A Blockchain and Machine Learning based Framework for Efficient Health Insurance Management

Adit Goyal
Dept. of CSE and IT, JIIT Noida
Noida, India
aditgoyal@hotmail.com

Anubhav Elhence
Dept. of EEE, BITS Pilani
Pilani, India
elhenceanubhav@gmail.com

Vinay Chamola
Dept. of EEE & APPCAIR, BITS Pilani
Pilani, India
vinay.chamola@pilani.bits-pilani.ac.in

Biplab Sikdar
Dept. of ECE, National University of Singapore
Singapore
bsikdar@nus.edu.sg

ABSTRACT

Having a health insurance is important for everybody, bearing in mind the increasing medical costs. Medical emergencies can have a severe financial and emotional impact. However, the current insurance system is very expensive and the claim settlement process is excessively lengthy, making it tedious. This results in policyholders not being able to successfully make a claim with their insurance company. In this paper, we focus on developing a fast and cost-effective framework based on blockchain technology and machine learning for the health insurance industry. Blockchain is capable of removing all third-party organisations by forming a smart contract, making the entire process more smooth, secure, and efficient. The contract settles the claim on documents submitted by the claimant. A ridge regression model is used for computing the premiums optimally, based on the total amount claimed under the current policy tenure, along with several other factors. A random forest classifier is applied for predicting the risk that helps in the computation of risk-rated premium rebate.

CCS CONCEPTS

• **Applied computing** → **Electronic funds transfer**; *E-commerce infrastructure*; • **Computing methodologies** → **Classification and regression trees**.

KEYWORDS

Insurance, Blockchain, Smart contract, Machine Learning, Random Forest

ACM Reference Format:
Adit Goyal, Anubhav Elhence, Vinay Chamola, and Biplab Sikdar. 2021. A Blockchain and Machine Learning based Framework for Efficient Health Insurance Management. In *The 19th ACM Conference on Embedded Networked Sensor Systems (SenSys '21), November 15–17, 2021, Coimbra, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3485730.3493685

SenSys '21, November 15–17, 2021, Coimbra, Portugal

ACM ISBN 978-1-4503-9097-2/21/11…$15.00
https://doi.org/10.1145/3485730.3493685

1 INTRODUCTION

With the advent of COVID-19 and various other rising instances of diseases, the cost of health care services has been increasing at an alarming rate. In the current insurance system, the small and consistent premium has taken a price-hike. The insurance industries spend a lot of money on administrative and marketing which comprises of commission to brokers, advertisements, brochures, etc., which are eventually recovered through the insurance premiums. There are a lot of issues in the classical insurance process. For instance, the insurer has to complete a lot of documentation and provide proof of the loss value. The broker causes more delays and expenditures. Knowledge sources are diversified and inconsistent, such that the insurer has to seek additional information. In this paper, we propose a blockchain-based solution to the above problems to provide health insurance services that can be cost-effective as well as efficient.

This paper aims to exploit blockchain technology to find solutions to our problem because of the various key features it pertains to by the use of smart contracts, automating and speeding up the tasks of client registration, policy issuance and claim processing. It makes use of a decentralised, distributed ledger that is capable of eliminating all third-party authentication. Fraud detection becomes easier as it is immutable and transparent. A transparent ledger of changes preserves the integrity of the document, which creates trust in the asset. The major contributions of our work are listed below:

- We propose a peer-to-peer network of hospitals and policyholders that will be bound in a smart contract. The smart contracts will also hold the claim histories of each user for the premium calculation of the next term.
- Ridge Regression is used to predict the premium. Claims are sent directly by the admitting hospital, which will be approved by the smart contract.
- Risk is predicted with the help of a random forest classifier. If the amount in the contract is sufficient for predicted claims, then a rebate may be given to all policyholders based on their predicted risk.

