Final Project

IST 687 – Introduction to Data Science

ANALYSIS AND PREDICTION OF HEALTH CARE COSTLY CUSTOMERS

GROUP 2:
Aditi Pala
Indra Ariunbold
Rujuta Mirajkar
Jaineel Parmar
Rithvik Segu

CONTENTS

I. INTRODUCTION

This project's main objective is to offer actionable insight and precisely estimate which individuals (clients) will be expensive. The dataset we used for this project includes healthcare expenses from an HMO (Health Management Organization).

Let's start this approach in steps.

a. *Packages*

- tidyverse – collection of R packages
- caret- build machine learning models
- ggplot2- primarily used for data visualization
- fastDummies – create dummy variables
- ggmap- functions to visualize spatial data and models
- shiny – to make shiny apps

b. *Importing data*

Here we are reading this csv file via url to access the data and saving in a 'hmo' to keep the data in structured way.

```
```{r}

#install.packages("tidyverse")
library(tidyverse)
hmo <- read_csv("https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv")
head(hmo)
dim(hmo)
#view(hmo)
```
```

Here we are reading the data using read_csv function.

	X	age	bmi	children	smoker	location	location_type	education_level	yearly_physical	exercise	married	hypertensi
1	1	18	27.900	0	yes	CONNECTICUT	Urban	Bachelor	No	Active	Married	
2	2	19	33.770	1	no	RHODE ISLAND	Urban	Bachelor	No	Not-Active	Married	
3	3	27	33.000	3	no	MASSACHUSETTS	Urban	Master	No	Active	Married	
4	4	34	22.705	0	no	PENNSYLVANIA	Country	Master	No	Not-Active	Married	
5	5	32	28.880	0	no	PENNSYLVANIA	Country	PhD	No	Not-Active	Married	
6	7	47	33.440	1	no	PENNSYLVANIA	Urban	Bachelor	No	Not-Active	Married	
7	9	36	29.830	2	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	
8	10	59	25.840	0	no	PENNSYLVANIA	Country	Bachelor	No	Not-Active	Married	
9	11	24	26.220	0	no	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	
10	12	61	26.290	0	yes	CONNECTICUT	Urban	No College Degree	No	Active	Married	
11	13	22	34.400	0	no	MARYLAND	Urban	Bachelor	No	Not-Active	Married	
12	14	57	39.820	0	no	MARYLAND	Urban	Bachelor	Yes	Not-Active	Married	
13	15	26	42.130	0	yes	PENNSYLVANIA	Urban	Bachelor	No	Active	Married	
14	16	18	24.600	1	no	PENNSYLVANIA	Country	No College Degree	Yes	Not-Active	Not_Married	

This is how the data looks like when viewed.

It has 7582 instances recorded and 14 attributes for each instance.

c. *Description of data*

- X (Integer) is an integer which has unique number for each person.
- Age (Integer) is an integer which contains the age of the person (at the end of the year).
- Location (Categorical) represents data about the name of the state in the United States where the person lived.
- Location Type (Categorical) contains description of the environment where the person has lived (urban or country).
- Exercise (Categorical) consists of data on exercise activities of a person in two categories:
  Not-Active - when the person did not exercise regularly during the year.
  Active - when the person did exercise regularly during the year.
- Smoker (Categorical) consists of data of if the person is smoker or not based on 2 types:
  Yes - if the person smoked during the past year.
  No - if the person didn't smoke during the year.
- BMI (Integer) is an integer which consists of the body mass index of the person. The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy.
- yearly_physical (Categorical) gives information on if the person has regular visits with doctor throughout the year based on 2 categories:
  Yes - when the person had a well visit (yearly physical) with their doctor during the year.
  No - when the person did not have a well visit with their doctor.
- Hypertension gives data about if the person has hypertension or not:
  0 - when the person did not have hypertension.
  1 – when the person had hypertension.
- Gender (Categorical) gives the gender of the person.
- education_level (Categorical) consists of a data about the amount of college education each person has based on below categories:
  No College Degree – if the person has no college degree at all.
  Bachelor – if the person has bachelor's degree.
  Master – if the person has master's degree.
  PhD – if the person has PhD degree.
- Married (Categorical) describes marital status of the person:
  Married – if the person is married.
  Not_Married – if the person is not married.
- num_children (Integer) gives number of children.
- Cost (Integer) is an integer which gives the total cost of health care for that person during the past year.

