CapstoneThree Final Report

HOW MUCH DO YOU HAVE TO PAY FOR HEALTH INSURANCE?

by Junko Takasawa

July 2021

In [2]:
```
%%html
<style>
table {float:left}
</style>
```

Problem Statement

USA TODAY wrote that according to the World Health Care Organization (WHO), health care costs are growing faster than the rest of the global economy.

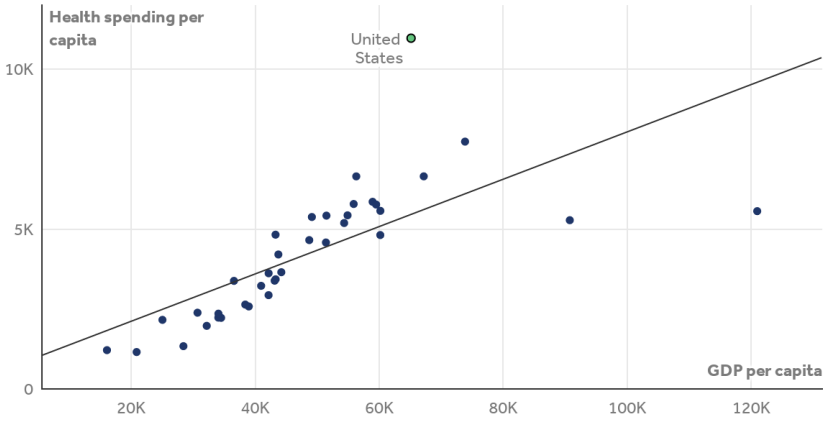
Every country has its unique political, economic, and social climate affecting its health care policies and spending. 24/7 Tempo reviewed health care expenditure data from the Organization for Economic Co-operation and Development (OECD), a group of 34 predominantly rich countries, because health spending is associated with a nation's wealth.

A third of OECD countries spend more than $2,000 per person each year on health care. The 12 countries with the highest health care costs, spend about twice that amount. The differences between countries is staggering, ranging from more than $8,000 per person in the country with the most expensive health care system to $541 in the OECD country with the lowest health care expenses per capita.

The United States has the highest healthcare expenditure per capita in the world. Wikipedia and other source show that it is more than $11,000 in 2019 (approx. $1,000/month) and is growing rapidly.

Rank	Country	2019 Expenditure/capita
1	USA	$11,072
2	Switzerland	$7,732
3	Norway	$6,647
4	Germany	$6,646
5	Austria	$5,851
6	Sweden	$5,782
7	Netherland	$5,765
8	Denmark	$5,568
9	Luxembourg	$5,558
10	Belgium	$5,428
11	Canada	$5,418
12	France	$5,376
13	Ireland	$5,276
14	Australia	$5,187
15	Japan	$4,823

GDP per capita and health consumption spending per capita, 2019 (U.S. dollars, PPP adjusted)



Notes: U.S. value obtained from National Health Expenditure data. Health consumption does not include investments in structures, equipment, or research.

Source: KFF analysis of OECD and National Health Expenditure (NHE) data

• PNG

Peterson-KFF Health System Tracker

HealthCare.gov lists some examples of how much certain health care may cost in the US :

- Fixing a broken leg can cost up to $7,500
- The average cost of a 3-day hospital stay is around $30,000
- Comprehensive cancer care can cost hundreds of thousands of dollars

This is the reason why having health insurance to protect you from high, unexpected costs like these is very important. It is even mandatory to have health insurance in some states, like California, Rode Island, Washington D.C., to name a few.

Unfortunately, health insurance is also very expensive. Thus it is very helpful to know how much you are likely to pay for insurance based on some fundamental features, such as age, gender, region you live in, and what factors are more likely to affect the insurance fee.

Data

Data source from kaggle : (https://www.kaggle.com (https://www.kaggle.com))
Links of datasheet : (https://www.kaggle.com/mirichoi0218/insurance/version/1 (https://www.kaggle.com/mirichoi0218/insurance/version/1))
Sampling methods : Random sampling

I downloaded the CSV file from kaggle, and imported it using pandas. This sample dataset contains 1338 rows of those insured with attributes of fundamental features.

Data Definition

- **age** - age of the insured
- **sex** - gender of the insured
- **bmi** - BMI (Body Mass Index) of the insured
- **children** - number of children of the insured
- **smoker** - smoking status of the insured
- **region** - region where insured lives in
- **charges** - annual insurance charge

Data Wrangling

The dataset is relatively clean, and there are no missing or undefined values in the dataset.

