‘What’s the Price?’ Predicting the Cost of Health Insurance, Given Various Factors

EXECUTIVE SUMMARY

One of the biggest issues plaguing health insurance is its high cost, which cannot be directly calculated by individuals beforehand. There are many factors that drive the increasing costs of health insurance, some of which are directly related to the personal attributes of the policy holder. This is a problem that affects the general population. With our dataset we obtained from a Kaggle data repository, we examined certain characteristics of individuals and performed analysis to determine which characteristics contributed most to health insurance costs. Once we determined these characteristics, we used regression models to predict insurance costs.

We determined the most significant characteristics that contribute to health insurance costs were the age of the policy holder, their body mass index (BMI), and whether they smoke or not. Body mass index is a value derived from the weight and height of a person, calculated using weight (in kilograms) divided by the square of the height (in meters). We used the data in the training set to train three models: linear regression, decision tree, and random forest. Then, we used the models to predict health insurance costs to compare their performances. We used R-squared and mean absolute error as our performance metrics. R-squared measures how well the model fits the data. R-squared values are between 0 and 1, where values closer to 1 indicate better performance. Mean absolute error measures the average difference between predicted and actual values. Mean absolute error are positive values, where values closer to 0 indicate better prediction. The random forest performed better than the other two models, with an R-squared of 0.78 and mean absolute error of 2833.14. We fine-tuned the parameters of the random forest model and prediction performance improved, with an R-squared of 0.86 and mean absolute error of 2022.03. These results indicate the random forest model fits 86% of our data and our predictions of health insurance costs are off by $2022.03 on average.

INTRODUCTION

Health insurance is a type of insurance that helps pay for medical expenses. It works by pooling resources and spreading the financial risk associated with major medical expenses across the entire population to protect those who require medical attention. Whether purchased privately or funded by the government, medical insurance is essential to provide protection against increasing medical costs. The Affordable Care Act of 2010 was designed to extend health insurance to those who could not afford it; however, costs continue to grow.

PROBLEM

With the cost of healthcare in the country today, the need for health insurance is stressed and overemphasized. An individual essentially needs health insurance to maximize their savings on healthcare. One of the biggest issues plaguing health insurance is its high cost, which cannot be directly calculated by individuals beforehand. There are many factors that drive the increasing costs of health insurance, some of which are directly related to the personal attributes of the policy holder. Our project is aimed to determine which personal attributes contribute most toward health insurance costs and use regression models to predict the cost based on these attributes.

METHODS

We obtained data from Kaggle that contained 1338 records of insured individuals, each providing information about the following variables:

- age – the age of the policy holder, ranging from 18 to 64 with the mean age of 39
- sex – binary variable indicating whether the policy holder is male (M) or female (F), contains 676 males and 662 females
- BMI – body mass index of the policy holder, ranging from 15.96 to 53.13 with a mean BMI of 30.66

- children – the number of children insured through the insurance policy, ranging from 0 to 5, with a mean of 1
- smoker – binary variable indicating whether the policy holder smokes (yes) or not (no), contains 274 that smoke and 1064 that do not smoke
- region – indicates where the policy holder resides: 324 from the northeast, 325 from the northwest, 364 from the southeast, and 325 from the southwest
- charges – the cost of health insurance in dollars, ranging from $1121.87 to $63770.43 with a mean of $13270.42

We converted the binary variables, sex and smoker, into binary labels, where female = 0 and male = 1 and no = 0 and yes = 1. We also converted the regions into encoded labels, where northeast = 0, northwest = 1, southeast = 2, and southwest = 3. We transformed the BMI variable from a continuous variable into encoded labels by binning the values into the following categories, as provided by the CDC:

- underweight = 0 , when BMI < 18.5
- normal = 1, when 18.5 ≤ BMI < 25
- overweight = 2, when 25 ≤ BMI < 30
- obese = 3, when BMI ≥ 30

The data contained records with 20 underweight BMIs, 225 normal BMIs, 386 overweight BMIs, and 707 obese BMIs. Finally, many insurance companies only charge for a maximum of three children, requiring no additional fee if more than three children are on the policy. As a result, we combined observations with 4 and 5 children with those of 3. The children variable can be thought of as a categorical variable, where children = 3 represents “3 or more” children. The final breakdown of value counts was: 574 with 0 children, 324 with 1 child, 240 with 2 children, and 200 with 3 or more children.

