The State University of New York - Plattsburgh
School of Business and Economics

Predictive Analytics - MGM 550A
Fall 2019

Analyzing Healthcare Cost using Predictive Analytics

Final Project Report

Submitted By:
Abhey Sharma
Danish Rai
Olajide Obatunwase

Submitted to:
Dr. Shakiba Enayati

Submission date:
November 11, 2019

Introduction

The U.S healthcare cost has maintained an upward trajectory over time and is projected to keep increasing. As a share of the nation's GDP, healthcare spending accounted for approximately 18 percent compared to 11 percent in comparable countries like the United Kingdom and Canada. In dollar terms, the cost of healthcare here is roughly double that of similar countries. Despite the Affordable Care Act reform program, there seems not to be a significant reduction in the trend. However, the objective of this project is to analyze the impact of potential drivers of US healthcare costs and make recommendations on possible strategies for reducing costs in the near future.

Previous Work

A few numbers number of stakeholders and professional bodies have analyzed US healthcare costs using trend analysis without recourse to quantitative research. However, a literature search indicated that very little quantitative research has been done to analyze healthcare costs using predictive analytics.

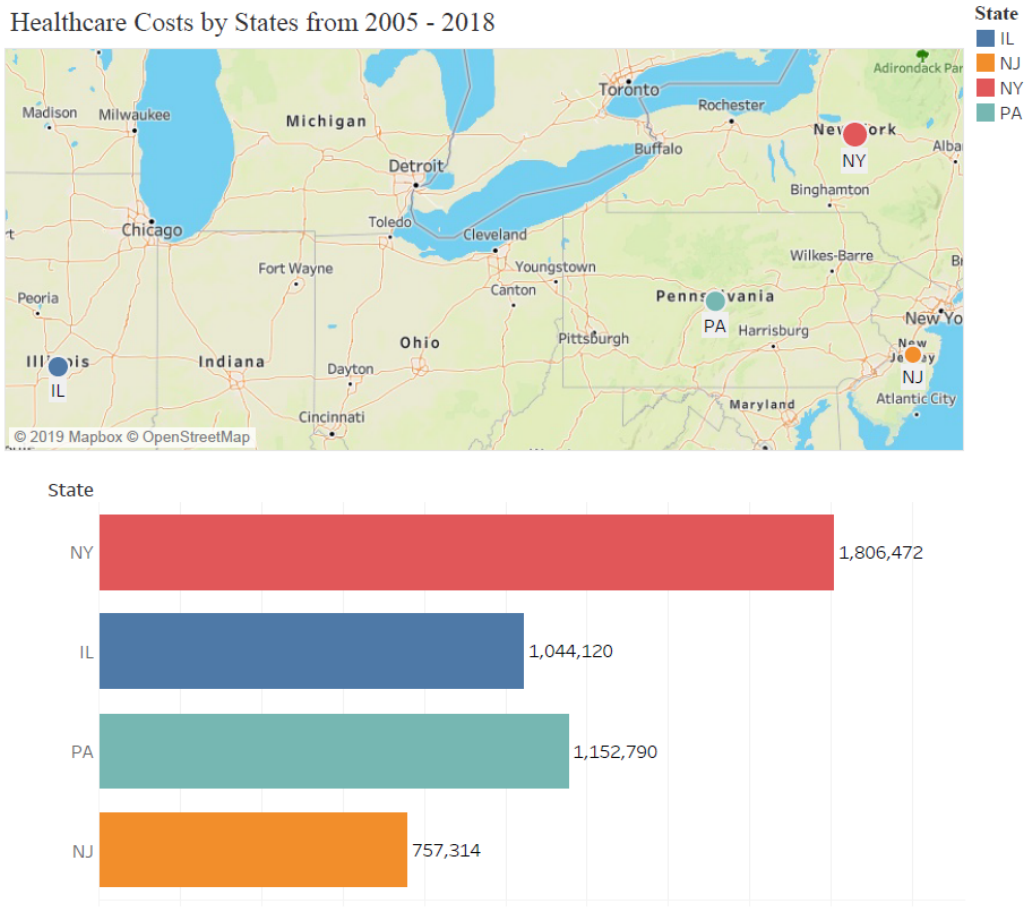
In 2018, the Society of Actuaries (SOA) with Kaiser Family Foundation (KFF) carried out a research on the topic: What Can We Do About the Cost of Health Care? The SOA is the world's largest provider of actuarial research and education and KFF is a nonpartisan source of analysis of current health policy issues, with a long-standing interest in how health spending growth affects government, employers, and consumers. The research work explores U.S Health cost drivers to include price and interest, population, aging. It was reported that 50 percent of the increase in U.S. expenditure from 1996 to 2013 was due to an increase in price and intensity while 23 percent accounted for the population. The report identified that using Technology in Direct Patient Care can lead to prevent and diagnose disease (like cancer) at an early stage.

Data and Methodology

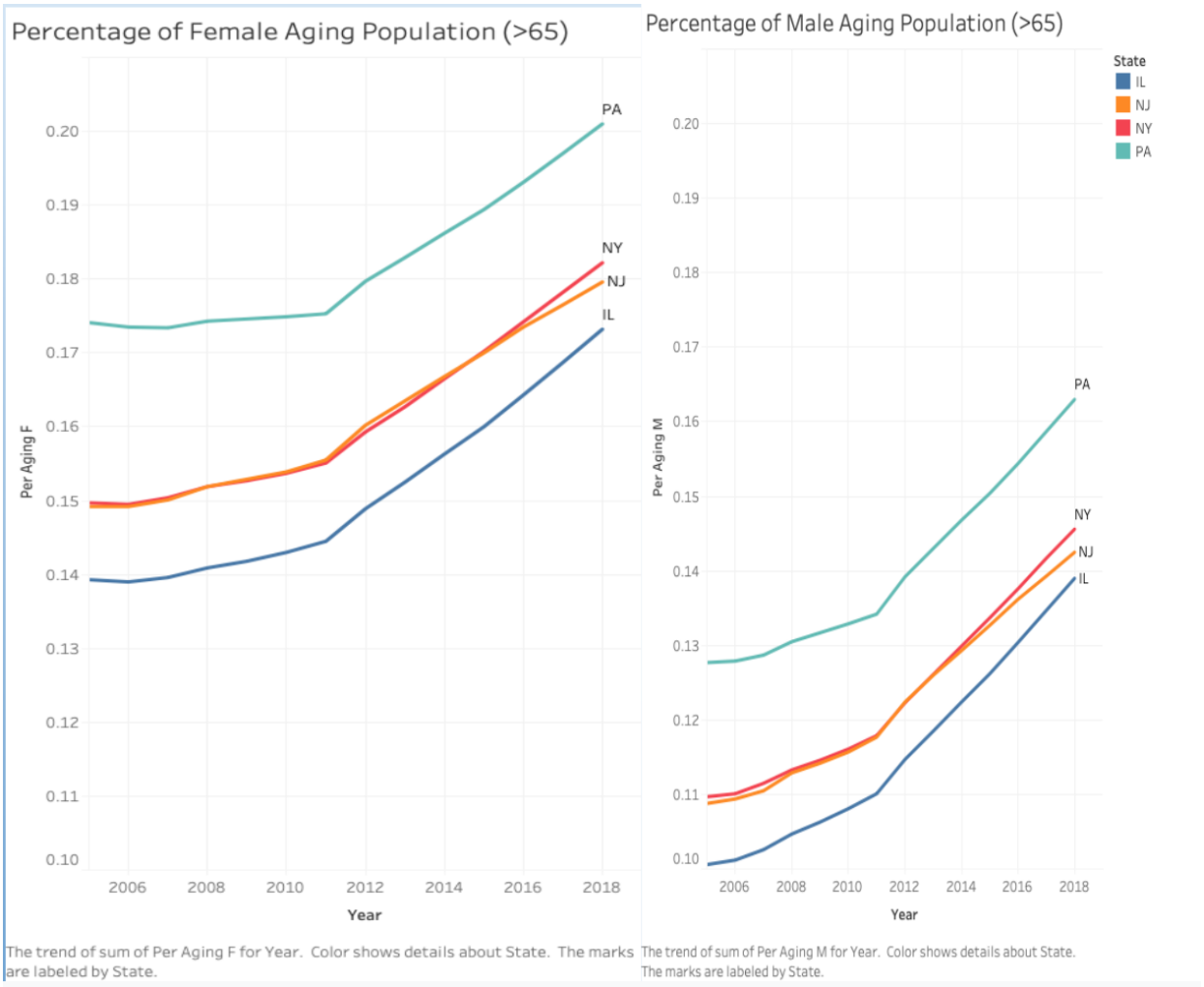
The analysis used from S&P Global (SNL) Market Intelligence, U.S. Bureau of Labor Statistics, Federal Reserve Economic. The project explores the potential predictors of Healthcare costs across 4 states, New York (NY), New Jersey (NJ), Pennsylvania (PA), and Illinois (IL) from 2005 to 2018 to include Population, Medicare and Medicaid claim incurred, Real Median Household Income, Percentage of Aging population(65 and over), Per Aging Female and Per Aging Male. The idea is to evaluate the impact of each variable on healthcare expenditure using statistical learning techniques and make future predictions of healthcare costs.

Graphical Analysis

Historically, the cost of healthcare has systematically maintained an upward trend across the 4 states of NY, NJ, PA, and IL. The state of New York, with the highest population growth, incurred the highest costs from 2005 to 2018 relative to NJ, PA and IL.



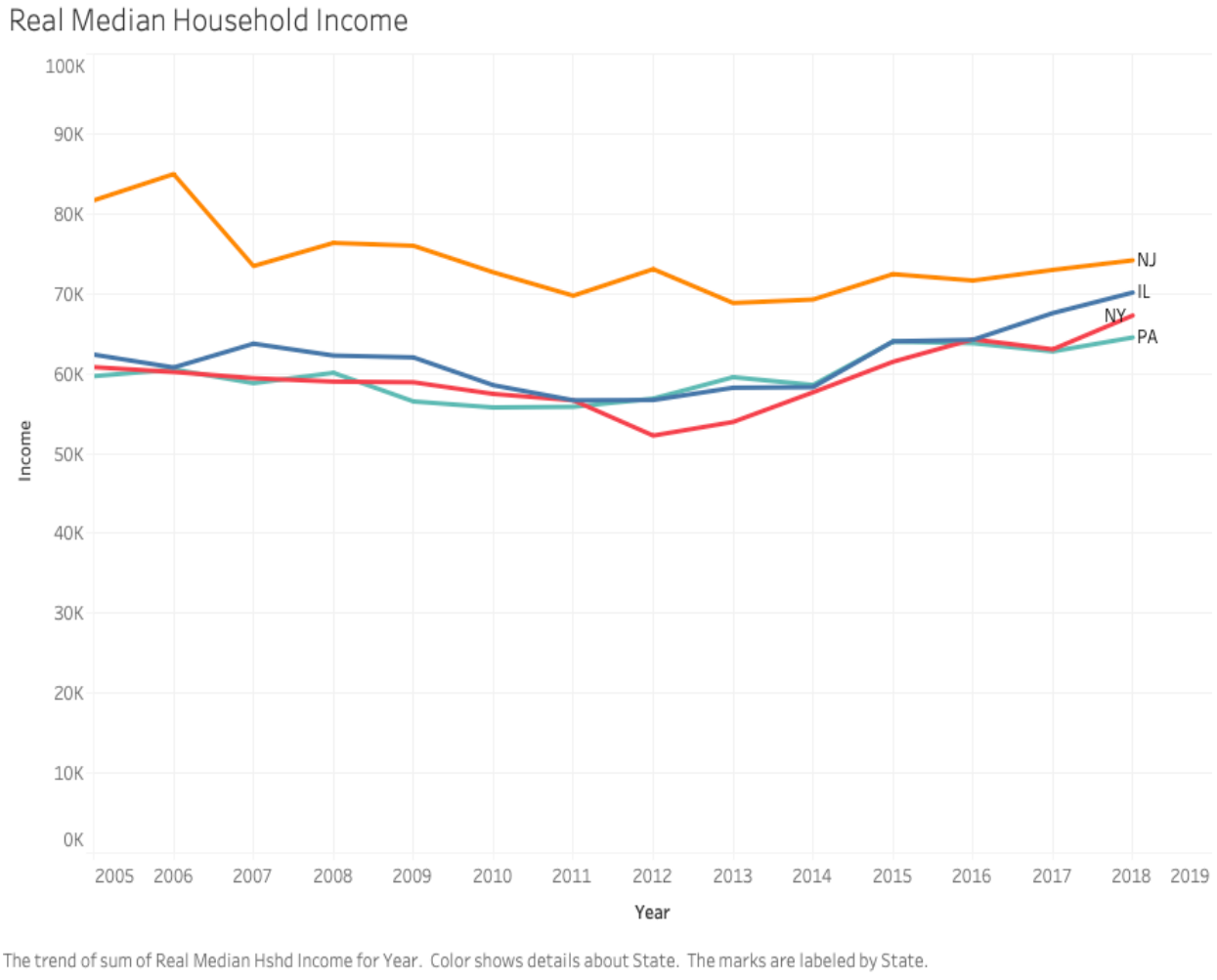
Comparing Percentage of Aging Population Female vs Male



It is important to understand the predictors in detail to perform our analysis. The line chart above explains the percentage of aging population Female vs Male. It can be observed that the female population makes a more significant portion in the chart as compared to the male population. PA has a larger share of female and male elderly population relative to NY, NJ and IL. The female population of PA is a little above 20% whereas the male represents about 16.2% of the aging population in the year 2018. The reason for the growth from the year 2005 to 2018 is due to the aging of the baby boomers (individuals born between 1946-1964) and consistently low fertility rate. The female population in NY and NJ have close call between the percentages. It can be said that for NY had about 18.3% of the female aging population and NJ had 18% for the year 2018. IL had about 17.4% for the same population. On the other hand, for male population, similar ranking can be observed in terms of the state as mentioned but NY stated with 11% and was at about 14.5% for the male aging population. NJ started at about 10.9% in 2005 and then was at

14.3% in 2018. Last but not the least Illinois, there was a huge surge from 10% in 2005 to 14% in 2018 for the male aging population.

