Vishal Kumar
Data Science Career Track
March 10, 2018

Table of Contents

1. Introduction

In the United States, health insurance marketplaces, also called health exchanges, are organizations in each state through which people can purchase health insurance. People can purchase health insurance that complies with the Patient Protection and Affordable Care Act (ACA, known colloquially as "**Obamacare**") at ACA health exchanges, where they can choose from a range of government-regulated and standardized health care plans offered by the insurers participating in the exchange.
ACA health exchanges were fully certified and operational by January 1, 2014, under federal law. First enrollment in the marketplaces started on October 1, 2013, and continued for six months.

Health insurance exchanges in the United States expand insurance coverage while allowing insurers to compete in cost-efficient ways and help them to comply with consumer protection laws. Exchanges are not themselves insurers, so they do not bear risk themselves, but they do determine which insurance companies participate in the exchange. An ideal exchange promotes insurance transparency and accountability, facilitates increased enrollment and delivery of subsidies, and helps spread risk to ensure that the costs associated with expensive medical treatments are shared more broadly across large groups of people, rather than spread across just a few beneficiaries. Health insurance exchanges use electronic data interchange (EDI) to transmit required information between the exchanges and carriers (trading partners).

1.1. Problem Definition

In this project, we have analyzed the data from the very first year of the health insurance marketplace inception, year 2014. We explored the data on how do plan rates and benefits vary across states, how do plan benefits relate to plan rates, how do plan rates vary by age, how do plans vary across insurance network providers? And any other related analysis on rate and benefits. We found some trends during the data exploration.

In the second part of this project (Machine Learning), we build a model that predicts the monthly premium rate of individuals and tobacco users. We compared the different regression model and found the best with higher prediction accuracy.

1.2. Potential Clients

This model can be used by anyone who wants to enroll in on the exchange and wants to know what would be his monthly premium based on coverage so that he/she can decide on the plans which would be most suitable for them. Any company which sells healthcare insurance and wants to know whether they can invest more in exchange or they should come out of the exchange based on the kind of environment and smog around this healthcare exchange program. For example the data shows that the last two year the number of companies in the exchange has dropped significantly around 45%. If this trend continues the marketplace would not survive and any company would not want to invest in the exchange.

This analysis can also be used by health insurance providers to improve their sales by comparing the similar plans with their competitors in particular state and specific areas in that states.

2. Data Acquisition

We acquired the datasets from two different sources, first from the kaggle.com which had datasets for the first three (2014, 2015 and 2016) years of data after the launch of health insurance marketplace and second from the National Bureau of Economic Research which has the last two (2017 and 2018) years of datasets. The links are below.

https://www.kaggle.com/hhs/health-insurance-marketplace
http://www.nber.org/data/cms-marketplace.html

CMS' Center for Consumer Information & Insurance Oversight produces the Health Insurance Marketplace Public Use Files (Marketplace PUF) to increase transparency in the Health Insurance Marketplace and to support benefit and rate analysis. The Marketplace, also known as the Affordable Insurance Exchange, is a key component of the Affordable Care Act (ACA). CMS describes the Marketplace PUF data in the following way:

The Marketplace PUF includes plan and issuer level information for certified Qualified Health Plans (QHPs) and stand-alone dental plans (SADPs) offered to individuals and small businesses through the Health Insurance Marketplace. The Marketplace PUF includes data from states participating in the Federally Facilitated Marketplaces (FFM), which include State Partnership Marketplaces (SPMs), and states whose State-based Marketplaces rely on the federal information technology platform for QHP eligibility and enrollment functionality. The Marketplace PUF does not contain any data on plans offered in states that established and operate their own Marketplace (State-based Marketplace) and do not rely on the federal platform for QHP eligibility and enrollment functionality, nor does it contain enrollment or claims data. Due to the limitations that this dataset does not contain any data about enrollment we cannot establish the relationship between the total number of insured and uninsured population state wise and in the entire USA. It would have been very interesting to verify the claims that after the establishment of marketplace or exchange the number of uninsured population has been decreased. We could have measured the number and percent and how the trend is over the years.

3. Data Wrangling and Cleaning

The downloaded data were not in the format best suited for our analysis. We needed to do some data cleaning and wrangling to bring the data to the desired format. First of all we had to merge all the datasets we downloaded from the two sources. First the dataframes were merged for the year 2017 and 2018 and then this new dataframes was merged with the dataset of first three years. After combining these two datasets, we had removed some columns which were not required for this analysis. Following are some common data problems we had, dealing with datasets.

- Missing data (NaN values)

- Outliers
- Duplicate rows

To check for missing data points we have used the following method.

1) Read the csv files into pandas dataframe.
2) Use **dataframe.info()** method to find if there is any missing data. In some cases the **dataframe.info()** method doesn't show the desire output for these kind of dataframes, we have also used **isnull()** method to find if there is any missing values in the columns.

We have used following options to deal with the missing values:

1. **fillna()** method: There are lot of options available with the fillna() method to deal with the missing values. For example:

   fillna(0) will replace all the NaN values with 0(zero)

2. **dropna()** method: to drop the entire data point if there are missing values in all the columns.

There are columns in the data frames which has floating numbers representing the amount with "$" sign as prefix. For our analysis we had to change the column as numeric so that we can perform mathematical calculation on the columns. To convert these kind of values to numerical values we had to get rid of the "$" sign. To do this we have used ***replace*** function on data frame column name and then pandas ***to_numeric*** function to convert the object datatype into numeric datatype. Similar kind of conversion has been done on Age column of the rate data frame, which has the values as object datatype and with the text values as "0-20" where we have to get rid of "-" and covert it to numerical values. Also RatingAreaId column has the similar values, like "Rating Area 10", in these types of values we had to get rid of space and then convert these areas into numbers. We needed to perform this data cleaning on these columns as these were part of prediction model using regression.