d. *Details of the dataset*

The dataset has 7582 rows and 14 columns. We explore them with str function:

```{r}
str(hmo)
```

```
spec_tbl_df [7,582 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ X              : num [1:7582] 1 2 3 4 5 7 9 10 11 12 ...
 $ age            : num [1:7582] 18 19 27 34 32 47 36 59 24 61 ...
 $ bmi            : num [1:7582] 27.9 33.8 33 22.7 28.9 ...
 $ children       : num [1:7582] 0 1 3 0 0 1 2 0 0 0 ...
 $ smoker         : chr [1:7582] "yes" "no" "no" "no" ...
 $ location       : chr [1:7582] "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA" ...
 $ location_type  : chr [1:7582] "Urban" "Urban" "Urban" "Country" ...
 $ education_level: chr [1:7582] "Bachelor" "Bachelor" "Master" "Master" ...
 $ yearly_physical: chr [1:7582] "No" "No" "No" "No" ...
 $ exercise       : chr [1:7582] "Active" "Not-Active" "Active" "Not-Active" ...
 $ married        : chr [1:7582] "Married" "Married" "Married" "Married" ...
 $ hypertension   : num [1:7582] 0 0 0 1 0 0 0 1 0 0 ...
 $ gender         : chr [1:7582] "female" "male" "male" "male" ...
 $ cost           : num [1:7582] 1746 602 576 5562 836 ...
```

And summary of the numerical columns is provided below:

```
      age             bmi           children          cost      
 Min.   :18.00   Min.   :15.96   Min.   :0.000   Min.   :    2  
 1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000   1st Qu.:  970  
 Median :39.00   Median :30.50   Median :1.000   Median : 2500  
 Mean   :38.89   Mean   :30.80   Mean   :1.109   Mean   : 4043  
 3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.:2.000   3rd Qu.: 4775  
 Max.   :66.00   Max.   :53.13   Max.   :5.000   Max.   :55715  
                 NA's   :78                                     
```

e. Handling missing values

We noticed that some rows contained missing values for variables age and bmi. Therefore, we used na_interpolation function to fill those empty cells.

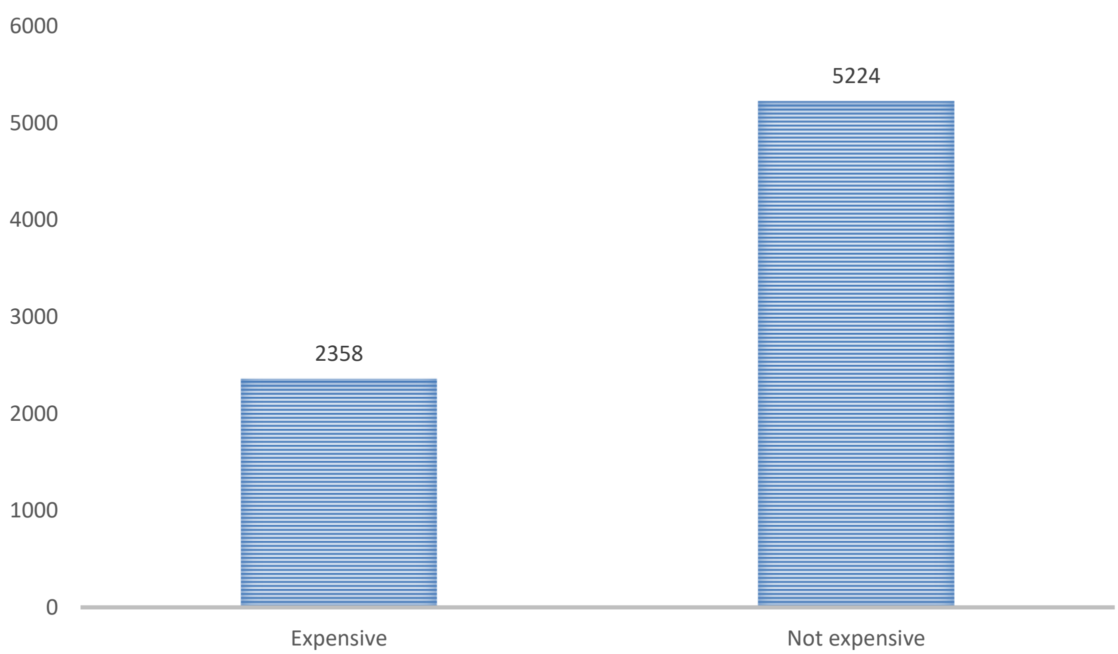
```
datafile$age <- na_interpolation(datafile$age)
datafile$bmi <- na_interpolation(datafile$bmi)
```