Line Chart of Real Median Household Income

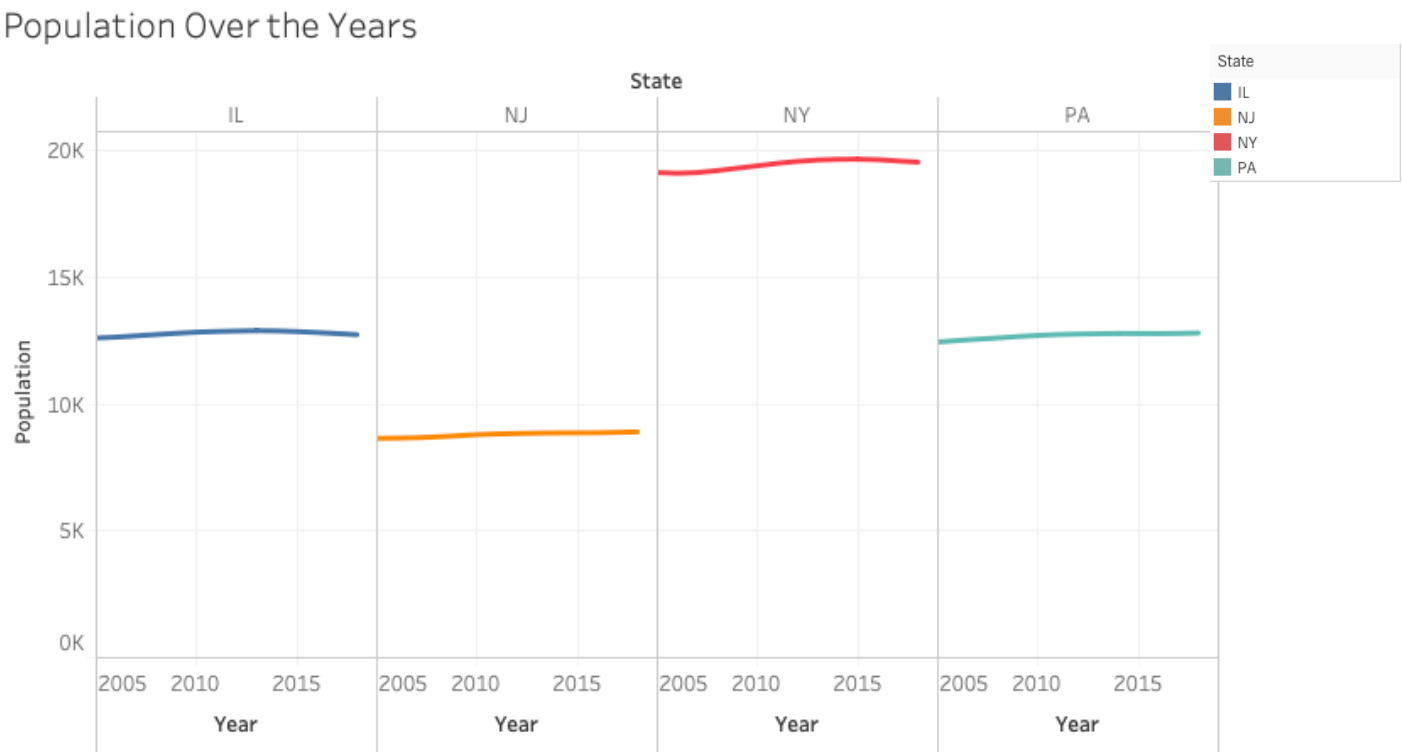


Real median household income measures the combined gross income of all members of a household. NJ has always topped the charts for Real median household income as compared to NY, IL, PA. NJ has seen a decrease from a little above $80,000 in 2005 to $74,094 in the year 2018. IL stands at $70,000 in 2018, NY $65,000 and PA at $64,000 and all three states have crossed each other at some point or the other from 2015 - 2018

Line Chart Per Capita Personal Income

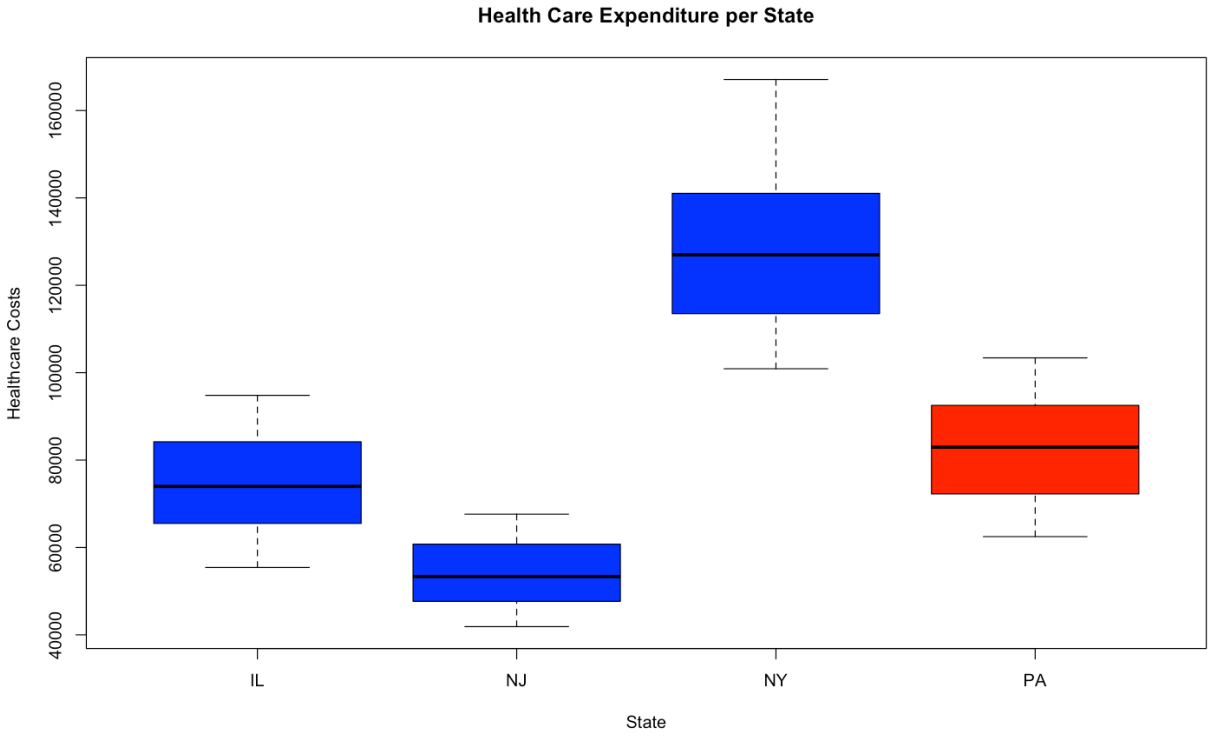
Per capita income measures the average income earned by each individual in a given area. In the year 2005, NJ was the best among all the four states by providing $45,000 average per year to its residents. However, over the years NY slowly picked up to NJ and as of 2018, now NY provides a slightly higher average of $69,000 as compared to NJ. IL and PA starting from the year 2009 have been going off of each other, pretty close in terms of their per capita personal income until 2018 at about $56000-$57,000. This describes the similarity between NY, NJ and IL and PA.

Line Chart of Population over the Years



The population is a crucial factor to consider related to projects of this nature. NY undoubtedly has the highest population in comparison from the year 2005 to 2018. The reason for the population is due to NY being the most famous destination for working, positive net migration rate as well as the presence of the well-known metropolitan cities in the United States. The trend is later followed by IL and PA and NJ is the least in terms of population.

Box for Health Care Expenditure per State



The plot shows the cost of healthcare per state with New York having the highest cost. The colors show the dominance of the democratic (blue) and republican (red) party for the respective states. NY is the highest in terms of healthcare cost due to higher taxes, and a lot of population. PA is the next in terms of high expenditure, followed by IL and NJ. From the box plot it can be easily seen the difference between NY and other states, especially NJ. The expenditure gap between those two states is very high.

Empirical Analysis

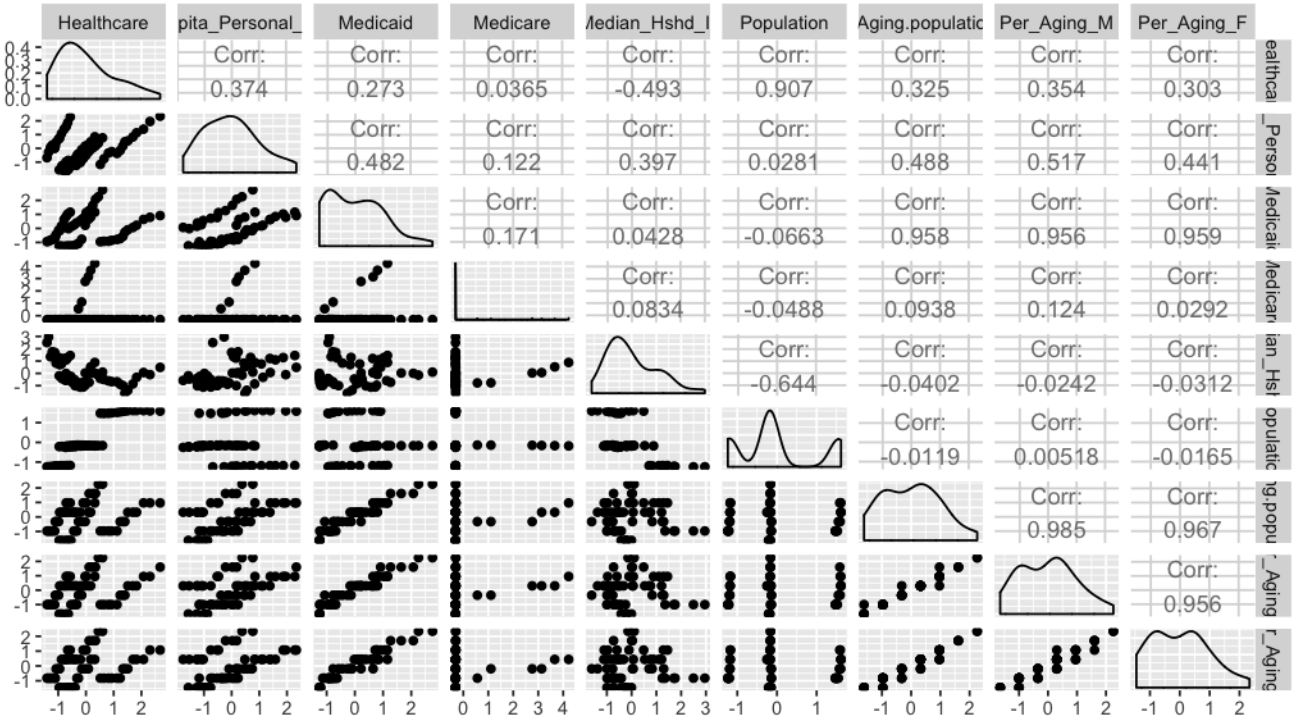
Principal Component Analysis (PCA)

The idea of PCA is to identify the variables that capture most of the information in predicting the healthcare costs. Information in this case relates to the sense of variability. With the principal component regression, the first five principal components capture over 90% of the total variation. This suggests that some of the predictors impact the direction of the response. However, there is no guarantee that the predictors will give the best direction for predicting the response. Fitting a least-square model might show otherwise. In this case, the PCA is a necessary condition but not sufficient condition to determine the direction of the response variable.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.21	1.31	1.05	0.93	0.52	0.22	0.2	0.15	0.11
Proportion of Variance	0.54	0.19	0.12	0.1	0.03	0.01	0.01	0.01	0.01
Cumulative Proportion	0.54	0.74	0.86	0.96	0.99	0.99	0.99	0.99	1
Weights per PC									
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
Per Capita Personal Income	0.3	-0.24	0.3	-0.63	0.09	-0.04	-0.49	0.31	-0.11
Medicaid	0.43	0.03	-0.13	0.17	-0.17	0.63	0.23	0.53	-0.04
Medicare	0.09	-0.15	0.74	0.59	-0.19	-0.06	-0.19	0.02	-0.02
Real Median Household Income	0.03	-0.71	-0.07	-0.18	-0.56	-0.03	0.3	-0.21	0.1
Population	-0.01	0.62	0.34	-0.37	-0.55	0.02	0.2	-0.11	0.07
Percentage of Aging population 65+	0.43	0.1	-0.17	0.1	-0.05	-0.45	-0.06	0.17	0.73
Per_Aging_M	0.44	0.09	-0.12	0.08	-0.07	-0.54	0.24	0.04	-0.65
Per_Aging_F	0.42	0.1	-0.25	0.1	-0.17	0.27	-0.5	-0.62	-0.08
Year	0.39	-0.06	0.35	-0.16	0.52	0.16	0.48	-0.39	0.14

The first principal component, from the analysis of the PCA, shows that healthcare costs is mostly influenced by Medicaid, Per Capita Personal Income, Percentage of Aging Population, Aging Male, Aging Female and Year. On the other hand, fitting the best subset least square model shows that Year, Population, Real Median Household Income and State are statistically significant in predicting healthcare costs.

