HEALTH INSURANCE PREMIUM PREDICTION USING BLOCKCHAIN TECHNOLOGY AND RANDOM FOREST REGRESSION ALGORITHMS

Ghosh Madhumita [1], Ravi Gor [2]

[1] Research Scholar, Department of Mathematics, Gujarat University, Ahmedabad-380009, India
[2] Department of Mathematics, Gujarat University, Ahmedabad-380009, India

Received 23 April 2022
Accepted 28 May 2022
Published 24 June 2022

Corresponding Author

Ghosh Madhumita,

madhumitaghoshfmg@gujaratuniversity.ac.in

DOI 10.29121/IJOEST.v6.i3.2022.346

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Blockchain technology is based on a sequence of blocks, where each block carries a certain amount of information. Medical records can be cryptographically secured in the health insurance ecosystem with blockchain technology. Here, blockchain technology model is used to create a user interface for storing data block wise. Also, Insurance premium is predicted using Support Vector Regression, Lasso Regression, Ridge Regression, Multiple Linear Regression and Random Forest Regression algorithms. Out of all these algorithms, Random Forest Regression algorithm gives the better result.

Keywords: Blockchain Technology, Supervised Learning, Uniform Resource Locator, Support Vector Regression, Lasso Regression, Ridge Regression, Multiple Linear Regression, Random Forest Regression, Health Insurance Data

1. INTRODUCTION

Due to the competitive environment, people are constantly under stress and suffer from physical and mental health problems. Therefore, it is important to purchase adequate health insurance plans for the treatment of physical and mental illness. Financial challenges can be avoided if a person has a health insurance policy at the time of medical treatment. So, today more and more people are realizing the importance of health insurance and opt to buy it.

How to cite this article (APA): Madhumita, G., and Gor, R. (2022). Health Insurance Premium Prediction Using Blockchain Technology and Random Forest Regression Algorithm. *International Journal of Engineering Science Technologies, 6*(3), 74-82. doi: 10.29121/IJOEST.v6.i3.2022.346

Now a day, machine learning and blockchain technology are used to run automated applications in every sector of the insurance industry. Machine learning and blockchain technologies can effectively store the medical history of patient, increase the access of medical records, increase the delivery of services related to patient care by efficiently designing effective algorithms.

Most important benefit of blockchain technology in the insurance industry is that, because of the immutable nature of the contracts stored on the blockchain, there is no need for any third party to act as an intermediary. Insurance companies can store data such as transactions and claims in a secure manner using blockchain technology that is virtually invulnerable. With the help of blockchain technology the insurance company can reduce the human error that occurs when updating records. excellarate (2021)

Machine learning is widely used across the insurance industry. It helps insurance companies in fraud detection, claim management, billing, and customer service (azure.microsoft.com). Many models are created using machine learning techniques to enhance customer service. These models are used for everything from claim registration to claim settlement. In addition, predictive model can be used to predict future claims and costs associated with claims accenture (n.d.)

2. LITERATURE REVIEW

Shreyas et al. (2016) implemented thirteen regression algorithms to predict the popularity of online articles. They compared the results obtained by all algorithms and then identified the top five algorithms such as Random Forest, Linear regression, Lasso, Ridge, and Nearest Neighbor giving the best results. These five models were identified based on the R2 score. Out of these five algorithms Random Forest Regression predicts better result with an accuracy of 88.8%. Shreyas et al. (2016)

Gururaj et al. (2019) used Linear Regression (LR) and Support Vector Machine (SVM) algorithms to predict the stock price. They compared both the algorithms and concluded that SVM performs better than LR. Both algorithms are compared by calculating MSE, MAE, RMSE and R-Squared. They also explained advantages and disadvantages of SVM and LR. Gururaj et al. (2019)

D'Costa et al. (2020) predicted the true value of cars by using machine learning algorithms. They divided the data into two parts training and testing. They applied Multiple Linear Regression algorithm to train and test the data. In this mode car model, fuel type, emission, mileage, and year of registration are taken as independent variables and car price taken as dependent variable. D'Costa et al. (2020)

Bajaj et al. (2020) used Machine Learning models such as Linear Regression, K-Neighbours Regressor, XGBoost Regressor, and Random Forest Regressor to forecast future sales of Big Mart Companies based on previous year's sales. The input criteria for the prediction are item weight, fat content, visibility, item kind, MRP, outlet establishment year, outlet size, and outlet location type. They also determine the precision of outcomes by calculating Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracy and concluded that the Random Forest Algorithm is the best of all, with a precision of 93.53%. Bajaj et al. (2020)