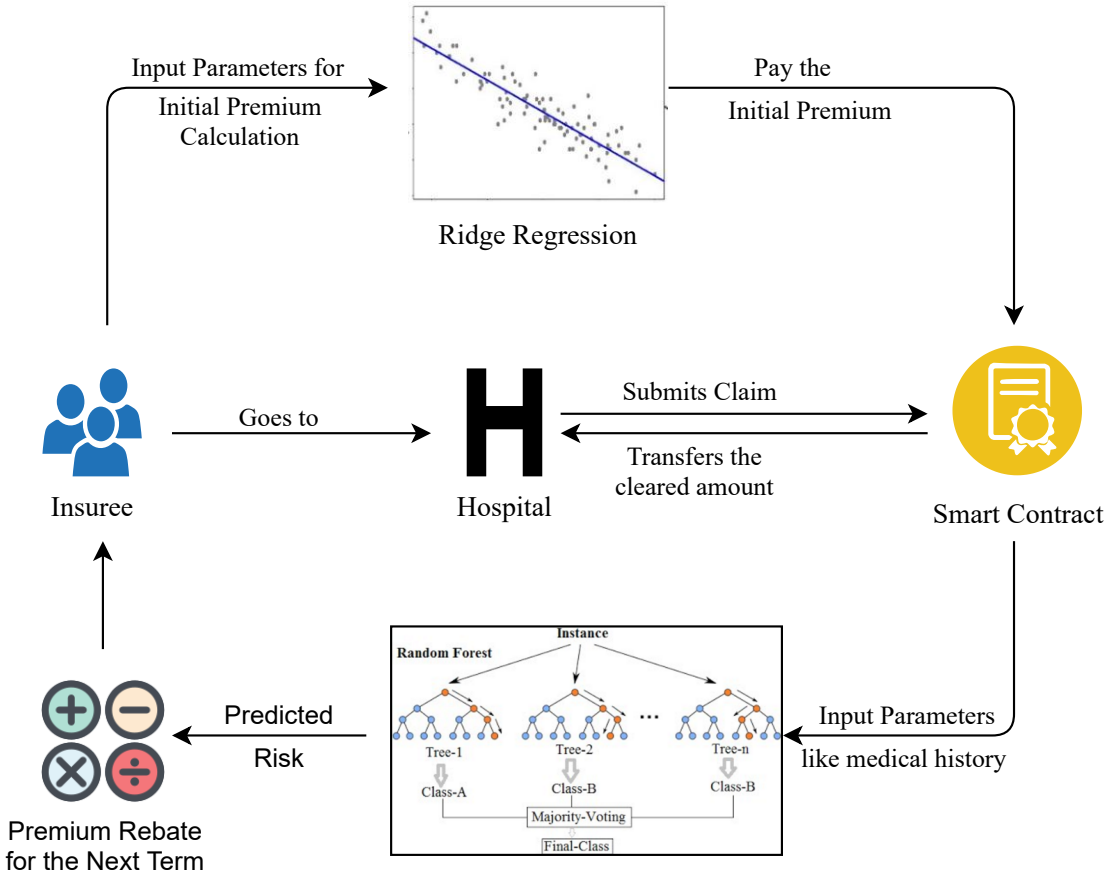

Figure 1: Flowchart of the proposed model.

- Furthermore, a risk-rated premium rebate model for premium calculation of the next term has also been proposed.
- Experimental results show that the proposed method is cost-effective and fast in comparison to the existing traditional methods.

2 RELATED WORK

In 2017, Etherisc, the German insurance company, collaborated in building an open-source framework focused on decentralized insurance applications: the Decentralized Insurance Framework [5]. Beenest and WeTrust are two US insurance companies that, by the end of 2017, developed a blockchain-enabled homeowners insurance system [16]. Raikwar and Mazumdar discussed exploring blockchain to remove the third-party organisation and automating insurance processes [18]. Peer-to-peer blockchain model has been used in various fields such as traffic jam management [8], biometric and security mechanisms in parking system [21], data marketplaces [22], data offloading [9] and smart cities [1, 7, 11].

