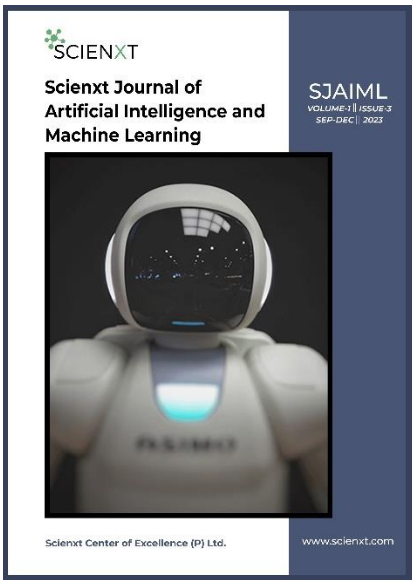

Scienxt Journal of Artificial Intelligence and Machine Learning
Volume-1 || Issue-3 || Sep-Dec || Year-2023 || pp. 1-9

Health insurance premium prediction by using machine learning

Madhu H. S
Department of AIML Jyothy Institute of Technology
Bengaluru, India

Pavithra Kumar D
Department of AIML Jyothy Institute of Technology
Bengaluru, India

Prof. Ramya B. N
Department of AIML Jyothy Institute of Technology
Bengaluru, India

**Corresponding Author: Madhu H. S*
Email: hsmadhu219@gmail.com

Abstract:

A large portion of the economy is devoted to paying for health care. Spending on healthcare accounts for around 30% of the GDP. In terms of both absolute spending and as a percentage of the economy, health spending in developed countries is the greatest. Through its Medicare program, the government foots a sizable percentage of the older population's medical costs. A significant load is placed on the exchequer by the rising cost of health care paired with the baby boomer generation's impending retirement and subsequent eligibility for Medicare. Therefore, it is imperative to use every available tool to limit health-related costs. In this study, we'll create a method for predicting medical costs using machine learning algorithms, which will help direct patients into affordable. The technology can also help policymakers identify which providers are often more expensive and, if required, take punitive action. The Random Forest Regression algorithm will be used in machine learning to forecast the cost of medical care. We also intend to test experiments using different machine learning models, like Gradient Boosted Trees and Linear Regression, on the same data and compare the outcomes. Early estimation of health insurance costs can help. Additionally, people may be vulnerable to being misled into paying for expensive health insurance that they don't need. Our research does not provide an exact amount required by any specific health insurance provider, but it does give a general sense of the cost a person may incur for their own health insurance.

Keywords:

Economy, Policy Makers, Machine Learning, Regression significant.

