Finding Machine Learning Models to Predict Health Insurance Costs

Akshat Srivastav

University of California – Los Angeles

Abstract

The aim of this paper is to find a Machine Learning model, or even a combination of many, that can predict the insurance charges of an individual with a reasonable accuracy. In this paper, I will present an *Introduction, Exploratory-Data-Analysis,* and *Fitting Models* section for a cohesive workflow.

Introduction

The ML-model is built using a dataset from Medical Cost Personal Datasets | Kaggle, which I believe is derived from *Machine Learning with R* by Brett Lantz. Based on variables such as age, sex, BMI, etc. the model can predict the Insurance Charges in USD.

Gaining a better understanding of the dataset is imperative before we begin understanding the process of creating an ML-model. Listed below are the names of the columns or features in the dataset.

i. age: age of the primary beneficiary

ii. sex: sex of the primary beneficiary

iii. BMI: Body mass index that provides an understanding of how heavy or light an individual is relative to their weight

iv. children: number of children or dependents covered by health insurance.

v. smoker: whether the primary beneficiary smokes

vi. region: the beneficiary's residential area: Northeast, Southeast, Southwest, or Northwest.

vii. charges: individual medical costs billed by the health insurance company.

In this paper, the variable *charges* is the response variable while the remaining features are the predictor variables. In other words, we predict *charges* using all the other variables.

Exploratory Data Analysis (EDA)

Carrying out an EDA is very important before we start building models. The aim of any EDA is to explore the dataset and the characteristics of its data through visualizations. EDA makes it easier to spot patterns and understand the relationships between dataset variables. In this exploration, I did not follow any protocol or workflow to carry out my EDA; I explored those aspects of the dataset that simply piqued my curiosity.

For this EDA, I relied on *Python's Seaborn* and *Matplotlib* libraries to create visualizations. My first question about this dataset was "What is the age distribution in this dataset?". A natural question to follow is "How are health insurance charges distributed?". Answers to both of questions are found in Figure 1 below.

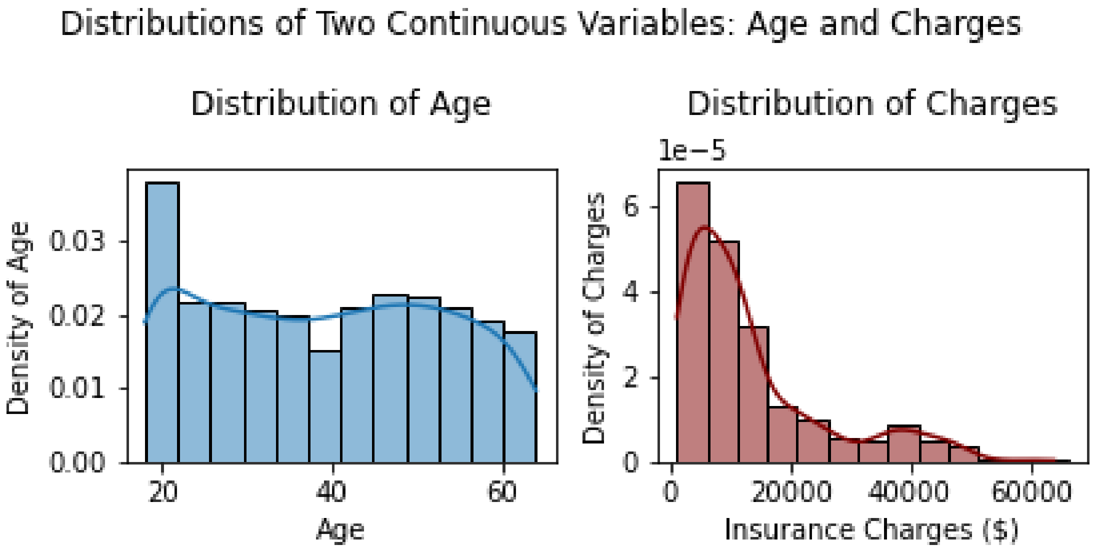


Figure 1

To visualize the distribution of two quantitative variables such as *age* and *charges*, I thought a histogram would be most appropriate. Using Python's *Seaborn* library, I created two histograms within the same subplot. Consider the distribution of age. There appears quite a high density of beneficiaries who are 18 to 19 years old. The density of age groups is roughly uniform at about 0.02. Shifting to charges, the distribution seems skewed to the right. Moreover, it appears that the median insurance cost or charges appears to be about $9000; to be precise, $9382.

Intuition has it that the older you get, the more expensive insurance becomes. Is that true? Look at Figure 2 for an answer.