After preparing our data, we created histograms of our numerical variables, as seen in *Figure 1*, and bar plots of our categorical variables, as seen in *Figure 2*.

Figure 1: Histograms of age and children

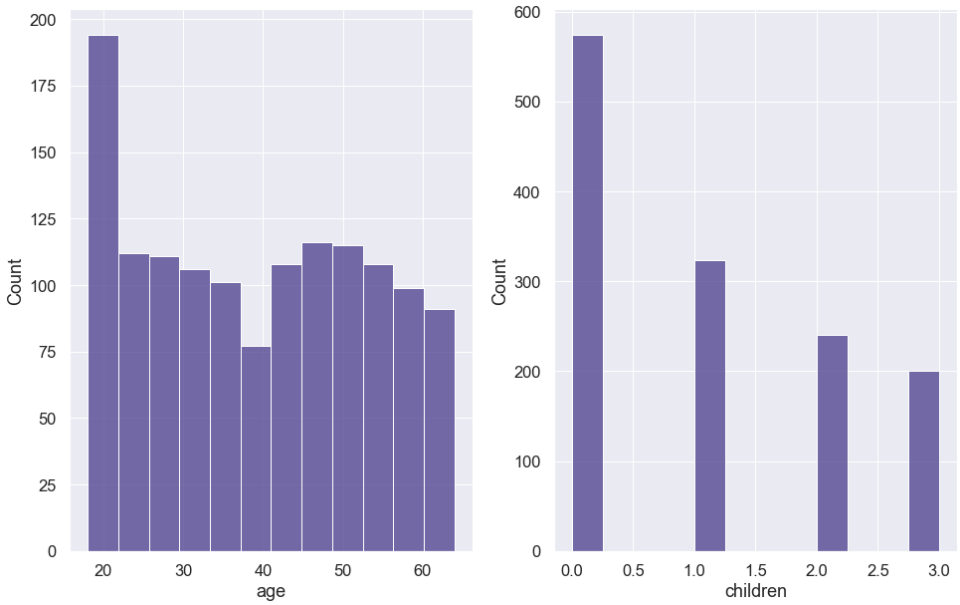
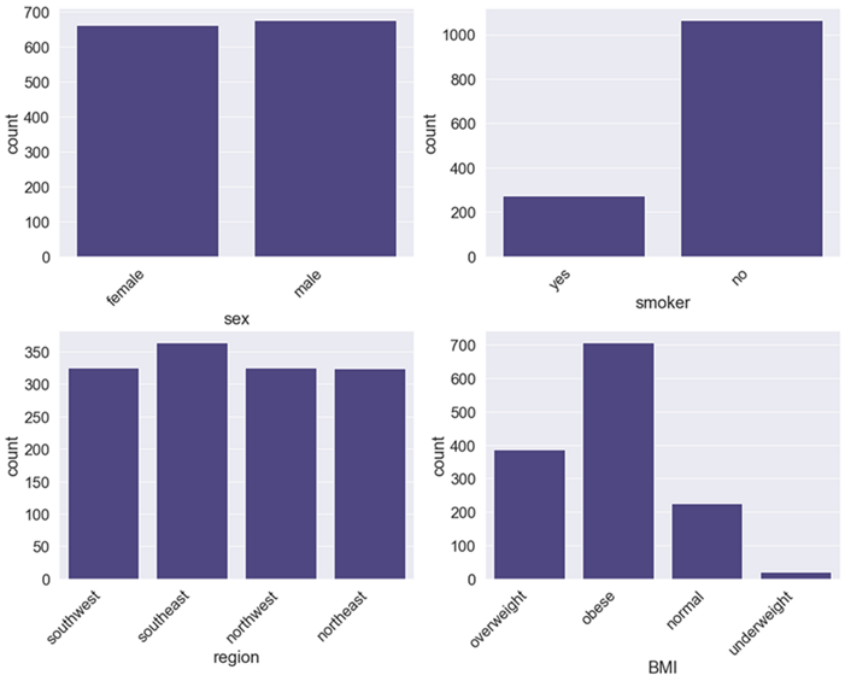


