Regression (Random Forest Regressor algorithm) Based Prediction App for Health Insurance Premium

Sindhiyadevi. T

Abstract:

Machine learning or AI widely used in many industries to see progress in business, revenue and automate their process, especially to predict the future step in business to avoid risk and loss. Now a days in every industry ML playing vital role, Finance, Image processing, Speech recognition, medical industry, Personal health fitness, algorithms to regulate business structure to get more revenue by targeting customer expectations. This paper represents a machine learning-based health insurance prediction system. This digital health insurance is a time-consuming process for policy holder and the insurer. This technology helps insurer to provide quick service to their customer. Meanwhile ML based digital insurance provide clients with accurate, quick, and effective health insurance coverage.

1. Problem Statement:

In India people have less aware about the medical insurance. Now a days most of the employers covering medical insurance of employee in concern to their family. But what about the status of remaining people? In comparison to other nations, India's government allocates only 1.5% of its annual GDP to public healthcare. On the other hand, over the past 20 years, worldwide public health spending has nearly doubled along with inflation, reaching US $8.5 trillion in 2019, or 9.8% of global GDP. In United States and many other countries health insurance is compulsory for every family, and privates are covering Dental insurance. India and many countries had massive impact during covid, many people could not afford to the medical expenses in India. COVID-19 accelerated the adoption of health insurance as mounting healthcare bills and the impact of prolonged lockdowns on the economy forced people to opt for insurance plans.

Consumers are now seeking an integrated and comprehensive health insurance and they are confused with companies and policies and their procedures, how to rise. Traditional insurance methods take massive amount of time to process with data and calculate the premium, it might exist with inaccurate premium amount, inconsistence with data and leads mis communication between insurers and insurance holder and cause distrust on the company name.

So, its mandatory for the insurance company to go in a digital way to connect with people easily and give solution in a quicker and effective way.

2. Market/Customer/Business Need Assessment

2.1 Customer needs

People are running to secure their life in all the way, usage of internet and smartphones made human life easier giving solutions in hand just like our neighbor. Chatbot is one of the great examples of AI giving handy tips and solutions. People want medical assistance to their parents and secured future to their children. In the competitive world they are looking for best options comparing their benefits with other better benefit in every way.

In every industry competition exists. Competition ensures the provision of better products and services to satisfy the needs of customers. It forces every company to focus and work on their quality and fast service to the customer.

2.2 Competition in Health Insurance

Traditional competition in health care involves one or more elements (e.g., price, quality, convenience, and superior products or services); however, competition can also be based on new technology and innovation. People are consulting doctor, physicians, and health insurance companies through chat by sitting in home. After covid the importance of technology become high.

In the modern world people are expecting quicker reply and answers from their insurance company about their premium, no. of years. They are expecting faster claim settlement. This is the main concern every policy holder has. Processing data in manual way takes longer period and increase the need of customer assistance to claim money and exits with customer dissatisfaction. Customers will also be urged to select a plan that meets their requirements rather than paying for services they may not use.

2.3 Solution and Assessment

The proposed AI app mitigate all the above problems and allow the insurance company to help their insurance holder to get all the information regarding their data, premium and new updates at any time and could maintain a strong relationship with their customer.

3. Target Specifications and Characterization:

There are plenty of health Insurance companies exits and many of them are successful, they are using the technology in a right way and being successful. Every startup or small health insurance could benefit through this AI based premium prediction app and can reach their customer easily.

This AI based premium amount predictive app helps insurer to provide the premium amount in a matter of second of receiving the customer information and can maintain user friendly relationship.

This app will be a handy, provides premium amount details regardless of senior citizen, above/below 40 just by entering the customer information.

4. External Search

Reference

[1] Keshav Kaushik 1, Akashdeep Bhardwaj 1, Ashutosh Dhar Dwivedi 2, * and Rajani Singh 2 Machine Learning-Based Regression Framework to PredictHealth Insurance PremiumsK *Environmental Journal of environmental Research and Public health*

[2] Mukund Kulkarni, Dhammadeep D. Meshram, Bhagyesh Patil, Rahul More, Mridul Sharma, Pravin Patange Medical Insurance Cost Prediction using Machine Learning *International Journal for Research in Applied Science and Engineering Technology*

[3] Aman Kharwal, Health Insurance Premium Prediction with Machine Learning, *The Clever Programmer*

5. Bench marking alternate products:

There are many data factors associated with premium prediction. Though the premiums are predicted, best algorithm will be the one who give the best fit line. The result will be compared to its actual value to find the accuracy of the output. Many regression algorithms are exits to predict the heath insurance premium. Some of the algorithms are Linear regression, Multiple linear regression, Gradient Boosting algorithms and decision trees.

