breast cancer development, the risk factors have been categorized into three groups of strong, moderate and minor risk factors. Factors such as age, family history, alcohol consumption etc. have been considered. Due to high imprecision and linguistic form of information it is difficult to analyze data and therefore they have used fuzzy-logic based analysis to handle the existing uncertainty. Based on this analysis, insurance companies can offer different premiums for different customer segments. They had concluded that carriers of certain kind of genes was a critical factor for deciding the risk of development of breast cancer.

* **Risk Prediction in Life Insurance- (Noorhannah Boodhun & Manoj Jayabalan) (2018)**

Their study was done to find and examine the factors that have an effect on insurance purchasing decision of customers. They developed a model from data of ~60,000 applicants having 21 features which describe the characteristics and nature of applicants like BMI, age, weight, family history, etc. They used algorithms like linear regression, random tree, REP Tree and Artificial Neural Network in their study to predict risk involved in life insurance policies. This model provided a very efficient method for life insurance business to classify the applicants for life insurance which is otherwise a very slow process that leads to people switching to different policies or not buying at all. Also, with the increase of data availability, automation of the process is necessary which is provided by the model for business advancement. It was found that REP Tree algorithm outperforms others with lowest MAE and RMSE of 1.5285 and 2.027 with CFS method, whereas linear regression with PCA method showed very satisfactory values of MAE and RMSE of 1.6396 and 2.0659 respectively.

* **A Study on Factors Affecting Customers Investment Towards Life Insurance Policies- (Babita Yadav, Anshuja Tiwari) (2012)**

They divided the applicants based on features like age, income, occupation etc. to analyze their behavior and its consequences on Insurance Sector. The main motive behind the study was that almost 70% of people’s lives are still uninsured and to give boost to the business development in this domain. They used various popular statistical algorithms like chi-square and correlation to analyze and to identify the most important factors for their hypothesis. Results of the study showed that people between the age of 30-40 are more likely to buy insurance. Also, company reputation, money back guarantee, low premium and kind of risk coverage attracted customers

* **Life Insurance Industry of India – Past, Present & Future (A Study of LIC of India-Shilpa Agarwal and A.K Mishra)**

This study was based on examining the status of LIC in pre and post liberalized era (LPG- Liberalization privatization and globalization in year 1991) as well as estimating the future trend in LIC business. LIC was formed in 1956 and became a mammoth in insurance industry. The data used was from the LIC website. They used Method Of least squares for examining the future trend of their business. Based on the year 2009, they calculated the trend value for year 2020 and it showed the business of LIC is in increasing trend. Till 2013 there were 52 insurance companies operating in India, of which 24 are in life insurance business and having a total share of 80.2%.

Some questions answered by previous researchers are:

* **Does age of the customer affect the chance of continual of premium payment or renewal of insurance?**
* Previous works have shown and suggested that increase in age leads to an increase in the renewal of insurance.
* **Does gender of the customer have an effect on insurance renewal or purchase?**
* Studies have suggested that females are less likely to buy or renew insurance than males.
* **Does type of area in which one lives affect the probability of insurance purchase?**
* Due to higher awareness and understanding of financial tools and easier access people living in urban areas have a much higher chance of insurance renewal as compared to those who live in rural areas.

**Objectives**

The aim of this study is to understand the premium payment pattern of customers of an Insurance company. For the study, we have customer data available primarily covering:

1. Customer demographic information e.g. Age, Income, Marital Status, residence area type etc.
2. Insurance policy and premium payment related information e.g. premium, renewal, sourcing channel etc.
3. Customer risk profile (risk score)

The objective is to predict the probability that a customer will default on premium payment, so that insurance agents can proactively reach out to the policy holders to follow up for the payment of premium. Simultaneously, it will also help understand customer demographics which are more likely to default and to price the premium amount in accordance to the same.

**Data Dictionary**

The dataset has 79853 records with total 17 different variables. The target or the dependent variable in the given dataset is “renewal”, which has values as 0 or 1. “0” indicates that customer has not renewed the premium and “1” indicates that customer has renewed the premium. The data is based on life insurance and has been collected between 2017 to 2019.