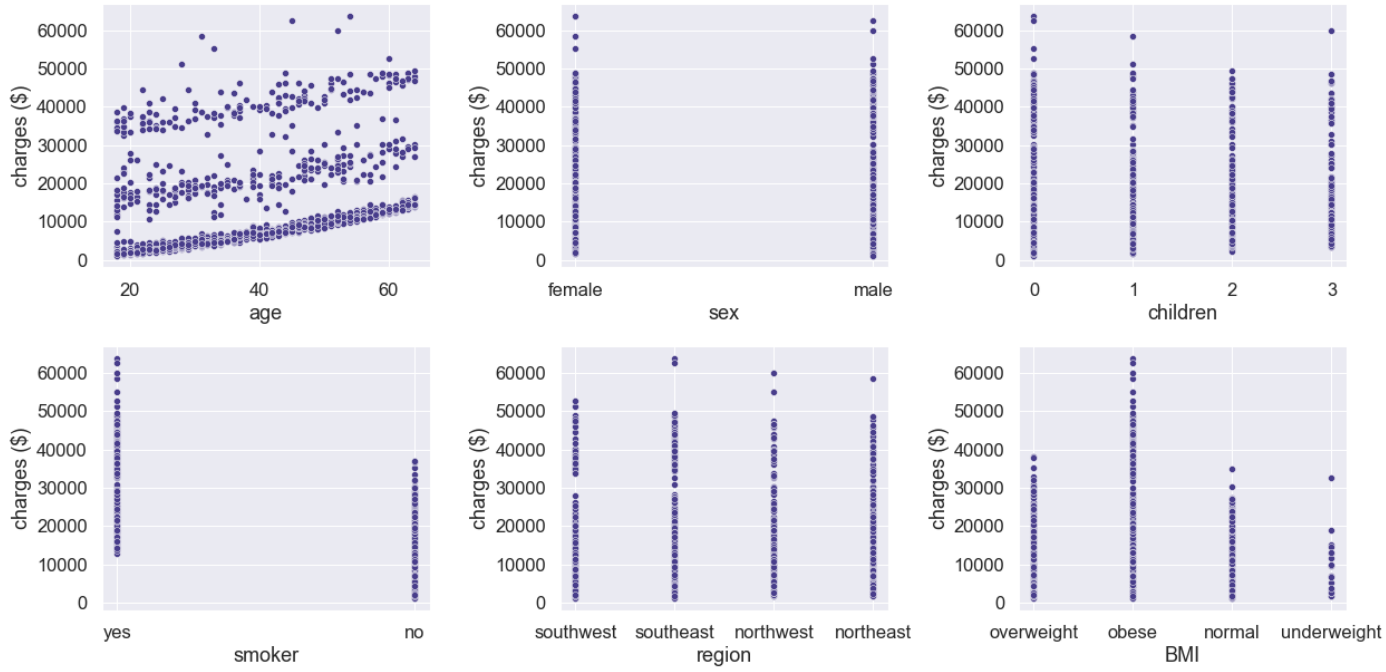
Figure 2: Bar plots of sex, smoker, region, and BMI



We also created scatterplots of the variables against charges, as seen in *Figure 3.* Based on the scatterplots, we observed there were three distinct groups of charges by age, with each group having slightly increasing charges as age increased. Charges appeared to be greater for policy

holders that smoke and have increased BMIs. The charges by gender, number of children, and region appeared to be consistent.

Figure 3: Scatterplots of all feature variables against the target variable charges



We created modified scatterplots by removed these three variables and included the smoker variable as a third variable, as seen in *Figure 4.* Based on these scatterplots, we observed that those who smoke are generally charged more, regardless of their age. We also observed that those who both smoke and have higher BMIs are charged more. We looked at BMI as it originally appeared in the dataset, as a continuous variable, and noticed a linear relationship between BMI and charges for smokers, as seen in *Figure 5*.

