Factors influencing Health insurance cost using Regression and Classification

Ajayi David

Saint Louis University

AA 5300 Final Project

Dr Micheal Fisher

March 10, 2023

Introduction

Overview

Health and wealth are related in many different ways which could be seen by the sheer cost of unhealthy habits. For instance, eliminating a regular smoking or junk food habit would save individuals thousands of dollars annually, plus interest. There are also savings over the long term for the rest of someone's life. The Centers for Disease Control estimates that a 10% weight loss could reduce an overweight person's lifetime medical costs by $2,200 to $5,300. Delaying the onset of diabetes can save thousands of dollars annually in increased medical costs.

Additionally, financial problems can affect a person's health status and vice versa. For example, overdue medical bills can result in physical symptoms of stress (e.g., migraines, insomnia, and anxiety) and/or delayed or inadequate treatment. Furthermore, high health costs could lead to a poor credit history and/or bankruptcy and reduced income available to save for retirement and other financial goals.

The health insurance dataset used in these focuses on these health factors as they influence the cost medical bills. The dataset was gotten from Kaggle and published by Brett Lantz. It could be found at the website below:

https://www.kaggle.com/code/ruslankl/health-care-cost-prediction-w-linear-regression/report

This dataset consists of 1338 rows and 7 Variables. The outcome variable is ‘Charges’ while others are predictor variables.

Objectives behind the collection of the data.

The dataset was collected with aim of prediction of medical insurance costs in relation to the people's lifestyle and health habits. This could be due to fact that the health and wealth relationship possessed by people in poor health often causes them to spend thousands of dollars on health care insurance costs, money that could otherwise have been invested in other profitable ventures.

Questions that audience interested in the dataset and its analyses might seek to see answered

Some questions to be answered are from this analysis are:

- What health factors have the greatest influence on individual medical costs?
- Does BMI play a significant role in health insurance costs?
- What region in the US spend the most on medical bills through health insurance?
- Does age significantly increase medical insurance costs?

Variables types ,roles (predictor or outcome) and description

	Variable	**Description**	**Type**	**Role**
	Age	Insurance contractor age, years	Numeric	Predictor
	Sex	Insurance contractor gender	String	Predictor
	BMI	:Body mass index	Numeric	Predictor
	Children	Number of children covered by health insurance	Numeric	Predictor
	Smoker	Smoking, [yes, no]	String	Predictor
	Region	Beneficiary's residential area in the US	String	Predictor
	Charges	Individual medical costs billed by health	Numeric	Outcome

		insurance		

Regression analytical techniques

Simple Linear Regression: Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable

Multiple Linear Regression: This involves extending the simple linear regression model so that it can directly accommodate multiple predictors.

Decision Trees: The regression decision trees (Random Forest, Bagged and Boosted) are used for the task of regression which can be used to predict continuous valued outputs instead of discrete outputs

Classification analytical techniques

Logistic Regression: Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

LDA : This is used if the paramater estimates for logistic regression tend to be unstable when the classes are well separated.

QDA: QDA is recommended if the training set is very large or if the assumption of a common covariance matrix for the K classes is not realistic

KNN: KNN attempts to estimate the conditional distribution of Y given X and then classify a given observation according to the highest estimated probability.

Analyses

Methods of analyses on the dataset, explanation and rationale behind using method

MLR	Will be used	The regression method can be used since we have multiple predictor variables
Random Forest	Will be used	The regression method can be used since we have a continuous outcome variable (Charges) from the insurance
Bagged	Will be used	The regression method can be used since we have a continuous outcome variable (Charges) from the insurance
Boosted	Will be used	The regression method can be used since we have a continuous outcome variable (Charges) from the insurance
Logistic Regression	Will be not be used	The Classification method will not be used given our outcome variable is continuous and not categorical

Clustering or dimensionality reduction/ Subsampling

Clustering is the assignment of variables to clusters while making sure that in different groups they are not similar. It is also considered an unsupervised task as it aims to describe the hidden structure of the variables. Additionally, dimensionality reduction is used to reduce variables under consideration, where each feature is a dimension that partly represents the objects. Therefore, they were not applicable to the insurance dataset given its relatively small size and the clarity of variables.

Furthermore, given subsampling is a method that reduces data size by selecting a subset of the original data it was not applicable to the dataset its relatively small size of 1338 rows and 7 Variables

Summary of results

Details of the validation method used, Model formulas and explanation of model selection

Method	Formula	Basis of Model selection and evaluation
MLR	k-fold CV, with repetitions, using Caret FitControl <- trainControl(method = "repeatedcv", number = 5, repeats = 3)	Repeated k-fold cross-validation was done to improve the performance of a machine learning model. This was achieved by splitting the training dataset into sub train and test sets. Therefore, the choice for evaluating performance of models is (Mean Absolute Error) MAE. Given it shows how accurate our predictions are and the amount of deviation from the actual values

