Project Report

On

Health Insurance -Cost Prediction

By coding strikers

A Sirisha

R Ananya

P Sree Lahari

P Skv Chaitanya

Y P Venkateswara Rao

Topics –

1 Title of the project - Health Insurance Cost Prediction

1.1 Introduction -

The United States' national health expenditure (NHE) grew 5.8% to $3.2 trillion in 2015 (i.e., $9,990 per person), which accounted for 17.8% of the nation's gross domestic product (GDP)[1]. In seeking to control these unsustainable increases in healthcare costs, it is imperative that healthcare organizations can predict the likely future costs of individuals, so that care management resources can be efficiently targeted to those individuals at highest risk of incurring significant costs[2]. Key stakeholders in these efforts to manage healthcare costs include health insurers, employers, society, and increasingly healthcare delivery organizations due to the transition from fee-for-service payment models to value-based payment models[3]. For any given individual, insurers generally have the most comprehensive information on healthcare costs as they pay for care delivered across various healthcare delivery organizations.

Predicting healthcare costs for individuals using accurate prediction models is important for various stakeholders beyond health insurers, and for various purposes[4]. For health insurers and increasingly healthcare delivery systems, accurate forecasts of likely costs can help with general business planning in addition to prioritizing the allocation of scarce care management resources. Moreover, for patients, knowing in advance their likely expenditures for the next year could potentially allow them to choose insurance plans with appropriate deductibles and premiums.

Despite the importance of healthcare cost prediction, to our knowledge there has been no review of the literature on this important topic. Therefore, we conducted a systematic literature review. Moreover, in order to enable a direct comparison of approaches on a common data set, we evaluated each of the identified approaches on a health insurance data set from the University of Utah Health Plans. We also evaluated additional state-of-the-art methods not previously evaluated in the literature.

1.2 Objective of research -

We constructed predictive models to predict the cost that a person can claim from an insurance company when he/she met an accident or falls ill.

The main outcome of the project is "basing on the effect of attributes on the health of an individual, we predict how much can he claim from the insurance policy".

1.3 Problem statement -

To predict the cost that a person can claim from an insurance company when he/she met an accident or falls ill. The prediction estimates the cost by considering the attributes, like age, sex, BMI, smoker, children and region.

1.4 Industry profile –

Health insurance in India is a growing segment of India's economy. The health situation and the provision of services vary considerably from one State to another. Although public health services in principle provide free basic health care to all, the care provided by most state health systems suffers from inadequate resources and poor management. In India, the health system mixes public and private providers to give the insurance.

Health Insurance covers the whole or a part of the risk of a person incurring medical expenses, spreading the risk over a large number of persons.

As a part of this ,every industry is providing a life time insurance to their employees for better prospective of their families.

2 Literature Review -

Adapting a search strategy from a previous systematic review[5], we searched Google Scholar and MEDLINE. The latest search was performed on February 21, 2017. We used a combination of the following search terms: healthcare cost prediction, medical claim cost, pharmacy claim cost; healthcare expenditure prediction; healthcare risk score prediction; and patient cost prediction.

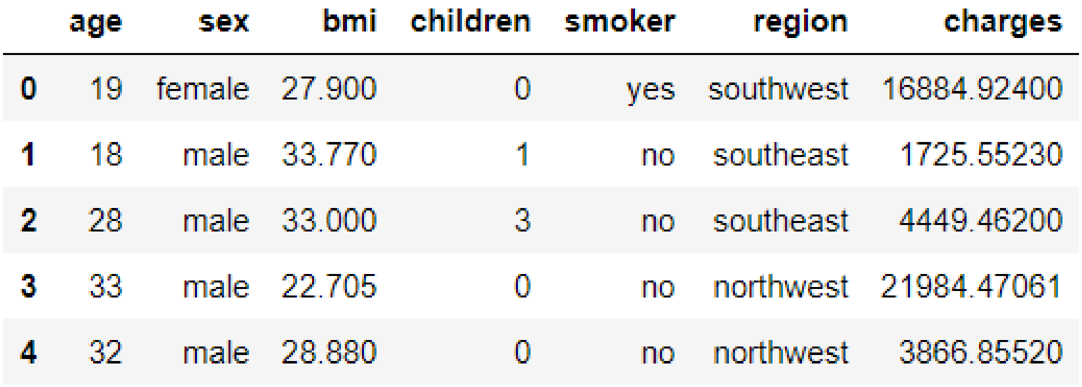
In conducting the systematic literature review, we sought to answer the following questions. Because the answer to the first question identified that using features of prior costs to predict future costs performed as well as or better than approaches that also used clinical data for cost prediction purposes, all subsequent questions were focused on approaches that used prior cost features to predict future costs (referred to henceforth as "cost on cost prediction").