II. DETERMINING THRESHOLD FOR EXPENSIVE VARIABLE

We established a criterion based on the person's location and location type in order to continue with the segregation of whether healthcare costs are expensive or not for a certain person. Average of cost as per both location and location type will be kept as threshold.

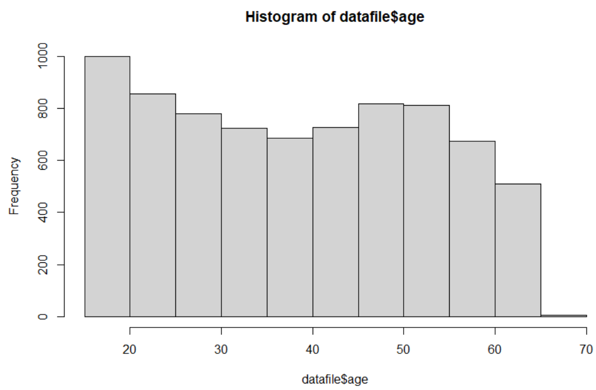
We added a new column called "avgcost" that will show the person's average cost based on their location and location type. Based on their location and location type, each person's average healthcare as per their location cost may vary. This average cost then considered as threshold deciding factor. Then, we built a new data frame that includes the newly calculated avgcost column together with all the columns from our original database.

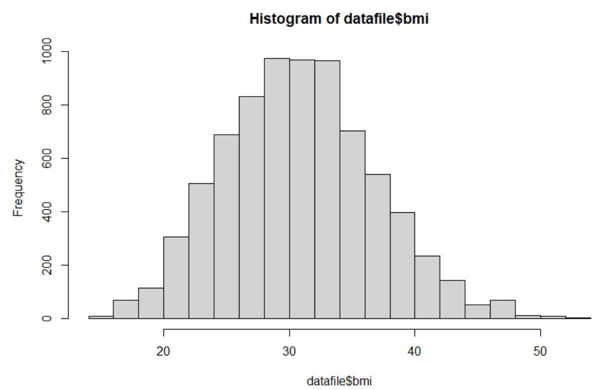
The calculated average costs for each location and location type will then be compared to the cost of healthcare person is paying. If a person's healthcare cost exceeds the computed average cost for their respective location, then their healthcare is expensive; otherwise, it is not expensive. Below is the result of our separation.



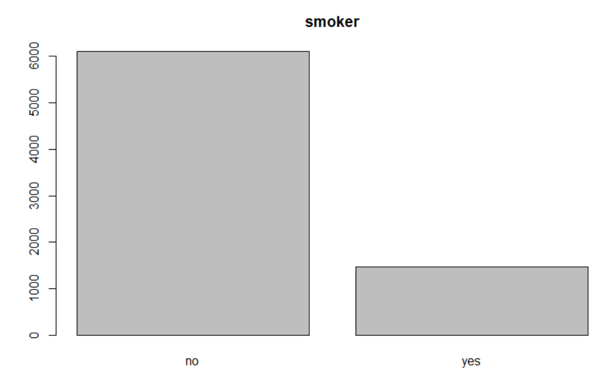
III. EXPLORATORY DATA ANALYSIS

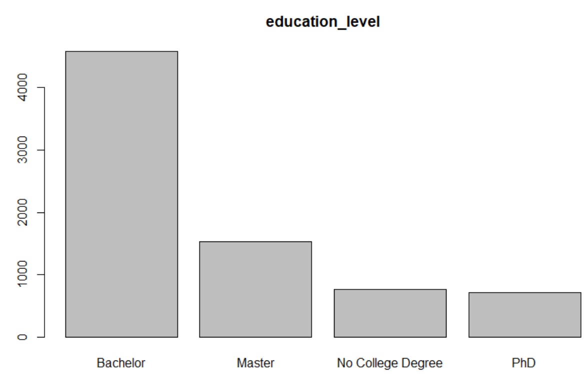
This is section we have performed graphical analysis of each columns and their combination. First of all, we should take a look at histograms of numerical values.