4. Data Exploration

Since the inception of health insurance marketplace in year 2014, Americans were offered total of approximately **7.2M** in last 5 years. Out of which **964** benefits were unique. Interestingly not all the 50 states have participated in the federal run Marketplace or Exchange. For example, California, New York.

States who are not participating in the federal run marketplace have their own state run exchange. There are actually only 39 states which are using federal run health insurance marketplace. Idaho moved to state run marketplace after the first year of enrollment. And state run marketplace of Kentucky was dismantled in year 2016 and joined federal run exchange. There are/were 40 states associated with federal run exchange since its inception.

The top benefits offered out of 7.2M benefits to the Americans year wise given in the below table.

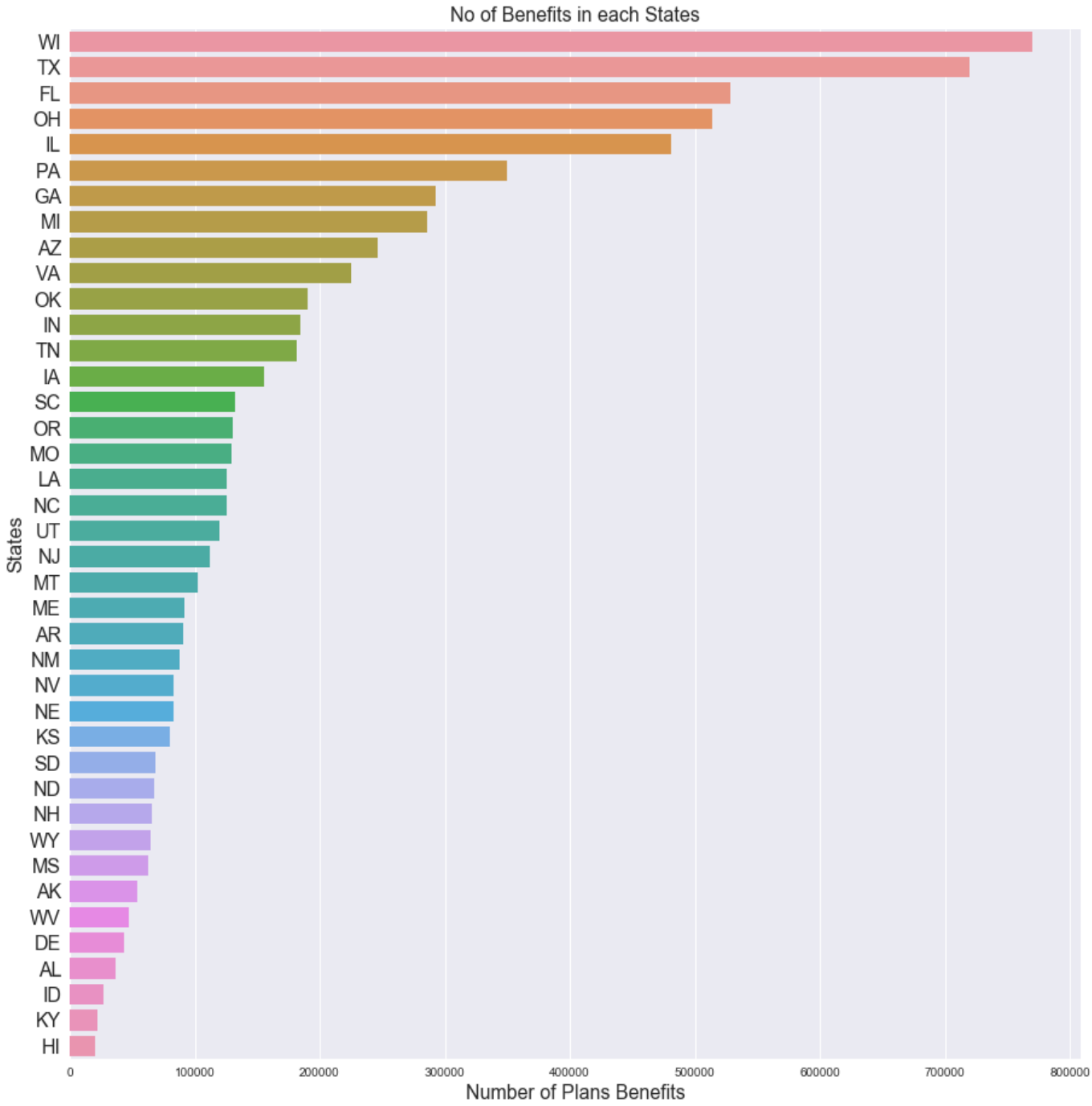
Table: Yearly - Top Benefits Offered

Business Year	Count	Unique	Top Benefit Name	Freq
2014	1164869	496	Major Dental Care - Adult	18719
2015	2079286	517	Orthodontia - Adult	31269
2016	1804253	429	Orthodontia - Adult	27389
2017	1324275	281	Basic Dental Care - Adult	21371
2018	829688	259	Orthodontia - Adult	13861