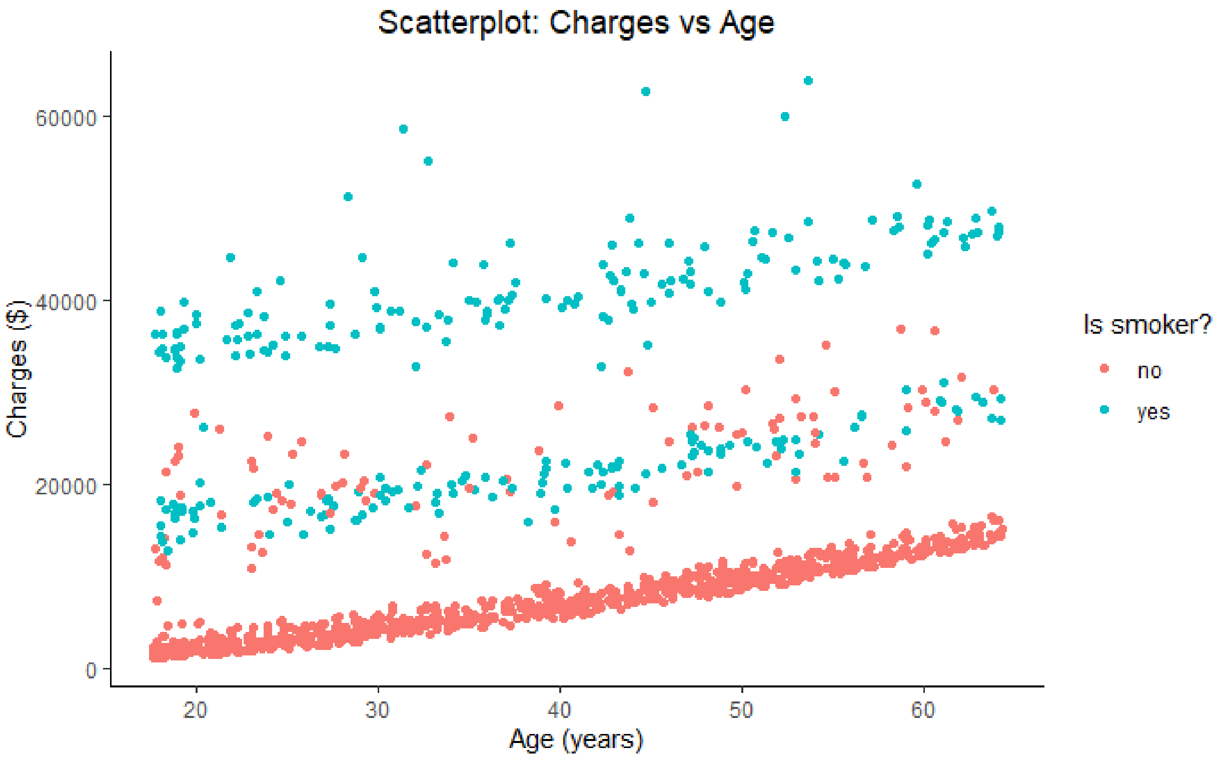


Figure 2

Although we discern a general trend whereby the charges tend to increase with the age of the primary beneficiary, notice we can see roughly three strata or layers. These strata, although parallel, appear separated based on the smoking status of the primary beneficiaries. Notice, that the lowest strata, representing the lowest insurance charges, is composed primarily of non-smokers. The middle strata, higher charges, is composed of smokers and non-smokers alike. However, the highest strata consist primarily of smokers. Figure 2 gives us a preemptive intuition that while the insurance charges increase with age, smoking status does have an important role to play in determining insurance charges, i.e., we must include the feature *smoker* in our model as a predictor.

While smoking makes an individual more vulnerable to poor health, obesity can also make a primary beneficiary more liable to insurance. As obesity is linked to heart disease, (Penn Medicine, 2019), we would expect that a higher BMI is linked to higher insurance charges.

Let us find out!

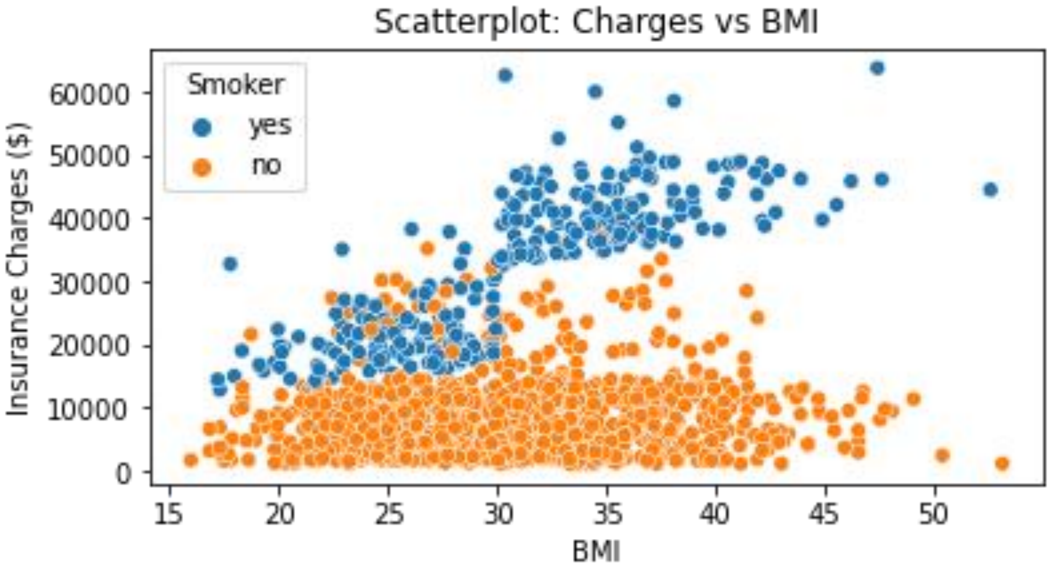


Figure 3

Figure 3 is a scatterplot depicting Insurance Charges on the y-axis and BMI on the x-axis. The correlation between Insurance Charges and BMI seems to depend on smoking-status. For non-smokers (orange dots), the Insurance Charges do not really change with BMI whereas for smokers (blue dots), higher BMI seems to lead to higher insurance charges.

