Performance Analysis of Supervised Learning for Medical Insurance Costs

By

Jinghui Yao, Qiaolin Liang, Yifan Wang, Ruizhi Xu, Haofei Lai, Yonghao Li

Word Count: 3593

Faculty of Information

University of Toronto

Abstract

Insurance companies devise health plans based on the insurer's personal conditions. Finding out the factors with higher impact on the insurer's medical expenditures can help the insurance company to generate more reasonable health plans and subsequently maximize revenue. In this paper, we explore influential factors contributing to each individual's annual medical expenditure. Using descriptive analysis and supervised learning including multiple linear regression and classification methods, we found that smokers tend to have a higher medical cost while other factors (such as region and obesity) are less impactful. Most importantly, we are able to visualize and predict the medical expenditures spent for each individual based on characteristic features, which might be of interest to insurance companies for future improvements to healthcare plans.

Introduction

Healthcare is an act of using necessary medical procedures to improve one's well-being, which may include surgeries, therapies and medications. These services usually are provided by a healthcare system that consists of hospitals and physicians. According to *Time*, currently the medical cost spent per person in the U.S is five times more than that in Canada, and the U.S administrative costs for healthcare came out to $812 billion in 2017. The amount spent in healthcare is significant, not only does each individual pay a lot for their insurance plans, the government has also been devoting resources to medical expenditures for a long time.

Under the healthcare law, insurance companies are not allowed to set the plan for each individual depending on their health, medical history and gender (Healthcare.gov). Therefore, we are interested in what factors contribute more to the medical costs of each individual. Our research questions include the following:

1. How much does obesity or smoking status affect medical expenditures?
2. What is the most significant factor that affects a patient's medical expenditures?
3. Does the difference in residential regions impact medical costs?
4. Can we predict personal medical expenditures using supervised learning models?

To understand more about the dataset, we first explore the relationships between the dependent variable (medical charges) and each independent variable (number of children, smoking status, BMI, etc.) to visualize the amount of influence of each IV. In order to further explore our research questions, we then build a multiple linear regression model to better understand the underlying relationships between all independent variables and the dependent variable medical cost. We also apply classification to explore our research questions as another approach. We use individual classifiers (including random forest, SVM, naive bayes, and logistic

regression) along with an ensemble model and decision tree model to thoroughly study the dataset. Additionally, we build a prediction function using the model with the highest accuracy to predict whether an individual's medical expenditures will be higher or lower than the mean of log-scale charges.

Literature Review

Obesity is the first factor that we are trying to examine whether it could have a direct impact on medical expenditures since it could lead to chronic illnesses which are common in this era. According to data from Medical Expenditure Panel Survey from 2005 to 2010, an average increase of 14.3 percent of medical expense, from $3070 to $3508, caused by obesity occurred in obese adult in the US. However, since people with obesity are not randomly distributed around the world, it is possible that other factors such as difficulty accessing healthcare providers or family situations also have an effect. A research using instrumental variable models to explore the causal effect of obesity on medical care costs suggest that there might be nonlinearities between the body max index (BMI) and medical expense (Biener, Cawley & Meyerhoefer, 2017).

Another factor we want to focus on is the smoking status as smoking puts people at risk of different ailments. Even though adult cigarette smoking prevalence has been decreasing for the past 50 years, it is still the leading preventable cause of disease in the US. Studies show that among all costs caused by smoking, medical expenses accounted for a large proportion. Besides, an estimated 5% to 14% of the annual healthcare expenditure falls into smoking related disease. The Annual Healthcare Spending Attributable to Cigarette Smoking built a logistic model and a linear model using data from 2006 to 2010 showing that 8.7% of the spending was used for smoking-attributable illness and most of this spending was paid by the public in the US. The report also explored other factors such as social characteristics and health-related behavior and the results showed that current smokers tend to be young, male, and unmarried, so these types of people might have higher healthcare expenses (Xu, Bishop, Kennedy, Simpson & Pechacek, 2015).

There is one special factor that is underestimated by researchers, the geographical factor. Victor Fuchs, Ph.D., who was invited to conduct a series of research to explore factors related to healthcare expenditures in the USA, pointed out that determinants could be divided into medical factors and non-medical factors which contain geographical information. He emphasized that by far, geography, or where a person lives, is explored by few researchers. In one of his reports, he indicated the existence of variation in the cost of medical care across different areas in the USA. In fact, big cities might have a higher average healthcare spending. Besides, areas with a high number of sick elderly people with higher death rates tend to have high spending. So generally speaking, there is a relationship among mortality, geography, and healthcare (Fuchs, 2011). However, this conclusion could only be used for elder people. But from the Health Care Expenditures per Capita by State of Residence ("Health Care Expenditures per Capita by State of

Residence", 2014), we could say that the average spending for residents in the northeast part of the USA are higher than the southwest part. More research needs to be done in this field. As a result, it is meaningful to examine the variance of medical spending among different areas of people for all ages.