The Main disadvantage of linear regression is (widely used method) it assumes the relationship between dependent and independent variable and fails to fit the complex dataset properly.

6. Applicable Regulations (Government and Environmental):

To develop this app, we will be using python open-source platform. While developing the app we will be using customers/policy holder data in a large amount, so its important to maintain data protection regulations.

7. Applicable Constraints

Below are the steps involving in developing the health insurance premium calculator.

- Collect and clean the data.
- Preparing data make it to understand by a Machine learning system.
- Feature engineering – which is a technique to create an additional feature adding two or more features from the raw data who miss reveal all the facts about the targeted label.4.
- Hiring AI and Machine learning service provider

Machine learning is the enabler technology, it is important to follow the proper step and execution for training and learning of models on algorithms, otherwise it will not work. Hence, it is mandatory to hire AI and Machine learning service providers and focus on their core competency.

8. Business Model (Monetization Idea)

This premium amount prediction app is very useful one to all the insurance company who expects to keep connection with the public with their updates.

We can provide this application to the insurance company in a subscription model and update the data and feature factors based on every company requirement. Meanwhile we can sell the complete app to the insurance company with feature to update the data set from the insurance company end and we can provide the lifetime service of maintenance. In future we are come up with any updated algorithm and to increase the quality we can ask the company to update their app with the updating fees.

We can even make a free public website to check the premium amount of a person/ family. In that we can keep some of the insurance company in the company option, so that the insurance company can join to this website with monthly subscription. If the Insurance company wish to continue in the list of company, they must pay every month.

The above feature helps insurance companies to get familiar with public. The same application method we can use for all the insurance like vehicle, life, Term insurance by training the dataset.

9. Concept Generation

We are using Python to develop this application.

The premium of a health insurance depends on various factors, which we call feature vectors, the predicting premium amount highly dependent with these factors. The feature factors are independent vectors which differ from each company to company. Some of the basic feature vectors are used by the companies, which are going to use to implement our application.

Name	Description
Age	Age of the client
BMI	Body Mass index
No of Children	Number of children the policy holder has
Gender	Male/Female
Smoker	Whether a client is a smoker or not
Region	Location of the client
Charges of the insurance premium	Medical cost the client pay

9.1 Import Libraries

To process these data's, we need to import some of the python libraries.

```
import numpy as np
import pandas as pd
data = pd.read_csv("Health_insurance.csv")
data.head()
```

9.2 Display of sample dataset

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

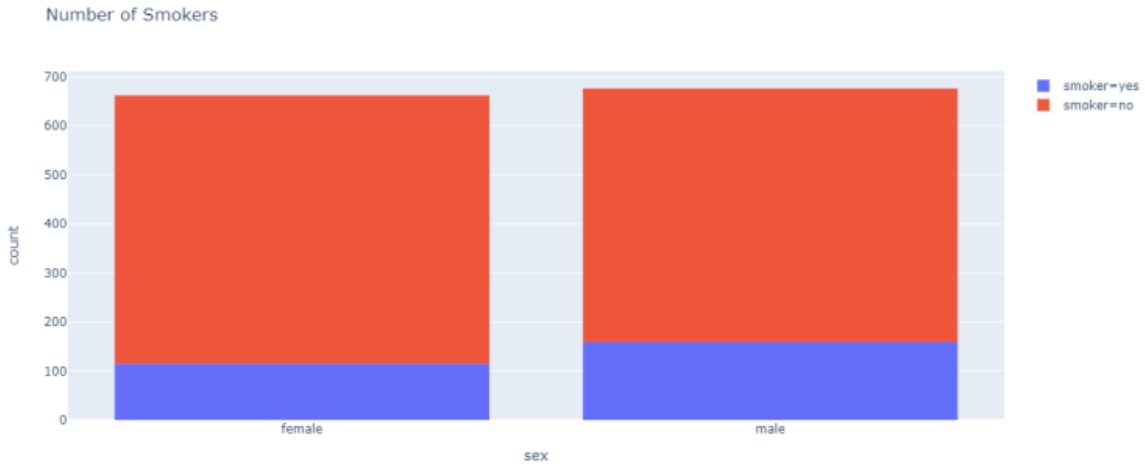
9.3 Sum of null value the dataset contains.