Correlation Analysis

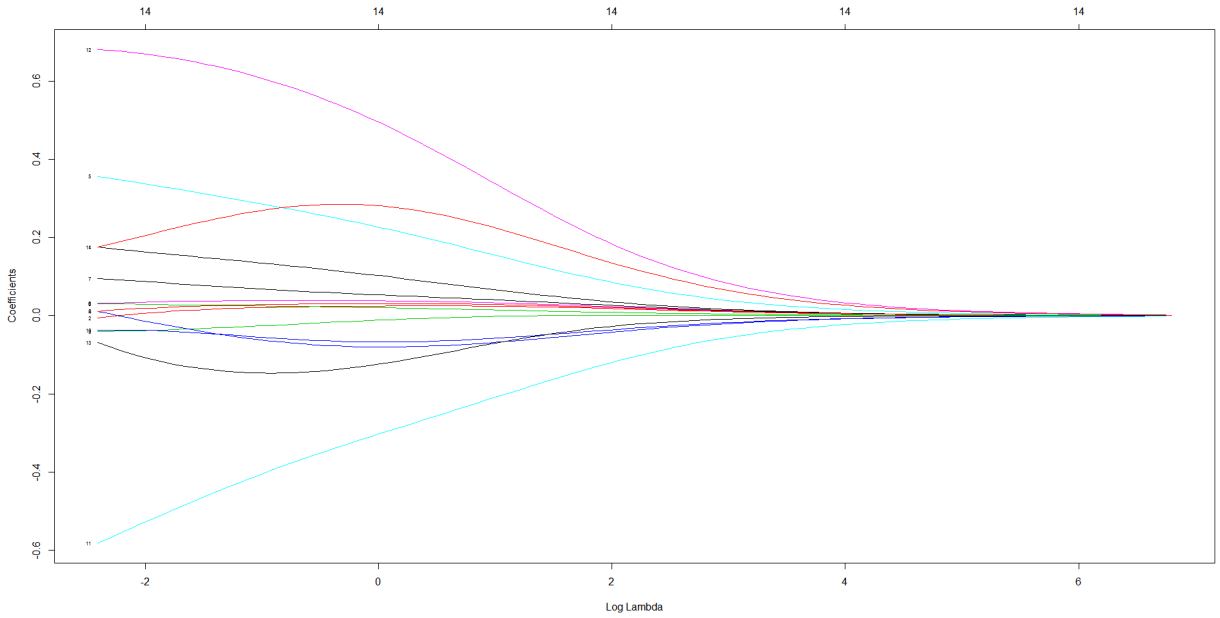


The correlation matrix plot shows the presence of correlation amongst some of the predictors. Aging Female is strongly correlated with Medicaid. Similarly, Per Aging Population is strongly correlated with Aging Female. These show there is information overlap (redundancy) between the variables.

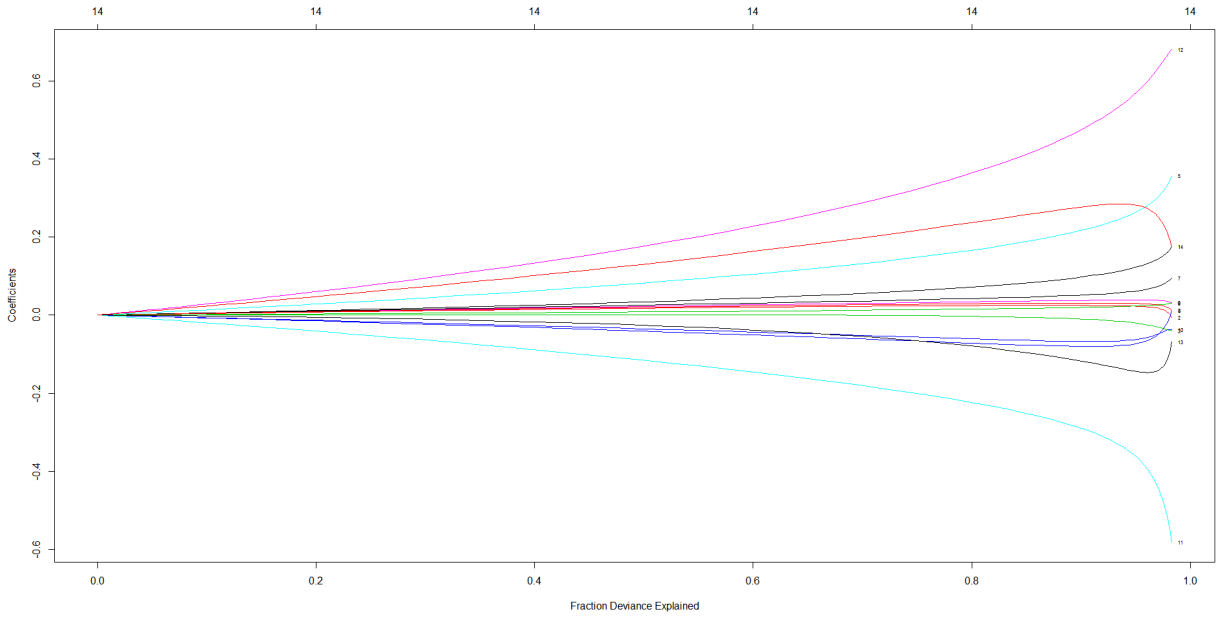
Variable Selection

The features selected for the model was done using ridge regression, lasso, and best subset approach while comparing the MSE error to determine the variables of interest.

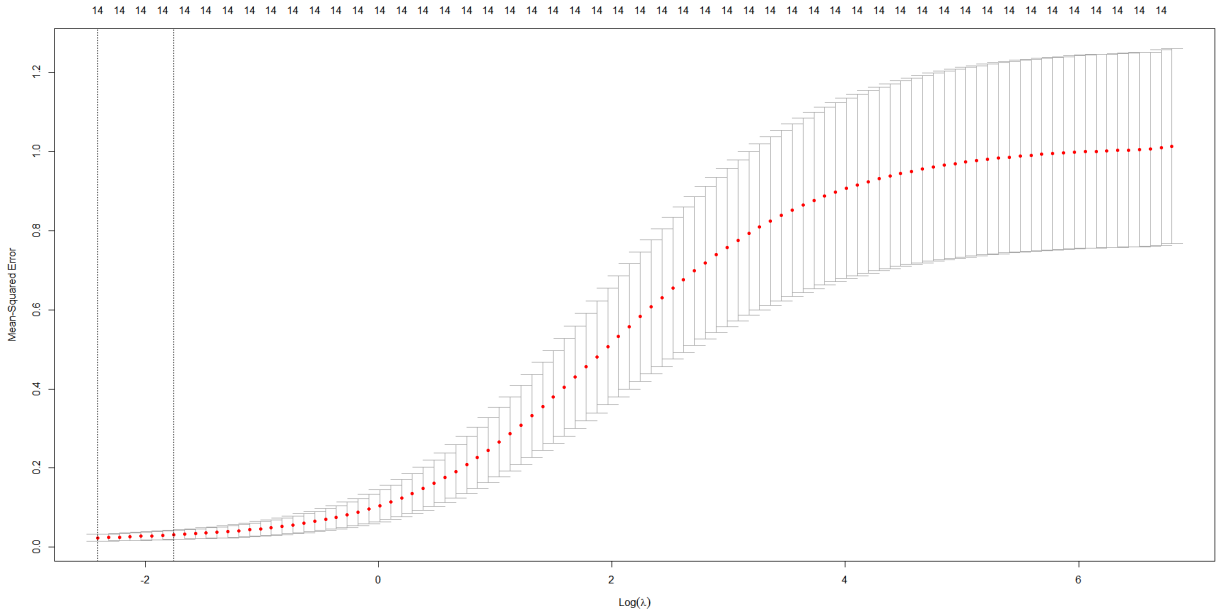
Ridge Regression



Increasing lambda 6 shrinks the coefficient to zero. For a relaxed log lambda value, the coefficients begin to increase, consequently, RSS for the coefficients are likely to increase. Increasing lambda helps to reduce (shrink) the size of the coefficients but not make them zero thereby reducing the variance.



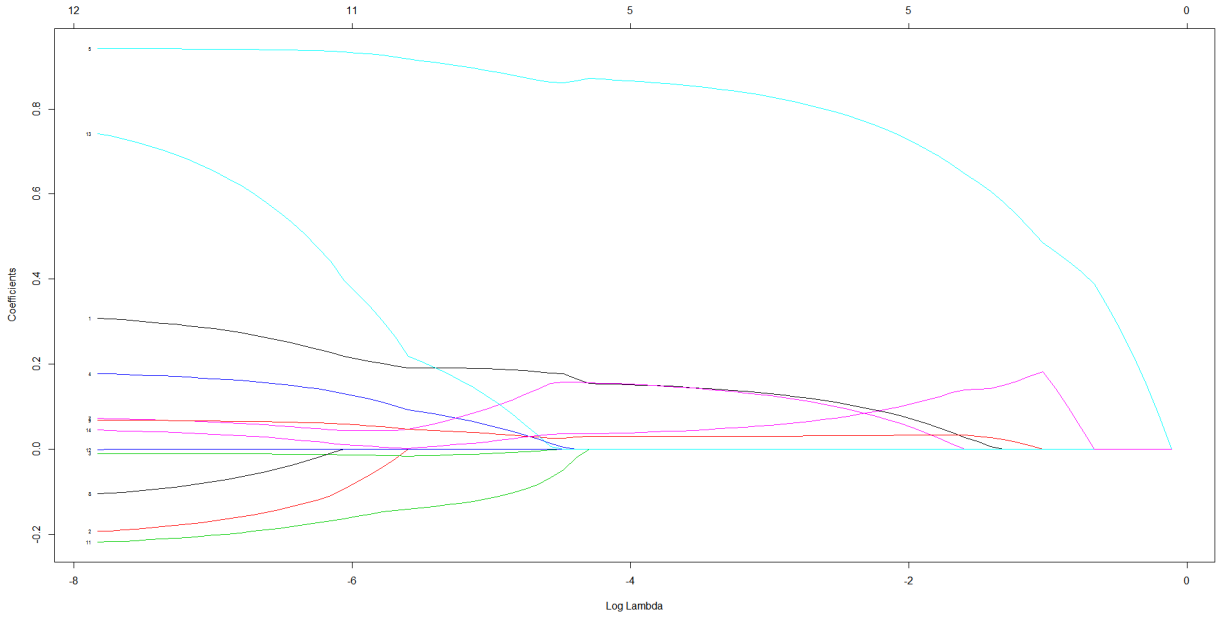
At deviance = 0.2, 20% of the variability is being explained with a slight increase in coefficients. However, at Deviance = 0.8, there is a sudden jump with the coefficient being highly inflated, indicating there might be overfitting in that region.