Generally speaking, factors with contribution to healthcare spending could be separated into four dimensions including medical factors like disability, chronic disease, or deadliest disease; socioeconomic factors like education, income, or occupation; health-related behavior like smoking and drinking; and geographic factors. If we look at these factors separately, we might easily make some findings. However, these factors might have a coeffect in terms of the influence of medical spending and the interpretation of outcome might need to be connected to the context provided, or there is no direct explanation on how the mixture of these factors can affect the medical spending. For instance, people with low socioeconomic status have more opportunity to have worse health which could imply a higher medical spending. But those with high income might put more money on medical care to enjoy a better life (Wunderlich, 2010). As such, a comprehensive consideration of various factors in this research is in need.

As the overall cost of healthcare is increasing each year and is expected to keep increasing, it is good for people to understand why they received such a high bill from the insurance company regarding healthcare. Apart from that, Victor also mentioned that understanding the influential factors of healthcare expenditures is important to the development of health policies (Fuchs, 2011). What's more, the result of the research could provide a hint for insurance companies to better set their price, as there are rising concerns towards increasing health insurance premiums that the average coverage for a family rose to $19,616 in 2018. An unacceptable and unaffordable insurance price might lead to patients avoiding going to a doctor (Probasco, 2019).

Dataset

Dataset source

Dataset[1] used in this study comes from the book "Machine Learning with R" by Brett Lantz, which provides an introduction to machine learning using R. The dataset is simulated on the basis of demographic statistics from the US census Bureau. It consists of 1338 respondents, where the age of primary beneficiary is from 18 to 64.

[1]Dataset source from:
https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

Measurements

- Independent Variables

Six independent variables used in this research include age of primamry beneficiary (Age), insurance contractor gender (Sex), body mass index(BMI), number of children covered by health insurance(Children), whether insurance contractor is a smoker (Smoker), and the beneficiary's residential area in US (Region). Gender is a nominal variable.

- Dependent Variable

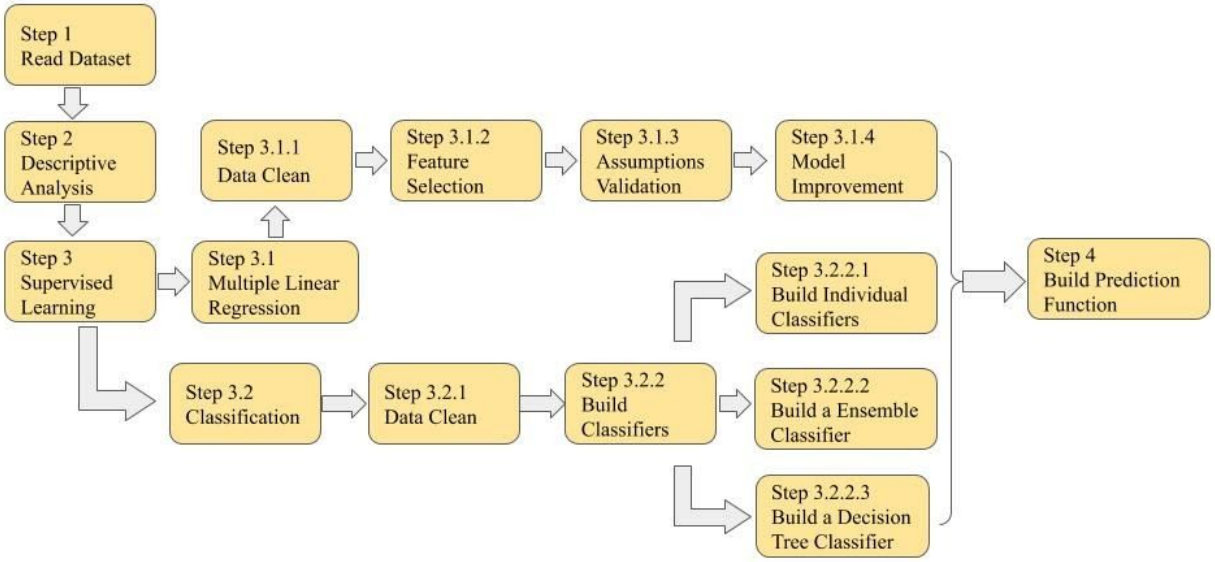
The dependent variable used in this research is the individual medical costs billed by health insurance (Charges), which is a continuous variable.

Table 1: Individual Attributes Description in Dataset

Attribute	Measurement	Date Type
Age	Independent	Numerical
Sex	Independent	Categorical (Male/Female)
BMI	Independent	Numerical
Children	Independent	Numerical
Smoker	Independent	Categorical (Yes/No)
Region	Independent	Categorical(Southwest/Northwest/Southeast/Northeast)
Charges	Dependent	Numerical

Approach

In our research, we first use descriptive analysis to explore demographic characteristics among all variables. In order to discover and predict the individual medical costs billed by health insurance (Charges) based on the value of all six variables, we plan to analyze the dataset according to supervised learning methods including multiple linear regression and classification. Last but not least, we develop a prediction function model according to the supervised learning methods.



Block Diagram 1: Steps of Research Approach

Step 1: Read Dataset

We import the dataset to R studio using the function *read.csv*.

Step 2: Descriptive Analysis

This step we are using *str()* and *summary()* to have a descriptive analysis of the dataset, followed by a box plot analysis. Throughout the box plot analysis we discovered that smoking status has a potential large impact on individual medical costs. Further analysis can be found in the result section.

