IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306024|

Medical Insurance Cost Prediction Using Machine Learning

Nihal Jalal [1], Alfiya Ibushe [2], Rohan Walvekar [3], Kumarsagar Dange [4]

Student, Department of Electronics and Telecommunication, NMCOE Peth, India[123]

Assistant Professor, Department of Electronics and Telecommunication Engineering, MCOE, Peth, India[4]

ABSTRACT: The prediction of medical insurance costs is a critical task for insurance companies and policyholders. Accurate predictions can lead to better risk management and more affordable insurance plans. This study explores various machine learning algorithms to predict insurance costs, including Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting Regression. Our findings indicate that the Gradient Boosting Regression algorithm performs the best in terms of accuracy. The dataset used comprises demographic and health-related information such as age, sex, BMI, number of children, smoking status, and region. Additionally, we have developed a localhost web application using Streamlit to make our predictive model accessible. This paper details the methodology, results, and implications of using Gradient Boosting for insurance cost prediction.

KEYWORDS: Insurance cost prediction, Gradient Boosting Regression, machine learning, healthcare analytics, regression analysis, Streamlit.

I. INTRODUCTION

Medical insurance costs are a significant concern for both insurance providers and policyholders. Accurate predictions of these costs can lead to better risk management and more affordable insurance plans. Insurance costs are influenced by a variety of factors, including age, sex, body mass index (BMI), smoking habits, and geographical location. Traditional statistical methods often fall short in capturing the complex, non-linear relationships among these variables.

In recent years, the insurance industry has increasingly turned to data analytics to enhance its decision-making processes. Machine learning (ML) and artificial intelligence (AI) provide sophisticated tools to analyze vast datasets and uncover patterns that were previously undetectable. These advanced techniques enable insurers to predict costs with higher accuracy, thereby reducing financial risk and optimizing pricing strategies.

To further enhance the usability and accessibility of our predictive model, we have developed a localhost web application using Streamlit. Streamlit is an open-source Python library that allows for the creation of interactive, data-driven web applications with minimal effort. This application enables users to input relevant factors and obtain predicted insurance costs in real-time, making the model practical for real-world applications.

This study investigates the use of various machine learning algorithms to predict insurance costs based on demographic and health data. By comparing the performance of several algorithms, including Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting Regression, we identify the most effective one for this task. The goal is to provide a robust predictive model that can be used to estimate insurance costs more accurately, benefiting both insurers and policyholders

II. BACKGROUND

The insurance industry relies heavily on data to assess risks and determine pricing. Actuarial science has traditionally used statistical methods to estimate insurance costs, but these methods can be limited in their ability to handle large datasets and complex relationships. Machine learning offers a promising approach to improve prediction accuracy by leveraging patterns in historical data. Techniques such as Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting Regression have been used in various domains for predictive modeling. This study aims to apply these techniques to the prediction of medical insurance costs.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306024|

III. LITERATURE REVIEW

Several studies have applied machine learning techniques to predict insurance costs. Linear Regression is often used due to its simplicity and interpretability, but it may not capture non-linear relationships. Support Vector Regression (SVR) provides more flexibility but can be computationally intensive and sensitive to parameter tuning. Random Forest Regression, an ensemble learning method, can handle non-linear relationships and interactions between variables, but may require extensive computational resources. Gradient Boosting Regression, which combines multiple weak learners to form a strong learner, has shown promising results in various regression tasks, including financial forecasting and medical cost prediction.

Previous research has demonstrated the potential of machine learning models to outperform traditional statistical methods in predicting insurance costs. However, there is a need for a comprehensive comparison of different machine learning algorithms on a standardized dataset to identify the most effective model. This study contributes to the literature by providing a detailed comparison of four popular machine learning algorithms for insurance cost prediction.

IV. METHODOLOGY

4.1 Data Description

We obtained the data set from the Kaggle website in order to calculate the cost of this model prediction. The data set is split into two categories: training data and test data, and it has seven attributes as listed in table 1. The majority of the data used is for testing, with just around 20% being used for training. It contains 1338 records with seven attributes.