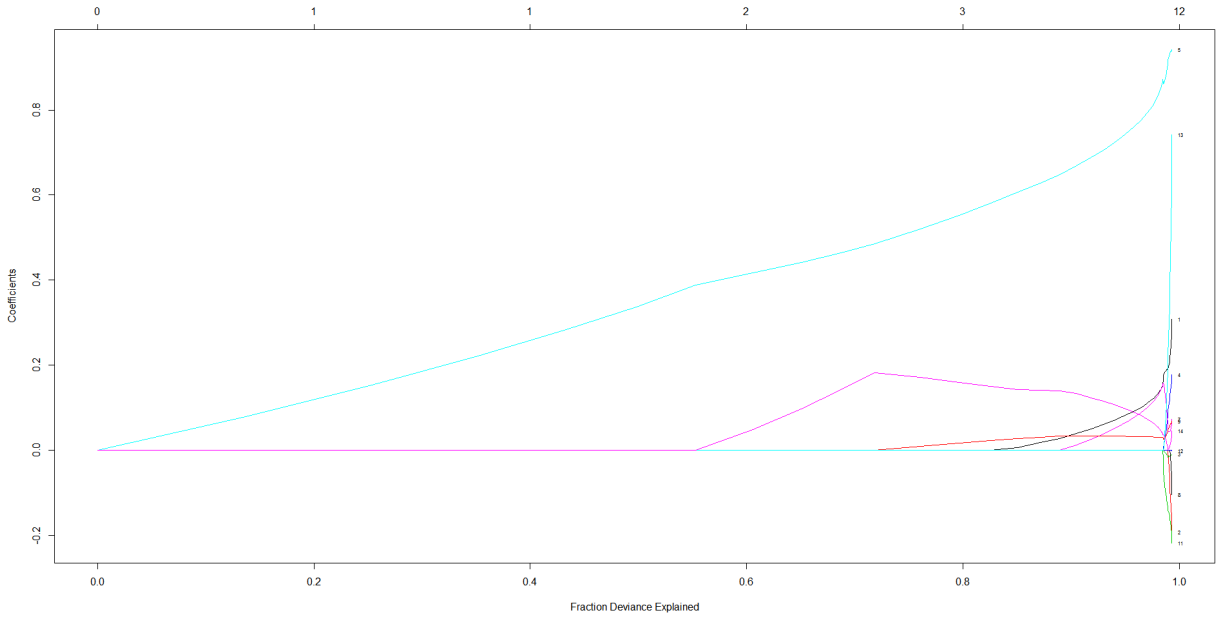
The plot of log lambda against the mean squared error shows the maximum error tolerance level for lambda to be 0.21. It is worthy of emphasis that the ridge regression performs better when the response variable is a function of many predictors (at least 20)

Lasso

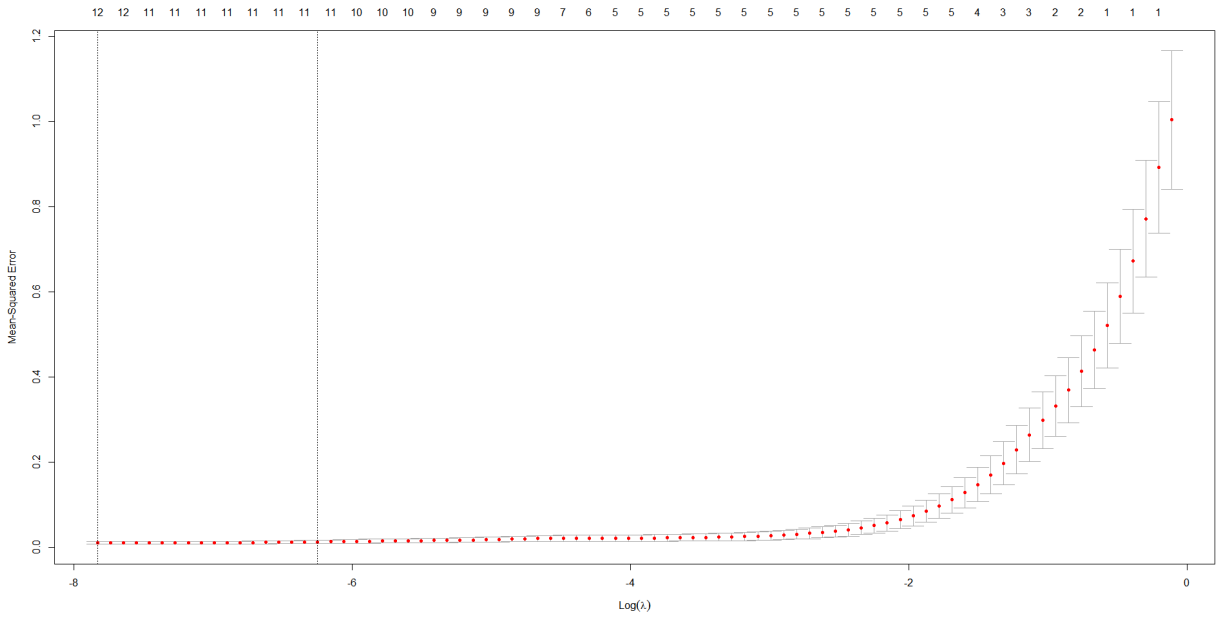
Unlike the ridge approach, using the lasso approach helps to reduce the number of variables by shrinking some of the variables to zero. Consequently, the resulting coefficient estimates are sparse.



The numbers on the horizontal line of the graph indicates the number of variables that are non-zero for a given lambda. For log lambda of size 4, the resulting number of variables is 4.



From 0.8, there is a jump in coefficients indicating the presence of overfitting.

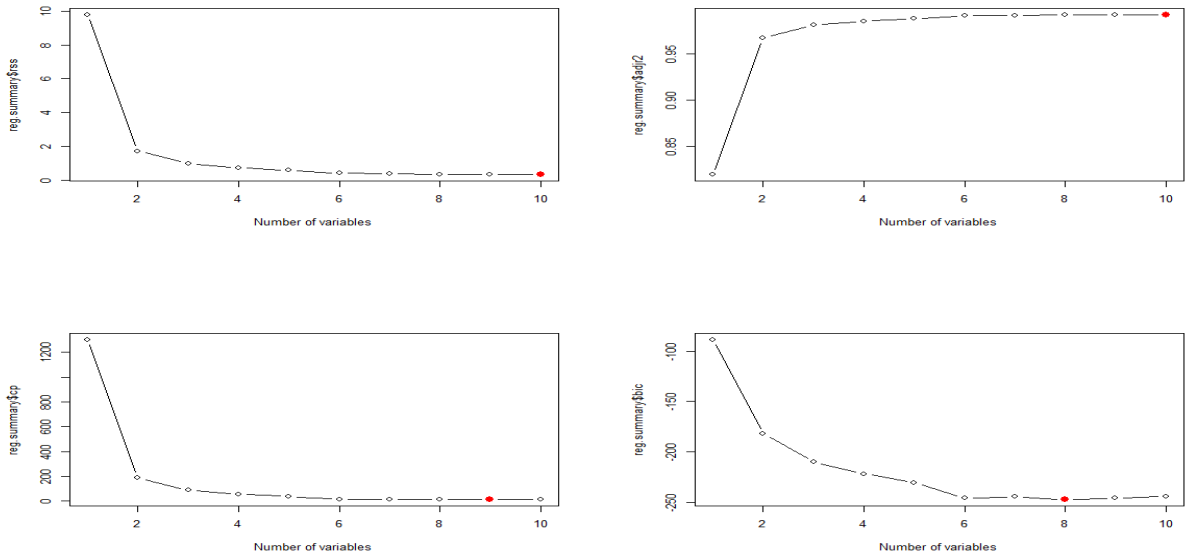


The plot of log lambda against the mean squared error shows the maximum error tolerance level for lambda to be is 0.4662 with some of the variables having a coefficient of zero.

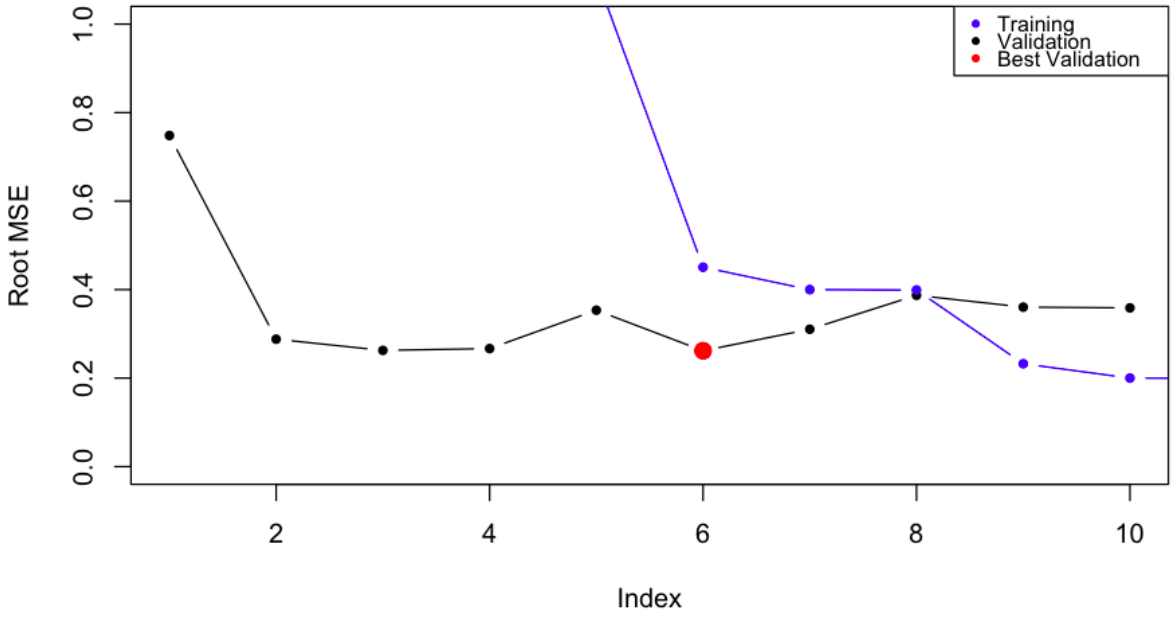
The Best Subset

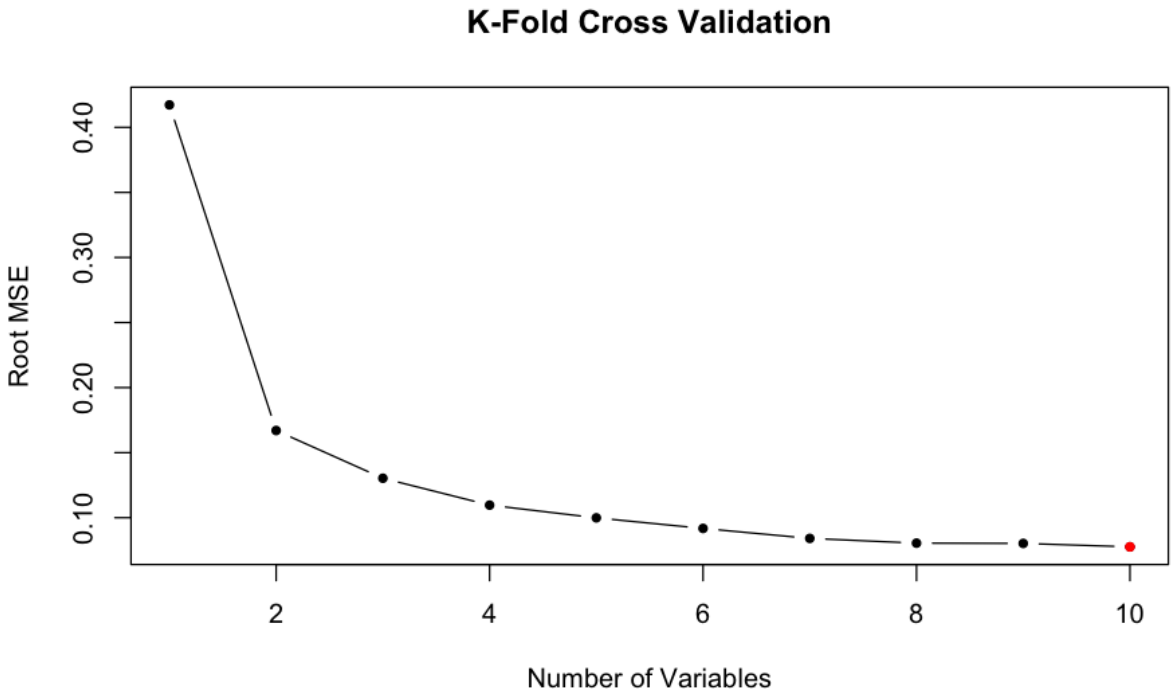
The choice of model was done using the best subset and using cross validation approach to choose the best model. The backward and forward approach was not used due to few predictors.

The plot displays the selected variables for the best model with a given number of predictors, ranked according to the RSS, Cp, Adjusted R square, and BIC. The model with the lowest BIC is the eight-variable model.



The selection of variables was done using the best subset selection approach on the full and trained data set. It turns out that best model under each data set contains six-variable with a low root MSE. The difference in each model are the Year and the Per Capita Personal Income. The training model has Year while the full model has Per Capita Personal Income. However, the best model contains six predictors.





K-fold cross validation shows different model sizes and their respective Root MSE. From the plot, a lower Root MSE starts from the six-variable model. The plot shows that the Root MSE does not change a lot after the seven-variable model. Hence, it suffices to choose seven-variable model contain, Population, Per Capita Personal Income, State, Year and Real Median Household Income.

