Predicting Medical Insurance Charges: Assessing Traditional and Advanced Methods on Nonparametric Data

Alexander Abusaa
Winnie Lu

California State University, East Bay
Dr. Joshua Kerr
May 10th, 2022

Introduction

When it comes to the American health insurance policy, it's predominantly characterized by exorbitant fees, and hidden clauses that only highlights the growing disparity among the socioeconomic classes. According to Dr. Shmerling from Harvard Health, this financial burden is often laced with surprise fees and conditions. For example, the cost of physician care and medicine might be covered in a particular surgical procedure, but the anesthesiologist would not be. Unfortunately, healthcare access is uneven as healthcare is tied to employment. This puts marginalized communities at risk. Oftentimes, for more significant procedures, health insurers require extensive forms and justification to be filled out. This does eliminate unnecessary expenses, but also discourages appropriate care deemed by the physicians. Estimating and predicting health care cost and coverage is advantageous to an extent.

Knowing which major factors influence the cost can help us curb costs in the future via preventive methods. By examining the factors related to personal medical costs, we aim to create models that characterize and predict the factors that influence the costs of health care. Additionally, these methods will allow us to gain insight into the complications of real-world observational data, as well as the relevant modeling techniques utilized in more opaque industries such as healthcare.

Data Description

The dataset being used contains patient data (both quantitative and categorical) and has a target variable of annualized medical costs billed to insurance in USD. The data was collected via demographic statistics from the U.S. Census Bureau, and subsequently curated for the book *Machine learning with R* by Brett Lantz. The dataset contains 1,338 observations of patients who carry an insurance policy, with multiple variables to measure the characteristics of each patient. Charges (our target variable) is a numeric variable of annual medical charges billed to a patient's insurance. Age is an integer variable of the age of the patient, with those above 64 being excluded from the dataset as they are qualified for government subsidized insurance. Sex is a 2-level factor variable of the patient's sex. BMI is a numeric variable of the patient's BMI. Children is an integer variable of the number of children or dependents covered by a patient's insurance plan. Smoker is a 2-level factor variable of the patient being a regular smoker or not.

Finally, region is a 4-level factor variable of the patient's region of residence in the US: northeast, southeast, southwest, or northwest.

Briefly examining the data shows that several dummy variables that may influence the insurance costs (sex, smoker, and region). We also see the basic summary statistics for all the predictors in our data, and note that besides the mean being greater than the median, there are no missing values across the predictors. To reflect on our summary statistics, our distribution is incredibly right skewed, even though a majority of the annual insurance expenses range from $0 to $15,000.

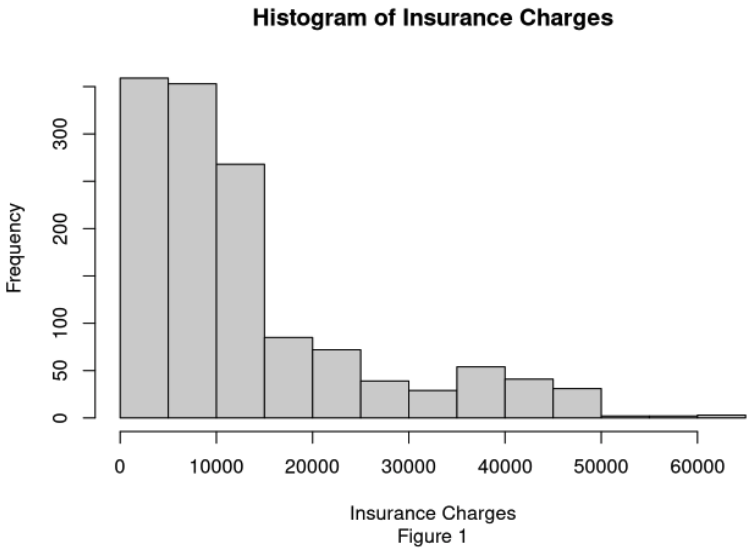
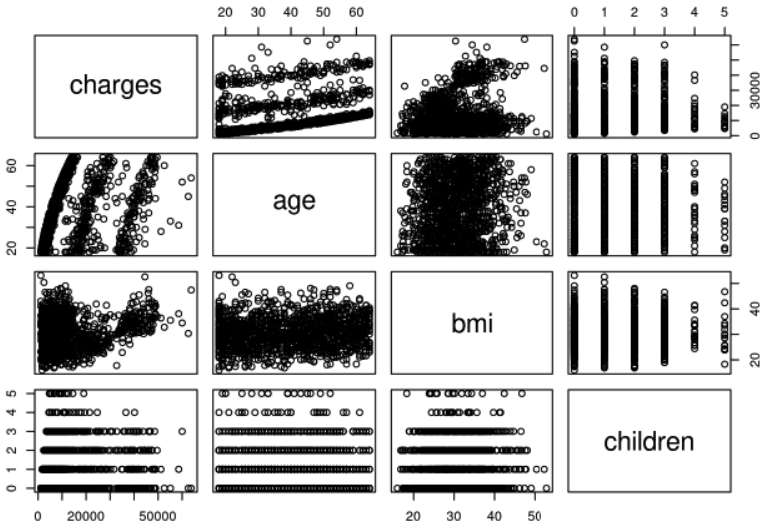