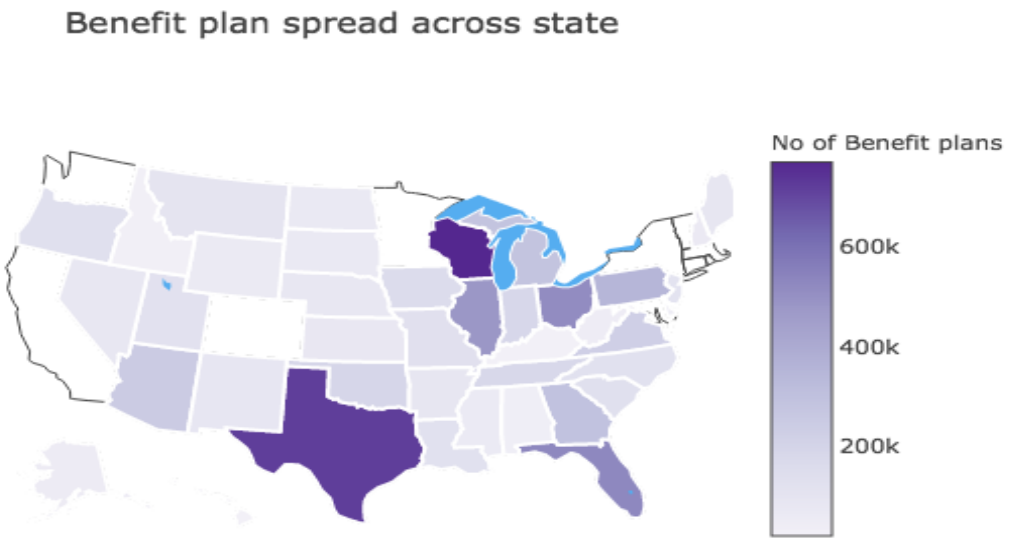
As we can see "Orthodontia - Adult" benefit is the most offered benefits to the American population in health insurance marketplace.

4.1 Plans Spread Across The States Since The Launch Of The Exchange

Now, let's see how these 7.2M benefits were offered across all the states since the inception of exchange.
Wisconsin, Texas, Florida, Ohio and Illinois are the top 5 states where most number of benefits is available to be chosen. Hawaii, Kentucky and Idaho are states sit at the bottom of the list. One reason for bottom three states is that, these states are not part of the exchange for all 5 years but either joined later or opted out after the initial years of marketplace inception.


No of Benefits in each States
WI
TX
FL
OH
IL
PA
GA
MI
AZ
VA
OK
IN
TN
IA
SC
OR
MO
LA
NC
UT
NJ
MT
ME
AR
NM
NV
NE
KS
SD
ND
NH
WY
MS
AK
WV
DE
AL
ID
KY
HI
States
0
100000
200000
300000
400000
500000
600000
700000
800000
Number of Plans Benefits

Let's put it on the USA map for better visual of the spread.



The darker the color, the more is the number of benefits available to that states. Complete white states are the states which are not participating in the federal run health insurance marketplace.

4.2 Individual Rates in Last 5 Years

Monthly premium for the health insurance plan is one of the biggest factors in choosing the right plan for the individual or family. Below is the list of 5 cheapest and costliest states in the health insurance marketplace. It means that the people of different states pay different premium for the similar or same plan.

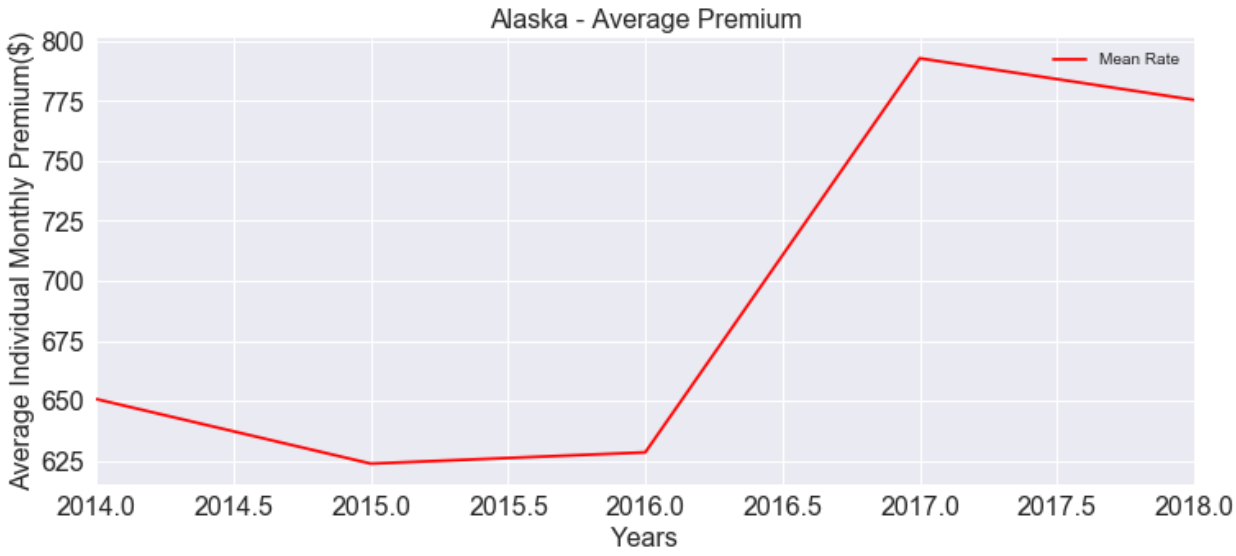
Monthly Average Rate(Individual)

5 Cheapest states			5 Costliest states	
State Code	Mean Rate		State Code	Mean Rate
MO	167.832877		AK	684.639431
MI	210.099972		WI	490.379170
TX	215.492503		IL	426.924091
AR	230.654717		NJ	415.198078
MS	247.00		SD	388.626110

Montana (MO) is the cheapest states in the United States in terms of paying monthly individual health insurance premium. On average people of Montana pay **$167.83** per month for individuals and people of Alaska (AK) pay whopping **$684.64** for the same or similar plan, which is more than **400%**. If we compare Montana with Wisconsin (WI), it is approximately **290%** more. The United States has the world's highest health care costs, and it sure looks like Alaska has the highest health care costs in the U.S. And the story is the same regarding health insurance, as the premiums for Alaskans on the exchanges for 2017 are at the top among the states. For further details on Alaska's high healthcare cost, please refer to the link below. The article talks about the high cost, consequences and remedies.

http://akcommonground.org/high-health-care-costs-in-alaska-facts-causes-consequences-and-remedies/

The graph below shows the average monthly premium of each year since the inception of health insurance marketplace in year 2014.



