Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/73/20240370

Health insurance claim amount prediction using statistical methods

Shiran Wen

New York University, 50 West 4th Street, New York, NY, 10012, USA

sw4389@nyu.edu

Abstract. Healthcare can be costly in the United States. In 2021, the average cost of health insurance for a family of four was $22,221 per year. The average cost of health insurance for an individual was $7,739 per year (Keiser Family Foundation). With the aim of helping the public know more about what factors contribute the most to the expensive cost of their health insurance, this study conducted the analysis to identify the top contributing factors that increase health insurance claim amounts. Various statistical and machine learning tools were used to answer the proposed question: data conversion, correlation, partial correlation, k-means clustering, principal component analysis, multiple regression, lasso regression, support vector machine, and random forest. Consistent results that BMI (Body Mass Index), blood pressure, and smoking are the three variables that contribute the most to the increase in health insurance claim amount were found.

Keywords: Pearson Correlation, Lasso Regression, Random Forest.

1. Introduction

Healthcare is a huge and growing industry in the United States. The total expenditure in the healthcare industry increased from $5259 per person in 1996 ($1.4 trillion) to $9655 per person in 2016 **(**$3.1 trillion). In 2016, an estimated 48.0% of healthcare spending was paid by private insurance, 42.6% by public insurance, and 9.4% by out-of-pocket payments [1]. Health Insurance in the United States also costs a fortune. In 2021, the average cost of health insurance for a family of four was $22,221 per year. The average cost of health insurance for an individual was $7,739 per year [2]. The importance of the healthcare industry and how tightly it is connected to people's everyday lives is apparent. If someone lives in the United States, it is likely that he or she will spend a significant portion of his/her annual income on health insurance.

Knowing that health insurance plays a significant role in people's lives here in the United States, the study aimed to do something that very few others have done explore health insurance on a deeper level. Only one similar study was conducted in 2020 by Bhardwaj and Anand to explore the same topic [3]. The result they received from their study was that age and being a smoker are the two biggest contributors to increasing health insurance claims. It provided people with some insights into the issue. However, they had limited independent variables to begin with. They only had five independent variables to choose as the biggest contributor to increasing health insurance claims. The lack of options might lead to inaccurate results. They also only had three regression models, which indicates the possibility that their analysis might not be comprehensive enough.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/73/20240370

With the aims of bringing more awareness to the issue of personal health, providing people a better understanding of what makes health insurance so expensive, and helping people find the best plan that fits their budget; This study analyzed how health, geographic, and demographic variables affect the health insurance claim amount an individual submits to the health insurance company to be reimbursed. The expectation was to figure out how changes in those different variables can affect the health insurance claim amount one submits, and if it is possible to come up with a predictive model that can predict an individual's claim amount given fixed values of independent variables. This study addressed those potential problems that existed in a similar study by providing ten independent variables to choose the biggest contributors to increased health insurance claims. It also utilized more than five regression models for the analysis to provide a comprehensive analysis. This research provides audiences with an understanding of which variables play bigger roles in affecting/predicting the health insurance claim amount. The hypothesis is that age, being a diabetic, BMI, and blood pressure will be the factors that contribute to the increase of health insurance claim the most.

2. Methods

2.1. Dataset Introduction and Data Cleaning

The dataset in this research was from Kaggle. It was uploaded by the user called "The Devastator". The original dataset contained 11 variables, including 10 independent variables and one dependent variable. They are Index, patient, Age, Gender, BMI, Blood pressure, Diabetic, Children, Smoker, Region, and Claim. "Claim" stands for the health insurance amount an individual submits to the insurance company for reimbursement. "Index" and "PatientID" were deleted from the dataset before the analysis even began, as they were identified as not being useful independent variables for the research, so they. Simple visualizations were then created to identify potential relationships between the independent variables and the dependent variable. However, these simple visualizations did not provide any established relationship between a single independent variable and the independent variable claim. So, the need for focusing on using multiple variables for data processing was recognized. In the dataset, four of the eight independent variables are categorical. All the categorical variables were converted into numerical variables to analyze the data more smoothly. For binary variables such as gender, they were directly converted to one and zero; where a value of one stands for being a male and a value of two stands for being a female. The variable "region" had four distinctive values and therefore cannot be directly converted. So, the virtual variable method was used, and the single variable region was replaced with four numerical variables. The sum of those four variables was always one, which means that knowing three of them can determine the remaining one. These four newly created variables from the original variable region have very strong internal relationships, and it would cause great difficulties for subsequent fitting if they were all directly retained. So, "region_4" was deleted, and "region_1", "region_2", and "region_3" were kept. Thus, two unneeded variables were successfully deleted, and eight mixed variables were transformed into 10 numerical variables, allowing for subsequent numerical calculations and fitting.