Step 3: Supervised Learning

Step 3.1: Multiple Linear Regression

Step 3.1.1: Data Clean for Linear Regression

In order to further develop the relationship between smoking status and individual medical costs, as well as relationships between each independent variable to prevent the multicollinearity problem, we plotted a correlation matrix to have a perspective.

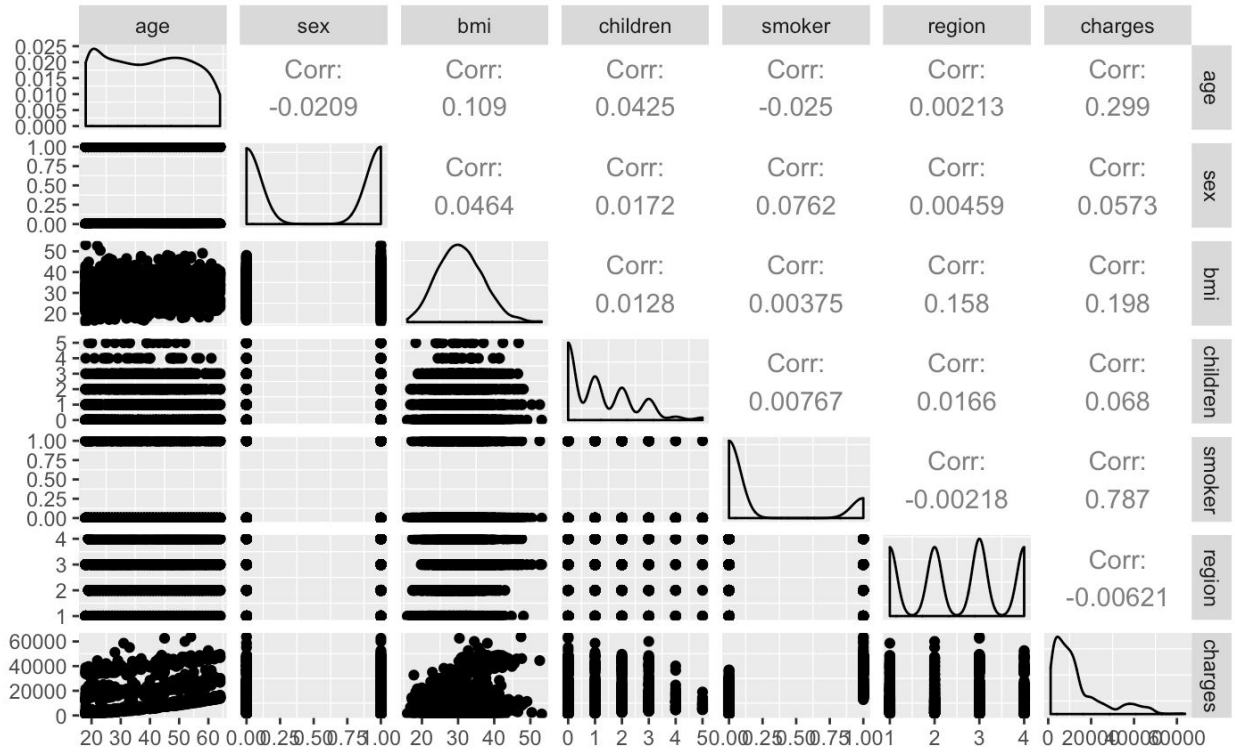


Diagram 2: Correlation Matrix

Diagram 2 not only shows that the smoking status has a strong positive relationship with individual medical costs but also indicates that none of the independent variables are highly correlated with one another, thus it is safe to say that the dataset does not have a multicollinearity problem.

Furthermore, we would like to ensure the skewness of the dependent variable matches the linear regression assumption. Thanks to function *skewness()*, we are able to find that the skewness score of the dependent variable is 1.51, which means the dependent variable is skewed right. Diagram 3 proved this conclusion.

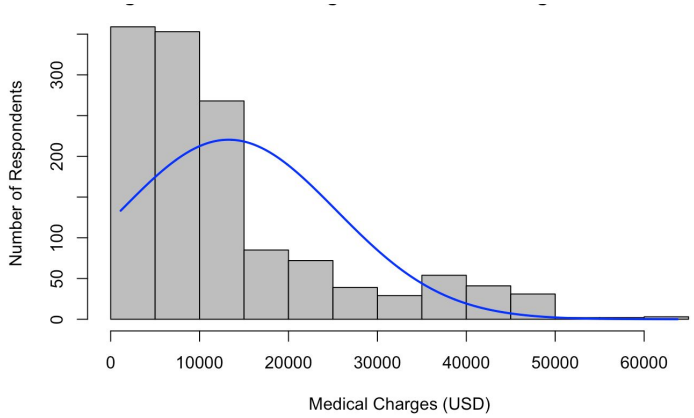


Diagram 3: Normal Histogram of Medical Charges Distribution

Hence, we used log-scale transformation to normalize the dependent variable.