To make a few attributes more legible for data analysis, I created the following additional attributes from "charges", "age", and "bmi".

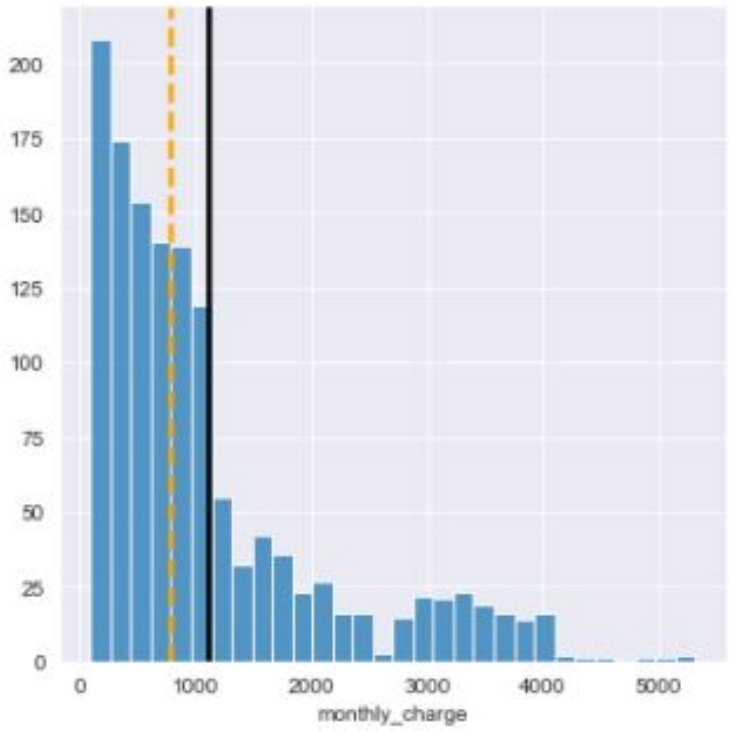
- **monthly_charge** : charges / 12 (months)
  * **charges** column is for annual premium charge. For analysis, monthly insurance charge is calculated.
- **age_group** : broke them to into groups of "under 20", "20's", "30's", "40's", "50's" and "over 60's".

- **weight_status** : grouped them into weight status "Underweight", "Normal", "Overweight" and "Obese" based on CDC Data for BMI standard chart for reference Data Source: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html (https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html).

BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Normal
25.0 – 29.9	Overweight
30.0 and Above	Obese

Exploratory Data Analysis

This sample dataset coinside with the average monthly health insurance fee found on various reliable sources referenced in the Problem Statement.

