An Ensemble Model to predict Health Insurance Premium using Machine Learning

Jubeena Rampal
Department of Computer Science & Engineering
IKG Punjab Technical University
Jubibarampal@gmail.com

*Prabhdeep Singh
Department of Computer Science & Engineering
Punjabi University, Patiala
Ssingh.prabhdeep@gmail.com

Dr. Rajbir Kaur
Department of Electronics & Communication Engineering
Punjabi University, Patiala
Rajbir277@yahoo.in

Dr. Kirandeep Singh
Department of Computer Science & Engineering
IKG Punjab Technical University
kdkirandeep@gmail.com

*Corresponding Author

Abstract

Health insurance is one of the most significant investments an individual makes every year. One-third of GDP is spent on health insurance, and everyone needs some level of health care. The healthcare premiums keep changing every year because of various factors such as medical trends, pharmaceutical trends, and political factors. There is a need to develop a mathematical model to predict premiums based on various parameters that impact the premiums. The premium rates are set by insurance providers using complex algorithms based on previous years' health care utilization and the total number of enrollments. In this paper, an ensemble-based regression model to predict future premiums is proposed. The proposed model is compared with the traditional four regression models. It is proved with various metrics that the proposed model always gives better results.

1. Introduction

It is a vital market for health insurance because one-third of GDP is invested in health insurance and everybody wants a certain level of healthcare. Health insurance is one of the most important investments every year made by individuals. This study seeks to identify mathematical models to forecast future premiums and to verify results with regression models. Medical costs resulting from injuries, incidents, and other medical reasons, without health insurance, are significantly costly; an individual is not required to pay for the entire medical cost of the treatment. Worldwide, there are several health care systems. Every year insurance rates adjust regardless of various causes, such as medical patterns, prescription developments, and policy considerations, etc., about which the consumer has little influence. The only choice for an individual is to carefully prepare potential expenses. There are noexisting tools that can predict future premiums based on historical data. Therefore, research is needed to find the premiums. The focus of this research is on predicting health insurance premiums based on individual market data for health insurance.

Although the health care system has undergone great uncertainty in recent years, the health care system is not completely lost for the following reasons:

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

1) The prices fluctuate based on the competition for the business position.

2) Safety from enhanced walks is provided by the government. Consumers receive their insurance premium tax credit.

Thus, a statistical model needs to be built to forecast premiums based on various parameters that influence premiums. The premium cost is calculated by insurance providers using sophisticated formulas dependent on use in health coverage and the cumulative amount of transactions in previous years.

2. Proposed Work

Various regression methods have been used in this paper to forecast potential premiums. Flowchart of the whole cycle is shown in Figure 1. The initial data for the learning phase are gathered and placed in the data collection or intake portion. Data may typically be in multiple formats, including organized and unstructured formats. These data can be collected in distributed format from Streaming APIs, Weblogs of cloud stores,or csv,.xls, or .json formats.

.