Figure 1



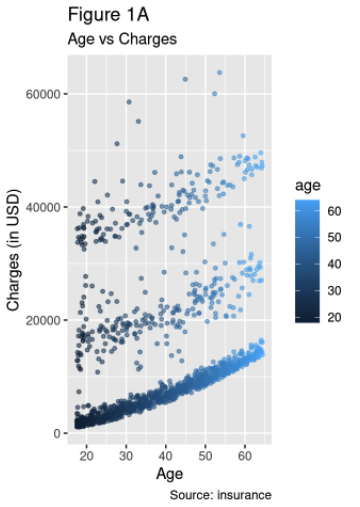


Figure 1A
Age vs Charges
Charges (in USD)
Age
Source: insurance
age

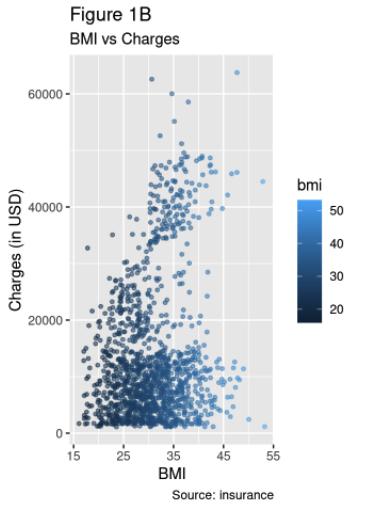


Figure 1B
BMI vs Charges
Charges (in USD)
BMI
Source: insurance
bmi

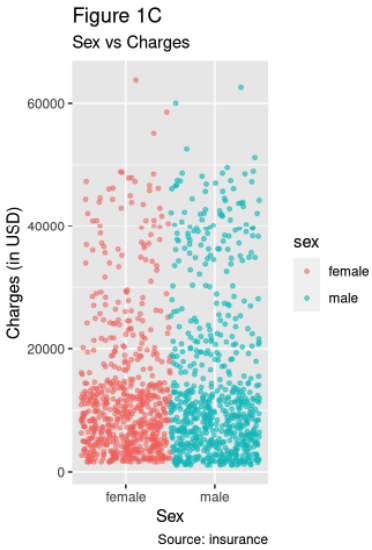


Figure 1C
Sex vs Charges
Charges (in USD)
female
male
Sex
Source: insurance
sex

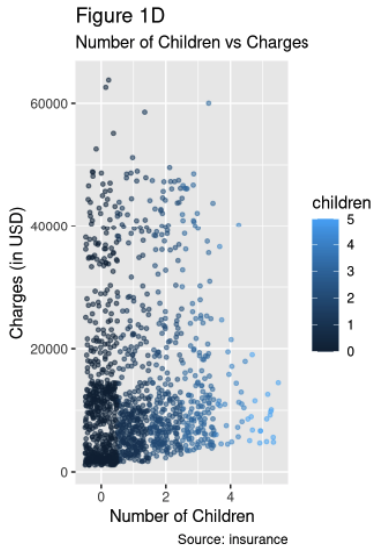


Figure 1D
Number of Children vs Charges
Charges (in USD)
Number of Children
Source: insurance
children

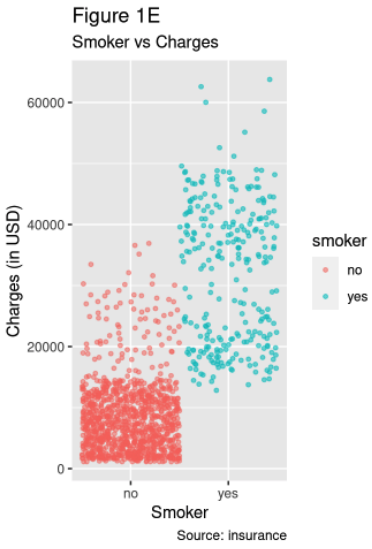


Figure 1E
Smoker vs Charges
Charges (in USD)
no
yes
Smoker
Source: insurance
smoker