Venkatesan et al. (2020) applied linear regression to discover the best strawberry growth production with the optimum water. The data was collected from September 2015 to May 2016. The 233 valid samples were separated into two groups: training and testing. 186 training sets are utilised in modelling, whereas 46 test sets are used to evaluate the model's prediction performance. Nutrition water, average temperature, humidity, and CO2 are used as independent variable. They also calculate R-squared, Root Mean Squared Error (RMSE) and P-value. Venkatesan et al. (2020)

Vali et al. (2020) studied the previous year's sales of a supermarket to predict their future sales. The Linear Regression Algorithm is applied for the prediction. Sales data from 2017 to 2019 are taken for this model. Data from 2017 and 2018 is used for training purposes, and data for 2019 is predicted. To calculate the accuracy of prediction, the actual data for the year 2019 was compared with the predicted data. Vali et al. (2020)

Rohith et al. (2020) used Decision Tree regression algorithm to predict crop price. Rainfall, Minimum Support Price, and Wholesale Price Index are taken as independent variables. The max depth parameter is used to reduce the complexity of model and size of tree. Also, Flask module is used to forecast the crop price through web application. Mean Square Error, Mean Absolute Error, and R-score are calculated to measure the performance of decision tree regression. Rohith et al. (2020)

Kausthub (2021) applied Multiple Linear Regression algorithm to predict sales related to commercials which were displayed in mainly three forms of media TV, Radio & Newspaper. He also noted the error prediction value by using yellow-brick library. He also calculated RMSE value to check the accuracy of model. Kausthub (2021)

Zhang (2021) used two methods to predict housing price. He analysed significant factors for input variable affecting on house prices. These factors were selected by using Spearman correlation coefficient method. Then Multiple Linear Regression model for housing price prediction was established. Also, conclude that Multiple Linear Regression model effectively predicts and analyse the housing price to some extent. Zhang (2021)

Dabreo et al. (2021) predicted Real estate prices using XGBoost, Random Forest, Decision Tree, and Linear Regression algorithms. Thirteen independent variables are used, such as the fraction of residential land zoned for plots larger than 25,000 square feet, the proportion of non-retail commercial acres per town, the full-value property-tax rate per $10,000, and so on. They also calculated the Root Mean Squared Error mean, Root Mean Squared Error standard deviation, and Mean Cross Validation Score to evaluate the model's performance. The XGBoost Regression machine learning algorithm came in first, followed by the Random Forest regression algorithm, and the Decision Tree came in third with a significant difference. Dabreo et al. (2021)

Bhavsar and Gor (2022) predicted restaurant ratings with the help of Machine Learning Model. Information such as Restaurant id, Country, categories for dining, cost, currency, online delivery option, aggregate rating, rating, votes were provided to the Artificial Neural Network model. The ratings were classified in 5 different categories form poor to Excellent. Results of three different optimizers Adam, Adamax and Nadam were compared, where Nadam shows best accuracy. Bhavsar and Gor (2022), Ghosh and Gor (2022)

3. REPRESENTATION OF BLOCKCHAIN TECHNOLOGY AND RANDOM FOREST REGRESSION MODEL

3.1. BLOCKCHAIN

Blockchain technology is defined as a chain of blocks that contains information It is difficult to change, hack or cheat information which is stored in a blockchain system excellarate (2021) Blockchain collects sets of information in a group, known as blocks (geeksforgeeks.org). Blocks have a specific storage capacity, when they are filled; they are closed and attached to a previously filled block, which forms a chain of data known as a blockchain.

Figure 1

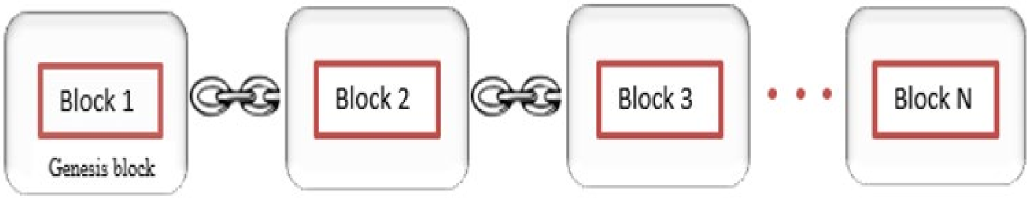


Figure 1 Structure of Blockchain

In a blockchain the first block is called the Genesis block. Each new block is linked to the previous block excellarate (2021)

3.2. REGRESSION

Regression is a supervised learning technique used when the output variable has a real or constant value, such as salary, weight etc Seely (2018) There are many algorithms are used to solve the classification and regression problem both Seely (2018) Here, Random Forest algorithm is used to solve regression problem.