1. Introduction:

The goal of this research is to help individuals understand the amount of money they may need for health insurance based on their personal health status. This can assist individuals in focusing more on the health- related aspects of insurance rather than the unnecessary ones. In the modern world, it is essential to have health insurance, and most people have a relationship with a public or private health insurance provider. The factors that influence insurance costs vary from company to company. Additionally, some people in rural areas may not be aware that the Indian government offers free health insurance to those who are below the poverty line. However, the process can be complex, and some rural residents either get private health insurance or make no investment at all. Additionally, people may be vulnerable to being misled into paying for expensive health insurance that they don't need. Our research does not provide an exact amount required by any specific health insurance provider, but it does give a general sense of the cost a person may incur for their own health insurance. This is a preliminary estimate and does not adhere to any particular company, so it should not be the only factor considered when choosing health insurance. Early estimation of health insurance costs can help individuals consider the required amount more thoughtfully. When a person can determine the

```
In [7]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Figure. 1: Data sets

2. Input data used:

The following article discusses a dataset that can be accessed on the Kaggle website for the purpose of training and testing. This dataset is saved in a CSV file and is well organized. It is available at the specified link for those interested in using it.

No. of columns=7

1338 rows total. Total number = 9366

In order to accurately predict the cost of health insurance, it is necessary to clean the dataset before applying regression algorithms. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role. Children's property was found to have little impact on the prediction, so it was removed from the input for the regression model to improve efficiency and accuracy.. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. This is a preliminary estimate and does not adhere to any company. These algorithms are designed to make classifications or predictions using statistical techniques, which can uncover key insights in data mining processes. The outcomes from these insights can be seen in the given figure 1 key growth indicators in businesses and applications, if used correctly. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. They will be able to make a more informed decision. Additionally, it may suggest.

3. Concept used:

3.1. Machine learning:

Machine Learning is a subset of computer science and AI that involves using data and algorithms to replicate the way that humans learn. These algorithms are designed to make classifications or predictions using statistical techniques, which can uncover key insights in data mining processes. The outcomes from these insights can have a significant impact on key growth indicators in businesses and applications, if used correctly. (S. Ramakrishnan, 2016) the data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role. The data shows that age and smoking status have the most significant impact on the amount of insurance, with smoking having the greatest effect. However, factors such as family medical history, BMI, marital status, and geography also play a role.

3.2. Linear regression algorithm:

Linear regression is a machine learning algorithm that is based on the concept of "supervised learning." It is used to predict the value of a dependent variable (y) based on the value of an independent variable (x).

Essentially, this means that linear regression is used to determine how closely a dependent variable is related to an independent variable, and then make predictions based on that relationship. (H. Goldstein, 2012)

This is a very useful tool for data analysis, as it allows analysts to understand complex patterns and relationships in data, and to make more accurate predictions about future outcomes.

3.3. Support vector machine algorithm:

SVM, or Support Vector Machine, is a widely used supervised learning algorithm for solving classification and regression problems, with a focus on classification in machine learning (T. Han, 2020)

The goal of SVM is to determine the best line or decision boundary that can separate a multi-dimensional space into different classes, enabling the efficient classification of new data points in the future. This optimal decision boundary is known as a hyperplane, which is created using extreme vectors and points called support vectors. SVM is a popular choice due to its ability to effectively classify data and handle high dimensional spaces.

This is a preliminary estimate. This optimal decision boundary is known as a hyperplane.

3.4. Random forest regression:

The Random Forest approach utilizing bootstrapping involves the use of multiple decision trees generated from the data and combined through ensemble learning techniques. This method often leads to accurate predictions and classifications by averaging the results of the randomly selected trees. (X. Zhu, C. Ying, J. Wang, 2021)

3.5. Gradient boosting algorithm:

Gradient boosting is a highly popular machine learning technique for analyzing tabular data sets. It is well-known for its ability to handle missing values, outliers, and large categorical values in the features, as well as its ability to detect nonlinear relationships between the target and the features. This makes it a powerful tool for data analysis and prediction (Douglas C Montgomery, 2012)

4. Training and prediction:

4.1. Training:

After the necessary data has been formatted and prepared, the model can begin its training and testing phases.