Does region affect insurance charges?

I was curious to see if there is geographical disparity in the insurance charges incurred by primary beneficiaries across the United States. To answer this question, I looked at the distribution of insurance charges by for each region. I also faceted the scatterplot in Figure 2 by region to see whether it affects the correlation between age and insurance charges in any way.

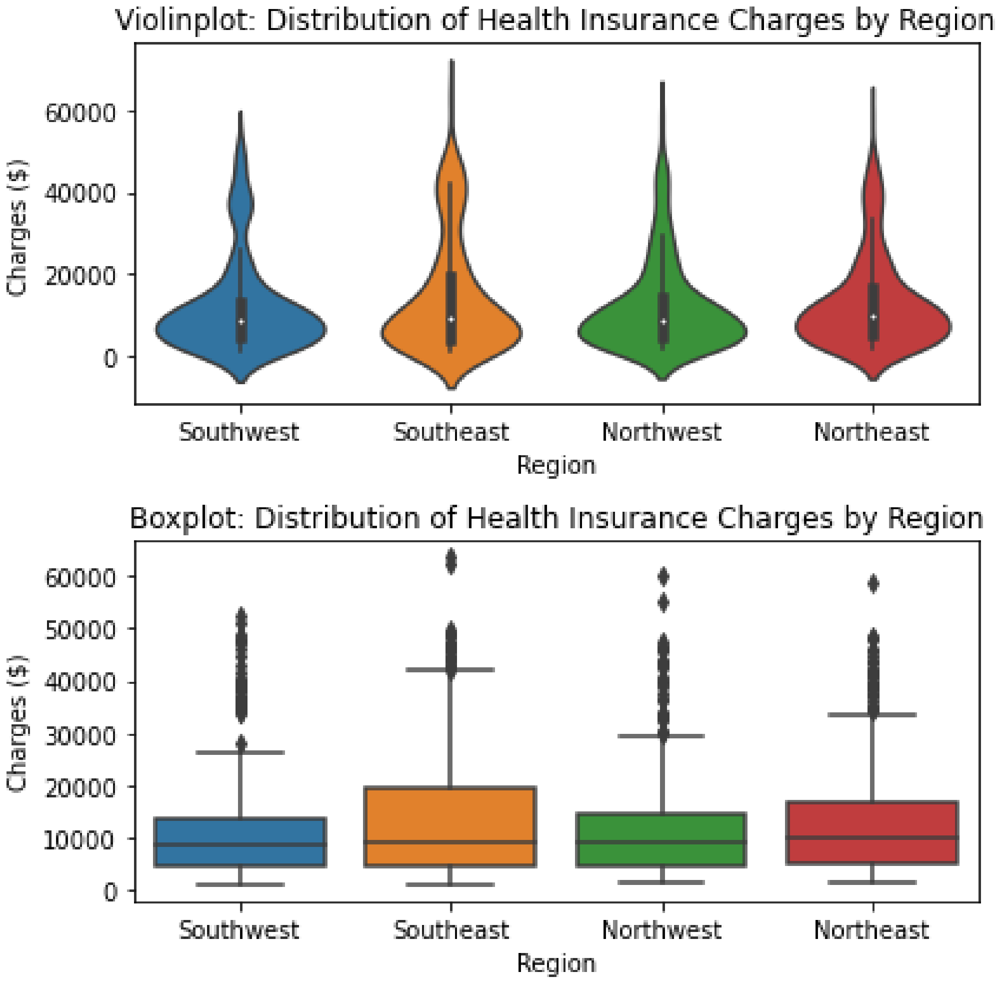


Figure 4

Figure 4 shows a Violin plot and Boxplot that demonstrate the distribution of insurance charges for each region. Although the third quartile differs from region to region, other characteristics of the distribution appear quite similar.

Figure 5 confirms the intuition that region does not really make a difference. The scatterplots, faceted by region, suggest that the correlation between charge and age are very similar across regions.