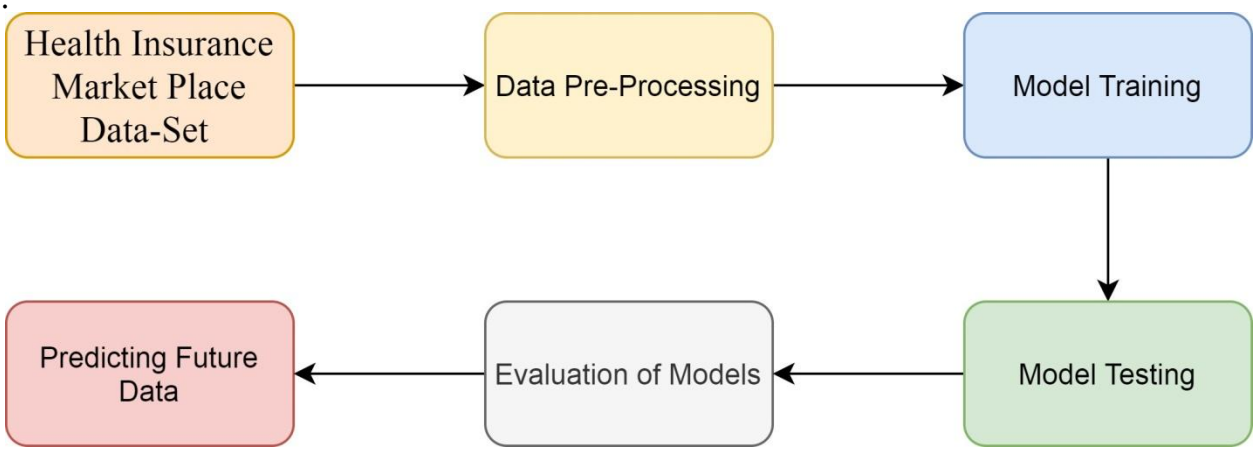

Fig 1: Flowchart of the Regression Process

Raw data from the health insurance market are obtained. The data was pre-processed and converted into specific models. Preprocessing consists of eliminating null values, contradictory values, and so on. The transition involves transforming strings and group variables into numerical values and eventually scaling them to the interval [0, 1]. The transformation includes The transformed data were conditioned on and evaluated with various frameworks such as the Decision Table, the Gaussian Method, the conditional regression, the Zero R, and the Ensembles Regression System. The training and testing process is iterated by changing parameters several times until the best accuracy has been achieved. Regression metrics in the evaluation model are measured. The strongest approach can be used to estimate potential values with metrics.

2.1 Health Insuranse Market Place Dataset

The data-set used in this study was prepared from the various Health Insuranse Market Place Dataset. The data format was avaiable in various formats with required attributes.Reports on the

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

health care business program were released quarterly. The syntactic data for the health insurance industry were known for our study. The first move in processing the data is to combine the required information into a single account. The initial files are pre-processed to eliminate incoherence and duplication. before entering the results. The combined data is divided into a set of 70 percent training and 30 percent validation remaining.

2.2 Data Pre Processing

The lack, incompleteness, or corruption of data leads to wrong results while performing functions such as count, average, average, and so on. Until doing some data review, such incoherence must be eliminated. The initial market place files have several columns, such as emitter addresses, names and deductible applicability, etc. that don't affect analysis and the columns have been removed from data. Text data is usually prone to spelling errors because of human errors; it needs therefore to be corrected. In the initial files, columns for state and county have several misspelled values that have been corrected by referring to one of the files.

As market participation for insurance companies is not mandatory, several new states and counties have been added and dropped in due course

2.2.1 Data Cleaning and Transformation:

Initial statistics provide different premium levels for infants, couples, and workers for varying ages. Such specific columns are turned into columns of the era, family size, and premium value. For instance, the original files have 36 columns, such as 'Premium adult individual age 21,' 'Premium adult individual age 27,' 'Couple+2, child, child 30,' etc. using column headers for preprocessing, has been transformed into four columns of age, family size, premium, and couple columns.

2.2.2 Categorical data Conversion:

Any of the factors are naturally categorical. A numerical representation was encoded with categorical details. The column age has values in the ranges named buckets for some period (for example, 0-21, 22-30, and so on). The average, minimum, or limit of the buckets is chosen for translating such variables into a single number. The average value is known for the age element. Normalization and standardization of graphical elements provide a clear pattern scale with all various data variables. All columns are normalized in the interval [0, 1] because many Learning Machine models allow the proper functionality of uniform data.

