International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.079 ISSN: 2581-4621

Insurance Claim Prediction using Machine Learning

B Sunil Kumar[1], P Tejaswee[2] ,A Sakunthala[3], N Sucharitha[4], J Yugandhar Naik[5]
Assistant Professor, student, Department of CSE, AITS, Tirupati

Abstract
Nowadays, data is extremely important and valuable in the insurance sector. Insurance policies aim to minimize or reduce the costs incurred due to different risks. Several factors affect insurance claim charges, which are taken into consideration when developing insurance policies. In the field of computational and applied mathematics, machine learning (ML) is a well-known research area. ML is a type of computational intelligence that can help address various challenges in a wide range of applications and systems by using historical data. However, there are some limitations to ML, and predicting medical insurance costs using ML approaches is still a problem that requires further investigation and improvement in the healthcare industry. To address this issue, this study employs a series of machine learning algorithms to develop a computational intelligence approach for predicting health insurance costs and the results of those algorithms were compared.
Keywords: Machine Learning, Prediction, Health Insurance, Claim, Healthcare industry, Risks, Data

I.Introduction

We live on a planet full of threats and uncertainty. Including People, households, durables, properties are exposed to different risks and the risk levels can vary. These risks range from risk of health diseases to death if not get protection, and loss in property or assets. But, risks cannot usually be avoided. Therefore, Insurance is one of the policies that either decreases or removes loss costs incurred by various risks [1]. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance [2]. A loss may occur when a client purchases a plan which may be claimed a less amount then their eligibility [3]. Various parameters or factors play an important role in estimating the insurance charges. If any factor is omitted or changed when the amounts are computed then, the overall policy cost changes. It is therefore very critical to carry out these tasks with high accuracy. So, the possibility of human mistakes are high so insurance agents also use different tools to calculate the insurance premium. Thus, ML is beneficial here. The most important advantage of Machine Learning (ML) to use in Insurance Industry is to facilitate data sets [4]. Machine learning is one of the advanced technologies which has a potential in finding patterns and developing functions based on the given information. Our goal is to predict insurance costs. In this, we used several models of regression, for example, Linear regression, Decision Tree regression, Random Forest regression and Gradient Boosting regression.

II.Related Work

In 2018, Muhammad rFauzan during this paper, the truth of XGBoost is applied to predict statements. Compare the output with the performance of XGBoost, a group of techniques e.g., AdaBoost, Random Forest, Neural Network. XGBoost offers higher Gini structured accuracy. mistreatment publically accessible urban center Seguro to Kaggle datasets. The dataset includes vast quantities of NaN values however this paper manages missing values by medium and median replacement. However, these simple, unprincipled strategies have additionally proved to be biased. They, therefore, target exploring the cubic centimeter methods that are extremely applicable for the issues of many missing values, such as XGboost [5].
Data mining (DM) and machine learning (ML) techniques are widely used for insurance cost prediction and medical fraud detection [6]. In [7], hierarchical Decision Trees and other ML models

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.079 ISSN: 2581-4621

are used for predictive analytics of healthcare costs. They also suggested that machine learning tools and techniques are critical in the healthcare sector and that they are exclusively used in the diagnosis and prediction of medical insurance costs. Similarly, the underwriting process and medical investigations necessary by the insurance firm to profile the applicants' risks can be difficult and costly [8].

Statistics has been in use in the insurance industry from the onset of the industry. There is a whole discipline for the use of statistics in the insurance industry known as Actuarial Science. With the massive increase in data processed by the industry, Predictive Analytics is coming into the limelight [9]. It encompasses data mining, predictive modelling, and machine learning techniques like classification, regression, clustering and outlier detection to make accurate and fast predictions about unforeseen events in the future using the current data [10].