The step to check whether the dataset contains any null values very important, because each factor impact in premium amount.

Age	0
Sex	0
BMI	0
Children	0
Smoker	0
Region	0
Charges	0

9.4 Is the person, Smoker?

This is the most important feature vector to check, the person who smokes may likely to have many health problems exits or in future compared to the person who doesn't smoke. The histogram chart visualization for smokers in the dataset.



9.5 Health Insurance Premium Prediction Model

Next step is training the machine learning using Random Forest Regressor method to predict the health insurance premium.

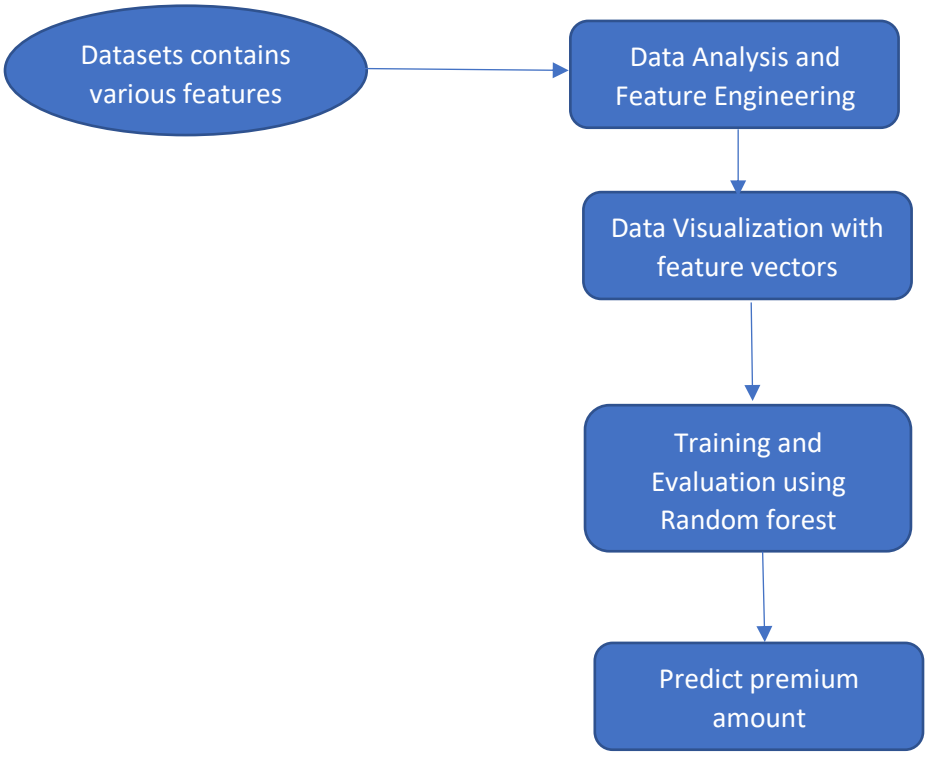
Split the dataset into training and test set, 80 percent of data will be trained and remaining 20 percent will be used for test.

```
from sklearn.model_selection import train_test_splitfrom sklearn.ensemble import RandomForestRegressor

xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size=0.2, random_state =42

forest = RandomForestRegressor()

forest.fit(xtrain, ytrain) forest = RandomForestRegressor()
```

10. Final Report Prototype

The objective of predicting insurance cost depends on various features mainly age, BMI, child number, the region of the person living, sex, and whether a client is smoker or not. These features contribute to our target variable prediction of insurance costs.

For evaluation datasets will be split into two parts training set and test. 80 percent of data will be trained to get more accurate result and 20 percent will be used for test.



10.1 Random Forest Regressor Algorithm

The following calculation will be done on the evaluation to test the predicted value accuracy, The Random Forest Regressor will give the best predicted value comparing to the Liner, Multiple and Decision Tree regression type.