2.3 Model Training

Once the training data is in an acceptable state to feed the model, the model will continue with its development and testing process. Throughout the training process, the model collection is the primary concern. It involves choosing the best simulation method or the right parameter settings for a particular application. Also, the choosing of a product applies to all methods as, in certain situations, various models were first evaluated and the better model result was chosen. The use of combinations of different models (known as ensemble methods) during the training process is often popular. Usually, this is a relatively straightforward method to operate a model based on a training dataset and assess its output on a test dataset (i.e. a sequence of data kept to validate the model that the model did not see at a training stage). This method is called cross-validation. The algorithm of learning trains a model with a named dataset. Only labeled training data sets can be used. Each observation in the training data set includes several input and output features. The dependent variable, also called the response variable, is the output. The input characteristics are

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

the independent variables, also known as predictive or indicator variables.It is apparent that the model proposed operates well on validation data and forecasts the values nearly.

2.4 Model Testing

The gold standard used to evaluate the model is given in the test data set. It is only used if a model (using the train and validation sets) is fully trained. The test set is usually used to compare competing models. The validation set is often used as a test set, but not good practice. In general, the test set is well curated. This includes randomized data representing the various groups the concept encounters as used in the real world.

2.5 Evaluation of Prediction Model

The primary goal of the prediction models is to generalize outside the examples range. This is very true because, whatever the variety of data in training, the same examples are very unlikely to be found during testing. The regression calculations are based on the training set, the validation set, and the evaluation set to assess the prediction models.The size and sign are checked for the coefficients obtained from multiple linear regression models. The larger value coefficients are more relevant to the model and vice versa. This assumes that the larger the cost and the greater the limit, the smaller the price.The tree structure is formed in the proposed regression from top to bottom using the importance of the feature. The root-knot is the family size, and the next hierarchy form reflects the linear regression of age, place, etc. did not consider the significance of places like the state and county that was defined by the proposed model.

2.6 Predicting future data

This is the final outcome of our model. Themodel is capable enough to predict the future value. In this paper the regression model is design after applying the model the numeric values are predicted. The Health Insurance premium are predicted at earlier stage.

3. Experiment Setup and Performance measure

This proposed model is implemented by using the weka which is evaluated with various parameters like Co-Relation Coefficient, Mean Absolute Error, Root Mean Squared Error Relative Absolute Error, and Root Relative Squared Error.

3.1 Performance measure

The evaluation of the machine learning regression model is an vital part. If tested with a metric say accuracy score, the model may offer you good results but it may provide bad results if calculated against certain measurements, such as logarithmic loss or other similar metrics. We use various different parameters as Co-Relation Coefficient, Mean Absolute Error, Root Mean Squared Error Relative Absolute Error, and Root Relative Squared Error to evaluate model.

3.1.1 Mean absolute error

Mean absolute error (MAE) is a probability parameter that meets the estimated absolute error failure value also known as the L1 loss rate. If y total is the expected value of the ith sample, and yi is the corresponding true value, the MAE approximation is defined as n input rows.

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

$(y,\bar{y})= 1n\Sigma|yi-\bar{yi}|$

Fig.2 shows that proposed ensemble regression model has 8.90 mean absolute error which is very less as compare to other traditional regression models.it is also results that ZeroR has the maximum error.

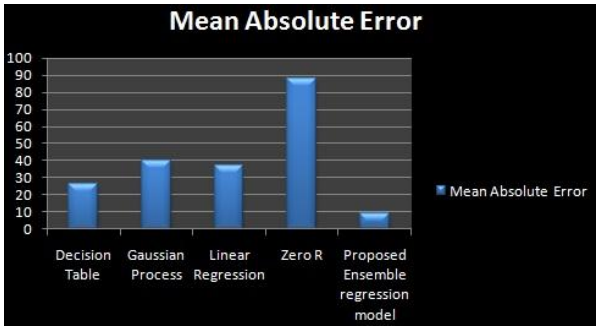

Fig 2: Mean Absolute Error

3.1.2 Mean square error

Mean square error is a risk metric that corresponds to the expected value of the square error loss (quadratic). If yi is the predicted value of the ith sample and yi is the corresponding true value, then the estimated mean squared error (MSE) over n input rows is defined.

$(y,\bar{y})= 1n\Sigma(yi-\bar{yi})2.$

As seen in figure 3, the proposed model had thelowest root mean squared error(36.35). In this race again ZeroR had the last position with maximum value.

