Predicting Health Insurance Costs Using Regression

Maddie Bauer

Table of Contents

PROBLEM STATEMENT

VIEW THE DATASET

METHODS USED

RESULTS

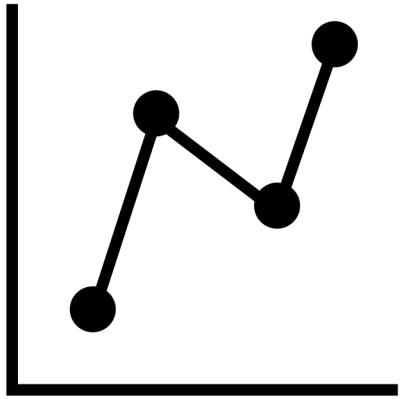
Problem Statement

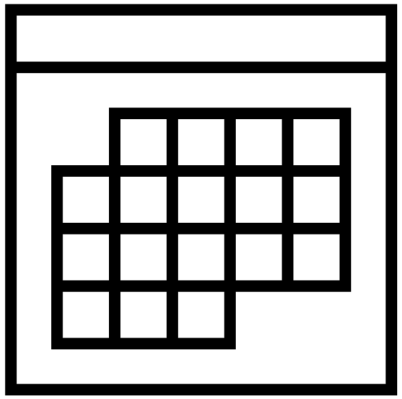
Health insurance is a must!

It is here to stay.

Illnesses, accidents, medications, and surgeries, can all add up in a person's yearly budget. By having health insurance, an individual is not liable for paying an entire medical bill.

A common application within the health insurance industry is **determining policy premiums for customers**.





The Goal

- To predict fair health insurance premiums for new customers by creating and implementing regression models with past customer data.