10.1.1 The MAE (Minimum mean Absolute error)

The Mean Absolute Error (MAE) is the difference between the original and forecast values obtained by averaging the absolute difference over the data set.

$$\text{MAE} = \frac{1}{N}\sum_{n-1}^{N}\left|y^{\Delta} - Y\right|$$

10.1.2 RMSE (Root Mean Square Error)

The RMSE of the disparity between the expected values and the real values is determined as the square root. For an accurate forecast, the RMSE must be low so there would be less variance among the expected values and the real values.

RMSE = Where N = Number of overall observations, = expected insurance fee values, y = real insurance fee values.

$$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(Y^{\Delta} - Y)^2}$$

10.1.3 R-Squared

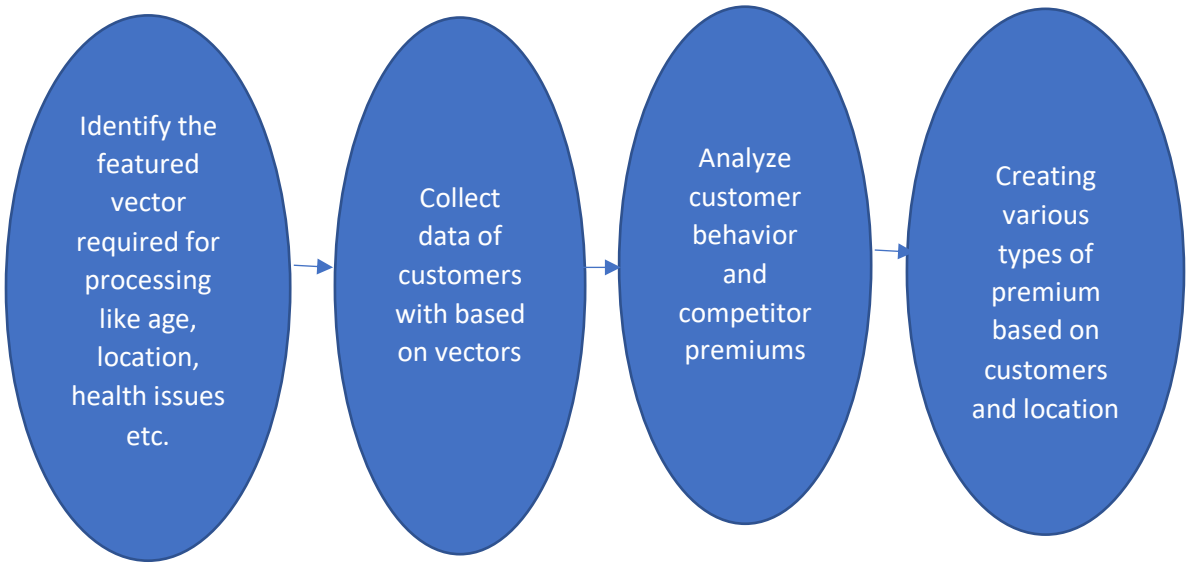
The more R-squared, the better the model output., and indicates that the model deviates less from real values. A R-squared score of 1 indicates that it suits perfectly.

$$R - Squared = \frac{Explained\ Variance}{Total\ Variance}$$

11.Conclusion

As now a days heath insurance plays a vital role in industry, AI based app helps analyze and evaluating large volumes of data to streamline and simplify health insurance operations and helps to retain their place in the market. This great idea helps both health insurance and policyholder to save the time and provide the accurate data. The single app can perform repeated actions with data set updates. The machine learning process works on historical data and address broad array of applications and systems. The app considerably reduces the amount of time individual invests in policymaking. It helps insurance company to give their attention more on customer.

Step3: Business Model:



1.Identify the feature vectors:

The first step is to identify the feature vectors involves calculating the premium amount, common feature vectors for health insurance calculation are age, gender, number of children, body mass index, location, health problems, smoker or not. Before enrolling the consumer, it is mandatory to take complete body checkup for the customer, it will avoid some future problems the insurance company will face.

2.Data collection:

We need to make a checklist based on the identified feature vectors. It is mandatory to fill those from the customer. Based on the feature vectors data of consumers needs to be collected.

3.Customer Segmentation:

Need to segment customers, based on the collected data there are many segmentations available to segment the customer. K-means clustering is one of the effective ways of segmentation to group customers.