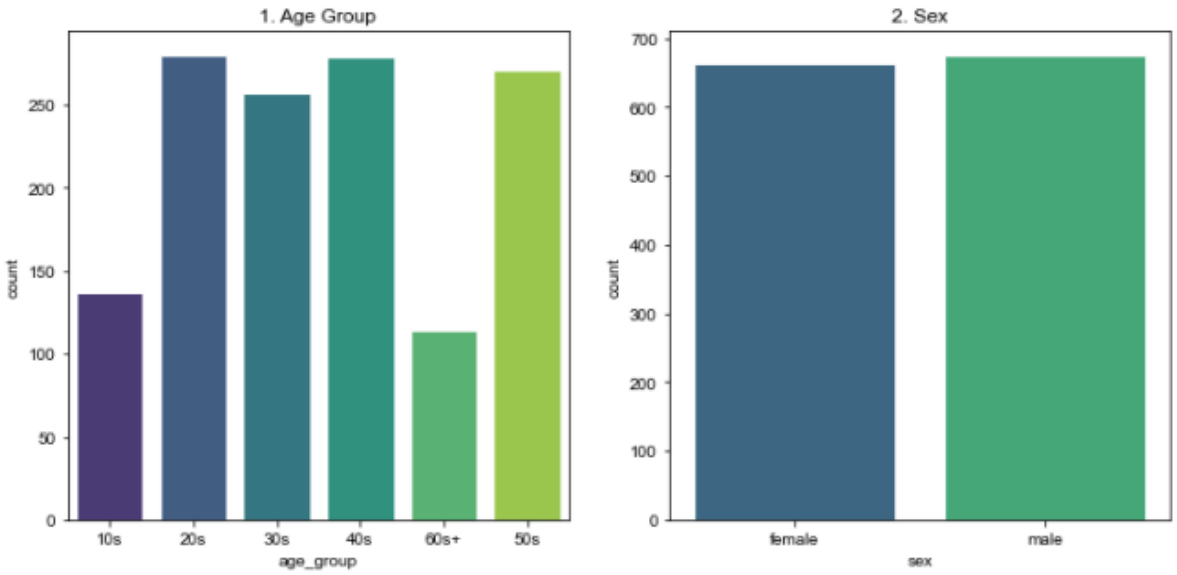


Average monthly fee (black line) : **$1,105.87**

Median (yellow line) : **$781.84**

Data Distribution

This data sample seems fairly well distributed among the age and gender groups.

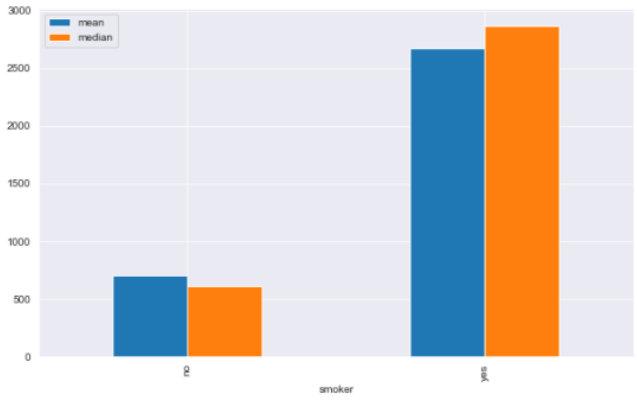


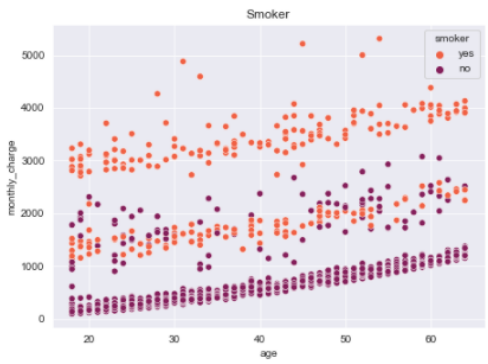
Except for 10's and 60's groups that are more likely not employed and are dependent member of other family members' health insurance policy, 20's, 30's, 40's, and 50's age groups all have around 250 sample data.

Gender groups seem equally distributed between male and female.

Feature - SMOKER

- EDA reveals that "smoker" status seems to have the most significant effect on the price of insurance.
- Left chart indicates that Mean (average) and Median of the insurance price for Smoker is almost five times higher than those who don't smoke.
- As shown on the chart on the right, a group of those who pay significantly higher insurance price in each group are dominantly smokers.



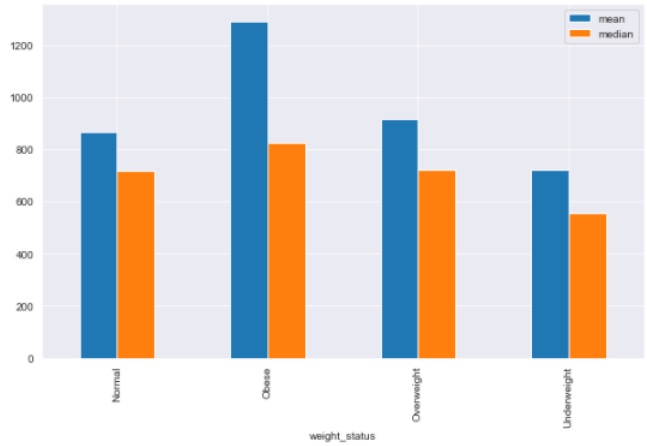


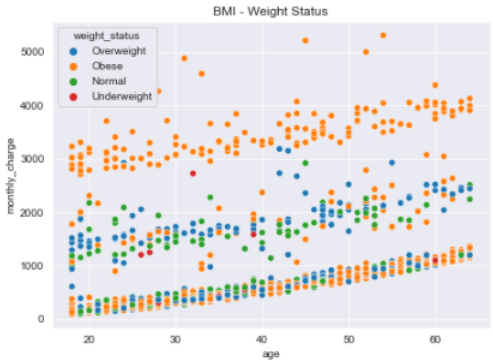
Health Insurance Monthly Fee by Smoking Status

Smoker?	Mean	Median
Yes	$2,670.85	$2,871.36
No	$702.85	$612.11

Feature - WEIGHT STATUS

- Second most important feature on the health insurance price seems to be the weight status.
- Left chart indicates that these weight status, "Underweight", "Normal", "Overweight", and "Obese", increases the health insurace price in that order.
- Just as what we saw for smoker, the chart on the right shows a group of those who pay significantly higher insurance price in each group are dominantly those in the "Obese" weight status.



