Health Insurance Claims Prediction Using Machine Learning

B.Kalyani[1], K.Lalitha Annapurna[2], A.Bhavani[3], B.Jhansi Vazram[4]

[1,2,3]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

[4]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

bandikalyani392@gmail.com[1],klalithaannapurna@gmail.com[2],ambatibhavani1111@gmail.com[3], jhansi.bolla@gmail.com[4]

ABSTRACT- The goal of this project is to create a predictive model for health insurance claims using machine learning methods. The Kaggle website provided the dataset for this study. The technology can also help policymakers identify which providers are often more expensive and, if required, take punitive action. To produce useful features for our machine learning models, we preprocess the data and engage in feature engineering. Following that, we assess a number of regression techniques utilizing metrics, including linear regression, random forest regression, and decision tree regression.

KEYWORDS: Health Insurance Claims, Machine Learning, Linear Regression, Random forest regression, Decision tree regression, Probability Prediction

1. INTRODUCTION

Machine learning, a fast evolving field of artificial intelligence, enables computers to automatically acquire knowledge through experience and improve over time without human input. Machine learning allows computers to examine massive volumes of data, spot patterns and trends, and then use that information to predict the future or make decisions. There are many useful uses for machine learning [1], including fraud detection, natural language processing, picture and speech recognition, and personalised recommendations. Machine learning is anticipated to have a significant impact on many businesses and facets of daily life as it develops.

A healthcare industry use of artificial intelligence called health insurance claim prediction using machine learning seeks to increase the precision and effectiveness of processing insurance claims. Machine learning models can be trained to recognise patterns and forecast the likelihood that a claim will be approved or refused by utilising historical data and prediction algorithms.

This can improve customer satisfaction, reduce fraud, and simplify the claims process for insurance companies. Machine learning can also assist medical personnel in identifying those with a likelihood of getting particular illnesses and offering them preventative therapy, improving patient outcomes.

The application of machine learning in health insurance claim prediction is becoming increasingly crucial for the industry to remain competitive and deliver high-quality care to patients as healthcare data continues to grow and get more complex.

The purpose of utilising machine learning to forecast health insurance claims is to increase the precision and effectiveness of processing insurance claims in the healthcare sector [2]. Scikit, Numpy, Pandas, and Tensorflow are a some machine learning software packages that can be used to create this system.

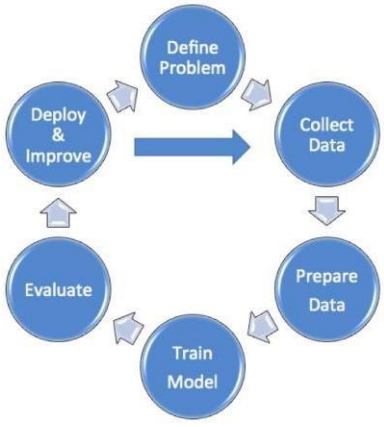

Fig.1. Workflow of Machine Learning

2. RELATED WORKS

Medical insurance claim prediction using ml algorithms is an active area of research due to its potential to improve the accuracy and efficiency of claim processing. This is a summary of a few recent research efforts in this area.

A systematic review of "Health Insurance Claim Prediction Using Machine Learning Algorithms" by S. S. Khan et al. This investigation performed a systematic analysis of 32 research that predicted health insurance claims using machine learning algorithms [7]. The most widely utilised algorithms, according to the authors, are decision trees, logistic regression, and neural networks.

The article "Predicting Health Insurance Claims Using Machine Learning Techniques: A Comparative Analysis" was written by D. P. Shukla et al. In order to anticipate health insurance claims, this study assessed the effectiveness of four machine learning algorithms In terms of accuracy, the authors discovered that random forest fared better than the other algorithms.

In "Health Insurance Claim Prediction Using Machine Learning Algorithms: A Comparative Analysis," S. P. Singh et al. Six machine learning algorithms were evaluated in this study for their ability to predict health insurance claims. In terms of accuracy, the team discovered that random forest and support vector machine fared better than the other algorithms.

The article "Health Insurance Claim Prediction Using Deep Learning Approaches" was written by S. K. Jha et al. In order to forecast health insurance claims, this study applied deep learning techniques. The convolutional neural network functioned more accurately than the long short-term memory, according to the authors.