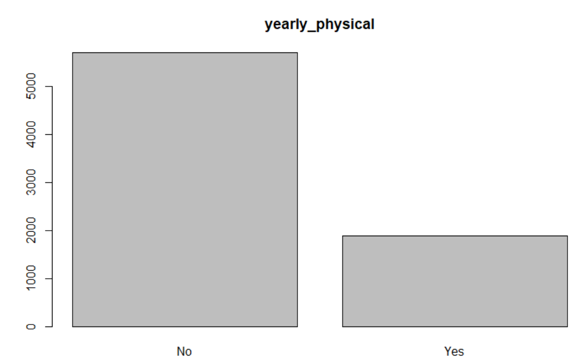


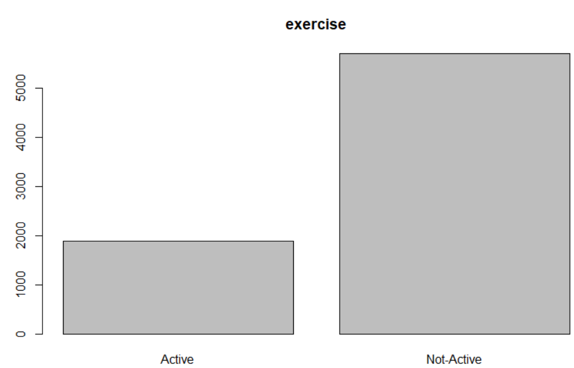


From these histograms we see that dataset contains people with normally distributed body mass index and uniform distributed age. This means that most of the people have average bmi, but different ages.



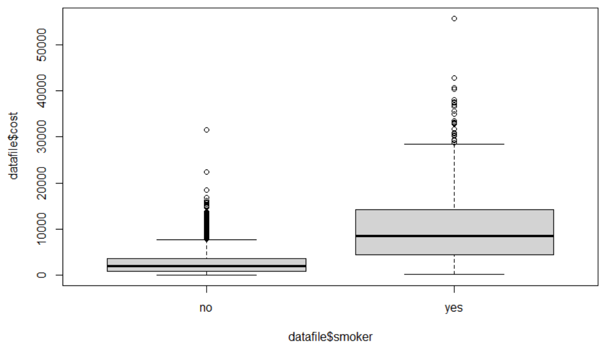


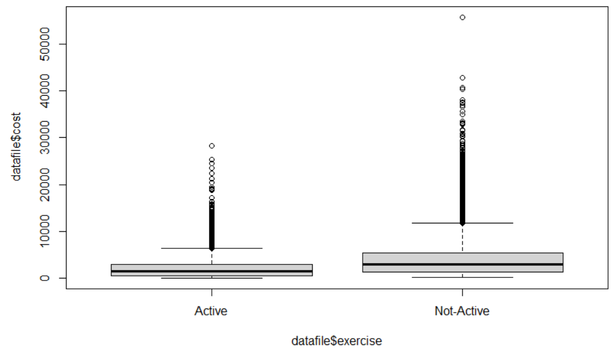




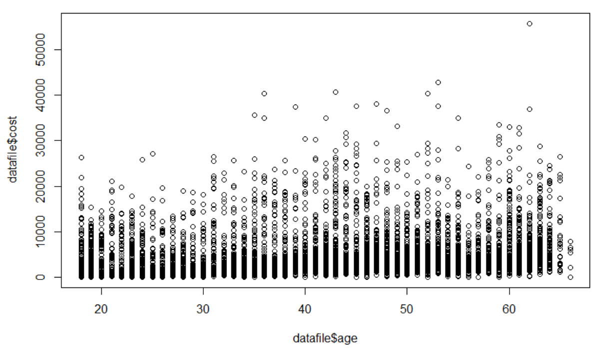
From the barplots we see that most of the people do not smoke, do not exercise, do not get yearly check-ups, and have Bachelor's degree.