Table 1: Overview of the Dataset

Attribute	Data Description
Age	The age of individual person
Sex	Sex of the person (Male, Female)
BMI	This is Body Mass Index
Children	Total number of children of the person have
Smoker	Whether the person is a smoker or not
Region	Where the person lives. Considering four regions (Southwest, Southeast, Northeast, Northwest)

	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338	1338.000000	1338.000000	1338	1338	1338.000000
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.207025	NaN	30.663397	1.094918	NaN	NaN	13270.422265
std	14.049960	NaN	6.098187	1.205493	NaN	NaN	12110.011237
min	18.000000	NaN	15.960000	0.000000	NaN	NaN	1121.873900
25%	27.000000	NaN	26.296250	0.000000	NaN	NaN	4740.287150
50%	39.000000	NaN	30.400000	1.000000	NaN	NaN	9382.033000
75%	51.000000	NaN	34.693750	2.000000	NaN	NaN	16639.912515
max	64.000000	NaN	53.130000	5.000000	NaN	NaN	63770.428010

There were 1338 rows and 7 columns in our data set. The charges variable, which has a float value, is our aim. Maximum number of individuals in our dataset range in age from 18 to 22.5, and the majority of them are male. Few have more than three children, and the majority of them have a BMI between 29.26 and 31.16. In this dataset, four main regions are taken into account: northeast, northwest, southeast, and southwest. The largest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke. We'll investigate our information to determine how the various factors are related. Our target column in this instance is "charges," which is dependent upon every other column.

4.2 Data Preprocessing

Data preprocessing steps included checking for null values, converting categorical variables into numerical values, and splitting the data into training and testing sets. The categorical variables 'sex', 'smoker', and 'region' were encoded using numerical values to facilitate model training. The dataset was then split into 80% training data and 20% testing data to evaluate model performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306024|

Table 2: Categorical to Numerical Conversion

Column Name	Before Conversion	After Conversion
sex	Male	1
	Female	0
Smoker	Yes	1
	No	0
Region	Southeast	1
	Southwest	2
	Northeast	3
	Northeast	4

4.3 Model Selection

Four machine learning models were chosen for comparison Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting Regression. These models were trained on the training set and evaluated on the test set. The implementations of these models were done using the Scikit-learn library in Python.

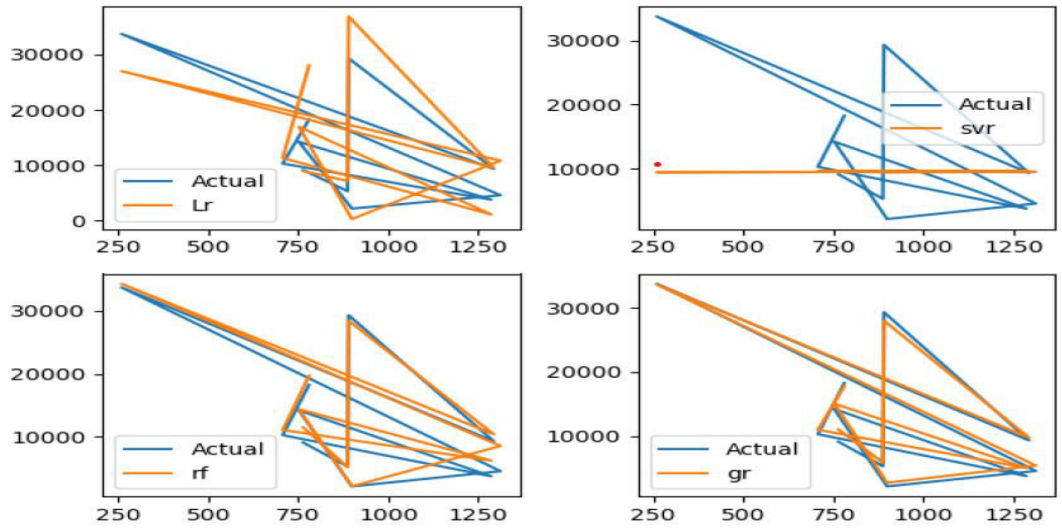

Fig.1 Comparison of actual value and predicted value for finding best algorithm

4.4 Evaluation Metrics

Model performance was evaluated using R-squared score and Mean Absolute Error (MAE). The R-squared score measures the proportion of variance in the dependent variable that is predictable from the independent variables. The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction.

The objective of this study is to predict insurance costs using factors such as age, gender, number of children, geographic location, BMI, and smoking status. These factors contribute to determining the cost of health insurance. The study employs several regression models to estimate insurance costs. The dataset is divided into two parts: 80% for model training and 20% for model testing. To evaluate the accuracy of each model in predicting costs, we calculate the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared value (R^2), and Mean Squared Error (MSE). By comparing these metrics across the models, we can determine which model provides the most accurate predictions.