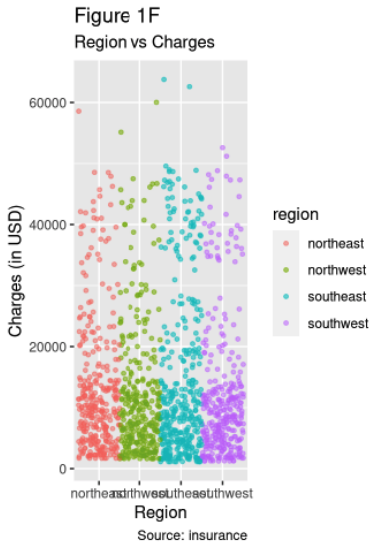


Figure 1F
Region vs Charges
Charges (in USD)
Region
Source: insurance
region
northeast
northwest
southeast
southwest

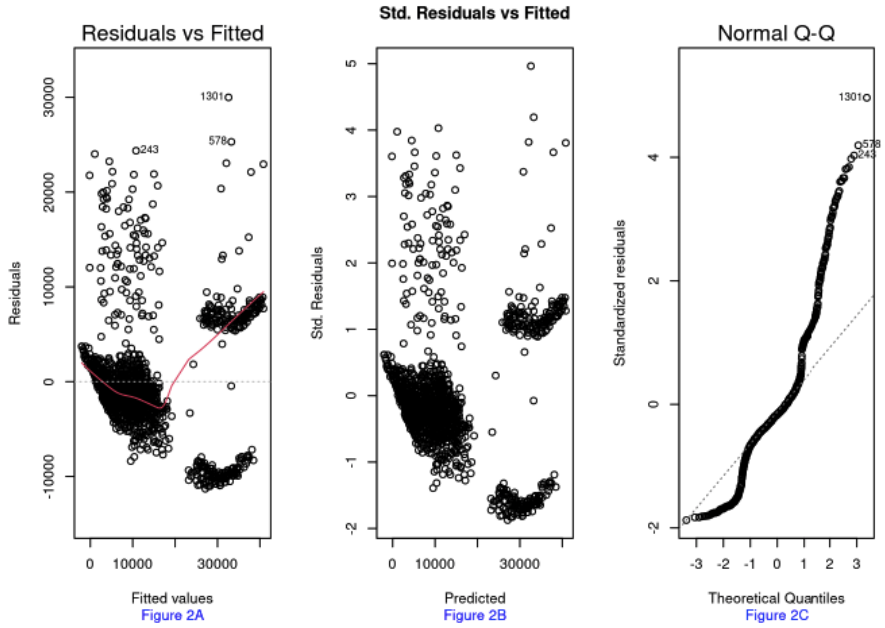
We start by visualizing our data to identify any obvious pattern between the predictors and response, or among the predictors themselves. Preliminary analysis in Fig 1A & 1B. show that age and bmi have a slight positive relationship with charges. As age and bmi increases, we see a general increase in charges. These are observations we would expect: generally, younger people are healthier than the elderly, and have less medical complications in comparison. A healthy individual has a BMI between 18.5 to 24.9, whereas a BMI over 29.9 is considered to be within the obese range, so this pattern is justified. Obese patients tend to be associated with coronoary diseases more so than non-obese patients (CDC, 2020). However, it is relevant to note that people with lower BMI still have high medical expenses.

An interesting observation in Figure 1D is that with an increasing amount of children per household, we see that charges appear to decrease. This finding contradicts what we would expect from insurance companies; as the common notion appears to be that having more children would compromise the family into buying more premiums or policies for them. It is notable to even say that with more children, there would be a greater chance of children inheriting or developing a poorer health condition than the older siblings. Fortunately, this isn't the case, as most insurance policies lay out a flat fee for spouses with 3 or more dependents under 21.

From Figures 1C and 1F, there appears to be no significant relationship between the sex of the policy holder and charges, and between region and charges. In contrast, smokers tend to pay more for their policies compared to non-smokers, as shown from Figure 1E. According to Dr. West, smokers are more likely to be associated with various health risks. Such factors include, but are not limited to the umbrella of heart diseases: myocardial infarctions, atherosclerosis, or respiratory distress: emphysema, bronchitis, etc, and even dental costs! (West, 2017). Therefore, smoking patients are paying more for certain premiums or paying more in terms of frequency in physician visits.

As previously mentioned, since our correlation matrix does not reveal any extreme collinearity among the predictors, we can move forward to identify factor(s), if any, that greatly influence the cost of medical charges.