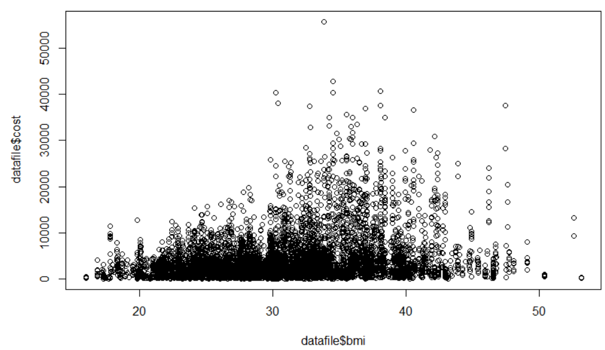
We also explored connection between categorical variables and cost:





We drew boxplots for all categorical variables and only effect we have on cost is whether that person is a smoker or not and whether he/she exercises. It seems to be these are very important variables in predicting the healthcare cost of a person.

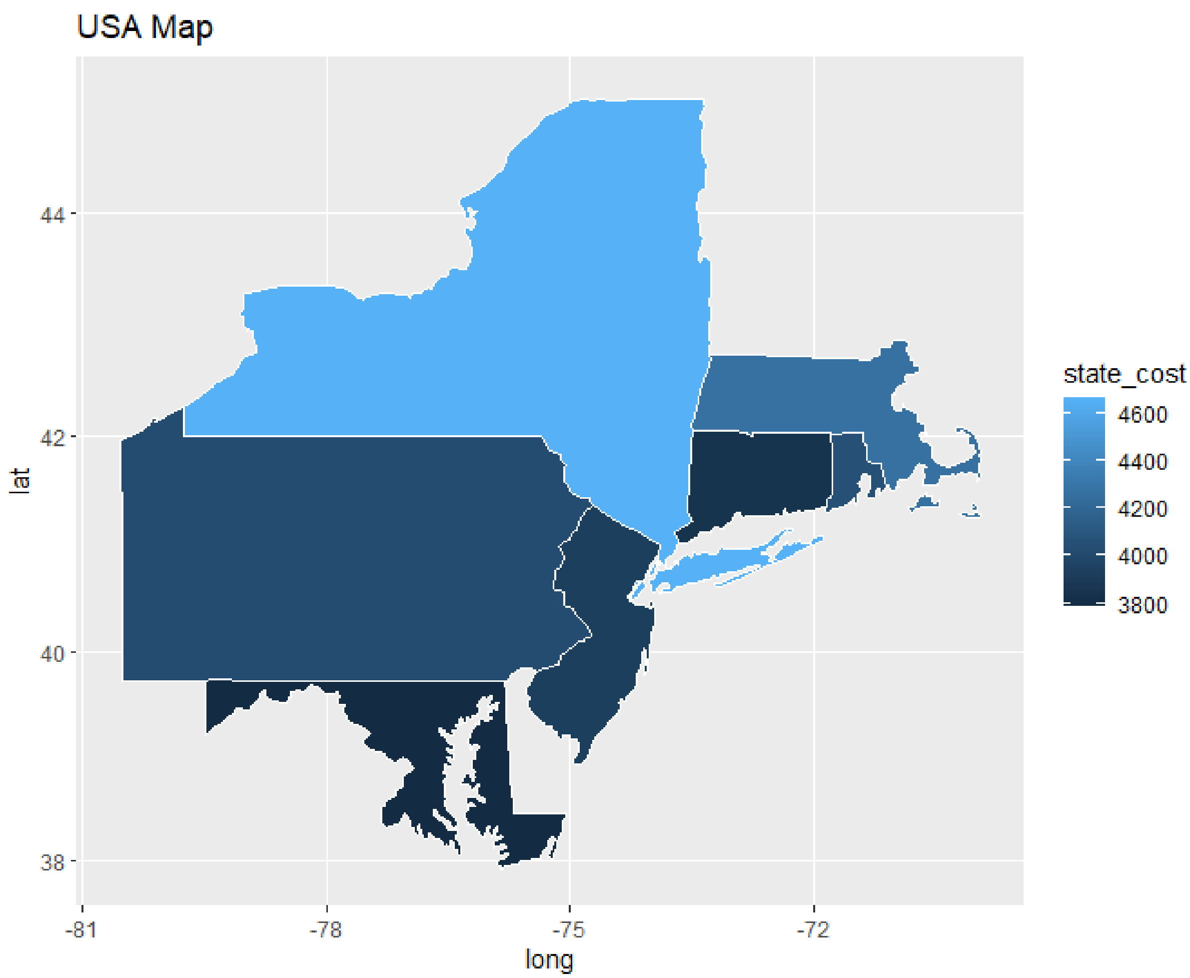




Exploring the connection between cost and numerical variables, show that there is a slight tendency for older people to spend more on health. However, for the bmi variable we have normally distributed people, therefore the cost is normally distributed as well.

IV. GEOGRAPHICAL ANALYSIS

Since we have location information for every person, we can illustration on maps.



For the average cost per person, we see that New York state has the highest average healthcare cost person and Maryland has the lowest.

V. BUILDING PREDICTIVE MODEL

In order to build binary classification model based on the variables that were given, we need to turn categorical variables into dummy variables. We do that with help of R package called fastDummies. After that we choose variables and split data into Train /70%/ and Test /30%/ subsets.

Next, we choose to train four models.

f. K-nearest neighbors

```
model_knn <- train(X_train, y_train, method='knn', tuneLength = 10,
                   trControl = trainControl(method = "cv"))
saveRDS(model_knn, "model_knn.rds")
```

```
> confusionMatrix(predictions,y_test)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1432  222
         1  125  495

               Accuracy : 0.8474
                 95% CI : (0.832, 0.862)
    No Information Rate : 0.6847
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6332

 Mcnemar's Test P-Value : 2.556e-07