4.Setting Premium:

There are various regression methods available to predict the premium. Need to set premium amount for each segment based on the regression method. In this business model we have approached random forest regressor based prediction, which give most accurate prediction than the linear and K-nearest models.

3.1 Various Business Models:

3.1.1. Subscription Business Model:

In the subscription business model, customers (here customer means insurance companies) pay a fixed amount of money on fixed time intervals to get access to our insurance amount calculating product.

This business model is highly recommended one for who developing insurance premium calculator app which can used by insurance companies as their profits are highly reliable to the customer, through this app they can provide perfect data to their customer in fast, and they can provide customer service 24/7 using this app. It creates some strong and user-friendly relationship with their customer.

Through subscription model they can maintain many customers (Insurance company). It highly increases their profit.

3.1.2 Retailer Business Model:

This is the most valuable business model most of the big companies to prefer. Here the app owner can sell his complete app to the customer for the fixed amount.

we can sell the complete app to the insurance company with feature to update the data set from the insurance company end and we can provide the lifetime service of maintenance. In future we are come up with any updated algorithm and to increase the quality we can ask the company to update their app with the updating fees.

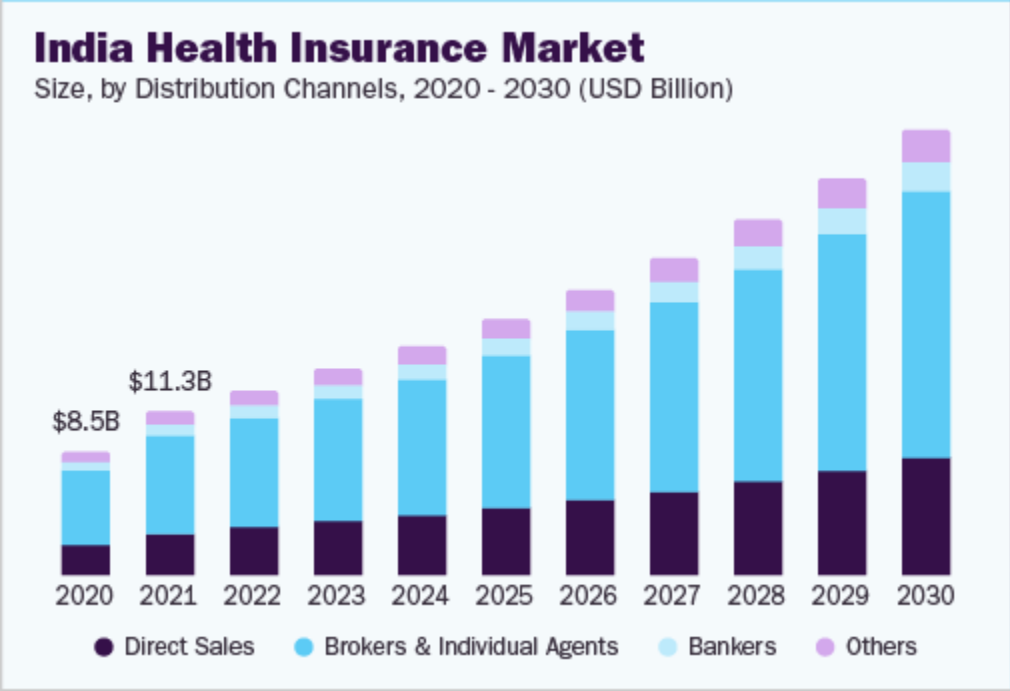
3.1.2 Pay-As-You-Go

Instead of charging a fixed fee, some companies may implement a pay-as-you-go business model where the amount charged depends on how much of the product or service was used. The company may charge a fixed fee for offering the service in addition to an amount that changes each month based on what was consumed.