Methods and Results



In our first model, $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$, and $H_A : \text{at least one } \beta_j \neq 0$. As a result, the overall F-test (p-value: < 2.2e-16) tells us that at least one predictor is not equal to another; at least one predictor is significant to our model. By examining the individual t-tests, the predictor, sex does not appear to be statistically significant. However, we see that age and smokers appear to greatly influence the charges. Since there are dummy variables present, one category of the variable is used as a reference. As a result, the estimates are interpreted relative to the reference. Thus, males have $131.30 less medical expenses each year relative to females and smokers cost an average of $23,848.50 more than non-smokers per year. We also see that the coefficient for the reference group, the northeast region, tends to have the highest average charges compared to the other 3 regions in our model. The adjusted R squared is 0.7494, which indicates that about 74.94% of the variability we see in insurance charges are explained by the predictors in our current model. Since this value is a reflection on our model, we can deem it not too bad. To conclude, we see high residuals, meaning high errors: over 50% of the errors fall within the first and third quartile: predictions were between $2848.10 over the true value and 41393.9 under the true value.

Obviously, this model isn't the best, seeing as how it violates our LINE assumptions: we see a discernible pattern in figure 2A and strong deviation of points from the line in figure 2B; violating our equal variance and normality assumptions.

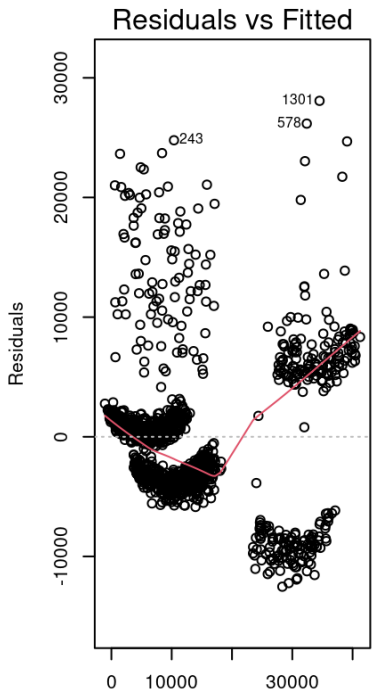


Figure 3A

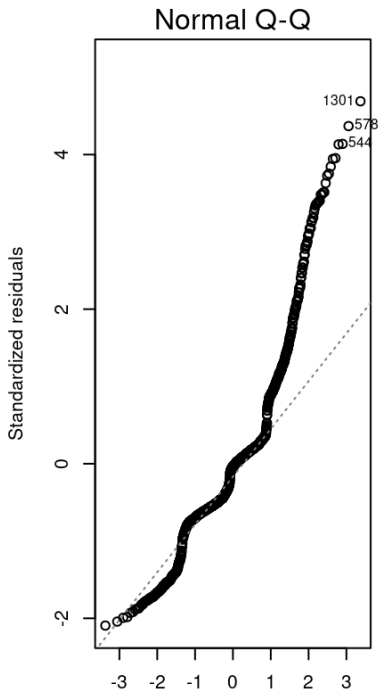


Figure 3B

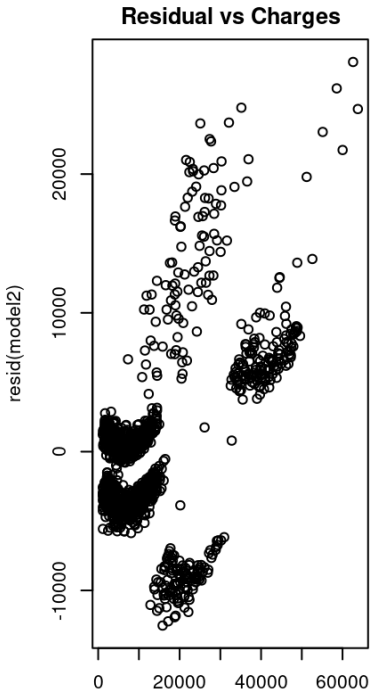
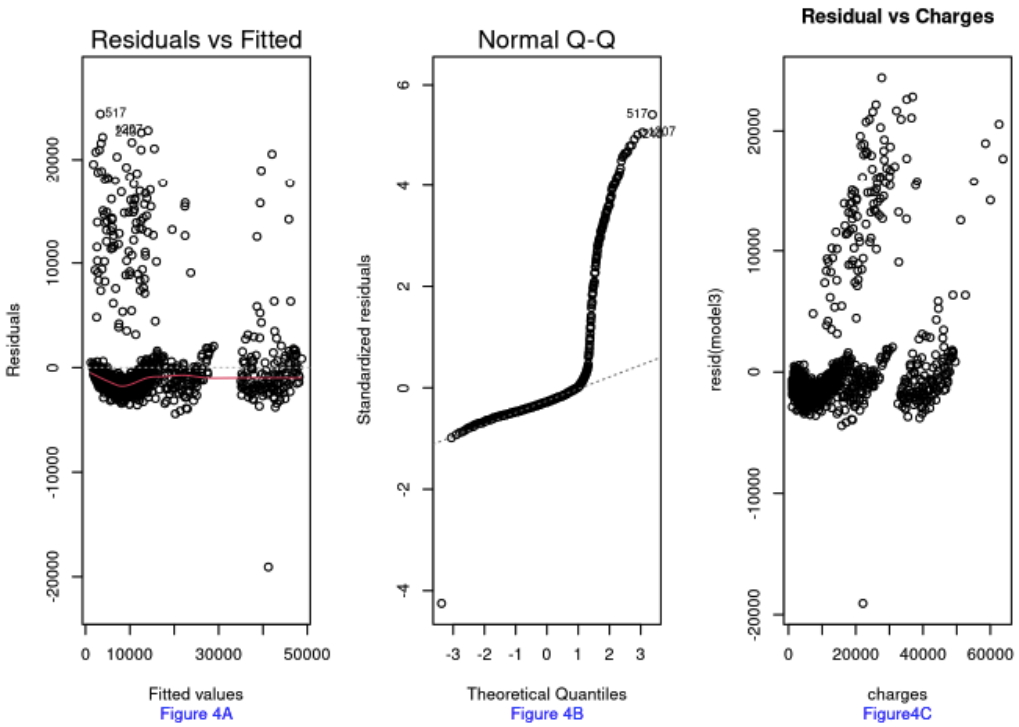