2.2. Pearson Correlation

Pearson correlation was run on the dataset. It was the first step of the research plan to figure out the relationships between independent variables (age, gender, BMI, etc.) and the dependent variable(claim), as the Pearson correlation coefficient provides a measure of the strength of linear association between two variables [4]. Correlation was treated as the first method to take a first glance at which independent variables play bigger roles in contributing to the dependent variable.

2.3. Partial Correlation

A more precise version of correlation called partial correlation was used as the second step to ensure the analysis's accuracy and a correct interpretation of the independent variables' contribution to the dependent variable. Partial correlation is able to hold the effects of a third variable constant when

DOI: 10.54254/2755-2721/73/20240370

identifying the correlation between two variables [5]. Partial correlation helps to eliminate the confounding variables that the data might have. Even though correlation does provide an estimation of the relationship between two variables, variables may be interconnected in the real world. The observed correlation between two variables could be influenced by a common third variable. Partial correlation provides a solution to this problem by measuring the relationship between two variables while controlling the effects of other variables, which efficiently removes the influence of confounding factors.

2.4. K-means Clustering

The third step was K-means Clustering. It was chosen with the aim of exploring its function of feature importance. The k-means algorithm treats all features fairly and sets the weights of all features equally when evaluating dissimilarity [6]. Clustering can indirectly highlight the importance of specific independent variables. The independent variables that significantly differentiate the clusters are likely to be more relevant in explaining the underlying patterns in the data.

2.5. Principal Component Analysis

Principal Component Analysis (PCA) was also utilized with the aim of reducing the dimensionality of the data and excluding the unnecessary independent variables. The PCA can help to define which combination of data best represents the concept that the research wants to measure. PCA technique reduces the dimensionality of the dataset and condenses those dimensions into a singular variable [7].

2.6. Multiple Regression

Multiple regression was run on the data. It was used to identify the relationship between one dependent variable with multiple related independent variables. It can be a great tool for prediction studies as it can help to develop the formula predicting the dependent variable with the fixed values of the independent variables [8].

2.7. Lasso Regression

To avoid the overfitting problem that multiple regression might present, Lasso Regression was also employed. Standard regression methods like multiple regression tend to generate models that would be overfit in terms of the number of variables to be included in the model, and also overestimate how well the model would perform in terms of utilizing the included variables to explain the observed variability. LASSO regression has the ability to identify the variables and their corresponding regression coefficients that would create a model that minimizes the prediction error. It is done by establishing a constraint on the model parameters, which limits the regression coefficients to zero. It forces the sum of the absolute value of the regression coefficients to be less than a fixed value (λ) [9]

2.8. Support Vector Machine and Random Forest

Lastly, learning algorithms SVM (support vector machine) and random forest were also incorporated in the dataset analyses to see new insights with results that are different from previous methods. SVM would provide the regression that best fits the data points while allowing for a certain margin of error. Random forest provides advantages like reduced overfitting, accurate predictions [10], and the ability to deal with various types of data that previous regression could not provide.

3. Results

As shown in Table 1 below, Correlation suggested that BMI, blood pressure, and smoking have a strong positive correlation with health insurance claims. The variables gender and children also seemed to have a positive correlation with gender, but their significance was a lot smaller.

DOI: 10.54254/2755-2721/73/20240370

Table 1. Correlation Between Independent Variables and the Dependent Variable

	Health Insurance Claim
Age	-0.0279
Gender	0.0597
BMI	0.2003
Blood Pressure	0.5313
Diabetic	-0.0784
Children	0.0661
Smoker	0.7872
Region_1	-0.0124
Region_2	-0.0263
Region_3	-0.0733
Health Insurance Claim	1.0000

In Table 2, partial correlation suggested very similar results. BMI, blood pressure, and smoking are still the three variables that affect the claim variable most. Age and children still had a positive correlation with the claim, while still having a much smaller effect compared to BMI, blood pressure, and smoking.