III. Dataset used

We utilized a dataset obtained from Kaggle for developing our prediction model. The dataset consisted of 1338 records with 7 attributes, namely 'age', 'gender', 'BMI', 'children', 'smoker', 'region', and 'charges'. The data was structured and stored in a CSV file. To prepare the model, we split the data into two parts, a training set, and a testing set. 80% of the total data was used for training the model, while the remaining 20% was used for testing. We used the training dataset to develop a predictor model for medical insurance cost, while the test set was used to evaluate the regression model's performance.

Name	Description
age	Age of client
gender	Male/Female
BMI	Body Mass Index
children	Number of children the client have
smoker	Whether a client is a smoker or not
region	Whether the client lives in southwest, northwest, southeast or northeast
Charges	Medical cost the client pay

Table 1: Dataset overview

IV. Regression Models

Regression is a statistical analysis technique that helps to understand the relationship between independent variables or features and a dependent variable or outcome. It is commonly used as a method for predictive modeling in machine learning, where an algorithm is employed to predict continuous outcomes.

A. Linear Regression

Linear regression is a widely used and straightforward Machine Learning algorithm for predictive analysis. It is a statistical technique that predicts continuous or numeric variables such as product prices, sales, age, and salary. Linear regression algorithm establishes a linear connection between one or more independent variables (X) and a dependent variable (y).

B. Decision Tree Regression

A decision tree is a type of model used in machine learning that builds classification or regression models in the form of a tree-like structure. This algorithm divides a given dataset into smaller and

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.079 ISSN: 2581-4621

smaller subsets while incrementally building an associated decision tree. The resulting tree consists of decision nodes and leaf nodes. Decision trees are effective in handling both categorical and numerical data, and are often used as a decision-making tool to model outcomes, input costs, and utility.

C. Random Forest Regression

Random Forest is a machine learning algorithm that uses an ensemble technique to perform both regression and classification tasks. It relies on the use of multiple decision trees, and a technique called Bootstrap and Aggregation (bagging). The idea behind Random Forest is to combine the predictions of multiple decision trees in determining the final output, rather than relying on individual decision trees.

D. Gradient Boosting Regression

Gradient boosting is an ensemble method that involves creating multiple weak models and combining them to achieve better performance overall. The key idea behind gradient boosting is to compute a sequence of simple trees, where each successive tree is built for the prediction residuals of the preceding tree. In this way, the algorithm gradually improves the model's predictions by minimizing the errors in each successive tree.

V.Results

The performance of each of the four algorithms was evaluated based on Mean Absolute Error (MAE) and R-Squared calculations. Table 2 presents the accuracy percentages of different attributes across all models. After comparing the performance of each of these machine learning algorithms, it was concluded that Gradient Boosting Regression and Random Forest Regression exhibited better performance as compared to the other algorithms. These models achieved an accuracy of 93.5% and 87%, respectively. The model that achieved the highest accuracy rate by utilizing all four attributes was chosen as the best model, which turned out to be Gradient Boosting Regression with an accuracy rate of 93.5%.

Algorithm	Mean Absolute Error (MAE)	R2_score (Training)	R2_score (Test)	Cross-validation
Linear regression	9194.0082	0.7414	0.7827	0.7445
Decision Tree regression	7342.8704	0.8201	0.8732	0.8334
Random Forest regression	4307.8631	0.8744	0.8792	0.8492
Gradient Boosting regression	2444.9258	0.8952	0.8986	0.8583

Table 2: Comparison of regression models

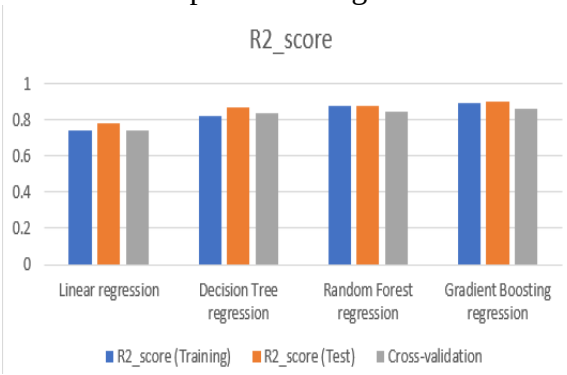

Fig 1: R-squared error graph

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Special Issue: 1 Volume No.7 April – 2023
DOI:10.46647/ijetms.2023.v07si01.079 ISSN: 2581-4621

