Bellevue University

Medical Cost Regression Analysis

Enkhchimeg Tsendnyam

List of Appendices

Background

Did you know the U.S. spends the most on health care out of all the other countries and spends twice as much as the other high-income countries? In 2019, the U.S. health care reached a staggering amount of $3.8 trillion. Divided by person, this cost equals to $11,582. Health insurance premiums and deductibles going up every year in the U.S. It is said that the national health care spending is estimated to reach $6 trillion by 2027. This is an issue that affects all of us as most of us who have health insurance pay for health insurance premiums every month. We are not even talking about the health insurance deductibles here which is the amount you pay out of pocket on top of the premiums. Did you also know that while the U.S. spends significant amount of money on health care, yet its citizens also have the worst health outcomes? The U.S. health care system is a very complex system and there is no easy answer to that. However, looking into how the health care system and insurance industry works in the U.S. will provide a glimpse to it.

The U.S. health care system relies on both public and private insurance. While the majority of people have private health insurance, primarily through an employer, many others obtain coverage through programs offered by state and federal governments. These state and federal government programs include Medicare, Medicaid and Civilian Health and Medical Program of the Department of Veterans Affairs (CHAMPVA) as well as care provided by the Department of Veterans Affairs and the military. Medicare is a federal program that helps to pay health care costs for people aged 65 and older and for certain people under age 65 with long-term disabilities. Medicaid is a government program for individuals with low income and resources. According to the 2021 U.S.

Department of Commerce report, in 2020, private health insurance coverage continued to be more prevalent than public coverage at 66.5 percent and 34.8 percent respectively out of all the insurers. As we can see most insurance is administered by private insurance companies. Despite that, individuals still pay for their health insurance, even if their employer subsidizes some of it. This is perhaps one of the reasons why health care is expensive in the U.S., though there are many other underlying reasons behind that. For example, according to the 2016, JAMA study findings, price of labor and goods including pharmaceuticals, and administrative costs appeared to be the major drivers of the overall cost between the U.S. and other high-income countries. This is true as doctors and specialist are paid more in the U.S. and administration costs are comprised of various activities from providers dealing with myriad regulations about usage, coding, and billing. Furthermore, because of the complexity of the U.S. health care system and the lack of any universal set prices for medical services, providers are free to charge what the market will bear. The amount paid for the same healthcare service can vary significantly depending on the payer whether it is through private insurance or through government programs, such as Medicare or Medicaid.

Business Problem

So, all this information brings up many more questions, but was curious to know how does U.S. insurance companies determine how much each everyone pays for insurance premiums? Could someone's smoking habit influence how much they pay for health care expenses? What about personal medical history such as person's age or body mass index (BMI)? Or what about where people live and or how many kids they have? To answer some of these questions, I worked on a predictive analysis data science project

using a medical care related dataset. Through this project, I looked at what's influencing how much people are paying for insurance premiums or medical expenses in general. Furthermore, I did a regression analysis and predicted how much insurance premiums will go up in the future. I used various machine learning regression methods and determined which method yields the most accurate results. The dataset that's being utilized for this project is "Medical Cost Personal Dataset" from Kaggle. This dataset was first published in data scientist Brett Lantz's 2013 book, *Machine Learning with R*. *Machine Learning with R* provides an introduction to machine learning using R and all of the datasets are currently in public domain. Medical Cost Personal Dataset comes in CVS format, and it provides insights into patients' personal health information as well as how much they got charged by their medical health insurance companies. This dataset has 1338 rows and 7 columns which are patients' age, sex, BMI, children, smoker, region, and charges.

```
#Display the data first 5 rows using head()
data.head(5)
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Some of the research questions that I explored and attempted to answer are:

- Does age have an impact on medical charges?
- Does BMI have an impact on medical charges?
- Does smoker status have an impact on medical charges? Compare medical charges for smokers vs. non-smokers
- Does gender have an impact on medical charges? Compare medical charges for male vs. female
- Does number of children that a patient has have an impact on medical charges?
- Does region have an impact on medical charges?