Predictive modelling has been used for premium calculation [3], and risk classification [20]. Regression has also been applied in the economic and financial fields, and several researchers use the power of regression in their decision-support processes [19, 23]. Risk premium functions and insurance pricing models have been proposed in various researches that have helped to make the system more accurate [12, 14].

For risk classification, ensemble methods like XGBoost, random forest or decision trees have been used in several frameworks because of the good results they deliver on multi-class classification [4, 15]. The authors of [13] used an ensemble random forest classifier for insurance data analysis, and the proposed framework outperformed other classification techniques like SVM (Support Vector Machine).

3 PROPOSED MODEL

Consider a person P who registers into the blockchain network with the required details. A regression model is applied to predict the premium for the health insurance policy. When P get admitted into a hospital H, H completes all the documentation for claim processing and automated claim settling is done. On completion of a term of the policy, a random forest classifier model is applied to calculate the risk of an individual, which is passed as a parameter to the risk-rated premium rebate mathematical model to calculate the new premium. The flowchart of the proposed model is shown in Figure 1.

Table 1: Input and output parameters in the dataset for premium prediction.

Input Parameters	**Description**
Age	Policyholder's Age.
Sex	Gender of Policyholder (male/female).
BMI	Body mass index (ratio of height to weight).
Children	Number of dependants.
Smoker	Whether a person has a smoking history or not (yes/no).
Region	Residential region of policyholder.
Output Parameter	**Description**
Premium Amount	Insurance premium for the policy.

3.1 Ridge Regression for Premium Prediction

In order to help the algorithm converge to linearly separable data and minimise overfitting, L_2 regularisation adds a penalising term to the squared error cost function. The cost function to be minimized with respect to a parameter, θ, is:

$$J(\theta) = || \ y - X\theta \ ||_2^2 + \alpha || \ \theta \ ||_2^2 \tag{1}$$

where X is a design matrix and α is a hyperparameter. Since a high variance was observed in the baseline model, ridge regression was considered to be appropriate.

The dataset being used here is the US Health Insurance dataset with 1338 rows of insured data and no missing or undefined values in the dataset [2]. For the initial prepossessing, we checked each feature of the dataset to convert categorical features into integer encoded vectors and boolean features into binary. The feature vector comprises of 6 elements after pre-processing. Moreover, the features were normalized to bring all the data to the same range. The details of input and output parameters are given in Table 1.

3.2 Random Forest Classification for Risk Calculation

On completion of a term, the risk of the policyholder is evaluated based on various parameters such as family history, insurance history, employment history and medical history (Table 2). A risk factor of an individual insured is calculated to provide a rebate in the next term's premium.

The dataset being used here is the Prudential Life Insurance Dataset with 59381 rows of data [17]. The data is split into training and testing sets, each with 53442 and 5939 rows, respectively. For pre-processing, we eliminated the features that were used solely to identify the beneficiaries. Categorical features were factorized to be represented in a numeric vector. Moreover, a new feature containing the number of medical keywords was introduced.

3.3 Risk-Rated Premium Rebate Model for Premium calculation of the next term

We use the concept of risk-rated (weighted) premium calculation to compute our rebate based on premium [6]. $\mathcal{P}_x$ is the premium of the current term, and the premium for the next term $E[x]$ has to be calculated. Risk x is covered under the medical policy, the claims that we get from x are distributed as a random variable, and its cumulative distribution function is given by $F(x)$.