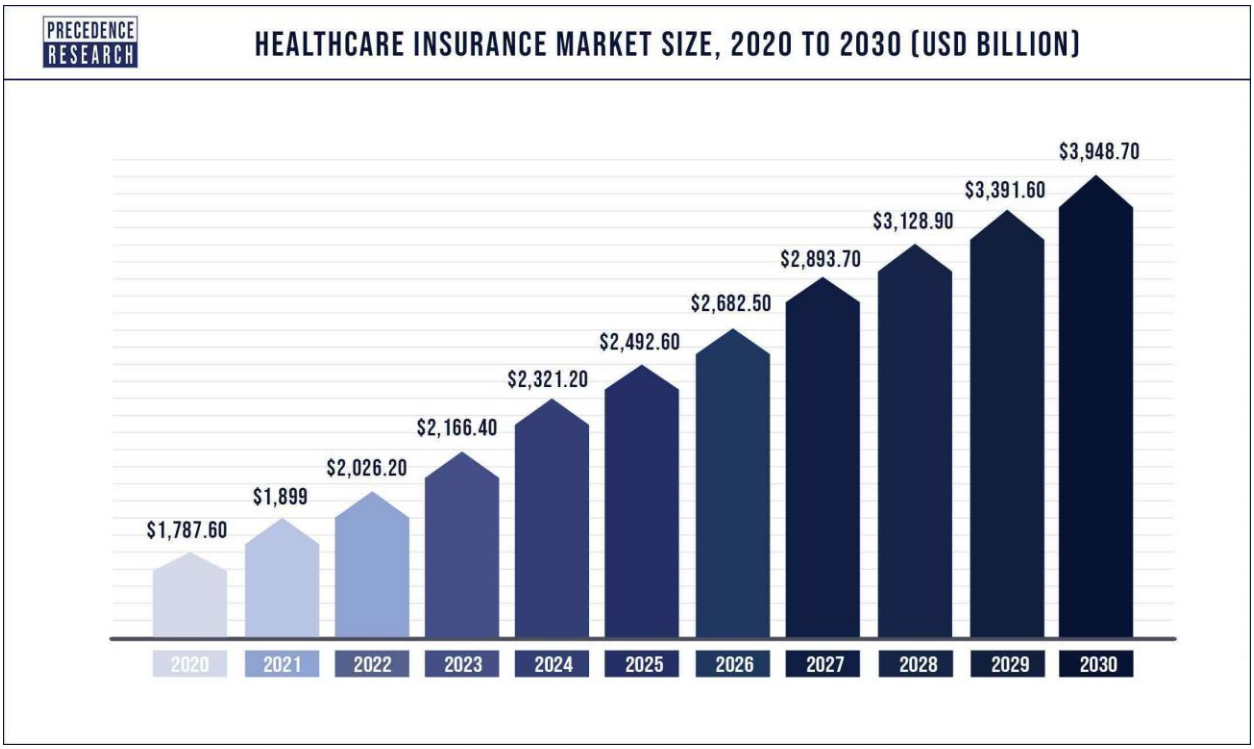
Here we can fix the price based on the number of feature vector the company want to use, More the feature vector the company want more the company should pay.

Step 4: Financial Modelling (equation) with Machine Learning & Data Analysis:

Health Insurance providers are the direct customer for our product. After covid the usage of health insurance was increased tremendous amount. In India everyone is facing challenges for medical benefits, so the number of medical insurance providers are increasing very year. Below chart explaining future health insurance market in India.



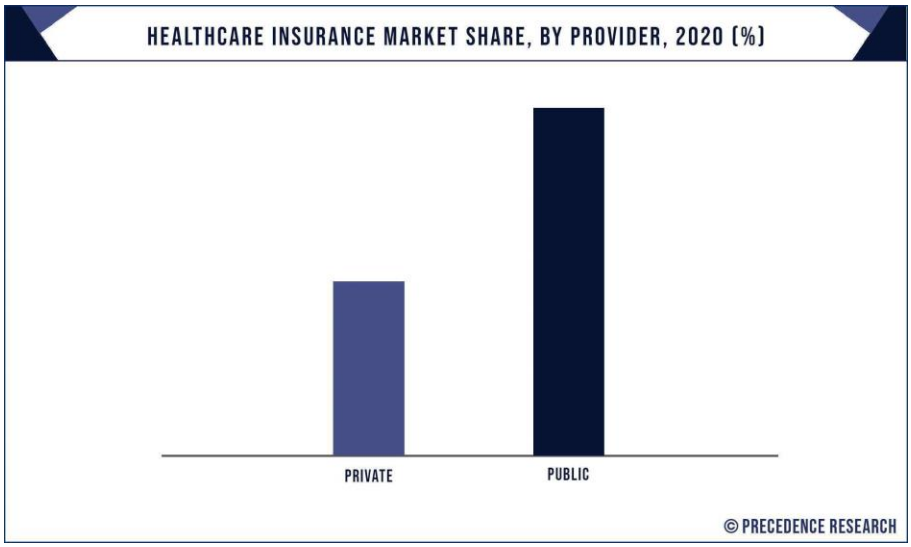
Global Health Insurance Market:



The global healthcare insurance market was estimated at USD 1,899 billion in 2021 and is expected to reach USD 3,948.7 billion by 2030, poised to grow at a compound annual growth rate (CAGR) of 7.6% during the forecast period 2021 to 2030.

