

**Predicting Health Insurance Costs**  
**Using Regression**

Maddie Bauer

Bellevue University

DSC 680: Applied Data Science

Professor Catherine Williams

August 11, 2021

## **Introduction & Problem Statement**

Health insurance is a market that will never die out as most people have (or should have) some level of health insurance. This has especially been expressed throughout the past year and a half due to the worldwide COVID-19 pandemic. Illnesses, accidents, medications, surgeries, etc. all can add up in a person's yearly budget, but by having health insurance, an individual is not liable for paying an entire medical bill. One of the many important tasks within the health insurance industry includes determining policy premiums for their customers. With predictive modeling, insurance companies can determine an accurate and fair policy premium based on a customer's behaviors and their attributes. By analyzing customer habits and behaviors over time, insurers are able to anticipate future behaviors and provide the right insurance products and policy premiums. This project will aim to predict future premiums through the use of regression models.

## **Data**

The dataset being used for this research paper was retrieved from Kaggle.com. It is a dummy dataset as real insurance data is not available to the public due to privacy laws. However, I am confident that the dataset I am using is comparative to real-world data because according to healthcare.gov, premiums are set by looking at only your age, location, tobacco use, individual vs. family enrollment and the plan category (healthcare.gov, n.d.).

The Kaggle dataset consists of the following variables: age, gender, BMI, children, tobacco use, region, and yearly charges. More information on each variable is described in table 1 below.

| <u>Variable</u> | <u>Description</u>   |
|-----------------|--|
| Age             | Age of client  |
| Sex             | Gender (Male/Female)   |
| BMI             | Body Mass Index  |
| Children        | Number of dependents   |
| Smoker          | Whether or not a client smokes (yes/no)                                |
| Region          | Where the client lives<br>(Northeast, Northwest, Southeast, Southwest) |
| Charges         | Yearly medical costs paid by client                                    |

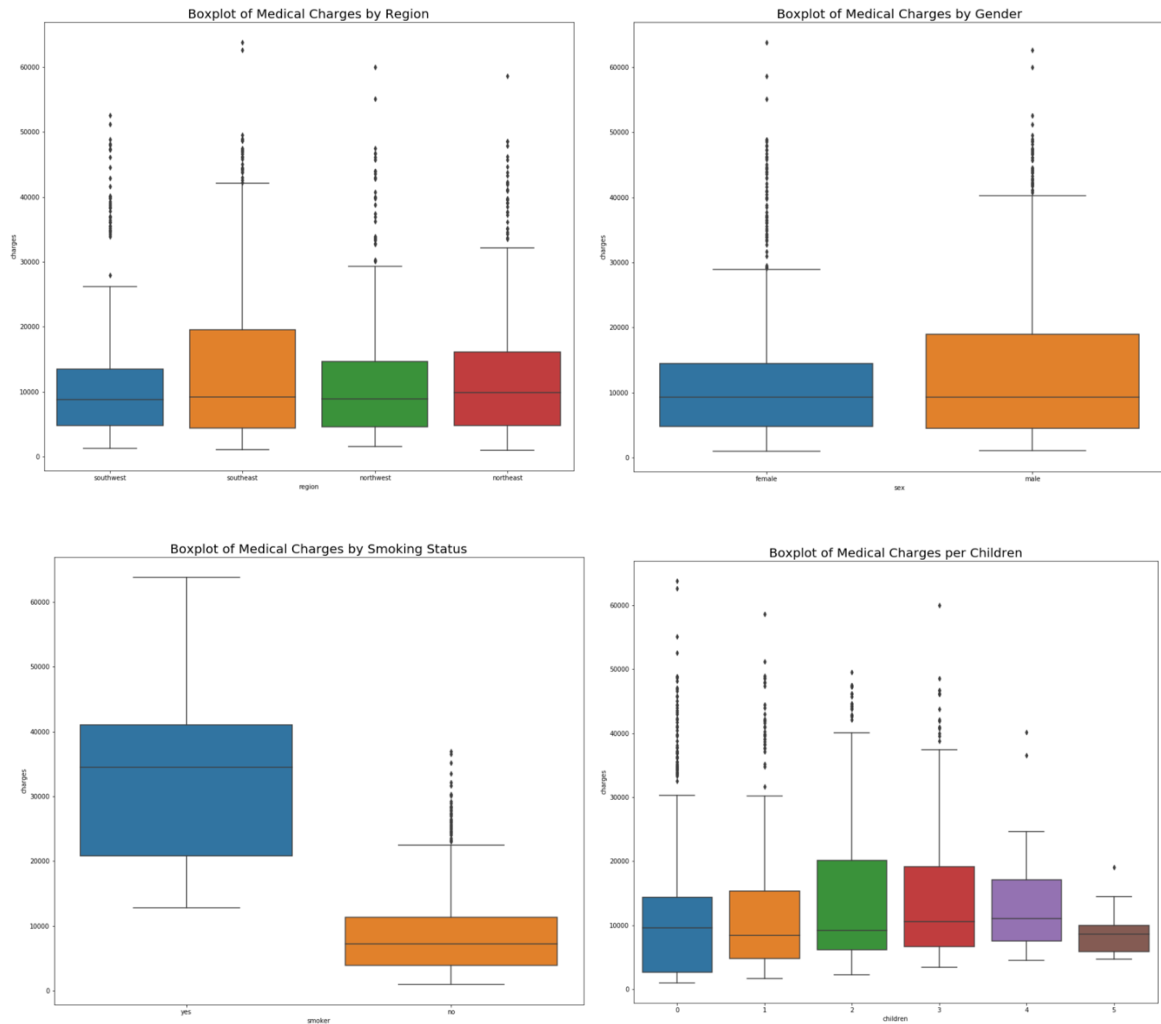
Table 1.

