www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May 2022, pp : 346-348

e-ISSN : 2583-1062
Impact Factor : 2.205

PREDICTING HEALTH INSURANCE COST USING MACHINE LEARNING REGRESSION MODELS

Prof. Sangameshwar Kawdi[*1], Gourishanker[*2]

[*1,2]Prof., Assistant Professor, Department of Information Science and Engineering,
Guru Nanak Dev Engineering College, Bidar, Karnataka, India.

[*2]UG SCHOLAR, Department of Information Science and Engineering,
Guru Nanak Dev Engineering College, Bidar, Karnataka, India.

ABSTRACT

Insurance is a policy that eliminates or decreases loss costs occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs. And we will compare the results of models, for example, Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, CART, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network. This paper offers the best approach to the Stochastic Gradient Boosting model .

Keywords: Insurance, regression, machine learning, k-Nearest, gradient boosting.

1. INTRODUCTION

The main aim of this paper is to identify or predict the nearest value of the health insurances of the citizens based on the collected data. This model ensures the predicted amount for the health insurance gives maximum accuracy to the people by implementing various different algorithms.+Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also people in rural areas are unaware of the fact that the government of India provide free health insurance to those below poverty line. It is very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance. Our paper does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance.

2. METHODOLOGY

Below listed are the different regression models which are used

Multiple Linear Regression, Decision Tree Regression,Gradient Boosting Regression.

2.1 Multiple Linear Regression

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable

Regression allows you to estimate how a dependent variable changes as the independent variable(s) change. Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable

2.2 Decision Tree Regression

Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The **Interior Nodes** represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems

2.3 Gradient Boosting Regression

It is one of the most powerful algorithms in the field of machine learning. Unlike, boosting algorithm, the base estimator in the gradient boosting algorithm cannot be mentioned by us.The base estimator for the Gradient Boost algorithm is fixed and i.e. Decision Stump. Like, AdaBoost, we can tune the n_estimator of the gradient boosting algorithm. However, if we do not mention the value of n_estimator, the default value of n_estimator for this algorithm is 100

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
Vol. 02, Issue 05, May 2022, pp : 346-348

e-ISSN : 2583-1062
Impact Factor : 2.205

Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss

3. MODELING AND ANALYSIS

Data Preparation & Cleaning

The data has been imported from kaggle website. The website provides with a variety of data and the data used for the project is an insurance amount data. The data included various attributes such as age, gender, body mass index, smoker and the charges attribute which will work as the label for the project. The data was in structured format and was stores in a csv file format. The data was imported using pandas library. The presence of missing, incomplete, or corrupted data leads to wrong results while performing any functions such as count, average, mean etc. These inconsistencies must be removed before doing any analysis on data. The data included some ambiguous values which were needed to be removed.

Training

Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed. During the training phase, the primary concern is the model selection. This involves choosing the best modelling approach for the task, or the best parameter settings for a given model. In fact, the term model selection often refers to both of these processes, as, in many cases, various models were tried first and best performing model (with the best performing parameter settings for each model) was selected.

Prediction

The model was used to predict the insurance amount which would be spent on their health. The model used the relation between the features and the label to predict the amount. Accuracy defines the degree of correctness of the predicted value of the insurance amount. The model predicted the accuracy of model by using different algorithms, different features and different train test split size. The size of the data used for training of data has a huge impact on the accuracy of data. The larger the train size, the better is the accuracy. The model predicts the premium amount using multiple algorithms and shows the effect of each attribute on the predicted value.

4. RESULTS AND DISCUSSION

We see that the accuracy of predicted amount was seen best i.e. 99.5% in gradient boosting decision tree regression. Other two regression models also gave good accuracies about 80% In their prediction. Fig 3 shows the accuracy percentage of various attributes separately and combined over all three models. Model giving highest percentage of accuracy taking input of all four attributes was selected to be the best model which eventually came out to be Gradient Boosting Regression.

	Linear Regression	Decision tree regression	Gradient boosting regressor
Age	8-13	2-13	2-20
Gender	0	0	0
Smoker	50-60	57-69	57-70
BMI	0-4	0-1	0
Age+Gender	0-15	0-1	0-15
Age+smoker	9-12	2-4	6-17
Age+BMI	0-9	0	0-11
Gender+smoker	59-65	56-69	59-67
Gender+BMI	2-10	0	0-3
Smoker+BMI	61-80	58-74	74-81

Fig.3 accuracy percentage of various attributes separately and combined over all three models

5. CONCLUSION

The health insurance data was used to develop the three regression models, and the predicted premiums from these models were compared with actual premiums to compare the accuracies of these models. It has been found that Gradient Boosting Regression model which is built upon decision tree is the best performing model. Various factors were used and their effect on predicted amount was examined. It was observed that a person's age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features. The effect of various independent variables on the premium amount was also checked. The attributes also in combination were checked for

better accuracy results. Premium amount prediction focuses on person's own health rather than other company's insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount.

6. REFERENCES

[1] B. Milovic and M. Milovic, "Prediction and decision making in health care using data mining," Kuwait Chapter of the Arabian Journal of Business and Management Review, vol. 1, no. 12, 2012.View at: Publisher Site | Google Scholar

[2] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation," in Proceedings of the AMIA Annual Symposium Proceedings, vol. 2017, American Medical Informatics Association, Washingdon, DC, USA, November 2017.View at: Google Scholar

[3] M. Kumar, R. Ghani, and Z. S. Mei, "Data mining to predict and prevent errors in health insurance claims processing," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 65–74, Washingdon, DC, USA, July, 2010.View at: Google Scholar

[4] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," BioMedical Engineering Online, vol. 17, no. 1, pp. 131–220, 2018.View at: Publisher Site | Google Scholar

[5] M. Iqbal and Z. Yan, "Supervised machine learning approaches: a survey," ICTACT Journal on Soft Computing, vol. 5, no. 3, 2015, https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/.View at: Google Scholar

[6] B. Panay, N. Baloian, J. A. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting health care costs using evidence regression," Multidisciplinary Digital Publishing Institute Proceedings, vol. 31, no. 1, p. 74, 2019.View at: Publisher Site | Google Scholar

[7] M. U. Ghani, T. M. Alam, and F. H. Jaskani, "Comparison of classification models for early prediction of breast cancer," in Proceedings of the International Conference on Innovative Computing (ICIC), Lahore, Pakistan, November.2019.View at: Publisher Site | Google Scholar

[8] K. Shaukat, F. Iqbal, T. M. Alam et al., "The impact of artificial intelligence and robotics on the future employment opportunities," Trends in Computer Science and Information Technology, vol. 5, no. 1, pp. 50–54, 2020.View at: Publisher Site | Google Scholar

[9] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in Proceedings of the IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE), pp. 1–6, Gold Coast, Australia, December 2020.View at: Publisher Site | Google Scholar

[10] T. M. Alam, M. M. A. Khan, M. A. Iqbal, W. Abdul, and M. Mushtaq, "Cervical cancer prediction through different screening methods using data mining," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, 2019.View at: Publisher Site | Google Scholar

[11] M. A. Fauzan and H. Murfi, "The accuracy of XGBoost for insurance claim prediction," International Journal of Advanced Software Computer Applications, vol. 10, no. 2, 2018.View at: Google Scholar

[12] B. S. Van, Flexible Imputation of Missing Data, CRC Press, Boca Raton, FL, USA, 2018.