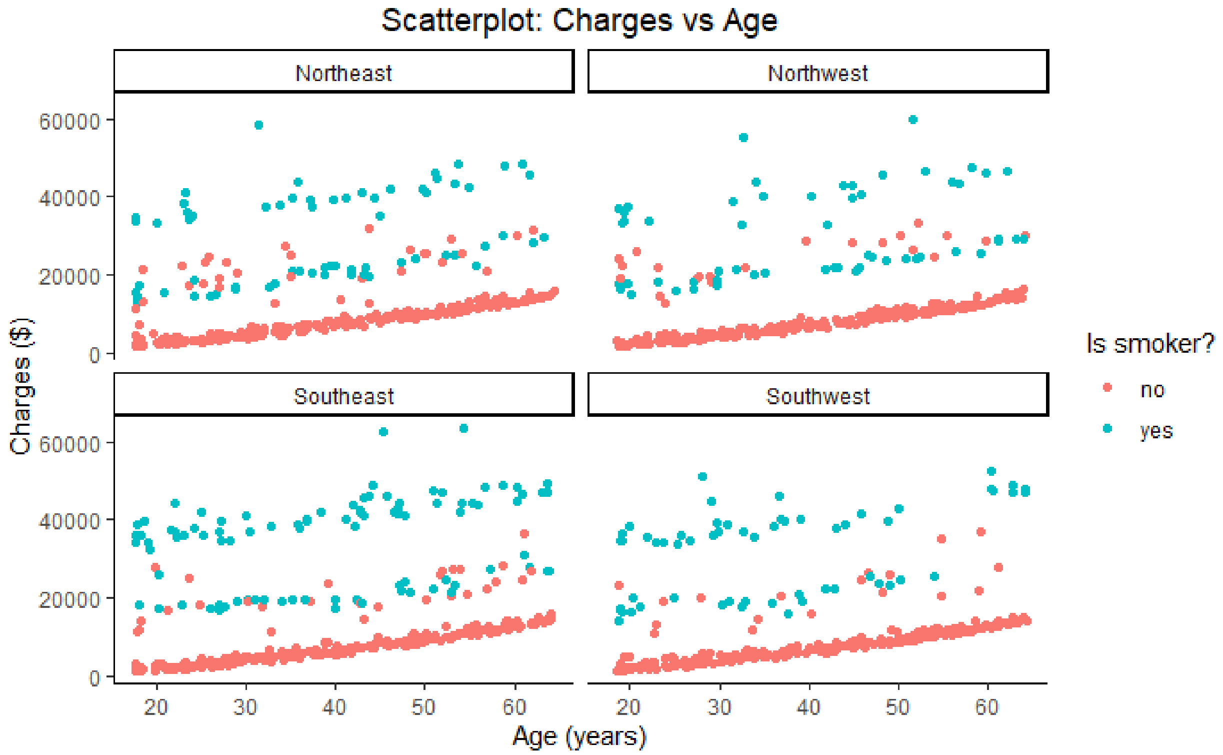


Figure 5

Does number of children affect insurance charges?

To answer the question above, we first look at the mean insurance charges and how they vary based on the number of children. We then visualize the interaction between smoking status and the number of children. To clarify, the number of children in this dataset represents the number of dependents covered on the insurance plan.

Figure 6 below shows the mean insurance charges in USD varying by the number of children. The figure suggests that as the number of children increases from 0 to 3, the mean insurance cost increases too. However, we see a decrease in the mean charges incurred as number of children increase from 4 to 5. Furthermore, this increase in the number of children is also characterized by a decrease in the variance, evidenced by the shorted error bars.