1. What are the types of healthcare cost prediction approaches reported in the literature?
2. What are the input features that have been used for cost on cost prediction?
3. What are the supervised learning methods that have been used for cost on cost prediction?
4. What are the performance measures and evaluation results for cost on cost prediction?

3 Data Collection -

The dataset has various attributes on which the target variable i.e., the cost depends. The various independent attributes in the dataset are age, sex, body mass index, children, smoker and region. The insurance depends on these independent attributes and based on these attributes the cost is predicted.

The dataset is viewed as



	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

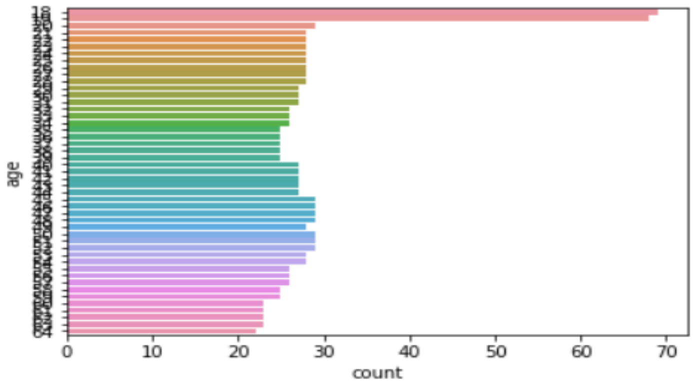
4. Methodology -

4.1 Exploratory Data Analysis -

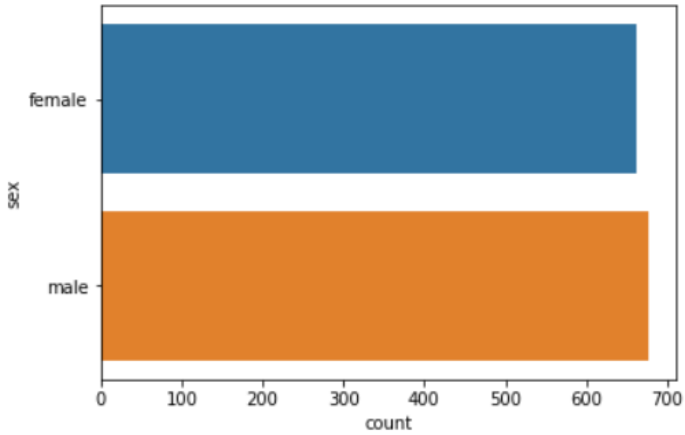
4.1.1 Figures and tables -

The data can be explored using various techniques such as count plot which is used to represent the count of unique value of an attribute.