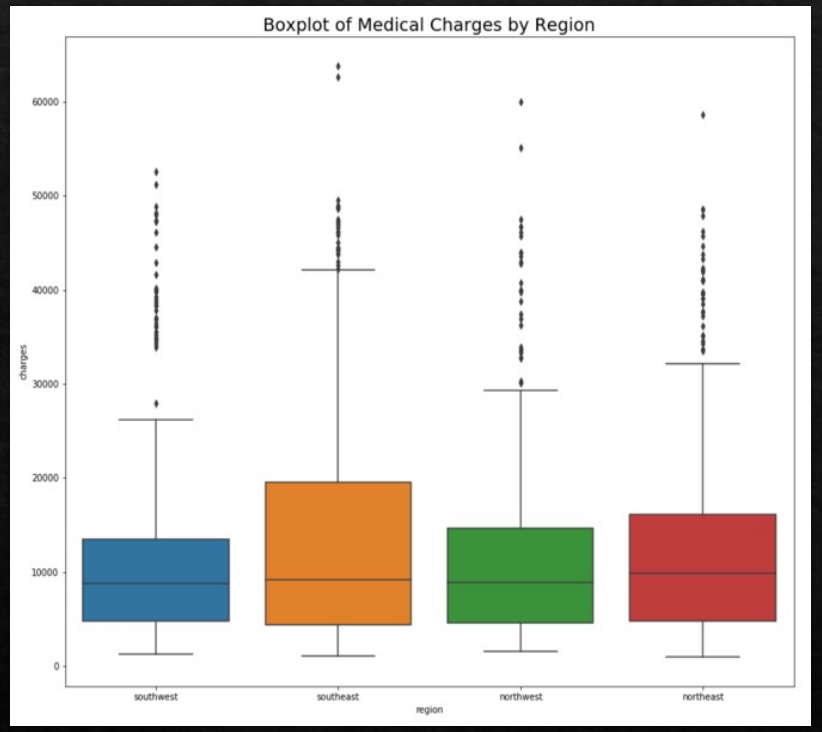
The Dataset

- Kaggle
  - 1,338
- Personal Data
  - 11
- Total: 1,349 data points

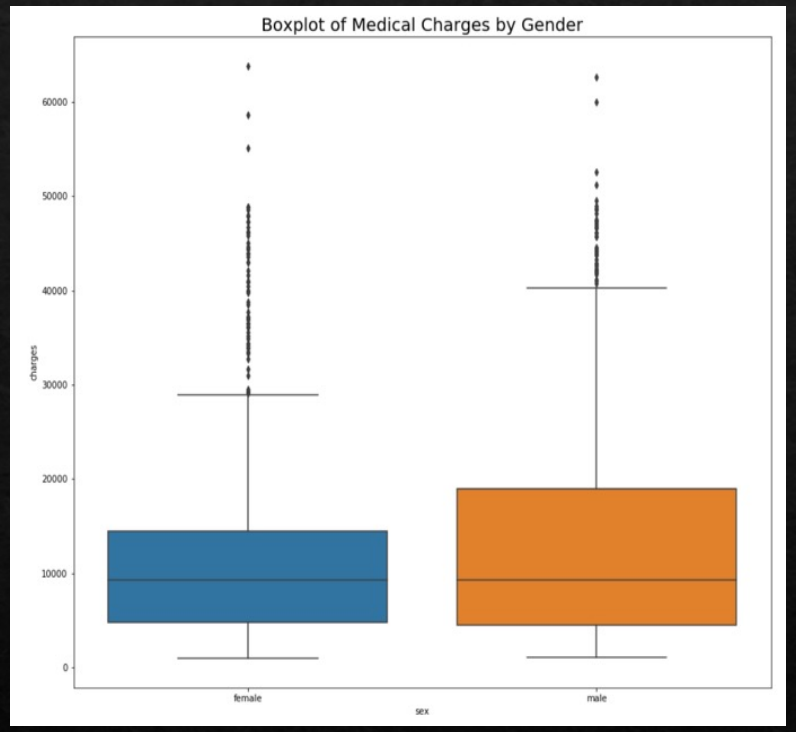
Variable	Description
Age	Age of client
Sex	Gender (Male/Female)
BMI	Body Mass Index
Children	Number of dependents
Smoker	Whether or not a client smokes (yes/no)
Region	Where the client lives (Northeast, Northwest, Southeast, Southwest)
Charges	Yearly medical costs paid by client