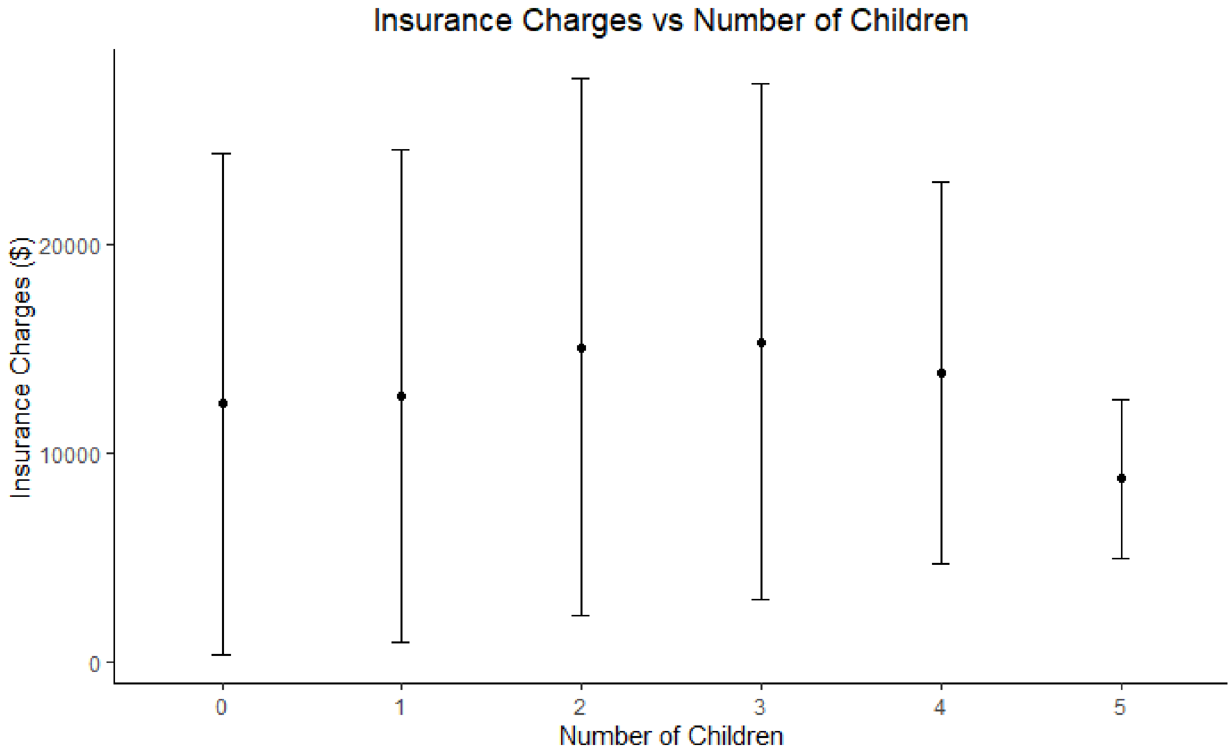


Figure 6

To learn more about the interaction between the smoking status and the number of children, consider the Violin plot below (Figure 7). Figure 7 demonstrates the distribution of insurance charges for the number of children, split by the smoking status. As a heads up, it appears that there is no data for insurance holders with five children who are smokers.

For each number of children, the distribution of charges for smokers has a greater median than that for non-smokers. Moreover, the overall distribution seems to have larger values for smokers compared to non-smokers. This shows us that the interaction between the two variables is significant as the smoking status, for each number of children significantly impacts the insurance charges incurred by policy holders.

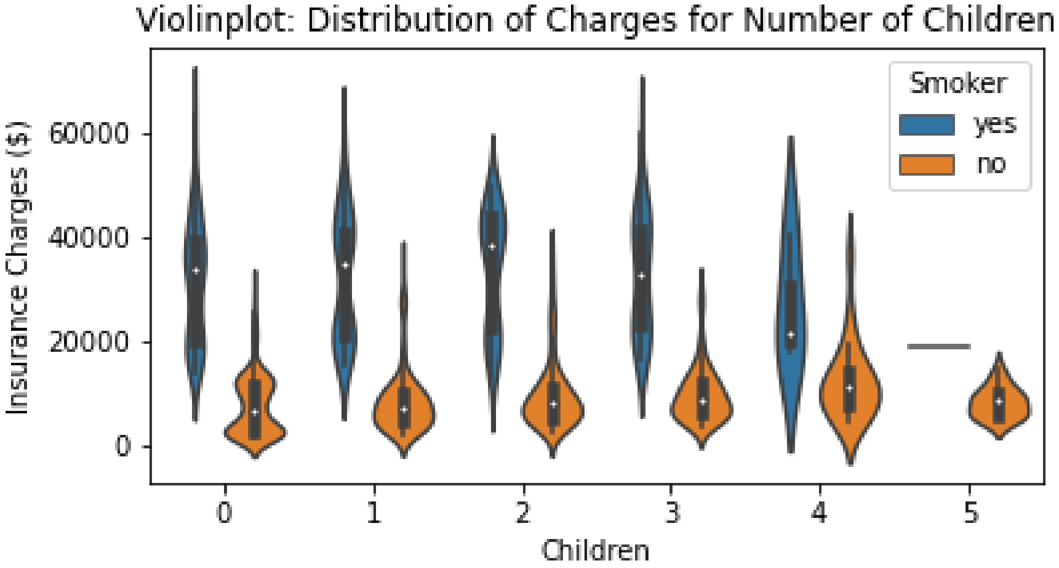


Figure 7

Does sex affect insurance charges?

To gauge the impact of sex on insurance charges, we plot the distribution of insurance charges for each sex in Figure 8. The distribution of insurance charges differs in the third quartile only as the lower quartile and the median are quite similar.

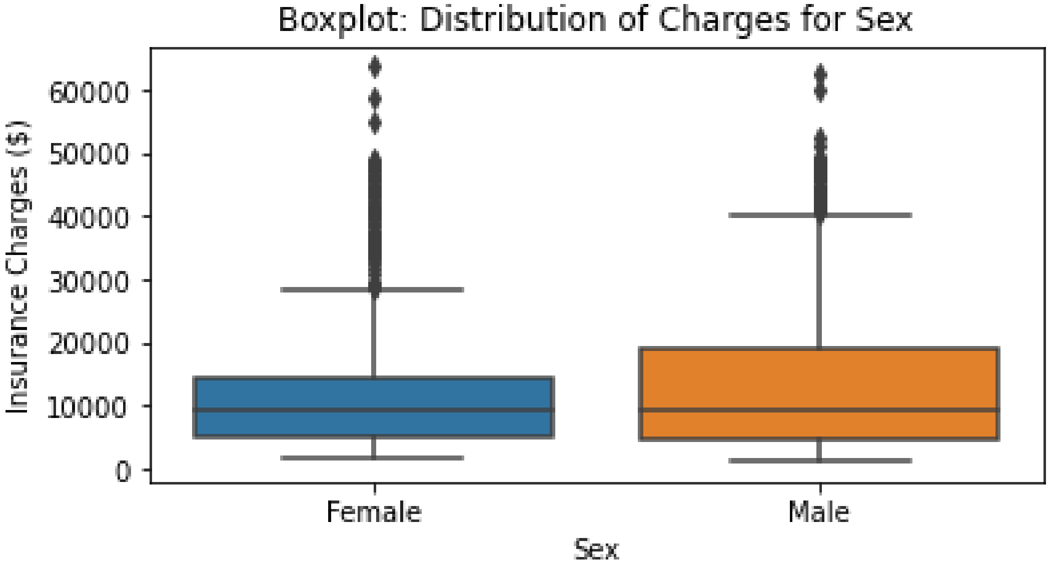


Figure 8

To get a more thorough idea about the effect of sex on insurance charges, we look at the scatterplot showing the correlation between charges and age, faceted on sex. We want to see whether sex affects the correlation between insurance charges and age. Refer to figure 9 below.