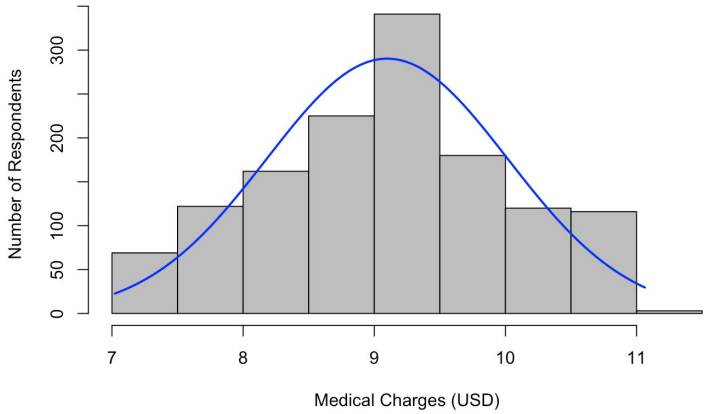


Diagram 4: Normal Histogram of Log-Scale Medical Charges Distribution

According to diagram 4, the dependent variable is in normal distribution after the log-scale transformation.

Step 3.1.2: Feature Selection

Before building the linear regression model, we implement backward feature selection to decide which variables should be included in the model. As a result (table 2), all predictors are significant at 0.05 alpha level, therefore we will keep all independent variables for the linear regression model.

Table 2: Preliminary Linear Model Summary after Backward Feature Selection

Coefficients	Estimate	P-Value
Intercept	7.031	< 2e-16
Age	0.035	< 2e-16
Sex-Male	-0.075	0.002
Bmi	0.013	2.42e-10
Children	0.102	< 2e-16
Smoker-Yes	1.554	< 2e-16
Region-Northwest	-0.064	0.068
Region-Southeast	-0.157	8.08e-06
Region-Southwest	-0.129	0.001
	Adjusted R-Squared	0.767
	Model P-Value	< 2e-16

The summary of the preliminary model illustrated adjusted R-squared is 0.767, which means about 76.7% of the variance within the dependent variable can be explained by the model. P-Value of the model is smaller than the significant level (alpha = 0.05), the whole model is significant.

Step 3.1.3: Assumptions Validation

The following four diagrams are used to validate linear assumptions of the preliminary model.

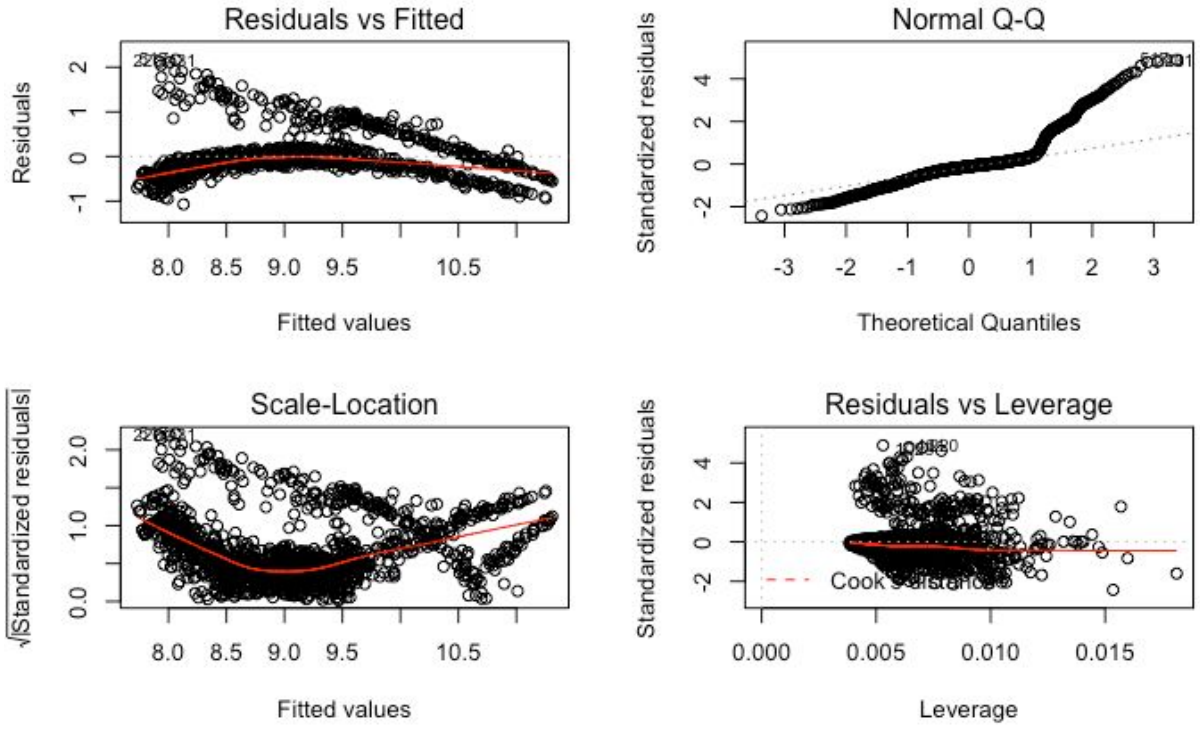


Diagram 5: Model Validation Plot

According to diagram 5, the residual vs fitted plot shows the linearity doesn't satisfy as the red line is not horizontal at roughly 0. Q-Q plot shows points on the top right partial did not follow the straight line, indicating that the normality assumption is not met. Additionally, the scale-location plot does not have a horizon line with randomly-scattered points, meaning that the model does not have a constant variance, further transformation of model is required. The residual vs leverage plot shows no outliers exceeding Cook's Distance, but there are influential points in the model.