- Is there any correlation between age & BMI?
- Is there any correlation between BMI & smoker status?
- What impacts (which variable) medical charges to go up the highest?

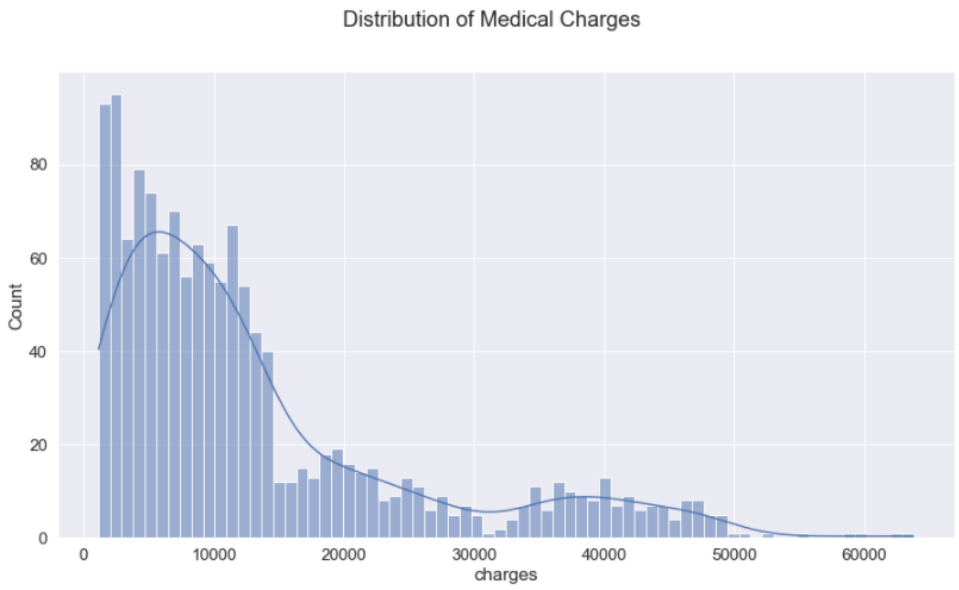
Assumptions

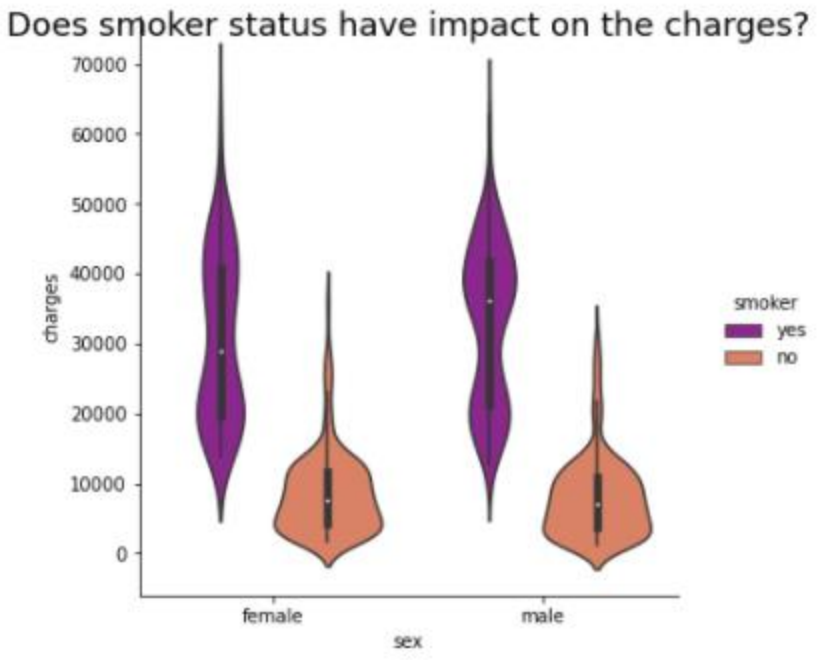
Since there are personal & patient history related variables included on this dataset, there are a few assumptions that can be made. Looking at the variables that are available in this dataset, some of my initial assumptions are:

- People who are older pay mode in medical expenses & people who are younger pay less in medical expenses
- People who have higher BMIs pay more in medical expenses
- People who smoke pay more in medical expenses

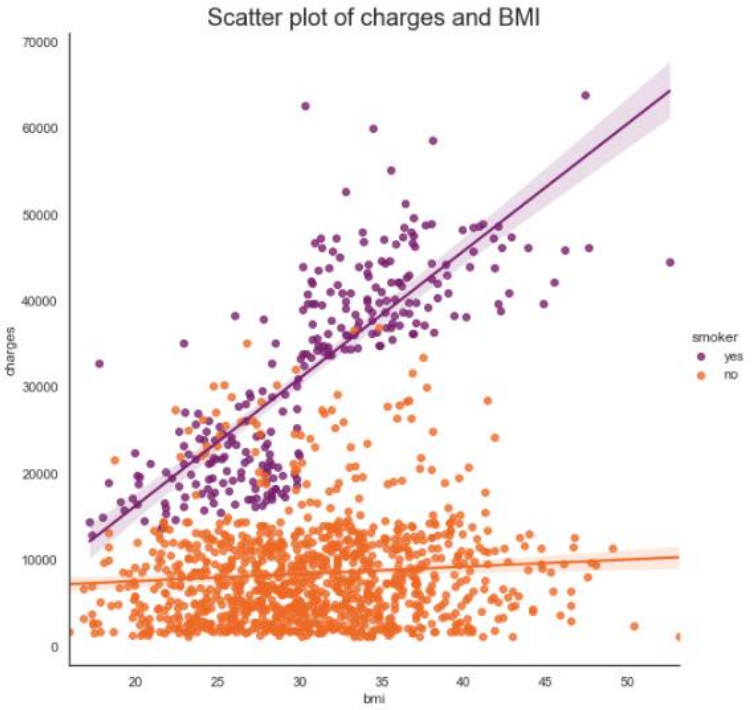
Data Explanation