Figure3C

Although there are clear violations to our equal variance and normality assumptions, we will not be using the box-cox method to remedy this. Currently, our goal is to identify the major influence(s) to our response, so our focus will be on the effects of charges via predictors. For the second model, we've dropped the predictors Region and Sex, since they appeared to be statistically insignificant. For this model, $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, and H_A : at least one $\beta_j \neq 0$. Given our F-statistic (p-value < 2.2e-16), at least one of our predictors is significant to our model, and our individual t-test shows that all predictors in our second model are significant. Among our current set of predictors, smoking and obesity are the predominant factors influencing the insurance charges. According to the CDC, obese patients have a BMI of over 30. We can include this into our model to show the effects of obesity on insurance charges. We can see that if a patient is a smoker, then the medical costs increase by 23819.41 dollars. Additionally, if a patient is obese, then they have increased medical cost of 2904.47 dollars. The

adjusted R. squared = 0.754, which shows an improvement compared to our first model, where the adjusted R.squared = 0.7494. However, we can see that our linear regression model is not the best model, because of the violations of normality and heteroscedasticity present (Fig3C). We require a different approach(es) for a better prediction.



Since we've previously shown that smoking and obesity are significant to our model, we test the significance of that interaction. Using the previous null and alternative hypothesis (with appropriate alterations to our current interaction addition), the overall F-statistic (p-value = < 2.2e-16), shows that at least one of our predictors is significant; this is confirmed by our individual t-test, where we see that all of the predictors, with the borderline exception of the obesity (BMI30), is significant. We're still going to include obesity into our model for comparative purposes: we want to see how obesity affects costs. Consequently, we can see that the interaction between obesity and smoking has a significant effect: smoking alone increases the costs by 13,412.40 dollars, while obese smokers increases costs by 19684.869 dollars. Although our assumptions are still violated, as shown from Figures 4A-C, it still shows a drastic improvement from our previous diagnostics (Figures 2A-C and Figures 3A-C). To support this final model as our "best" model, we contrast the adjusted R.squared values from the previous models to our current one. Our previous adjusted R.squared = 0.7494, and 0.754 from models 1

and 2, respectively, while our current adjusted R.squared = 0.8614, so approximately 86.14% of the variability is explained by the predictors in our third model.

While we were able to achieve a satisfactory fit with our linear regression models, we were not able to validate the least squares assumptions using model diagnostics: indicating that our linear models are not as generalizable or as robust as we desire. In order to improve upon these models we must first consider the fundamental properties of our modeling techniques. Linear regression, as its name implies, is predicated on the assumption that the relationship between our response and predictor variables is linear. The linearity of this relationship is also relevant to the calculation of the least squares estimates that parameterize the model itself. Given the irregularities in our diagnostic plots (see figures 2, 3, and 4, A through C) we may consider statistical methods that do not rely on strictly linear relationships to make predictions. In this case we will be conducting our analysis via a non-parametric method: as we believe the relationships within our data may not all belong to a known, parameterized distribution.