4.3 Plan Benefits and Rates

Plans in the Health Insurance Marketplace are presented in 4 "metal" categories: Bronze, Silver, Gold, and Platinum. ("Catastrophic" plans are also available to some people). The below table represent the total cost sharing in each category when we take any medical service.

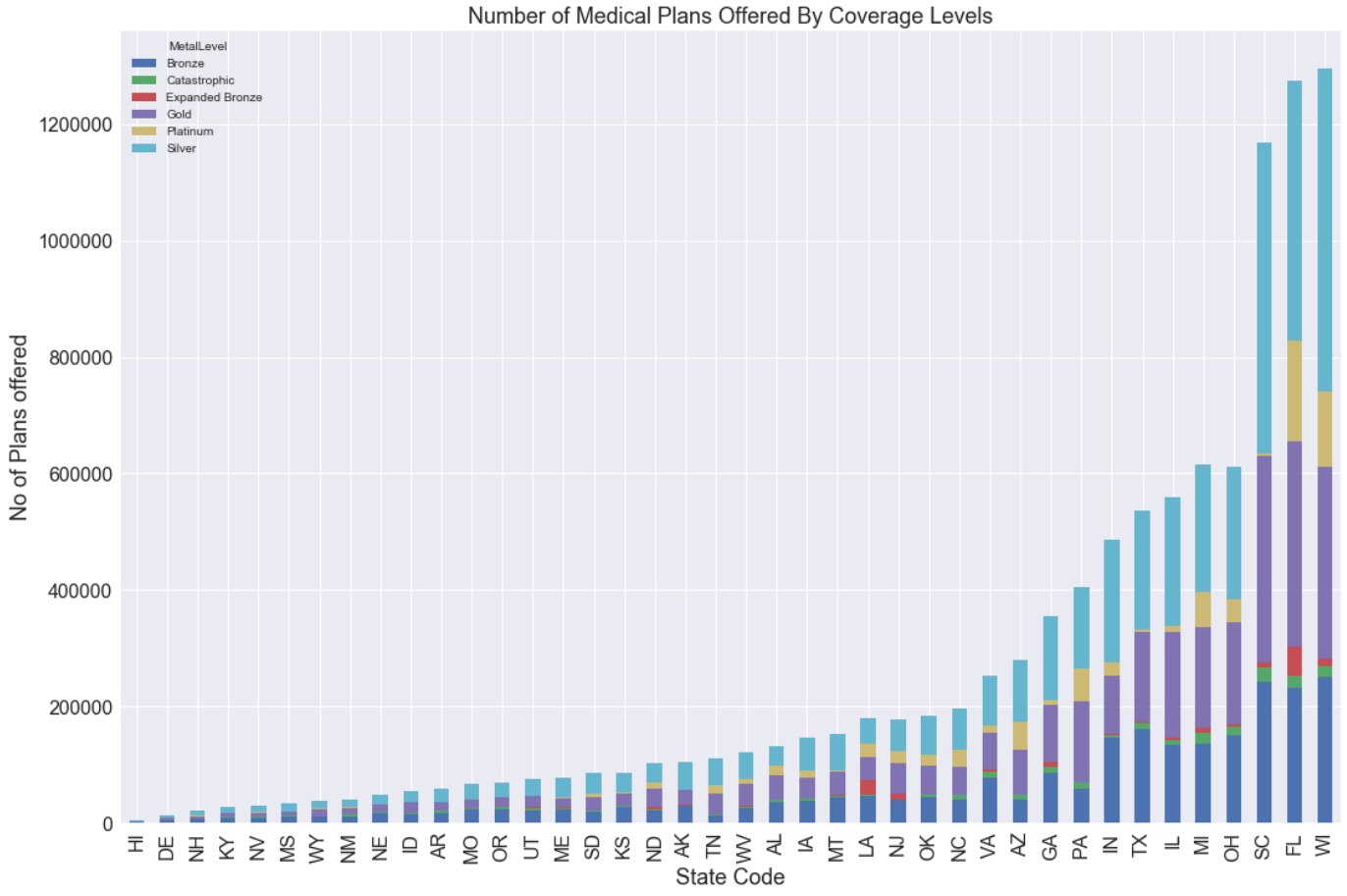
Plan Category	The insurance company	You pay
Bronze	60%	40%
Silver	70%	30%
Gold	80%	20%
Platinum	90%	10%

Catastrophic health plans: Catastrophic health insurance plans have low monthly premiums and very high deductibles. They may be an affordable way to protect you from worst-case scenarios, like getting seriously sick or injured. But you pay most routine medical expenses yourself. For more details on it and other metal level please refer to the below link.