Regression Analysis

Analysis Prior to Variable Selection

Summary from R-Code

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -79.17746   41.71423  -1.898 0.064410 .
Per_Capita_Personal_Income                0.32089    0.07056   4.548 4.39e-05 ***
Medicaid                                 -0.13384    0.06552  -2.043 0.047228 *
Medicare                                  0.02093    0.01988   1.053 0.298334
Real_Median_Hshd_Income                   0.18798    0.03171   5.929 4.64e-07 ***
Population                                3.26792    0.78273   4.175 0.000143 ***
Percentage.of.Aging.population.65.and.over. -0.02478    0.08570  -0.289 0.773881
Per_Aging_M                               0.08650    0.07724   1.120 0.268958
Per_Aging_F                              -0.10418    0.06615  -1.575 0.122598
Year                                      0.03949    0.02068   1.910 0.062843 .
StateNJ                                   2.15025    0.80386   2.675 0.010529 *
StateNY                                  -3.96407    1.35752  -2.920 0.005552 **
StatePA                                   0.77420    0.15157   5.108 7.10e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.085 on 43 degrees of freedom
Multiple R-squared:  0.9944,    Adjusted R-squared:  0.9928
F-statistic: 630.7 on 12 and 43 DF,  p-value: < 2.2e-16
```

Diagnostic Plot

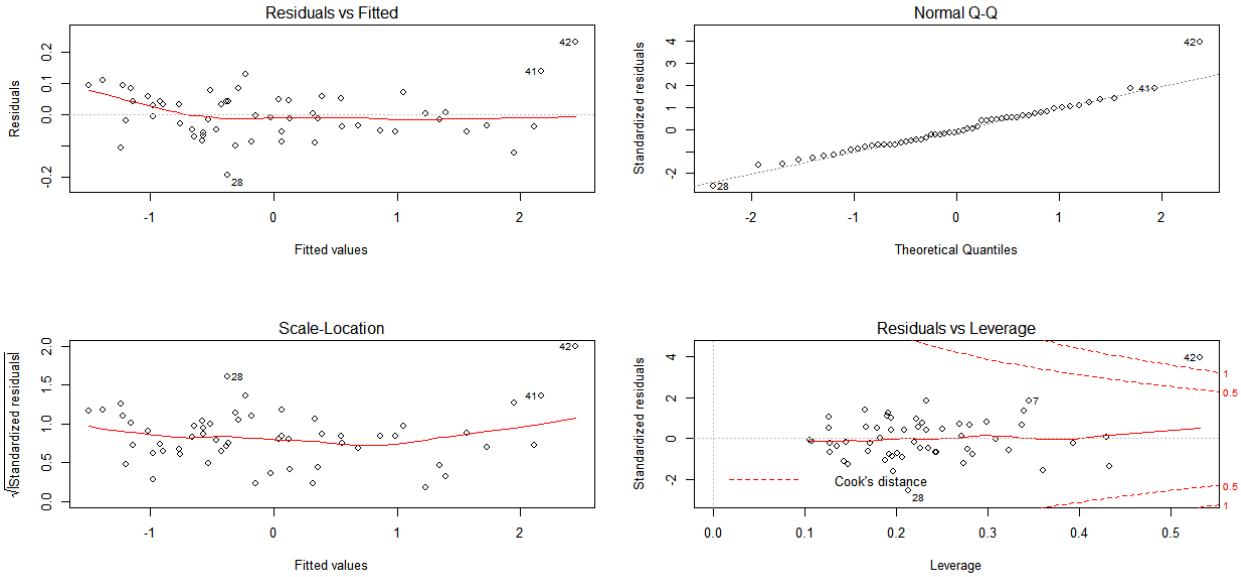


Fig. 1.0

The diagnostic plot shows that the residuals have no non-linear pattern and the model follows a normal distribution. The residuals are approximately normal with few points away from the line towards the lower and upper tail. The scale distribution plot does not indicate any presence of heteroskedasticity which implies the variance of the error term is constant (homoscedasticity). The plot of the residual vs fitted shows that the observation point is an outlier (the 42-observation point). However, this does not seem to impact the model after fitting the model.

Data Partitioning

The training set has all the observations for 2005 to 2013 and the validation set has all the observations from 2014 - 2018.

Fitting Model after Variable Selection

Response Variable: Healthcare

Predictors: Population, Per Capita Personal Income, State, Year and Real Median Household Income.

Summary from R-Code

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -83.23782   22.37891  -3.719 0.000887 ***
Year                        0.04154    0.01109   3.744 0.000830 ***
Population                  4.06495    0.60958   6.668 3.09e-07 ***
Per_Capita_Personal_Income  0.07396    0.04998   1.480 0.150107
Real_Median_Hshd_Income     0.07008    0.02540   2.759 0.010105 *
StateNJ                     3.41432    0.64628   5.283 1.28e-05 ***
StateNY                    -5.30725    1.03411  -5.132 1.93e-05 ***
StatePA                     0.40988    0.03089  13.271 1.34e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0456 on 28 degrees of freedom
Multiple R-squared:  0.9978,    Adjusted R-squared:  0.9972
F-statistic:  1790 on 7 and 28 DF,  p-value: < 2.2e-16
```

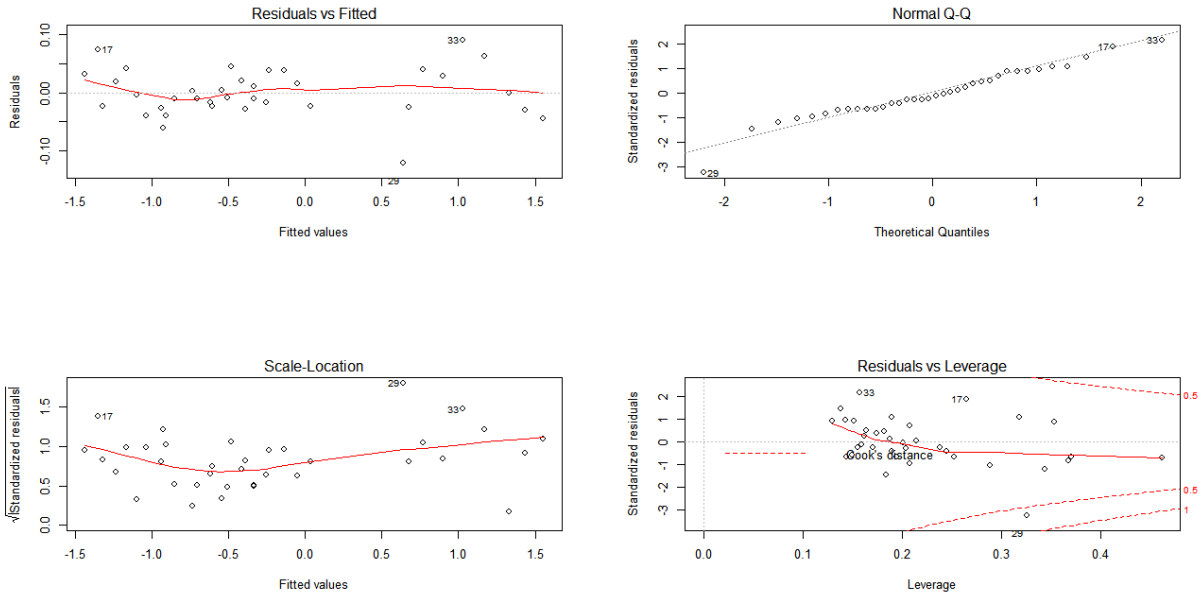
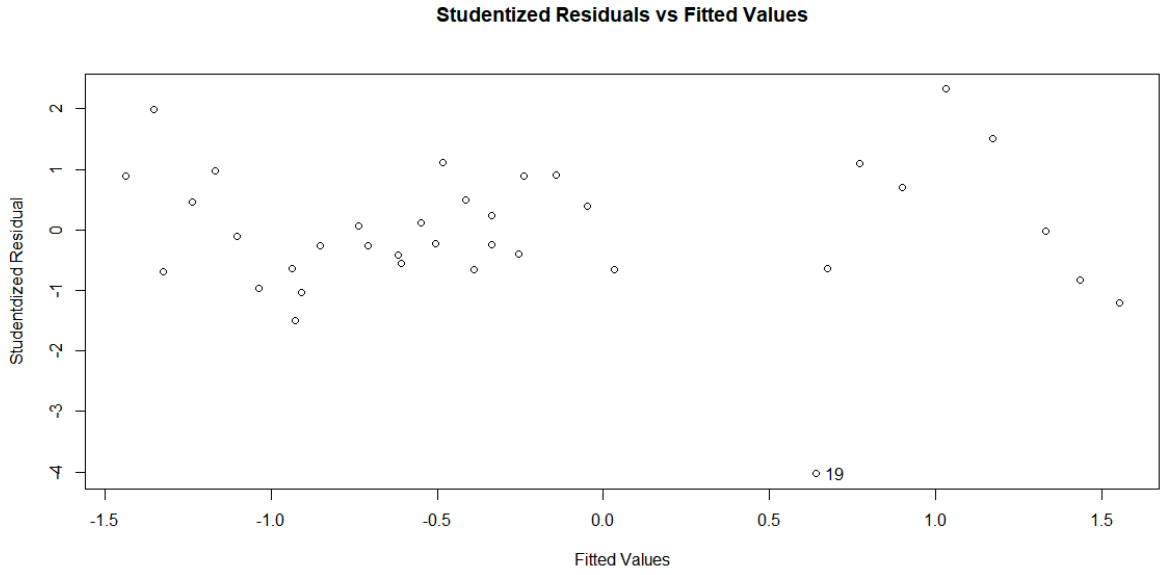


Fig. 1.4

The diagnostic plot after variable selection shows a better distribution of the residuals as compared to the diagnostic plot figure 1.0. However, mostly it stays the same. This is expected as the final model uses 7 out of 10 variables. The only difference in the plot can be found in the

Residual vs Leverage plot as it now shows a potential leverage observation 19 of train data set as compared to the 42nd observation of the entire data set in Figure 1.0.

Testing for Leverage



The studentized plot confirms that 19th observation is an outlier and a leverage point.

Fig. 1.5

Diagnostic plot after removing the outlier

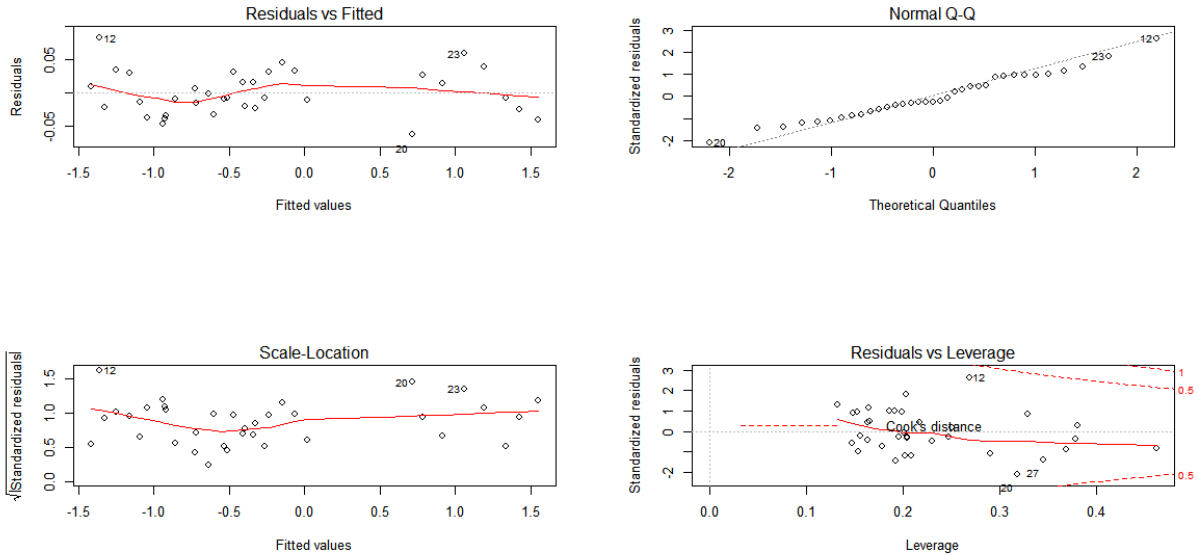


Figure 1.6

All the plots remain the same except in the Residual vs Leverage, the Cook's distance has no point under it.

Evaluating the Final Model using the Cross-Validation Approach

The scaled test data was predicted using the model created using the training data set and computed the Mean Squared Error (MSE). For the chosen model, the scaled Test MSE is 0.054 and Training MSE is 0.0016. Interestingly, after we removed the outlier, the test MSE increases to 0.0734. The RSE on unscaled is $28471.36 (in millions) and the training error is $19521.67 (in millions). It is expected that the training error is lower than test error as the model is trained on more observations.