By utilizing non-parametric methods we will enter the domain of statistical and machine learning methods. Fortunately there exists a wealth of information and pre-existing research, making these methods highly accessible. The particular method chosen for our subsequent analysis is the random regression forest model. Relative to other statistical learning methods the random forest is simple to implement and tune, as well as capable of producing robust results with minimal variance or risk of overfitting. In addition to not requiring the same operational assumptions as linear regression, the random forest method is able to effectively learn from the data by averaging the results from a large number of regression trees: each of which represents a bootstrapped sample of randomly selected data and predictors.

The R packages *caret* and *randomForest* were selected for our analysis as, in addition to being simple to implement and tune, they can be used to retrieve detailed diagnostics and plots, respectively. Across both packages a random regression forest was run with the default parameters of 500 total trees and a 10-fold cross validation. The diagnostic criteria being used for our random forest are the root mean square error (RMSE) and the R squared values. The *caret* package provided the results and summary statistics for the diagnostic criteria of 3 separate random forests: one in which 2 variables were sampled at each branch (node) of a tree, one in which 5 variables were sampled, as well as one in which 8 variables were sampled. Note that

each level of a factor-type variable is considered its own predictor. The model returned by the *randomForest* package was a single random regression forest in which 2 variables were sampled at each tree's branch. The *randomForest* package also returned the RMSE, R squared value, and variable importance plot of its random forest (Figure 5).

Between the 3 random forests conducted through the *caret* package, the forest with 5 variables sampled at each branch yielded the best results: with an RMSE of 4,614.325 dollars and an R squared value of approximately 0.849. The *randomForest* package's results sampling 2 variables at each branch were quite close: with an RMSE of 4,682.99 dollars and an R squared value of 0.85. It should be noted that as RMSE is a quadratic measure of average error it is sensitive to points with relatively high errors (e.g. anomalies, outliers, etc…). This will come into consideration when observing a plot of the predicted values of this model over a theoretical 1:1 prediction line (Figure 6).

Figures 5 and 6 (below) are the *randomForest* package's variable importance plot and plot of predicted values, respectively. In figure 5 we can see each predictor variable ranked in order of importance: which is quantified as the percent increase in mean absolute error (MAE) observed as a particular variable's values are permuted. It is immediately apparent that a patient's smoking status was highly important when predicting annual medical charges. Likewise the policyholder's age and BMI were also found to be very important in predictions. The remaining predictors had very little importance in predicting medical charges. The distribution of predictions can be seen below in figure 6.

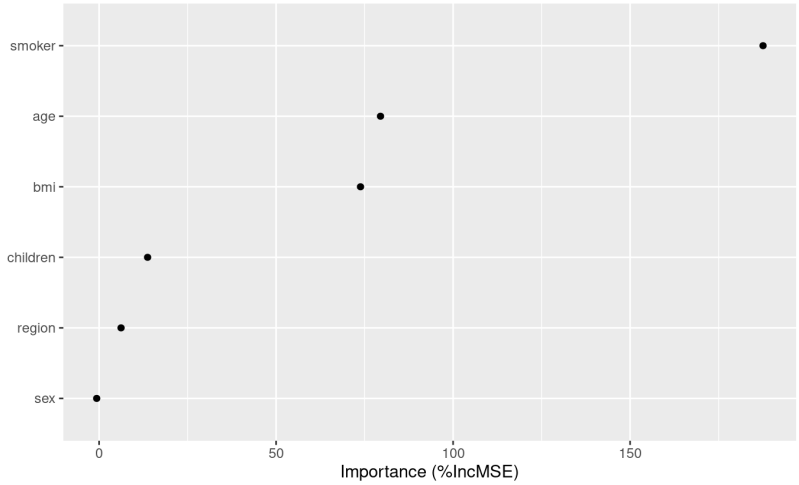


Figure 5: Variable Importance Plot

Despite the high magnitude of the RMSE of our models we can observe that the range of predicted values is more than adequate. A significant proportion of predicted values appear quite close to, and very nearly on, the 1:1 line. Though there are 2 wide clusters that appear below the 1:1 line: indicating that the model had a tendency to under-predict medical charges under certain circumstances. That being said, however, the number of points along the 1:1 line far exceeds that of those below the line: with these under-predicted values likely being responsible for the high RMSE value.