https://www.healthcare.gov/choose-a-plan/plans-categories/

4.3.1 Medical Plans Offered By Coverage Levels

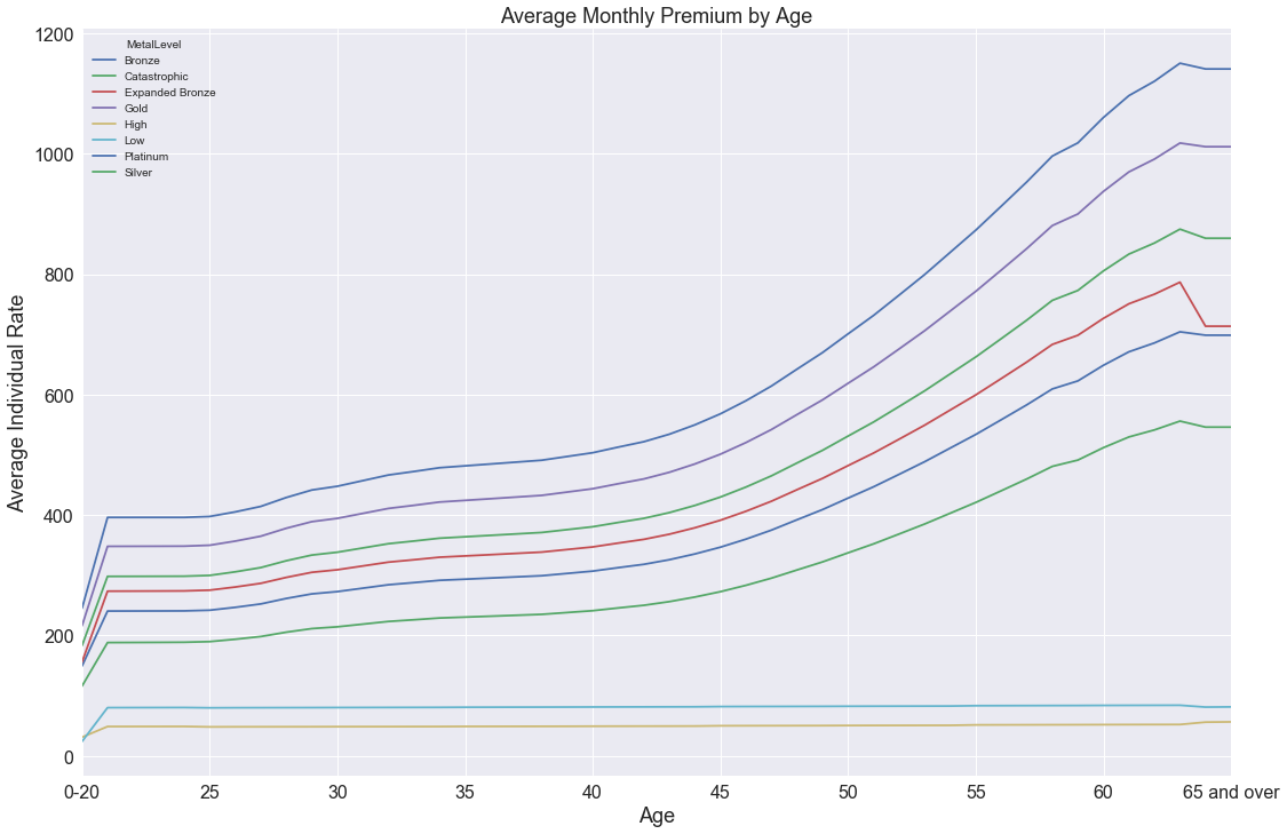
Silver plan is the most offered and popular plan compared to other metal plan. Wisconsin (WI) is the state which offers more number of plans than any other participating state. Florida (FL) is very close to WI. Three states (South Carolina, Florida, and Wisconsin) offered more than 1 million plans to its population since the inception of ACA or Obamacare.



We do not find Texas in this list of top 5 because states who offers more number of plans because lot of plans offered in Texas are dental plans and we have not taken dental plan into consideration for this coverage level analysis.

4.3.2 Average Health Insurance Cost by Age

While a large portion of our health insurance costs will depend upon the amount of coverage our plan provides (Bronze, Silver, Gold or Platinum) as well as where we live, we can examine how our age affects health insurance costs by looking at how the premiums for a single plan scale based on age. Fortunately for consumers, the federal regulations regarding individual health insurance set guidelines on how ACA compliant plans can adjust their rates based on the age of the policyholder.