Let the current total amount in the smart contract be C. On the scale of $1-8$, a risk factor equal to 1 corresponds to the claims of the full amount insured (let the amount insured be $\mathcal{I}$) and risk factor 8 corresponding to no claim. The total amount spent on expected claims would be $\mathcal{T}$:

$$\mathcal{T} = \sum_{i=1}^{n} \frac{8 - r_i}{7} \mathcal{I}, \tag{2}$$

where n is the number of policyholders registered in the contract, and r_i is the risk associated with individual i. Next, we calculate the extra revenue (ER) that is being collected in the contract:

$$ER = C + Amt - \mathcal{T}, \tag{3}$$

where Amt is the total amount collected from premiums if no rebate is given to those particular individuals whose term of policy needs to be re-insured for the current scenario. The risk-rated premium is calculated for each individual, and then a rebate is provided from the extra revenue (ER) being collected from the policyholders. The risk-adjusted premium based on Esscher's principle is given by [10]:

$$\mathcal{P}_x = \int_0^\infty [1 - F(x)]^{1/r} dx. \tag{4}$$

The risk-rated premium rebate is:

$$\mathcal{P}_r = \mathcal{P}_x - \frac{r}{R} ER. \tag{5}$$

This will be shown on the dashboard of the policyholder to continue their policy.

4 NUMERICAL ANALYSIS AND RESULTS

4.1 Simulation Settings

Blockchain is used to make the peer-to-peer network, and a smart contract is deployed on the blockchain network as the first node that cannot be changed. Smart contracts are written in python language using the SmartPy library, which can be deployed on the Ethereum network. To improve the performance of the ridge regression model, we used the polynomial features of $degree = 7$ and took the value of α to be 3 to obtain the best results. The training and testing set was split into a 3:1 ratio for ridge regression.

In the risk calculation model, the training and testing sets are randomly split in a 9:1 ratio. The number of trees in the forest is set as 50, the number of features to consider for splitting is 75, the minimum number of samples at the leaf node is 8, the maximum depth of the tree is 50, the minimum number of samples required to split an internal node is 40, and bootstrap is kept as False. All the other parameters are kept at their default value.

Table 2: Input and Output parameters in the dataset for risk prediction.

Input Parameters	**Description**
Age	Age of the policyholder.
Height	Height of the policyholder
Weight	Weight of the policyholder
Body Mass Index (BMI)	BMI of the policyholder
Employment Information	Set of normalised variables pertaining to employment history
Insured Information	Set of normalised variables pertaining to current insurance information
Insurance History	Set of normalised variables pertaining to previous insurance information
Family History	Set of normalised variables relating to family history of the policyholder
Medical History	Set of normalised variables relating to medical history of the policyholder
Medical Keywords	Set of variables related to presence or absence of a particular disease
Output Parameter	**Description**
Risk Factor	Target variable that predicts the risk on a scale of $1-8$.

Table 3: Comparison of classification models for calculating the risk.

Model	Accuracy Score • Training set • Testing set	Root Mean Square Error • Training set • Testing set	Mean Absolute Error • Training set • Testing set
XGBoost	0.28 0.27	2.5 2.63	2.4 2.18
Random Forest	0.71 0.59	2.03 2.26	0.92 1.19

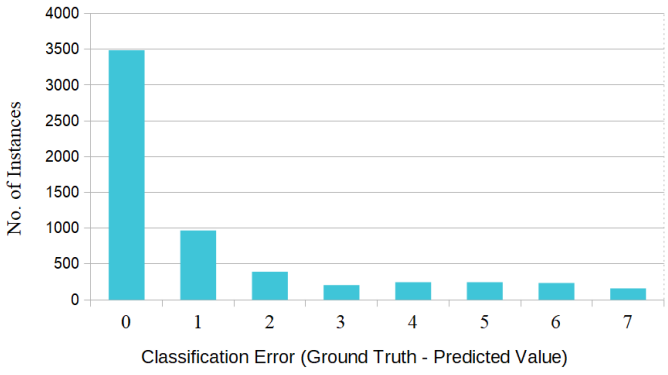

Figure 2: Histogram density plots to estimate the difference between predicted and actual values on the testing set for the random forest classifier.