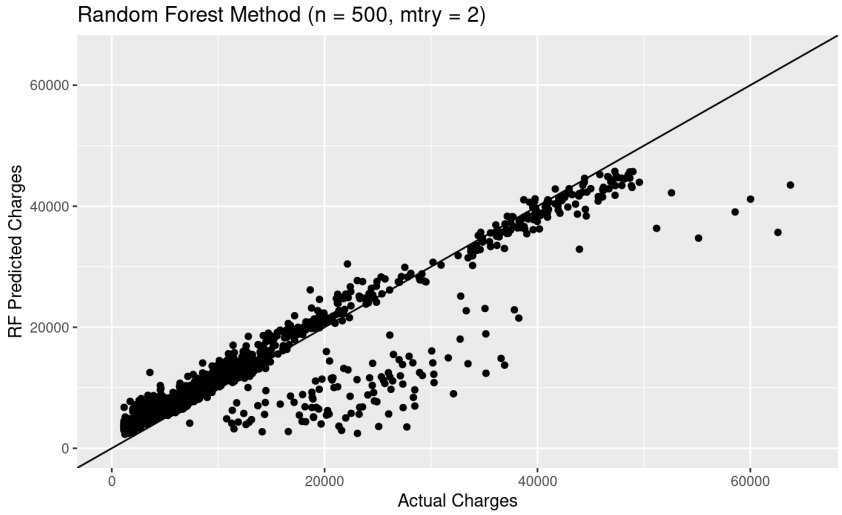


Figure 6: Plot of Predicted Values

Conclusion

From both methods, we were able to reach our goal: we've successfully identified the major contributors to insurance premiums: older, obese smokers are responsible for the pattern of surging insurance charges we've previously stated. With these conclusions, we can suggest certain preventative care, to an extent, as an attempt to curb the charges. Since aging is inevitable, we can focus on developing a healthier diet that would lower LDL cholesterol, as it does contribute to obesity (Cabezas, Elte, Klop, 2013). According to the KFF, most states can charge up to 50% for a person who uses tobacco products. This certainly can prevent people from purchasing tobacco products, but there are some exceptions. For instance, California prevents the tobacco surcharge. But to reiterate, there are increased risks and diseases attached to smoking which would, in the long run, only increase costs.

In addition to patient health and insurance charges, our results also have implications in the areas of statistical and machine learning. We were able to explore advanced modeling techniques and consider the implications different models may carry for different types of data. In scenarios such as in this project, wherein the data may contain excessive noise or not have obvious interactions, non-parametric methods are quite capable of producing reasonably accurate predictions despite not operating under the assumptions of traditional linear regression. While the field of statistical learning is constantly evolving in terms of scope and complexity, this project provides a solid foundation of the reasoning and underlying mechanisms of many techniques found within it.