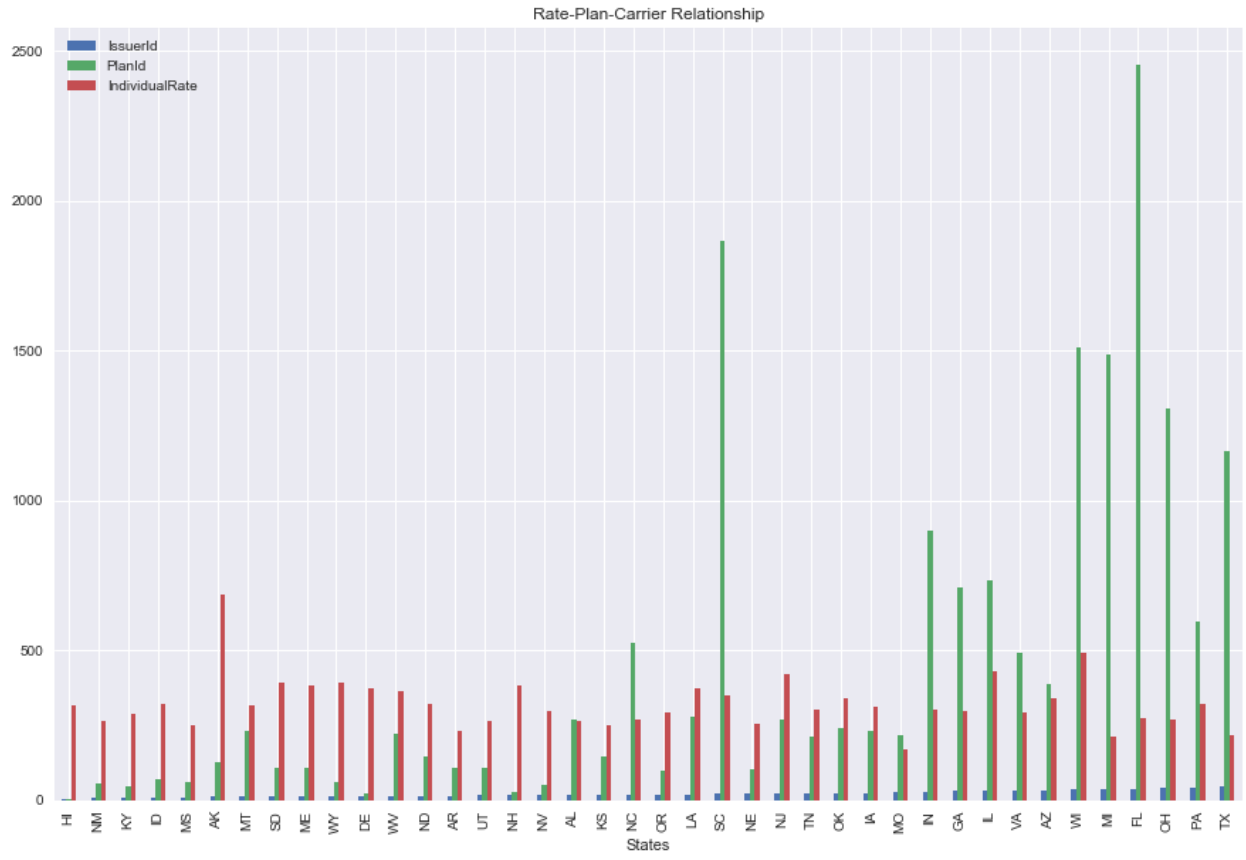


In most states, the base rate for a plan is calculated using a 21 year old policyholder in mind. This rate is then adjusted according to the age of the consumer. States using the federal age and premium guidelines will see anyone under the age of 21 all treated equally, with monthly costs coming out to a little more than 63% of the base rate for a 21 year old. Health Insurance rates go up as a policyholder gets older, with the largest increases coming after age 50. This reflects the higher expected share of health care costs that older Americans are expected to utilize.

At the high end of the age range, those consumers 64 and older have their premiums capped at 3 times the premiums of the 21 year old base rate. As you can see from the data, the largest changes in health insurance costs take place between the ages of 50-60, where premiums go from 1.78x the 21 year old rate to 3x by age 64.

4.3.3 Carrier, Plans And Rates

When the number of issuer (companies which provides health insurance), is less then premium is high and where the number of issuer is high the premium is low. This way it provides more competitive rates to the consumers, else there will be monopoly of fewer issuers and the consumer will have to pay more to the insurance company.
The below group chart is the visualization of the same theory.



In most of the states the above theory is true. Even if we do not consider Alaska as there are other extreme factors which contributes to the high monthly premium. We had expected that the red bar on the right side will be taller than the left side as the number of issuers are ascending from left to right on the above group chart. But this is not the case. Also we tried to find if there is any relation between the number of plans offered and the number of issuers. We would expect that when the number of issuers is more, then the offered plans will also be high but again from the chart above this is not conclusive. Further investigation and hypothesis can be done to dig more about the relationship between carrier, plans and rates.

5. Predictive Model

We will build this model to predict **Individual Rates** and **Individual Tobacco Rates** using regression models and will compare results and performance of these models. Before we start modeling let's first outline some important aspect of the predicting individual and individual tobacco monthly premium rate.

How premiums are set

Under the health care law (ACA), insurance companies can account for only 5 things when setting premiums.

1. Age: Premiums can be up to 3 times higher for older people than for younger ones.
2. Location: Where you live has a big effect on your premiums. Differences in competition, state and local rules, and cost of living account for this.
3. Tobacco use: Insurers can charge tobacco users up to 50% more than those who don't use tobacco.
4. Individual vs. family enrollment: Insurers can charge more for a plan that also covers a spouse and/or dependents.
5. Plan category: There are five plan categories – Bronze, Silver, Gold, Platinum, and Catastrophic. The categories are based on how you and the plan share costs. Bronze plans usually have lower monthly premiums and higher out-of-pocket costs when you get care. Platinum plans usually have the highest premiums and lowest out-of-pocket costs.
For more details on this please refer to the link below.
https://www.healthcare.gov/how-plans-set-your-premiums/

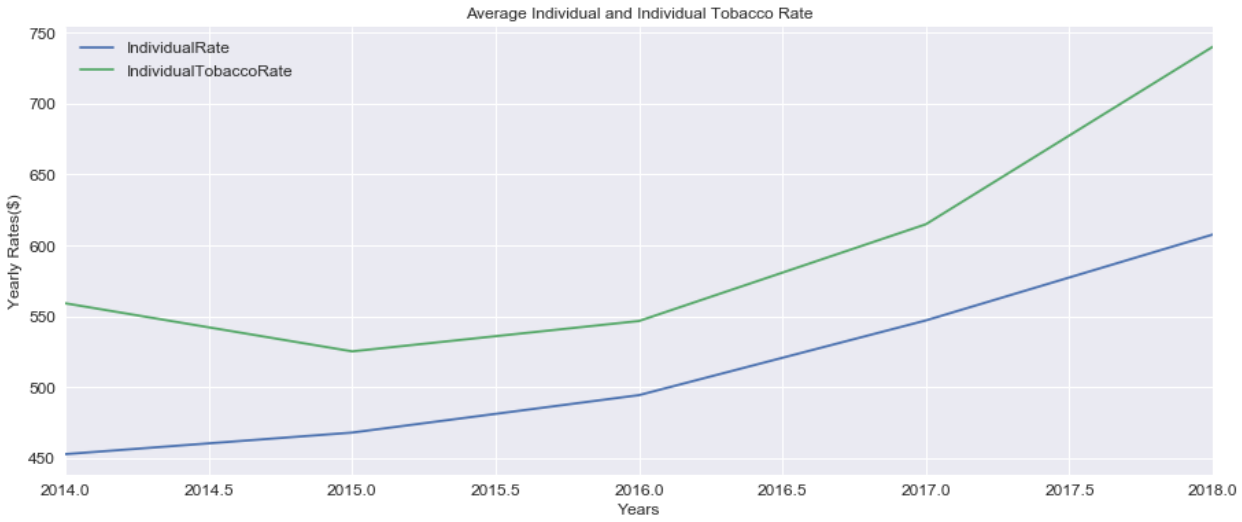
5.1 Data Pre-Processing

Before we start feeding the data into the machine learning algorithm, we need to bring the data into desire form. This includes merging of the two datasets which has rates and plans attributes so that we get the plan category field which is one of the independent variable for the prediction. We had also removed the two dental plans (Low and High) from the data set. We also had to change the data types of all the predictor variables. To change the metal plans values into numerical values we have used **OneHotEncoder.** To achieve this we have used pandas ***get_dummies()*** function.