Figure 4: Modified 3-D scatterplots

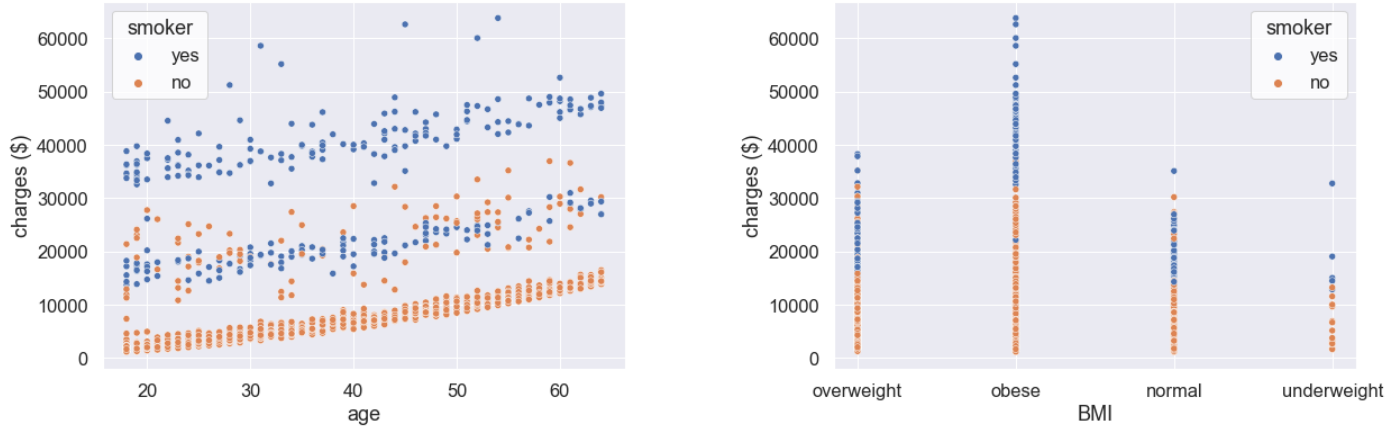
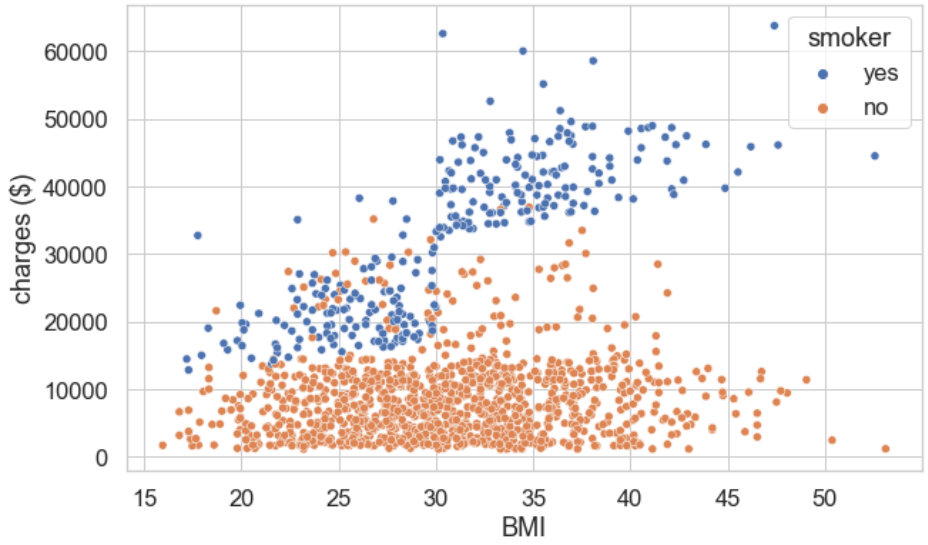