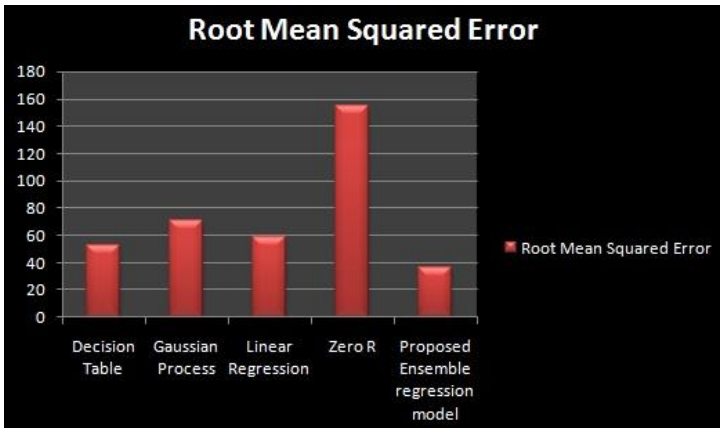

Figure 3: Root Mean Squared error

3.1.3 Root Relative Squared Error

The RRSE is equivalent to what it would have been if a simple predictor were used. This basic indicator is the average of actual values. Therefore, the relative squared error takes the total squared error and normalizes it by splitting the basic predictor 's total squared error. By taking the square root of the relative square error, the error is reduced to the same dimensions as expected.As shown in figure 4. in contrast to other models, the proposed model won with the value 23.49 of Root Relative Squared Error to complete the fitting of the model.

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

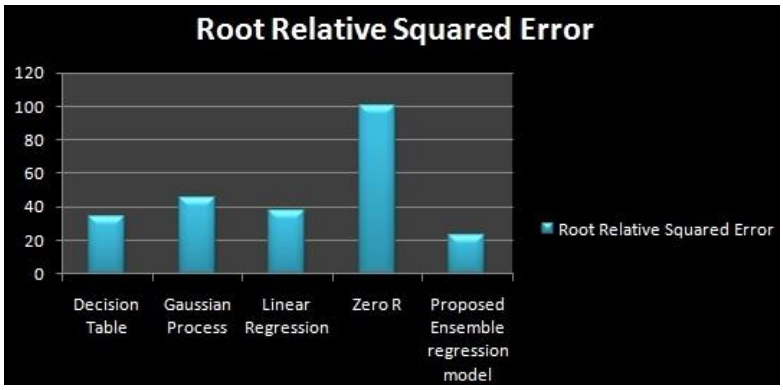

Figure 4: Root Relative Squared Error

3.1.4. Relative Absolute Error

Relative Absolute Error (RAE) is a calculation of predictive model efficiency. It's used mainly in machine learning, data mining, and operations management. RAE should not be confused with relative error, a general indicator of precision or accuracy for instruments like clocks, rulers, or scales. The Relative Absolute Error is expressed as a ratio, contrasting a mean (residual) error with a negligible or naive model. A rational model (one that performs better than a trivial model) should result in less than one ratio.Figure 5 shows that proposed model always perform better on relative absolute error with the lowest value of 10.16.