Method	**Formula**	**Basis of Model selection and evaluation**
Random Forrest	k-fold CV, with repetitions, using Caret FitControl <- trainControl(method = "repeatedcv", number = 5, repeats = 1)	Repeated k-fold cross-validation was done to improve the performance of a machine learning model. This was achieved by splitting the training dataset into sub train and test sets. Therefore, the choice for evaluating performance of models is (Mean Absolute Error) MAE. Given it shows how accurate our predictions are and the amount of deviation from the actual values

Method	**Formula**	**Basis of Model selection and evaluation**
Boosted	k-fold CV, with repetitions, using Caret FitControl <- trainControl(method = "repeatedcv", number = 5, repeats = 1)	Repeated k-fold cross-validation was done to improve the performance of a machine learning model. This was achieved by splitting the training dataset into sub train and test sets. Therefore, the choice for evaluating performance of models is (Mean Absolute Error) MAE. Given it shows how accurate our predictions are and the amount of deviation from the actual values

Method	Formula	Basis of Model selection and evaluation
Bagged	k-fold CV, with repetitions, using Caret FitControl <- trainControl(method = "repeatedcv", number = 5, repeats = 1)	Repeated k-fold cross-validation was done to improve the performance of a machine learning model. This was achieved by splitting the training dataset into sub train and test sets. Therefore, the choice for evaluating performance of models is (Mean Absolute Error) MAE. Given it shows how accurate our predictions are and the amount of deviation from the actual values

Conclusion

MLR Results

Method	**RMSE**	**Rsquared**	**MAE**
MLR	0.5324427	0.7638333	0.3580949

Variables	**MLR Variable Improtance**
smoker yes	100.000
age	36.422
bmi	20.160
children	5.002
regionsouthwest	3.843
regionsoutheast	3.079
regionnorthwest	1.618
sexmale	0.000

Decision Trees Results

Method	RMSE	Rsquared	MAE
Random Forest	0.3755054	0.8160649	0.2116961
Boosted	0.3652181	0.8253147	0.2040897
Bagged	0.3755054	0.8160649	0.2218860

	Random Forest	Boosted	Bagged	Aggregate Variable Importance
Smoker yes	100.0000	100.00000	100.0000	100.0000
bmi	24.4792	22.93034	24.4792	23.96291333
age	20.8584	17.30435	20.8584	19.67371667
children	2.2527	1.28005	2.2527	1.928483333
regionnorthwest	0.1337	0.09938	0.1337	0.12226
regionsouthwest	0.0000	0.08804	0.0000	0.029346667
sexmale	0.1330	0.05754	0.1330	0.107846667
regionsoutheast	0.1076	0.00000	0.1076	0.071733333

Description of Best Model results

Firstly, we discovered our predictors have no variances which were zero or close to zero

Given the results shown in table above, among three models (MLR, Random Forest, Bagged and Boosted) the best performing model is the **Boosted Tree Model** as it has the minimum Mean Absolute Error) MAE value in **(0.2040897)** implying it possesses the least errors and deviations from the perfect model.

The importance statistics/variation in parameter estimates associated with the model imply to a decision-maker.

Furthermore, from the tables above it is seen the predictors that contribute the most to the reduction of residual sum of squares. The results gotten showed that the variable **'(Smoker yes)'** is the most important while second most important is the variable **'age'** and the third most important is the variable **'bmi'.** This could be attributed to their influence on reducing the error in their respective models.

Theoretically, Health care costs for smokers at a given age are as much as 40 percent higher than those for nonsmokers. This is because smokers tend to suffer more from a large variety of diseases thereby incurring higher medical costs. Furthermore, we could also deduce from the results that age plays a significant role in the amount of medical bills accrued by individuals (Barendregt, 1997). This because even nonsmokers live longer and could incur more health costs at advanced ages. Finally, from various studies carried out it has been observed that obesity has become a major public health crisis in the United States. Research to date has consistently demonstrated a correlation between obesity and higher medical costs for a variety of U.S.

subpopulations and specific categories of care. This points to why the **'bmi'** variable serves as one of the most important factors in the model.

Deeper insights into how the dataset can help a decision-maker who is associated with the context within which the dataset was collected

Finally, decision makers associated with this data set would be average human being given we all strive to live healthy and also minimize cost spent on medical insurance bills. Therefore, people should endeavor to exercise regularly, reduce or avoid smoking, and manage other risk factors which would culminate in them spending far less money on medications and medical costs.

References

Barendregt, J.J., Bonneux, L. and van der Maas, P.J. (1997) “The health care costs of smoking,” *New England Journal of Medicine*, 337(15), pp. 1052–1057. Available at: https://doi.org/10.1056/nejm199710093371506.

Cawley J;Biener A;Meyerhoefer C;Ding Y;Zvenyach T;Smolarz BG;Ramasamy A; (no date) *Direct medical costs of obesity in the United States and the most populous states, Journal of managed care & specialty pharmacy*. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/33470881/