Hybrid Machine Learning Methods for Health Insurance Claim Prediction, S. Sharma et al. In order to anticipate health insurance claims, this study developed a hybrid machine learning strategy that includes decision trees, k-nearest neighbours, and random forests.

The hybrid technique performed more accurately than the individual algorithms, according to the authors.

3. MATERIALS AND METHODOLOGY

The dataset was downloaded in csv format from Kaggle. There are 6 columns of prediction features in this dataset's 1338 records. Both categorical and continuous data are present in this dataset. The dataset's characteristics are listed below.

Features	Representation
Age	Age of the patient
Sex	Gender of the patient
Children	Number of children to the patient
B.M.I	Body mass index of the patient
Region	Residential area of the patient
Smoker	Smoking habits of the patient
Charges	Medication cost of the patient

TABLE.1. Features of Dataset

There are no missing or null values in the insurance dataset that was used for this study. Hence, there is not much preprocessing to be done [8]. However, to turn the categorical data into continuous data, we applied label encoding. Next, we must use different regression models, including linear regression, random forest regression, and decision tree regression, to train our dataset. Several evaluation metrics, including Mean Squared Error (MSE), R2 score, and Root Mean Squared Error (RMSE), are used to compare the training dataset (MSE). The most accurate model will be selected to forecast the user's health insurance claims.

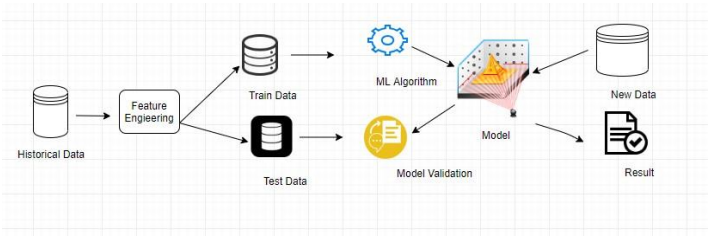

Fig.2. Workflow of insurance cost prediction

To better comprehend the dataset, we have done a graphical depiction. Here are some distributions of Age, BMI in Fig.3. and distributions of Children, Charges in Fig.4.

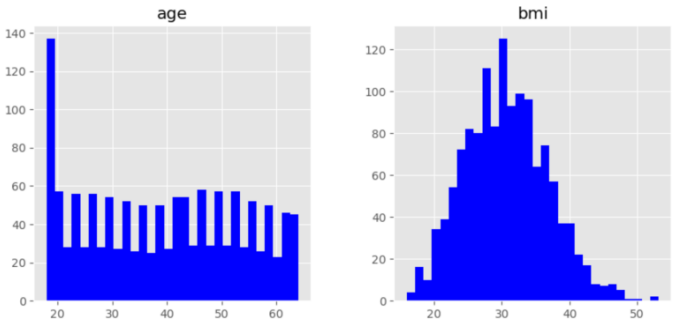

Fig.3. Distributions of Age and BMI

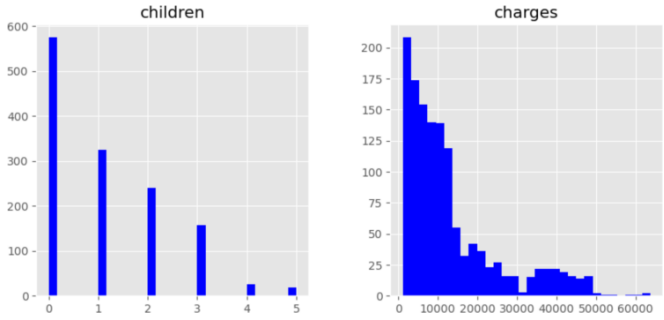

Fig.4. Distributions of Children and Charges

4. REGRESSION METHODS

Regression methods are used in statistics to examine the connection of both a dependent variable and one or even more independent variables [3]. Regression analysis is a powerful tool that can shed light on the underlying connections and patterns that the information reveals when used in research, forecasts, and decision-making.

To train the dataset for this system, we used three regression techniques: linear regression, random forest regression, and decision tree regression.