Below is the list of variables along with the description and categorization:

| Variables | Description | Type |
| --- | --- | --- |
| Id | Unique customer ID | Continuous |
| perc\_premium\_paid\_by\_cash\_credit | % of the premium paid by cash payments | Continuous |
| age\_in\_days | Age of the customer in days | Continuous |
| Income | Income of the customer | Continuous |
| Count\_3-6\_months\_late | Number of times premium was paid 3-6 months late | Continuous |
| Count\_6-12\_months\_late | Number of times premium was paid 6-12 months late | Continuous |
| Count\_more\_than\_12\_months\_late | Number of times premium was paid more than 12 months late | Continuous |
| Marital Status | 0 indicates that customer is Unmarried and  1 indicates that customer is Married | Indicator |
| Veh\_owned | Number of vehicles owned (1-3) | Indicator |
| No\_of\_dep | Number of dependents in the family on the customer(1-4) | Indicator |
| Accomodation: | 0 indicates that customer has rented the accommodation and 1 indicates that customer has owned the accommodation | Indicator |
| Risk\_score | Risk score of customer | Continuous |
| no\_of\_premiums\_paid | Number of premiums paid till date | Continuous |
| sourcing\_channel | Channel through which customer was sourced (A/B/C/D/E) | Indicator |
| residence\_area\_type | Residence type of the customer (Rural/Urban) | Indicator |
| premium | Premium amount | Continuous |
| renewal | 0 indicates that customer has not renewed the premium and 1 indicates that customer has renewed the premium | Indicator |

**Data Source**

The data used in this study has been collected from secondary sources. The data has been provided by The University of Texas at Austin (UT Austin) and Great lakes Institute of Management, Chennai.

**Methodology and Modelling**

In this study we will first do data pre-processing, exploratory data analysis, data normalization and outlier treatment. Data variables are normalized to avoid any one variable overshadowing the model and to make sure data remains uniform. This will help in getting a brief idea on the data we are working with and we will also be able to check for any missing values or exceptions. Also, we will be able to check the presence of dependency and correlation among variables. For the modelling part, we will use logistic regression and random forest and then we will compare the model performance measures to find out which is better suited for this study to arrive at the right conclusions.

**Logistic Regression**

Logistic regression is a predictive analysis technique. It is used to predict a classification problem. It is easier to implement and makes no assumptions about the distribution but in this model multi-collinearity among variables can affect the outcome. The steps of this model are:

* Analyze the base data vis-à-vis the modified data and test if the modification is adding value to the model.
  + Base data has individual columns for count for late payment (3\_6 months, 6\_12 months, >12 months) and modified data has single aggregated column for count of late payments.
* Run Logistic Regression function on data and observe the significant variables
* Re-Run Logistic Regression function with significant variables
* Build the prediction model
* Use test data to analyze the model

Some equations that are used are:

* Activation Function -Sigmoid function= sig(x)=1/(1+e-x)
* Cost/Error function=−​1\*​∑ ​​y\*logy’​+(1−y​)\*log(1−y’​) where y and y’ values are taken for limits i=1 to i=output size. Here y’ is the scalar value in the model output and y is the corresponding target value for all i’s*.*

Reason for the choice of this cost function is that it is differentiable which is needed for almost all the optimizers like gradient descent to optimize the weights which restricts us from using discrete values error function like ​∑ (yp - ya) where yp is the predicted class and ya is the actual class as it can’t be optimized. Moreover ‘-ve‘ sign is to maximize the probability by minimizing the los function . Decreasing the cost will increase the maximum likelihood.

**Random Forest**

Decision Trees generally tend to overfit data and are usually very sensitive to change in data. Random Forest is a technique used for better accuracy. It randomly selects observations and specific features to build multiple decision trees and then average the results (i.e. means for a continuous variable) or use the most popular prediction (i.e. modes for a classification variables) across all the trees for a robust prediction. The general procedure of using multiple trees to obtain better performance is called ensemble learning. The steps for this model are:

* Generate random forest using ‘renewal’ as dependent variable and others as independent variable
* Predict values and assess model performance

**Hypothesis**

From the knowledge of previous studies and some market trends we expect the following effects that variables will have on insurance renewal.

* Age will have a positive effect and will be a significant factor- It is a common trend and

has been observed by various studies that as a customer gets older, they are much more likely to pay their insurance premiums due to various reasons such as higher health risks, more financial stability or larger number of dependents in the family.