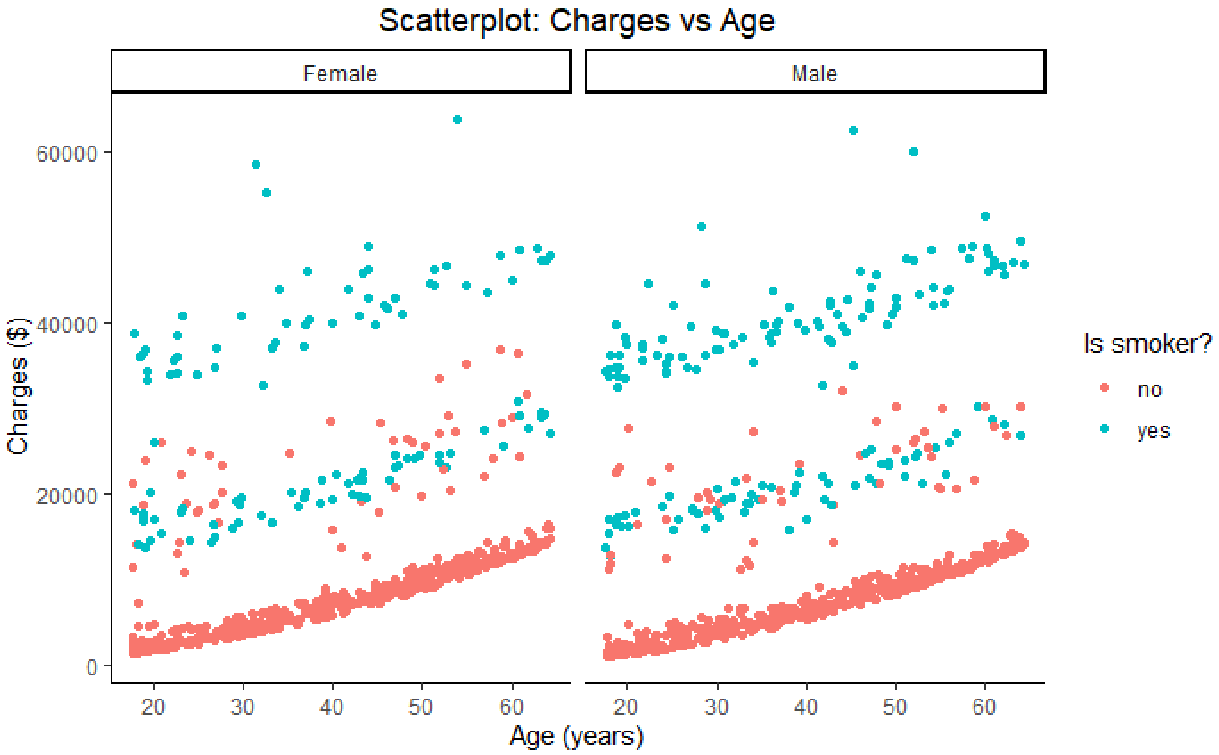


Figure 9

Both scatterplots look very identical. The correlation between charges and age seems unaffected by gender. Although the insurance charges for males appear to be slightly lower across all three strata compared to that of males, I do not think this difference is significant enough to say that *sex* affects the insurance charges incurred.

Finally, before discussing our Machine Learning models, it might be a good idea to look at the correlation between all the numerical variables. Refer to figure 10 below which encodes a heatmap using a correlation matrix for the numerical variables.

