ZKG INTERNATIONAL

ISSN: 2366-1313

ADVANCED INTELLIGENCE HEALTH INSURANCE COST PREDICTION USING RANDOM FOREST

V. SAI SRINIVAS[1],D.PUSHPALATHA[2], G. SARATHKUMAR[3], CH. KAVITHA[4], D.HARSHITHKUMAR[5].

[1]Assistant Professor,CSE,Chalapathi Institute of Technology,Guntur, India
[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: In the domains of computational and applied mathematics, soft computing, fuzzy logic,and machine learning (ML) are well-known research areas. ML is one of thecomputational intelligence aspects that may address diverse difficulties in a wide rangeof applications and systems when it comes to the exploitation of historical data.Predicting medical insurance costs using ML approaches is still a problem in thehealthcare industry that requires investigation and improvement. To address this problem,a study was conducted that provides a computational intelligence approach for predictinghealthcare insurance costs using a series of machine learning algorithms.The proposed research approach uses various regression models, including LinearRegression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regress or, Multiple Linear Regression, and k-Nearest Neighbors.For this purpose, a medical insurance cost dataset was acquired from the KAGGLErepository, and machine learning methods were used to show how different regressionmodels can forecast insurance costs and to compare the models' accuracy.The results showed that the Stochastic Gradient Boosting (SGB) model outperformedthe others with a cross-validation value of 0.858 and an RMSE value of 0.340, providing86% accuracy.

1. INTRODUCTION

The goal of this project is to allows a person to get an idea about the necessary amountrequired according to their own health status. Later they can comply with any health insurancecompany and their schemes & benefits keeping in mind the predicted amount from our project.This can help a person in focusing more on the health aspect of aninsurance rather than the futilepart. Health insurance is a necessity nowadays, and almost every individual is linked with agovernment or private health insurance company. Factors determining the amount of insurancevary from company to company. Also people in rural areas are unaware of the fact that thegovernment of India provides free health insurance to those below poverty line. It is very complexmethod and some rural people either buy some private health insurance or do not invest money inhealth insurance at all. Apart from this people can be fooled easily about the amount of theinsurance and may unnecessarily buy some expensive health insurance. Our project does not givethe exact amount required for any health insurance company but gives enough idea about theamount associated with an individual for his/her own health insurance. Prediction is prematureand does not comply with any particular company so it must not

ISSN: 2366-1313

be only criteria in selection of ahealth insurance. Early health insurance amount prediction can help in better contemplation of theamount needed. Where a person can ensure that the amount he/she is going to opt is justified. Alsoit can provide an idea about gaining extra benefits from the health insurance.

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. Thesedangers contain the risk of death, health, and property loss or assets. Life and wellbeing are thegreatest parts of people's lives. But, risks cannot usually be avoided, so the world of finance hasdeveloped numerous products to shield individuals and organizations from these risks by usingfinancial capital to reimburse them. Insurance is, therefore, a policy that decreases or removesloss 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 orquantify the amount covered by this policy and the insurance charges which must be paid for it.Arious variables estimate these charges. Each factor of these is important. If any factor is omittedwhen 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 theinsurance premium. ml is beneficial here.l may generalize the effort or method to formulate thepolicy. These ml models can be learned by themselves.The model is trained on insurance datafrom the past. The requisite factors to measure the payments can then be defined as the modelinputs, Then the model can correctly anticipate insurance policy cost .This decreases human effortand resources and improves the company's profitability. Thus the accuracies can be improvedwith ml. our objective is to forecast insurance charges in this article. The value of insurance feesis based on different variables. As a result, insurance fees are continuous values. the regression isthe best choice available to fulfill our needs. We use multiple linear regression in this analysis sincethere are many independent variables used to calculate the dependent(target) variable. For this study,the dataset for cost of health insurance is used. Preprocessing of the dataset done first. Then wetrained regression models with training data and finally evaluated these models based on testingdata. in this article, we used several models of regression, for example, multiple linear regression,generalized additive model, Svm, rf, decision tree (cart), xgboost, k-nearest neighbors, stochasticgradient boosting, and deep neural network. It is found that the stochastic gradient boostingprovides the highest accuracy with an r-squared value of 85. 8295. The key reason for this studyis to include a new way of estimating insurance costs.

2. LITERATURE SURVEY

A Comprehensive Analysis of Healthcare Bigdata Management, Analyticsand Scientific Programmingby Gupta, S., & Tripathi, P

This paper is an endeavour toward comprehensive report on healthcare

ISSN: 2366-1313

bigdata.In the field of healthcare the major goal of the big data analytics is to model, predictand inference, classification, clustering, regression, and other generic approaches to beExploited.

Predicting Motor Insurance Claims Using Telematics Data—XGBoostversus Logistic Regressionby JessicaPesantez-Narvaez

This study compared the relative performances of logistic regression andXGBoost approaches for predicting the existence of accident claims using telematicsdata. These findings showed that logistic regression is a suitable model given itsinterpretability and good predictive capacity.

