Machine Learning to Forecast the cost of Health Insurance

Keerthana Prathap
PG Scholar
Department of Computer Applications
Amal Jyothi Collge of Engineering,
Kanjirapally,Kerala
keerthanaprathap@mca.ajce.in

Bijimol TK
Assistant professor
Department of Computer Applications
Amal Jyothi Collge of Engineering,
Kanjirapally, Kerala
tkbijimol@amaljyothi.ac.in

***Abstract*— It is a significant difficulty for the insurance industry to charge each customer a premium that is reasonable for the risk that they represent. The insurer's management results and financial statements are significantly obstructed by how accurately the claim amount is projected. Estimating the cost of claims in an insurance company is a real problem that needs a more precise and automated solution. A lot of factors influence the cost of claims based on health characteristics such as age, sex, weight, BMI, hereditary disorders, the number of dependents, blood pressure, smokers, region, regular exercise, and others. The cost of health insurance is analyzed and projected by insurance companies using a variety of techniques. This study demonstrates the ability of various models to predict insurance premiums.**

Keywords*— *Insurance, Regression, SVM, Random Forest, Gradient Boosting

I. INTRODUCTION

The insurance policy, which establishes the privileges that the insurer is rightfully obligated to pay, is a agreement between the insurer and the policyholder. The insurer guarantees to reimburse losses brought on by risks enclosed by the policy language in return for an upfront payment or cost known as the premium . In order for insurance firms to accurately evaluate or quantify the amount covered by this policy and the insurance premiums that must be paid for it, it becomes crucial to consider the price of insurance in people's lives. These costs are estimated using several variables. Each of these factors is significant. When the quantities are calculated, the policy is altered if any factor is left out. Therefore, it is essential that these responsibilities be carried out precisely. Insurance companies employ experts in this field since mistakes made by humans can happen. Additionally, there are several tools to determine the insurance premium. The model is trained using historical insurance data. The model can then accurately predict insurance policy costs by using the necessary elements to measure the payments as its inputs. This reduces the need for labor and other resources and increases business profitability. In this they read the data then data analysis and preparation done. Data analysis includes summary statistics, correlation, relationships etc. Insurance premiums are based on a variety of factors. Insurance premiums are therefore continuous quantities. Regression is the finest option that can meet our needs. The dataset for health insurance premium costs is utilized for this investigation [2].

First, the dataset underwent pre-processing. Next, we used training data to create regression models, and testing data to assess these models. Insurance companies can utilize linear regression with interaction to forecast claim amounts for their clients.

II. RELATED WORKS

The price of health insurance is predicted using a variety of machine learning algorithms. The use of machine learning algorithms such as decision trees, support vector machines (SVM), linear regression, etc. A statistical method used for predicting outcomes is linear regression. A supervised machine learning technique called SVM is frequently used for regression and classification. Another supervised machine learning method is the decision tree algorithm, which continuously divides given data into subsets based on predetermined criteria at each row until the final result is created.

Michael Chernew et. al[1] Creativities intended at reducing the uninsured rate must address the mounting pressure on coverage costs. Probit regression and instrumental variable methods are used to estimate the relationship between rising local health insurance costs and the declining prevalence of people having any type of health insurance coverage while controlling for a wide range of financial, demographic, and policy covariates.

Xiaoqun et.al [2] This study analyses and forecasts residential health insurance cost data using a support vector machine model that is built on K-means. To confirm the model's efficacy, we employ the five model assessment metrics ME, RMSE, MAE, MPE, and MAPE. This paper will offer some suggestions for the upcoming development of cost-control measures for health insurance in order to further advance the cause of health insurance.

Bardwaj et.al[3] The medical information to estimate the cost of an individual's insurance. Three models with names like Decision Tree and Multiple Linear Regression, Decision tree Regression and Gradient Boosting Regression has taken to evaluate and contrast the effectiveness of these algorithms.

Sailaja et.al [4] A Stacking Regression model would be advantageous for health insurers to use to forecast individual insurance costs. This belief based on suggestion graphs states that a person can reduce their insurance costs by depressing their BMI, relocating, or transitioning to a non-smoker.

