An Analysis of Medical Insurance Costs

William Frauen and Jack Miller

December 2022

Introduction

In the United States, the significant rise over the past few decades in healthcare costs that Americans collectively pay each year has opened up frequent debate regarding healthcare policy. Additionally, with this rise in costs, insurance expenses comprise a significant portion of many American budgets, and individuals may be concerned about how much they are likely to end up paying for medical insurance given their characteristics. As a result, it is be helpful to understand what factors influence how much an individual pays for medical coverage, to what degree those factors impact insurance bills, and how these factors might interact with each other.

Our aim will be to explore how medical insurance charges relate to the characteristics of an individual such as age, gender, smoking status, among others. We will seek to create a model for predictive purposes, as well as draw inference from the resulting fit. We will try to answer questions such as: What types of people in the United States may be more heavily burdened with medical costs? What habits are influential on a person's cost of medical care, and does that impact vary with other factors? What strategies should policy makers employ in order to help reduce the burden of medical coverage for their communities?

The Data Set

Our data set is sourced from Kaggle, originally from Machine Learning with R by Brett Lantz (Link to data: https://www.kaggle.com/datasets/mirichoi0218/insurance). The data set contains 1,338 observations of 7 variables. Response variable: charges. Predictor variables: age, sex, bmi, children, smoker, region.

Charges: annual medical cost billed to patient, in USD

Age: age of patient, in years

Sex: sex of patient (male or female)

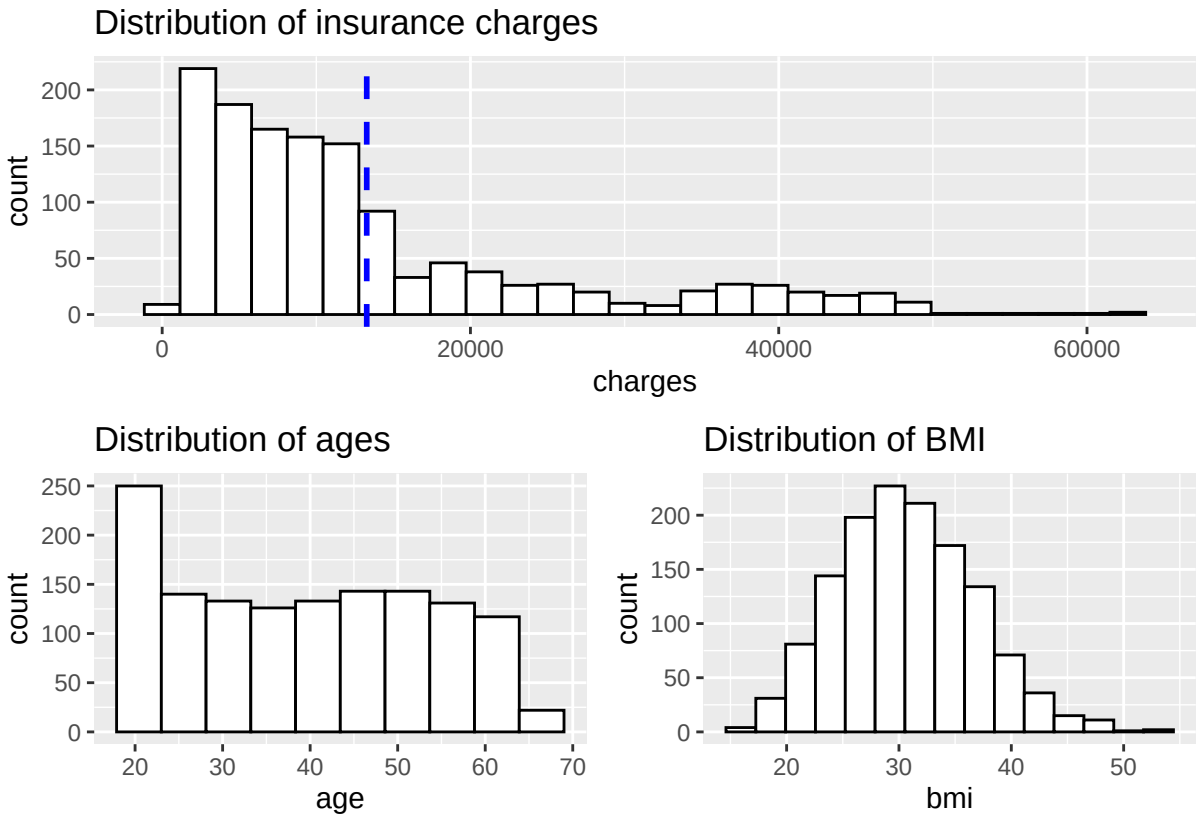
BMI: measure of body mass index of patient