Then we split the dataset into training and test datasets. I have split the data as per the years, first four years from 2014 to 2017 as training dataset and the year 2018 dataset se the target dataset.

5.2 Regression Model

Let's have a look at the two target variables, i.e. **Individual** and **Individual Tobacco** rates, how the average monthly rates have been doing since the inception of health insurance marketplace. Below line graph shows the same.



As we can see, in year 2015 the average monthly premium of the individual tobacco rate decreased and after that each year it has only increased. Total increase in both the rates is about 30% since the inception of the health insurance marketplace. There is a big increase in both the rates in year 2017 and then 2018. These changes are due to the fact that the new administration wants to repeal "Obamacare" and hence the insurance companies are not very confident and want to keep themselves as far as possible with the exchange.

5.2.1 OLS (Ordinary Least Square Method)

In statistics, **ordinary least squares (OLS)** or linear least squares is a method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being predicted) in the given dataset and those predicted by the linear function. Geometrically this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression line – the smaller the differences, the better the model fits the data. The resulting estimator can be expressed by a simple formula, especially in the case of a single regressor on the right-hand side.

The OLS estimator is consistent when the regressors are exogenous, and optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated. Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation

when the errors have finite variances. Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator.

First we had implemented the OLS method to predict the individual monthly rate. The result for individuals and individual tobacco users are **64%** and **62.24%** respectively on training dataset. When compared with the true value of the test dataset the results were very disappointing at the about **25%** and **27%** approximately.

5.2.2 skit-learn Linear Regressor - LinearRegression()

This is the linear model with the "sklearn" package in python for data science. The principle algorithm is same as the OLS method and the result were also exactly same for both individual and individual tobacco rate. So let's move on to the next model.

5.2.3 Decision Tree - Individual Rate

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called **classification trees**; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called **regression trees**. We have used this regression tree in our prediction model.

We have used ***DecisionTreeRegressor()*** to predict the IndividualRate and IndividualTobaccoRate and will compare with the Linear model to check if the model improves in predicting the monthly rates. Below is the table of predictor variable's importance and the score on training dataset.

Predictor Variable Importance	
Predictors	**Importance**
Age	0.827148
RatingAreaId	0.012227
Tobacco	0.000346
MetalLevel_Bronze	0.066977
MetalLevel_Catastrophic	0.036383
MetalLevel_ExpandedBronze	0.000357
MetalLevel_Gold	0.00553
MetalLevel_Platinum	0.008546
MetalLevel_Silver	0.042487
Total Score	**0.730541028**

Note: The Total score at the bottom of the table is not the sum of all the variables above, rather the score is the model accuracy on training dataset which is slightly more than **73%** which is better than **64%** in OLS method.

We have to check the score on data which is not seen by the model i.e. test data.

Score on test data: 0.256188218491

The accuracy is **25.61%** on the test data, which is similar to the linear model. On unseen data performance of the model is very poor.

Cross Validation

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.
In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. So for k-fold we will have k scores and then we take average of those score to measure the accuracy.

We implemented 5 fold cross validation on the training dataset to train the model in much better way for Individual Rate.

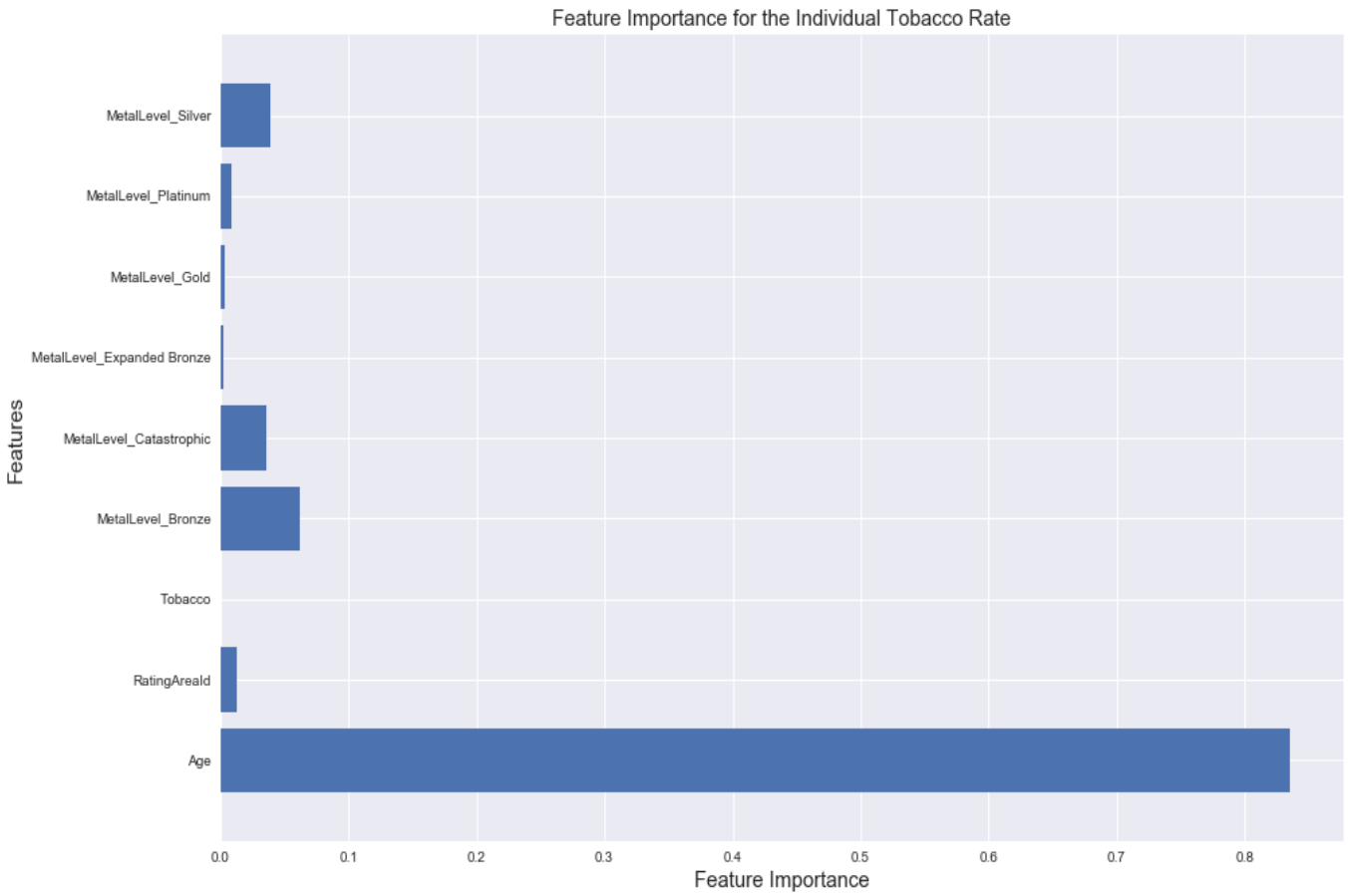
```
Score on test data with 5 fold cross validation:
[0.70417789 0.74342826  0.72860679  0.74188463  0.73561604]

Average Score:  0.730742723257
```