The Dataset Continued…

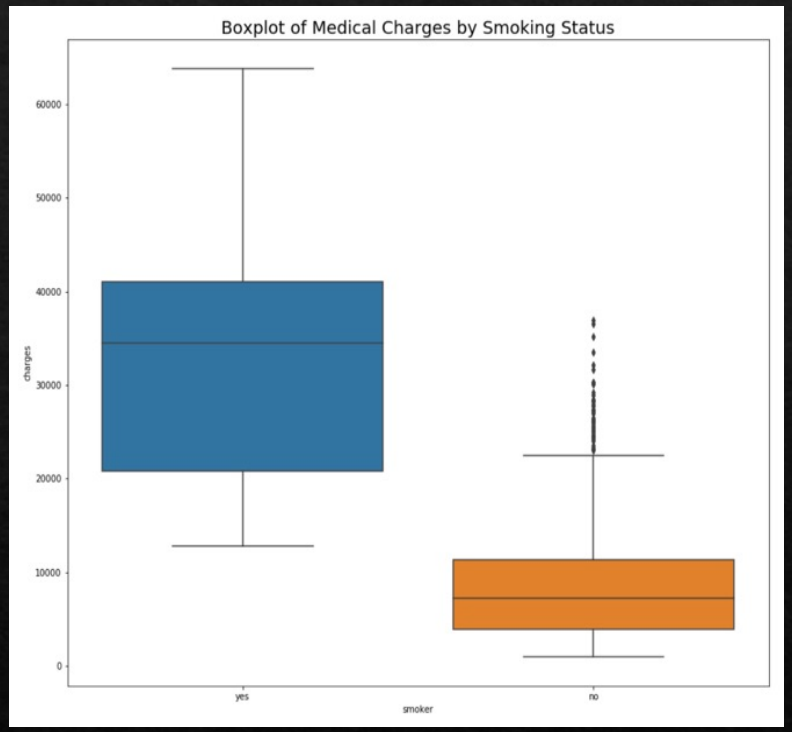
charges



region



sex



smoker

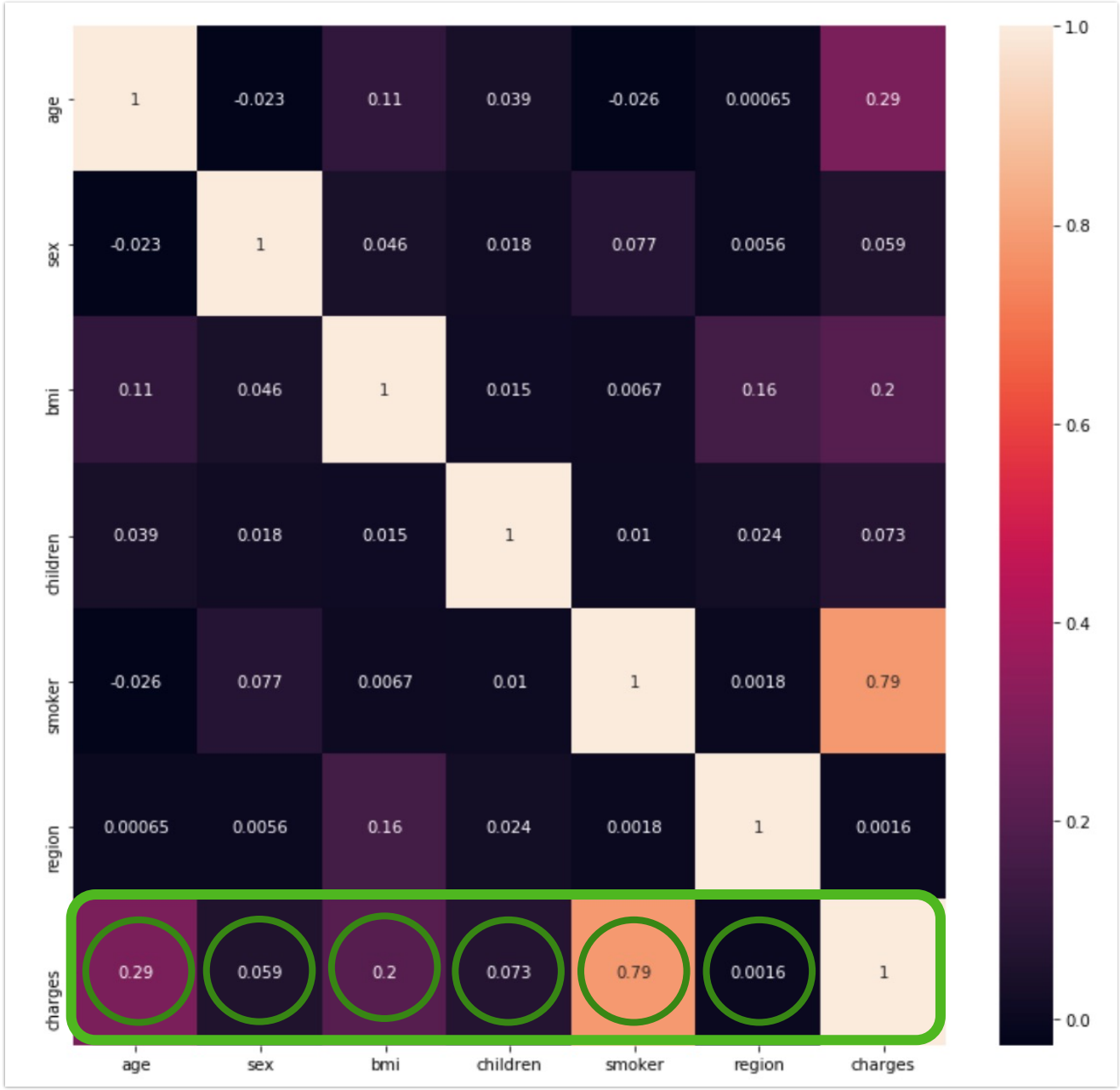
One-Hot Encoded Variables

Variable	Encoded
Sex	Female = 0 Male = 1
Smoker	Yes = 1 No = 0
Region	Northwest = 1 Southeast = 2 Southwest = 3 Northeast = 4

Correlation with Target Variable

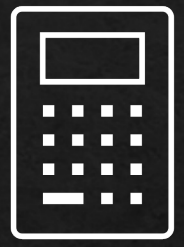
- Smoking has a high correlation at 0.79
- Age has a moderate correlation at 0.29
- BMI has a moderate correlation at 0.2

- Number of children & gender have small/low correlation
- Region has basically no correlation



Regression Analysis

- Predictive method used to explore the relationship between a dependent variable and the independent variable(s)
- Used to forecast or make predictions with new input values
- Two different regression models were created for this project
- The data was split into training (80% of data) and testing sets (20% of data)



Regression Analysis

Linear Regression

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

- $\hat{Y}$ 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
- b_1 through b_n are the estimated regression coefficients
- Each regression coefficient represents the change in Y in relation to a one-unit change in the respective independent variable while holding all other independent variables constant