Step 3.1.4: Linear Model Improvement

As we mentioned in earlier steps, the dataset has potential influential points and the dependent variable needs further transformations. In this step, we will eliminate these two problems and refit the model to try to get a better result.

For the dependent variable, we applied box-cox transformation. In specific, we used *boxcox()* function and found that when lambda equals 2, the log-likelihood is the highest. We then use 2 as the dependent variable exponent.

Then we removed influential points according to the difference in fit method which suggested we exclude any data points having an absolute value of its DFFITS value greater than cut point 0.174. The equation of how we got 0.174 is shown in diagram 6.

$$2 * \frac{\sqrt{(p+1)}}{(n-p-1)}$$

Diagram 6: Equation of Cut Point, n is number of observations and p is number of predictors

Step 3.2: Classification

Step 3.2.1: Data Clean

In order to better apply classification methods to our dataset, we need to transform factor IVs into numeric formats. For the target variable (Charges), as mentioned above, we continue to use log-scale transformation because it produces a normalized distribution and is more scientific for classification. After log transformation, the number of cases below and above mean is similar to each other, so we use the mean value after log-scale transformation as the cut-off for categorizing charges. We calculated the exponent of the mean after log-scale transformation and found out the cut-off line of high and low charges is $8,943.29. The value transformations are shown below:

Table 3 : Value Transformation

Variable Name	Original Value	New Value
Sex	Male/Female	1/0
Smoker	Yes/No	1/0
Region	Northeast/Northwest/Southeast/Southwest	1/2/3/4
Charges	>=8943.29/<8943.29	Charge Level: High/Low

Step 3.2.2: Build Classifiers

Step 3.2.2.1: Build Individual Classifiers

The dataset is randomly divided into training and testing sets with a ratio of 70:30.

Then we build individual classifiers to see their performance. The modeling algorithms we used for individual classifiers are random forest, SVM, naive bayes and logistic regression. We used a train control of 3 repeats of 5-fold cross-validation for all these individual classifiers.

Step 3.2.2.2: Build an Ensemble Classifier

In order to improve the prediction performance of our model, we continue to build an ensemble of all previous models. The correlation matrix is shown below:

Table 4: Correlation Matrix of Algorithms

Modeling Algorithms	Random Forest	SVM	Logistic Regression	Naive Bayes
Random Forest	1	0.73	0.58	0.32
SVM	0.73	1	0.78	0.55
Logistic Regression	0.58	0.78	1	0.44
Naive Bayes	0.32	0.55	0.44	1

For correlation analysis, we assume that correlation higher than 0.75 is considered as strongly correlated. According to the correlation matrix, we found out that the correlation between SVM and logistic regression is 0.78, which is higher than 0.75 and can be considered as strongly correlated. Since the accuracy of logistic regression is lower, we will remove it to build an updated ensemble model.

Step 3.2.2.3: Build a Decision Tree Classifier

To better visualize the impact of each column on medical costs, we also built a decision tree model. To better suit the model, the following transformation for the remaining continuous variables was applied:

Table 5: Value Transformation

Variable Name	Original Value	New Value
Age	<=30 / >30 and <60 / >=60	Younger Age / Middle Age / Older Age
BMI	<18.5 / >=18.5 and <25 / >=25 and < 30 / >=30	Underweight / Healthy / Overweight / Obese

Then we built the model using decision tree algorithm and visualized the tree structure.

Step 4: Build Prediction Function

We then compared the performances of all models and built a prediction function based

on the model with the best performance.

Results

Descriptive Analysis

In this process, we explore the dataset from multiple perspectives. First, we use *str()* and *summary()* function which gives us some information about the dataset including min, max, mean, median, 1st qt, and 3rd qt. From the summary statistics table (table 6) of the dataset, we can see that the respondents are between 18-64 years old, having 0-5 children, with annual medical costs between $1,122 and $63,770 (mean value is $13,270). Also, by using *is.na()* function and summarizing its output, we discover that the dataset does not contain any missing value.

Table 6: Individual Attributes Description in Dataset

Numeric Columns	Min	Median	Mean	Max
Age	18	39	39.21	64
Bmi	15.96	30.4	30.66	53.13
Children	0	1	1.905	5
Charge	1,122	9,382	13,270	63,770
Categorical Columns				
Sex	2 levels, male count: 676, female count: 662.			
Smoker	2 levels, no count: 1064, yes count: 274			
Region	4 levels, northeast: 324, northwest: 325, southeast: 364, southwest: 325			

We then use the ggplot function to build two boxplots that can help us understand our independent variables better. The first boxplot (diagram 7) is the number of children vs. healthcare costs. To make it easier to plot, we made the numeric column children into factor by using *as.factor()* function.