3.3. RANDOM FOREST

Random Forest (RF) algorithm can be used for classification and regression both. This algorithm is an ensemble of decision tree. The prediction of the random forest is based on the predictions of each individual tree Vali et al. (2020) In this paper Random Forest regression is applied to predict the future sales.

Steps involved in Random Forest Regression: Biau (2012)

Step-1 Select the sample randomly from the training data set.

Step-2 Apply the decision tree algorithm individually on the collected sample.

Step-3 Calculation of decision tree. Coursera (n.d.)

1) Start with the root node, which contains the complete data set.
2) Find the best attribute using Attribute Selection Measure (ASM).

Two popular techniques for ASM

Information Gain:

$$Information\ Gain = Entropy(s) - [(Weighted\ Avg) * Entropy(each\ Feature)]$$

Where, s = total number of samples

$$Entropy = -\sum_{i=1}^{n} p_i * \log{(p_i)}$$

Gini Index:

$$Gini\ index = 1 - \sum_{i=1}^{n} (p_i{}^2)$$

3) Divide the root node into subsets that contain possible values for the best attributes.
4) Generate the decision tree node, which contains the best attribute.
5) Recursively make new decision trees using the subsets of the dataset created in step-iii. Continue this process until a stage is reached where you cannot further classify the nodes.

Step-4 Calculate the average of the predictions made by output of the individual decision tress.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t$$

where, T = decision trees in the Random Forest,

yˆ_t = predictions made by each decision tree.

4. METHODOLOGY USED IN THE PAPER

The data of Health insurance has been collected from Kaggle is shown in Table 1 In this model age, sex, BMI, and children are taken as an independent variable to predict an insurance premium.

Table 1

Table 1 Health Insurance dataset (kaggle.com)

Age	Sex	Bmi	Children	Smoker	Region	Expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southwest	1725.55
28	male	33	3	no	southwest	4449.46
33	male	22.7	0	no	northwest	21984.46
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southwest	3756.62
46	female	33.4	1	no	southwest	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northwest	6406.41
60	female	25.8	0	no	northwest	28923.14
25	male	26.2	0	no	northwest	2721.32
62	female	26.3	0	yes	southwest	27808.73
23	male	34.4	0	no	southwest	1826.84
56	female	39.8	0	no	southwest	11090.72
27	male	42.1	0	yes	southwest	39611.76
19	male	24.6	1	no	southwest	1837.24
52	female	30.8	1	no	northwest	10797.34
23	male	23.8	0	no	northwest	2395.17
56	male	40.3	0	no	southwest	10602.39
30	male	35.3	0	yes	southwest	36837.47

Age: age of person having policy

Sex: gender of person having policy (female=0, male=1) (kaggle.com)

BMI: index of body weight (*kg/m^2*) using the ratio of height to weight, ideally 18.5 to 25 (kaggle.com)

Children: Total number of children (kaggle.com)

Charges: medical costs billed by health insurance for particular individual

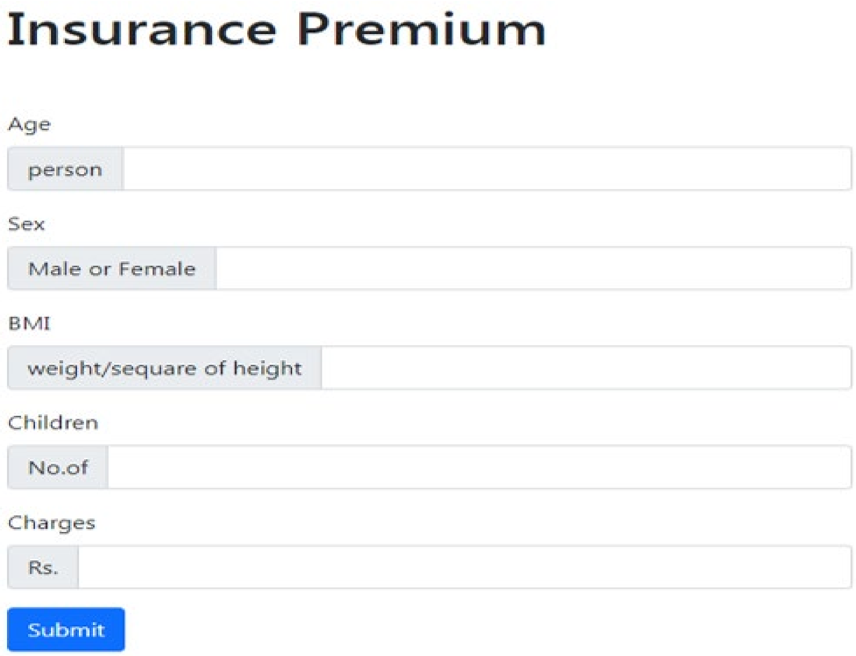
Here, two different techniques Blockchain and Regression are used to store and predict health insurance premium.