C. Jyothsna et.al [5] Regression models used in this work included multi-linear, decision trees, random forests, and gradient boosting. Gradient Boosting has the highest

DOI: 10.5281/zenodo.7996987
ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

accuracy of all the approaches, scoring an impressive 87 percent after comparing the accuracies. Finally, the Telegram chatbot is programmed with instructions to converse with the user and calculate the insurance premium using the optimal model.

III. METHODOLOGY

3.1 Dataset

The dataset is from US Dataset, which consists of 1339 patient's data who are of the age 18 or older. The dataset used is in CSV format. There are ten features in total. Each feature is defined as given below

age : Age of the policyholder (Numeric)

sex: Gender of policyholder (Categoric)

weight: Weight of the policyholder (Numeric)

children: No of children

bmi: Body mass index

no_of_dependents: Number of dependent persons on the policyholder (Numeric)

smoker: Indicates policyholder is a smoker or a non-smoker (non-smoker=0;smoker=1) (Categoric)

charges: The amount claimed by the policyholder (Numeric)

diabetes: Indicates policyholder suffers from diabetes or not (non-diabetic=0; diabetic=1) (Categoric)

regular_ex: A policyholder regularly exercises or not (no-exercise=0; exercise=1) (Categoric)

region: region where person reside

3.2 Data Pre-processing

Real-world data is unstructured, which means that it may contain missing values, noisy data, and/or inconsistent data. No quality results may be found if the data quality is poor. To produce high-quality results, the data must be preprocessed. The data is preprocessed using cleaning, integration, alteration, reduction, and discretization. Making the data more appropriate for data mining and analysis in terms of time, cost, and quality is crucial.

3.3 Data Cleaning

Data cleaning is the procedure of eliminating or varying data that is wrong, missing, redundant, duplicated, or formatted incorrectly in order to prepare it for analysis. In the fields of data management, analytics, and machine learning, data cleaning is vital. One of the fundamental components of fundamental data science is data cleaning.

3.4 Splitting of Dataset

The complete dataset is separated into a training dataset and a test dataset after pre-processing and data cleaning. The remaining 20% of the data are utilized for testing, and the remaining 80% are used for training. The relevant data variable is used to hold the training and test data, which is then used for future prediction.

3.5 Find accuracy

The accuracy of each method is then assessed. Specifically, each algorithm was compared after the training and test results of four algorithms were analyzed. So, we proceed with the algorithm that has the highest level of accuracy to create the model. According to the information presented above, the gradient Boosting Regressor is the best model for this dataset. It is best practice to train our model on the complete dataset before going into production.

IV. MODELLING

- Linear Regression

A machine learning algorithm created on supervised learning is linear regression. It performs a regression operation. Regression uses independent variables to model a goal prediction value. It frequently serves to regulate how factors and forecasting interact. The task of predicting a dependent variable's value (y) based on an independent variable is carried out using linear regression (x). Therefore, x (the input) and y (the output) are found to be linearly related by this regression technique (output). Thus, the term "linear regression" was coined.

- SVR

In classification issues, Support Vector Machines (SVMs) are well known. However, the application of SVMs in regression has received less attention. The term "Support Vector Regression" is used to describe these models (SVR).

- Random Forest regression

It is an ensemble learning method and is the development of using multiple models, trained over the same data, averaging the results of each model finally finding a more powerful predictive result. There is a group of predictors called an ensemble, hence this technique is called Ensemble Learning. The random forest has a lower generalization error than a single decision tree due to its randomness, decreasing the models' variance.

- Gradient boosting

Classification and regression tasks are between the applications of a machine learning technique termed gradient boosting. It offers a weak prediction model in the method of a collection of prediction models. The resulting method, known as gradient-boosted trees, often outperforms random forest when a decision tree is the weak learner.