            Sensitivity : 0.9197
            Specificity : 0.6904
         Pos Pred Value : 0.8658
         Neg Pred Value : 0.7984
             Prevalence : 0.6847
         Detection Rate : 0.6297
   Detection Prevalence : 0.7274
      Balanced Accuracy : 0.8050

       'Positive' Class : 0
```

g. Support vector machines

```
model_svm <- train(X_train, y_train, method='svmLinear',preProcess=c("center", "scale"))
saveRDS(model_svm, "model_svm.rds")
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1454  272
         1  103  445

               Accuracy : 0.8351
                 95% CI : (0.8192, 0.8501)
    No Information Rate : 0.6847
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5921

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9338
            Specificity : 0.6206
         Pos Pred Value : 0.8424
         Neg Pred Value : 0.8120
             Prevalence : 0.6847
         Detection Rate : 0.6394
   Detection Prevalence : 0.7590
      Balanced Accuracy : 0.7772

       'Positive' Class : 0
```

h. Logistic Regression

```
model_lb <- train(X_train, y_train, method='LogitBoost',preProcess=c("center", "scale"))
saveRDS(model_lb, "model_lb.rds")
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1336  188
         1  221  529

               Accuracy : 0.8201
                 95% CI : (0.8037, 0.8357)
    No Information Rate : 0.6847
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5885

 Mcnemar's Test P-Value : 0.1136

            Sensitivity : 0.8581
            Specificity : 0.7378
         Pos Pred Value : 0.8766
         Neg Pred Value : 0.7053
             Prevalence : 0.6847
         Detection Rate : 0.5875
   Detection Prevalence : 0.6702
      Balanced Accuracy : 0.7979

       'Positive' Class : 0
```