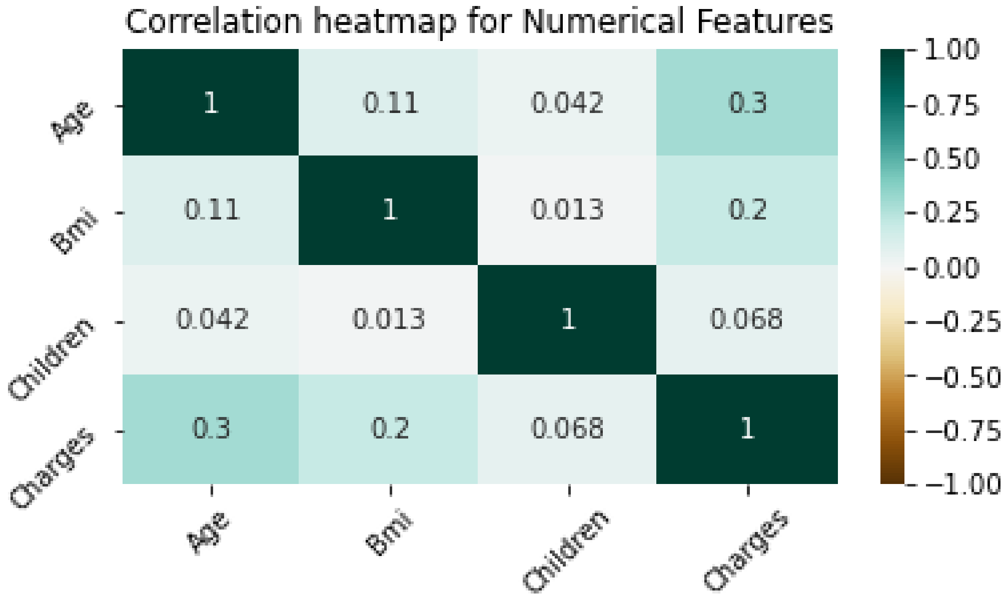


Figure 10

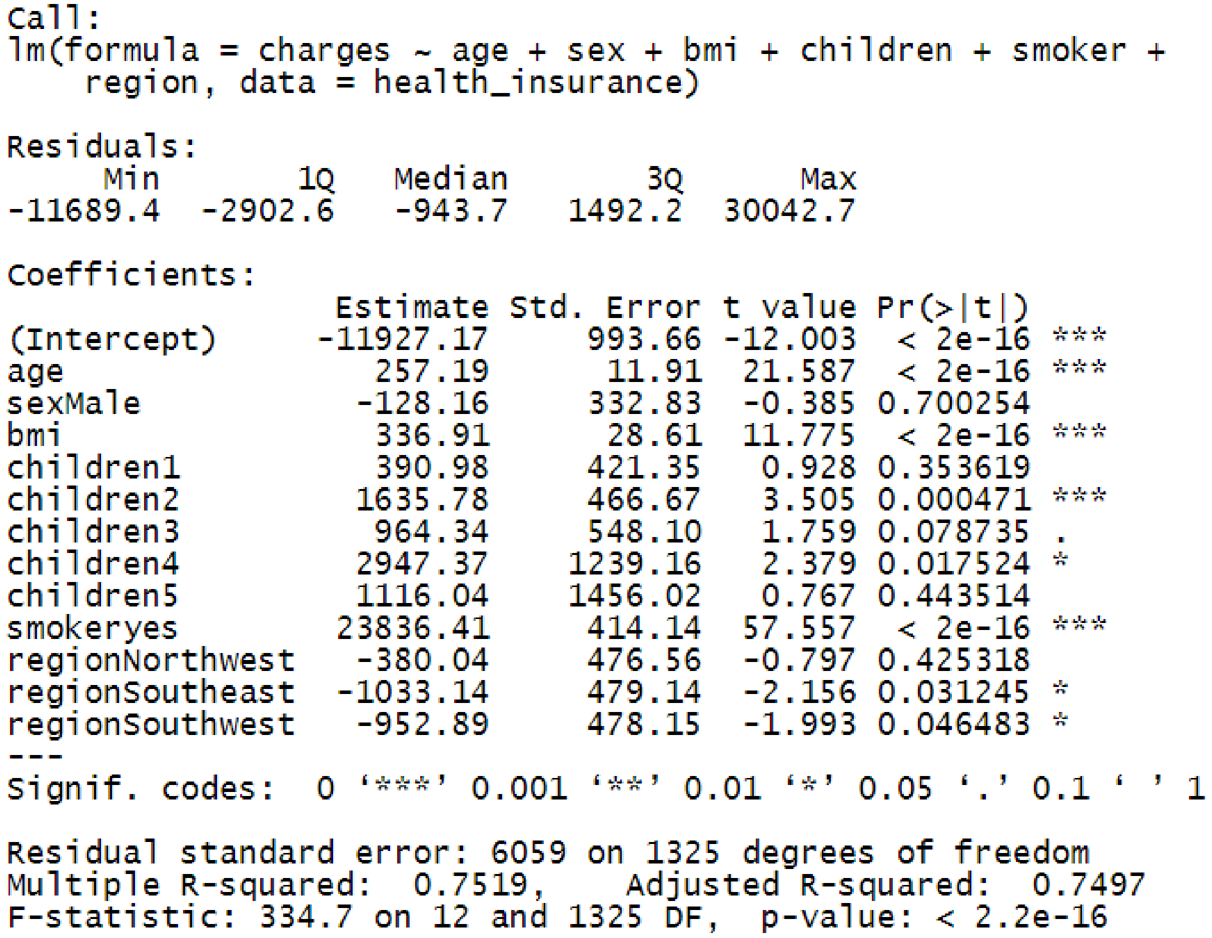
Summary of Machine Learning Models

Ordinary Least Squares Regression

For more details on code refer to the *health_insurance_eda.Rmd* notebook. The idea behind ordinary least squares is to minimize the *RSS* or the *Residual Sum Squared.* In this case, when we want to model the variable *charges* based on other predictor variables our model can be written in the form below.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{3i} + \ldots + \beta_p x_{pi} + e_i$$