* Income will have a significant and positive impact-With larger incomes people are more likely to be able to pay their premiums. They are financially secure and are also looking for opportunities to use their money to protect their family and their own interests.
* Count of late payments will have a negative impact on the probability of renewal- A customer can be late on their premium payments for various reasons such as loss of feel for the need of insurance, unable to pay premiums due to sudden unavoidable expenditures or unemployment etc. Increase in the number of late payments is indicative of the fact that due to some reason the customer is unable to pay their premiums and is thus more likely to not renew their insurance.
* Marital status will have a positive impact but not a significant one- Along with the increased chance of higher dependents in the family there is a sense of commitment and want of protecting their loved ones and studies have suggested that a married person is more likely to buy or renew insurance than an unmarried person.
* Vehicles owned will not be a significant factor- The number of vehicles owned can be a very deceptive for companies trying to figure out potential customers who might renew. It is very possible that a customer with a good or high income has no vehicles due to personal reasons but has a good chance to renew their insurance and a customer has multiple vehicles but has a low chance to renew their insurance. There is no general trend with the effect of vehicles owns and hence it is highly difficult and unreliable to use it as predictive measure.
* Number of dependents will have a positive effect but not a significant one as again it is a difficult measure to use to draw conclusions due to multiple different possibilities in similar situations.
* Risk score generally has a positive impact on the purchase of insurance i.e people at higher risks are more likely to buy insurance however the given data does not clarify the method of its calculation and its impact on credit worthiness of a customer
* Number of premiums paid and premium amount will have a positive and significant effect as they both generally indicate how committed a customer is in renewing their insurance.
* Percentage of premium paid in cash will have a negative effect-Paying high percentage in cash is generally indicative of the customer not having a long term commitment towards insurance or running low on funds and soon might not be able to pay the premium.
* Residence area type will have a negative impact-As insurance awareness and financial stability is lesser in rural areas compared to urban areas it is generally observed that customers living in urban areas are more likely to purchase/renew insurance.
* Accommodation will not have a significant impact-It is difficult to decide if the impact is positive/negative as whether a customer has rented or owned the accommodation often has very less to do with insurance renewal decision.

**Based on previous studies and observations , we are suggesting below correlation: Renewal=f(Age(+ve), Income(+ve), Count of late payments(-ve), Marital Status(+ve), Number of dependents(+ve), Risk score(+ve), Percentage of premium paid in cash(-ve), Number of premiums paid (+ve), Premium(+ve), Residence area type(-ve)).**

**References**

* **Base paper- Ma, Y. (2020) Prediction of Default Probability of Credit-Card Bills. *Open Journal of Business and Management*, 8, 231-244**
* [**https://www.scirp.org/journal/paperinformation.aspx?paperid=97459**](https://www.scirp.org/journal/paperinformation.aspx?paperid=97459)
* **International Journal of Marketing, Financial Services & Management Research Vol.1 Issue 7, July 2012, ISSN 2277 3622**
* [**http://indianresearchjournals.com/pdf/ijmfsmr/2012/july/9.pdf**](http://indianresearchjournals.com/pdf/ijmfsmr/2012/july/9.pdf)
* [**https://www.sciencedirect.com/science/article/pii/S1532046412000883**](https://www.sciencedirect.com/science/article/pii/S1532046412000883)
* **Boodhun, N., Jayabalan, M. Risk prediction in life insurance industry using supervised learning algorithms. *Complex Intell. Syst.* 4, 145–154 (2018)**
* [**https://link.springer.com/article/10.1007/s40747-018-0072-1#citeas**](https://link.springer.com/article/10.1007/s40747-018-0072-1#citeas)
* **Johnson, Eric J., John Hershey, Jacqueline Meszaros, and Howard Kunreuther. "Framing, probability distortions, and insurance decisions." Journal of risk and uncertainty 7, no. 1 (1993): 35-51.**
* **Neter, John, and C. Arthur Williams Jr. "Acceptability of three normative methods in insurance decision making." Journal of Risk and Insurance (1971): 385-408.**
* **Beck, Thorsten, and Ian Webb. "Economic, demographic, and institutional determinants of life insurance consumption across countries." The World Bank Economic Review 17, no. 1 (2003): 51-88.**