Fig 2: Mean Absolute Error (MAE) graph

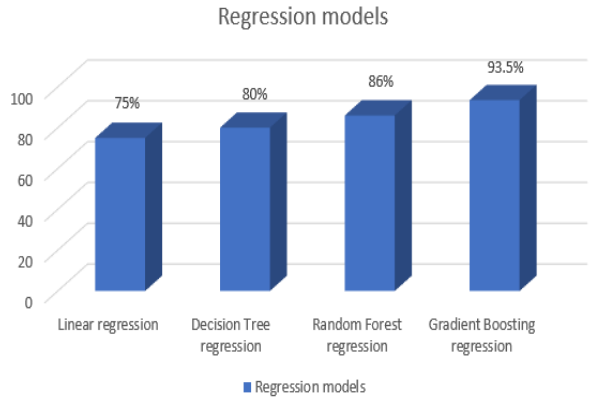

Fig 3: Accuracy rate of regression models

Conclusion

Machine learning (ML) is a vital component of computational intelligence that has the potential to solve a range of problems in various applications and systems by leveraging historical data. In this study, we employed different machine learning regression models to forecast health insurance charges based on specific attributes using the medical cost personal dataset obtained from Kaggle.com. We examined the effect of various factors on the predicted amount and identified that a person's age and smoking status were the most influential attributes across all algorithms applied. Features that had no impact on the prediction were removed from the dataset. Our focus was on predicting the premium amount based on a person's health rather than the insurance company's terms and conditions. We compared the predicted premiums generated by the models with the actual premiums to assess the accuracies of the models. We discovered that Gradient Boosting Regression was the most efficient model, exhibiting an accuracy of 93.5%. Hence, we recommend the use of Gradient Boosting Regression for estimating insurance costs as it provides better performance than other regression models.

VI.References

[1] Mukund Kulkarni, Dhammadeep D. Meshram, Bhagyesh Patil, Rahul More, Mridul Sharma and Pravin Patange, “Medical Insurance Cost Prediction using Machine Learning”, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue XII Dec 2022.

DOI:10.46647/ijetms.2023.v07si01.079 ISSN: 2581-4621

[2] Nidhi Bhardwaj & Rishabh Anand, "Health Insurance Amount Prediction," International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 9 Issue 05, May-2020.
[3] Sujith Thota, Kotha Vishnu Sai and P Swarnalatha, "Machine Learning Implementation for Health insurance", International Journal of Advanced Trends in Computer Science and Engineering, ISSSN 2278-3091, Volume 10, No.3, May- June 2021.
[4] Rama Devi Burri, Ram Burri, Ramesh Reddy Bojja & Srinivasa Rao Buruga, "Insurance Claim Analysis Using Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4, April 2019.
[5] 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
[6] L. S. Chen and J. C. Chen, "Using data mining methods to detect medical fraud," in Proceedings of the 2020 International Conference on Management of e-Commerce and e-Government, pp. 89–93, Jeju Island, South Korea, July 2020.
[7] A. Tike and S. Tavarageri, "A medical price prediction system using hierarchical decision trees," in Proceedings of the IEEE International Conference on Big Data (Big Data), pp. 3904 –3913, IEEE, Boston, MA, USA, December 2017.
[8] N. Boodhun and M. Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," Complex & Intelligent Systems, vol. 4, no. 2, pp. 145–154, 2018
[9] Pal, D., Mandana, K. M., Pal, S., Sarkar, D., & Chakraborty, C. (2012). Fuzzy expert system approach for coronary artery disease screening using clinical parameters. KnowledgeBased Systems, 36, 162–174. 10.1016/j.knosys.2012.06.013
[10] Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks – a review. Journal of King Saud University - Computer and Information Sciences, 31(4), 415–425. 10.1016/j.jksuci.2017.12.007.