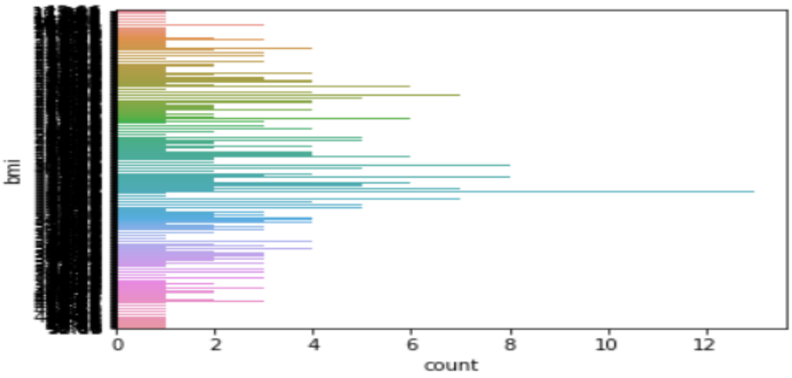
4.1.1.1 Count plot for Age



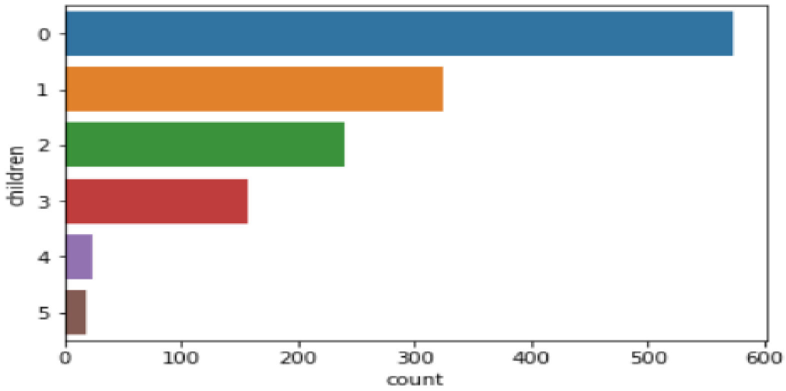
4.1.1.2 Count plot for Sex



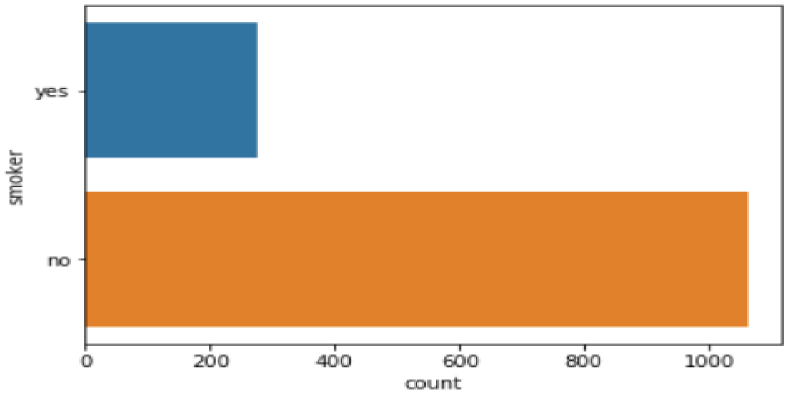
4.1.1.3 Count plot for BMI



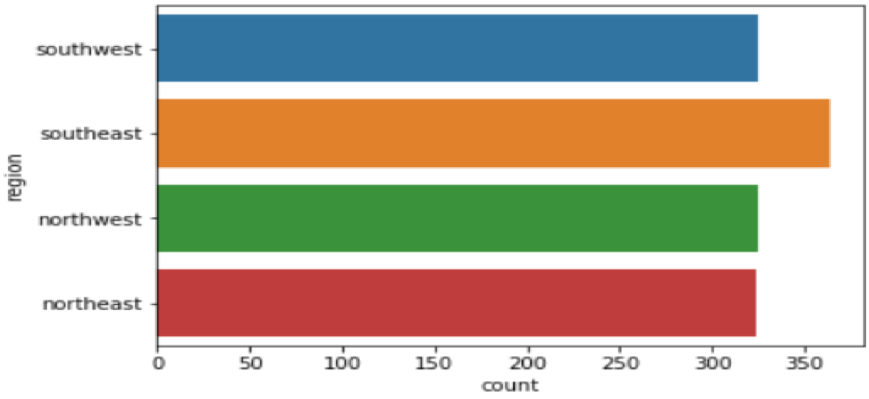
4.1.1.4 Count plot for Children



4.1.1.5 Count plot for Smoker



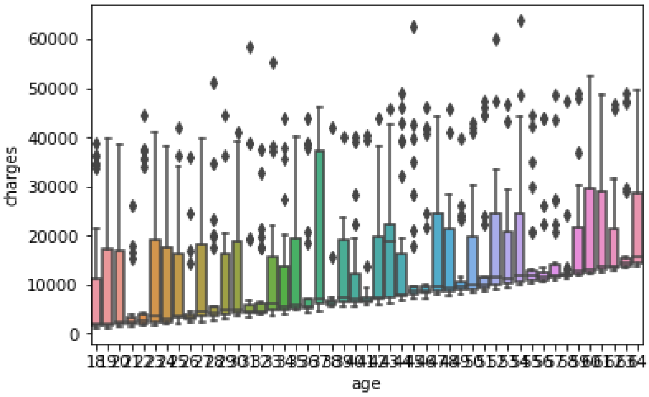
4.1.1.6 Count plot for Region



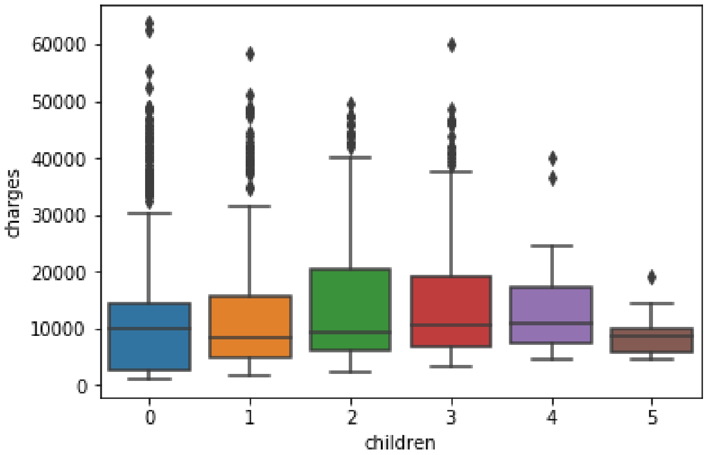