Children: number of children of patient

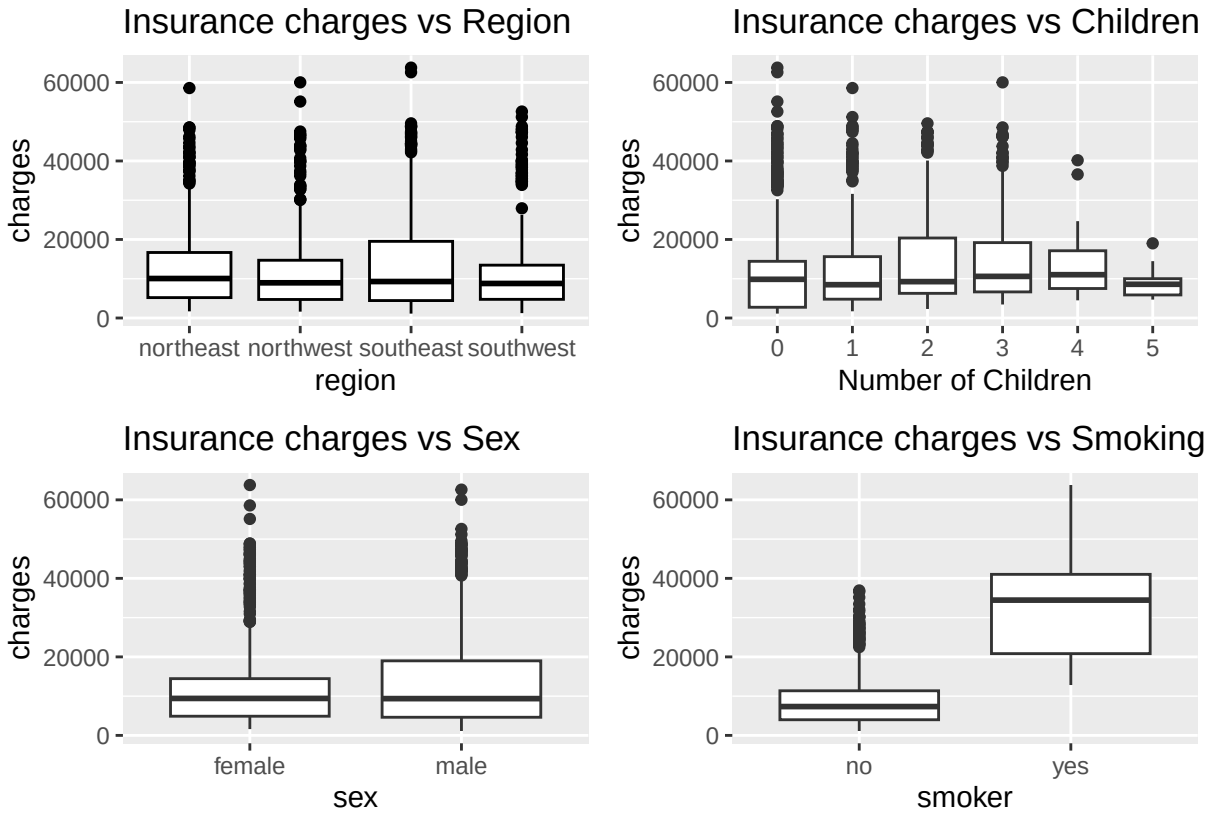
Smoker: whether the patient is a smoker (yes or no)

Region: region of the United States that the patient is from (Northeast, Northwest, Southeast, Southwest)