A key focus during the training phase is choosing the appropriate model for the task at hand. This may involve deciding on the optimal modelling strategy or determining the best parameter values for a particular model (V. Roth, 2014)

In some cases, this process is referred to as model selection because various models may be tested and the one that performs the best, is ultimately chosen, which is created using extreme vectors and points called support vectors.

4.2. Prediction:

The model used for predicting the insurance sum for health was based on the relationship between certain features and the label. The accuracy of this prediction was determined by the extent to which the expected value matched the actual amount.

In order to improve the accuracy, the model employed various characteristics, methods, and train-test split sizes. It was found that the amount of data used for training had a significant impact on the accuracy, with a larger train size leading to better results.

The model also employed multiple algorithms in order to forecast the premium amount, and showed how each attribute affected the outcome (Kaggle, Regression data)

5. Results:

The following results can be seen in Prediction:

5.1. Linear regression algorithm:

The accuracy of the Linear Regression Algorithm is 78.334 %.(Bertsimas, M.V. otter, 2018)

5.2. Support vector machine:

The accuracy of the Support Vector Machine Algorithm is 7.229%

5.3. Random forest regression:

The accuracy of Random Forest Regression is 87.006% (Stucki, Finland, 2019). It was found that the amount of data used for training had a significant impact on the accuracy, with a larger train size leading to better results (Kenward, J.A., 2019)

5.4. Gradient boosting algorithm:

The Accuracy of the Gradient Boosting Algorithm is 87.776%.

From the Fig. 2 we can see that the best optimum Algorithm for the Amount prediction is Gradient Boosting Algorithm with the highest accuracy of 87.776 %.(H. Demirtas, J. Stat Soft.)

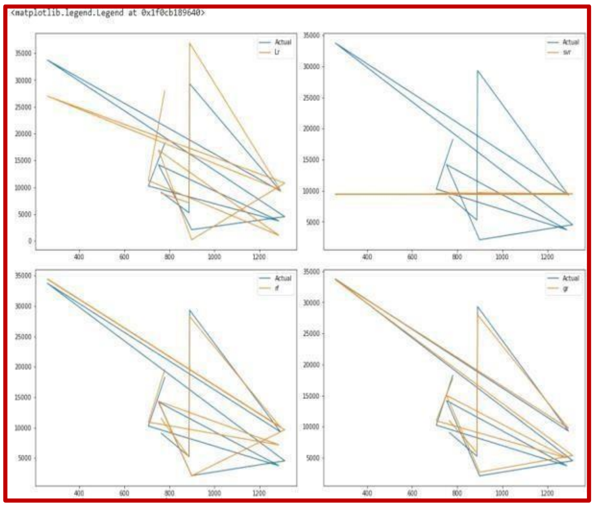

Figure. 2: Actual and predicted price

6. Conclusion and future scope:

6.1. Conclusion:

It was found that Gradient Boosting Decision Tree Regression had the highest accuracy rate for predicting the amount, with a score of 87.776%. (Tian Jinyu , 2019) While linear regression and random forest were able to make correct predictions about 80% of the time, Support Vector Machine did not perform well and was not considered a reliable predictor in this case (G. Reddy, S. Bhattacharya,2020).

When all four attributes were considered, Gradient Boosting Regression was determined to be the best model due to its high accuracy rate. The accuracy of this prediction was determined by extent to which the expected value matched the actual amount.

6.2. Future scope:

The use of the Random Forest algorithm allows for the introduction of unpredictability in the feature selection process, which can improve prediction accuracy (Ostertagova, 2012).

In order to assess the scalability of the system, it would be beneficial to test it on a dataset with at least a million records in the future. Distributed frameworks like Spark and Hadoop can be utilized to handle large amounts of data and enhance the scalability of the system.

Currently, the algorithm is being trained and tested using thousands of records (Donald W. Marquardt, 2012).

7. References:

(1) Kaggle Medical Cost Prediction datasets Kaggle Inclusion Kaggle.com

(2) An emerging trend of big data analytics with heatlh insurance in our country (2016, February). IEEE

(3) H. Goldstein, W. Browne and J. RasBash, “Multilevel modelling of medical data,” Statistics in Medicine, John Wiley and Sons, vol. 21, no. 21, pp. 3291–3315, 2012.

(4) T. Han, A. Siddique, K. Khayat, J. Huang and A. Kumar, “An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete,” Construction and Building Materials, vol. 244, pp. 118–271, 2020.

(5) X. Zhu, C. Ying, J. Wang, J. Li, X. Lai et al., “Ensemble of ML-kNN for classification algorithm recommendation,” Knowledge-Based Systems, vol. 106, pp. 933, 2021.

(6) Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining, “Introduction to linear regression analysis”, John Wiley & Sons, vol. 821, 2012.

(7) V. Roth, “The generalised LASSO”,” IEEE Transactions on Neural Networks”, vol. 15, pp – 16 28, 2014.

(8) Medical Cost Prediction Dataset, [Online]. Available: https://www.kaggle.com/hely333/eda-regression/data

(9) Bertsimas, M. V. Bjarnad´ottir, M. A. Kane, J. C. Kryder, R. Pandey,

S. Vempala, and G. Wang, Operations Research, vol. 56, no. 6, pp. 1382– 1392, 2018.

(10) Stucki, O. “Predicting the customer churn with machine learning methods: case: private insurance customer data” Master’s dissertation, LUT University, Lappeenranta, Finland, 2019.

(11) Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, Carpenter, J. R. (2019). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMI, 338L.

(12) H. Demirtas, "Flexible Imputation of Missing Data", J. Stat. Soft., vol. 85, no. 4, pp. 1–5, Jul. 2018. Available: DOI: 10.18637/ jss. V 085. B 04.

(13) G. Reddy, S. Bhattacharya, S. Ramakrishnan, C. L. Chowdhary, S. Hakak et al., "An ensemble-based machine learning model for diabetic retinopathy classification," in 2020 Int. Conf. on Emergig Trends in Information Technology and Engineering, IC-ETITE, VIT Vellore, IEEE, pp. 1–6, 2020.

(14) Tian Jinyu, Zhao Xin et al., "Apply multiple linear regression model to predict the audit opinion," in 2009 ISECS International Colloquium on Computing, Communication, Control, and Management, IEEE, pp.1–6, 2019.

(15) Ostertagova et al.," Modelling using Polynomial Regression", vol. 48, pp. 500-506, 2012.

(16) Donald W. Marquardt, Ronald D. Snee et al.," Ridge Regression in Practice", " The American Statistician", vol. 29, pp – 3-20, 2012.