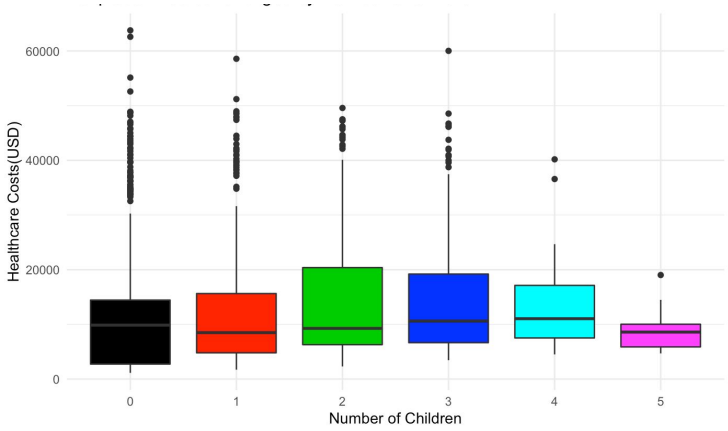


Diagram 7: Boxplot of Medical Charge by Number of Children

From the plot result we can see when having children numbers 2 and 3 the healthcare cost is typically higher while having children number 0 and 5 the healthcare cost is usually lower. The plot also indicates that six categories all have potential outliers. The second box plot (diagram 8) indicates clearly that respondents who smoke have an average healthcare cost that is almost 4 times higher than those who do not smoke.

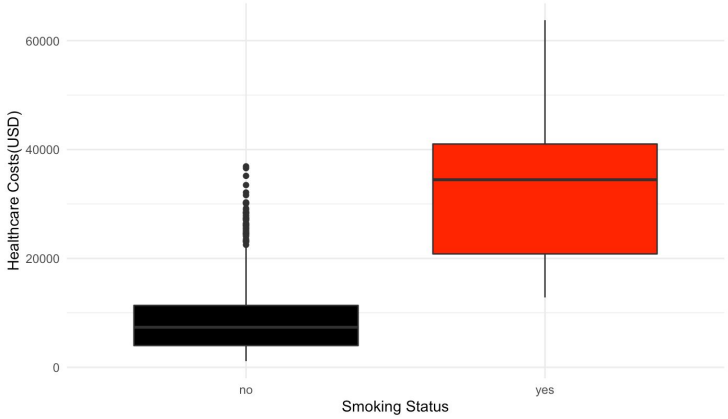


Diagram 8: Boxplot of Medical Charge by Smoking Status

Multiple Linear Regression Result

According to our approach, we can refit the model using the fixed dataset, all independent variables are being included in the improved model.

Table 7: Improved Linear Model Summary

Coefficients	Estimate	P-Value

Intercept	42.6005	< 2e-16
Age	0.6898	< 2e-16
Sex-Male	-1.3126	2.57e-06
Bmi	0.2123	< 2e-16
Children	1.8461	< 2e-16
Smoker-Yes	30.0797	< 2e-16
Region-Northwest	-1.4388	0.0003
Region-Southeast	-2.4408	1.58e-09
Region-Southwest	-1.9565	1.03e-06
	Adjusted R-Squared	0.9105
	Model P-Value	< 2.2e-16

In table 7, we can see that the model has reference groups set at “Female” at sexual, “No” at smoker status, and “Northeast” at region. The adjusted R-squared has dramatically improved to 0.9105, meaning that about 91.05% of variance within dependent variables can be explained by the improved model. The p-value of the model is smaller than the alpha level of 0.05, which indicates significance. However, as we redo the model validation using the new regression model, some of the assumptions are still not satisfied.