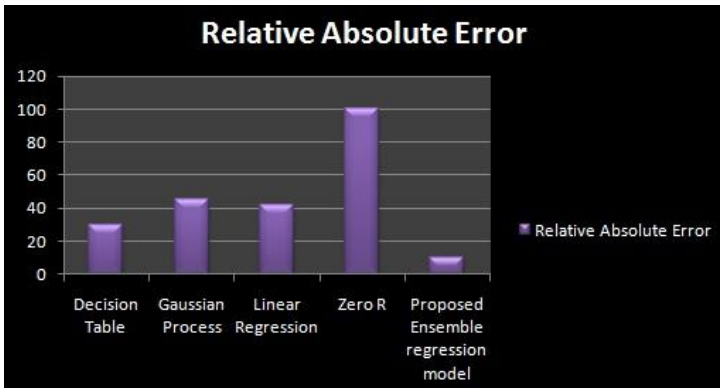

Figure 5: Relative Absolute Error

3.1.5 Correlation Coefficient

The correlation coefficient is a statistical estimate of the relationship intensity between two variables' relative movements. Values range from -1.0 to 1.0. A measured number greater than 1.0 or less than -1.0 indicates a correlation calculation error. A -1.0 correlation shows a perfect negative correlation, while a 1.0 correlation shows perfect positive correlation. Figure 6 is displaying the correlation coefficient. They have proven how the correlation coefficient is highest with the proposed model therefore model is providing better results than the available methods

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

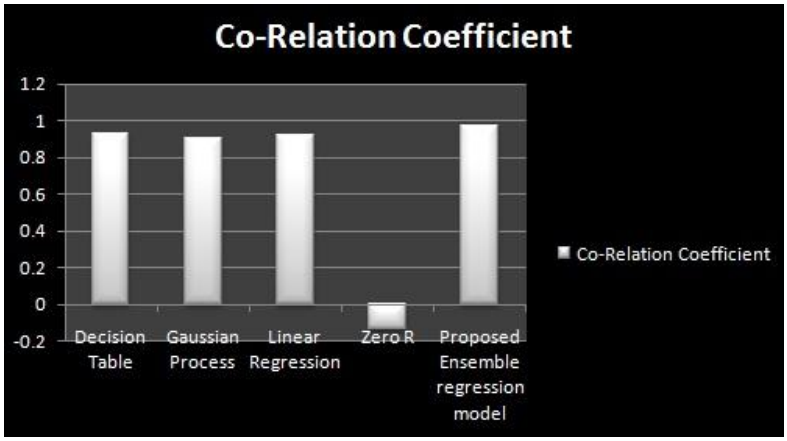

Figure 6: Correlation Coefficient

The following table 1 shows the Comparison of Regression metrics.

Table 1: Comparison of Regression metrics

Regression Model	Co-Relation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error
Decision Table	.93	26.04	52.83	29.70	34.13
Gaussian Process	.90	39.77	70.61	45.37	45.62
Linear Regression	.92	36.96	58.45	42.17	37.76
Zero R	-.14	87.65	154.76	100	100
The proposed Ensemble regression model	.97	8.90	36.35	10.16	23.49

4. Conclusions

This paper provides an ensemble regression model for data from the health insurance market, compared to four regression models. The proposed model is the best performing model. The models will be used to estimate the value of the data obtained in the coming years. If the model is built on a larger dataset with several years, the accuracy of the model is expected to increase further. In the future, a two-sided web application, client-side application, and server-side application can be built. Predictions dependent on program choices may be presented and contrasted on the consumer side. The server handles data training and customer prediction queries. The server should have the necessary computational power and libraries.