Automating Car Insurance Claims Using Deep Learning Technique by Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S

This paper analyses the car insurance by using Deep learning Techniques.This system takes images of the damaged car as input and gives relevant informationlike the damaged parts and provides an estimate of the extent of damage (no damage,mild or severe) to each part. This serves as a cue to then estimate the cost of repairwhich would be used in deciding insurance claim amount.

3. EXISTING SYSTEM

Health-Care insurance prediction using linear regression analysis: Linearregression is used to predict data. Linear regression is used for modeling the relationship betweena scalar dependent variable y and one or more explanatory variables denoted x. There areadvances in this field, but the limitations remain the same. Simple Linear Regression is the onewhere only one explanatory variable is used.

Disadvantages:

1. Considers only two columns of the dataset for analysis.
2. The open value and close value are considered.
3. But the accuracy given is not satisfactory.

4. PROPOSED SYSTEM

Health-Care insurance prediction using random forest regression technique:Random Forest is a Supervised Learning algorithm that uses ensemble learning method forclassification and regression. Random forest is a bagging technique and not a boosting technique.The trees in random forests are run in parallel. There is no interaction between these trees whilebuilding the trees. It operates by constructing a multitude of decision trees at training time andoutputting the class that is the mode of the classes (classification) or mean prediction (regression)of the individual trees. A random forest is a meta-estimator (i.e. it combines the result of multipleprediction) which aggregates many decision trees, with some helpful modifications. The number of features that can be split at each node is limited to some percentage of the total (which is knownas the hyper parameter). This ensures that the ensemble model does not rely too heavily on anyindividual feature, and makes fair use of all potentially predictive features. Each tree draws arandom sample from the original data set when generating its splits, adding a further element ofrandomness that prevents over fitting.

Advantages:

1. It is one of the maximum correct studying algorithms available. For many recordssets, it produces a exceedingly correct classifier.
2. It runs efficaciously on huge databases.
3. It can cope with hundreds of enter variables without variable deletion.
4. It offers estimates of what variables are crucial with inside the classification.
5.It generates an inner independent estimate of the generalization blunders because the Woodland constructing progresses.
6.It has an effective method for estimating missing data and maintains accuracy when a largeproportion of the data are missing.

ISSN: 2366-1313

5.SYSTEM ARCHITECTURE

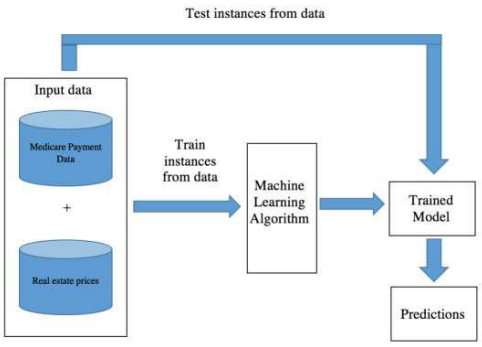

Fig 1: System architecture

6. IMPLEMENTATION

Our framework is a combination of various techniques or methods like:

Data collection, Data Preprocessing, Data partition, Select Features, Train Models, EvaluateModels, Tune Hyper parameters, Deploy Models.

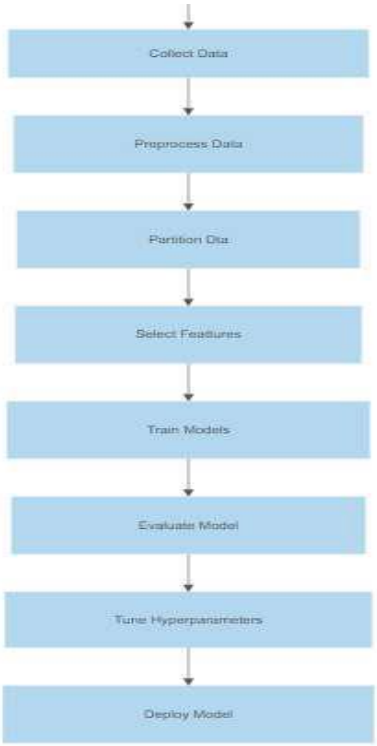

Fig.2. WorkflowDiagram

Data Collection:

This step involves gathering relevant data that will be used to build the health insurance costprediction model. The data can be obtained from various sources such as medical records, surveys,and insurance claims data. In this step, it is important to ensure that the data is relevant, complete,and accurate.

Data Preprocessing:

The collected data may not be in a suitable format for analysis. Therefore, this step involvescleaning and transforming the data to make it suitable for analysis. The data may need to be checkedfor missing values, duplicated data, and outliers. Categorical variables may also need to beconverted to numerical data using techniques such as one-hot encoding or label encoding.

Data Partitioning:

ISSN: 2366-1313

This step involves splitting the preprocessed data into two sets: the training dataset and the testingdataset. The training dataset is used to train the model, while the testing dataset is used to evaluatethe performance of the model. A typical ratio of 70:30 or 80:20 for training and testing, respectively,is often used.

Feature Selection:

In this step, the most relevant features for the model are selected. This involves analyzing thecorrelation between the features and the target variable (health insurance cost). Features thathave a strong correlation with the target variable are considered more important and areincluded in the model. However, it is important to avoid including features that are highlycorrelated with each other, as this can lead to overfitting.