Here, Y_i refers to *charges* and β_0 represents the intercept. Note that the ith subscript refers to the ith data point and e_i represents the ith random error. The parameters β_0 and β_i are obtained by minimizing the *RSS*. Using these least square estimates of β_0 and β_i, we formulate the above equation and predict the insurance charges. When I included all the variables, I got a multiple-R-squared value of 0.75. Refer to figure 11 below that shows the summary of the lm() function call.



```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
    region, data = health_insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11689.4  -2902.6   -943.7   1492.2  30042.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11927.17     993.66 -12.003  < 2e-16 ***
age                257.19      11.91  21.587  < 2e-16 ***
sexMale           -128.16     332.83  -0.385 0.700254
bmi                336.91      28.61  11.775  < 2e-16 ***
children1          390.98     421.35   0.928 0.353619
children2         1635.78     466.67   3.505 0.000471 ***
children3          964.34     548.10   1.759 0.078735 .
children4         2947.37    1239.16   2.379 0.017524 *
children5         1116.04    1456.02   0.767 0.443514
smokeryes        23836.41     414.14  57.557  < 2e-16 ***
regionNorthwest   -380.04     476.56  -0.797 0.425318
regionSoutheast  -1033.14     479.14  -2.156 0.031245 *
regionSouthwest   -952.89     478.15  -1.993 0.046483 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6059 on 1325 degrees of freedom
Multiple R-squared:  0.7519,	Adjusted R-squared:  0.7497
F-statistic: 334.7 on 12 and 1325 DF,  p-value: < 2.2e-16
```

Figure 11

I then used the *Bayes Information Criterion* for selecting the relevant variables and found that the features *age, BMI, children,* and *smoker* were most relevant. Including only these variables increased the R^2 value to 0.77. However, 0.77 is not satisfactory so I tried to use L1 and L2 regularization.

L1 (Lasso) and L2 (Ridge) regularization

Lasso Regression or L1 regularization is an effective method to deal with overfitting. The idea is that the L1 regularization adds a penalty term to the loss function that generally causes the coefficients to be smaller. Since, we found that only *age, BMI, children,* and *smoker* were relevant all models from now are fit only using these predictor variables. Using *sklearn.linear_model* we created an instance of *Lasso()* and fit this to our data. We then used *GridSearchCV* to tune our parameters, especially *alpha* that controls the penalty for overfitting. After tuning our parameters Lasso regression gave us an R^2 score of 0.794. Following a similar process for L2 regularization, we obtained an R^2 of 0.798.

SVR (Support Vector Regressor) and Linear SVR

The SVR is the regression equivalent of SVM (Support Vector Machine). I used the *Gaussian* kernel, which uses a *radial basis function*. Here, it was important to tune the *C* parameter that controls the minimum margin. After tuning the *C* parameter, we found an R^2 of about 0.8956. Repeating a similar process for LinearSVR, which is essentially an SVR with a linear kernel, we found an R^2 of only 0.65.

KNeighborsRegressor

The *KNeighborsRegressor* is the regression equivalent of *KNN* (K Nearest Neighbors). In this model, the target is predicted using targets associated with the nearest neighbors. To fit this model, we instantiate *KNeighborsRegressor()* from *sklearn.neighbors* and choose *n_neighbors*, *algorithm,* and *weights.* The parameter *n_neighbors* determines how many nearest points or neighbors will be used to predict the target for a test point, while the *weights* parameter determines how each point is weighted. For instance, you can have uniformly weighted points, i.e., points in the same neighborhood have uniform weights. Otherwise, the points can be weighted by the inverse of their distance, so closer points will have a greater influence. The argument *algorithm* determines the algorithm used to compute the nearest neighbors of the point. After doing a *GridSearchCV*, to tune the parameters we find the optimum parameters to be the following.

i. *n_neighbors: 9*

ii. *weights: uniform*

iii. *algorithm: auto*

With these parameters, the *KNearestRegressor* produces an R^2 of 0.881.

Random Forest Regressor

A Random Forest regressor uses ensemble methods for regression. It is a supervised learning algorithm that uses many decision trees to output the mean prediction of the different trees. With this algorithm, the parameter *n_estimators* determines the number of decision trees to use. The R^2 value with *n_estimators = 100* is 0.867.

Ensemble of Methods

I created an ensemble of regression methods using *sklearn's StackingRegressor* class. We use stacking as an ensemble learning technique to combine many multiple regression models. There is a higher level and a lower level of estimators. All the estimates of low-level models are passed into the high-level model. For our low-level estimators, we use *GradientBoostingRegressor*, *SVR, KneighborsRegressor*, and *RandomForestRegressor.* For our high-level estimator, we use *Lasso* regression. Using this ensemble, we obtain an R^2 of 0.902.

Bibliography

Penn Medicine. (2019, March 25). *Three Ways Obesity Contributes to Heart Disease*. Retrieved from Penn Medicine: https://www.pennmedicine.org/updates/blogs/metabolic-and-bariatric-surgery-blog/2019/march/obesity-and-heart-disease