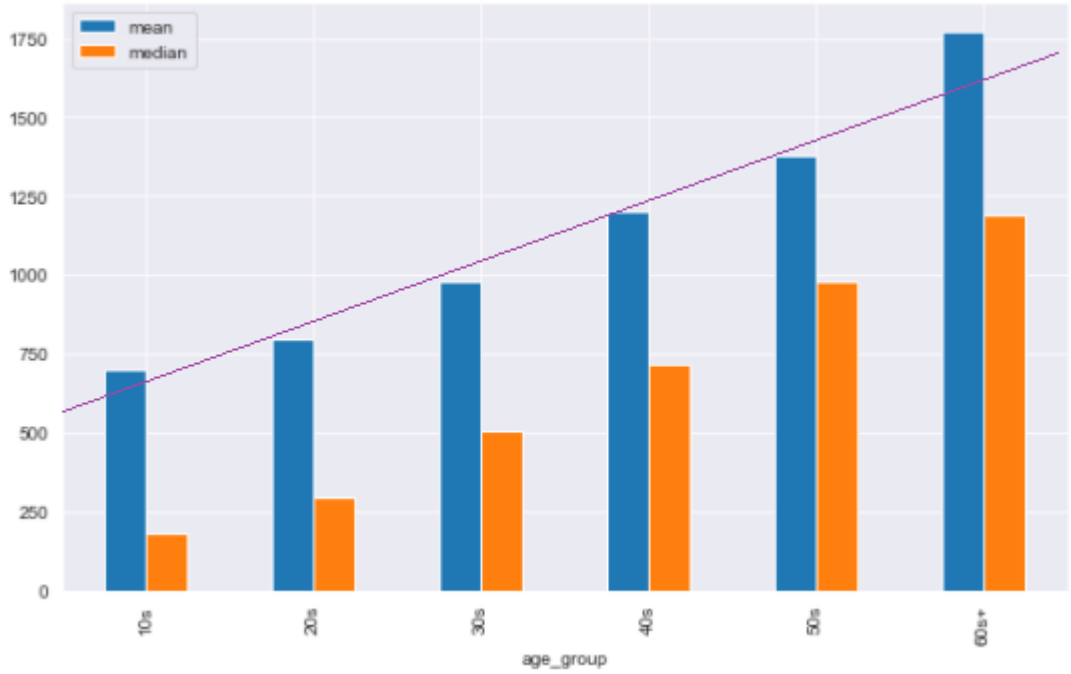


Health Insurance Monthly Fee by Weight Status

Weight Status	Mean	Median
Obese	$1,290.96	$824.60
Overweight	$917.23	$721.61
Normal	$867.07	$717.04
Underweight	$721.46	$553.38

Feature - AGE

- As expected, health insurance fee goes as the age goes up.
- Increase rate is relatively even.
- In this data sample, there is no insured over 70 years old. With that in mind, there seems to be a larger increase from 50's to 60's. However, this is somewhat expected.