4.2 Statistical Techniques and Data Visualization –

- Data is visualized using Box Plots.
- **Box Plot** is a simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.

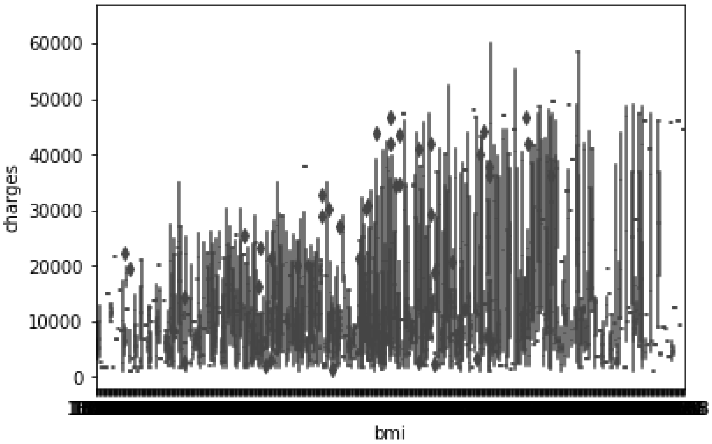
Relation between Age and Charges:



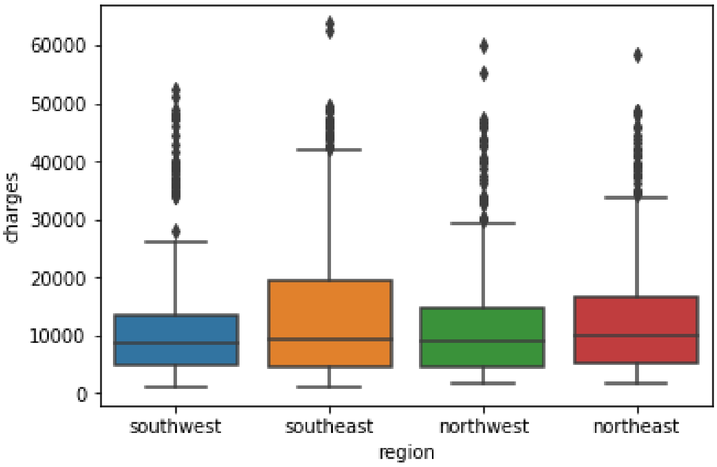
Relation between Children and Charges:



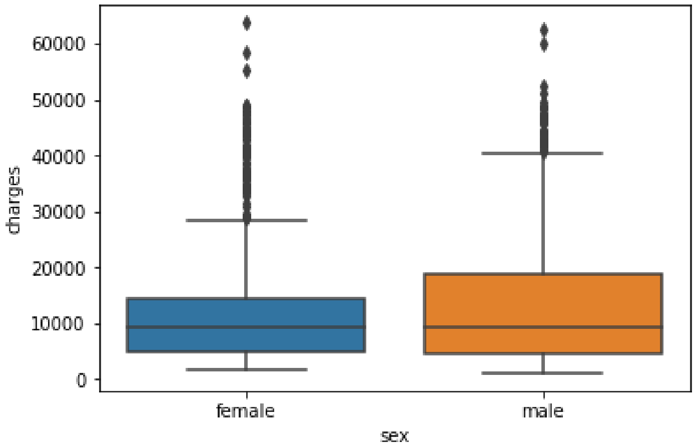
Relation between BMI and Charges:



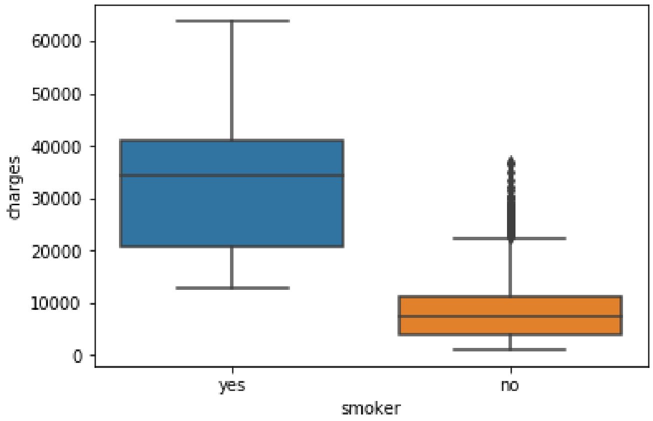
Relation between Region and Charges:



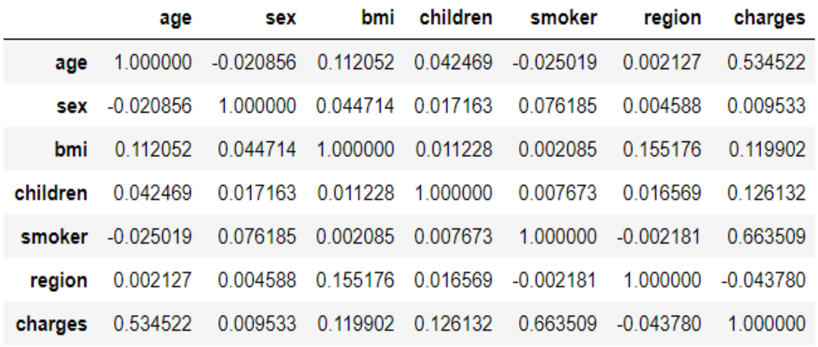
Relation between Sex and Charges:



Relation between Smoker and Charges:



Correlation among the attributes:



	age	sex	bmi	children	smoker	region	charges
age	1.000000	-0.020856	0.112052	0.042469	-0.025019	0.002127	0.534522
sex	-0.020856	1.000000	0.044714	0.017163	0.076185	0.004588	0.009533
bmi	0.112052	0.044714	1.000000	0.011228	0.002085	0.155176	0.119902
children	0.042469	0.017163	0.011228	1.000000	0.007673	0.016569	0.126132
smoker	-0.025019	0.076185	0.002085	0.007673	1.000000	-0.002181	0.663509
region	0.002127	0.004588	0.155176	0.016569	-0.002181	1.000000	-0.043780
charges	0.534522	0.009533	0.119902	0.126132	0.663509	-0.043780	1.000000

4.3 Data Modelling using Supervised ML techniques –

The modelling technique that is used to predict the cost is Random Forest Regression.

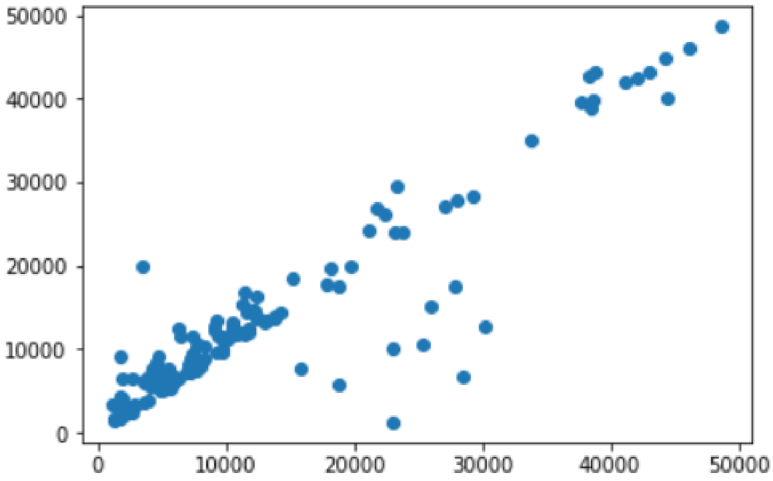
Random Forest Regression is a supervised learning algorithm. It is one of the most used algorithms as it can be used for both classification and regression problems.