4.5 Model Performance

The performance of the four models on the test set is summarized as below.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306024|

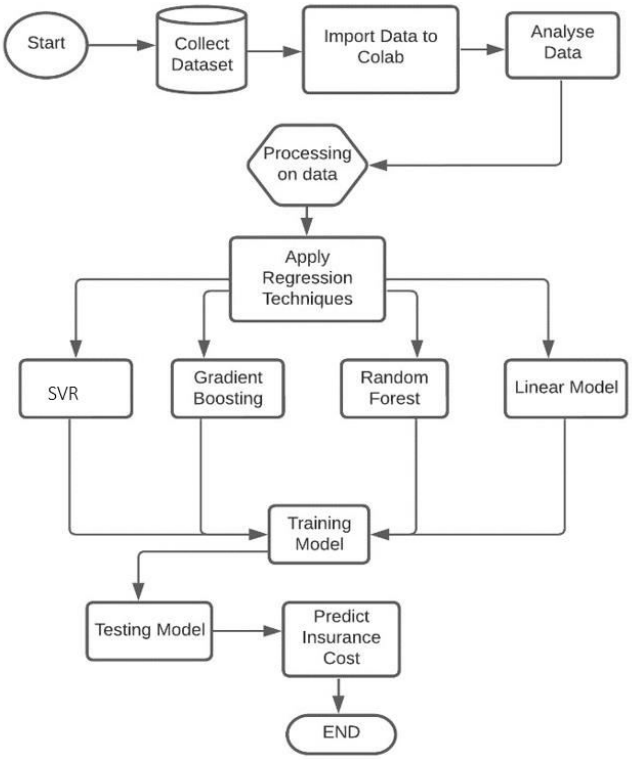

Fig.2 Flow Chart of Medical Insurance Cost Prediction System

Table 3: R-squared score and Mean Absolute Error (MAE).

Model	R-squared	MAE
Linear Regression	0.783	4186.51
Support Vector Regression	-0.072	8592.43
Random Forest Regression	0.866	2472.92
Gradient Boosting Regression	0.878	2447.95

4.6 Best Model

Gradient Boosting Regression outperformed the other models with the highest R-squared score (0.878) and the lowest MAE (2447.95). The results indicate that Gradient Boosting Regression is better at capturing the complex relationships between the input features and the target variable

4.7 Streamlit Web Application

To make the predictive model accessible and user-friendly, we developed a localhost web application using Streamlit. Streamlit allows for rapid development of interactive web applications with Python. The application enables users to input variables such as age, sex, BMI, number of children, smoking status, and region to obtain predicted insurance costs instantly. This interactive tool enhances the practical utility of our model, allowing users to explore different scenarios and their potential impact on insurance costs.

V. RESULTS & DISCUSSION

The results from our analysis indicate that the Gradient Boosting Regression model is the most suitable for predicting insurance costs among the models tested. Its superior performance can be attributed to its ability to handle both linear and non-linear relationships in the data effectively. This model's high accuracy suggests it can be reliably used in

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306024|

practical applications to forecast insurance costs based on the given factors such as age, sex, number of children, location, BMI, and smoking status.

Furthermore, the deployment of this model using Streamlit for creating a web application allows users to input their personal data and receive a predicted insurance cost instantly. This makes the model not only powerful but also accessible and user-friendly for real-world use.

In conclusion, while simpler models like Linear Regression and Support Vector Regression have their advantages in terms of interpretability, the Gradient Boosting model's performance demonstrates the importance of using advanced techniques for complex predictive tasks. This study provides a robust framework for accurately predicting insurance costs, which can be beneficial for both insurance companies and policyholders.