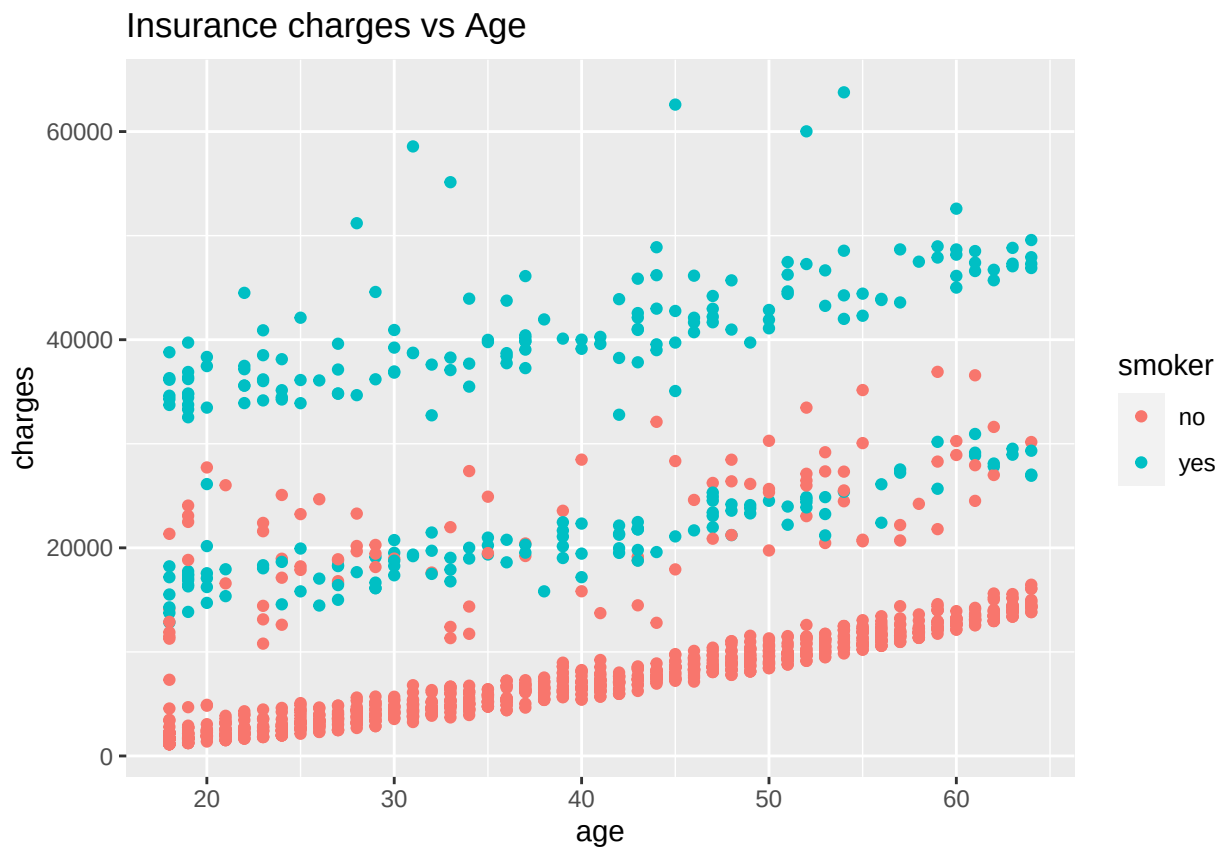
Exploratory Data Analysis



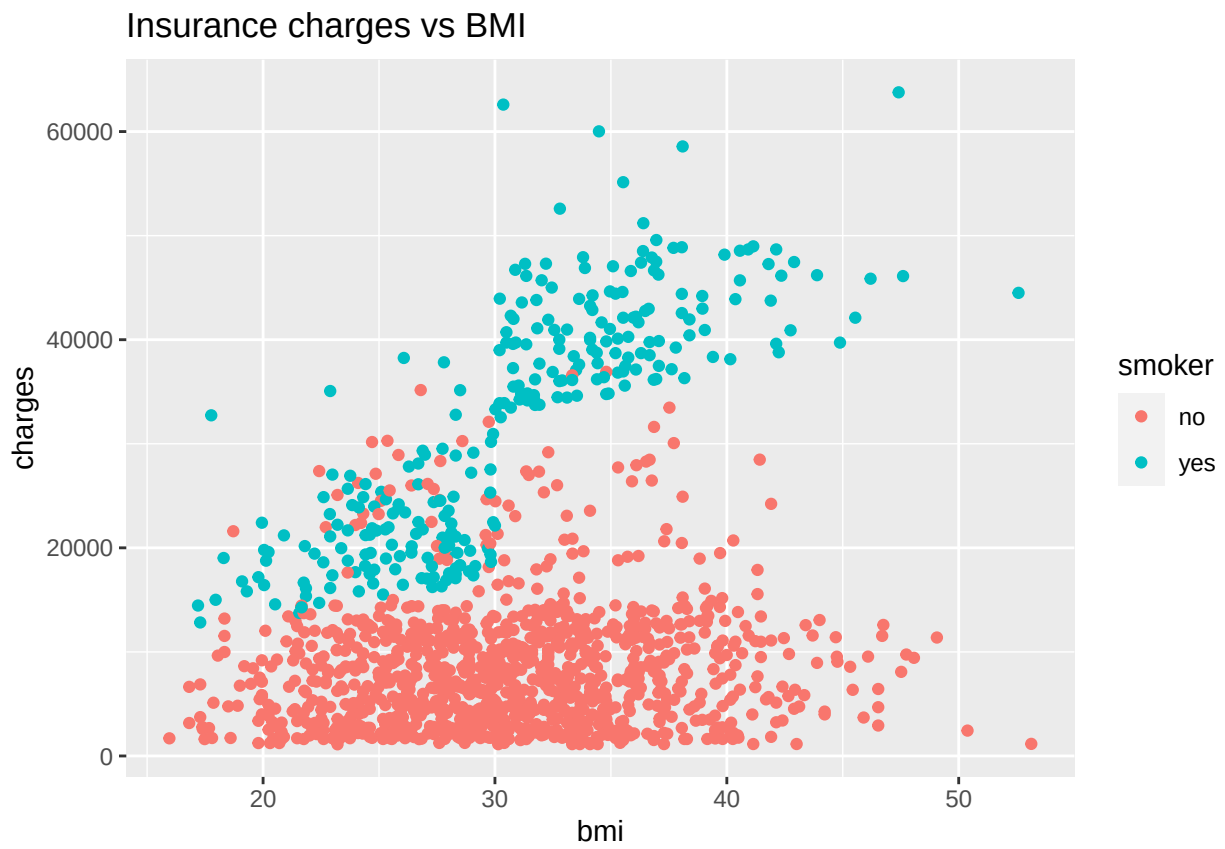
We first take a look at the distribution of the 1,338 insurance charges we have in our data set that serve as our response, as well as the distribution of ages and BMI, two potentially important predictors. Charges range from a low of \$1,122 to a high of \$63,770, have a mean of \$13,270, a median of \$9,382, standard deviation of \$12,110, and we can see significant positive skew in its histogram. The ages in our data set range from 18 to 64, centered at a mean and median of 39, and we notice a relatively larger amount of the youngest individuals and a smaller amount of the oldest individuals. In between, the distribution is more or less flat for intermediate age ranges. The BMI values in our data set appear roughly normally distributed, with a mean of 30.7 and standard deviation of 6.1.



We now wish to take a look at how the insurance charges are distributed across a number of our predictor variables. We first note that insurance charges appear to have similar distributions across all four United States regions Northeast, Northwest, Southeast, and Southwest. Each group is centered at roughly the same charge, each with a fair number of high outliers among them. Plotting insurance charges vs number of children, we notice a very slight increase in the median charge generally as we increase the number of children. Plotting insurance charges vs sex, we observe very similar distributions between males and females. And in the last plot we can see that individuals who smoke appear to have significantly higher insurance bills compared to non-smokers. Taking this all together, we expect region and sex to not be very significant predictors for an individual's annual insurance bill. Number of children may have a moderate effect, and we expect smoking to be a significant factor.