Health Insurance Monthly Fee by Age Group

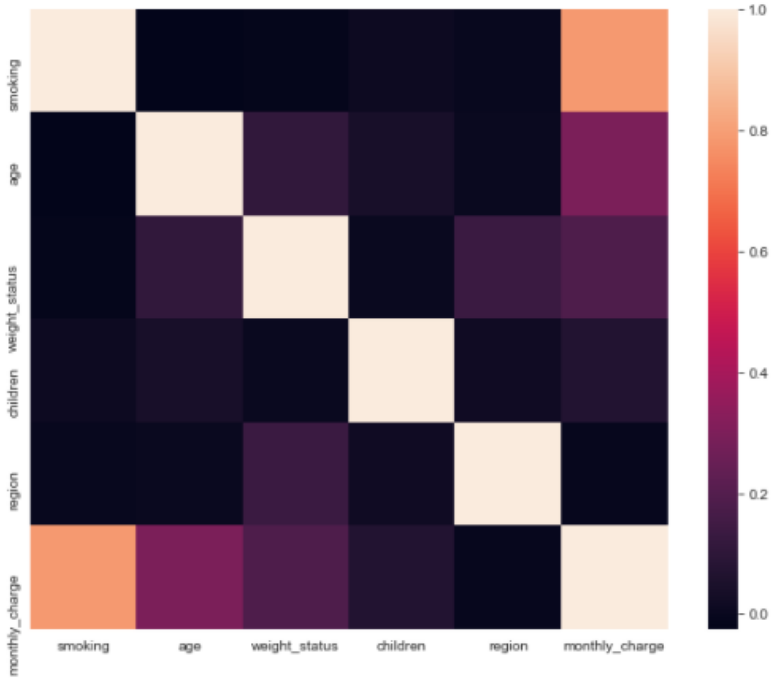
Age Group	Mean	Median
10's	$700.61	$178.17
20's	$796.81	$296.68
30's	$978.23	$506.87
40's	$1,199.93	$717.04
50's	$1,374.60	$977.470
60's	$1,770.66	$1,187.950

Feature - Others

- For other features like Region and Children, I did not see any noticeable trend or pattern.

Findings of which features influence the charge in this EDA can be confirmed with this correlation chart.
Most influencial features are "smoking", "age", and "weight status". "(number of) children" and "region" seem to have little to none impact on the charges.

Correlation of Features



Top 3 Feature Importance

1. smoking
2. age
3. weight status

Machine Learning - Predictive Models

Models Evaluated:

Since my goal is to predict health insurance fees depending on a few fundamental criterias, I chose regression models which describe the relationship between variables by fitting a line to the observed data. Regression allows us to estimate how a dependent variable changes as the independent variable(s) change.

1. Linear Regression
2. KNN Regression
3. Random Forest Regression
4. Gradient Boosting Regression

Data Pre-processing:

Split sample dataset into Train Data (80%) and Test Data (20%), then they are normalized for them to be applied to the models.

Data	Count	percentage(%)
Full data	1,388	100%
Train data	1,070	80%
Test data	268	20%

Normalized Features:

smoker - 0 : no (non-smoker), 1 : yes (smoker), 2 : no
gender - 0 : male, 1 : female

Fitting Data to Each Model:

I fitted Train Data and Test Data to each models, and projected the Accuracy Score and Mean Squared Error respectively.

1. Linear Regression

> Linear Regression is one of the most basic types of regression in machine learning. Linear regression comprises a predictor variable and a dependent variable related to each other in a linear fashion.

Accuracy Score and Mean Squared Error (MSE)

Accuracy	MSE
0.7999	221,027.18

2. KNN Regression

> KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood.

Accuracy Score and Mean Squared Error (MSE)

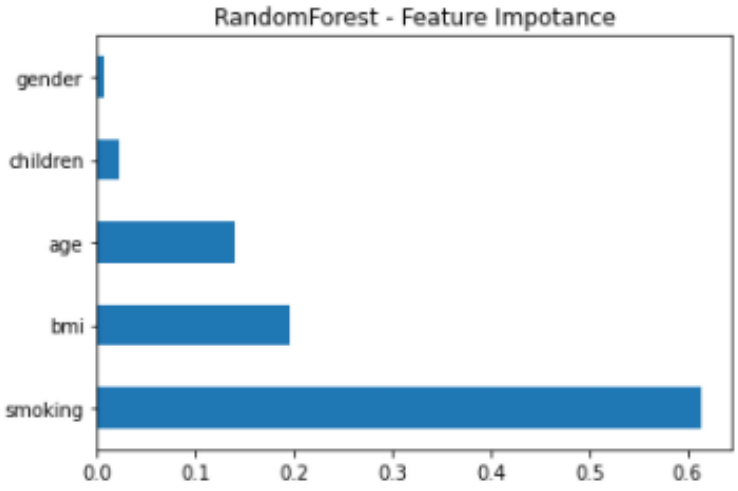
Accuracy	MSE
0.3162	755,535.86

3. Random Forest Regression

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Generally, Random Forests produce better results, work well on large datasets, and are able to work with missing data by creating estimates for them. However, they pose a major challenge that is that they can't extrapolate outside unseen data.