On test data performance of the model has increased significantly, from **25.61%** to **73.07%**, which is great outcome.

5.2.4 Decision Tree - Individual Tobacco Rate

We expected the model to behave exactly the same way as of the Individual rate. i.e. with simple decision tree model we would get a very low performance score on test data and will have little better performance than the linear models we had built above. The model have performed little bit better on training dataset, from **62.24%** to **71%** approximately.
Before we implement the cross validation on the decision tree model let's look at the importance of the predictor variables in the below bar graph.



As we can see that "Age" plays very important role in deciding the monthly rate of the insurance plan. The second most important feature is "Metal Level".

The performance on test data was found to be about **32.57%** which is very low and hence we implemented 5 fold cross validation on the model similar to the Individual Rate. The performance improved significantly to **73.73%**. Below is the score with 5 fold cross validation.

```
Scores for 5 fold cross validation: [ 0.70279793  0.72901925
0.74179106  0.76607729  0.74718976]

The Average Score: 0.73737505663
```

Random Forest Regressor

Random forests or **random decision forests** are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or

mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.
We tried to improve the model using ***RandomForestRegressor()*** on both the models for individual and individual tobacco rate and check if the model predicts better. Unfortunately the model showed no to very little improvement.

6. Limitations

Though the dataset is huge, the dataset is not complete, or we can say that those data were omitted to maintain the security about the issuers (companies which sell health insurance in the marketplace). This dataset is all about what was available to the consumers but how consumers have reacted to the offer would have been more interesting. So, actual enrolled people and the population data would have made this analysis much better.

The dataset does not tell anything about the **quality** of the medical service provided.

Memory limitation: During the course of this project, the dataset I had was huge more than 5gb in total and many times I got the low memory warning on my local machine, and could not run the entire project at once. So I have divided the actual code into 2 different note book, EDA (Explanatory Data Analysis) and Machine Learning.

Since the datapoints are in millions, I could not plot the regression line.

7. Other Data and Future Work

As mentioned in the limitations section, we have got the enrollment data for all the states. We would try to merge the two datasets and would do some interesting analysis on the Health Insurance Marketplace. The entire concept of exchange was to make sure that all the Americans are covered under the health insurance. Would try to find if these insurance are really helping people getting the kind of medical services they need. Once we start digging about all this information, this analysis part of this project would be more helpful to the consumers and for the issuers.
In machine learning part would like to scale the model to predict couples, and couples with dependents. There are almost 5 to 7 kinds of rates which can be predicted. We would also like to improve the hyper-parameter of the proposed model to predict various rates.

8. Conclusion

The dataset is huge, there can be very vast analysis be done on this dataset, like state wise, county wise, year wise etc., but we have kept the analysis as simple as possible and try to build the model which would predict the individual rate and individual tobacco rate.

8.1 EDA

During the EDA (Exploratory Data Analysis) we found that not all the States in the U.S participate in the Federal run health insurance marketplace or exchange. There are states which were part of this exchange in the beginning but later moved to state based exchange. And there were states which were having State based exchange but later moved on to the federal run exchange. We have also seen the states who are large consumers of medical services like Wisconsin, Texas and Florida to name few.

We have analyzed how the rates are spread across all over the U.S. What are the cheapest and costliest states to live in, in terms of healthcare.

We had try to find the relation among the number of issuers, number of benefit plans offered and the monthly premium rates across the U.S. This analysis was not very conclusive though.

8.2 Machine Learning Models

We tried to predict Individual Rate and Individual Tobacco Rate from the datasets we had. In the beginning of this part I have used a plot to show how the average monthly rate kept increasing over the years since the inception of Healthcare Insurance Marketplace. When you compare the monthly individual rate of the years 2017 and 2018, you will find this huge increase in rates. This is not the similar increase what you will find if you compare year 2016 and 2017. The increase is big in this case too. But it is not as big as last year.

The insurance company decides the rate based on the 5 criteria as explained above. My calculation and model is based on those criteria. I have used the small sample from the dataset where my training data was taken from year 2014 and 2015, test data from year 2016. That model gave me better result of around 84% accuracy when used in decision tree and cross validation, which I believed to be good as compared to the Linear model which was around 63-64%.

Decision tree with cross validation of 5 fold gave us better performance on test data compared to Linear mode. So I think **Decision Tree** with **cross validations** better bet in prediction of continuous values. **GridSearchCV** helps find us the maximum depth of the decision tree at which the model will result in optimal performance.