Plotting insurance charges across ages and coloring the points by smoking status, we can see that there appears to be a linear trend between charges and ages, for which, although the effect of age on charges does not appear to be significantly affected by smoking status, we can still see that smokers have a somewhat constant increase in charges vs non-smokers when laid out this way.



Plotting insurance charges across individual's BMI's and coloring the points by smoking status, we can see that there roughly appears to be a weak linear trend between insurance charges and BMI, but only when accounting for an individual's smoking status. We notice that, especially for smokers compared to non-smokers, we see higher charges as we increase BMI. Indeed, however, we can see that the distribution of blue and red dots are far from perfectly linear, but we deem this approximation an appropriate trade-off between model prediction power and interpretability. In terms of what this means for our modeling, we will introduce an interaction term between smoking status and BMI to account for this interaction effect between these two explanatory variables.

Model Exploration

We first split our 1,338 observations randomly into a training set of size 1,071 (~80%) and a test set of size 267 (~20%). We also construct a matrix for input to the glm functions for future ridge and lasso regressions that includes the interaction term between smoking status and BMI that we wish to include as discussed earlier. We explore ordinary least squares, ridge regression, lasso regression, pruned tree, bagged trees, random forest, and gradient boosting regression methods as possible fitting methods for the task of predicting insurance charges.

OLS: We utilize all of the variables (age, BMI, smoking status, sex, number of children, region), as well as the interaction term between BMI and smoking status, to predict insurance charges for our training data. We choose to include all of the variables as we will explore shrinkage and variable selection through means of ridge and lasso regression. With this OLS fit, we then make the predictions on the test data and compare it to the true test responses. We achieve Root Mean Square Error of \$5,117 and Mean Absolute Error of \$3,023 with OLS.

Ridge: We perform cross-validation to find the optimal value of λ in our model fitted with all of the variables

in addition to the interaction term between BMI and smoking status. We achieve RMSE of $5373 and MAE of $3441 on the test set.

Lasso: We perform cross-validation to find the optimal value of λ in our model fitted with all of the variables in addition to the interaction term between BMI and smoking status. We achieve RMSE of $5120 and MAE of $3029 on the test set.

Pruned Tree: We fully grow out the unpruned tree, and then perform cross-validation to determine the optimal tree size, which was the fully unpruned tree with 5 tree nodes. We achieve RMSE of $5563 and MAE of $3480 on the test set.

Bagged Trees: We perform bagging with 500 trees and achieve RMSE of $5343 and MAE of $2818 on the test set.