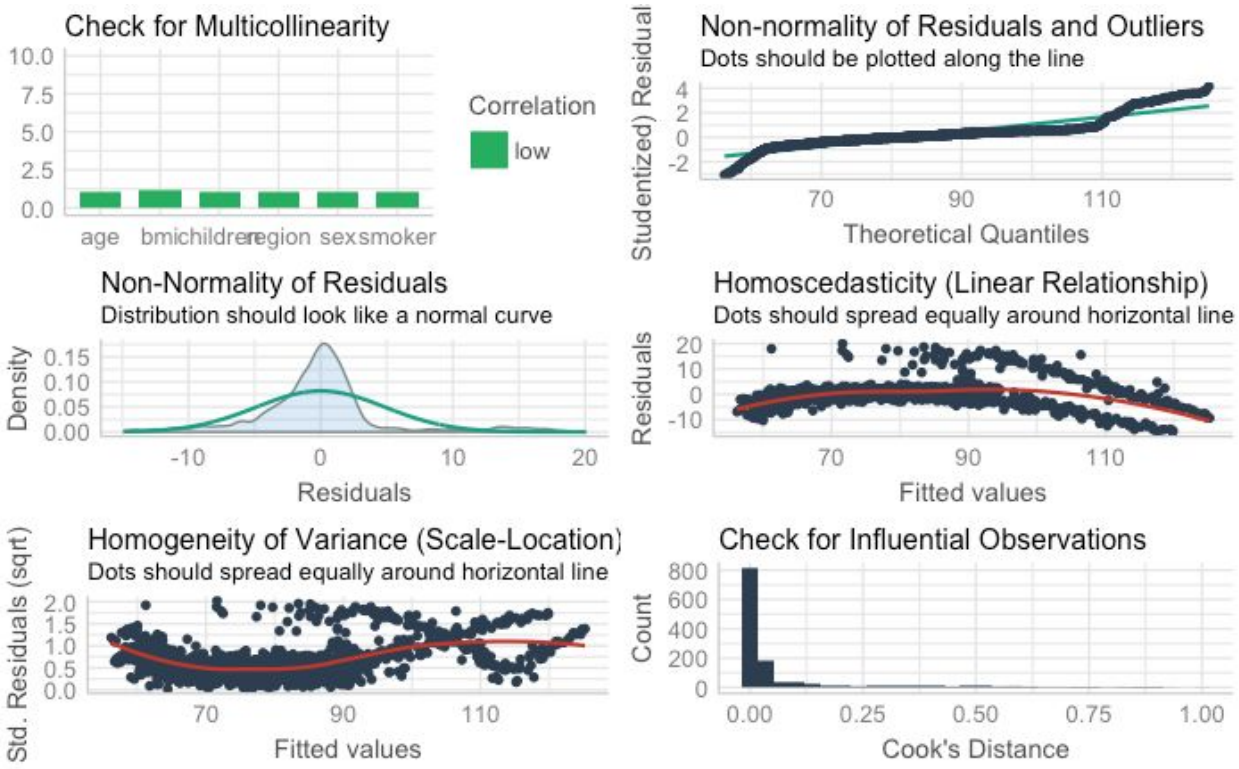


Diagram 9: Improved Model Validation Plot

Diagram 9 shows that the red line in the homoscedasticity plot is still not horizontal even after several transformations. The residual points seem to follow a weird downward curve, suggesting that they do not satisfy the constant variance assumption. Because of this finding, though linearity and normality assumptions are met, we suspect that there are non-linear relationships within the dataset, rendering the linear regression models to be less than useful.

Classification Result

According to Approach Step 3.2.2.1, the accuracies of random forest, SVM, naive bayes and logistic regression are 0.9353, 0.9328, 0.893, 0.8955 respectively. Random Forest classifier has the highest accuracy among the four.

According to Approach Step 3.2.2.2, the ensemble model consisting of random forest, SVM and naive bayes algorithms has an accuracy of 0.9428 and F1 score of 0.9390.

According to Approach Step 3.2.2.3, the accuracy of the decision tree model is 0.7662. Below is the visualized tree structure.

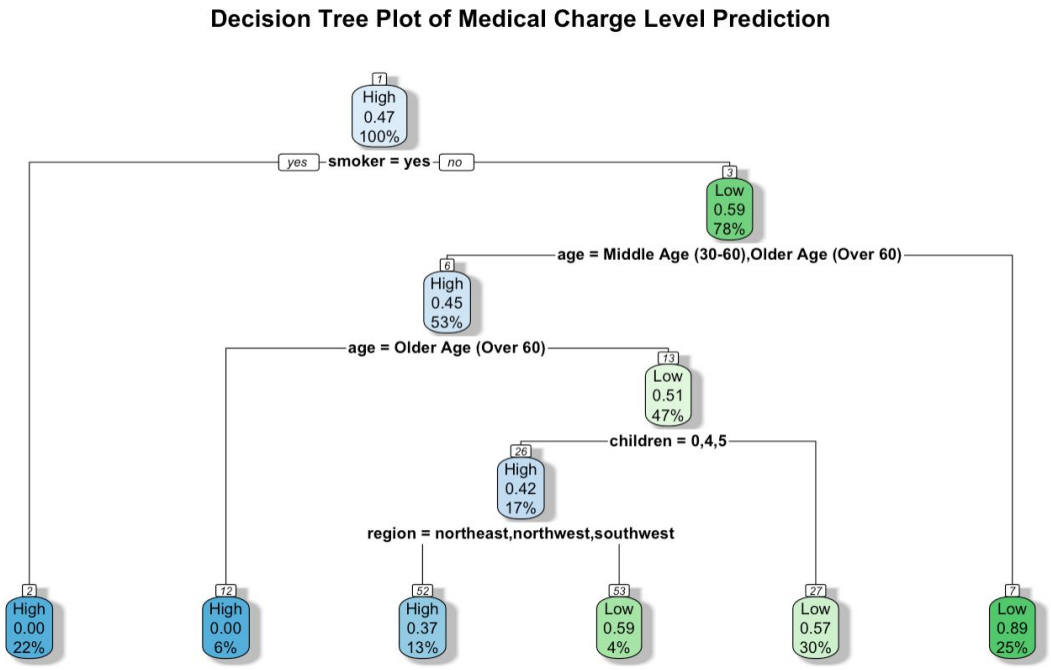


Diagram 10: Decision Tree Structure

According to Diagram 10, we can see that the root node is whether smoking or not, which means it has the highest information gain and it is the factor that impacts most in terms of cost differences. People who don't smoke tend to have a lower medical cost. Additionally, age is also considered as an influential factor such that younger people tend to have a lower medical cost. From the tree structure, we can see that region can also make a difference that people who live in the southeast tend to have a lower medical cost, but this may not be accurate enough due to the lack of number of samples.