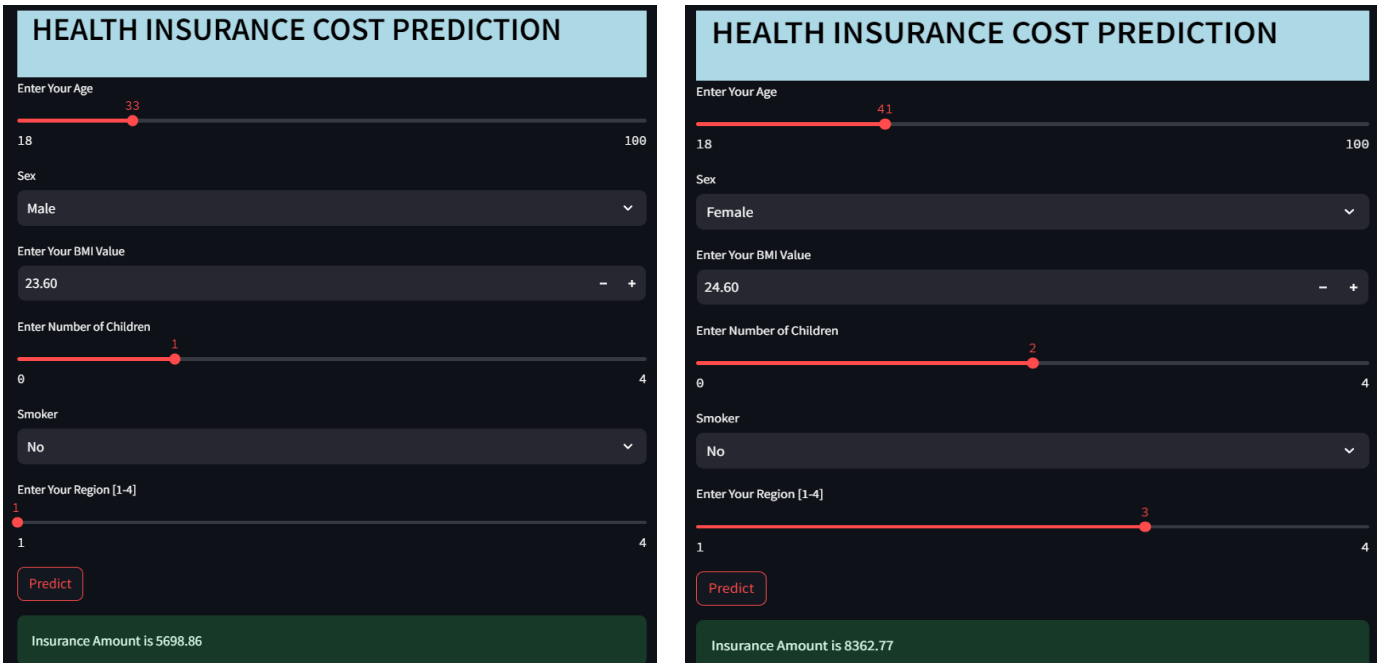

Fig.3 Resulted Webpage by using streamlit

In this Figure 3 we can see how medical insurance cost prediction system predict the cost.

VI. CONCLUSION

This study demonstrates that Gradient Boosting Regression is an effective tool for predicting medical insurance costs. Its high accuracy makes it a valuable asset for insurance companies looking to enhance their pricing strategies. By accurately predicting insurance costs, companies can better manage risk and offer more competitive pricing to policyholders.

Future work may explore the integration of additional features such as medical history and lifestyle factors to further improve prediction accuracy. Additionally, the application of other advanced machine learning techniques, such as deep learning and neural networks, may provide further insights into the complex relationships influencing insurance costs.

REFERENCES

1. Kajia Muller. "The Identification of Insurance Fraud – An Empirical Analysis Working Papers on Risk Management and Insurance," Institute of Electrical and Electronics Engineers (IEEE), June 2013.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306024|

2. Krittika Dutta, Satish Chandra. “A Data Mining Based Target Regression-Oriented Approach to Modelling of Insurance Claims,” Institute of Electrical and Electronics Engineers (IEEE), 2021.
3. G. Kowshalya. “Predicting Fraudulent Claims in Automobile Insurance,” Institute of Electrical and Electronics Engineers (IEEE), 2018.
4. Riya Roy Thomas. “Detecting Insurance Claims Fraud Using Machine Learning Techniques,” Institute of Electrical and Electronics Engineers (IEEE), 2017.
5. J. H. Lee. “Pricing and Reimbursement Pathways of New Orphan Drugs in South Korea: A Longitudinal Comparison,” Multidisciplinary Digital Publishing Institute, International Journal of Research Publication and Reviews (IJRPR), 2021.
6. Gupta, S., & Tripathi. “An Emerging Trend of Big Data Analytics with Health Insurance in India,” Institute of Electrical and Electronics Engineers (IEEE), February 2016.
7. J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz. “Predicting Motor Insurance Claims Using Logistic Regression,” Institute of Electrical and Electronics Engineers (IEEE), June 2019.
8. M. Hanafy and O. Mahmoud. "Predict Health Insurance Cost by Using Machine Learning and DNN Regression Models," International Journal of Innovative Technology and Exploring Engineering, 2021.
9. Hassan C.A., Iqbal J., Hussain S., AlSalman H., Mosleh M.A.A., Sajid Ullah S. A. “Predicting Medical Insurance Cost,” Institute of Electrical and Electronics Engineers (IEEE), 2021.
10. Piet de Jong, Gillian Z. Heller. “Linear Models in Insurance, Including Rating and Other Actuarial Applications,” Machine Learning for Insurance Pricing, 2008.
11. Monique M. Elseviers, Liesbeth Van Camp, Bert De Turck. “A Predictive Model for Health Insurance Costs Using Claims Data, Applying Various Statistical and Machine Learning Techniques,” 2017.
12. Edward W. Frees, Richard A. Derrig, Glenn G. Meyers. “Predictive Modeling Techniques and Their Applications in Actuarial Science, with a Focus on Insurance,” 2014.
13. Goldburd, Marshall, Khare, Anshul, Tevet, Dan. “Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques,” 2016.
14. Agency for Healthcare Research and Quality. “A Predictive Model for Health Costs Using Claims Data,” 2008.
15. Cost Prediction Data Set: https://www.kaggle.com/code/maverickss26/regression-modellng-using-insurance-dataset/input.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details