First user interface (web page) is created by using blockchain technology. With the help of webpage, data can be stored block wise, and this data cannot be hacked by anyone. This data is secured in blockchain. Then different machine learning algorithms are applied on this data to predict the premium of health insurance and all the results obtained by algorithms are compared. In these models age, sex, BMI, and children taken as an independent variable to predict charges of health insurance. 70% of data are used for training and 30% of data are used for testing purpose Bhavsar and Gor (2022)

5. RESULT AND DISCUSSION

Information is filled into the user interface and submitted. After submission the block is created with the hash function. In such a way blocks are connected one by one and create a blockchain.

Figure 2



Insurance Premium

Age

person

Sex

Male or Female

BMI

weight/sequare of height

Children

No.of

Charges

Rs.

Figure 2 User Interface of Insurance Premium

Data set cannot be taken directly for regression because of null values and some attributes which have unnecessary information, or which decreases the accuracy of model Bhavsar and Gor (2022) To improve the accuracy and speed of model we have to remove the null values and attributes like region and smoker parameters from dataset.

After cleaning the data, correlation is checked between the variables. The RBF kernel is employed among all kernels in the Support Vector Regression algorithm. The maximum depth 5 and random state 13 are used in the Random Forest Regression. Then, the premium is predicted by using Support Vector Regression, Lasso Regression, Ridge Regression, Multiple Linear Regression and Random Forest Regression. Among these algorithms Random Forest Repression gives better result.

Table 2

Table 2 Mean Absolute Error and R-Squared of regression algorithms

Algorithms	Mean Absolute Error	R-squared
Support Vector Regression	8817.232	0.1146
Lasso Regression	8858.1903	0.1146
Ridge Regression	8858.8802	-0.0668
Multiple Linear Regression	4307.8631	0.7991
Random Forest Regression	2444.9258	0.8954

Consequently, the error between actual charges and predicted charges are calculated by Mean Absolute Error (MAE) method. To check the accuracy of model R-squared value is also calculated. The performance of various methods based on Mean Absolute Error and R-squared value is depicted in Figure 3 and Figure 4 The R-squared value obtained by Random Forest Regression.