Predicting Healthcare Cost for 2019

We estimated the values of our predictors for 2019 and predicted the Healthcare cost for each state. We got the following predicted values for each state.

```
      PA       NJ       NY       IL
61354.21 47092.29 88927.68 61297.71
```

After adding and subtracting the error of $28,471.36, following the threshold for 2019 Healthcare cost.

PA: ($32882.85M, $89,825.57M)

NJ: ($18,620.93M, $75,563.65M)

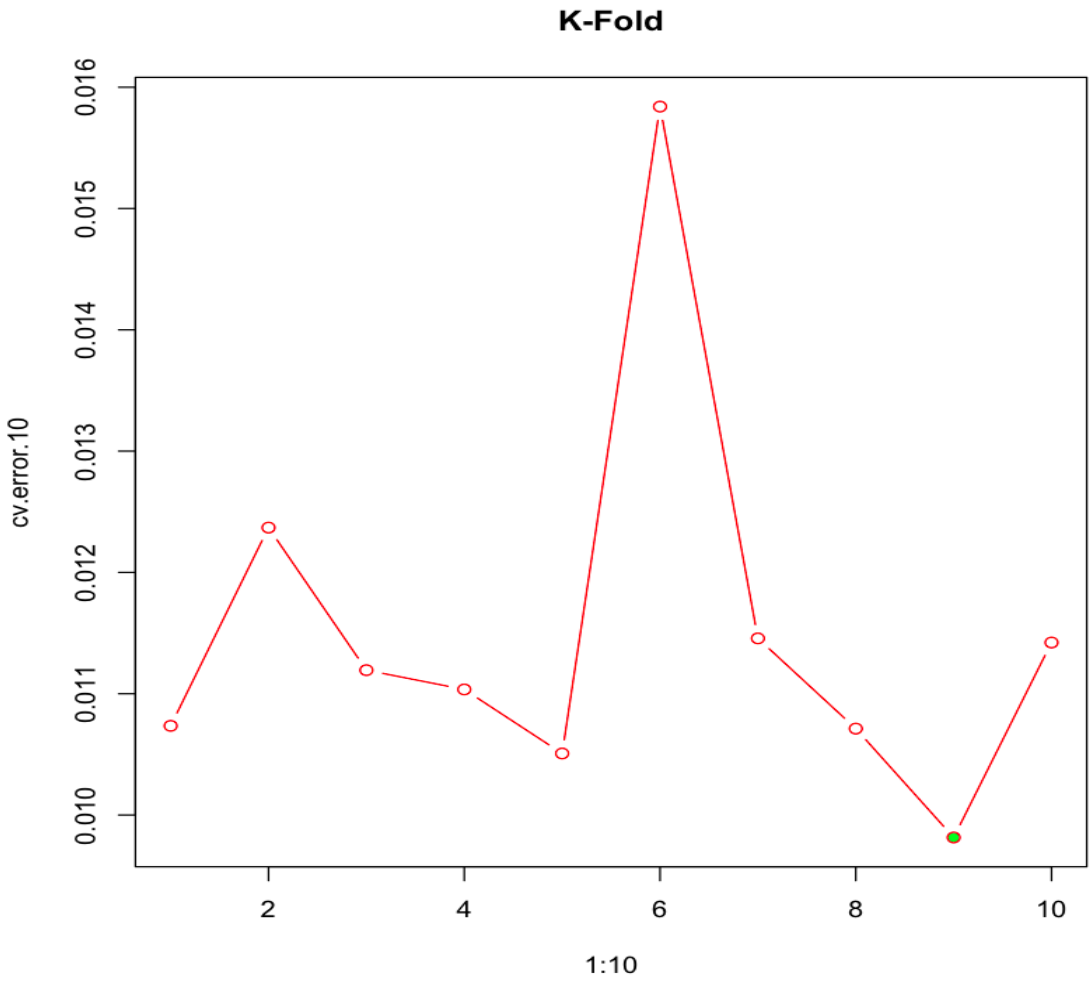
NY: ($60,456.32M, $117,399.04M)

IL: ($32,826.35M, $89,769.07M)

As per our model, we expect the cost of healthcare to fall in between the threshold.

LOOCV vs K-Fold

Using leave one out cross validation (loocv), the mean squared error is 0.0123 which is 90198.8 on the scaled data set. LOOCV takes one value as test and uses the rest as training data set, which means such a high error for test is expected.



Using another approach called K-fold, which divides the data in k different fold and uses one entire fold as test data and rest as training. We do this on 10 folds for our model.

Using the 10-fold approach, we can see that we are able to achieve a lowest test error of 0.0098, highlighted in green on the chart, on the unscaled data which is lower than LOOCV (0.0123) and the validation set approach (0.054). This shows that the model is capable of calculating the Healthcare cost with the lowest MSE of 0.0098.

Classification

For classification, Logistic, Linear Discriminant Analysis and KNN classifiers were used to achieve a model with the lowest misclassification error. We first create 2 classes i.e. Healthcare cost greater than 50th percentile, it will be classified as High, otherwise low.

For 2 Classes: High and Low

Logistic

R Code:

```
ClassTrainIndex= ht.df2$Year<2014
ClassTrainData= ht.df2[ClassTrainIndex,]
ClassTestData= ht.df2[!ClassTrainIndex,]

glm.fit=glm(HighLow~Year+Population+Per_Capita_Personal_Income+Real_Median_Hshd_Income+State,
            data = ht.df2,subset=ClassTrainIndex,family = "binomial")#Logistic regression
glm.prob=predict(glm.fit,newdata = ClassTestData,type="response")
glm.pred=ifelse(glm.prob>0.5,"High","Low")
table(glm.pred, HighLow[!ClassTrainIndex])
mean(glm.pred != HighLow[!ClassTrainIndex])#Misclassification Error
mean(glm.pred ==HighLow[!ClassTrainIndex])#Model Accuracy
```

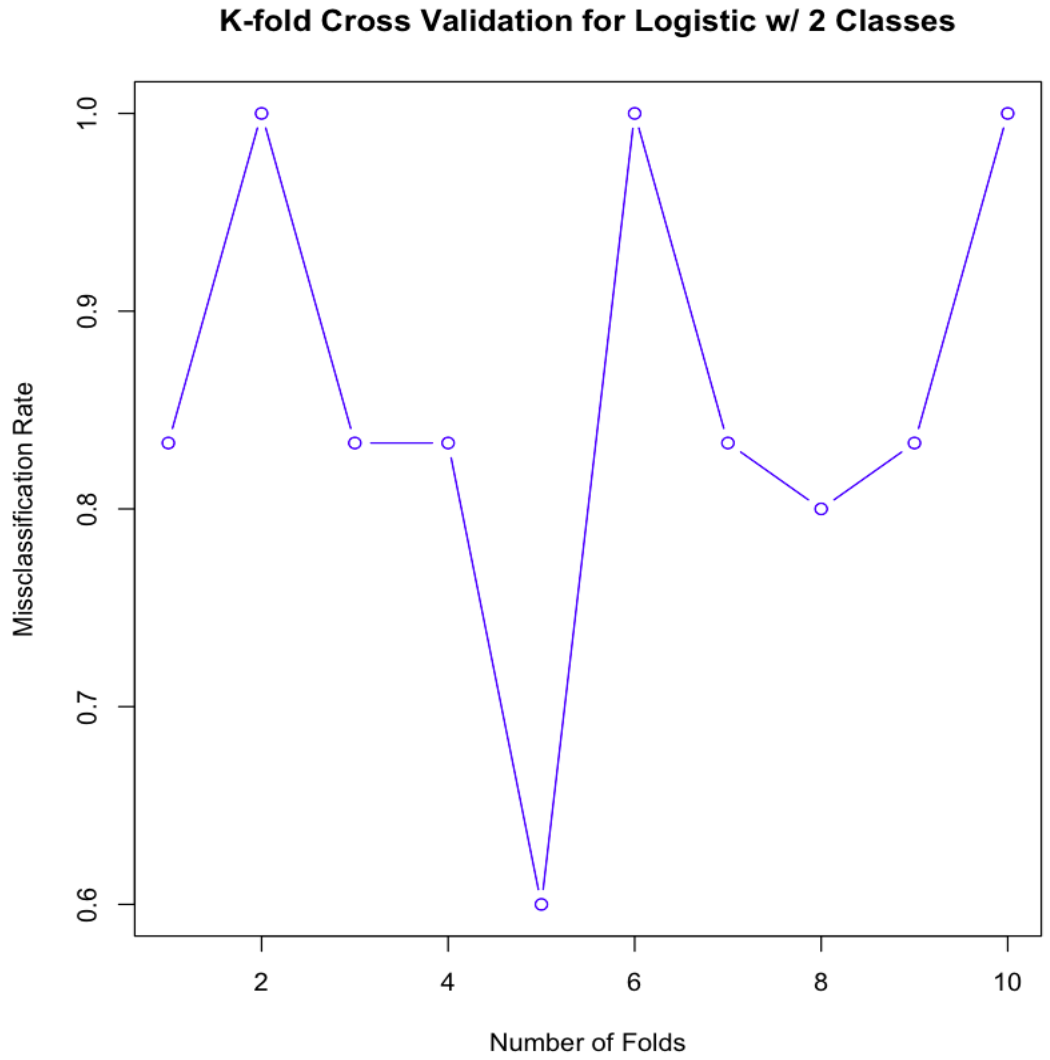
Confusion Matrix (Output from R)

```
 glm.pred High Low
     High    7   4
     Low     8   1
```

Similar to the linear regression, we fit the model on the training data set. For logistic regression, we used a threshold of 0.5. Any predicted value higher than 0.5 gets classified as High and Low otherwise. To evaluate the model, we computed the misclassification error i.e. 0.40 with an

accuracy of 0.60. This is because our model has a high number of False Positives and False Negatives.

K-Fold for Logistic



Using the K-fold cross validation approach, we can see that our logistic model is only capable of providing 0.6 as the lowest test error. This means that if we were to use logistic model, we will only be correct 40% of the time.

Linear Discriminant Analysis

R code

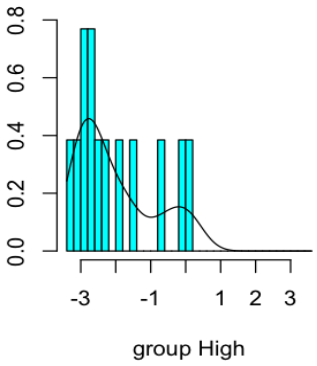
```
lda.fit = lda(HighLow~Year+Population+Per_Capita_Personal_Income+Real_Median_Hshd_Income+State,
              data = ht.df2, subset = ClassTrainIndex)
lda.fit$counts
lda.fit$prior
lda.pred = predict(lda.fit, newdata= ClassTestData)
testclass=ifelse(lda.pred$posterior[,1]>0.5,"High", "Low")
test.Healthcost = HighLow[!ClassTrainIndex]
table(testclass, test.Healthcost)
mean(testclass != test.Healthcost)#Misclassification Error
mean(testclass == test.Healthcost)#Model Accuracy
```

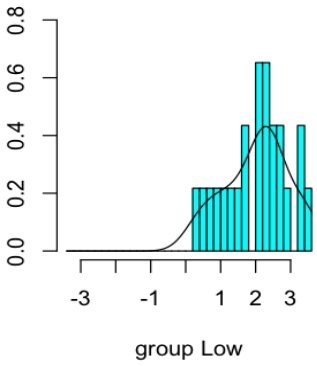
Confusion Matrix

```
          test.Healthcost
testclass High Low
     High   14   2
     Low     1   3
```

In Linear Discriminant Analysis, we set the threshold of the posterior to classify any probability higher than 0.5 as High and Low otherwise. The misclassification error for LDA is 0.15, which is lower than Logistic Regression. This means that the LDA model is able to classify the healthcare cost correctly 85% of the time. This is seen in the confusion as there only few false positives and false negatives.

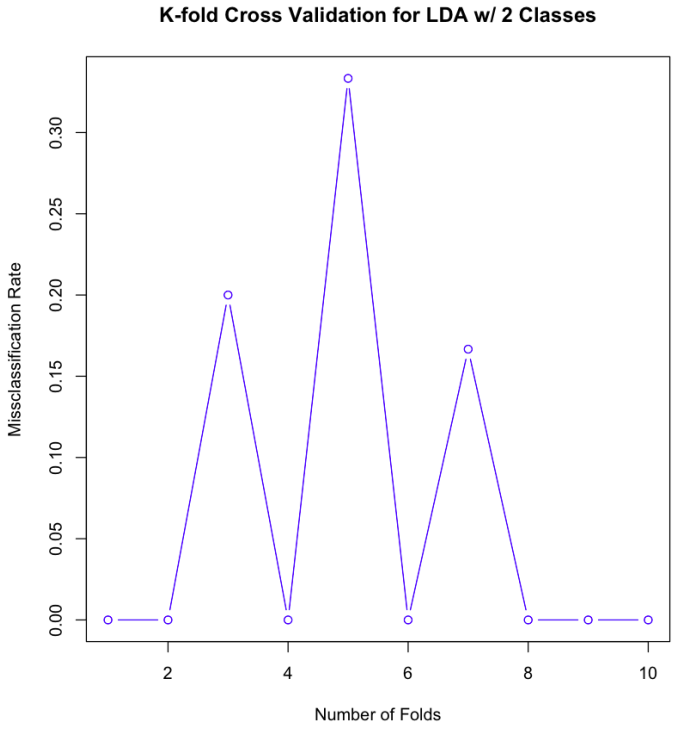
Plotting the Class for LDA





There is no overlap between high and low. This suggests that using LDA will yield a high prediction accuracy. From the confusion matrix, the LDA model is able to classify the healthcare costs 85% of the time.