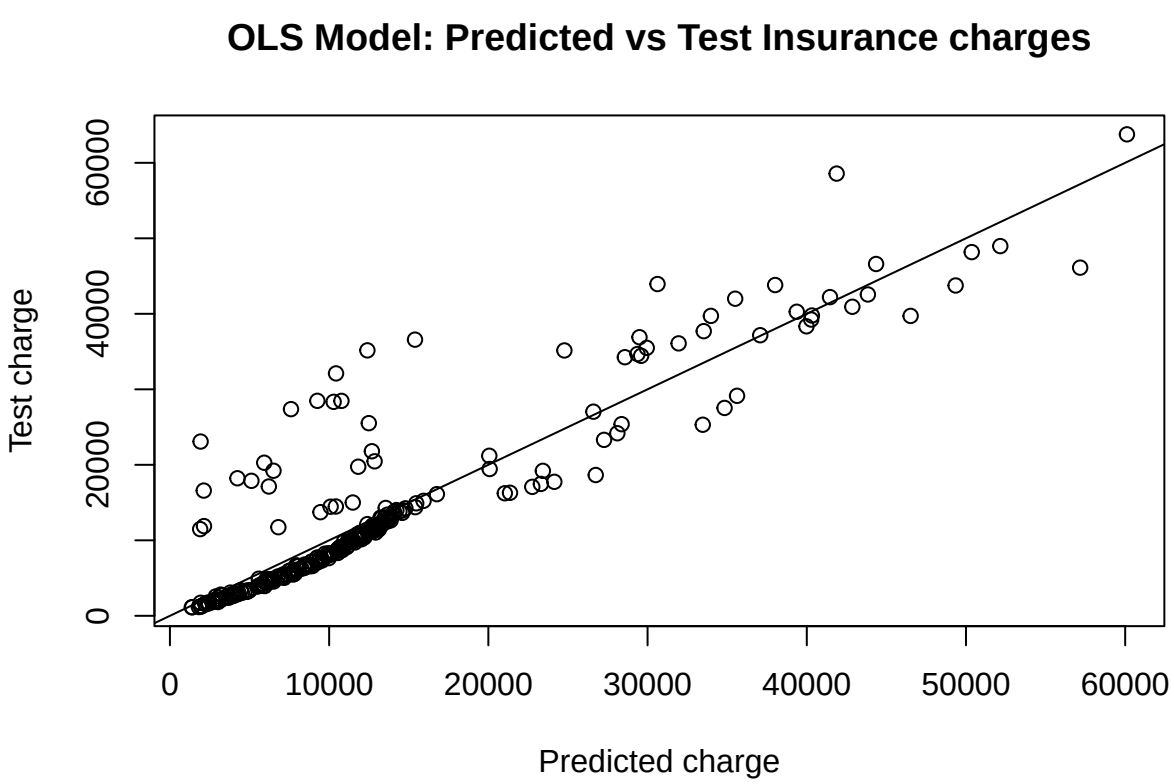
Random Forest: We create a random forest with 500 trees, with m = 2 predictors as candidates at each split. We achieve RMSE of $5273 and MAE of $2991 on the test set.

Gradient Boosting: We perform gradient boosting with 1000 trees and an interaction depth of 2. We achieve RMSE of $5159 and MAE of $2827 on the test set.

Model Selection

Regression Method	RMSE	MAE
OLS	5117	3023
Lasso	5120	3029
Gradient Boosting	5159	2827
Random Forest	5273	2991
Bagged Trees	5343	2818
Ridge	5373	3441
Pruned Tree	5563	3480

From the above table we can see the Root Mean Squared Error and Mean Absolute Error that we achieved with each method of regression. We see that our Ordinary Least Squares regression model performs best in terms of RMSE with $5,117, and has an MAE of $3,020. Meanwhile, our Bagged Tree regression model does best in terms of MAE with $2,818, and has an RMSE of $5,343. While in certain instances we may be interested in reducing MAE, here we are more interested in reducing the squared errors. So, we select our OLS regression model as our model of choice for predicting insurance charges. Additionally, as a linear model, we are able to easily interpret the effect of each explanatory variable on our response variable charges. We can see that our Lasso and Boosted Tree models gave nearly as good test performance compared to the OLS fit. Though the simple nature of the OLS model proved to generalize relatively well to the test data, so it ends up on the top of our list and we will explore it further.



From the above plot we can see the true test charges plotted against the predicted test charges fitted by our OLS model. We roughly see the points scattered fairly close around the 45 degree line, with our predictions seeming like a reasonable fit and fairly accurate fit. A residuals vs leverage plot did not indicate any influential outliers in our data set. Our RMSE of $5,117 tells us that $5,117 is a typical distance between our predicted charges and the true charges for the test data. Our MAE of $3,023 tells us that the average (absolute) error we have between our predicted charges and the true charges for the test data is $3,023.