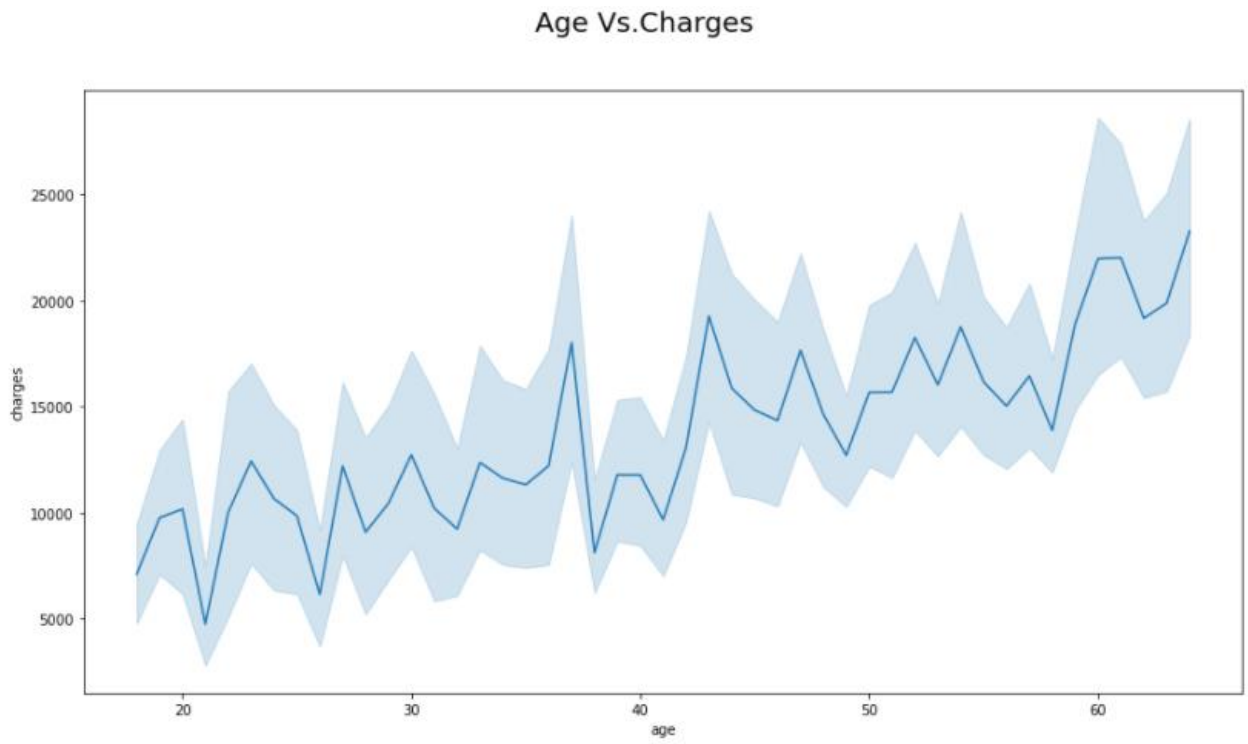
The data explanation methods for this project included doing statistical analysis as well as exploratory data analysis. Basic data describing methods revealed the type of data, count of unique values, mean, min, max and distributions. Through EDA, I was able to explore all the continuous and category variables' frequencies by visualizing subplots using both **Matplotlib** and **Seaborn** libraries. For example, below chart shows the frequency of the medical charges and as you can see there were high frequencies of charges under less than $10,000 for these patients. Furthermore, I was able to prove that there are some correlations between people' personal health attributes and medical charges through various graphs. For example, when it comes to differentiating the medical charges for