K-Fold Cross Validation for LDA with 2 Classes



The above plot shows that if the data is folded in 10 different folds, seven out of ten times we get a 100% accuracy rate. This could be a sign of overfitting the model.

KNN Classifier

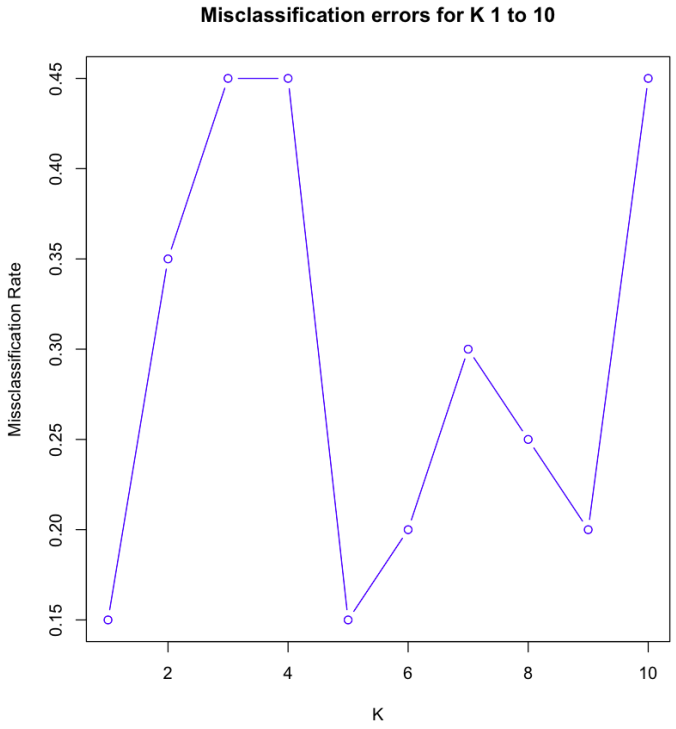
R code:

```
set.seed(1)
knn.pred=knn(Xlag[ClassTrainIndex,],Xlag[!ClassTrainIndex,],HighLow[ClassTrainIndex],k=5)
table(knn.pred,HighLow[!ClassTrainIndex])
mean(knn.pred!=HighLow[!ClassTrainIndex])#MisclassificationError
mean(knn.pred==HighLow[!ClassTrainIndex])#Model Accuracy
```

Confusion Matrix

```
knn.pred High Low
    High    9   0
    Low     6   5
```

Misclassification error for K 1:10



Contrary to the Logistic and LDA, when K = 4, the KNN classifier gives a misclassification error of 0.30. Interestingly, for Seed =1, there are no false positives and 6 false negatives. This is better than logistic regression but worse than LDA. For K=5, we are able to achieve a misclassification rate of .15, similar to LDA. However, there is a concern for overfitting as the lower the value of K, the lower the chances are of overfitting the model.

Overall, for 2 classes Linear discriminant analysis provides the best accuracy. We will model LDA for 3 classes to reduce the variance and try to increase the bias by only a little bit. This is to get better accuracy and avoid overfitting the model.

LDA with three Classes

Creating the class: Any value of healthcare costs greater than the 75th quartile will be labeled high in R. Any value below the 25th quartile, will be labeled as Low. Anything in between will be labeled as Moderate.

R code:

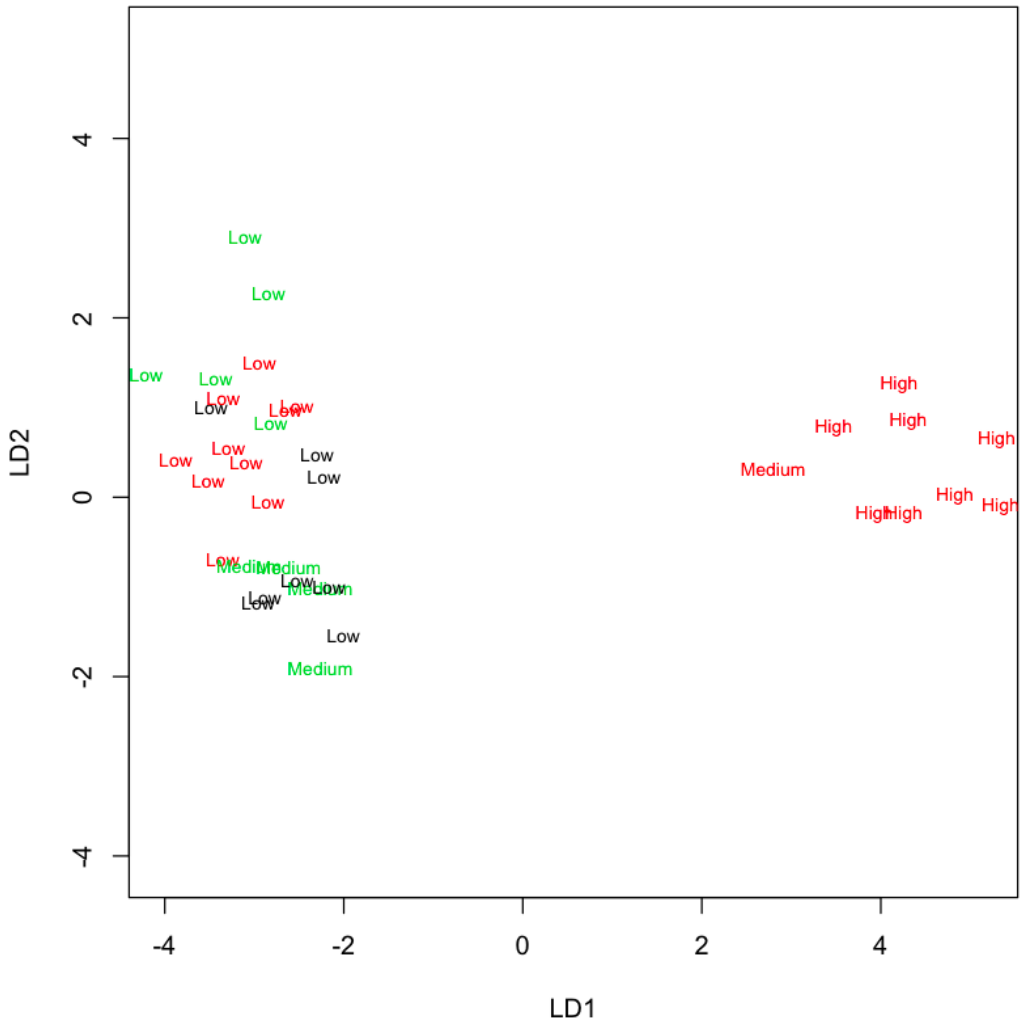
```
##LDA for 3 classes
lda.fit2 = lda(Health.cost~Year+Population+Per_Capita_Personal_Income+Real_Median_Hshd_Income,
               data = ht.df3, subset = train.x)
lda.fit2$counts
lda.fit2$prior
lda.pred2 = predict(lda.fit2, newdata= test.x)
test.Healthcost = ht.df3$Health.cost[!train.x]
table(lda.pred2$class, test.Healthcost)
mean(lda.pred2$class != test.Healthcost) #Misclassification Error
mean(lda.pred2$class == test.Healthcost) #Model Accuracy
```

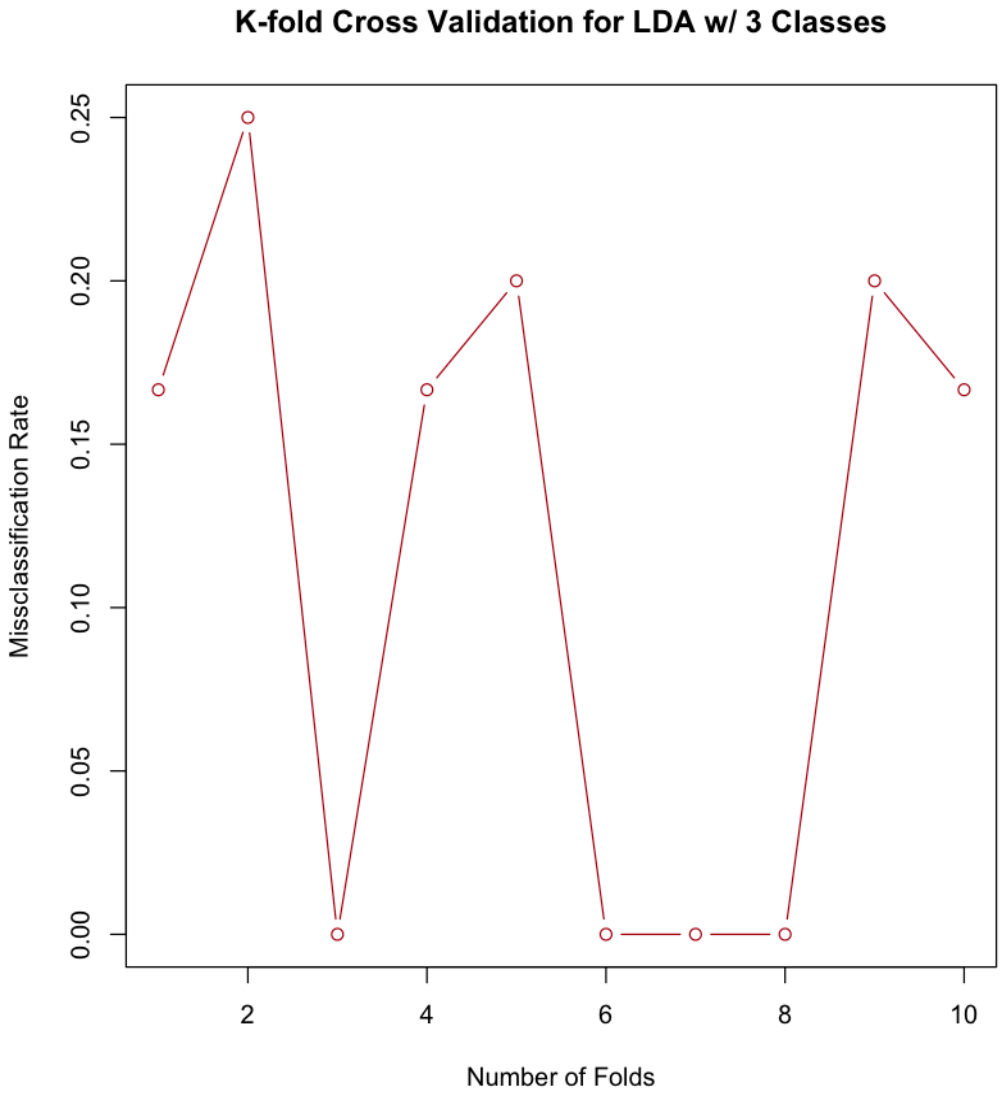
Confusion Matrix

```
        test.Healthcost
         High Low Medium
  High      5   0      0
  Low       0   4      0
  Medium    1   1      9
```

The LDA model with 3 classes successfully predicts with an accuracy of 90% and misclassification rate of 10%. We look at the chart to understand how that is achieved. The plot shows that there are no instances when High and Low overlaps. There are only a few instances when Low and Medium overlap with each other, resulting in high accuracy.