Figure 3

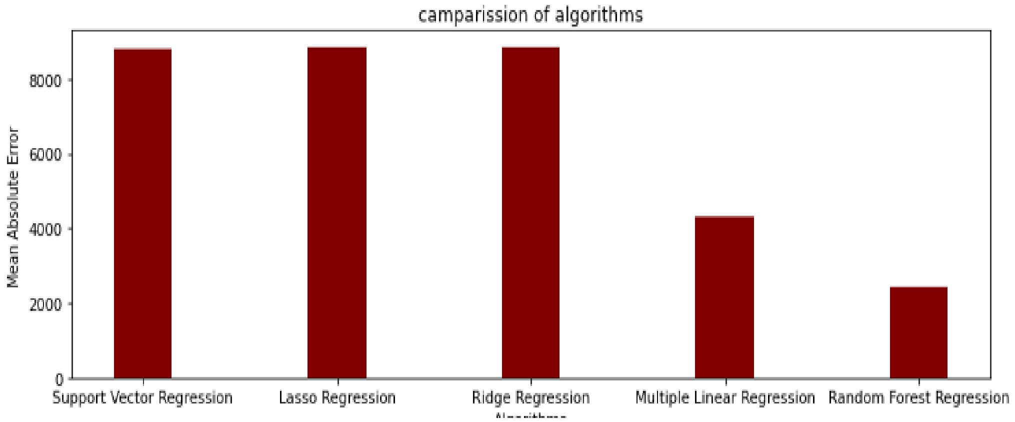


Figure 3 Mean squared error graph

Figure 4

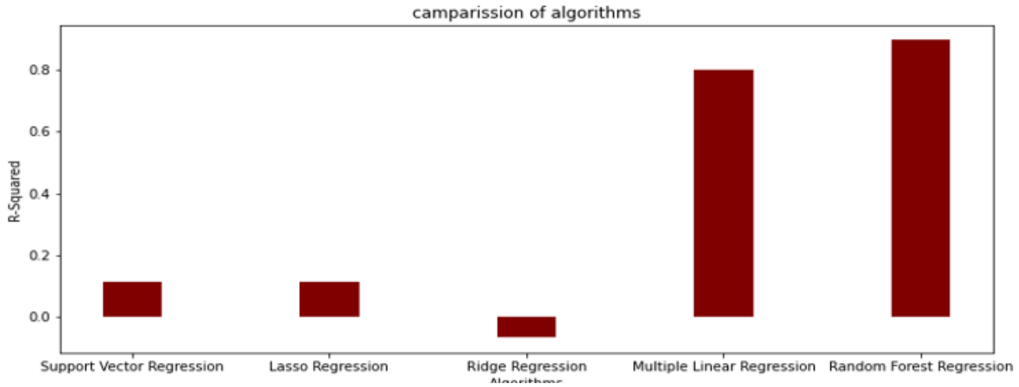


Figure 4 R-Squared error graph

6. CONCLUSION

The health insurance premium data is stored and predicted by using Blockchain Technology and Random Forest Regression algorithm. Here, five supervised learning regression-based algorithms were used to predict the premium, namely Support Vector Regression, Lasso Regression, Ridge Regression, Multiple Linear Regression and Random Forest Regression. For building the model four features have been taken which affect the premium price. The performance of five algorithms has been calculated in terms of MAE and R-Squared. Hence, the compared result concludes that Random Forest Regressor gives better result.

In future, this type of problems can be solved with other supervised learning techniques.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

accenture. (n.d.). Machine Learning in Insurance

Bajaj, P. Ray, R. Shedge, S. Vidhate, S. & Nilkumar. (2020). Sales Prediction using Machine Learning Algorithms. International Research Journal of Engineering and Technology, 7(6).

Bhavsar, S. & Gor, R. (2022). Predicting Restaurant Ratings using Back Propagation Algorithm. International Organization of Scientific Research Journal of Applied Mathematics (IOSR-JM), 18(2), 5-9.

Biau, G. (2012). Analysis of a Random Forests Model. Journal of Machine Learning Research, 13, 1063-1095.

Coursera. (n.d.). Medical Insurance Premium Prediction with Machine Learning.

D'Costa, L. D'Souza, A. Abhijith, k. & Varghese, D. (2020). Predicting True Value of Used Car using Multiple Linear Regression Model. International Journal of Recent Technology and Engineering, 8(5). https://doi.org/10.35940/ijrte.E1010.0285S20

Dabreo, S. Rodrigues, S. Rodrigues, V. & Shah, P. (2021). Real Estate Price Prediction. International Journal of Engineering Research & Technology, 10(4).

David, D. (2020). Random Forest Classifier Tutorial : How to Use Tree-Based Algorithms for Machine Learning.

excellarate. (2021). Blockchain in Insurance : Use Cases and the Way Forward.

Ghosh, M. & Gor, R. (2022). Short Message Service Classifier Application using Naïve Bayes algorithm (In Press). International Organization of Scientic Research Journal of Computer Engineering (IOSR-JCE).

Gururaj, V. Shriyaand, V. & Ashwini, K. (2019). Stock Market Prediction using Linear Regression and Support Vector Machines. International Journal of Applied Engineering Research, 14, 1931-1934.

Hanafy, M. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models.

javatpoint. (2022). Decision Tree Classification Algorithm.

kaggle. (n.d.). Medical Cost Personal Datasets.

Kharwal, A. (2021). Health Insurance Premium Prediction with Machine Learning.

Kausthub, K. (2021). Commercials Sales Prediction Using Multiple Linear Regression. International Research Journal of Engineering and Technology, 8(3).

Rohith, R. Vishnu, R. Kishore, A. & Chakkarawarthi, D. (2020). Crop Price Prediction and Forecasting system using Supervised Machine Learning Algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 9(3).

Ronaghan, S. (2018). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark.

Seely, S. (2018). Eight use cases for machine learning in insurance.

Shreyas, R. Akshata, D. Mahanand, B. Shagun, B. & Abhishek, C. (2016). Predicting Popularity of Online Articles using Random Forest Regression. Institute of Electrical and Electronics Engineers, 1-5. https://doi.org/10.1109/CCIP.2016.7802890

Vali, M. Sankeerthana, K. Naveen, B. & Vishal, N. (2020). Prediction of Online Sales using Linear Regression. International Journal of Creative Research Thoughts, 8(2).

Venkatesan, S. Sathishkumar, V. Park, J. Shin, C. & Cho, Y. (2020). A Prediction of Nutrition Water for Strawberry Production using Linear Regression, International Journal of Advanced Smart Convergence. 132-140.

Zhang, Q. (2021). Housing Price Prediction Based on Multiple Linear Regression. Hindawi Scientific Programming, (3), 1-9.