In addition to this dataset from Kaggle, I also collected data from my close family and friends who did not mind sharing it with me so that I could further my learning. This means that there are more data points altogether so that we can ultimately make better predictions with the model. The only inconsistency was that a few of my friends and family members had a hard time remembering their BMI from their most recent doctor's visit. In the case that they couldn't remember their exact BMI, they estimated to the best of their knowledge.

The full dataset consists of 1,349 data points. Age, BMI, children and charges are numerical whereas sex, region and smoker are categorical. There are no missing values within the dataset and no outliers to be concerned of.

## Methods

The overall plan with this dataset is to explore the relationships among the variables and determine which features best effect the yearly medical charges. After merging the two datasets, both univariate and bivariate analysis were conducted through data visualizations. Below we can see the spread of several independent variables along with their impact on the target variable, charges.



Based on the boxplots above, medical charges are not highly affected by the region nor is gender. However, smokers tend to pay a significant amount more than non-smokers. People with two children tend to pay a little bit more than people with four or five children.

Next, I one-hot encoded the categorical variables so they could be used in a machine learning model. Table 2 shows the breakdown of how the categorical variables were encoded.

| <u>Variable</u> | <u>Encoded</u>   |
|-----------------|--|
| Sex             | Female = 0<br>Male = 1   |
| Smoker          | Yes = 1<br>No = 0  |
| Region          | Northwest = 1<br>Southeast = 2<br>Southwest = 3<br>Northeast = 4 |

Table 2

I also checked the correlation between variables and specifically looked for how strong the relationship between each independent variable and the target variable was. Figure 1 shows the correlation matrix below. The following conclusions can be made in relation to the target variable:

- Smoking has a high correlation with charges
- Age and BMI have a moderate correlation with charges
- Number of children and gender have little effect on charges
- Region has basically no effect on charges

Since the region's correlation with charges is very small ( $<0.002$ ), I decided to proceed without this variable for the regression models.



Figure 1

Regression analysis is a predictive method used to explore the relationship between a dependent (target) variable and the independent (predictor) variables. Regression is typically used to forecast or make predictions with new input values (Regression Analysis, n.d.). For this analysis, two different regression models were created to estimate health insurance costs based on the five independent variables. By using regression models, we can then predict future health insurance costs for new customers. Before any models were built, the data was split into training (80%) and testing sets (20%) so that we can evaluate how well each model performs.

The first model built for this analysis was created using Linear Regression. Specifically, I used multiple linear regression as there are five independent variables. Below is the equation for multiple linear regression in which  $\hat{Y}$  represents the predicted outcome,  $b_0$  is the value of Y when all the independent variables ( $x_1$  through  $x_n$ ) are equal to zero, and the  $b_1$  through  $b_n$  are the estimated regression coefficients.

$$\hat{Y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

Each of the regression coefficients represent the change in Y in relation to a one-unit change in the respective independent variable while holding all other independent variables constant (LaMorte, 2016).

The second model built for this analysis was created using Random Forest Regression. A random forest consists of a collection of decision trees that are built on random samples with different policies for splitting a node. Decision trees are best described as being similar to tree diagrams (see figure 2 below) and are used to implement a sequential decision process (R, 2020). Starting at the top, or the root node, a decision is made and one of the two branches below it is selected. Each node within

the tree is a decision or rule. You continue down the tree following the nodes until you reach the final leaf, which represents the target value.

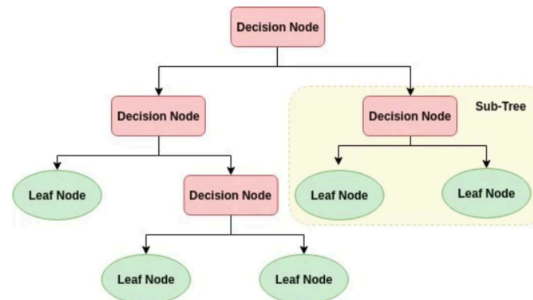


Figure 2 (Giraud, 2020).

Some advantages of random forests include that they handle outliers well and are indifferent to non-linear variables (Giraud, 2020). This helps limit overfitting and ultimately predicts the target variable more accurately (Giraud, 2020).