Model Training:

This step involves training the Random Forest model on the training dataset using the selectedfeatures. The Random Forest algorithm is an ensemble learning method
That constructs multipledecision trees and combines their predictions to make a final prediction. The trees are constructed by randomly selecting a subset of features for each tree and splitting the data based on the selectedfeatures**.**

Model Evaluation:

Once the model is trained, it is evaluated on the testing dataset to determine its performance.This step involves calculating metrics such as mean absolute error, mean squared error, and Squared.These metrics help to determine how well the model is performing and whether it isgeneralizing well to new data.

Hyper parameter Tuning:

The performance of the Random Forest model can be further improved by tuning thehyper parameters. Hyper parameters are parameters that are set before the training process begins,and they can have a significant impact on the performance of the model. Some commonhyper parameters that can be tuned include the number of trees, maximum depth, and minimumsamples per leaf.

Model Deployment:

Once the model has been optimized and its performance has been evaluated, it can be deployedto predict the health insurance cost for new patients. This involves providing the necessary inputs the model (e.g., age, sex, BMI, smoking habits, etc.) and obtaining a predictedhealthinsurancecost.

7. SCREEN SHOTS

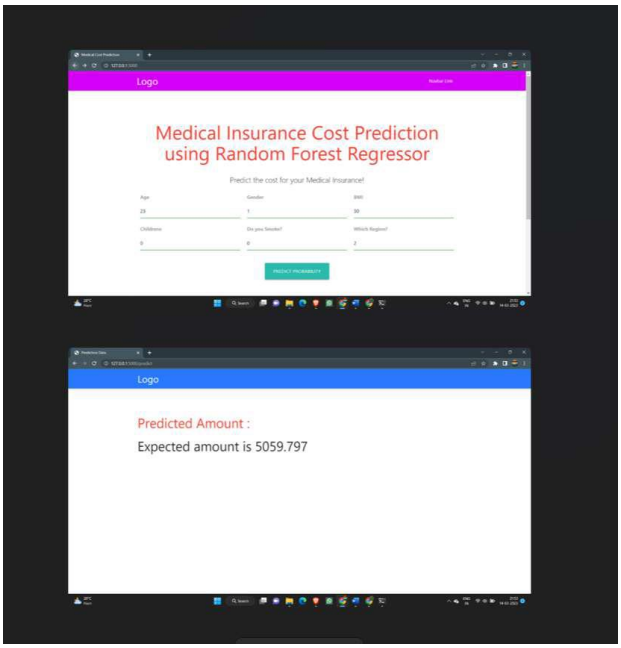

Fig. Output of predicted amount

8. CONCLUSION

In the field of health insurance, machine learning is well-suited to tasks that are.often performed by people at a slower speed. AI and machine learning can analyze and evaluatinglarge volumes of data in order to streamline and simplify health insurance operations. The impactof machine learning on health insurance will save time and money for both policyholders andinsurers.

ISSN: 2366-1313

AI will handle repetitive activities, allowing insurance experts to focus on processes thatwill improve the policyholder's experience. Patients, hospitals, physicians, and insuranceproviders will benefit from ML's ability to accomplish jobs that are currently performed by peoplebut are much faster and less expensive when performed by ML. When it comes to exploitinghistorical data, machine learning is one component of cognitive computing that may addressvarious challenges in a broad array of applications and systems. Forecasting health insurancepremiums is still a topic that must be researched and addressed in the healthcare business. In thisstudy, the authors trained an ANN-based regression model to predict health insurance premiums.The model was then evaluated using key performance metrics, i.e., RMSE, MSE, MAE, r2, andadjusted r2. The accuracy of our model was 92.72%. Moreover, the correlation matrix was alsoplotted to see the relationship between various factors with the charges. This domain of insuranceprediction has not been fully explored and requires thorough research.

9. FUTURE SCOPE

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

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 Cyberecurity (ICICCS-INBUSH) (pp. 64-69). IEEE.

2. Kaggle Medical Cost Personal Datasets. Kaggle
Inc.https://www.kaggle.com/mirichoi0218/insurance.

3. Pesantez-Narvaez, J., Guillen, M., &Alcañiz, M. (2019). Predicting motor insurance claimsusing telematics data—XGBoost versus logistic regression. Risks, 7(2), 70

4. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September).Automating Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE FifthInternational Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

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 clinicalresearch: potential and pitfalls. Bmj, 338

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

8. Fauzan, M. A., &Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction.Int. J. Adv. Soft Comput. Appl, 10(2).

9. Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobileinsurance. In 2018 Second International Conference on Inventive Communication andComputational Technologies (ICICCT) (pp. 1338-1343). IEEE.

10. Kayri, M., Kayri, I., &Gencoglu, M. T. (2017, June). The performance comparison ofmultiple linear regression, random forest and artificial neural network by using photovoltaic andatmospheric data. In 2017 14th International Conference on Engineer

ISSN: 2366-1313

ing of Modern ElectricSystems (EMES) (pp. 1-4). IEEE