K-fold Cross Validation for LDA with 3 Classes

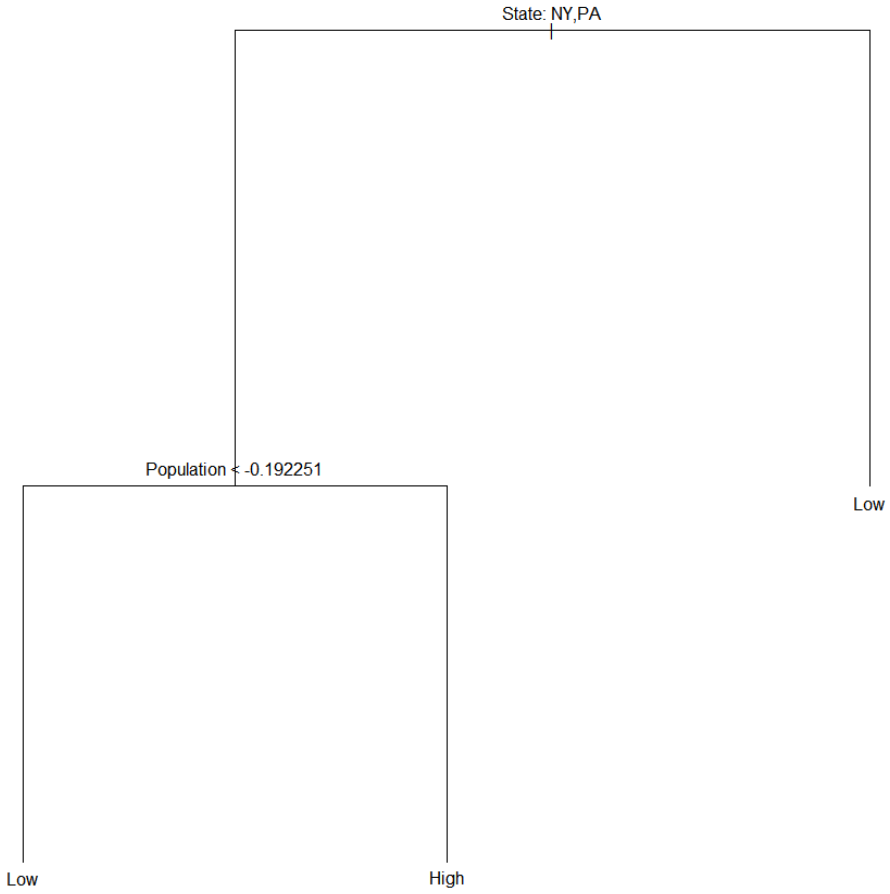




The plot shows that the model is able to predict 100% accurately in four out of Ten folds. This is much better when compared to LDA with 2 classes when it was classifying correctly seven folds out of ten. The chances of overfitting are lowered by a high margin when we divide and classify our data on 3 levels rather than 2.

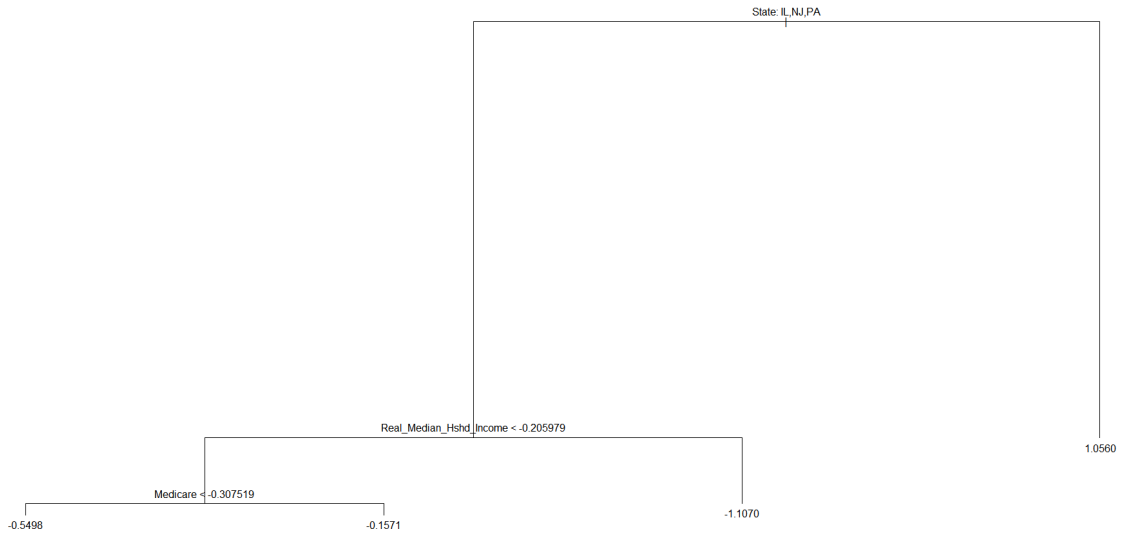
Decision Tree Analysis

Classification Tree



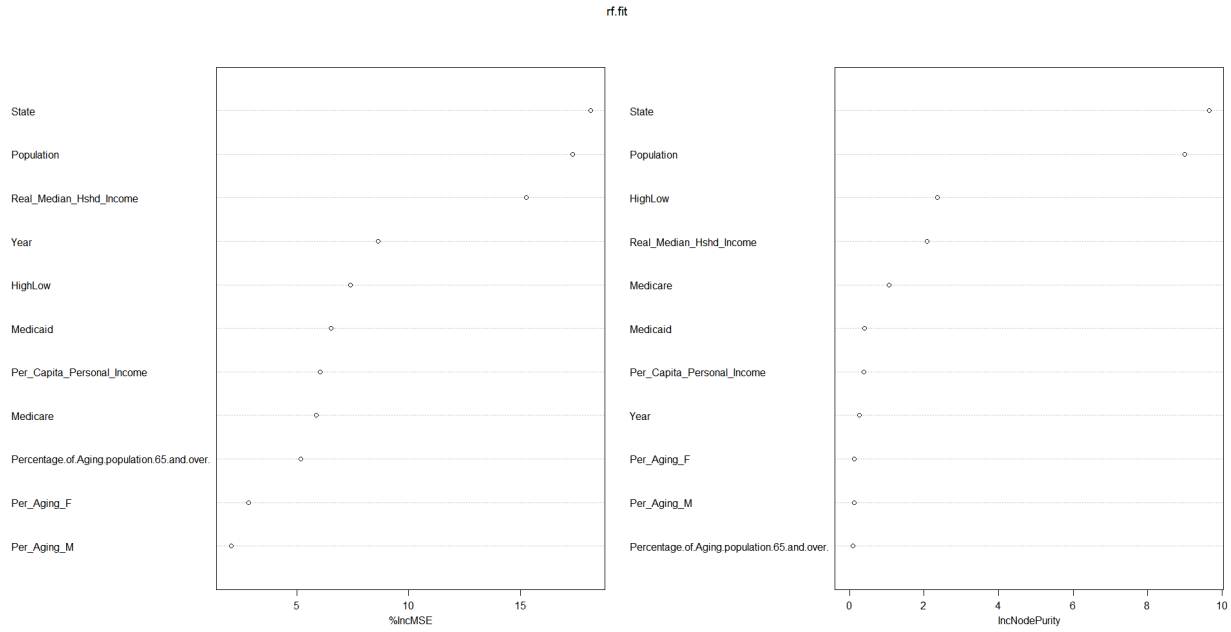
From the classification tree, State and Population appear to be the two variables for predicting whether the healthcare cost will by high or low. The misclassification error rate is 0.25 for both pruned and unpruned trees.

Regression Tree



Three variables are used in plotting the tree; namely State, Real Median Household Income and Medicare. The test root MSE is 1.063326 indicating that the reg tree model.

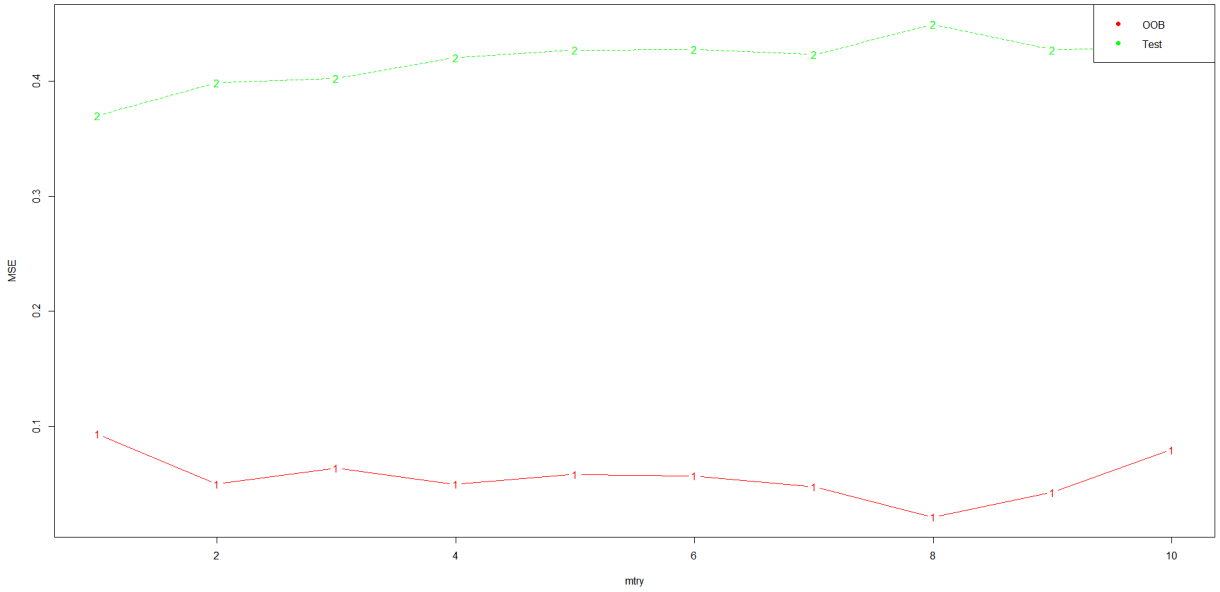
Bagging and Random Forest



For the bagging approach, all the predictors were used in fitting the tree. Just like the classification tree analysis, State and Population are the two most important variables in predicting healthcare costs. Test MSE associated with the unpruned bagging is 0.6312 relative to regression tree

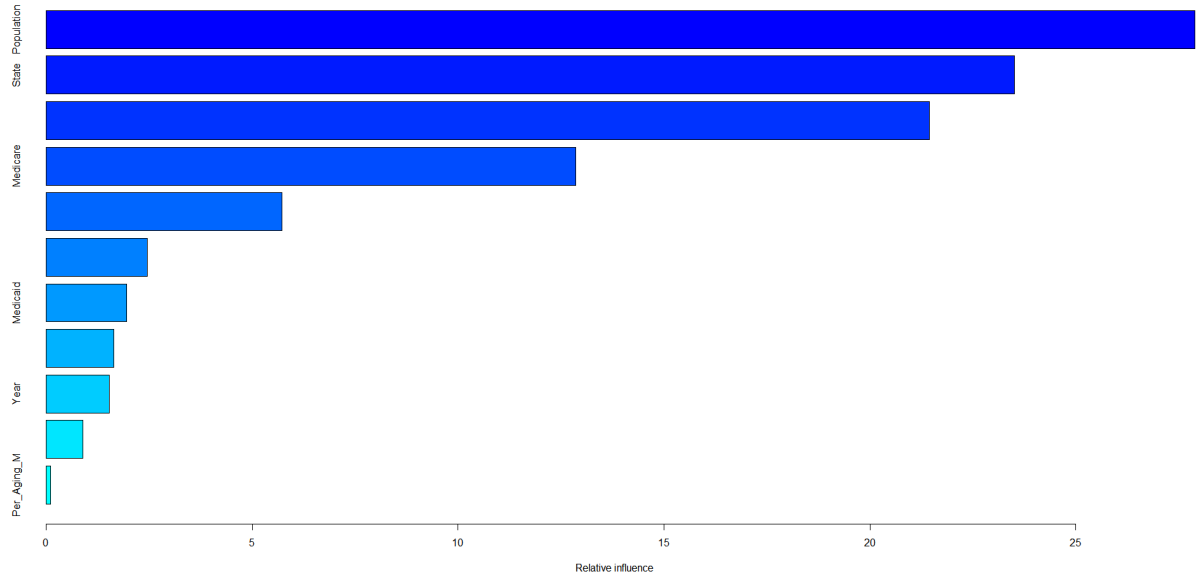
1.063326. However, pruning the tree results in error estimate of 0.6128 relative to unpruned tree of 0.6312 . With random forest, six predictors are selected for building the tree. The test MSE for random forest is 0.5498, lower than unpruned bagging but higher than pruned bagging.

Classification for Bagging



The test error is displayed as a function of the number of trees. The remaining one-third of the observations not used to fit the tree are referred to as the out-of-bag (OOB) observations and its error is lower relative to the MSE. The resulting OOB error is a valid estimate of the test error for the bagged model, since the response for each observation is predicted using only the trees that were not fit using that observation.

Boosting



Population and State are by far the two most important variables with test MSE error of 0.0111. Overall, boosting yields the smallest error relative to regression, bagging and random forest.

Classification for Boosting

The classification for boosting was done using the complete dataset and it yields an error rate of 0.006539531. The classification error rate for boosting yields a lower error rate relative to classification for boosting using OOB.

References

Federal Reserve Economic Data: FRED: St. Louis Fed. (n.d.). Retrieved from https://fred.stlouisfed.org/.

Home. (n.d.). Retrieved from https://www.spglobal.com/marketintelligence/en/.

U.S. Bureau of Labor Statistics. (2019, December 6). Retrieved from https://www.bls.gov/.