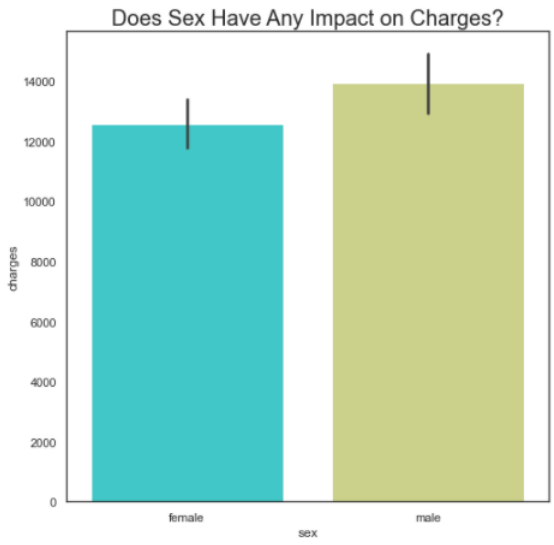


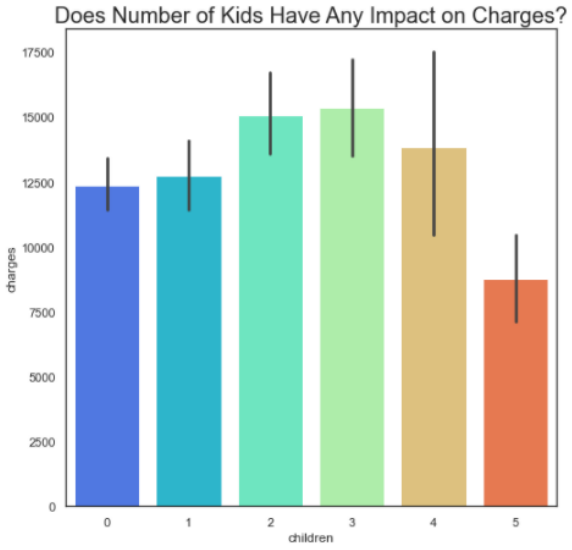


smokers vs. non-smokers, I used a violin chart. As shown on the above chart, we can clearly see that the medical charges were higher for smokers for both male and female patients. One interesting finding that was not on my original research questions, but I found are how people who smoke also tend to have higher BMIs. As you can see on the scatter plot on the right, if you are smoker and have a high BMI, your likelihood of high medical expenses is high. Similar to the smoker status and BMI index, I was able find some differences in medical charges based on people's sex, age as well as how many children they have.