A. *RANDOM FOREST REGRESSION*

The random forest regression approach works by building a number of decision trees, each trained on a different subset of data that is chosen at random [4]. Using the values of the independent variables as their basis, the trees are made to divide the data into more manageable and homogeneous groupings . The final forecast is the average of all individual tree predictions. Each tree makes a prediction based on the mean or mode of the dependent variable inside its terminal nodes [9]. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The ability of random forest regression to handle high-dimensional datasets and datasets with a lot of linked variables is one of its key benefits.

In order to choose features and understand models, the algorithm can determine which variables are most crucial for prediction [5]. Also, compared to conventional regression models, random forest regression is less prone to overfitting, which can be problematic when working with complicated data.

Random forest regression does have certain drawbacks, though. When dealing with huge datasets, the approach can be computationally demanding [6]. Moreover, as the technique needs complete data for training, it is not suitable for datasets containing incomplete data.

5. RESULTS

The performance of the several regressors were analysed in order to pick the most effective fundamental algorithm and establish the correctness and reliability of the model's deployment. The efficiency of the regressor can be explained using the estimation coefficient and the Root Mean Squared Error (RMSE) (r2 score).

MODEL	R2 SCORE	RMSE	ACCURACY
Linear Regression	0.7694415	6144.199	76 %
Random Forest Regression	0.8525963	1876.602	85 %
Decision Tree Regression	0.7293854	6299.100	72 %

TABLE.2. Comparison of models

Comparing the multiple regression models used in Table 2 for predicting the insurance claim revealed that the larger the r-squared value of the models. Hence, the R-squared number gives the most accurate representation of how variable the dependent variables is in relation to the surrounding mean.

With an R2 value of 0.8525963, the Random Forest Regression model was determined to be the best model.

6. CONCLUSION AND FUTURE WORK

The expense of the health policy has been estimated as precisely as possible using the predictive models described here. To create better medical facilities, this is allegedly highly beneficial for the healthcare organisation. The study made use of Decision Tree Regression, Random Forest Regression, and Linear Regression methodologies. When all the results were compared, Random Forest Regressor had the greatest R2 score.

Future research, however, can focus on exploring a number of additional deep learning applications. The technique that can converge the changes and do multiple framework can be improved using improved approaches from machine learning and other fields. Although most forecast errors involve a significant financial claim transaction, updating some of them can generally benefit insurance firms.

7. REFERENCES

[1] "Prediction of Insurance Claim Severity Loss Using Regression Models," R. M. Ogunnaike and D. Si, 2017, pp. 233-247.

[2] "Modeling frequency and severity of claims with the zero-inflated generalised cluster weighted models," Insur. Math. Econ., vol. 94, pp. 79–93, September 2020; doi: 10.1016/j.insmatheco.2020.06.004..

[3] "A complete overview and analysis of generative models in machine learning," Computer Science Review, vol. 38, no. 100285, 2020, doi: 10.1016/j.cosrev.2020.100285.

[4] "Nuclei segmentation in cell pictures using fully convolutional neural networks," Int. J. Emerg. Technol., vol. 11, no. 3, pp. 731–737, 2020. S. S. Rautaray, S. Dey, M. Pandey, and M. K. Gourisaria.

[5] Evaluation of Technology Adoption Models and Theories to Assess Readiness and Appropriate Usage of Technology in a Corporate Organization by T. Dube, R. Van Eck, and T. Zuva 10.36548/jitdw.2020.4.003

[6] T. J. Layton, "Imperfect risk adjustment, risk preferences, and sorting in competitive health insurance markets," J. Inf. Technol. Digit. World, vol. 02, no. 4, pp. 207–212, 2020. 56, 259–280, J. Health Econ., 2017, doi: 10.1016/j.jhealeco.2017.04.004.

[7] Jayasree, M., and M. G. Chandrasekhar (2020). Machine Learning Methods for Predicting Health Insurance Claims. The International Conference on Communication and Signal Processing (ICCSP) will take place in 2020. (pp. 0426-0431). IEEE.

[8] P. Awasthi, P. Sharma, and S. (2020). Machine Learning for Predictive Analysis of Health Insurance Claims. The eleventh ICCCNT (International Conference on Computation, Communication, and Networking Technologies) will take place in 2020. (pp. 1-6). IEEE

[9] Albarrak, A. M., Elshaikh, E. A., and Basheer, M. A. (2021). A thorough review of the prediction of health insurance claims using machine learning techniques. 45(1), 1-14 in Journal of Medical Systems.