Appendix

```
insurance <- read.csv("insurance.csv")
head(insurance)
summary(insurance)
sum(is.na(insurance))
hist(insurance$charges, main = "Histogram of Insurance Charges", sub = "Figure 1", xlab = "Insurance Ch
```

```
pairs(charges ~ age + bmi + children, data=insurance)

round(cor(insurance[, -c(2, 5, 6)]), 4)


library(ggplot2)
```

```
a <- ggplot(insurance, aes(age, charges)) +
  geom_jitter(aes(color = age), alpha = 0.5, size = 1, width = .5) +
  labs(subtitle="Age vs Charges",
       y="Charges (in USD)",
       x="Age",
```

```
       title="Figure 1A",
       caption = "Source: insurance")


b <- ggplot(insurance, aes(bmi, charges))+
    geom_jitter(aes(color = bmi), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="BMI vs Charges",
       y="Charges (in USD)",
       x="BMI",
       title="Figure 1B",
       caption = "Source: insurance")

grid.arrange(a, b, ncol = 2)
```

```
s <- ggplot(insurance, aes(sex, charges))+
    geom_jitter(aes(color = sex), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Sex vs Charges",
       y="Charges (in USD)",
       x="Sex",
       title="Figure 1C",
       caption = "Source: insurance")

c <- ggplot(insurance, aes(children, charges))+
    geom_jitter(aes(color = children), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Number of Children vs Charges",
       y="Charges (in USD)",
       x="Number of Children",
```

```
       title="Figure 1D",
       caption = "Source: insurance")


grid.arrange(s, c, ncol = 2)
```

```
sm <-ggplot(insurance, aes(smoker, charges))+
    geom_jitter(aes(color = smoker), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Smoker vs Charges",
       y="Charges (in USD)",
       x="Smoker",
       title="Figure 1E",
       caption = "Source: insurance")

r <-ggplot(insurance, aes(region, charges))+
    geom_jitter(aes(color = region), alpha = 0.5, size = 1, width = .5) +
    labs(subtitle="Region vs Charges",
       y="Charges (in USD)",
       x="Region",
       title="Figure 1F",
       caption = "Source: insurance")

grid.arrange(sm, r, ncol = 2)
```

```
model1 <-lm(charges ~ age + sex + bmi + children + smoker + region, data=insurance)
summary(model1)

par(mfrow=c(1,3))
plot(model1, 1) + title( sub ="Figure 2A", col.sub = "blue")
plot(predict(model1), rstandard(model1), xlab="Predicted", ylab="Std. Residuals") + title(main = "Std.
plot(model1, 2) + title(sub ="Figure 2C", col.sub = "blue")
```

```
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
model2 <-lm(charges ~ age  + bmi + bmi30 + children + smoker, data=insurance)
summary(model2)


par(mfrow=c(1,3))
plot(model2, 1) + title(sub = "Figure 3A", col.sub = "blue")
plot(model2, 2) + title(sub = "Figure 3B", col.sub = "blue")
```

```
plot(resid(model2) ~ charges, data=insurance) + title(main = "Residual vs Charges", sub="Figure3C", col
```

```
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
model3 <-lm(charges ~ age  + bmi + bmi30 + children + smoker + smoker:bmi30, data=insurance)
summary(model3)

par(mfrow=c(1,3))
plot(model3, 1) + title(sub = "Figure 4A", col.sub = "blue")
plot(model3, 2) + title(sub = "Figure 4B", col.sub = "blue")

plot(resid(model3) ~ charges, data=insurance) + title(main = "Residual vs Charges", sub="Figure4C", col
```

```
library('lsr')
library('e1071')
library('MASS')
library('Metrics')
library('ggplot2')
library('dplyr')
library('lattice')
library('caret')
library('randomForest')
library('tidyverse')
library('vip')
```

```
insurance <- read.csv("insurance.csv", header = TRUE)
insurance$sex <- as.factor(insurance$sex)
insurance$smoker <- as.factor(insurance$smoker)
insurance$region <- as.factor(insurance$region)
```

```
#10-fold cross validation: 9 training, 1 validation fold
#Cross validation w/ repetitions
crossval <- trainControl(method = "repeatedcv", number = 10, classProbs = FALSE)
```

```
#Regression Tree
set.seed(63)
regtree <- train(charges ~ ., data = insurance, method = "rpart",
                 metric = "RMSE", trControl = crossval)
```

```
#Random Forest
set.seed(63)
randfor <- train(charges ~ ., data = insurance, method = "rf",
                 metric = "RMSE", trControl = crossval)
```

```
summary(resamples(list(RT = regtree, RF = randfor)))
```

```
set.seed(700)

randfor2 <- randomForest(charges ~ ., data = insurance, importance = TRUE)
randfor2
```

```
plot(c(1: 500), sqrt(randfor2$mse), xlab = "Number of Trees", ylab = "RMSE",
     type = "l")
```

```
vip(randfor2, num_features = 6, geom = "point", include_type = TRUE)
```

```
pred_frame <- data.frame(
  Actual = insurance$charges,
  RF_Predicted = randfor2$predicted
)

ggplot(pred_frame, aes(x = Actual, y = RF_Predicted)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  xlab("Actual Charges") + ylab("RF Predicted Charges") +
  ggtitle("Random Forest Method (n = 500, mtry = 2)") +
  xlim(0, 65000) + ylim(0, 65000)
```

Bibliography

[1] *Assessing Your Weight*. (2022, May 3). Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/index.html

[2] Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, 08(04), 341–357.

[3] Bujokas, E. (2021). *Regression Tree in Python from Scratch*. Towards Data Science. Retrieved from https://towardsdatascience.com/regression-tree-in-python-from-scratch-9b7b64c815e3

[4] *Can I be charged higher premiums in the Marketplace if I smoke?* (2020, July 15). KFF. https://www.kff.org/faqs/faqs-health-insurance-marketplace-and-the-aca/can-i-be-charged-higher-premiums-in-the-marketplace-if-i-smoke/

[5] Harvard T.H. Chan School of Public Health. (n.d.). *Obesity Consequences - Health Risks*. Obesity Prevention Source. Retrieved from https://www.hsph.harvard.edu/obesity-prevention-source/obesity-consequences/health-effects/

[6] Klop, B., Elte, J., & Cabezas, M. (2013). Dyslipidemia in Obesity: Mechanisms and Potential Targets. *Nutrients*, *5*(4), 1218–1240. https://doi.org/10.3390/nu5041218

[7] Lantz, B. (2013). *Machine learning with R*. Packt Publishing. https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv

[8] Shmerling, R. H., MD. (2021, July 13). *Is our healthcare system broken?* Harvard Health. https://www.health.harvard.edu/blog/is-our-healthcare-system-broken-202107132542\

[9] West, R. (2017). Tobacco smoking: Health impact, prevalence, correlates and interventions. *Psychology & Health*, *32*(8), 1018–1036. https://doi.org/10.1080/08870446.2017.1325890

[10] *Who's included in your household*. (2021). HealthCare.Gov. https://www.healthcare.gov/income-and-household-information/household-size/