The performance of all classification method can be seen below:

Table 8: Performance of Each Model

Modeling Algorithm	Accuracy	F1 Score	P-value
Random Forest	0.9353	0.9316	< 2e-16
SVM	0.9328	0.9288	< 2e-16
Naive Bayes	0.893	0.8938	< 2e-16
Logistic Regression	0.8955	0.8945	< 2e-16

Ensemble(random forest, SVM, naive bayes)	0.9428	0.9390	<2.2e-16
Decision Tree	0.7662	0.7267	<2.2e-16

Based on results from all the classification approaches above, the ensemble model which consists of random forest, SVM and naive bayes modeling algorithms has the best performance, so we will use this model for prediction function.

Prediction Function Result

According to Approach Step 4, we then built a function using the ensemble model so that further predictions can be done. For example, a 45-year-old smoking female with 2 children living in the southeast region, who has a BMI score of 30, is predicted to spend more than $8,943.29 USD in medical costs (categorized as High level of medical charges).

```
> # Example: Using Ensemble model because of its high accuracy
> pred_med_level(stack_rf, 45, 1, 30, 2, 1, 3)
[1] "A person with age: 45, sex (male=1, female=0): 1, bmi: 30, children: 2, smoker (yes=1, no=0): 1, region in US
(northeast=1, northwest=2, southeast=3, southwest=4): 3 is predicted to have a High medical charge level."
```

Diagram 11: Example of Prediction Function Results

Conclusions

After studying the dataset using descriptive analysis and supervised machine learning, we can conclude that smoking status is the most influential factor for our dependent variable: individual medical costs. Although the linear model does not work, the correlation plot and decision tree can provide evidence to complete our research. Additionally, our classification models also answers our research question in that obesity has less influence on medical costs than other factors. This can be attributed to the health law which states that health should not be a factor when insurance companies design health plans. For the region factor, since the decision node of it is toward the end of our decision tree (whose accuracy is only around 77%), further investigations are necessary to determine whether people's physical living locations impact medical expenses.

For generalization purposes, insurance companies can use the decision tree visualization along with our prediction function to develop a program in which individual's expected levels of medical expenditures can be identified using their characteristics. With our models, health insurance companies might be able to further develop their pricing system for customers in an effort to maximize revenues.

References

Abigail Abrams (2020). Time. The U.S Spends $2,500 Per Person on Health Care Administrative Costs. Canada Spends $550. Here's Why. Retrieved from https://time.com/5759972/health-care-administrative-costs/

Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Joseph Larmarange (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.R-project.org/package=GGally

Biener, A., Cawley, J., & Meyerhoefer, C. (2017). The High and Rising Costs of Obesity to the US Health Care System. *Journal Of General Internal Medicine*, 32(S1), 6-8. doi: 10.1007/s11606-016-3968-8

Daniel Lüdecke, Dominique Makowski and Philip Waggoner (2020). performance: Assessment of Regression Models Performance. R package version 0.4.4. https://CRAN.R-project.org/package=performance

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3. https://CRAN.R-project.org/package=e1071

David Robinson and Alex Hayes (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.4. https://CRAN.R-project.org/package=broom

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1. https://CRAN.R-project.org/package=skimr

Fuchs, V. (2011). *What Factors Affect Health Care Expenditures and Health?*. Robert Wood Johnson Foundation. Retrieved from https://www.rwjf.org/en/library/research/2011/04/what-factors-affect-health-care-expenditures-and-health-.html

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Hadley Wickham and Lionel Henry (2020). tidyr: Tidy Messy Data. R package version 1.0.2. https://CRAN.R-project.org/package=tidyr

Health Care Expenditures per Capita by State of Residence. (2014). Retrieved 29 March 2020, from https://www.kff.org/other/state-indicator/health-spending-per-capita/?activeTab=map&currentTimeframe=0&selectedDistributions=health-spending-per-capita&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

HealthCare gov. How insurance companies set health premiums. Retrieved from https://www.healthcare.gov/how-plans-set-your-premiums/

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret

Probasco, J. (2019). Why Do Healthcare Costs Keep Rising?. Retrieved 29 March 2020, from https://www.investopedia.com/insurance/why-do-healthcare-costs-keep-rising/

Salvatore Mangiafico (2020). rcompanion: Functions to Support Extension Education Program Evaluation. R package version 2.3.25. https://CRAN.R-project.org/package=rcompanion

Torsten Hothorn, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651--674.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wunderlich, G. (2010). *Improving health care cost projections for the Medicare population.* Washington, D.C.: National Academies Press.

Xu, X., Bishop, E., Kennedy, S., Simpson, S., & Pechacek, T. (2015). Annual Healthcare Spending Attributable to Cigarette Smoking. *American Journal Of Preventive Medicine*, 48(3), 326-333. doi: 10.1016/j.amepre.2014.10.012

Zachary A. Deane-Mayer and Jared E. Knowles (2019). caretEnsemble: Ensembles of Caret Models. R package version 2.0.1. https://CRAN.R-project.org/package=caretEnsemble