International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Health Insurance Amount Prediction using Machine Learning

Mogili Siva Naga Raju[1], Nelapudi Anandavalli[2], Sayed Farook Ali Abdul Kharem[3]

MR.K. Sanjeevaiah (Guide)

Department of Computer Science & Engineering
Hyderabad Institute of Technology And Management

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, Random Forest Regression, Decision tree.

Key Words: Health Insurance, Regression, ML

1.INTRODUCTION

We are on a planet full of threats and uncertainty. People, households, companies, properties, and property are exposed to different risk forms. And the risk levels can vary. These dangers contain the risk of death, health, and property loss or assets. Life and wellbeing are the greatest parts of people's lives. But, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using financial capital to reimburse them. Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks. Concerning the value of insurance in the lives of individuals, it becomes Important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it. Various variables estimates these charges.

Each factor of these is important. If any factor is omitted when the amounts are computed, the policy changes overall. It is therefore critical that these tasks are performed with high accuracy. As human mistakes are could occur, insurers use people with experience in this area. They also use different tools to calculate the insurance premium. Ml is beneficial here. ML may generalize the effort or method to formulate the policy. these ml models can be learned by themselves. The model is trained on insurance data from the past. The requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance policy costs.

This decreases human effort and resources and improves the company's profitability. Thus, the accuracies can be improved with ml. our objective is to forecast insurance charges in this article. The value of insurance fees is based on different variables. As a result, insurance fees are continuous values. The regression is the best choice available to full fill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used.

Pre-processing of the dataset done first. then we trained regression models with training data and finally evaluated these models based on testing data. In this article, we used several models of regression, for example, multiple linear regression, decision tree (cart), and Random Forest Regression. It is found that the Random Forest Regression provides the highest accuracy with an r-squared value of 87.07829. the key reason for this study is to include a new way of estimating insurance costs.

2.LITERATURE SURVEY

RELATED STUDY

Machine learning is helpful for a variety of situations. The prediction of dependent variable values from independent variables is one of the uses of this methodology. The objective of the study is too predictive the insurance cost based on age, BMI, child number, the region of the person living, Gender, and whether a client is smoking or not. These features contribute to our target variable prediction of insurance costs.

For the measurement of the cost of insurance, several regression models are applied in this study. The dataset is split into two sections. One part for model training and the other part for model evaluation or testing. In this study, the data set is separated into two-part the first part is called training data and the second called test data, training data makes up about 80 percent of the total data used, and the rest for test data. Every one of these models is trained with the training data part and then evaluated with the test data. And we used Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating these models.

SYSTEM STUDY

For the measurement of the cost of insurance, several regression models are applied in this study. The dataset is split into two sections. One part for model training and the other part for model evaluation or testing. In this study, the data set is separated into two-part the first part is called training data and the second called test data, training data makes up about 80 percent of the total data used, and the rest for test data. Every one of these models is trained with the training data part and then evaluated with the test data. And we used Mean absolute error (MAE), root mean squared error (RMSE) and R-squared As a standard for evaluating these models.

3.IMPLIMENTATION

MODULES

Segregating the project into different modules can be useful to achieve clarity while implementing each module. By this process we can have a clear idea and when we have changes in a particular module it stays in the respective module and the process can be easy.

DATA COLLECTION

We have collected data from Kaggle website. We have 6 attributes which will help us to prepare a model for predicting the insurance amount. It has a total of 1338 data information. Given below are the input and output features of the data.

Input Features

1. Age: Age in years
2. Gender: Gender (1 = male; 0 = female)
3. BMI: Body Mass Index, objective index of body weight (kg/ m^2) using the ratio of height to weight

4. Children: Number of children

1. Value between 0-5
2. Smoker: Smoking or not(1=yes,0=no)
3. Region: Area they belong to.
4. Value 1: southwest
5. Value 2: southeast
6. Value 3: northwest
7. Value 4: northwest

Output Features

• The output will be the prediction value.

The data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model.

4.DATA PREPROCESSING

Data pre-processing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Preprocessing for this approach takes 4 simple yet effective steps. They are:

1.Attribute selection

2. Cleaning missing values

3. Training and Test data

4. Feature Scaling

5. Evaluating the data

- Age Numeric [18 to 64; mean=39.207025]
- BMI Numeric [15.96 to 53.13; mean=30.66]
- Gender Male or Female
- Children Numeric [0 to 5; mean=1.0949]
- Smoker Yes or no
- Region southwest, southeast, northwest, or northeast
- Charges (target variable) Numeric [1121.873 to 63770.428; mean=13270.422]

5.CONCLUSION

The research uses various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle.com. The Random Forest Regression offers the best efficiency, with an RMSE value of 0.07291, an MAE value of 0.036628, and an accuracy of 87.078291.RandomForest Regression can therefore be used in the estimation of insurance costs with better performance than other regression models. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data.

6.FUTURE SCOPE

In the future we can easily extend the model by using different machine learning models. Not only the machine learning models we can use other prediction models also. In the present model we have usen only a few attributes. We can also increase change the attributes and extend the project.

7.REFERENCES

1. Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.

2. Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus

3. logistic regression. Risks, 7(2), 70

4. Ms.sonal patil, Mr.Mayur Agrawal, Ms.Vijaya R. Baviskar "Efficient Processing of Decision Tree using ID3 & improved C4.5 Algorithm", International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1956-1961.

5. Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.

6. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

7. Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

8. Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.

9. Kayri, M., Kayri, I., & Gencoglu, M. T. (2017, June). The performance comparison of multiple linear regression, random forest by using photovoltaic and atmospheric data. In 2017 14th International Conference on Engineering of Modern Electric Systems (EMES) (pp. 1-4). IEEE.

10. Denuit, Michel & Hainaut, Donatien & Trufin, Julien. (2019). Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions. 10.1007/978-3-030-25820-7.

11. Breiman, Leo. 2001. ―Random Forests.‖ Machine Learning 45 (1). Springer: 5–32.

12. Aler, R., Galván, I.M., Ruiz-Arias, J.A., Gueymard, C.A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. In Solar Energy vol. 150, pp. 558-569.

13. Volkovs, M., Yu, G. W., & Poutanen, T. (2017). Content-based neighbor models for cold start in recommender systems. In Proceedings of the Recommender Systems Challenge 2017 (pp. 1-6).