Random Forest Regression

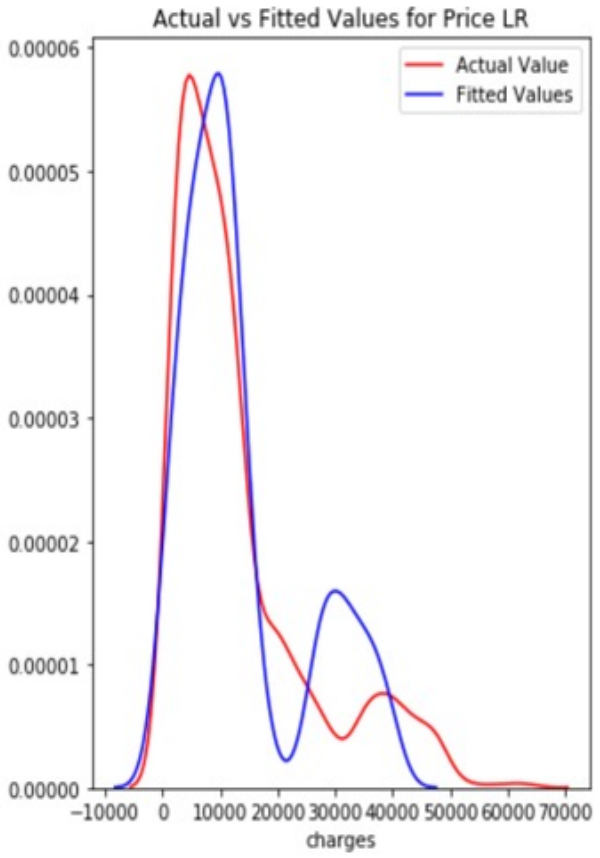
- A collection of decision trees that are built on random samples with different policies for splitting a node
- Used to implement a sequential decision process
- Each node within the tree is a decision or rule
- You continue down the tree following the nodes until you reach the final leaf, which represents the target value

By using regression models, we can then predict future health insurance costs for new customers.

Results

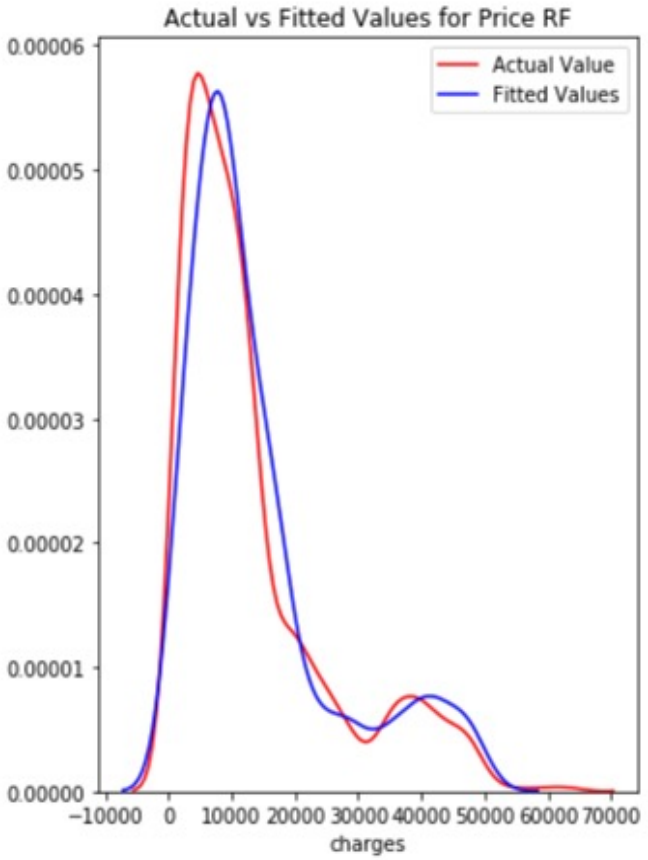
- R Squared or R^2
  - Calculation that gives us the error between the actual values from the test set (red line) and the predicted values from the model (blue line)

Linear Regression



R^2 = 74.4%

Random Forest Regression



R^2 = 96.5%

Using The Model

- **New Customer**
  - 30-year-old
  - Male
  - Doesn't Smoke
  - BMI of 24
  - No Children

Cost Prediction = $3,403.30

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 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.

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
- 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

Thank you!