Accuracy Score and Mean Squared Error (MSE)

Accuracy	MSE
0.8911	120,261.95

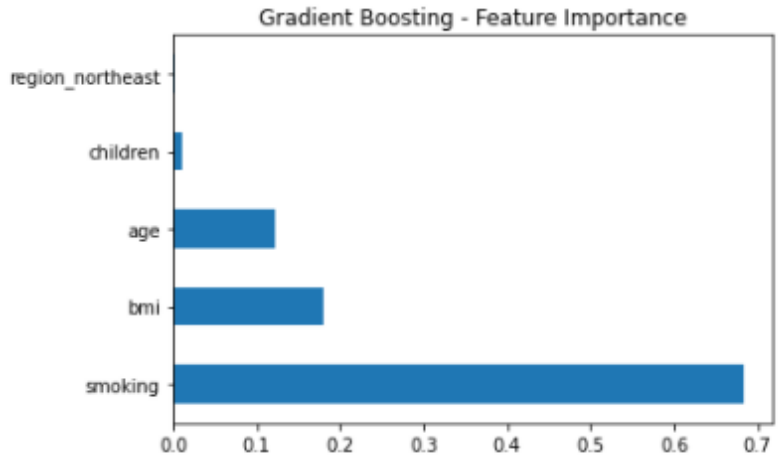


4. Gradient Boosting Regression

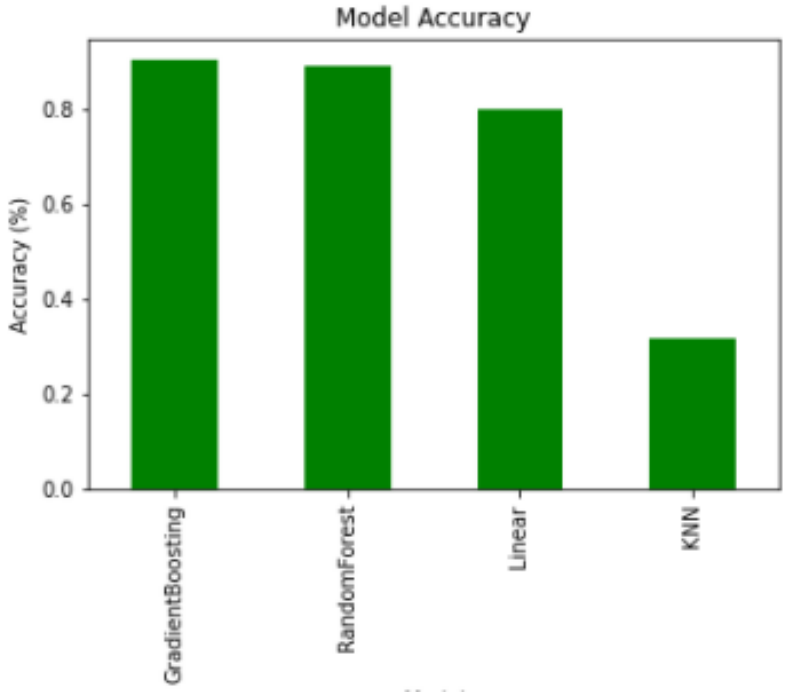
Gradient Boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

Accuracy Score and Mean Squared Error (MSE)

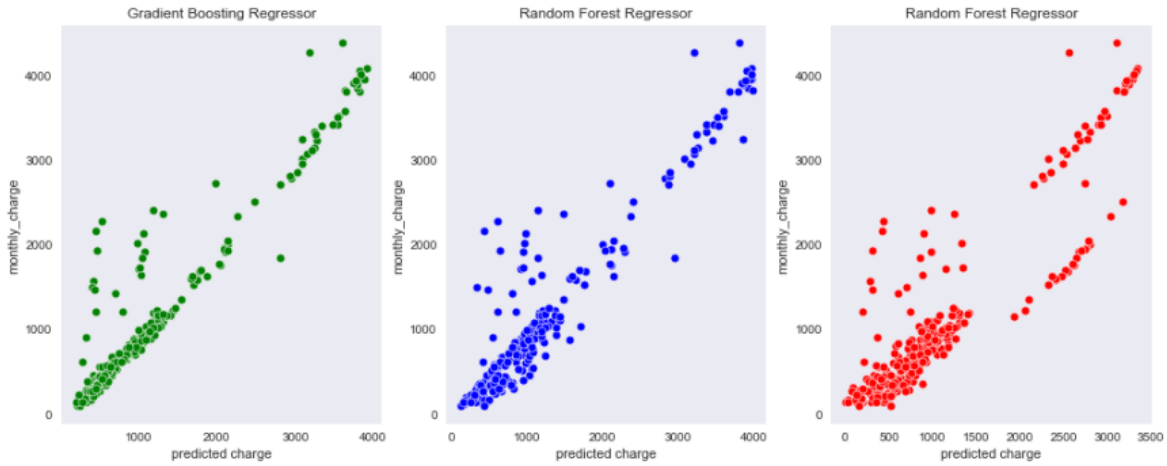
Accuracy	MSE
0.9013	109,038.20



Of all the machine learning models applied, Gradient Boosting is the model that can make the best prediction.