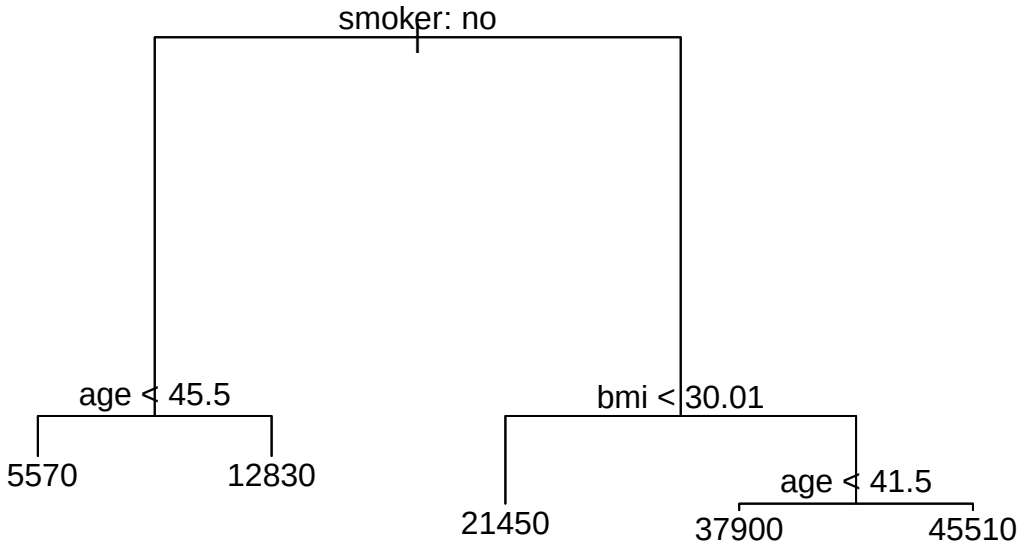
term	estimate	std.error	statistic	p.value
(Intercept)	8665	359	24.11	0.000
scale(age, scale = FALSE)	267	10	25.51	0.000
scale(bmi, scale = FALSE)	22	29	0.79	0.432
smokeryes	23878	361	66.20	0.000
children	511	124	4.13	0.000
regionnorthwest	-371	418	-0.89	0.374
regionsoutheast	-1186	421	-2.82	0.005
regionsouthwest	-897	422	-2.13	0.034
sexmale	-511	294	-1.74	0.082
scale(bmi, scale = FALSE):smokeryes	1459	59	24.72	0.000

The coefficients of our Ordinary Least Squares fit to our training data are seen above. We can thus write our model as:

$$\hat{\text{charge}} = 8665 + (267)(\text{age} - 39) + (23878)\mathbf{1}_{\text{smoker:yes}} + (22)(\text{bmi} - 30.5) + (1459)\mathbf{1}_{\text{smoker:Yes}} * (\text{bmi} - 30.5) + (511)\text{children}$$
$$+ (-371)\mathbf{1}_{\text{region:NW}} + (-1186)\mathbf{1}_{\text{region:SE}} + (-897)\mathbf{1}_{\text{region:SW}} + (-511)\mathbf{1}_{\text{sex:Male}}$$

Findings

With our OLS model, we have mean-centered the variables age with mean 39, and BMI with mean 30.5. So, our intercept tells us that for an individual who is female, does not smoke, has no children, lives in the Northeast, is 39 years old, and has a BMI of 30.5, we expect their annual medical insurance bill to be roughly $8,665, on average. For each one year increase in the age of the individual, holding all else constant, we expect the annual medical insurance bill to increase by roughly $267, on average; or an increase of $2,670 each decade. We expect an individual who smokes to be paying on average $23,878 more annually in medical insurance bills compared to non-smokers, holding all else constant. For each additional child that an individual has, holding all else constant, we expect the annual insurance bill to increase by $511, on average. Compared to individuals living in the Northeast, we expect those living in the US Southeast to pay $1,186 less, and those in the Southwest to pay $897 less annually on medical insurance, on average, all else held constant. The difference between the Northeast and Northeast was found not to be significant, as was the effect of sex on insurance charges. For non-smokers, the effect of BMI on insurance is not significant. But for smokers, we found that for each one unit increase in an individual's BMI, holding all else constant, the expected medical charge increases by 22 + 1,459 = $1,481, on average.