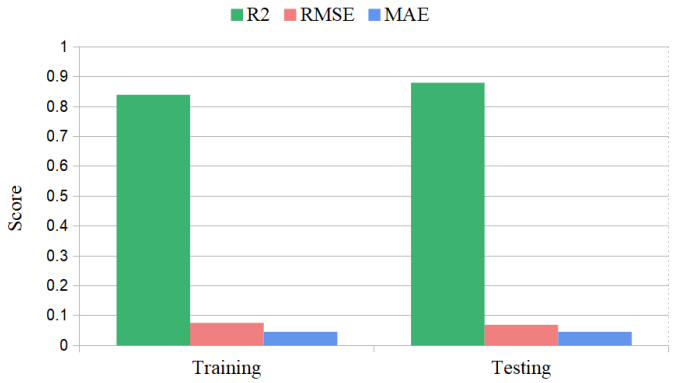

Figure 3: Performance Metrics of the Ridge Regression Model.

4.2 Performance Evaluation

The histogram density plot for the random forest classification is shown in Figure 2. For premium prediction, the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R^2 scores have been calculated. The above-mentioned values have been reported for the training and testing sets in Figure 3. Out of the various regression models tested, ridge regression with the use of interaction terms gave the least RMSE and the highest R^2 value. In the risk classification problem, we calculate the accuracy, RMSE, and MAE to assess the performance of our model, as shown in Table 3 (all values were calculated on results in the range [1, 8]). The low accuracy is attributed to a large number of classes. The predicted classes fall in close range of the actual value (Figure 2), which is shown by the low MAE.

5 CONCLUSION

The two major problems in the current insurance system, slow-moving and expensive has been solved in the proposed blockchain-based health insurance model. The results show that our proposed model is reliable, cost-efficient and fast. The whole procedure of claim processing with a fewer number of mediators and smart

contracts helps to make the procedure fast, and risk-rated premium rebates help to motivate people with lesser claims to continue with their policy. It also stopped the extra revenue being generated in the current insurance system by distributing them in rebates rightly based on the risk. Results show that out of the various multi-class classification machine learning techniques, random forest classifier gave more accurate results.

6 ACKNOWLEDGMENT

This work was supported by SERB's ASEAN - India Collaborative R&D scheme (Project Grant File no. CRD/2020/000369).

REFERENCES

[1] Tejasvi Alladi, Vinay Chamola, Reza M. Parizi, and Kim-Kwang Raymond Choo. 2019. Blockchain Applications for Industry 4.0 and Industrial IoT: A Review. *IEEE Access* 7 (2019), 176935–176951. https://doi.org/10.1109/ACCESS.2019.2956748

[2] Anirban Datta. [n.d.]. US Health Insurance Dataset. https://www.kaggle.com/teertha/ushealthinsurancedataset accessed 31 July 2021.

[3] M. David. 2015. Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance* 20 (2015), 147–156.

[4] Kartika Chandra Dewi, Hendri Murfi, and Sarini Abdullah. 2019. Analysis Accuracy of Random Forest Model for Big Data – A Case Study of Claim Severity Prediction in Car Insurance. In *2019 5th International Conference on Science in Information Technology (ICSITech)*. 60–65. https://doi.org/10.1109/ICSITech46713.2019.8987520

[5] Etherisc. 2017. Etherisc—Decentralized Insurance Protocol to Collectively Build Insurance Products. https://etherisc.com/ accessed 09 August 2021.

[6] Edward Furman and Ričardas Zitikis. 2008. Weighted premium calculation principles. *Insurance: Mathematics and Economics* 42, 1 (2008), 459–465.

[7] Vikas Hassija, Vinay Chamola, and Sherali Zeadally. 2020. BitFund: A blockchain-based crowd funding platform for future smart and connected nation. *Sustainable Cities and Society* 60 (2020), 102145.

[8] V. Hassija, V. Gupta, S. Garg, and V. Chamola. 2020. Traffic Jam Probability Estimation Based on Blockchain and Deep Neural Networks. *IEEE Transactions on Intelligent Transportation Systems* (2020), 1–10.