This can also be confirmed by the following charts that the actual and predicted charges with these models indicates that Gradient Boosting shows the strongest linear relationship, thus most accurate of all.



Recommendations

As explained in the problem statement, health care costs are growing faster than the rest of the global economy. And the United States has the highest health care expendiature per capita of all countries.

The top three factors that influcense health care cost are: **"Smoking"**, **"bmi (weight management)"**, and **"age"**.

Average monthly charge for smoker is almost four times more than that of non-smokers, and those who are in the **Obese** weight group have to pay approx. 35% more than those in the **Normal** weight group.

Smoker	Average
smoker	$2,670.85
non-smoker	$702.85

Weight Group	Average
Obese	$1,290.96
Overweight	$917.23
Normal	$867.07
Underweight	$721.47

Age Group	Average
10s	$700.61
20s	$796.81
30s	$978.23
40s	$1,199.93
50s	$1,374.60
60s	$1,770.67

Although we can not control our **age**, we can certainly try avoiding/refraining from two other big contributing factors, **smoking** and **weight management**.

People are well aware that smoking causes some types of cancer and lung related diseases. But CDC reports that smoking leads to disease and disability and harms nearly every organ of the body. More than 16 million Americans are living with a disease caused by smoking. For every person who dies because of smoking, at least 30 people live with a serious smoking-related illness. Smoking causes cancer, heart disease, stroke, lung diseases, diabetes, and chronic obstructive pulmonary disease (COPD), which includes emphysema and chronic bronchitis. Smoking also increases risk for tuberculosis, certain eye diseases, and problems of the immune system, including rheumatoid arthritis.

As for Obesity, a list of illnesses gets even longer, and it includes not only regular sickness, but also mental illness, low quality of life, and all-causes of death.

Almost everyone gets some kind of sickness at some point in his/her life, and there are sicknesses that we have little control of. However, we can definitely minimize the chance of getting sick by avoiding certain bad habits and maintaining the healthy life.

We may also get involved in accidents that will require medical attention. Therefore, what may even be most important is to have a good health insurance coverage.

Being healthy is one of the most fundamental and yet most important things in all of our lives, and avoiding such bad habits will lower our chance of getting sick, as well as lowering the healthcare charges to secure the medical attention when in need.

References:

Kaggle - Dataset https://www.kaggle.com/mirichoi0218/insurance (https://www.kaggle.com/mirichoi0218/insurance)

USA TODAY
https://www.usatoday.com/story/money/2019/04/11/countries-that-spend-the-most-on-public-health/39307147/ (https://www.usatoday.com/story/money/2019/04/11/countries-that-spend-the-most-on-public-health/39307147/)

Wikipedia
https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita (https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita)

The Peterson Center on Healthcare and KFF (Kaiser Family Foundation)
https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#item-spendingcomparison_gdp-per-capita-and-health-consumption-spending-per-capita-2019 (https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#item-spendingcomparison_gdp-per-capita-and-health-consumption-spending-per-capita-2019)

Statista
https://www.statista.com/statistics/184955/us-national-health-expenditures-per-capita-since-1960/ (https://www.statista.com/statistics/184955/us-national-health-expenditures-per-capita-since-1960/)

High Cost
https://www.healthcare.gov/why-coverage-is-important/protection-from-high-medical-costs/ (https://www.healthcare.gov/why-coverage-is-important/protection-from-high-medical-costs/)

CDC (Smoking)
https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm (https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm)

WebMD
https://www.webmd.com/diet/obesity/obesity-health-risks#1 (https://www.webmd.com/diet/obesity/obesity-health-risks#1)

CDC (Obesity)
https://www.cdc.gov/healthyweight/effects/index.html (https://www.cdc.gov/healthyweight/effects/index.html)

Special Thanks to my mentor at Springboard, David Lara Arango.