While the OLS regression fit, which we use if we are trying to accurately predict an individual's annual medical insurance bill, offers us a straightforward interpretation of the model coefficients, additionally as a byproduct of our model exploration, we found that the tree diagram outputted above by our Pruned tree model offers a very useful and easy to use flowchart for quick-and-dirty medical insurance cost estimates that any individual can pick up and use. It uses the variables smoking status, age, and BMI (all of which an individual can quickly input) in the tree and has 5 terminal nodes, meaning there are 5 values of estimates that this tree can provide. While the Pruned tree has a higher RMSE and MAE of $5,563 and $3,480, respectively, this is a fairly modest sacrifice in accuracy for the ease of use of this flowchart. The tree can be used as follows: Start at the top, if the answer is "yes" to the expression in the label then we head down

the left direction, otherwise we go to the right, and the values at the bottom indicate the predicted annual insurance charge that you end up with. For example, if the individual is a smoker ("no" to non-smoker), has a BMI of 35 ("no" to bmi < 30.01), and is 40 years old ("yes" to age < 41.5), then we predict their annual medical insurance bill to be roughly \$37,900, on average.

Explanatory Variable	Relative Influence
smoker	56.36
bmi	22.57
age	15.52
region	2.57
children	2.42
sex	0.56

An additional insight gained from our exploration of models was the information on the relative influence of each variable that we observed with our Gradient Boosted model. With this, we see that an individual's smoking status is the single most influential factor in determining medical insurance charges. BMI comes next (through its interaction with smoking status as discussed earlier), followed by the age of the individual. US region and number of children appear to very little influence, and sex appears to have nearly none.

Conclusions

Putting our findings together, we see that the most important factors in determining an individual's annual medical insurance bill among those explored in this report are smoking status, BMI, and age. Our OLS model that we chose tells us that smoking has a large and significant impact on medical insurance charges. Indeed, smokers are expected to be paying roughly north of \$23,000 more versus non-smokers, all else held constant. For smokers concerned about higher insurance bills, they should pay extra attention to their BMI, as every one-unit increase in BMI for them corresponds to an almost \$1,500 increase in charges. Meanwhile, non-smokers need not worry about a higher BMI translating to higher medical insurance charges. Among all individuals, on the other hand, an increase in age corresponds to a approximately linear increase in insurance charges, at a rate of roughly \$267 per year, all else held constant. Thus, the individuals most at risk of paying high medical insurance bills are those who are older, have larger BMIs, and smoke. An older, more overweight individual who smokes who wants to minimize their insurance charge should thus focus on trying to reduce their BMI and quit smoking if possible. For policy makers who want to alleviate medical costs that the public faces, while people's age can't be decreased by policy of course, they should pursue policies that help reduce obesity, such as fitness and healthy eating campaigns, and smoking prevalence, such as smoking informational/awareness campaigns. These factors of obesity and smoking prevalence, especially when tackled together, is likely to help to reduce the overall medical insurance charges individuals face on average.

Meanwhile, the number of children had a fairly limited impact on an individual's expected charge, and the factors of geographic US region and sex were found to have had hardly any significant impact on insurance bills. First, we note that it appears neither sex should be worried about receiving any meaningfully increase in their expected bill as a result of their gender. We also observe that individuals living in Northern regions of the US (NE + NW) are expected to be paying roughly \$1,000 more than those in Southern regions (SE + SW), with the differences among North and South regions not being significantly different. Nevertheless, it appears this impact on insurance charges is at best small, less than our RMSE for OLS; so individuals moving around the United States need not worry about facing higher medical costs based on where they plan to live. However, one limitation to these findings is that these region variables do not have very high resolution, in the sense that we do not have more detailed and potentially helpful information such as the specific state the individual is living in, or whether it is a rural or urban environment, or whether the community is more left or right leaning politically. When broken down by these factors, it is possible that the geographic location of an individual might become an important factor in determining medical charges; though here we can

only say that at the scope of these 4 US regions, geographic location does not appear to be a significant determinant of medical insurance bills.

With our OLS model, on the test set we achieve an RMSE of $5,117 and MAE of $3,023. Charges in our data set range from roughly $1,000 to $64,000, so a typical deviation of around $5,000 gives a decent, albeit somewhat rough, estimate of an individual's insurance charges, though at the benefit of needing only a smaller number of readily producible inputs. As stated before, we have also produced an easy-to-use tree diagram for an individual trying to estimate annual insurance costs that sacrifices a slight decrease in accuracy for a very interpretable form.