Figure 5: Scatterplot with BMI as a continuous variable



Looking at the data, it appeared the most significant variables contributing to health insurance charges were age, BMI, and whether the policy holder smoked or not. However, we used a decision tree regression model to determine feature importance. First, we split our data into training, validation, and test sets, with the training set containing 50% of the data and the remaining 50% split evenly between the validation and test sets. Next, we trained the decision tree regression model on our training data, with BMI in categorical form. The results were as follows:

Variable	smoker	BMI	age	children	region	sex
Significance	62.3%	16.5%	15.3%	2.9%	1.6%	1.4%

We repeated the process but with BMI as a continuous variable as it originally appeared in the data. The results were as follows:

Variable	smoker	BMI	age	children	region	sex
Significance	61.9%	20.9%	13.8%	1.6%	1.1%	0.8%

BMI as a continuous variable has more significance than as a categorical variable. It also lessens the significance of region, children, and sex. Therefore, we selected smoker, BMI as a continuous variable, and age as the variables to use in our regression models.

Using the selected variables – smoker, BMI, and age – we trained and fit linear regression, decision tree, and random forest models to the validation data. We decided to use the R-squared and mean absolute error metrics to evaluate the performance of our models. R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. The mean absolute error (MAE) measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. Using these metrics to evaluate the models, we obtained the following results:

Model	**R-squared**	**MAE**
Linear Regression	0.708	4163.34
Decision Tree	0.683	3239.07
Random Forest	0.78	2833.14

The random forest regression model performed better than the other two models. Next, we concatenated our training and validation data and performed a cross-validation grid search to choose the hyper-parameters for our random forest regression model. Using the grid search to find the values that resulted in the best mean absolute error score, the following parameters were selected:

- criterion = ‘mae’
- max_depth = 4
- max_features = None
- min_samples_leaf = 3
- min_samples_split = 4
- n_estimators = 200

RESULTS

Using these parameters, we trained the random forest with the training and validation data and fit it on our test data. We obtained the following results:

- R-squared: 0.86
- MAE: 2022.03

The results using the fine-tuned random forest shows increased performance in both performance metrics. The resulting R-squared value indicates that 86% of our data fit the random forest regression model. The mean absolute error indicts our prediction is off by $2022.03 on average.

CONCLUSION

Based on the regression models we used, we saw improved prediction when using the fine-tuned random forest model. However, there may be other models that increase prediction performance. Our analysis and predictions were limited to the variables in our dataset. There are many other factors that may contribute to the costs of health insurance, besides the personal attributes of the policy holder. Some of these factors may include, but are not limited to, the following:

- whether a spouse is on the plan
- whether the policy holder has secondary health insurance
- whether insurance coverage is provided by an employer, Medicare, Medicaid, Affordable Care Act, etc.
- which insurance company is providing coverage
- the type of insurance plan, such as high or low deductible
- whether those insured have chronic diseases or illnesses

REFERENCES

Bihari, Michael. (2020, December 8). “What Determines the Cost of a Health Insurance Plan?” Verywell Health. Retrieved February 14, 2021, from www.verywellhealth.com/cost-of-health-insurance-1738623.

Choi, M. (2018, February 21). Medical Cost Personal Datasets. Retrieved December 02, 2020, from https://www.kaggle.com/mirichoi0218/insurance

Defining adult overweight and obesity. (2020, September 17). Retrieved February 14, 2021, from https://www.cdc.gov/obesity/adult/defining.html

Fernando, Jason. (2020, November 18). “R-Squared.” Investopedia. Retrieved March 02, 2021, from www.investopedia.com/terms/r/r-squared.asp.

How does the size of my family impact my insurance cost? (n.d.). Retrieved February 14, 2021, from https://www.bcbsm.com/index/health-insurance-help/faqs/topics/buying-insurance/family-size-impact-cost.html

How health insurance marketplace® plans set your premiums. (n.d.). Retrieved February 14, 2021, from https://www.healthcare.gov/how-plans-set-your-premiums/

Understanding health Insurance Costs: Premiums, deductibles & more. (n.d.). Medical Mutual. Retrieved February 14, 2021, from https://www.medmutual.com/For-Individuals-and-Families/Health-Insurance-Education/Health-Insurance-Basics/Understanding-Costs.aspx