## Results

After creating and fitting the models discussed above, each model was then scored by calculating its R squared value. R squared is a calculation which gives us the error between the actual values from the test set and the predicted values from the model. Visually, you can see how each model did in table 3 below. The random forest regressor performed the best with an R squared score of 96.5% whereas the linear regression model only had an R squared score of 74.4%.

Based on the results above, we would choose to use the random forest regressor moving forward as it performed better. We can predict a new customer's insurance cost by inputting their age, gender, BMI, number of children and smoking status into our random forest regressor model we created. For example, the insurance cost for a 30-year-old male who doesn't smoke with a BMI of 24 and no children is predicted to be \$3,403.30 using our model.

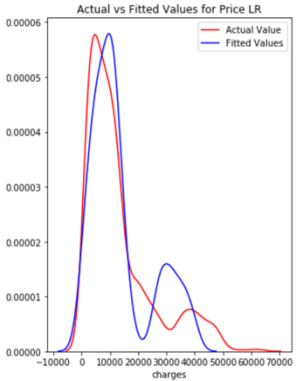
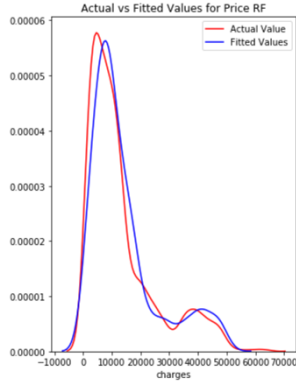
| <u>Model</u>             | <u>Accuracy</u> | <u>Plot</u>   |
|--------------------------|-----------------|---|
| Linear Regression        | 74.4%           |  <p>The plot for Linear Regression shows the distribution of actual values (red line) and fitted values (blue line) for 'charges'. The x-axis ranges from -10000 to 70000, and the y-axis ranges from 0.00000 to 0.00006. The fitted values curve is smoother than the actual values curve, capturing the general shape of the distribution with a primary peak around 10,000 and a secondary peak around 35,000.</p> |
| Random Forest Regression | 96.5%           |  <p>The plot for Random Forest Regression shows the distribution of actual values (red line) and fitted values (blue line) for 'charges'. The x-axis ranges from -10000 to 70000, and the y-axis ranges from 0.00000 to 0.00006. The fitted values curve closely follows the actual values curve, capturing the distribution's peaks around 10,000 and 35,000 with high fidelity.</p>                                |

Table 3

## Conclusion

While the healthcare industry is constantly evolving, I think it is safe to say it will always be around. Individuals cannot afford to pay all of their medical expenses up front, especially when accidents or surprising diagnoses occur. Thoughtful use of predictive analytics has only recently allowed insurance companies to improve their premium pricing accuracy, create customized plans for individuals and their families and help build stronger relationships with their customers. By analyzing customer behaviors and attributes overtime, we can accurately predict future premiums through regression analysis.



## **Acknowledgments**

I would like to thank Kaggle for providing the dataset used in this project. Also, I would like to thank our Professor Catherine Williams from Bellevue University as well as the students of DSC680 for their continuous support throughout this course. Finally, I would like to express my gratitude for Wikipedia, KDNuggets.com, Medium.com and their countless authors for their research and articles on the subject matter. This report would not have been possible without their contributions and sharing of information online.

## References

Giraud, A. (March 21, 2020). Quick intro to random forest. Retrieved August 5, 2021 from <https://towardsdatascience.com/quick-intro-to-random-forest-3cb5006868d8>

Healthcare.gov. (n.d.). How insurance companies set health premiums. Retrieved August 5, 2021 from <https://www.healthcare.gov/how-plans-set-your-premiums/>

Kumar, A. (June 22, 2020). Random forest for prediction. Retrieved August 5, 2021 from <https://towardsdatascience.com/random-forest-ca80e56224c1>

LaMorte, W. (May 31, 2016). The multiple linear regression equation. Retrieved August 5, 2021 from [https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713\\_multivariablemethods/bs704-ep713\\_multivariablemethods2.html](https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/bs704-ep713_multivariablemethods2.html)

Prasad, R. (March 3, 2020). Machine learning for beginners. Retrieved August 5, 2021 from <https://betterprogramming.pub/machine-learning-for-beginners-predicting-insurance-costs-using-linear-regression-40989645dfa3>

R, A. (March 8, 2020). The basics of decision trees. Retrieved August 5, 2021 from <https://medium.datadriveninvestor.com/the-basics-of-decision-trees-e5837cc2aba7>

Regression Analysis. (n.d.). [https://en.wikipedia.org/wiki/Regression\\_analysis](https://en.wikipedia.org/wiki/Regression_analysis)

## Questions:

1. Why keep the sex variable in your models with such a low correlation?
2. Does it seem plausible that the region doesn't affect charges?
3. Are there other machine learning regression models that could have been used?
4. What made you choose linear regression and random forest regression?
5. Use your models to predict the medical charges for the same person. How close are their results?
6. What are other areas in which insurance companies may use machine learning?
7. How long can this model be used?
8. How often would you recommend updating the model? Monthly, quarterly, yearly, etc?
9. What other applications can insurance companies use machine learning?
10. Are there other calculations that could be used to evaluate a model?