The Random Forest model is a type of additive model that makes prediction by combining decisions from a sequence of base models. More formally we can write this class of models as

$g(x) = f_0(x) + f_1(x) + f_2(x) + ...$

where the final model **g** is the sum of all simple base models f_i.

The plot for target variable training and testing data is visualized as



5. Findings and Suggestions -

Findings -

- Age plays a major role in the cost of a premium for health insurance. Younger people get less insurance, as the age increases their insurance also gets increased.
- Basing on the gender females get less insurance than males.
- Basing on the BMI, person with normal weight gets highest insurance, then the people with underweight, then the people with overweight, and the least insurance is for the people with obesity.
- When a person has no children or more than 3 children the insurance claiming will be less.
- Basing in the habit, if a person smokes he will get less insurance than the non-smokers.
- Basing on the climatic conditions, if it is adverse the health may deteriorate so that the insurance claimed by the person will be more and vice versa.

Suggestions to Government -

- Recognising health insurance as a separate line of business.
- Reduce capital requirement for health insurers from Rs 100 crore to Rs 30-50 crore.
- Introduce capital monitoring and product level norms for health insurance.
- Accreditation and benchmarking of health providers. There should be some quality standards and protocols to follow.

- Invest in training doctors, providers, health economists, cost accountants, epidemiologists, hospital managers, record keepers in computerisation etc.
- Reform public health system by decentralising autonomy and invest more to ensure standards

Suggestions to Insurance Companies -

- Creating more awareness regarding health insurance.
- Strong underwriting and claims management.
- Review of Mediclaim to cover 'existing illness', if possible with a higher premium.
- Introduction of new products for different market segments.
- Offer products for specific treatments to profitable segments.

Suggestions to Hospitals -

- Accreditations and standardizations of the tariff as far as possible for similar pattern of healthcare providers.
- Regular orientation to the doctors regarding health insurance.
- Concept of negotiable doctor's fees should be discouraged as far as possible.

Suggestions to Health insurance customers -

- Taking health policy at very young age and covering all members of the family.
- Customers should be fully aware of the various health coverages available.
- Customers should know about the various health insurance schemes and companies providing these schemes.
- The attitude of customers should be always towards the preventive health care.
- Customers should take decisions relating to the features of the policy, sum assured, premium paid, persons covered, after careful analysis.
- They must be aware of the conditions and exclusions in the policy

6 Conclusion -

The literature indicates that the preferred approach to healthcare cost prediction is cost on cost prediction using supervised learning methods. Empirical analysis of alternate approaches using data from a single health insurer found that gradient boosting provides the best cost on cost prediction models in general, with ANN providing superior performance for higher cost patients.

"Health coverage to all" should be the motto ofthe health insurance sector. There should be easy access to healthcare facilities and cost control measures should be in place. Health insurance is going to develop more in the current liberal economic scenario. But, a completely unregulated or very less regulated health insurance sector may concentrate only on those who have the ability to pay for the insurance cover. So, the challenge is in helping the benefits percolate to the economically weaker sections of the population. Transparent and accountable government and non-government participation should be encouraged. Developing and marketing social health insurance schemes through cooperatives and rural association would go a long way in benefiting the vast unorganized employment sectors currently neglected under the existing schemes. Also a thorough revamp of schemes like ESIS and CGHS is necessary for them to be more purposeful and efficient.

If the government, service provider, health care industry and the health insurance customers can incorporate all these suggestions given in the study, then the concept of health insurance will reach new heights in the near future and Mother India will be definitely, the most healthiest nation in the world

7 References -

1. The Centers for Medicare & Medicaid Services (CMS) DoHaHS, United States. National Health Expenditure Data 2016. Available from: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics- Trends-and-Reports/NationalHealthExpendData/index.html.

2. Duncan I, Loginov M, Ludkovski M. Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. North American Actuarial Journal. 2016;20(1):65–87. [Google Scholar]

3. Burwell SM. Setting value-based payment goals--HHS efforts to improve US health care. 2015[PubMed] [Google Scholar]

4. Lahiri C, Agarwal N. Predicting healthcare expenditure increase for an individual from medicare data. Proceedings of the ACM SIGKDD Workshop on Health Informatics. 2014 [Google Scholar]

5. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. Bmj. 2005;330(7482):68. [PMC free article] [PubMed] [Google Scholar]