References

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

1. Xie, Y., Schreier, G., Chang, D.C., Neubauer, S., Liu, Y., Redmond, S.J. and Lovell, N.H., 2015. Predicting days in hospitals using health insurance claims. IEEE Journal of biomedical and health informatics, 19(4), pp.1224-1233.
2. Miyano, T., Tsutsui, T., Seki, Y., Higashino, S., and Taniguchi, H., 2005. Prediction of care class by local additive reference to prototypical examples. *IEEE Transactions on Information Technology in Biomedicine*, *9*(4), pp.502-507.
3. Ting, C.H., Mahfouf, M., Nassef, A., Linkens, D.A., Panoutsos, G., Nickel, P., Roberts, A.C. and Hockey, G.R.J., 2009. Real-time adaptive automation system based on identification of operator functional state in simulated process control operations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *40*(2), pp.251-262.
4. Yang, C., Delcher, C., Shenkman, E. and Ranka, S., 2018. Machine learning approaches for predicting high cost high need patient expenditures in health care. *biomedical engineering online*, *17*(1), p.131.
5. Boodhun, N. and Jayabalan, M., 2018. Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, *4*(2), pp.145-154.
6. Cheng, L., Shi, Y. and Zhang, K., 2020. Medical treatment migration behavior prediction and recommendation based on health insurance data. *World Wide Web*, pp.1-20.
7. Jödicke, A.M., Zellweger, U., Tomka, I.T., Neuer, T., Curkovic, I., Roos, M., Kullak-Ublick, G.A., Sargsyan, H. and Egbring, M., 2019. Prediction of health care expenditure increase: how does pharmacotherapy contribute?. *BMC health services research*, *19*(1), p.953.
8. Young, J.B., Gauthier-Loiselle, M., Bailey, R.A., Manceur, A.M., Lefebvre, P., Greenberg, M., Lafeuille, M.H., Duh, M.S., Bookhart, B. and Wysham, C.H., 2018. Development of predictive risk models for major adverse cardiovascular events among patients with type 2 diabetes mellitus using health insurance claims data. *Cardiovascular diabetology*, *17*(1), p.118.
9. Bahuguna, P., Guinness, L., Sharma, S., Chauhan, A.S., Downey, L. and Prinja, S., 2020. Estimating the Unit Costs of Healthcare Service Delivery in India: Addressing Information Gaps for Price Setting and Health Technology Assessment. *Applied Health Economics and Health Policy*, pp.1-13.
10. Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S., 2019. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, *6*(1), p.54.
11. Zhu, Y., Wu, H. and Wang, M.D., 2019, May. Feature Exploration and Causal Inference on Mortality of Epilepsy Patients Using Insurance Claims Data. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (pp. 1-4). IEEE.
12. Ren, Y., Zhang, K. and Shi, Y., 2019, November. Survival Prediction from Longitudinal Health Insurance Data using Graph Pattern Mining. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1104-1108). IEEE.

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23

13. Singh, K.D. and Sood, S.K., 2020. Optical fog-assisted cyber-physical system for intelligent surveillance in the education system. *Computer Applications in Engineering Education*, *28*(3), pp.692-704.
14. Kaur, P., Singh, P. and Singh, K., 2019. AIR POLLUTION DETECTION USING MODIFIED TRAINGULAR MUTATION BASED PARTICLE SWARM OPTIMIZATION.
15. Sood, S.K. and Singh, K.D., 2018. Identification of a malicious optical edge device in the SDN-based optical fog/cloud computing network. *Journal of Optical Communications*, *1*(ahead-of-print).
16. Gupta, V., Singh Gill, H., Singh, P. and Kaur, R., 2018. An energy efficient fog-cloud based architecture for healthcare. *Journal of Statistics and Management Systems*, *21*(4), pp.529-537.
17. Singh, N., Singh, P. and Kaur, R., 2019. Design and Development a Hybrid Classifier to Improve Lung Cancer Diagnosis. *Journal of the Gujarat Research Society*, *21*(15), pp.323-328.
18. Xue, Y., Harel, O. and Aseltine, R., 2019, July. Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data. In *2019 13th International conference on Sampling Theory and Applications (SampTA)* (pp. 1-4). IEEE.
19. Umemoto, K., Goda, K., Mitsutake, N. and Kitsuregawa, M., 2019, April. A Prescription Trend Analysis using Medical Insurance Claim Big Data. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (pp. 1928-1939). IEEE.
20. Wang, H.Y., Hsieh, C.H., Wen, C.N., Wen, Y.H., Chen, C.H. and Lu, J.J., 2016. Cancers screening in an asymptomatic population by using multiple tumour markers. *PloS one*, *11*(6), p.e0158285.
21. Le Nguyen, T. and Do, T.T.H., 2019, August. Artificial Intelligence in Healthcare: A New Technology Benefit for Both Patients and Doctors. In *2019 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1-15). IEEE.

D.O.I : 10.46528/DRSRJ.2020.V10I08N02.23