As shown on the first chart, as people get older their medical expenses tend to go up. One of the fe surprising findings were males tend to have higher medical expenses compared to females and people who have larger number of kids, especially if you have 4 kids, the amount of money you spend on medical expenses are much higher than

people who have fewer kids or people who have 5 kids. On the other hand, when it comes to region, there were no significant geographical differences on how much people are paying for medical expenses.

Methods/Analysis

In order to get my dataset ready for predictive modeling, data cleaning and preprocessing steps were performed. There were no missing values in the dataset and there were some outliers in variables such as medical charges and BMI. My initial plan was to not remove any outlier since I already have a small dataset to start with. I encoded my categorical variables*: sex, smoker, and region* and split my dataset into training and test sets using 70:30 ratio. The target variable that is being predicted is medical *charges* a total of five different supervised regression machine learning techniques were applied. These include:

- ❖ Linear Regression
- ❖ Ridge Regression
- ❖ Random Forest Regression
- ❖ Ransac Regression
- ❖ Gradient Boosting Regressor

The training set was normalized and fit to each model and medical charges were predicted using each model.

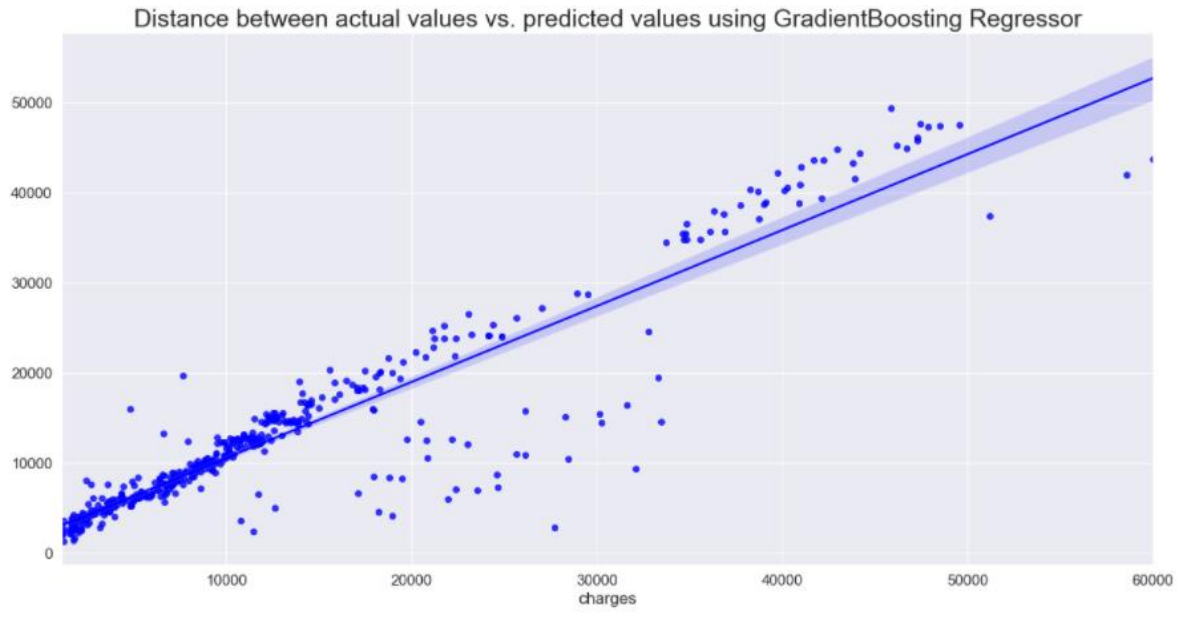
Model Evaluation/Conclusion

Model Evaluation metrics used for evaluating the results are R2, MSE, RSME, & MAE.