V. RESULT

Gradient Boosting Regression was ultimately determined to be the best model since it provided the highest accuracy rate when all four attribute were considered

```
from sklearn import metrics
```

```
score1 = metrics.r2_score(y_test,y_pred1)
score2 = metrics.r2_score(y_test,y_pred2)
score3 = metrics.r2_score(y_test,y_pred3)
score4 = metrics.r2_score(y_test,y_pred4)
```

```
print(score1,score2,score3,score4)
```

```
0.7830609963486657 -0.07243713495572446 0.8562668884936885 0.875325053644595
```

DOI: 10.5281/zenodo.7996987
ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam

```
[ ] s1 = metrics.mean_absolute_error(y_test,y_pred1)
    s2 = metrics.mean_absolute_error(y_test,y_pred2)
    s3 = metrics.mean_absolute_error(y_test,y_pred3)
    s4 = metrics.mean_absolute_error(y_test,y_pred4)
```

```
print(s1,s2,s3,s4)
```

```
4203.845939096965 8597.056304044343 2635.2103155738805 2494.9130745876632
```

Choice with gradient boosting by taking into account factors like age, sex, BMI, children, smoking, geography, the number of dependents, diabetes, and frequent exercisers, a regressor is used to forecast the cost of a specific location.

```
data = {'age' : int(input('Enter age:')),
        'sex' : int(input('Enter sex:')),
        'bmi' : int(input('Enter bmi:')),
        'children' : int(input('Enter no of  children:')),
        'smoker' : int(input('Enter smoker or not:')),
        'region' : int(input('Enter region:')),
        'no_of_dependents' : int(input('Enter no_of_dependents:')),
        'diabetes' : int(input('Enter diabetic or not:')),
        'regular_ex' : int(input('Enter regular_ex or not:'))}
```

```
Enter age:30
Enter sex:1
Enter bmi:39
Enter no of  children:1
Enter smoker or not:0
Enter region:1
Enter no_of_dependents:1
Enter diabetic or not:0
Enter regular_ex or not:1
```

```
df = pd.DataFrame(data,index=[0])
df
```

	age	sex	bmi	children	smoker	region	no_of_dependents	diabetes	regular_ex
0	30	1	39	1	0	1	1	0	1

```
gr = GradientBoostingRegressor()
```

```
gr = GradientBoostingRegressor()
gr.fit(X,y)
```

```
GradientBoostingRegressor()
```

```
new_pred1 = gr.predict(df)
print("Medical Insurance cost for New Customer is : ",new_pred1[0])
```

```
Medical Insurance cost for New Customer is :  5035.598693325743
```

VI. CONCLUSION

Four regression models are assessed for data on distinct health insurance. The health insurance data was used to create the regression models, and their predicted premiums were associated to the actual premiums to measure their accuracy. The most successful model was discovered to be the Gradient Boosting Regression model.

Gradient boosting has been shown to be one of the best strategies for building prediction models. The accuracy of the model is inclined by a number of factors, including the algorithm that was chosen. Data preprocessing is one such component. It is necessary to eliminate redundant and null values in order to increase efficiency. As this study has demonstrated, feature selection is essential for increasing accuracy and reducing runtime. The selection results in the creation of the model.

Using the data that will be acquired in the following years, the models can be used to predict the premium. This can facilitate cooperation between consumers and insurance companies to achieve better and more health-focused insurance prices. To boost effectiveness, we can raise the dataset's parameters. The dataset will benefit from having variables like blood pressure, city, job, diet habit, and corona detected or not.

VII. REFERENCE

[1] Michael Chernew, David M. Cutler, Patricia Seliger Keenan, “Increasing Health Insurance Costs and the Decline in Insurance Coverage”, 2005, Vol.40, Issue 4, 1021-1039

[2] L.xiaoqun and L. Run,”An Improved k mean Clustering Model Based on Support Vector Machine for Health Insurance Cost Prediction," 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI), 2022,pp.521-525

[3] Nidhi Bardwaj, Rishabh Anand, “Health insurance Amount prediction”,2020, Vol. 9

[4] N. V. Sailaja, M. Karakavalasa, M. Katkam, D. M, S. M and D. N. Vasundhara, "Hybrid Regression Model for Medical Insurance Cost Prediction and Recommendation," 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), 2021

[5] C. Jyothsna, K. Srinivas, B. Bhargavi, A. E. Sravanth, A. T. Kumar and J. N. V. R. S. Kumar,"Health Insurance Premium Prediction using XGboost Regressor," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022

DOI:10.5281/zenodo.7996987
ISBN: 978-93-5906-046-0@2023, Dept. of Computer Applications, Amal Jyothi College of Engineering Kanjirappally, Kottayam