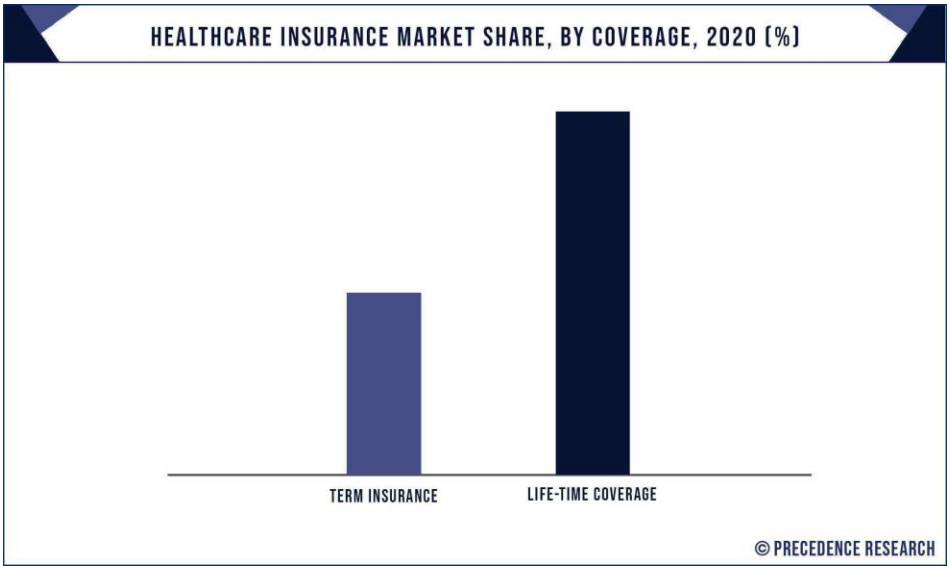
The global healthcare insurance market is primarily driven by various factors such as the rising prevalence of chronic diseases, high healthcare costs, increased disposable income, constant new product launches by the numerous players, rising awareness regarding health insurance, and improved services related to claim settlement.

Provider Insights:



the private segment is estimated to be the most opportunistic segment during the forecast period. The private players are coming up with improved healthcare services and premium options for the customers. Moreover, certain inefficiencies associated with the public sector healthcare insurance companies are overcome by the private market players, and hence is gaining more attention in the market. Moreover, they offer higher benefits as compared to that of the public segment.

Coverage Type Insights:



Financial Equation

$$\text{GPM} = \frac{Netsales - cost\ of\ product\ sold}{Net\ Sales} \times 100$$

Where,

GPM = Gross profit Margin; COGS = cost of product sold

Operating Profit Margin:

Operating profit is a slightly more complex metric, which also accounts for all overhead, operating, administrative, and sales expenses necessary to run the business on a day-to-day basis. By dividing operating profit by revenue, this mid-level profitability margin reflects the percentage of each dollar that remains after payment for all expenses necessary to keep the business running.

The formula for operating profit margin,

$$\text{OPM} = \frac{Operating\ Income}{Revenue} \times 100$$

Where, OPM = operating income.

Net Profit Margin:

The net profit margin reflects a company's overall ability to turn income into profit. Net Profit Margin reflects the total amount of revenue left over after all expenses and additional income streams are accounted for This includes not only COGS and operational expenses as referenced above but also payments on debts, taxes, one-time expenses or payments, and any income from investments or secondary operations.

$$\text{NPM} = \frac{R - CGS - OPE - OTE - I - T}{R} \times 100$$

NPM = Net profit margin; R = Revenue; CGS = Cost of goods sold; OPE = Operating system;

OTE = other Expense; I = Interest; T =Taxes