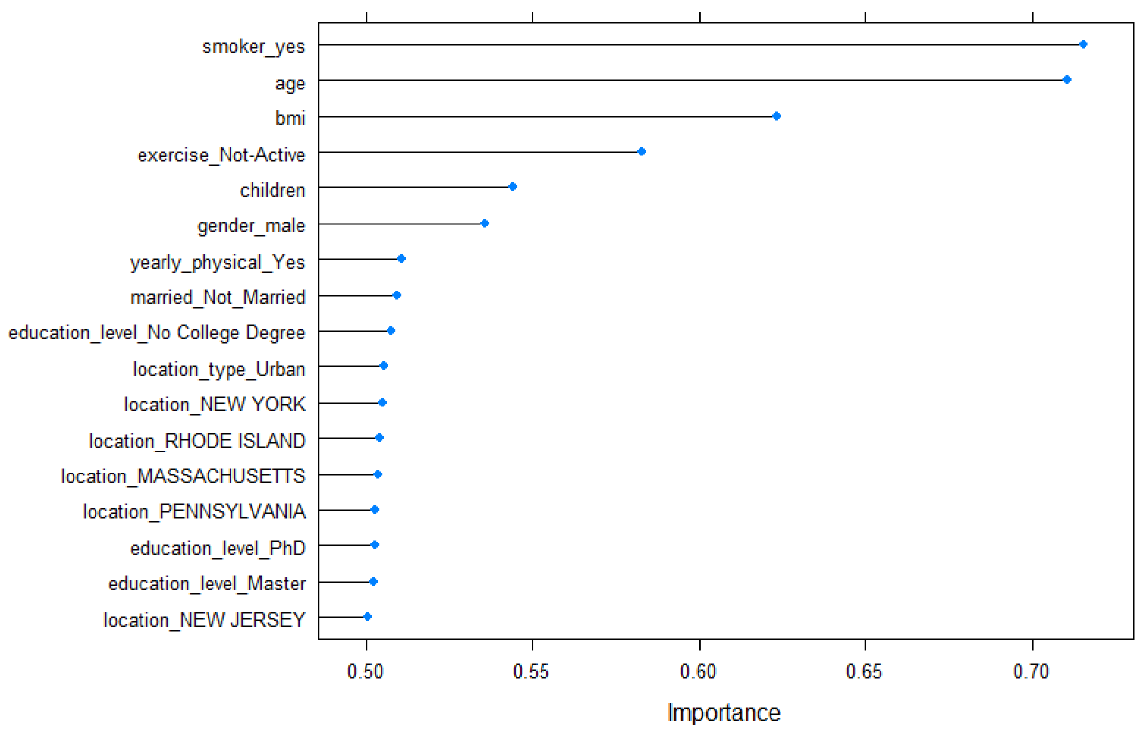
i. Neural Network

```
model_nnet <- train(X_train, y_train, method='nnet',tuneLength = 2,
                     trace = FALSE,
                     maxit = 100)
saveRDS(model_nnet, "model_nnet.rds")
```

```
> confusionMatrix(predictions,y_test)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1438  243
         1  119  474

               Accuracy : 0.8408
                 95% CI : (0.8251, 0.8556)
    No Information Rate : 0.6847
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6133

 Mcnemar's Test P-Value : 1.015e-10

            Sensitivity : 0.9236
            Specificity : 0.6611
         Pos Pred Value : 0.8554
         Neg Pred Value : 0.7993
             Prevalence : 0.6847
         Detection Rate : 0.6324
   Detection Prevalence : 0.7392
      Balanced Accuracy : 0.7923

       'Positive' Class : 0
```

j. Model accuracy results

After training models, we have our feature importance plot:



And model summary table is as follows:

Model	Accuracy	Sensitivity	Specificity
K-nearest neighbors	84.74%	91.97%	69.04%

SVM	83.51%	93.38%	62.06%
Logistic Regression	82.01%	85.91%	73.78%
Neural Network	84.08%	92.36%	66.11%

So, in order to test the models, we predicted values from the sample data /20 rows of new data/. And as a result, we chose Neural Network model to be best one to make predictions and used it in out Shiny app.

VI. CONCLUSION

From both exploratory analysis and predictive analysis, we see that the most important variables in deciding whether a person will have a high health care cost is whether they smoke, exercise, their age and bmi. Therefore, to make recommendations we need to focus on those areas.

We have following recommendations:

- Since smoking is the most important variable, we need to identify smokers and help them quit smoking in various ways, such as by recommending programs, patches and other solutions.
- The likelihood of people with high bmi and older age to go for an exercise is usually low, therefore we need to identify people who could benefit the most from daily exercising and recommend possible at-home routines, or fitness facilities close to their home.
- Yearly check-ups are also important in detecting health problems early on and reduce high costs in the future. Thus, we need to remind especially at-risk people to get check ups done.