[9] Vikas Hassija, Vikas Saxena, and Vinay Chamola. 2020. A mobile data offloading framework based on a combination of blockchain and virtual voting. *Software: Practice and Experience* (2020).

[10] Antonio Heras, Beatriz Balbas, and José Luis Vilar. 2012. Conditional tail expectation and premium calculation. *ASTIN Bulletin: The Journal of the IAA* 42, 1 (2012), 325–342.

[11] Ayten Kahya, Anusha Avyukt, Gowri S Ramachandran, and Bhaskar Krishnamachari. 2021. Blockchain-enabled Personalized Incentives for Sustainable Behavior in Smart Cities. In *2021 International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–6.

[12] N. Krečar, F. E. Benth, and A. F. Gubina. 2020. Towards Definition of the Risk Premium Function. *IEEE Transactions on Power Systems* 35, 2 (2020), 1085–1098.

[13] Weiwei Lin, Ziming Wu, Longxin Lin, Angzhan Wen, and Jin Li. 2017. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access* 5 (2017), 16568–16575. https://doi.org/10.1109/ACCESS.2017.2738069

[14] H. Mao and K. Ostaszewski. 2010. Pricing insurance contracts and determining optimal capital of insurers. In *2010 IEEE International Conference on Industrial Engineering and Engineering Management*. 1–5.

[15] Widya Fajar Mustika, Hendri Murfi, and Yekti Widyaningsih. 2019. Analysis Accuracy of XGBoost Model for Multiclass Classification - A Case Study of Applicant Level Risk Prediction for Life Insurance. In *2019 5th International Conference on Science in Information Technology (ICSITech)*. 71–77. https://doi.org/10.1109/ICSITech46713.2019.8987474

[16] GLOBE NEWSWIRE. 2018. Bee Token Partners With WeTrust To Strengthen Insurance Offering In New Home-Sharing Network Powered by Blockchain. https://www.globenewswire.com/en/news-release/2018/01/26/1313129/0/en/Bee-Token-Partners-With-WeTrust-To-Strengthen-Insurance-Offering-In-New-Home-Sharing-Network-Powered-by-Blockchain.html accessed 09 August 2021.

[17] Padraic. [n.d.]. Prudential Life Insurance Assessment. https://www.kaggle.com/c/prudential-life-insurance-assessment/data accessed 5 August 2021.

[18] M. Raikwar, S. Mazumdar, S. Ruj, S. Sen Gupta, A. Chattopadhyay, and K. Lam. 2018. A Blockchain Framework for Insurance Processes. In *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. 1–4. https://doi.org/10.1109/NTMS.2018.8328731

[19] S Salcedo-Sanz, L Cuadra, A Portilla-Figueras, S Jiménez-Fernández, and E Alexandre. 2012. A review of computational intelligence algorithms in insurance applications, Statistical and Soft Computing Approaches in Insurance Problems. *Statistical and Soft Computing Approaches in Insurance Problems* (2012), 1–50.

[20] Peng Shi, Xiaoping Feng, and Anastasia Ivantsova. 2015. Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics* 64 (2015), 417 – 428. https://doi.org/10.1016/j.insmatheco.2015.07.006

[21] A. Waheed and P. Venkata Krishna. 2020. Comparing Biometric and Blockchain Security Mechanisms in Smart Parking System. In *2020 International Conference on Inventive Computation Technologies (ICICT)*. 634–638.

[22] Ronghua Xu, Gowri Sankar Ramachandran, Yu Chen, and Bhaskar Krishnamachari. 2019. Blendsm-ddm: Blockchain-enabled secure microservices for decentralized data marketplaces. In *2019 IEEE International Smart Cities Conference (ISC2)*. IEEE, 14–17.

[23] Hujia Yu and Jiafu Wu. 2016. Real estate price prediction with regression and classification. *CS 299 (Machine Learning) Final Project Reports* (2016).