Table 2. Partial Correlation Between Independent Variables and the Dependent Variable

	Health Insurance Claim
Age	0.0282
Gender	-0.0025
BMI	0.3011
Blood Pressure	0.3370
Diabetic	-0.0214
Children	0.1259
Smoker	0.7540
Region_1	-0.1410
Region_2	-0.1022
Region_3	-0.0907
Health Insurance Claim	1.0000

What's important to notice from k-means clustering is that the variables that have more differentiated clusters are BMI, blood pressure, and smoking, which align very well with the results from partial correlation. It confirms the thoughts that BMI, blood pressure, and being a smoker are the 3 variables that are more significant in contributing to the increase in health insurance claims. Unfortunately, PCA was not able to reduce the dimensionality of the data and exclude the unnecessary x variables. The reason was attributed to the low correlation between features. Similar to earlier results, multiple regression suggested that BMI, blood pressure, and being a smoker strongly positively contribute to the increase of claims. The multiple regression further confirmed the findings from partial correlation, clustering, and PCA. What's interesting is that children and age also appear to be two variables that are significant in contributing to the increase in health insurance claims. It was not expected to see such a strong influence that these two variables could have on health insurance claims earlier. Especially for children, which has a 692.05 coefficient. Similar results were received from lasso regression, which also indicates that children are a variable that has a high correlation with health insurance claims with a high coefficient. The results from the random forest and support vector machine are shown in Table 3. By comparing the coefficients obtained by these different methods, it is found that although there are many

DOI: 10.54254/2755-2721/73/20240370

subtle differences, the overall trend is consistent. For example, the influence of the variable "smoker" is very significant in several methods. The influencing factors of those "region" variables are basically the same as well.

Table 3. The result of Random Forest

	IncNodePurity
Age	8942535658
Gender	1381792947
BMI	0.3011
Blood Pressure	0.3370
Diabetic	-0.0214
Children	0.1259
Smoker	0.7540
Region_1	-0.1410
Region_2	-0.1022
Region_3	-0.0907

4. Discussions

The results from this study were consistent. It showed that BMI, Blood Pressure, and smoking are the three biggest contributors to the increase in health insurance claims. Even though certain other factors like age, gender, and the number of children also had positive contributions to increasing health insurance claims, their impacts were relatively negligible. Different from the results that the earlier similar study generated in 2020 by Bhardwaj and Anand that age and being a smoker are the biggest contributors to increasing health insurance, this study got the result that BMI, blood pressure, and being a smoker are the factors that contribute to the increase of insurance claim the most.

The strength of this study is that this study included more independent variables. This study included eight independent variables for the analysis, whereas the previous one included only five. The lack of independent variables in the previous study might prevented it from conducting a more comprehensive analysis of the issue, which is considering more independent variables in the formula. The second strength might be that the previous study only had three regression analyses. This study included seven different ways of analysis, which implies that this study provided analysis that is more multi-angel and muti-dimensional and thus provides results that are more relatable for people in real life to consider.

The limitation of this study is that it also only covered a limited number of independent variables to consider as the biggest contributor to increasing health insurance claims. There could be many more variables that actually affect the health insurance claim rate that this study was not able to include in consideration. Future research on this issue should include as many related independent variables as possible to ensure the comprehensiveness of the study and to get more meaningful insights into the issue. Knowing that BMI, blood pressure, and being a smoker are the three main contributors to increasing health insurance claims, people can adjust their lifestyles accordingly. If they want to keep their insurance claim amount in control, they should pay more attention to these factors. Health insurance, on the other hand, can also be more aware of which population they should pay extra attention to when determining what prices, they should charge. Insurance companies should be sensitive to individuals with high values in those three contributing factors to health insurance claim amounts to ensure the company's profitability.

References

[1] Dieleman, Joseph L., et al. "US Health Care Spending by Payer and Health Condition, 1996-2016." *Jama Network*, 2020

[2] Published: Nov 10, 2021. "Section 1: Cost of Health Insurance - 9805." *KFF*, 10 Nov. 2021,

DOI: 10.54254/2755-2721/73/20240370

[3] *Health Insurance Claim Prediction Using Artificial Neural Networks*, www.researchgate.net/publication/342605306_Health_Insurance_Claim_Prediction_Using_Artificial_Neural_Networks. Accessed 17 Sept. 2023.
[4] Sedgwick, Philip. "Pearson's Correlation Coefficient." *The BMJ*, British Medical Journal Publishing Group, 4 July 2012
[5] Brown, Bruce L., and Suzanne B. Hendrix. *Partial Correlation Coefficients*, 2005
[6] Blum, A.L., et al. "Developing a Feature Weight Self-Adjustment Mechanism for a K-Means Clustering Algorithm." *Computational Statistics & Data Analysis*, North-Holland, 15 Mar. 2008.
[7] Caterina, Francisco, and Urdinez Labrín. "Principal Component Analysis: 15: R for Political Data Science: CAT." *Taylor & Francis*, Taylor & Francis, 17 Nov. 2020.
[8] "Multiple Regression." *Google Books*, Google, Accessed 17 Sept. 2023.
[9] Ranstam, J, and J A Cook. "Lasso Regression." *OUP Academic*, Oxford University Press, 7 Aug. 2018.
[10] Rigatti, Steven J. "Random Forest." *Allen Press*, Allen Press, 1 Jan. 2017.