- R2 – coefficient of determination; proportion of the variance in the dependent variable that is predictable from the independent variables
- MSE – mean squared error; estimates the average squared difference between the estimated values and the actual value

- RSME – root mean squared error; measure of the differences between values predicted by the model
- MAE – mean absolute error; measure of error between paired observations expressing the same phenomenon

Using five different machine learning models resulted in five different R2, MSE, RSME, & MAE metrics. When it comes to determining how well the model fits the dependent variables, Random Forest Regression model excelled with the highest R2 of 97%, followed by Gradient Boosting Regressor with R2 of 90%. In other words, when Random Forest Regression model is deployed, 97% of the medical charges can be explained by the variations in the dependent variables such as age, sex, BMI, smoker status, location and number of children. However, when it comes to determining how close the forecasests are to actual values, Gradient Boosting Regressor did the best, with the lowest RSME of 4,302. Since RSME is an absolute measure of fit an describes how accurately the model predicts the medical charges, it is the most important criteria for this project.



Limitations/Challenges/Ethical Assessment

Even though I had high accuracy score, I wanted to see if I could improve the results. I was not able to improve the model evaluation metrics any higher than my initial baselines. I tried removing the outliers in variables such as *BMI* and medical charges or removing *region* variable, but it made no difference in terms of the model performance. Also, I had challenges finding background information on this dataset. The fact that I had no insights into when this data was collected, how it was collected and what these medical charges were related to made it difficult to tie the results back to the background of this paper, which was about finding out why U.S. health care system and insurance costs so high.

Future Uses/Additional Applications

I was able to do everything I planned to do on this project and find meaningful results. For small datasets like this one, I believe the regression and tree-based models were the most appropriate. In the future, I would like to further improve my analysis by checking on additional metrics such as residual errors. Also, when it comes to ethical aspects of how health care and insurance system operates in the U.S., other peer reviewed scientific research and assessments that should be explored and considered in the future.

Recommendations(Q&A)

After looking at the correlation matrix and various other graphs/chart, we can confidently say that person's BMI, smoker status, sex, age, number of children they have has an impact on their medical expenses.

- Does age have an impact on medical charges? **Yes**

- Does BMI have an impact on medical charges? **Yes, especially when it's combined with smoking**
- Does smoker status have an impact on medical charges? **Yes**
- Does gender have an impact on medical charges? **Yes**
- Does number of children that a patient has have an impact on medical charges? **Yes**
- Does region have an impact on medical charges? **No**
- Is there any correlation between age & BMI? **No, nothing conclusive**
- Is there any correlation between BMI & smoker status? **Yes**
- What impacts (which variable) medical charges to go up the highest? **Smoker status**

References

U.S. Census Bureau, Keisler-Starkey , K., & Bunch, L. N., Health Insurance Coverage in the United States: 20201–32 (2021).

Choi, M. (2018, February 21). Medical Cost Personal Datasets. Kaggle. Retrieved February 27, 2022, from https://www.kaggle.com/mirichoi0218/insurance

Team, T. I. (2022, February 8). *6 reasons healthcare is so expensive in the U.S.* Investopedia. Retrieved March 5, 2022, from https://www.investopedia.com/articles/personal-finance/080615/6-reasons-healthcare-so-expensive-us.asp

Why is healthcare so expensive in the United States? Consumer Watchdog. (n.d.). Retrieved March 5, 2022, from https://consumerwatchdog.org/news-story/why-healthcare-so-expensive-united-states

Irene Papanicolas, P. D. (2018, March 13). *Health Care Spending in the United States and other high-income countries*. JAMA. Retrieved March 5, 2022, from https://jamanetwork.com/journals/jama/article-abstract/2674671

Wikimedia Foundation. (2022, January 28). *Health Care Finance in the United States*. Wikipedia. Retrieved March 5, 2022, from https://en.wikipedia.org/wiki/Health_care_finance_in_the_United_States#cite_note-Census2016-2

Kurama, V. (n.d.). Regression in machine learning: What it is and examples of different models. Built In. Retrieved February 27, 2022, from https://builtin.com/data-science/regression-machine-learning

How to perform feature selection for regression data. Machine Learning Mastery. (2020, August 18). Retrieved February 27, 2022, from https://machinelearningmastery.com/feature-selection-for-regression-data/